(**Note:** To appear **i**n V. Solovyev & V. Polyakov (eds.) (2005) Text Processing and Cognitive Technologies, No 11, Moscow: MISA, pp. 337-346. In case of any discrepancy with the printed version, the printed version will be the 'authorized' version.)

SYLLABLE COMPLEXITY AS A FUNCTION OF WORD COMPLEXITY

August Fenk, Gertraud Fenk-Oczlon, Lisa Fenk

ABSTRACT

The present study examines a specific version of Menzerath's first law: The more syllables per word, the fewer phonemes per syllable. The search for the best fitting function revealed the best fits for a model of exponential decay in two of three single languages as well as in a crosslinguistic computation with each of 33 different languages represented by one data pair: mean number of syllables per word (x) and mean number of phonemes per syllable (y). Despite of pronounced differences between languages as to their syllable complexity, the regression analysis showed a rather uniform curve progression of the decay from the maximum of syllable complexity in short words to the minimum in long words. The asymptotic values of this decay differ only slightly, and those differences obviously depend on the respective languages' variation of syllable complexity and their inventory of simple CV-, VC- and V-syllables. This offers a rather new and crosslinguistic view on the interplay between Menzerath's first law and his second law which implicitly postulates a restricted variation of syllable complexity in words containing a higher number of syllables. The exponential function obtained can be identified as a version of Altmann's generalized mathematical formulation of Menzerath's first law .The results are, moreover, discussed in terms of a relatively "constant" and economic flow of linguistic information.

KEYWORDS

Syllable complexity, word complexity, Menzerath's law, systems theory, information theory, regression analysis

1 INTRODUCTION

The most unassuming "definition" of the complexity of a certain system or unit refers to the number of elements or components of this unit (a); more "complex" conceptions may also take into account the complexity of the components (b), the number of different component types (Changizi 2001), the number of possible interactions between the parts (Simon 1996), the number of different rules determining these interactions, etc. This paper confines itself to the relatively simple levels (a) and (b) – higher complexity means a higher number of (more complex) components – and analyzes the complexity of syllables as a function of the complexity of words.

Menzerath, in his classical study from 1954, already investigates the relation between two measures of word complexity – word complexity in terms of number of syllables and in terms of number of phonemes. His first law (Menzerath I in Menzerath (1954: 100)) says that the "relative" number of phonemes (y) decreases with an increase of the number of syllables per word (x). In Menzerath **y denotes the number of phonemes per word**: the size of words as measured in phonemes is growing slower than their size as measured in syllables. But Menzerath I implies that, with an increasing number of syllables per word, the number of phonemes per syllable decreases. This deduction or transformation (I') interpreting **y as the number of phonemes per syllable** allows to use "syllable complexity" as a separate variable

and as a basic and universally applicable measure that can be related to other variables such as word complexity (present study) or complexity of sentences.

Menzerath's second law (II) postulates a rather restricted margin of deviation of the number of phonemes in those words containing a higher number of syllables. This implies again: The more syllables per word, the lower the variability of syllable complexity (II').

In a previous study (Fenk & Fenk-Oczlon (1993: 12)) we have extracted the relevant data concerning the German language from Menzerath's Table 8 (1954: 96). The first result of the statistical analysis (with SPSS) was a negative linear correlation between the number of syllables per word (word complexity x) and the number of phonemes per syllable (syllable complexity y): a correlation coefficient of $r = -0.766$ ($p < 0.05$), i.e. a determination coefficient ($R^2 = RSQ = r^2$) of 0.587. Grading the coordinate x (number of syllables per word) logarithmically resulted in an r^2 of 0.842, and admitting a quadratic function resulted in a further increase ($r^2 = 0.876$) of the "correlation". ("Correlation" in the broader sense of the word.)

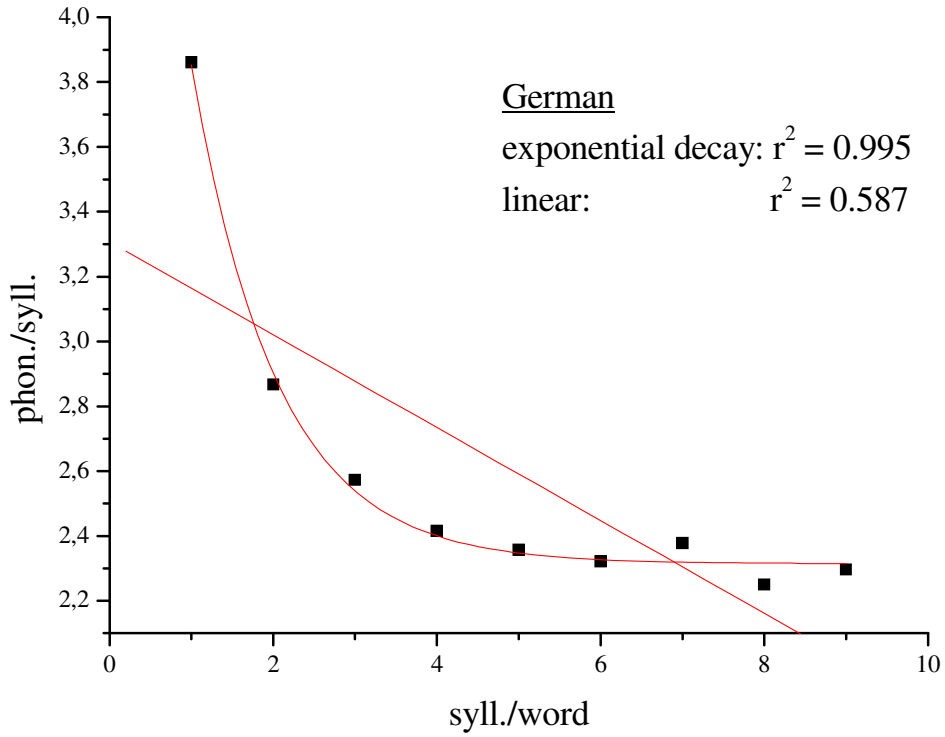
The finding of a negative correlation between syllable complexity and word complexity indicates some sort of balancing effects in order to maintain a rather invariant flow of linguistic information. The details of this trade-off may differ from language to language (see Section 3) with respect to typological differences in the word formation, in the variability of syllable complexity, etc. But the principle of such balancing effects in order to avoid a cognitive overload (and an overload of articulatory mechanisms) is regarded as a universal essential in natural language systems.

2 QUESTIONS AND ASSUMPTIONS

The present study uses a different statistical program (Origin 7) in order to find the best fitting functions. This search starts, in Section 3, on the single language level and with those data sets available in Menzerath (1954: 96, 114) concerning German and Italian and in Altmann & Schwibbe (1989: 56) concerning Indonesian. We assume that there are better fitting functions than found so far, and not only on the single language level, but also in crosslinguistic computation. These better fitting functions should allow more precise explanations of the trade-off between word complexity and syllable complexity. Thus, a more precise knowledge of the respective function will have some impact on both the theory of language variation and the understanding of universal economy principles in communication and cognition.

3 THE SINGLE LANGUAGE LEVEL: GERMAN, INDONESIAN, ITALIAN

In German and Indonesian the model of exponential decay showed the best fit from all the models investigated (linear, polynomial, etc.). Figures 1 and 2 offer a direct comparison between the exponential and the linear model, and Figure 2 of the cubic model as well. The model of exponential decay shows the highest fit in German and the lowest fit in Italian. In Italian the most probably best fitting function is cubic.


German
exponential decay: r2 = 0.995
linear: r2 = 0.587
phon./syll.
syll./word
4,0
3,8
3,6
3,4
3,2
3,0
2,8
2,6
2,4
2,2
0
2
4
6
8
10

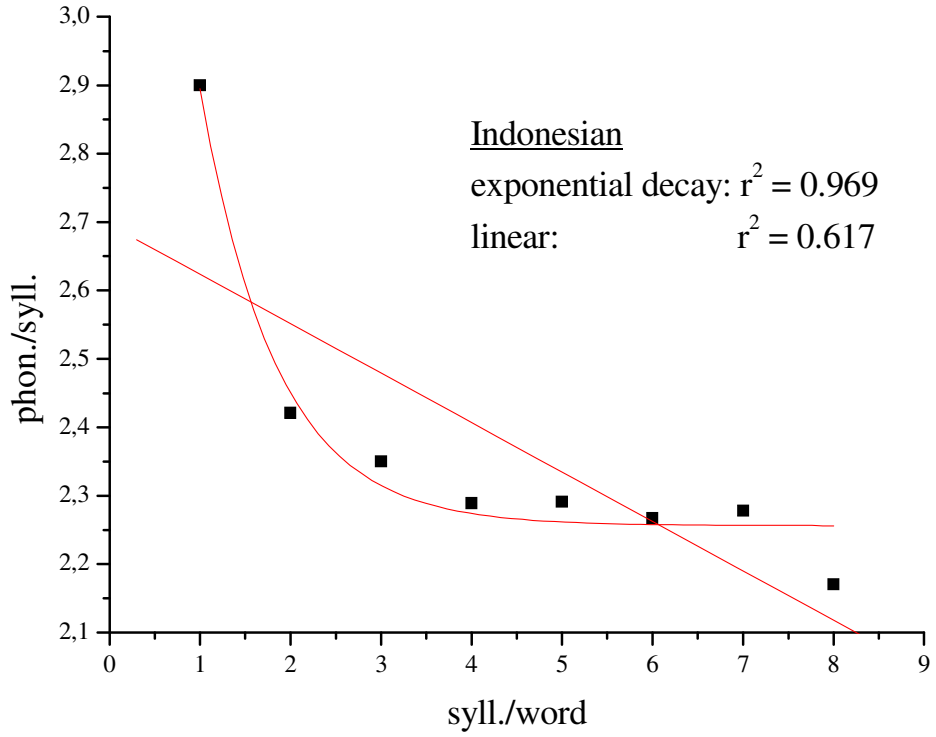
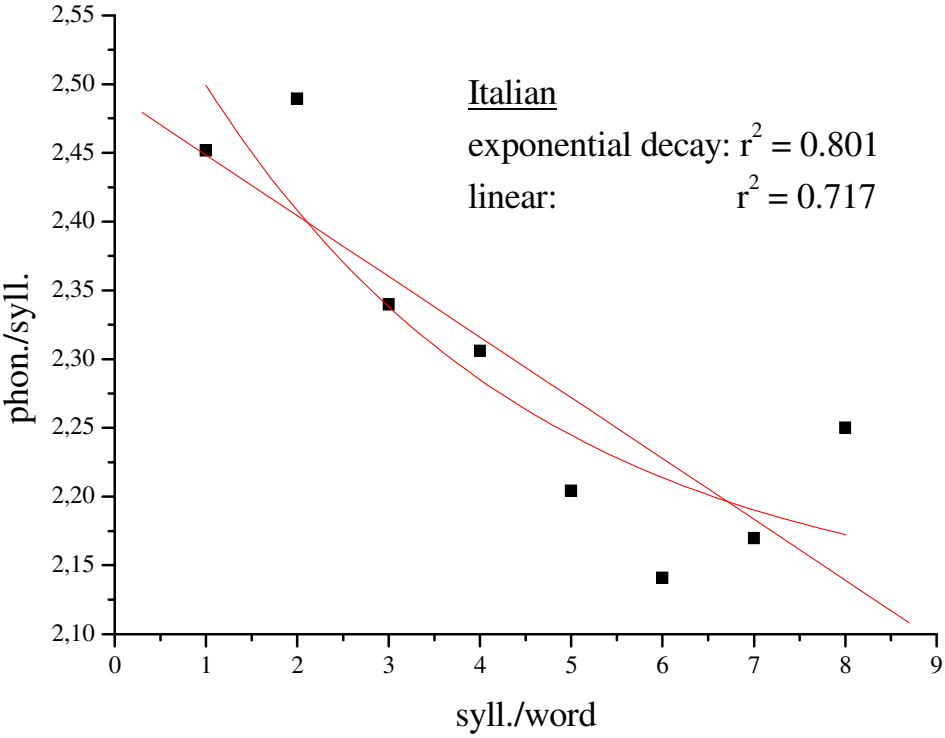


Figure 1. Syllable complexity as a function of word complexity, and a comparison between the fits with the exponential and the linear model in German and Indonesian.



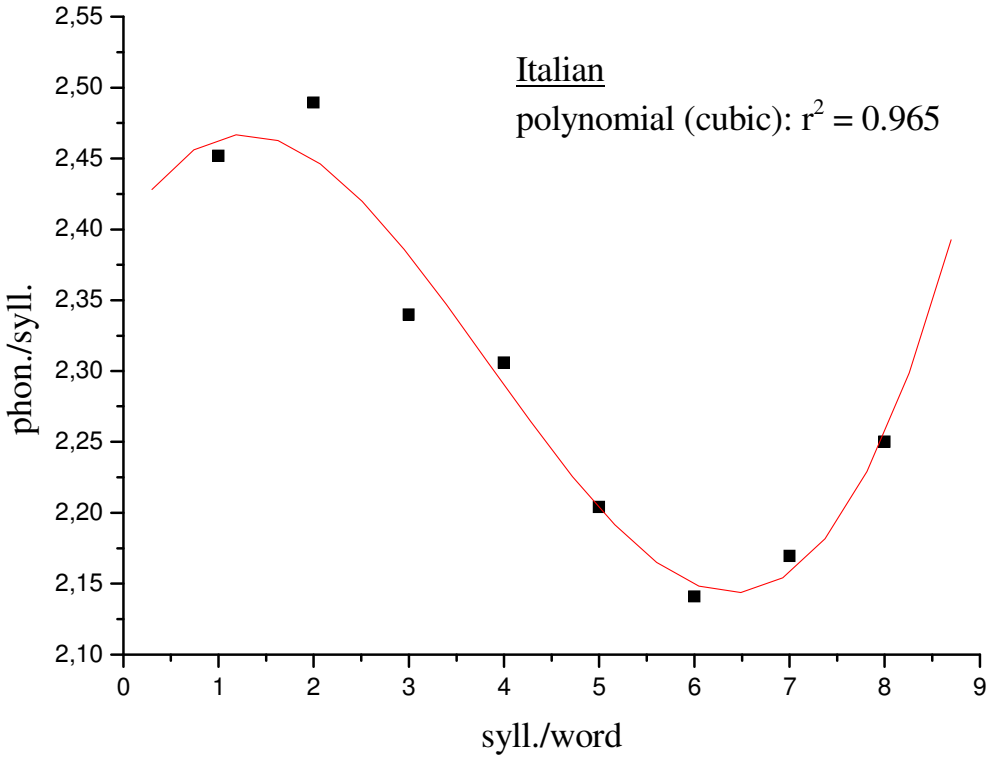


Figure 2. A comparison between the fits with the exponential and the linear model (upper panel) and the cubic model (lower panel) in Italian.

4 AGGREGATING DATA FROM THREE SINGLE LANGUAGES

We expect that averaging the data of these three languages would result in both a smoother curve and a better fit of the model of exponential decay. Actually, in the averaged data from German, Indonesian, and Italian (Figure 3) the model of an exponential decay showed a better fit (r^2 = 0.996) than in anyone of these single languages.

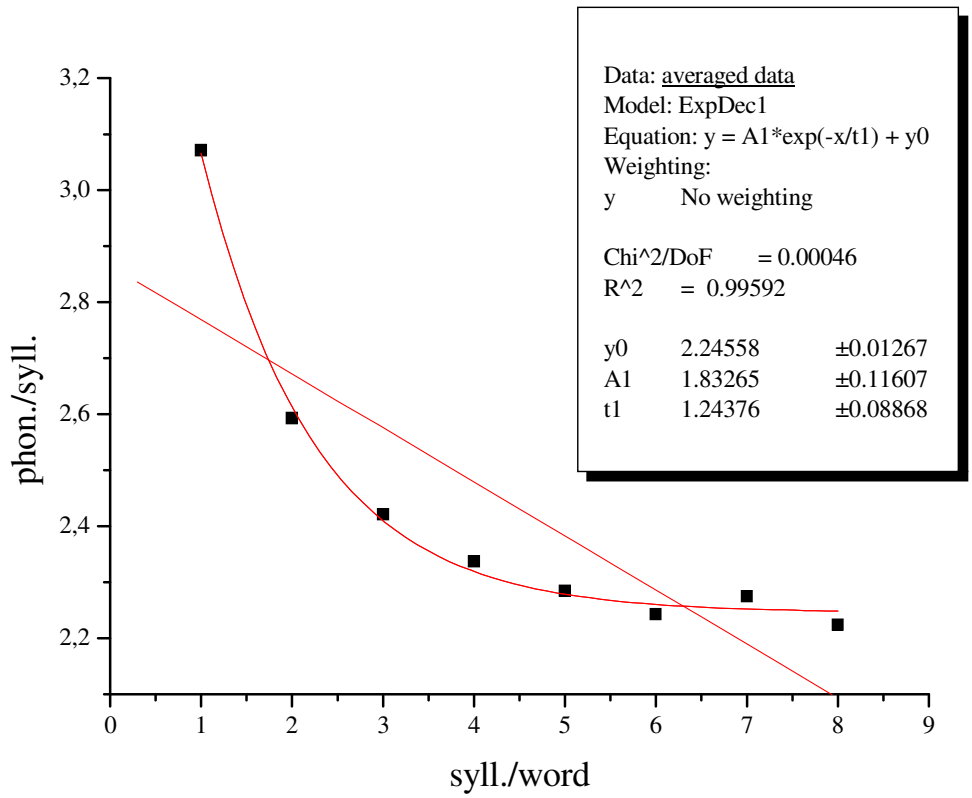


Figure 3. A comparison between the linear and the exponential model in the mean values from German, Indonesian, and Italian.

5 A CROSSLINGUISTIC PERSPECTIVE: ONE DATA-PAIR PER LANGUAGE

In order to test the assumption that a language's mean number of syllables per clause depends on this language's mean number of phonemes per syllable, we correlated these two variables and found a highly significant negative correlation between the complexity of sentences in terms of syllables and the complexity of syllables in terms of phonemes (Fenk-Oczlon & Fenk (1985)). This is in line with the most general form of Menzerath's first law: "the bigger the whole, the smaller the parts" (Menzerath (1954:10) in word-for-word translation). But here the bigger unit is the clause or sentence, instead of the word, and the domain of this new law is not the intra-language but the crosslinguistic analysis: The regression is not analyzed within a single language (or within data aggregated from single languages; Section 4), but across a sample of different languages, each of them represented by a single data-pair, i.e. its mean number of syllables per clause and its mean number of phonemes per syllable.

If the first characteristic value is the mean number of syllables per word, this corresponds to our regularity Menzerath I', but now on the level of crosslinguistic computation. A previous study (Fenk & Fenk-Oczlon (1993: 19), n = 29 languages) has revealed a negative correlation between the number of syllables per word and the number of phonemes per syllable. The coefficient of determination r^2 was 0.204 in the case of a linear function and 0.258 in the case of the best fitting quadratic function. What happens if we continue the search for the best fit by testing, in addition, the model of exponential decay in the meanwhile bigger sample of languages (Table 4 in Fenk & Fenk-Oczlon (1993) and its continuation in Fenk-Oczlon & Fenk (1999)?

5.1 DATA FROM 33 DIFFERENT LANGUAGES

Each data point in the range diagram below represents one single language – the mean of syllables per word (x) and the mean of phonemes per syllable (y). The sample of 33 languages includes 18 Indo-European and 16 non-Indo-European languages. Along the x-axis the extreme data points represent Chinese with one syllable per word and Korean with 2.648 syllables per word. Along the y-axis the extreme values originate from Japanese with 1.876 phonemes per syllable and Dutch with 2.973 phonemes per syllable. The statistical results of our previous study (Fenk & Fenk-Oczlon (1993)): $r = -0.45$ ($p < 0.001$); $r^2 = 0.20$ in this linear regression and 0.26 in a quadratic regression. The respective results in our 1999-study with a somewhat extended sample were $r = -0.54$ ($p < 0.001$) and $r^2 = 0.29$.

In the case of our exponential regression (see Figure 4) the fit is again noticeably higher: $r^2 = 0.37$.

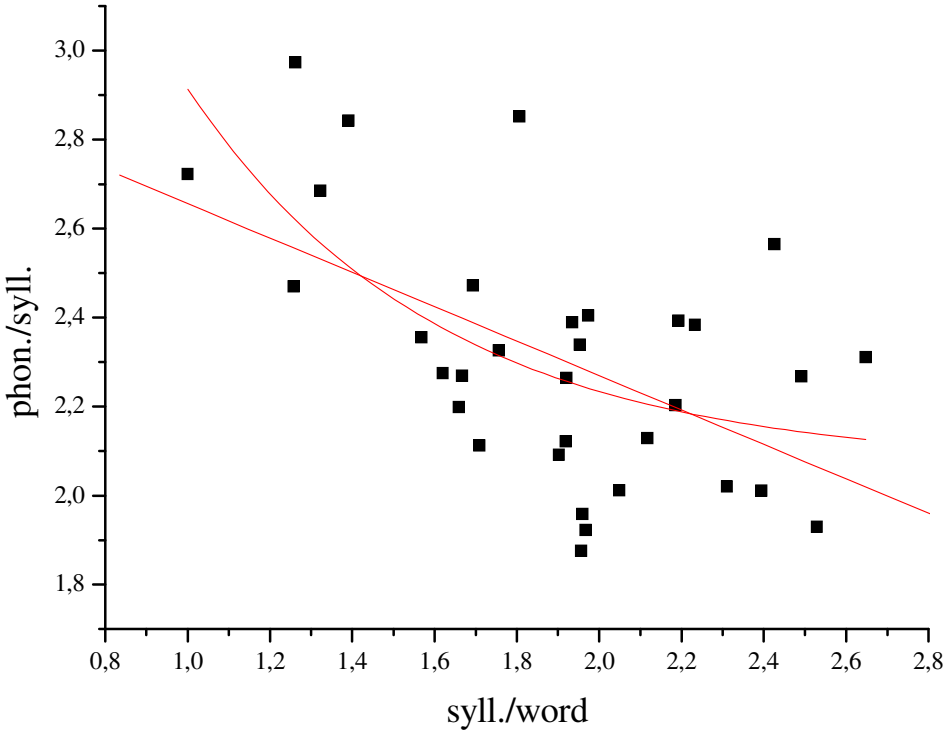


Figure 4. Characteristic values of 33 different languages and a comparison between the exponential ($r^2 = 0.372$) and the linear model ($r^2 = 0.295$).

5.2 DATA FROM FUCKS

Fucks (1956) analyzes texts from 9 different languages. But those characteristic values relevant in our context are available only from English (Shakespeare's "Othello"), German (Goethe's "Wilhelm Meister"), Greek, and Latin (presented from left to right in our Figure 4).

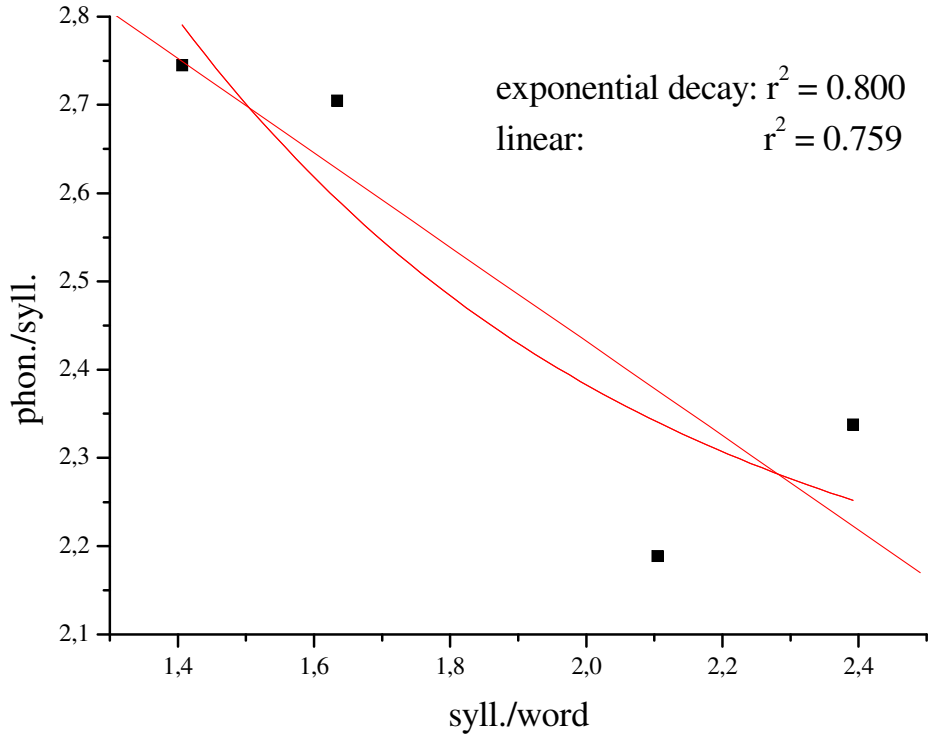


Figure 5. The "fits" in texts of four different languages (data from Fucks 1956).

6 INTERPRETATIONS, AND THE ROLE OF MENZERATH'S II. LAW

All four columns of Table 1 show a decrease of the values regarding the mean number of phonemes per syllable from German to Italian. This decrease is dramatic in the range of variation (column 3) and relatively moderate in the lowest values (column 2) and in the asymptotic values (column 4). In crosslinguistic computation (last line in Table 1) the highest value of 2.973 represents, as already mentioned, Dutch; in this computation German achieves a value of 2.843.

Table 1. Extreme values (columns 1 and 2) of the mean number of phonemes per syllable (basic data from Menzerath (1954)), the range of variation and the asymptotic values y_0 obtained in the application of the model of exponential decay.

	highest $\bar{y}$	lowest $\bar{y}$	range of $\bar{y}$	asympt. y_0
German	3.861	2.250	1.611	2.314
Indonesian	2.900	2.171	0.729	2.256
Italian	2.489	2.141	0.348	2.115
means of 33 languages	2.973	1.876	1.097	2.069

We may summarize: As compared with the pronounced typological differences regarding the maxima of syllable complexity (in monosyllabic words in German and Indonesian, in disyllabic words in Italian) there are relatively moderate differences concerning (a) the curve progression of the decay from the maximum or the "starting point" to the lower limits and concerning (b) those lower limits of the decay.

Ad (a) The model of the exponential decay does not achieve the best fits in any single language (see Italian in Figure 2) but seems to be the best model for most of the instances in single language analysis (Figure 1) and a suitable approximation in crosslinguistic computation as well (Figures 4 and 5).

Ad (b) A short but rather metaphorical explanation interpreting the "remaining" differences as regards the lower limits (columns 2 and 4 in Table 1) and connecting Menzerath's I.law with his II. law: Menzerath's laws I and II are effective in every language, and a high intralinguistic variation of syllable complexity is the presupposition for both, a pronounced decrease of the complexity of syllables in longer words (Menzerath I') and a pronounced decrease of the variability of syllable complexity in longer words (Menzerath II'). A virtual language having exclusively simple CV syllables simply could not follow these regularities, while languages showing a high and highly variable syllable complexity have good conditions to reflect Menzerath's laws. Maybe they have even better conditions to follow similar regularities on the level of the duration of articulation, i.e. to reduce the duration of syllables and the variability of this duration in words composed of a higher number of syllables. Because those languages seem to incline to stress-timed rhythm (Fenk-Oczlon & Fenk (2004)). But those languages having rather complex syllables and a relatively small inventory of very simple CV-, VC- and V- syllables will have more "difficulties" in finding enough very small components when producing words with an increasing number of components.

We tend to suggest that Menzerath I' and, indirectly, also II' serves the principle of a rather constant or invariant flow of linguistic information. The crosslinguistic comparison in Figure 6 shows an almost ideal proportionality function between the mean length of words in syllables and in the information (in bits) transmitted by these words. The proportionality function, i.e. the linear function forced through the zero point of the coordinates, is almost identical with the best fitting linear function. (An r^2 of 0.854 in both cases and no difference within the first five decimal places.) And even the fit of the seemingly best fitting non-linear model, the Boltzmann function, is not that much better ($r^2 = 0.887$). A higher number of syllables per word obviously goes hand in hand with a proportionately higher informational content and presumably also with a proportionately longer duration of articulation.

This postulate of a constant flow of linguistic information means, in crosslinguistic as well as in intralinguistic comparison, that phonemically poorer syllables transmit a proportionately smaller amount of information. This view is, rather indirectly, supported by or at least consistent with yet unpublished and extremely high correlations between word length in syllables and word length in phonemes: r = 0.999 in German and in Indonesian and 0.998 in Italian. (The relevant coefficient in a crosslinguistic computation was 0.841 in a sample of 33 languages.)

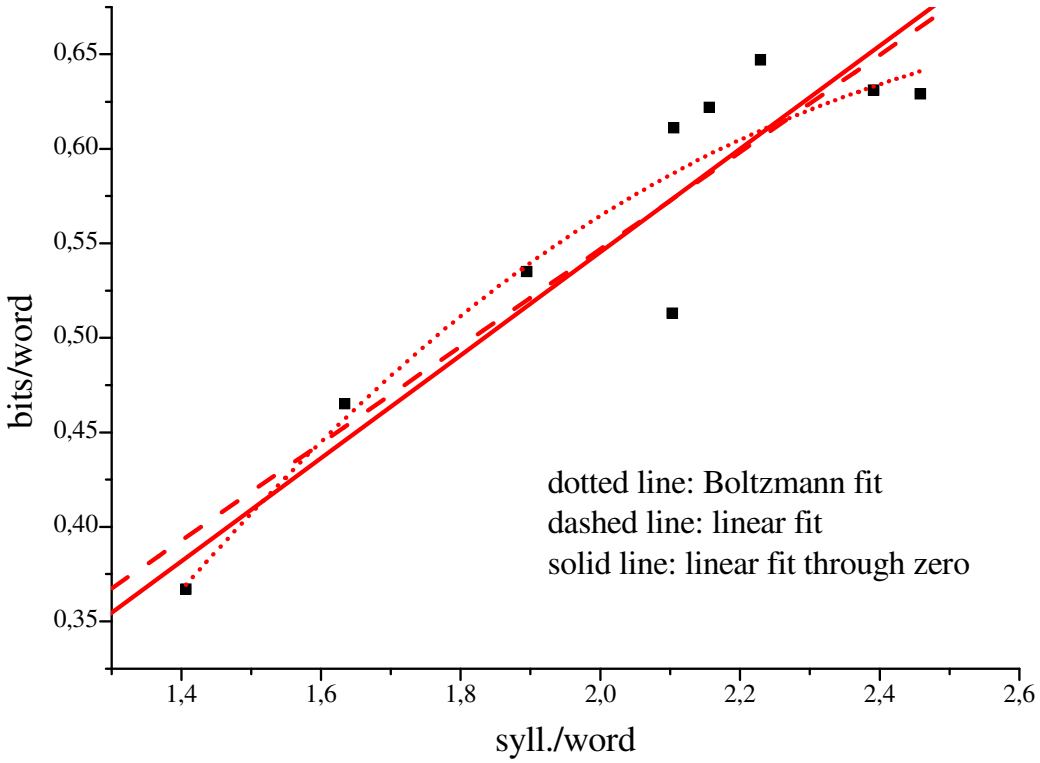


Figure 6. The words' informational content (in bits) as a function of their length in syllables (basic data from Fucks (1956:10)).

7 CONCLUSIONS

Menzerath's laws I and II and our specifications I' and II' describe universal facts about language. These facts can be related to economy principles in information processing – "information" in the sense of information theory – forced by the constraints of our cognitive capacities.

A second conclusion refers to a mathematical formulation of Menzerath I suggested in Altmann (1980; cf. Altmann et al. (1982:537)):

$y'/y = b/x + c$, with y as the size of the constituents, x as the size of the construct, and b and c as constants.

The above mentioned authors present the general solution of this differential equation, i.e. $y = ax^b e^{cx}$, and the special case $y = ae^{cx}$ when b = 0. The best fitting function of exponential decay, as described in the present study, is, if we assume a negative sign of c, equivalent with this special case with the exception of an additional asymptotic value y_0: $y = ae^{cx} + y_0$. This asymptotic value is determined by the respective languages' inventory of simple syllables and their mean syllable complexity.

REFERENCES

1. Altmann, G. , & Schwibbe, M. H. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Hildesheim: Olms.
2. Altmann, G. (1980). Prolegomena to Menzerath's law. *Glottometrika*, 2, 1-10.
3. Altmann, G., Beöthy, E., & Best, K.-H. (1982). Die Bedeutungskomplexität der Wörter und das Menzerathsche Gesetz. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 35, 537-543.
4. Changizi, M.A. (2001). Universal Scaling Laws for Hierarchical Complexity in Languages, Organisms, Behaviors and other Combinatorial Systems. *J. theor. Biol. 211*, 277-295.

5. Fenk, A. & Fenk-Oczlon, G. (1993). Menzerath's Law and the Constant Flow of Linguistic Information. In R.Köhler & B.Rieger (Eds.), *Contributions to Quantitative Linguistics,* 11-31. Dordrecht: Kluwer.
6. Fenk-Oczlon, G. & Fenk, A. (1985). The Mean Length of Propositions is 7 Plus Minus 2 Syllables - but the Position of Languages within this Range is not Accidental. In G. D'Ydewalle (Ed.), *Cognition, Information Processing, and Motivation. XXIII International Congress of Psychology*. (Selected/revised papers), 355 – 359. Amsterdam: North-Holland, Elsevier Science Publishers B.V.
7. Fenk-Oczlon, G. & Fenk, A. (1999). Cognition, Quantitative Linguistics, and Systemic Typology. *Linguistic Typology 3*, 151-177.
8. Fenk-Oczlon, G. & Fenk, A. (2004). Systemic Typology and Crosslinguistic Regularities. In V. Solovyev & V. Polyakov (Eds.), *Text Processing and Cognitive Technologies. International Conference "Cognitive Modeling in Linguistics 2004", Paper Collection, N9*, 229-234. Moscow: MISA
9. Fucks, W. (1956). Die mathematischen Gesetze der Bildung von Sprachelementen aus ihren Bestandteilen. *Nachrichtentechnische Fachberichte 3*, 7-21.
10. Menzerath, P. (1954). Die Architektonik des deutschen Wortschatzes. Bonn: Dümmler.
11. Simon, H.A. (1996 [1962]). The Architecture of Complexity: Hierarchic Systems. *Proceedings of the American Philosophical Society 106* (Dec.1962), 467-482.