

# Menzerath's law for the smallest grammars

*Łukasz Dębowski*

## 1 Introduction

The aim of this article is to develop a discussion of Menzerath's law from the point of view of information theory. More precisely, we shall seek for links between the law and the recently abstracted mathematical problem of the smallest grammar (Kieffer & Yang 2000, Charikar et al. 2005).

The Altmann-Menzerath law is a general statement about the natural language constructions which says:

$$\text{The longer a construction, the shorter are its constituents.} \quad (1)$$

For example, the average number of phonemes in a syllable of a word having  $n$  syllables decreases with  $n$  (Menzerath 1928, Altmann 1980). Analogical regularity was observed for a number of other linguistic entities.

Similarly to Zipf's law, law (1) is not restricted to human language. It has also been observed in DNA and in the social behavior of animals (Altmann & Schwibbe 1989). Such omnipresence needs explanation. Possibly, Zipf's and Menzerath's laws may be given many good explanations depending on the context where they are found. Three simple models by Mandelbrot (1953), Simon (1955), and Miller (1957) inspired multiple explanations of Zipf's law. Less work, however, was done on conceiving analogical models for Menzerath's law (Köhler 1989). Despite some similarity of wording, law (1) has nothing to do with regression towards the mean (Galton 1886).

A new idea for explaining Zipf's and Menzerath's laws comes with the concept of the smallest grammar for a text (Dębowski 2004d). Consider context-free grammars that generate exactly one string, such as

$$G = \left\{ \begin{array}{l} A_0 \mapsto A_2 A_1 A_3, \text{\_dear\_Gabriel} A_1 \\ A_1 \mapsto A_4 A_2 A_4 \\ A_2 \mapsto A_3 \text{\_to\_you} \\ A_3 \mapsto \text{Happy\_birthday} \\ A_4 \mapsto !\_ \end{array} \right\}, \quad (2)$$

where  $A_0$  is the start symbol. Grammars generating only one string will be called *admissible*, after Kieffer & Yang (2000).

A given string can be generated by many distinct admissible grammars. We can try discriminating between them on the basis of their length. For string  $g_i$  consisting of terminal and nonterminal symbols, let length  $|g_i|$  be the number of symbols, e.g.  $|A_3\_to\_you| = 8$ . Kieffer & Yang (2000) defined the length of a grammar as the total length of its right-hand sides,

$$|G| := \sum_{(A_i \mapsto g_i) \in G} |g_i|. \quad (3)$$

For example, we have  $|G| = 45$  for grammar  $G$  in (2). An admissible grammar is called the smallest grammar for a string if it generates the string and has the minimal length. For any string, an example of the smallest grammar can be computed in the exponential time. Nevertheless, there exist fast algorithms which find some local minima of  $|G|$  or local minima of the length of grammar's binary encoding (Charikar et al. 2005). These algorithms return grammars of diverse properties, which we shall call together vaguely "locally smallest grammars".

Long before Kieffer & Yang (2000), Wolff (1980), Nevill & Manning (1996), and DeMarcken (1996) computed some locally smallest grammars for corpora of texts in English and Chinese. They observed that many nonterminals of the grammars can be interpreted as the occurrences of syllables, morphemes, words, and fixed phrases. Although these grammars are not sufficient for computational linguists' needs, we think that they could inspire fruitful collaboration between quantitative linguists and information theorists.

In parallel to researching quantitative laws for deeply rethought formal linguistic entities, we propose investigating the analogical laws for nonterminals of locally smallest grammars. There are two reasons for this:

1. Locally smallest grammars can be computed for any string. One can compare structural complexity of very different strings: human texts, music, DNA, deemed extraterrestrial messages, expansions of  $\pi$ ,  $e$ , etc.
2. Some properties of locally smallest grammars can be analyzed mathematically. Linguists can interest specialists in probability calculus and theoretical computer science and possibly get some interpretations.

In the following two sections, we contribute to the sketched research program. In section 2, we define an analog of Menzerath's law for admissible

grammars. In section 3, we check if the analog holds for two texts in English and Polish and two random strings of the same length.

## 2 Analog of Menzerath's law

Consider text  $T$  which is a concatenation of words  $w_i$ ,  $T = w_1w_2\dots w_{N(T)}$ . Each word type  $W_i$  is a concatenation of syllables  $s_{ij}$ ,  $W_i = s_{i1}s_{i2}\dots s_{iN(W_i)}$ , and each syllable type  $S_j$  is a string of phonemes  $a_{jk}$ ,  $S_j = a_{j1}a_{j2}\dots a_{jN(S_j)}$ .

Treat the symbols for phonemes as terminals and other symbols as non-terminals. Then we can construct an admissible grammar for the text, in the form of

$$G^{\text{ideal}} = \left\{ \begin{array}{l} A_0 \mapsto w_1w_2\dots w_{N(T)} \\ W_i \mapsto s_{i1}s_{i2}\dots s_{iN(W_i)}, \quad i = 1, 2, \dots, V_W \\ S_j \mapsto a_{j1}a_{j2}\dots a_{jN(S_j)}, \quad j = 1, 2, \dots, V_S \end{array} \right\}, \quad (4)$$

where  $V_W$  is the number of word types and  $V_S$  is the number of syllable types. The average number of phonemes in a syllable of an  $n$ -syllable word is

$$C(n) = \frac{1}{n} \cdot \frac{\sum_{\substack{i: W_i \text{ has } n \text{ direct constituents} \\ i: W_i \text{ has } n \text{ direct constituents}}} \sum_{j=1}^n N(s_{ij})}{\sum_{\substack{i: W_i \text{ has } n \text{ direct constituents}}} 1}. \quad (5)$$

Menzerath's law holds for the text if  $C(n)$  decreases with  $n$ . Now, let  $G$  be an arbitrary admissible grammar. We can write it in the form of

$$G = \left\{ \begin{array}{l} A_0 \mapsto w_1w_2\dots w_{N(T)} \\ W_i \mapsto s_{i1}s_{i2}\dots s_{iN(W_i)}, \quad i = 1, 2, \dots, V_W \\ S_j \mapsto a_{j1}a_{j2}\dots a_{jN(S_j)}, \quad j = 1, 2, \dots, V_S \\ \dots \end{array} \right\}, \quad (6)$$

where  $W_i$  are nonterminals appearing at least once in string  $w_1w_2\dots w_{N(T)}$ , and  $S_j$  are nonterminals appearing at least once in some string  $s_{i1}s_{i2}\dots s_{iN(W_i)}$ . Strings  $w_1w_2\dots w_{N(T)}$ ,  $s_{i1}s_{i2}\dots s_{iN(W_i)}$ , and  $a_{j1}a_{j2}\dots a_{jN(S_j)}$  may consist of both terminals and nonterminals while the final three dots in (6) stand for the possible remaining rules in the grammar.

We will say that Menzerath's law holds for grammar  $G$  if certain function  $C(n)$  constructed for  $G$  decreases with  $n$ . We require that  $C(n)$  satisfy (5) for

$G = G^{\text{ideal}}$  and we use the following extension of  $C(n)$  for other grammars: For  $G$  like in (6), we define  $C(n)$  by formula (5) where  $N(s_{ij}) := 1$  for any terminal symbol  $s_{ij}$ . (Other extensions may be worth considering.)

Menzerath's law holds for some grammars representing human texts, such as  $G^{\text{ideal}}$ . According to experiments by Wolff, Nevill & Manning, and DeMarcken, many rules of  $G^{\text{ideal}}$  are identical to the rules of certain locally smallest grammars for the same text. So we can ask interdisciplinary questions:

*Let algorithm  $\mathcal{A}$  compute a locally smallest grammar  $\mathcal{A}(T)$  for string  $T$ , i.e., some grammar which generates  $T$ .*

1. *(to empirical researchers)* Let  $G$  be the output of  $\mathcal{A}$  for a generic human text. Does Menzerath's law hold for  $G$  just as it does for  $G^{\text{ideal}}$ ? Are functions  $C(n)$  for  $G$  and  $G^{\text{ideal}}$  similar?
2. *(to theoretical computer scientists)* Is Menzerath's law for grammars  $\mathcal{A}(T)$  a tautology, i.e., does it hold for all input strings  $T$ ? *(implicitly: Is there a trivial explanation for Menzerath's law in human language?)*
3. *(to probabilists)* Does Menzerath's law hold for all grammars  $\mathcal{A}(T)$ , where strings  $T$  are the typical outcomes of a fixed kind of a stochastic process? *(implicitly: Is Menzerath's law evidence for or against a particular stochastic model of language production?)*

The answers may depend on algorithm  $\mathcal{A}$  and extension  $C(n)$ . We cannot refute the possibility that given some algorithms  $\mathcal{A}$  and extensions  $C(n)$ , Menzerath's law for  $\mathcal{A}(T)$  is a tautology. Such a case would not be singular. Recently, we have recognized a tautology in the linguistically important inequality dealt by Dębowski (2005). The inequality reads

$$|G| - V_v \leq (V_v + V_\tau)^2, \quad (7)$$

where  $G$  is an *irreducible* grammar – see Kieffer & Yang (2000), section 3.2, for the definition –,  $V_v$  is the number of *all* nonterminal types in  $G$ , and  $V_\tau$  is the number of terminal types. Each irreducible grammar satisfies (7) since any concatenation of two symbols can occur in its right-hand sides only once.

There exists an irreducible smallest grammar for every string. What happens if the string produced by the grammar is some human text counting  $N$  characters (terminals)? Assuming that language production has constant entropy rate  $h > 0$  (Shannon 1950), we obtain  $|G| \geq hN / \log N$  from inequalities  $|G| \log |G| \geq hN$  and  $|G| \leq N$  (Dębowski 2005). On the other hand, (7)

implies  $V_v + V_\tau \geq -1 + |G|^{1/2}$ . Combining the inequalities yields

$$V_v + V_\tau + 1 \geq [hN/\log N]^{1/2}. \quad (8)$$

Guided by Wolff and his successors, we could speculate that  $V_v$  or  $V_v + V_\tau$  is proportional to the number of distinct words in the text. Then, we would recognize in (8) the well known power-law inequality for vocabulary growth (Kuraszkiewicz & Łukaszewicz 1951, Guiraud 1954, Herdan 1964).

Dębowski (2005) derived a power-law inequality for  $V_v$  as an effect of a similar hypothetical power-law inequality for excess entropy of human narration (Hilberg 1990). Such inequality for excess entropy would be a symptom of long memory in the narration. Nevertheless, (8) is true even for an algorithmically random ( $\sim$  „structureless”) string  $T$  with entropy rate  $h > 0$ . Even for such a string, the smallest irreducible grammar is very different from trivial grammar  $G = \{A_0 \mapsto T\}$ . We hope that this paradox will inspire the theory of grammar-based coding, initiated by Kieffer & Yang (2000).

### 3 Experiments

Let us present a short survey of several statistics for locally smallest grammars. We considered four input texts of similar length: *Gulliver's Travels* by Jonathan Swift (in English), *W pustyni i w puszczy* by Henryk Sienkiewicz (in Polish), and the unigram model versions of both novels (i.e. the strings of characters tossed according to their frequencies in the novels). The texts were turned into lower case. Spaces were preserved but punctuation deleted.

For each text we computed two grammars: a longest matching grammar (LMG) defined by Kieffer & Yang (2000) and a biased LMG (BLMG), which is our modification. LMG is built from the trivial grammar  $G = \{A_0 \mapsto T\}$  by iterative introduction of rules  $A_n \mapsto \gamma$ , where  $\gamma$  is the longest substring with  $|\gamma| > 1$  appearing at least twice in grammar's right-hand sides. The rules are added until there is no such string. LMG is irreducible and obeys (7)–(8).

In contrast, BLMG is built from the trivial grammar by iterative introduction of rules  $A_n \mapsto \gamma$ , where  $\gamma$  is the longest substring with  $|\gamma| > \lg n + 2$  appearing  $> (|\gamma| + 1)/(|\gamma| - \lg n - 2)$  times. The final grammar  $G$  minimizes locally the length of a naive binary representation of  $G$  rather than  $|G|$  itself.

In Table 1, we resume basic statistics of the texts and grammars. Parameters of (B)LMG clearly distinguish the original novels from the unigram

Table 1: Statistics of the texts and their grammars

	plain text					LMG			BLMG		
	$N_\tau$	$V_\tau$	$N_W$	$V_W$	$L^{>1}$	$ G $	$V_V$	$P$	$ G $	$V_V$	$P$
<i>Gulliver's Travels</i>											
original	561k	30	105k	8k	54	147k	32k	0.86	326k	11k	0.550
unigram	561k	30	105k	51k	11	272k	54k	0.71	558k	135	0.005
<i>W pustyni i w puszczy</i>											
original	616k	39	101k	18k	69	176k	38k	0.84	394k	10k	0.470
unigram	616k	39	100k	56k	9	316k	61k	0.69	615k	99	0.003

**Key:** 1k = 1000;  $N_\tau$  ( $V_\tau$ ) = number of character tokens (types);  $N_W$  ( $V_W$ ) = number of word tokens (types) meant as space-to-space strings;  $L^{>1}$  = length of the longest repeated string;  $|G|$  = length of the grammar;  $V_V$  = number of nonterminal types;  $P$  = parsing rate, i.e.,  $P = 1 - N_{0\tau}/N_\tau$ , where  $N_{0\tau}$  = number of terminal tokens in the start rule of the grammar.

model texts. Parsing rate  $P$  and nonterminal vocabulary  $V_V$  of BLMG are about 100 times smaller for the random texts than for the nonrandom ones. In contrast, LMG nonterminal vocabulary for unigram texts is almost twice as big as for the novels. This is not puzzling in view of (8) since entropy rate for the latter texts is less.

In Figures 1 and 2, we present the graphs of functions supposed to satisfy Menzerath's and Zipf's laws in the ideal case. In the upper plots,  $C_{(B)LMG}(n)$  are computed for (B)LMG as defined in the previous section. The baseline is

$$C_{\text{ideal}}(n) = \frac{1}{n} \cdot \frac{\sum_{i: \gamma_i \text{ has } n \text{ disjoint vowel clusters}} |\gamma_i|}{\sum_{i: \gamma_i \text{ has } n \text{ disjoint vowel clusters}} 1}, \quad (9)$$

where  $\gamma_i$  are consecutive space-to-space strings in the input text. Vowel clusters are defined operationally as clusters of letters  $\text{ieaouyęąó}$ .

The lower plots in Figures 1 and 2 depict rank-frequency distributions. Value  $f_{\text{ideal}}(r)$  is the frequency of  $r$ -th ranked space-to-space string appearing in the input text. Value  $f_{(B)LMG}(r)$  is the frequency of  $r$ -th ranked nonterminal appearing in string  $\gamma$ , where  $A_0 \mapsto \gamma$  is the start rule of (B)LMG.

The respective plots for *Gulliver's Travels* and *W pustyni i w puszczy* are similar. There is, however, a huge difference between the plots for LMG and BLMG.  $C_{\text{LMG}}(n)$  for the original novels does not decrease. Other functions  $C_{\dots}(n)$  decrease for  $n < 8$  – they obey Menzerath's law in that range.

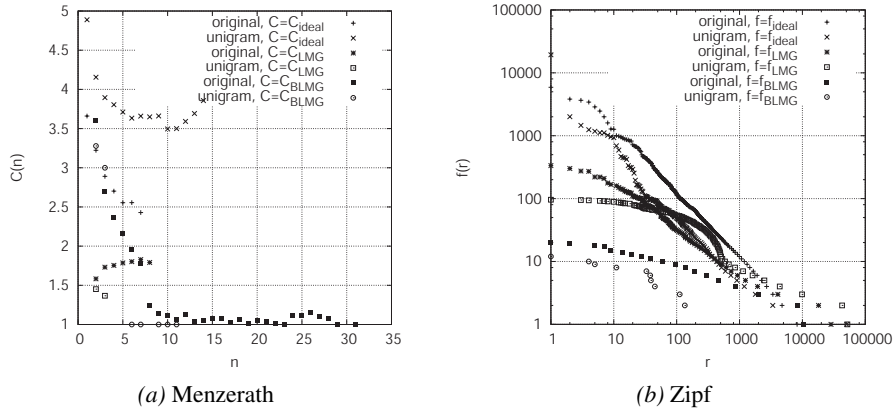


Figure 1: Menzerath's and Zipf's laws for *Gulliver's Travels*

As for the rank-frequency distributions, the tail of  $f_{ideal}(r)$  exhibits Zipf-Mandelbrot power-law. In this case, random texts do not differ from nonrandom ones, as noticed by Miller (1957). On the other hand, functions  $f_{LMG}(r)$  and  $f_{BLMG}(r)$  do not exhibit the power-law in the tail. The plot of  $f_{LMG}(r)$  in log-log scale is close to a straight line in the middle range for the original novels but it consists of two large humps for the unigram texts. One lesson from the presented data is that even simple statistics of two locally smallest gram-

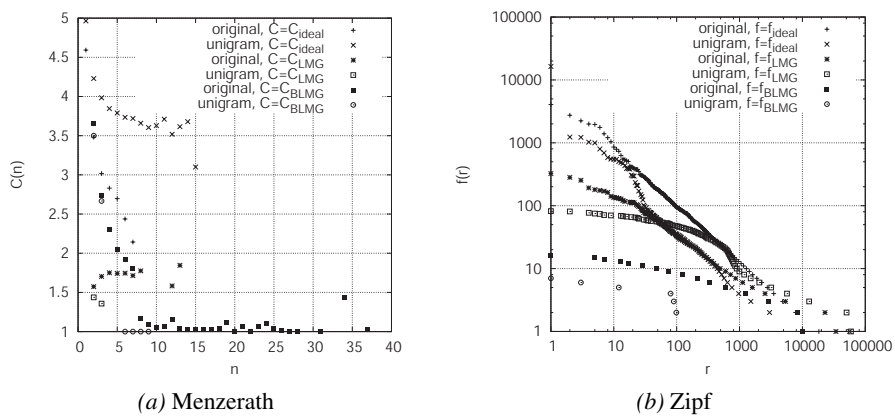


Figure 2: Menzerath's and Zipf's laws for *W pustyni i w puszczy*

grams for the same text can be very different. On the other hand, variation of the statistics across different texts is much smaller. It would be good to check the same statistics for more texts and more kinds of admissible grammars.

## References

- Altmann, Gabriel  
 1980 “Prolegomena to Menzerath’s law”. In: *Glottometrika 2*. Bochum: Brockmeyer, 1–10.
- Altmann, Gabriel; Schwibbe, Michael H.  
 1989 *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Hildesheim: Olms.
- Dębowski, Łukasz  
 2005 “On Hilberg’s law and its links with Guiraud’s law”. Preprint.
- Galton, Francis  
 1886 “Regression Towards Mediocrity in Hereditary Stature”. In: *Journal of the Anthropological Institute*, 15; 246–263.
- Guiraud, Pierre  
 1954 *Les caractères statistiques du vocabulaire*. Paris: Presses Universitaires de France.
- Herdan, Gustav  
 1964 *Quantitative Linguistics*. London: Butterworths.
- Hilberg, Wolfgang  
 1990 “Der bekannte Grenzwert der redundanzfreien Information in Texten – eine Fehlinterpretation der Shannonschen Experimente?”. In: *Frequenz*, 44; 243–248.
- Kieffer, John C.; Yang, Enhui  
 2000 “Grammar-based codes: A new class of universal lossless source codes”. In: *IEEE Transactions on Information Theory*, 46; 737–754.
- Köhler, Reinhard  
 1989 “Das Menzerathsche Gesetz als Resultat des Sprachverarbeitungsechanismus”. In: Altmann, Gabriel; Schwibbe, Michael H. (Hg.), *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Hildesheim: Olms, 108–112.
- Kuraszkiewicz, Władysław; Łukaszewicz, Józef  
 1951 “Ilość różnych wyrazów w zależności od długości tekstu”. In: *Pamiętnik Literacki*, 42(1); 168–182.
- Mandelbrot, Benoit  
 1953 “An informational theory of the statistical structure of languages”. In: Jackson, Willis (Ed.), *Communication Theory*. London: Butterworth, 486–502.

- de Marcken, Carl G.  
1996 *Unsupervised Language Acquisition*. Dissertation, Massachusetts Institute of Technology.
- Menzerath, Paul  
1928 "Über einige phonetische Probleme". In: *Actes du premier Congrès international de linguistes*. Leiden: Sijthoff.
- Miller, George  
1957 "Some effects of intermittent silence". In: *American Journal of Psychology*, 70; 311–314.
- Moses, Charikar; Lehman, Eric; Lehman, April; Liu, Ding; Panigrahy, Rina; Prabhakaran, Manoj; Sahai, Amit; Shelat, Ami  
2005 "The Smallest Grammar Problem". In: *IEEE Transactions on Information Theory*, 51; 2554–2576.
- Nevill-Manning, Craig G.  
1996 *Inferring Sequential Structure*. Dissertation, University of Waikato.
- Shannon, Claude  
1950 "Prediction and entropy of printed English". In: *Bell System Technical Journal*, 30; 50–64.
- Simon, Herbert A.  
1955 "On a class of skew distribution functions". In: *Biometrika*, 42; 425–440.
- Wolff, J. Gerard  
1980 "Language acquisition and the discovery of phrase structure". In: *Language and Speech*, 23; 255–269.