(**Note**: to appear 2008 in: M. Miestamo, K. Sinnemäki and F. Karlsson (eds.) *Language complexity: typology, contact, change,* pp. 43-65.. Amsterdam/Philadelphia: John Benjamins. In case of any discrepancy with the printed version, the printed version will be the 'authorized' version.)

Complexity trade-offs between the subsystems of language

Gertraud Fenk-Oczlon & August Fenk

Starting from a view on language as a combinatorial and hierarchically organized system we assumed that a high syllable complexity favours a high number of syllable types, which in turn favours a high number of monosyllables. Relevant crosslinguistic correlations based on Menzerath's (1954) data on monosyllables in 8 languages turned out to be statistically significant. A further attempt was made to conceptualise "semantic complexity" and to relate it to complexity in phonology, word formation, and word order. In English, for instance, the tendency to phonological complexity and monosyllabism is associated with a tendency to homonymy and polysemy, to rigid word order and idiomatic speech. The results are explained by complexity trade-offs rather between than within the subsystems of language.[1]

1. Hierarchy and complexity in the language system

In his famous article on "The Architecture of Complexity: Hierarchic Systems", Herbert A. Simon (1996; originally 1962) called the attention of systems theory to hierarchy as a central scheme of organized complex systems:

> Thus my central theme is that complexity frequently takes the form of hierarchy and that hierarchic systems have some common properties independent of their specific content. Hierarchy, I shall argue, is one of the central structural schemes that the architect of complexity uses. (Simon 1996:184)

By a "complex system" he means "one made up of a large number of parts that have many interactions." (pp.183f) "In a hierarchic system one can distinguish between the interactions *among* subsystems, on the one hand, and interactions *within* subsystems – that is, among the parts of those subsystems – on the other." (p.197) Later in his article he qualifies the complexity of a structure as critically depending "upon the way in which we describe it" (p.215).

In Gell-Mann's (1995) conceptualisation of *complexity* the description of a structure becomes the crucial point. But he argues that the algorithmic information content (AIC) of such a description is, e.g. because of its context dependency, an inappropriate measure for complexity. "Effective complexity", he says, "refers not to the length of the most concise description of an entity (which is roughly what AIC is), but to the length of a concise description of a set of the entity's regularities." High "effective complexity" in the sense of Gell-Mann amounts to a relatively high number of regularities: It becomes near zero in "something almost entirely random" as well as in "something completely regular, such as a bit string consisting entirely of zeroes." It "can be high only in a region intermediate between total order and complete disorder". In this respect Gell-Mann's conceptualisation differs from classical information theory. In information theory as well as in Kolmogorov complexity (cf. Juola 1998), highest complexity is attributed to random order.

A rather recent systems theoretical article by Changizi (2001) studies first of all the development of communication systems. In the abstract of his article he states

> general laws describing how combinatorial systems change as they become more expressive. In particular, /…/, increase in expression complexity (i.e. number of expressions the combinatorial system allows) is achieved, at least in part, by increasing the number of component types. (Changizi 2001:277)

As an example for the development of "human language over history" he takes English and studies "how the number of word types in a language increases as the number of sentences increases" (p.281) and concludes that "increasing expressivity in English appears to be achieved exclusively by increasing the number of word types" (p.283). Notice that in his terminology each "entry in the dictionary is a different word type" (p.281).

The following criteria of *complexity* may be ordered in the sense of increasing degrees or levels of complexity:

(a) The number of components: That there are entities that can form a bigger entity or can be described as components of this bigger entity is the minimum requirement for complexity. But it is a matter of terminology if such a 2-step organisation – a unit and its elements – already should be considered a minimal hierarchy.

(b) The number of components of the components, i.e. the complexity of the components: This criterion refers to a real hierarchy of at least 3 steps.

(c) The number of component types (Changizi 2001): The existence of different types of components makes the entity more complex, irrespective of whether or not it is really hierarchically organized.

(d) The number of possible interactions between the components (Simon 1996): The higher a system's complexity with respect to (a), (b), and (c), the higher its complexity with respect to (d).

(e) The number of rules determining these interactions, i.e. the number of rules necessary for a concise description of these interactions (Gell-Mann 1995): The higher (a) – (d), the higher the possible number of rules determining the interactions within and between the components.

But if we focus on the system "human language", we can make out two different types of structuring of this system:

1. Within language, or below the superordinate concept *language*, we may discriminate between different subsystems – or rather levels of description? – such as phonology, morphology, syntax, and semantics. This structure is not very distinctive. Semantics, for instance, intrudes on all other subsystems of language.
2. A rather "technical" hierarchy may differentiate, even if ending at the sentence level, between at least five hierarchical steps in the sense of the above criterion b: phonemes, syllables, words, clauses, sentences. The elements of the lowest level are the phonemes, and each unit at a higher level n is, in principle, a complex or composition of units of the level n - 1 and is in turn an element of units at the level n + 1. But a unit on level n can be identical with a unit on level n – 1, as is the case in monophonemic syllables, monosyllabic words, monoclausal sentences, and, depending on the definition of *clause*, even one-word clauses. Nor should we forget the argument that not all of these divisions are equally clear-cut in any language (c.f. the division of sentences or clauses into words) and at any level: The syllable can – unlike the phoneme (c.f. Ladefoged 2001) – occur as an independent entity and is – unlike the morpheme, for instance – an easily countable component of bigger entities.

Obviously, these two hierarchies cannot be fully compatible. But when assigning dimensions of linguistic complexity to different subsystems of language (in Table 1) we encounter some metric parameters belonging to the rather technical hierarchy. The selection of complexity facets in Table 1 is of course "biased" by the findings to be discussed in the chapters below; for instance when taking the number of syllables per word as an indicator of the morphological complexity of these words. Especially in the extremes – monosyllables on the one hand, extremely long words on the other – the n of syll./word will be an excellent predictor of the words' morphological complexity. From that it follows that languages having rather short words will need more words for encoding a certain semantic unit. These considerations are supported by our crosslinguistic correlation (d) in Section 2.

Table 1: Relates some dimensions of linguistic complexity to certain subsystems of language.

subsystems	*facets of linguistic complexity*
phonology	size of phonemic inventory syllable complexity (= n of phon./syll.) n of syllable types
morphology	complexity in word structure (n of morphemes and n of syllables per word) n of morphological cases, gender distinctions etc. opaqueness of morphological forms
syntax	rigid (?) word order hypotactic constructions
semantics	n of meanings per expression (homonymy, polysemy)

2. Crosslinguistic correlations

2.1 Previous results

A previous study by the authors (Fenk & Fenk-Oczlon 1993)[2] revealed a set of significant and mutually dependent crosslinguistic correlations between the four variables number of phonemes per syllable (syllable complexity), number of syllables per word, number of syllables per clause and number of words per clause:

(a) The more syllables per word, the fewer phonemes per syllable.
(b) The fewer phonemes per syllable, the more syllables per clause.
(c) The more syllables per clause, the more syllables per word.
(d) The more syllables per word, the fewer words per clause.

Correlation (a) is a crosslinguistic version of a law originally found by Menzerath (1954: 100) in German: "The relative number of sounds decreases with an increasing number of syllables [per word]" (our translation). Additional calculations admitting higher order (quadratic, cubic, logarithmic) functions resulted – for obvious reasons – in higher determination coefficients (Fenk & Fenk-Oczlon 1993, e.g. 18f) than the linear correlations. The whole set of correlations a – d was confirmed in a later study (Fenk-Oczlon & Fenk 1999) with an extended sample of 34 languages, 18 Indo-European and 16 non-Indo-European.

In the case of Menzerath's law on the single-language level the best fit (e.g. a determination coefficient of .995 for German) could be achieved when using the model of exponential

decay in order to describe the syllable complexity (n of phonemes per syllable) as a function of the number of syllables per word (Fenk, Fenk-Oczlon, & Fenk 2006). That this function of an exponential decay can be shown (same article, p.332) as a special case of Altmann's (1980) mathematical generalization of Menzerath's original law is not that surprising. Altmann's law, often referred to as the "Menzerath-Altmann law" or even, as already in Altmann (1980), as "Menzerath's law", is so general that Meyer (2007) could specify a mere stochastic mechanism generating such relations "in ensembles of hierarchically structured entities of whatever kind."

Summarizing the results of our previous work we may say two things. Firstly, as a plausible consequence[3] of the negative correlations of the syllable complexity (A) with both the number of syllables per word (B) and the number of syllables per sentence (C), we expected and indeed found a positive correlation between B and C (Fenk & Fenk-Oczlon 1993:18). This is one of several arguments for the mutual dependency of the correlations and for a systemic view of language variation. Our crosslinguistic correlations can best be understood as balancing effects between subsystems of language. Such balancing effects seem to provide a crosslinguistically rather constant size (duration) of clauses and of mono-clausal sentences.

Syllable complexity seems to play a key role within this system (Table 2). It interacts, first of all, with other metric properties. (In each column of Table 2 the first 4 rows including the headings "paraphrase" the correlations a – d mentioned above.) Syllable complexity is also significantly associated with word order (Fenk-Oczlon & Fenk 1999:163f) and almost significantly with the number of cases, which is in turn significantly associated with adposition order (Fenk-Oczlon & Fenk 2005:81). These two studies offer some further statistical as well as theoretical arguments concerning the last 3 rows in Table 2. And syllable complexity also interacts, more or less directly, with semantics, as we will argue in Section 3.

Table 2: Associations between syllable complexity and some other metric and non-metric properties.

high syllable complexity	*low syllable complexity*
low n of syllables per word	high n of syllables per word
low n of syllables per clause	high n of syllables per clause
high n of words per clause	low n of words per clause
VO order	OV order
low n of morphological cases	high n of morphological cases
prepositions	postpositions
cumulative case exponents	separatist case exponents
stress timed	syllable timed
fusional or isolating morphology	agglutinative morphology

2.2. New assumptions

A high syllable complexity in a certain language requires a rather large phonemic inventory. (A high syllable complexity can only be achieved by large initial and final consonant clusters. In languages showing comparable degrees of freedom in the combinatorial possibilities of consonants, those having a larger inventory of consonants will incline to bigger consonant clusters.) A relatively high syllable complexity is in turn a precondition for a high variability of syllable complexity and therefore also for a high number of syllable types; a high number of syllable types will not be possible in a language with a maximum of, let us say, two

phonemes per syllable. And a high number of syllable types is a precondition for a large inventory of monosyllabic words. Facing such implications of combinatorial possibilities we have to be aware of two methodologically relevant aspects:
On the one hand it is plausible to assume that a certain system realizes a good deal of its combinatorial possibilities. On the other hand, the concrete system will hardly ever realize the total of cogitable possibilities. In the system "language", as we know, the phonotactic possibilities realized fall below the number of combinatorial possibilities. Thus, one cannot mathematically determine a language's mean number of phonemes per syllable from a given size of its phonemic inventory or its number of syllable types from a given mean or maximum syllable complexity or its number of monosyllables from a given number of syllable types. Instead, we always depend on empirical investigation if we want to determine the extent to which combinatorial possibilities are realized.

Thus our chain of preconditions and requirements (first paragraph in this section) should be reformulated in the sense of a chain of statistically testable tendencies. The data regarding syllable complexity (X), the number of syllable types (Y) and the number of monosyllabic words (Z) could be put together from Menzerath's (1954) descriptions of eight different languages or could be calculated from these data. But it should be noted that this number of syllable types is the number of syllable types realized in the monosyllabic words of the respective languages. If it is true that especially monosyllables "use" a considerable range of the spectrum of syllable types, or – in rather analytic/fusional languages – almost this whole spectrum, then this value is acceptable at least as a good indicator for the number of syllable types[4]. Thus, the following hypotheses can be evaluated:

Hypothesis I: The higher a language's number of syllable types, the higher its number of monosyllabic words.
Hypothesis II: The higher a language's syllable complexity, the higher its number of syllable types.
Hypothesis III: The higher a language's syllable complexity, the higher its number of monosyllabic words.

In a second step we tried to find data generated by one author and one method regarding the size of the phonemic inventory (W) in the given sample of 8 languages. Determining the number of phonemes is generally problematic (Bett 1999), and one of the most problematic areas is the analysis of phonetic diphthongs (Maddieson 1984:161). Spanish, for example, seems to have no diphthongs in Ladefoged (2001) but has 5 diphthongs according to Campbell (1991). Campbell offers exact numbers regarding consonants and monophthongs in our 8 languages but remains incomplete as to the number of diphthongs. Thus, in testing the first chain in our above arguments (first sentence in this Section), we decided to take the sum of consonants and monophthongs as a value indicating the size of the phonemic inventory:

Hypothesis IV: The bigger a language's phonemic inventory, the higher its syllable complexity.

2.3. Evaluation and results

Our hypotheses do not imply any specific assumption regarding the shape of the respective regressional functions, so that the standard correlation (Pearson's product-moment correlation) and the corresponding test of significance was the first choice. If this application of the linear model reveals significant coefficients, this means a confirmation of the hypotheses in a very rigid examination, because further appropriate curve-fitting procedures

will necessarily result in higher, never in lower determination coefficients. In the following we will compare the results of the application of the linear model with the results of the data-driven search for better fitting models. From those "simple" (c.f. Mulaik 2001) models tested the model of exponential growth achieved the best fit in all of the following evaluations (see Figures 1-3).

Hypothesis I predicts a crosslinguistic correlation between the number of syllable types and the number of monosyllables. Or, to put it the other way round: A large inventory of monosyllabic words will go hand in hand with a large inventory of different syllable types: V; CV, VC; CCV, CVC, VCC;…CCVCCCC; … CCCCVCCCC. This hypothesis is in line with the argument that, in the system "language", complexity trade-offs will happen between rather than within its subsystems. A high number of syllable types reflects a high (variation of) phonological complexity while a high number of monosyllables reflects a low complexity in word formation – a low complexity in terms of n of syllables and, indirectly, in terms of morphemes as well.

Table 3: The frequency of monosyllables consisting of different numbers (1, 2, … 8) of phonemes. (Data assembled from Menzerath 1954)

n of phon. per monosyllable	1	2	3	4	5	6	7	8
1 English	14	326	2316	2830	1199	161	8	2
2 German	9	114	645	962	444	69	2	
3 Romanian	8	81	480	474	135	19	1	
4 Croatian	6	42	353	273	42	1		
5 Catalan	11	94	285	265	25			
6 Portuguese	9	84	177	51	4			
7 Spanish	10	59	115	70	9			
8 Italian	4	50	32	7				

Table 4: Data collected or calculated from Menzerath's descriptions of eight different languages.

	X n of phon. per monosyllable		Y n of syllable types realized in monosyllables	Z n of monosyllables
	X_{mean}	X_{max}		
1 English	3.787	8	43	6856
2 German	3.861	7	35	2245
3 Romanian	3.591	7	16	1198
4 Croation	3.427	6	12	717
5 Catalan	3.293	5	11	680
6 Portuguese	2.868	5	9	325
7 Spanish	3.034	5	17	263
8 Italian	2.452	4	8	93

Table 3 assembles relevant data from Menzerath's typological descriptions of eight different languages (Menzerath 1954: 112 – 121) regarding the frequency of monosyllables of different complexity. From the rows in Table 3 one can calculate the mean n of phonemes per monosyllable in any single language. Taking Italian as an example: 1x4 + 2x50 + 3x32 + 4x7 = 228 divided by the cumulative frequencies (4 + 50 + 32 + 7 =) 93 is 2.452. Table 4 compares this mean syllable complexity (X_{mean}) with other data by Menzerath. It allows a direct test of our Hypotheses I – III in the sense of crosslinguistic correlations. The first result[5] was a clear confirmation of Hypothesis I:

Hypothesis I: The more syllable types (in monosyllables), the more monosyllables. r_{yz} = + *.895 (p < .01).* This means a determination coefficient of .801. The determination coefficient achieved with an exponential growth model (Figure 1) is .978. Figure 1 also highlights the essential contribution of the 2 Germanic languages (English and German, cf. Table 4) to that regression.

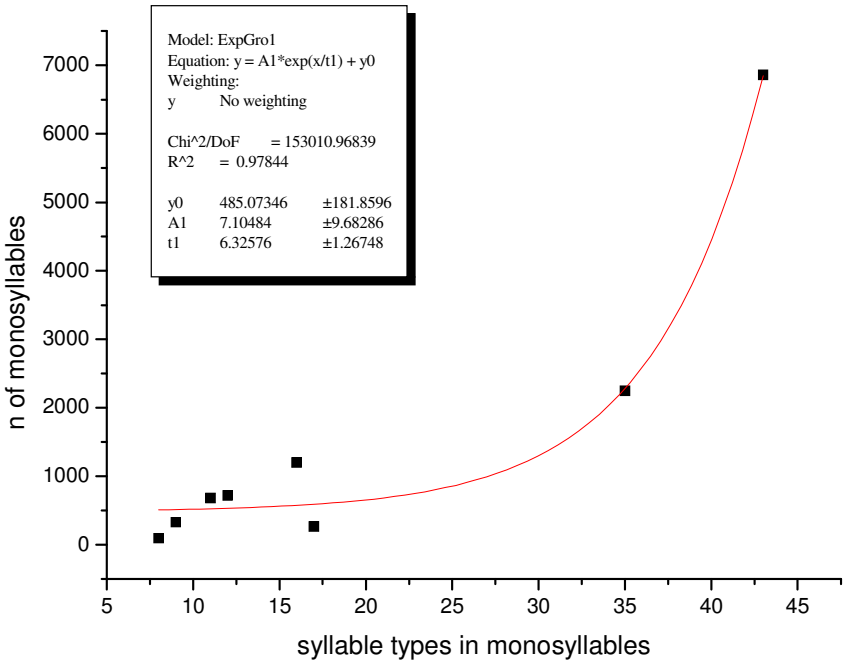

Figure 1: The regression regarding Hypothesis I with the exponential growth model.

Hypotheses II and III both concern the parameter of syllable complexity. The examination of Hypothesis II again revealed significant correlations.
Hypothesis II: The higher the mean number of phonemes per syllable, the more syllable types. r_{xy} = + *.76 (p < .05)*
Hypothesis II' The higher the maximum number of phonemes per syllable, the higher the number of syllable types. r_{xy} = +. *835 (significant, p < .01)*
This means a determination coefficient of .578 or .697 respectively. The determination coefficients achieved with a model of exponential growth (Figure 2) are .820 or .801. This means that in the seemingly most appropriate regressional model the mean syllable complexity is a better predictor of the number of syllable types than the maximum syllable complexity.

Because of different phonotactical possibilities in the respective languages, such correlations cannot be predicted or explained by mere combinatorial effects. In the case of a trivial relation between the mean number of phonemes per monosyllable (X_{mean} in Table 4) and the number of syllable types (Y) a regression r_{xy} should be "perfect"; German for instance should have, as compared with English, either a lower value in X_{mean} or a higher value in Y. The very same can be said if we take, instead of the **mean** number of phonemes, the **maximum** number (X_{max}) of phonemes: In the case of a trivial relation it would not be possible to find 3 Romanic languages with an equal maximum of 5 phonemes per syllable showing different numbers of syllable types ranging from 9 in Portuguese to 17 in Spanish, or to find 2 languages (Romanian, German) with an equal maximum of 7 phonemes per syllable but 16 vs 35 syllable types (X_{max} in Table 4 and right panel in Figure 2).

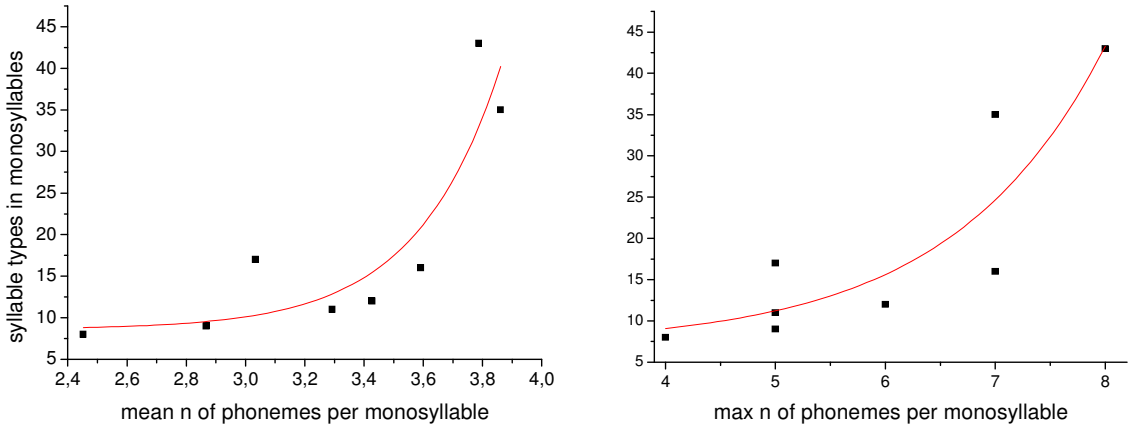

Figure 2: Regressions regarding Hypothesis II. The independent variables are mean syllable complexity (left panel) and maximum syllable complexity (right panel).

The inferential step from syllable complexity (X) to the number of monosyllabic words (Z) is not trivial either: Of those 3 languages with an equal value for maximum syllable complexity (X_{max} in Table 4), Spanish has indeed by far the highest number of syllable types (17) but by far the lowest number of monosyllables: 263 as compared with 680 in Catalan. Notably, Catalan also shows the highest syllable complexity of those 3 languages. The relevant Hypothesis III is not only the remaining link in our chain of arguments. It was, moreover, inspired by our significant correlation (a): "The higher the number of phonemes per syllable, the lower the number of syllables per word". The absolutely lowest number of syllables is of course realized in the monosyllabic word so that there is only a small step to predicting an association between high syllable complexity and a strong tendency to produce monosyllabic words.
The result of the statistical examination was an almost significant correlation when the n of monosyllables was correlated with the mean syllable complexity and a significant correlation when correlated with the maximum syllable complexity:

Hypothesis III: The higher the mean number of phonemes per syllable, the more monosyllabic words. $r_{xz} = + .64$ *(*$p < .1$*)*
Hypothesis III': The higher the maximum number of phonemes per syllable, the higher the number of monosyllables. $r_{xz} = + .807$ *(*$p < .05$*)*
This means a determination coefficient of .410 or .651 respectively. The determination coefficients achieved with a model of exponential growth (Figure 3) are .536 or .980.

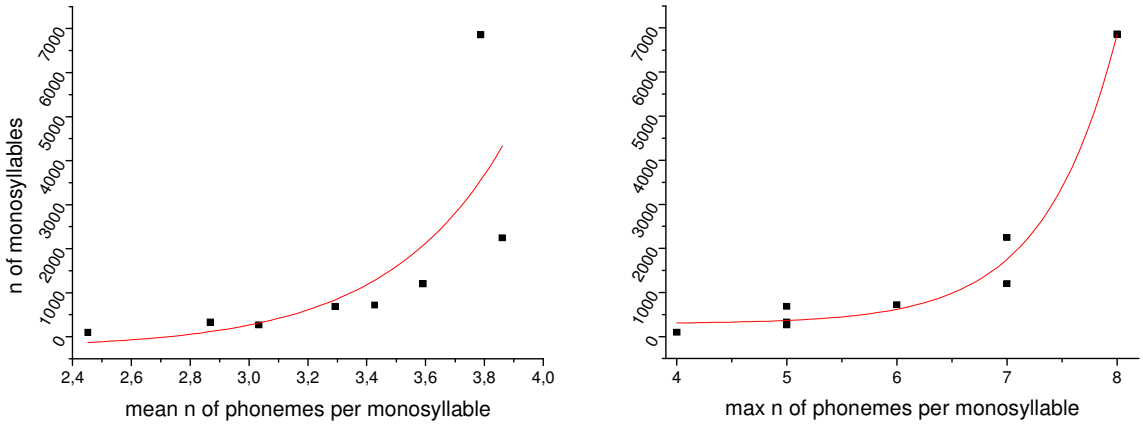

Figure 3: The regressions regarding Hypothesis III, with X_{mean} (left panel) and X_{max} (right panel) as the independent variables.

In the evaluation of the remaining Hypothesis IV we counted the numbers of phonemes, excluding diphthongs, in our 8 languages as reported by Campbell (1991): English 35, German 40, Romanian 28, Croatian 34, Catalan 30, Portuguese 31, Spanish 25, and Italian 29. The correlation of these numbers with mean syllable complexity (X_{mean} in Table 4) was not too far from being significant and showed, as expected, a positive sign: *The bigger the phonemic inventory, the more phonemes per syllable.* r_{wx} = + .622 ($p < .1$). This means a determination coefficient of .387. The determination coefficient achieved with the exponential growth model was .411. When the syllable complexity figured as the independent variable, this determination coefficient was .650. This might be seen as a further indication (cf. Fenk-Oczlon & Fenk 2005, Fenk & Fenk-Oczlon 2006) for the central role of syllable complexity in language variation.
The linear correlations of the segmental inventory (W) with the other variables in Table 4 (X_{max}, Y, and Z) showed positive signs as well: + .573, + .635 ($p < .1$), and + .508.

2.4 Discussion

What we could show significantly in a small sample of 8 Indo-European (1 Slavic, 2 Germanic, 5 Romanic) languages is a new set of mutually dependent, crosslinguistic correlations between monosyllabism, number of syllable types, and syllable complexity. (All the correlations of the phonemic inventory size with other variables showed the expected signs, and two of them were almost significant.)

In the case of a negative correlation between e.g. syllable complexity and number of syllables per clause one may easily identify a balancing mechanism. But how can we state such a balancing effect in view of our three positive correlations? Our answer to this question is that both a high number of phonemes per syllable and a high number of syllable types mean or reflect high phonological complexity, while the tendency to produce many monosyllables reflects low complexity in word structure.

2.4.1 A short excursion into diachrony
If we consider the repertoire of monosyllables as the system in question and the syllable types as its component types and take a look at the diachronic changes of English from this

perspective we find developments that conform both Changizi's claims (see Section 1) and our model of complexity trade-offs: A comparison of the Beowulf Prologue in Old English (OE) with its translation into Modern English (ME) shows a remarkable increase of monosyllables from 105 in OE to 312 in ME and a concomitant increase of the mean syllable complexity from 2.63 phonemes in OE to 2.88 in ME.
We could not determine the exact number of syllable types in OE, but higher syllable complexity favours a higher number of syllable types (see our significant correlations II and II' in Section 2.3). Other indications for a higher number of syllable types in ME than in OE are: The loss of final segments as in *guest* versus *gas-tir*, *horn* versus *hor-na* (examples from Lehmann 1978) resulted in an increase of final consonant clusters and therefore also of syllable types. And while the initial consonant clusters seem to be similar in OE and ME, Modern English shows a higher number of complex final clusters such as *strands* /strændz/ CCCVCCC, *glimpsed* /glɪmpst/ CCVCCCC.

2.4.2 Where the trade-offs happen and why they do not indicate an equal overall complexity of languages

All of our previous and present results indicate that in the system "language" one meets complexity trade-offs between rather than within the subsystems, and that within a subsystem one may even observe a diacronic increase in many parameters of complexity.
This corresponds to the results reported by Maddieson (1984) concerning phonology. He showed convincingly that languages with a large consonant inventory also tend to have a large vowel inventory (p.17). A lower number of manner contrasts was not found to be compensated by a higher number of place contrasts of stops and fricatives (p.18), and the often stated assumption that a small phonemic inventory is compensated by more complex suprasegmentals (i.e. tone, stress) could not be confirmed either. Maddieson's analysis of 56 languages showed on the contrary that languages with simpler segmental inventories tend to have less elaborated suprasegmental properties (p.21). He also mentions a positive though rather weak and insignificant correlation between segmental inventory size and syllable inventory size. In a recent study (Maddieson 2006:110f) with an extended sample of languages he found a significant difference: the most consonants in the inventory of languages with complex syllable structures, the least in those with simple syllable structures. Our positive correlation between phonemic inventory and syllable complexity is, though not significant, in line with these findings by Maddieson.
Menzerath's (1954) law points to complexity trade-offs on the intra-language level, and our results (Fenk & Fenk-Oczlon 1993 and present study) point to such trade-offs in crosslinguistic comparison. Such crosslinguistic trade-offs or balancing effects gave rise to the attractive idea of something like an equal overall complexity in all our natural languages. We agree with those arguments (e.g. Miestamo, this volume) saying that there is no possibility to verify such a hypothetical equality and would like to stress the fact that this idea of equality is in no way supported by those correlations pointing to balancing effects. Let us illustrate this with an example with only two parameters:
A low-budget University department regularly records the number of printouts and copies per individual member. (Some are proud of their "productivity", others of their "economy".) A "cross-subject" negative correlation between the number of printouts and copies indicates a balancing effect: The more copies, the fewer printouts, and vice versa. But this correlation does not at all mean that all the members of the department achieve the same sum of copies plus printouts. It is fully compatible with some members producing both far more printouts and far more copies than others. We may conclude: Not even very clear crosslinguistic balancing effects can be interpreted in the sense of an equal overall complexity in the respective languages. And in language we have not only two and very clearly defined

parameters, but different subsystems whose complexity is measured by different, more or less well defined parameters.

3. Conceptualizations of "semantic complexity" and a look at Pidgin languages

Pidgin and Creole languages are often supposed to be the world's simplest languages (e.g. McWhorter 2001): They show a small lexicon and a low complexity in phonology, morphology, and syntax. We will argue that this is to some degree compensated for by semantic richness of the expressions, i.e. by what we call high "semantic complexity". According to this conceptualization, a large proportion of expressions encoding a large repertoire of different meanings, i.e. a tendency to homonymy and polysemy, may be regarded as indicating high "semantic complexity".
But we have to admit that such a concept of *semantic complexity* goes beyond the scope of a rather technical concept of hierarchical steps of complexity. The different meanings that can be assigned to a certain verbal expression can hardly be viewed as the parts of this expression, nor can the expression, strictly speaking, be viewed as a complex of its possible meanings.[6] And relevant expressions need not in any case be viewed as "poly-functional", i.e. as having "inherently" many meanings that are triggered in certain contexts: Gil (2005) observed a "high availability of apparently associational interpretations" of expressions in isolating languages with basic SVO word order. He views the prominence of such an "associational semantics" as a characteristics of simplicity rather than complexity. But the scope of perspectives on semantic complexity seems to be broader (Raukko 2006).

What are the alternatives if one denies "semantic complexity" as operationalized above? Maybe it is generally misleading to talk about *meaning(s)* as something that can be ascribed to an isolated expression instead of something that comes about by the context and context-dependent associations. The last two of the following conceptualisations try to copy with this problem:

1. If one takes a "word" as a unit of a particular sound pattern plus a particular meaning, then the number of words has to be identified – at least in face of the unrelated meanings in homonyms – with the number of different meanings instead of the number of different sound patterns. This would mean a relatively large and highly variable adaptive lexicon in Pidgin languages instead of a restricted lexicon.
2. We relate "semantic complexity" to Simon's (1996) complexity criterion "number of interactions between components". Taking the word or the lexical morpheme as the component, it may be argued that the "different meanings" of this component come about by those linguistic contexts admitting the occurrence of this certain component.[7] These linguistic contexts represent the possible interactions with other components.
3. We say that "semantic complexity" is not an inherent property of the external symbol system "language" but has to be allocated "in the heads" of its users.[8] The efficient use of a Pidgin language demands, even more than a standardized and highly "overlearned" language, high context sensitivity, awareness of the situational context as well as intuitive and fast associative checks and decisions. Using a Pidgin language is more like sailing than driving a heavy motorboat.

And while it may be possible to count or to estimate the number of meanings per expression in highly standardized languages with their large volume lexica and canonical assignments of literal and figural meanings to a certain expression, this will be almost impossible in languages that incessantly produce new meanings and more or less metaphorical applications of their verbal expressions.

Pidgin languages show a tendency to reduce the phonological complexity of expressions from both the substratum language and the superstratum language, a tendency that often results in homophony. An example reported by Todd & Mühlhäusler (1978:11): In some idiolects of Cameroon Pidgin the word *hat* has acquired four different meanings. This homophone originates from a simplification of English phonemes and syllable structures. It is the result from a sound merger of the English words *heart*, *hot*, *hurt*, and *hat*.
Pidgin languages have a rather small vocabulary from the very beginning, and amalgamations of the sort illustrated above are likely to keep the small lexicon small – at least at the sound level (see alternative 1 above)! The single entries into a small lexicon generally incline to both a frequent use and polysemy. (The association between frequency and polysemy is a well known phenomenon since Zipf 1949). This inclination to homonymy and polysemy and to an accumulation of more or less "figural meanings" goes hand in hand with or is favoured by the creation of non-conventionalized ad-hoc metaphors and by the adoption of idioms from both the superstratum and the substratum language.
Polysemy and especially homonymy, because of the semantically more distant values associated with one word (c.f. Raukko 2006:358), may be regarded as contributing to "semantic complexity": In order to be effective, the language counterbalances its simplicity in the lexicon with complexity in semantics, i.e. in a creative and flexible accumulation of context-specific meanings. Viewing our alternative possible operationalizations one might allocate complexity not in the external symbol system but "in the heads" of the users and in their "mental navigation" between the cultural backgrounds of superstratum and substratum language. All this means higher cognitive costs, in particular if one assumes that homonyms and polysemous words have to be stored and memorized together with different possible contexts. And it would boil down to compensatory effects not between or within subsystems of language but, instead, between external and internal, i.e. mental representations.
Pidgin languages are languages in statu nascendi and do not show the "overall complexity" of more developed languages. Nevertheless, we can state at least two balancing effects within Pidgin languages. Firstly, the relatively small lexicon and the low complexity in phonology go hand in hand with a high "semantic complexity", i.e. with a tendency to homonymy, polysemy, and non-conventionalized metaphors. Semantic complexity in this sense will grow tremendously in more or less conventionalized expressions of Pidgin languages reflecting extremely different cultural backgrounds. The following examples by Todd & Mühlhäusler (1978) may illustrate this:

Cameroon Pidgin (p11):
(1) *Wash bele*
wash belly
'the last child'
Tok Pisin (pp 24f):
(2) *Han bilongan I nogut*
Hand POSS is not good
'she is menstruating'
(3) *karim lek*
to carry legs
'a form of courtship in the new Guinea highlands'

Secondly, the relatively small lexicon and the low complexity in phonology go hand in hand with a higher complexity of words in terms of n of syllables. As compared with English, which is the superstratum of many Pidgin languages, those Pidgin languages exhibit a higher proportion of bisyllabic words (Hall 1966, Heine 1973). This tendency to bisyllabic words means, unlike an increase in "semantic complexity", a higher complexity in a rather technical

sense of the term. As syllable structures are simplified, the number of syllables per word increases. (See correlation (a) between n of phonemes per syllable and n of syllables per word in Section 2.)

4. Low complexity in word structure – high semantic complexity and idiomatic speech?

How can semantic complexity, as conceptualised above, be localised within the language system? We think that low morphological complexity especially in word structure favours a higher semantic complexity – e.g. a tendency to homonymy and polysemous expressions encoding a higher number of senses – which in turn requires and favours a rather formulaic speech, i.e. stable fragments of speech that allow a quick identification of the context-relevant meaning. Therefore the context should be stored and memorized together with the homonymous and polysemous words. A high proportion of idioms and a tendency to formulaic speech increases the cognitive costs in the acquisition of the respective language. As to these new assumptions we will refer to statistical information supporting our arguments but cannot offer crosslinguistic *correlations* in the sense of inferential statistics. Instead we will substantiate these assumptions by taking English as an example and by a contrastive comparison between English and Russian.

4.1 English as an example

English has a high number of monosyllabic words, roughly 8000 according to Jespersen (1933). And a high number of very short, i.e. monosyllabic words, is associated with a high number of syllable types (our correlation I) and with a large phonemic inventory. All these are indications pointing to a high phonological complexity in English. Monosyllabic words are not suitable for coding many morphological categories. This amounts to a low morphological complexity.

Languages with a high number of monosyllabic words tend, moreover, to have a higher number of homonyms. According to Jespersen (1933) there are about four times more monosyllabic than polysyllabic homonyms: "The shorter the word, the more likely is it to find another word of accidentally the same sound". Homonymy affects of course also "parts of speech" and most of the "grammatical homophones" such as *love* (verb, noun) or *round* (noun, adjective, adverb, preposition, verb) are again monosyllables. Because of the well known association between frequency and polysemy on the one hand and frequency and shortness on the other, polysemy should also be a frequent phenomenon in monosyllabic words. Both homonymy and polysemy may be viewed, as already mentioned, as dimensions of semantic complexity.

4.1.1 English phrasal verbs - monosyllabic and idiomatic!

Phrasal verbs may be considered as a special case of idioms. They consist of a verb in connection with an adverb and/or preposition such as *get by, get along with, get in, get over, act on, act up.* Phrasal verbs are idiomatic, because their meaning cannot be derived from the meaning of each word separately. The verb as well as the adverb or preposition forming the phrasal verb are often polysemous. But in combination they are quite unambiguous, despite the fact that the phrasal verb may again have more than one idiomatic meaning, as e.g. in *go for* and *set off.*

In a short analysis of a collection of 1406 English phrasal verbs[9] we found that 1367 or 97 % of the verbs that were part of the phrasal verb construction were monosyllabic. (39 phrasal verbs included a bisyllabic verb and only one was found with a trisyllabic verb.)

Phrasal verbs have to be memorized holistically. One has to know rules concerning their separability, e.g. *add up* (separable) versus *get around* (inseparable). In case of separability one has to know in addition that a pronominal direct object must always be put between verb and preposition.

4.2 A short comparison between English and Russian

The tendency to short words, to polysemy and homonymy and to idiomatic speech becomes a distinct characteristic of English especially in comparison with Russian. Relevant data counted and reported by Polikarpov (1997) perfectly match with our model:

Word length:	Russian words are on average 1.4 times longer than English words
Polysemy:	English words have on average 2.7 meanings, Russian words only 1.7
Homonyms:	in English at least 2 000, in Russian about 500
Idioms :	in English roughly 30 000, in Russian about 10 000

Do the tendencies to idiomatic speech, and to rigid word order in general, mean a higher or a lower complexity? This is the central question to be discussed in the following section.

5. Final discussion

The tendency to idiomatic speech marked the endpoint of several of our chains of arguments so far. Taking English as an example: high phonological complexity - low morphological complexity - high semantic complexity - rigid word order and idiomatic speech. What we have avoided so far, because of conflicting positions between the first and second author, is an answer to the question whether rigid word order and idiomatic speech indicate high or low complexity. Let us start with the first author's position that is illustrated in Figure 4:
High phonological complexity, e.g. a large number of syllable types, is associated with a tendency to monosyllabism. Monosyllabic words are not suitable for coding many grammatical morphemes (i.e. isolating morphology); this means low morphological complexity and a low number of morphological rules. Furthermore, monosyllabism is strongly associated with lexical and part-of-speech ambiguity. This means high semantic complexity. To keep the language system efficient, grammatical ambiguity requires or favours rigid word order. Lexical ambiguity requires collocations for resolving homonymy and polysemy, etc. All this results in a higher word order complexity: More word order rules, more lexical collocation rules, formulaic speech, and idioms such as e.g. phrasal verbs.

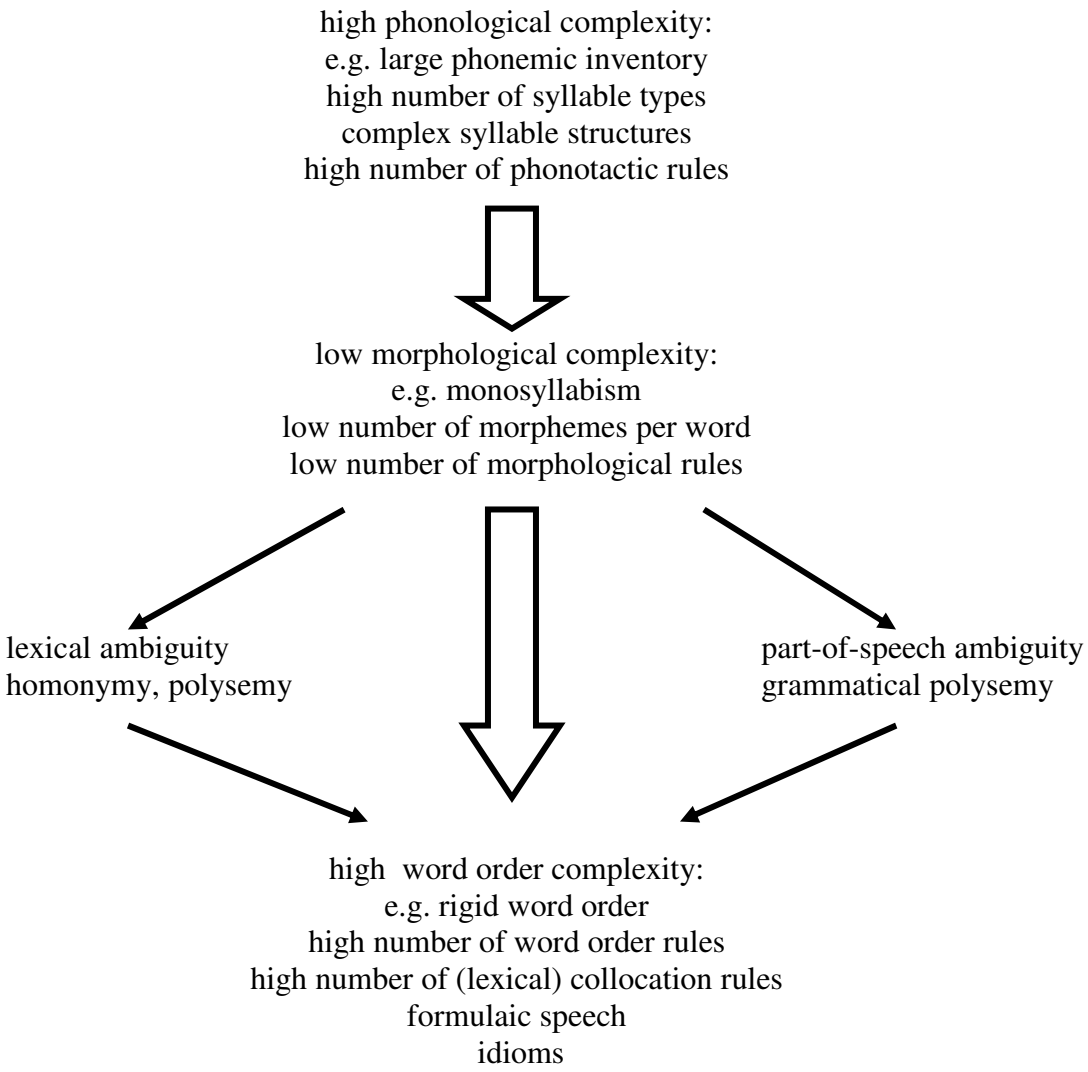

Figure 4: Illustrates complexity trade-offs between the subsystems phonology, morphology, semantics, and syntax.

Information theory, however, offers a somewhat different approach. This approach is, in the opinion of the second author, fully compatible with the descriptions of balancing effects in the earlier sections of this paper. All the following properties contribute to "rigid word order" in a more general sense: formulaic speech, including idiomatic speech and phrases, as well as "rigid word order" in a more specific sense, i.e. in the sense of those rules we know e.g. from English: The verb always comes after the subject; in questions one has to put the auxiliary before the subject, etc.

Most linguists would attribute high complexity to rigid word order, at least to rigid word order in the more specific sense of the term. But according to the "logic" of the balancing effects hypothesized above it should be associated with low complexity: The meaning of the individual words that constitute an idiom differs from the dictionary definitions of these

words and depends on the meaning of the whole of which it is a part. In other words: This whole is a "prefabricated" (Wray & Perkins 2000:1) series of individual words, and the contribution of the individual word is specified by this series. Given a sufficient familiarity of the idiom, no other meanings will be associated with the individual words: The prefabricated linguistic context effectively "selects" or specifies the meaning within the phrase. Thus we may say that the high redundancy - or low complexity? - of highly overlearned idioms is a very selective filter that allows high complexity in word semantics.

This view is actually consistent with information theory. Rigid word order boils down to high redundancy, high predictability, low informational content. Given a fragment of a rather redundant series it is relatively easy to anticipate or to reconstruct the whole. The number of errors that are made by native speakers in the guessing game technique (Shannon 1951) will be rather low in redundant series. This technique can only measure, though this was not intended by Shannon, the subjective information, i.e. the information a text contains for the specific guessing subjects. A text that is highly redundant (for a highly competent speaker) is without any doubt simple (for the highly competent speaker), and not at all complex.

A low level of complexity of the sequence of elements and "supersigns" may come about by a very low number of rules. Very few and very simple rules (e.g. exclusively CV-syllables, bisyllabic words and VO order!) may lead to an extremely high redundancy of the string. In such a language we will observe low complexity in both the string and the production rules (the "source"). But a text of a comparably high redundancy and low complexity may also be produced by a huge number of less restrictive rules. In such a language the text remains simple and redundant for the highly competent speaker, but the "production system" may be regarded as highly complex according to Gell-Mann's criterion "number of rules". Complexity in this sense will show in the process of language acquisition, especially in Second Language Acquisition (SLA), and in the attempts of linguists to extract the phonological, morphological and syntactical rules of the respective language. A third virtual language may also have a huge number of rules, being highly complex in this sense, but these rules are in part less rigorous, more concerned with aesthetic and pragmatic principles that allow considerable freedom in sentence construction, while other rules of this language may be rigorous but only of "local" effectiveness. (Higher freedom in word order does not mean randomness and does not necessarily mean a lower number of rules determining word order. This virtual language requires, on the contrary, many additional "stylistic" rules that decide in cases of conflicts.) Such a language is complex in both the production rules and the "surface".

From this point of view, the trade-off is between "semantic complexity" on the one hand and "word order complexity" on the surface on the other, independent of the number of rules contributing to word order complexity: High complexity in word semantics requires or favours low complexity (high rigidity, high redundancy) in word order and vice versa. This would suggest the following modification of the succession illustrated in Figure 4: high phonological complexity - low morphological complexity - high semantic complexity - *low* complexity (high rigidity) in word order.

6. Concluding remarks

Functional explanations of all the complexity trade-offs reported or hypothesized above refer to economy principles in the usage of language. The essence of such explanations (e.g. Fenk-Oczlon & Fenk 1999, 2002) is a tendency of natural languages to keep the size of clauses and the information flow within these clauses rather constant. This tendency forces complexity trade-offs between those units (syllables, monosyllables and polysyllabic words, clauses and "mono-clausal" sentences…) analyzed in terms of phonology, morphology, and syntax. Our new set of significant crosslinguistic correlations indicates such balancing effects. It suggests

trade-offs between facets of phonological complexity and morphological complexity but by no means supports, as could be demonstrated at the end of Section 2.4, the idea of an equal overall complexity in natural languages.

Taking Pidgin languages as an example, an attempt was made to conceptualise *semantic complexity* and to relate it to complexity in phonology and morphology. As to the simplicity versus complexity of rigid word order – "rigid word order" in a broader sense – we could at least outline two arguable positions for a necessary future discussion. Both lines of argumentation and all our other empirical and theoretical arguments suggest the view of complexity trade-offs between rather than within the subsystems of language.

---

[1] We would like to thank the referees and editors for their helpful suggestions.

[2] The main database for this statistical reanalysis originates from a quasi-experimental study by Fenk-Oczlon (1983): Native speakers of 27 typologically different languages were instructed to translate a set of 22 German "mono-clausal" sentences into their mother tongue and to determine the n of syllables of each of the sentences produced. The written translations (in facsimile on the pages 104 – 182 of this study) allowed to enumerate the n of words per sentence. The number of phonemes was determined with the help of the native speakers and grammars of the respective languages. The experiments were continued (29 languages in our 1993-study, 34 in the 1999-study). Due to research grants (Fulbright and University of Klagenfurt) the size of the sample will arrive at 65 in the near future.

[3] In the case of two given correlations r_{xy} and r_{xz} with the same sign (a positive sign in both correlations or a negative sign in both correlations) a third "correlation" between Y and Z, i.e. any coefficient r_{yz} different from zero, will rather show a positive sign. In cases of different signs in the two given correlations we rather have to expect a negative sign in the third correlation. The higher the correlations r_{xy} and $r_{xz,}$, and the higher therefore the determination coefficients r^2 and the variance explained by them, the higher the plausibility of the assumption of a correlation r_{yz}. For a more detailed discussion of this sort of statistical reasoning see Fenk-Oczlon & Fenk (2005) and Fenk & Fenk-Oczlon (2006).

[4] In languages showing a high proportion of monosyllables we encounter, apart from monosyllabic function words, also a high number of monosyllabic content words. Especially languages with such a manifest tendency to monosyllabism will show a tendency to isolating techniques (not necessarily the other way round), high syllable complexity and high variability of syllable types.

[5] This significant result could already be presented at the 2005 Helsinki Symposium on Approaches to Complexity in Language.

[6] In this respect we fully agree with Meyer (in press). He questions (in footnote 8) several applications (e.g. in Cramer 2005) of Menzerath's law: "Can the different meanings of a polysemous lexeme really be treated as 'constituents' of the lexeme?"

[7] This conceptualisation is inspired by ideas of Wittgenstein, the others are more or less in line with constructivistic or otherwise mentalistic conceptualisations. A separate article would be necessary to investigate such attributions.

[8] According to Evans (2005:34) "semantic structure derives from and mirrors conceptual structure /…/ Hence linguistic polysemy reflects complexity at the level of mental representation".

[9] (http://usingenglish,com), 22.02.2006

References

Altmann, G. 1980. Prolegomena to Menzerath's Law. *Glottometrika* 2: 1-10.

Bett, S. (1999). Can we pin down the number of phonemes in English? *Newsletter* April 1999 http://victorian.fortunecity.com/vangogh/555/Spell/phon-inv-art2.html, 12.10.05.

Campbell, G. L. 1991. *Compendium of the World's Languages*. London: Routledge.

Changizi, M.A. 2001. Universal scaling laws for hierarchical complexity in languages, organisms, behaviors and other combinatorial systems. *Journal of Theoretical Biology* 211: 277-295.

Cramer, I.M. 2005. Das Menzerathsche Gesetz. In *Quantitative Linguistik/Quantitative Lingusitics*, R. Köhler, G. Altmann, and R.G. Piotrowski (eds), 659 – 688. Berlin/New York: Walter de Gruyter.

Evans, V. 2005. The meaning of *time*: polysemy, the lexicon and conceptual structure. *Journal of Linguistics* 41: 33-75.

Fenk, A. and Fenk-Oczlon, G. 1993. Menzerath's Law and the constant flow of linguistic information. In *Contributions to Quantitative Linguistics,* R. Köhler and B. Rieger (eds), 11-31. Dordrecht: Kluwer Academic Publishers,

Fenk, A. and Fenk-Oczlon, G. 2006. Crosslingusitic computation and a rhythm-based classification of languages. In *From Data and Information Analysis to Knowledge Engineering,* M. Spiliopoulou. R. Kruse, C. Borgelt, A. Nürnberger, and W. Gaul (eds), 350 – 357. Berlin/Heidelberg: Springer.

Fenk, A., Fenk-Oczlon G., and Fenk L. 2006. Syllable complexity as a function of word complexity. In *The VIII-th International Conference "Cognitive Modeling in Linguistics"*, Vol. 1, V. Solovyev, V. Goldberg, and V. Polyakov (eds), 324 – 333. Kazan: Kazan State University.

Fenk-Oczlon, G. 1983. *Bedeutungseinheiten und sprachliche Segmentierung. Eine sprachvergleichende Untersuchung über kognitive Determinanten der Kernsatzlänge.* Tübingen: Gunther Narr.

Fenk-Oczlon, G. and Fenk, A. 1999. Cognition, quantitative linguistics, and systemic typology. *Linguistic Typology* 3: 151–177.

Fenk-Oczlon, G. and Fenk, A. 2002. The clausal structure of linguistic and pre-linguistic behavior. In *The Evolution of Language out of Pre-Language,* T. Givon and B. F. Malle (eds), 215-229. Amsterdam/Philadelphia: John Benjamins.

Fenk-Oczlon, G. and Fenk, A. 2005. Crosslinguistic correlations between size of syllables, number of cases, and adposition order. In *Sprache und Natürlichkeit. Gedenkband für Willi Mayerthaler*, G. Fenk-Oczlon and C. Winkler (eds), 75–86. Tübingen: Narr.

Gell-Mann, M. 1995. What is Complexity? *Complexity* 1: 16–19.
www.santafe.edu/~mgm/**complexity**.html

Gil, D. (2005). How complex are isolating languages? Symposium "Approaches to Complexity in Language", University of Helsinki, August 24-26, Abstract 23-24.
http://www.ling.helsinki.fi/sky/tapahtumat/complexity/complexity.shtml

Hall, R.A. 1966. *Pidgin and Creole Languages.* Ithaca: Cornell University Press.

Heine, B. 1973. *Pidgin-Sprachen im Bantu-Bereich*. Berlin: Dietrich Reimer.

Jespersen, O. 1933. Monosyllabism in English. *Linguistica* In *Selected Writings of Otto Jespersen (no year)*, 574-598. London: George Allen & Unwin LTD.

Juola, P. 1998. Measuring linguistic complexity: the morphological tier. *Journal of Quantitative Linguistics* 5(3): 206-13).

Ladefoged, P. 2001. *Vowels and Consonants: An Introduction to the Sounds of Language*. Oxford: Blackwell.

Lehmann, W. 1978. English: A characteristic SVO Language. In *Syntactic Typology*, W. Lehmann (ed), 169- 222. Sussex: The Harvester Press.

McWhorter, J. H. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology* 5(2/3): 125-166.

Maddieson, I. 1984. *Patterns of Sounds*. Cambridge: Cambridge University Press.

Maddieson, I. 2006. Correlating phonological complexity: Data and validation. *Linguistic Typology* 10: 106-123.

Menzerath, P. 1954. *Die Architektonik des deutschen Wortschatzes*. Hannover/Stuttgart: Dümmler.

Meyer, P. 2007. Two semi-mathematical asides on Menzerath-Altmann's law. In *Exact methods in the study of language and text: dedicated to Professor Gabriel Altmann on the occasion of his 75th birthday,* P. Grzybek & R. Köhler (eds). New York: Mouton de Gruyter.

Meyer, P. in press. Normic Laws in Quantitative Linguistics. In P. Grzybek (ed), *The Science of Language. Structures of Frequencies and Relations.*

Miestamo, M. This volume. Grammatical complexity from a cross-linguistic point of view.
Mulaik, S.A. 2001. The curve-fitting problem: An objectivist view. *Philosophy of Science,* 68: 218 – 241.
Polikarpov, A.A. 1997. Some factors and regularities of analytic/synthetic development of language system. Paper presented at the XIII Int. Conference on Historical Linguistics, 10-17 August Düsseldorf. http://www.philol.msu.ru/~lex/articles/fact_reg.htm, 13.02.06.
Raukko, J. 2006. Polysemy as Complexity? In *A Man of Measure: Festschrift in Honour of Fred Karlsson on his 60th Birthday*. *SKY Journal* 19*:* 357-361.
Shannon, C. E. 1951. Prediction and entropy in printed English. *Bell Syst. Techn. J.* 30: 50-64.
Simon, H. A. 1996 (1962). The architecture of complexity: hierarchic systems. In *The Sciences of the Artificial*. H.A. Simon, 183-216. Cambridge, Mass.: MIT Press.
Todd, L. and Mühlhäusler, P.1978. Idiomatic expressions in cameroon Pidgin English and Tok Pisin. *Papers in Pidgin and Creole Linguistics* 1: 1-35.
Wray, A. & Perkins, M.R. 2000. The functions of formulaic language: an integrated model. *Language & Communication* 20: 1-28.
Zipf, G, K. 1949. *Human behaviour and the principle of least effort. An introduction to human ecology.* Cambridge, Mass: Addison-Wesley.