Gabriel Altmann and Thorsten Roelcke*

Morphological complexity of the word

Abstract: Although there is much linguistic work concerning morphological complexity of words no study tries to scale their physical form. In this paper firstly we present a scaling of word's morphological complexity considering reduplication, compounding, derivation, inflection and suppletivism. Secondly we show some quantitative analyses of morphological complexity with respect to arc, motif, and distances of words in a Slovak poem as an example.

Keywords: morphology, complexity, arc, motif, distance

DOI 10.1515/glot-2015-0002

1 Introduction

In quantitative linguistics one tries to operationalize and quantify every language property. Since properties have degrees, it must be, in principle, possible to quantify also the morphological complexity of words. There is a number of procedures both general and applied to language, used for different purposes, mostly for typology (Anderson 2012, Bane 2008; Bickel, Nichols 2005; del Prado Martín, Kostić, Baayen 2004; Gell-Mann 1995; Goldsmith 2001, 2006; Juola 1998, 2007; Lempel, Ziv 1976; McWhorter 2001; Nichols 2007, 2010; Pellegrino, Coupé, Marsico 2007; Shosted 2006; Siegel 2004; Vulanović 2007) but none of them tries to scale directly the physical form of the word. A quite simple way is to count the number of morphemes making up the word, however, this does not express the complexity but rather the length of the word measured in terms of morpheme numbers. In order to express the complexity one must perform a scaling of morphemes and morphological procedures. This can be based on criteria taken from grammar, semantics, or from the history of language development, or from all together. It is almost sure that every linguist would perform this scaling differently, not only because languages differ in their morphology but also because of

***Corresponding author: Thorsten Roelcke:** Fachgebiet Deutsch als Fremdsprache, Technische Universität Berlin, Sekr. HBS 2, Hardenbergstraße 16–18, D–10623 Berlin.
E-mail: roelcke@tu-berlin.de
Gabriel Altmann: Stüttinghauser Ringstraße 44, D–58515 Lüdenscheid.
E-mail: RAM-Verlag@t-online.de

different conceptions of grammar and the aim of the study. Needless to say, in every grammar there are cases, which are not unequivocally ascribable to individual classes.

It is to be noted that an analogical scaling could be set up also for parts-of-speech or other word classes, or for valencies of verbs, or for the degree of predication, though nobody tried to do it so far. This is caused by the fact that there are many criteria which could (or must) be used in order to be applicable to all languages. Nevertheless, one could begin with one language and evaluate some texts.

Here, our aim is to perform a preliminary scaling of words according to their grammatical status and the involved morphological procedures, in order to be able to rewrite a text in numbers expressing the degree of morphological complexity of words. The study of morphological complexity is a problem attracting linguists since generations (and schools) but there is no attempt at scaling it, as far as we know. In the present contribution we shall try to set up a preliminary scaling which can be completed by phenomena occurring in other languages.

2 Scaling morphological complexity

We start from the conjecture that in the evolution of language (also in language learning) the following procedures are introduced one after another: *stem* < *reduplication* < *compounding* < *derivation* < *inflection* < *suppletivism*. Thus the scaling can lean against this sequence (cf. Table 1).

The simplest morphological entity existing in all languages is a *stem* or *root*, hence it obtains degree 1. In the evolution of language which according to G. K. Zipf (1949) performs grand cycles, first the roots are made more complex, then they are simplified again and a new cycle begins. Root is part of each word belonging to whatever word class.

The simplest construction is the *reduplication* adding nothing new to the root. One can ascribe it degree 2. Reduplications exist perhaps in all languages, e.g. Indonesian *kadang-kadang* (sometimes) and the plural of nouns (or rather "many ...") has the form of reduplication. However, there are reduplications with variation of some phonemes, e.g. Indonesian *bolak-balik* (to and fro), Slk. *krížom-krážom* (through and through), or also written together, e.g. Hungarian *tarka-barka* (topsy-turvy). Thus the simple reduplication obtains degree 2.1, that with variation 2.2, that written together (*hophop*) degree 2.3 and that written together but with variation 2.4. Variation itself could be further scaled according to its extent (one phoneme, two phonemes, etc.). Nevertheless, there will be a number of cases whose degree must be decided separately e.g. Indonesian *gelap gulita*

(dark) which may be considered also compound though *gulita* does not occur in the dictionary as a separate item.

The next level is compounding, i.e. concatenation of two stems which are "felt" as a whole or can be statistically (cf. Ziegler, Altmann 2002: 92ff.) or grammatically declared as belonging together. Since we analyze only written texts, we can evaluate the strength of cohesion of compound parts using the written form. Fan and Altmann (2007) distinguished 29 different cohesion levels which can be subdivided in five classes and scaled within these classes. We use the examples but rescale the domain.

Level 3 contains compounds whose components are interchangeable, i.e. have the weakest cohesion. But even with **interchangeability** there are slight differences. At level 3.1 are the compounds with blank, e.g. technical expressions *rubber scrap* = *scrap rubber* or *filter candle* = *candle filter*. At level 3.2 we have hyphenized interchangeable compounds, e.g. German *schwarz-weiß* = *weiß-schwarz*; at level 3.3. those which have in addition to the hyphen a fugue e.g. *mathematic**o**-physical* = *physic**o**-mathematical*. At level 3.4. there are interchangeable stems written together and joined by a juncture, e.g. Slovak *červen**o**biely* = *biel**o**červený* (red-white). At level 3.5 there are phrasal constructions which are not yet petrified. In German one can say *Kultur- und Naturschutz* or *Natur- und Kulturschutz* (culture and nature protection). Different languages (those not using alphabetic script) can mark the compounds in a quite different way.

The compounds whose elements are **not interchangeable** but are written with a **blank** can stay at level 4. They signal a stronger cohesion than those at level 3. Fan and Altmann (2007) mention the following categories: 4.1. There is no juncture between the components, e.g. *space flight, fall velocity,* Indonesian *kereta api* (train = car fire). (Needless to say, if one of the components of the compound has further subcomponents, they must be evaluated, e.g. in *space flight*, both stems, their compounding and the derivation in *flight* must be evaluated). 4.2. Phrasal constructions considered already compounds e.g. German *immer und ewig*, French *come il faut*. 4.3. The components are joined by a preposition but the first component does not have inflection, e.g. *Bruno von Kuckucksheim*. 4.4. The components are joined by a preposition or a case marker and there is internal inflection possible, e.g. *velocity(ties) of fall, King(s) of England*, French *chemin(s) de fer*, Italian *faro(fari) a raggio largo*. German *Haus(Häuser) der Jugend*. 4.5. The first component has an internal inflection signalling some grammatical category, e.g. German *fahrbahr**er** Trockenlöscher* (trolley-mounted dry-powder dispenser) in which *-er* in *fahrbahrer* marks nominative, singular, masculine. 4.6. The first component has a morphological juncture, e.g. *techn**ical** worker, fall**ing** velocity, scient**ific** paper.* Needless to say, level 4 can be further refined by cases from other languages.

5. Many compounds are written with **hyphen** which demonstrates a stronger cohesion. Here we can distinguish 5.1. Hyphenized compound without juncture, e.g. *Baden-Württemberg*. 5.2. Hyphenized phrases containing no preposition, e.g. *forget-me-not, pen-and-ink test, ready-to-operate*. 5.3. Compound with phonological or morphological juncture, e.g. *ethno-musical*. 5.4. Prepositional constructions allowing internal inflection, e.g. *mother-in-law* (pl. *mothers-in-law*). 5.5. Compound with internal inflection, e.g. G. *Saure-Gurken-Zeit*, gen. *Sauren-Gurken-Zeit*, but *Saure* itself contains a grammatical marker.

6. A still stronger cohesion have compounds **written together** (without blank or hyphen). Here we can distinguish five degrees: 6.1. Compounds made up by reduplication e.g. *Bonbon, Filmfilm* which are neologisms but can be considered also as belonging to degree 2. 6.2. Compounds made up without any change of components, e.g. *Hausmeister* (janitor), *redbreast*. 6.3. Compounds with internal phonological or morphological juncture, e.g. G. *Therm**o**meter, Kind**er**garten*, Slovak *bel**o**hlavý*. 6.4. Compounds allowing internal inflection, e.g. G. *Hohepriester* (high priest), gen. *Hohe**n**priesters*. 6.5. Phrases which became compounds e.g. G. *Habenichts* (a person having nothing), *Tunichtgut* (good-for-nothing).

7. The strongest cohesion in compounding can be supposed if one or more components of the compound are **abbreviated** in some way. Here we can set up a scale as follows: 7.1. One of the components gets shorter, e.g. G. *Pauschbetrag* < *Pauschalbetrag* (lump sum), *zum* < *zu dem* (to the), *Vergissmeinnicht* < *vergiss meiner nicht* (forget-me-not), *Kurlaub* < *Kur Urlaub*, Toba Batak *karetapi* < *kareta api* (train). 7.2. Both components get shorter, e.g. G. *am* < *an dem, im* < *in dem*, *Kripo* < *Kriminalpolizei, Josta-Beere* < *Johannisbeere* + *Stachelbeere*.

7.3. One of the components is maximally abbreviated to its first letter, e.g. G. *S-Bahn* < *Stadtbahn, U-Boot* < *Unterseeboot*. 7.4. Both (all) components are maximally abbreviated: *USA, EU, CIA*. For many compounds of this level the etymology is not known to a „normal" speaker, e.g. G. *Montag*, E. *Tuesday*, etc. One must decide whether they can be considered monomorphematic or inserted in some of the levels. Mostly they are names of days or months.

A still stronger cohesion is displayed by **derivation** which arose later in the evolution. A component may loose its independence but not completely. Its status may approach that of an affix. Hence we obtain different degrees *of derivation*: 8.1. Quasiderivation with uncertain status of one of the components, e.g. G. *Schulwesen*, or *Flugzeug*. Both affixoids *-wesen* and *-zeug* exist also independently with their own meanings but in compounds they merely play the role of an affix. 8.2. Clitic written either together with the word or joined with a hyphen or written separately, e.g. Indonesian *-kah, -lah*, Japanese *-ka*. It depends on the consent in modern grammar whether something is considered clitic or not, e.g. in Czech *se, si* written separately. There is a number of such morphemes in lan-

guages and there is no general prescription. 8.3. Affixes that exist also independently, e.g. as prepositions or detachable prepositions like in German (*vor-*, *auf-*, *durch-*, ...), negation in Slovak (*nie* > *ne-*) and Hungarian (*nem* > *ne-*). 8.4. Not detachable affixes like G. *be-*, *ge-*, *ent-*, *-heit*, E. *-ity, -ness*, Indonesian *-an, ber-*, representing pure derivation. 8.5. If several affixes can be attached to the root, then the affix at the last position may obtain value 8.51, the last but one 8.52, etc. This sub-scaling may be relevant for strongly agglutinating languages. In the Hungarian word *legmegszentségteleníthetetlenebb* the individual affixes may contribute in different degree to the complexity (the root is *szent*); the first and the last affix belong together and express the superlative, making the scaling still more complex.

Inflections expressed by special morphemes or variation of roots are the last level displaying the strongest morphological complexity. They usually express grammatical categories. Zero morphemes are not counted. This level obtains the degree 9. Gradation is considered inflection (*tall – taller – tallest*). Slavic and Hungarian infinitives are considered here inflections, just as all other verbal forms. However, even here a scaling is possible if we consider introflection a greater change of the word than e.g. final inflection. Hence the German *Väter*, *Mütter* (fathers, mothers) would be a stronger effect than *Kinder* (children). If we consider the phonetic form, then even in *child – children* there is an introflection + inflection just as in the German dat. pl. *Vätern*. A more specific gradation may be made with respect to the distinction of inflection and introflection in combination with this of regular and irregular verbs (e.g. in German): Here we may define an increasing complexity (Roelcke 2011: 41–43): 9.1 regular weak inflection (adding inflectional morphemes): *sagen / sagte / gesagt*; 9.2 regular strong inflection resp. introflection (e.g. varying root's vocalism in German "Ablaut"): *sprechen / sprach / gesprochen*; 9.3 irregular inflection resp. introflection (e.g. varying root's vocalism and consonantism in German): *denken / dachte / gedacht*.

The highest level is that of **suppletivism**. Here nothing recalls the original stem, e.g. G. *Mensch – Leute, gut – besser*, It. *buono – meglio – ottimo*. Hence suppletivism obtains degree 10. Some personal pronouns in plural and possessive pronouns may be considered suppletivisms. In many languages the forms of the verb *be* are suppletive.

A survey of all these degrees is presented in Table 1. Needless to say, in other languages one may find further intermediate complexities which can be inserted in this scaling. And in languages having non-alphabetic script a quite different scaling may result. We do not take into account the capability of the stem to form derivates, compounds or take on inflections (checking a whole corpus) and are not concerned with the information conveyed by individual entities.

Table 1: Scaling morphological complexity of words

Degree	Type	Example
1	**Stem/Root**	*house*
	Reduplication	
2.1	simple reduplication with hyphen	Indon. *kadang-kadang*
2.2	reduplication with phonetic variation with hyphen	Indon. *bolak-balik*
2.3	reduplication written together	*hophop*!
2.4	Reduplication written together with variation	Hungarian: *tarkabarka*
	Compounding with interchangable stems	
3.1	interchangeable stems with blank	*rubber scrap* = *scrap rubber*
3.2	hyphenized stems, interchangeable	G. *schwarz-weiß* = *weiß-schwarz*
3.3	hyphenized, interchangeable stems and a juncture	*mathematic**o**-physical*
3.4	compounds written together with interchangeable stem and a juncture	Sk. *modr**o**biely* = *biel**o**modrý*
3.5	phrasal, interchangeable stems with blank	G. *Kultur- und Naturschutz* = *Natur- und Kultuaschutz*
	What not interchangeable stems	
4.1	Compounds with simple apposition	*space flight, fall velocity,* Indonesian *kereta api* (train = car fire)
4.2	Phrasal constructions considered already compound	G. *immer und ewig*, F. *come il faut*
4.3	Compound with preposition, one element not inflected/inflectable	G. *Bruno von Kuckucksheim*
4.4	Components joined by a preposition or case marker, internal inflection possible	*velocity (ties) of fall,* G. *Haus der Jugend*.
4.5	The first component with a morphological juncture	*techn**ical** worker, fall**ing** velocity, scienti**fic** paper*
4.6	The first component inflected	*G. fahrbahr**er** Trockenlöscher*
	Hyphenized compounding	
5.1	Hyphenized compound without juncture	G. *Baden-Württemberg*
5.2	Hyphenized phrases containing no preposition	*forget-me-not, pen-and-ink test, ready-to-operate*
5.3	Hyphenized compound with phonological or morphological juncture	*ethn**o**-musical*
5.4	Hyphenized compound with internal inflection	G. *Saur**e**-Gurken-Zeit* gen. *Saur**en**-Gurken-Zeit*
5.5	Hyphenized prepositional constructions allowing internal inflection	*mother-in-law* (pl. *mothers-in-law*)

Table 1 (cont.)

Degree	Type	Example
	Compounds written together	
6.1	Reduplicated neologism	*Bonbon, Filmfilm* (see 2.3)
6.2	Two stems without change	G. *Hausmeister, redbreast*
6.3	With internal phonological or morphological juncture	*Therm**o**meter,* G. *Kind**er**garten* Slovak *bel**o**hlavý*
6.4	Compounds allowing internal inflection	G. *Hohepriester* (high priest), gen. *Hohe**n**priesters*
6.5	Phrases which became compounds	G. *Habenichts* (a person having nothing), *Tunichtgut* (good-for-nothing)
	Abbreviated compounds	
7.1	One of the components gets shorter	G. *Pauschbetrag* < *Pauschalbetrag* (lump sum); *zum* < *zu dem* (to the); *Kurlaub* < *Kur Urlaub*
7.2	Both components get shorter	G. *im* < *in dem*, *Kripo* < *Kriminalpolizei, Josta-Beere* < *Johannisbeere* + *Stachelbeere*
7.3	One of the components is maximally abbreviated to its first letter	G. *S-Bahn* < *Stadtbahn, U-Boot* < *Unterseeboot*
7.4	Both (all) components are maximally abbreviated	*USA, EU, CIA*
	Derivation	
8.1	Quasiderivation with uncertain status of one of the components	G. *Schulwesen*, *Flugzeug*
8.2	Clitic written either together with the word or joined with a hyphen or written separately	Indonesian *-kah, -lah*, Japanese *-ka*, Czech *se*
8.3	Affixes that exist also independently	Detachable prefixes in German; prefixes made of prepositions in Slavic
8.4	Not detachable affixes	G. *be-, ge-, ent-, -heit* E. *-ity, -ness*
8.5	Scaling of affixes according to the distance from the root	E. *Luther/an/ism*
9	**Inflection**	
9.1	Regular weak inflection (adding inflectional morphemes)	*houses,* Sl. *robiť* (to do) G. *sagen / sagte*
9.2	Regular strong inflection (e.g. varying root's vocalism: german "Ablaut")	G. *sprechen / sprach*
9.3	Irregular inflection (e.g. varying root's vocalism and consonantism)	G. *denken / dachte*
10	**Suppletivism**	*good* vs. *better* G. *ist* vs. *war*

Problems will appear in every language having much synthetism. For example, the infinitive – if it exists – can be considered stem, stem + derivation or stem + inflection. Frequently, the nominative of nouns (mostly masculine) has a zero morpheme, hence it can be considered pure root but in other genders having an overt morpheme it is in any case a root with inflection. Languages having a strong degree of analytism are easier to process.

Needless to say, the compounding procedure having the form $R_1 - R_2$ has the complexity 1+5.1+1, i.e. the two roots obtain the value of 1 each and the compounding procedure the value 5.1, hence 7.1.

If one applies this preliminary scale to a text, one obtains a sequence of numbers representing morphological complexity of words, which can be further processed. In many cases one meets problems which must be solved by decision or a new degree must be inserted.

Since using this scaling every complex word is a sum of degrees, in short texts it will not be possible to search for the distribution of complexities. In longer texts perhaps a continuous function will be found expressing some trend but this is a task for future research. Preliminarily, a rank-frequency distribution does not give much sense.

Languages may be "specialized", e.g. having many isolated roots in texts and one or two very frequent formations. In those languages both the time series of complexity measures and the Köhlerian motifs will differ from these constructions in strongly synthetic languages. One should not reduce the above scale, on the contrary, looking at other languages it should be further refined.

3 Data: "Aby spriesvitnela" (Eva Bachletová)

In order to illustrate the scaling procedure we take a Slavic language and analyse the Slovak poem *Aby spriesvitnela* by Eva Bachletová. The poem is written without bound rhythm and is not rhymed. The analysis is presented in Table 2. In the second column one finds the degrees of individual components of individual words; in the third column the added values are presented.

The sequence obtained from these complexities is

C = [18.3, 10, 10, 1, 19, 18.3, 10, 43.4, 18.4, 10, 1, 10, 1, 18.3, 1, 10, 1, 19, 1, 10, 10, 10, 8.4, 26.7, 1, 27.3, 19, 18.3, 18.4, 10, 9.4, 1, 17.7, 10, 10, 10, 10, 9.3, 18.3, 10, 1, 18.3, 26.7, 18.4, 1, 9.4, 10, 10, 35, 1, 1, 25.7, 10, 35, 1, 10, 10, 9.4, 18.3, 17.7]

Evidently, the probability distribution or the Fourier series of these values would be very irregular or complex. But one can compute some indicators.

Table 2: Morphological complexity of words in a Slovak poem

Nemám rada bielu	8.3+1+9, 1+9, 1+9	18.3, 10, 10,
dnes je prízrakom chladu	1, 10+9, 8.3+1+9, 1+9	1,19,18.3, 10,
znecitlivenia	8.3+8.3+1+8.4+8.4+9	43.4,
konečného verdiktu	1+8.4+9, 1+9	18.4, 10,
nad človekom	1, 1+9	1, 10,
nad pocitom	1, 8.3+1+9	1, 18.3,
nad láskou.	1, 1+9	1, 10,
Dnes je tu iná biela	1, 10+9, 1, 1+9, 1+9	1, 19, 1, 10, 10,
biela obrazovky	1+9, 1+8.4+9	10, 18.4,
počítača	8.3+1+8.4+9	26.7,
tam nahadzujeme	1, 8.3+1+9+9	1, 27.3,
svoje vnemy	10+9, 8.3+1+9	19, 18.3,
čiernymi linkami	1+8.4+9, 1+9	18.4, 10,
rýchlo a bezpečne	1+8.4, 1, 8.3+1+8.4	9.4, 1, 17.7,
kreslíme životy	1+9, 1+9	10, 10,
slovami,	1+9	10,
ktoré navždy	1+9, 8.3+1	10, 9.3,
zmenili bielu	8.3+1+9, 1+9	18.3, 10,
odviedli nás	1, 8.3+1+9	1, 18.3,
od základných farieb	1, 8.3+1+8.4+9	26.7,
bytia.	1+8.4+9	18.4,
A možno stačí jedna	1, 1+8.4, 1+9, 1+9	1, 9.4, 10, 10,
nenapísaná veta	8.3+8.3+1+8.4+9, 1	35, 1,
aby „novodobá“	1, 1+6.3+1+8.4+9	1, 25.7,
biela spriesvitnela.	1+9, 8.3+8.3+1+8.4+9	10, 35,
Lebo čistá – biela krehkosť	1, 1+9, 1+9, 1+8.4	1, 10, 10, 9.4,
prichádza potichu …	8.3+1+9, 8.3+1+8.4	18.3, 17.7,

The empirical mean and the variance of the complexities in the poem are

$$\overline{C} = 12.5733, \quad Var(C) = 87.8715, \quad n = 60$$

hence an asymptotic comparison with other texts of the author, other text sorts or other languages would be very simple. It is to be noted that the mean of complexity is here relatively high. For some other poetic texts by E. Bachletová we obtain

Iba neha: $\overline{C} = 8.7291$, $Var(C) = 67.9101$, n = 141
Moje určenie: $\overline{C} = 10.0160$, $Var(C) = 63.6919$, n = 145
Hľadanie odpovedí: $\overline{C} = 11.8972$, Var(C) = 70.8303, n = 36
Dielo Stvoriteľa: $\overline{C} = 11.3609$, Var(C) = 55.2225, n = 133
Sila ľudského ducha: $\overline{C} = 10.7490$, Var(C) = 65.8801, n = 353

In order to state whether the variability of $\overline{C}$ is characteristic only for the author, we may compare pairwise the means using the t-test with $n_1 + n_2 - 2$ degrees of freedom according to the formula

$$t = \frac{\overline{C}_1 - \overline{C}_2}{\hat{\sigma}_{\overline{C}_1 - \overline{C}_2}} \tag{1}$$

where

$$\hat{\sigma}_{\overline{C}_1 - \overline{C}_2} = \sqrt{\frac{\sum_{i=1}^{n_1}(C_{i1} - \overline{C}_1)^2 + \sum_{i=1}^{n_2}(C_{i2} - \overline{C}_2)^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{2}$$

and n_1, n_2 are the numbers of words in the given texts. For the difference of means of *Aby spriesvitnela* and *Iba neha* we obtain $t = 2.90$ with 199 degrees of freedom which is significant; for *Aby spriesvtnela* and *Moje určenie* we obtain $t = 1.98$ with 203 DF which is at the boundary of significance; etc. As can be seen, the mean word complexity is not constant for a language, it varies significantly even in the texts of the same author written in the same text sort. Hence it is more realistic to work with intervals but this will be possible only when a great number of texts have been processed.

For a prosaic work *Čas pre nádych vône* we obtain the vector

C(*Čas pre nádych vône*)
= [1,10,10,18.3,10,10,1,10,1,26.6,9.3,8.3,1,10,1,1,18.3,1,1,10,10,10,1,10,9.4,19,1,10,
1,10,18.4,10,1,1,1,43.9,10,10,1,19,17.3,1,1,1,26.7,18.3,10,10,9.4,18.3,1,10,9.3,10,1,
10,1,10,10,1,10,27.3,10,1,9.3,10,18.3,18.3,9.4,18.4,10,10,10,19,10,19,9.4,10,1,19,
10,1,19,10,1,25.7,1,19,10,1,10,1,1,10,1,10,1,10,17.4,9.4,1,1,26.7,1,1,18.3,19,9.4,10,
18.4,10,1,10,10,10,9.4,18.3,8.2,1,26.4,1,18.4,26.4,10,17.7,10,17.3,1,10,1,10,1,18.4,
18.7,8.2,19,10,10,10,18.3,1,9.3,1,9.3,1,9.4,1,10,1,10,9.3,1,19,1,1,19,19,1,9.4,18.4,1,
10,1,18.4,1,10,1,1,1,1,19,10,18.3,10,9.3,10,10,26.7,1,35,18.4,18.4,1,1,10,10,1,18.4,
10,18.4,1,19,10,19,10,19.4,18.3,18.3,1,18.4,10,1,10,10,10,10,1,19,1,10,19,10,10,1,
19,10,10,18.4,10,18.4,1,10,1,10,10,1,9.4,10,1,1,10,9.4,1,10,18.4,10,1,19,8.2,19,1,
10,10,1,19,19,1,19,18.3,10,19,10,1,18.4,10,10,10,10,19,18.4]

Its mean is $\overline{C} = 9.9396$, $\mathrm{Var}(C) = 57.0922$, $n = 260$, hence it lies between the poetic works. Preliminarily, the mean morphological complexity of Slovak seems to lie between 8 and 12.

4 Arc

If we join the subsequent degree values with straight lines, we obtain an irregularly oscillating curve which could not be captured by a simple Fourier function. It represents rather a fractal line occupying more of the two-dimensional space than a straight line. Both the size of jumps from one value to the next and the length of the sequence of straight lines as well as the distances between equal values and Köhler's motifs (see below) may be used for the characterisation of a text, author, text sort or language. In sciences one uses indicators like Hurst exponent, Minkowski sausage, Lyapunov exponent, etc. (cf. Hřebíček 2000). Since we have short texts, for our purposes it is sufficient to compute the relative length of the arc which sufficiently expresses the extent of oscillation (cf. Popescu, Mačutek, Altmann 2009). Let c_i $(i = 1,2,...,n)$ be the individual complexity degrees of words in a text of length n. Then the arc length between two points is defined as the Euclidean distance

$$L_i = [(c_i - c_{i+1})^2 + 1]^{1/2} \tag{3}$$

and the complete arc from $i = 1$ to $i = n$ as

$$L = \sum_{i=1}^{n-1} [(c_i - c_{i+1})^2 + 1]^{1/2}. \tag{4}$$

The mean arc is simply

$$\bar{L} = \frac{L}{n-1} \tag{5}$$

and the empirical variance of the arc is defined as

$$Var(L) = \frac{1}{n-1} \sum_{i=1}^{n-1} (L_i - \bar{L})^2. \tag{6}$$

The variance of the mean is given as $Var(\bar{L}) = Var(L)/(n-1)$. Thus both the characterisation and the comparison of the morphological complexity of texts is possible using an asymptotic normal test.

For the texts analysed above, we obtain the values presented in Table 3.

As can be seen, the mean arc of the prosaic work lies between the poetic works. Hence the arc of morphological complexity is not text-sort dependent.

Table 3: Arc length in Slovak texts

Text	n	$\bar{L}$	Var(*L*)
Aby spriesvitnela	60	10.8329	86.1980
Iba neha	141	9.7363	55.4771
Moje určenie	145	10.1724	53.5280
Čas pre nádych vône (Pr)	260	9.2904	46.7143
Hľadanie odpovedí	36	10.3139	39.9226
Dielo Stvoriteľa	133	8.7261	40.5231
Sila l'udského ducha	349	10.0733	47.4222

Further, the mean arc is not dependent on text size. If we take the overall mean of all these texts, we obtain [60(10.8329) + 141(9.7363) + 145(10.1724) + 260(9.2904) + 36(10.3139) + 133(8.7261) + 249(10.0733)]/(60 + 141 + 145 + 260 + 36 + 133 + 349) = 9.7516 which may represent the general arc of morphological complexity of words in Slovak measured in the above way.

5 Motifs

Motifs are secondary entities analogical to those in music. In the linguistic literature (cf. Beliankou, Köhler, Naumann 2013; Köhler 2006, 2008a, b; Köhler, Naumann 2008, 2009, 2010; Mačutek 2009; Sanada 2010) they are defined as sequences of non-decreasing values of some property. For example in the sequence (2, 4, 3, 6, 7, 5, 2, 2, 3) we have the motifs (2,4), (3,6,7), (5), (2,2,3). These motifs have, again their own properties like *length* expressed by the number of elements in them, *range* expressed by the difference of the first and the last element, mean and other statistical properties. Motif is a kind of abstraction, a unit of a higher level constructed on the basis of lower level properties. Evidently one can continue in creating ever more abstract properties because lengths, means, ranges, etc. may be the basis of further sequences.

Here we shall analyse only the complexity sequence of the prosaic text *Čas pre nádych vône* for which we obtain the motifs as follows:

[(1,10,10,18.3), (10,10), (1,10), (1,26.6), (9.3), (8.3), (1,10), (1,1,18.3), (1,1,10,10,10), (1,10), (9.4,19), (1,10), (1,10), (18.4), (10), (1,1,1,43.9), (10,10), (1,19), (17.3), (1,1,1,26.7), (18.3), (10,10), (9.4,18.3), (1,10), (9.3,10), (1,10), (1,10,10), (1,10,27.3), (10), (1,9.3,10,18.3,18.3), (9.4,18.4), (10,10,10,19), (10,19), (9.4,10), (1,19), (10), (1,19), (10), (1,25.7), (1,19), (10), (1,10), (1,1,10), (1,10), (1,10,17.4), (9.4), (1,1,26.7), (1,1,18.3,19,9), (4,10,18.4), (10), (1,10,10,10), (9.4,18.3), (8.2), (1,26.4), (1,18.4,26.4),

(10,17.7), (10,17.3), (1,10), (1,10), (1,18.4,18.7), (8.2,19), (10,10,10,18.3), (1,9.3), (1,9.3), (1,9.4), (1,10), (1,10), (9.3), (1,19), (1,1,19,19), (1,9.4,18.4), (1,10), (1,18.4), (1,10), (1,1,1,1,19), (10,18.3), (10), (9.3,10,10,26.7), (1,35), (18.4,18.4), (1,1,10,10), (1,18.4), (10,18.4), (1,19), (10,19), (10,19), (4,18.3,18.3), (1,18.4), (10), (1,10,10,10,10), (1,19), (1,10,19), (10,10), (1,19), (10,10,18.4), (10,18.4), (1,10), (1,10,10), (1,9.4,10), (1,1,10), (9.4), (1,10,18.4), (10), (1,19), (8.2,19), (1,10,10), (1,19,19), (1,19), (18.3), (10,19), (10), (1,18.4), (10,10,10,10,19), (18.4)]

This subdivision yields, for example a smooth distribution of motif lengths. The individual length (= number of elements in a motif) are presented in Table 3. In modelling the distribution we start from the unified theory (cf. Wimmer, Altmann 2005) and conjecture that in any grammar there is a regularity of motif length representation in text. In a language having complex morphology, the lengths are not equal but controlled in such a way that the next length class is proportional to the previous, i.e. $P_x = g(x)P_{x-1}$, where P_x is the probability of classes with length x and $g(x)$ is a proportionality function. The proportionality function must contain the diversifying force of the writer who wants to express his style, the controlling force of the hearer (or the regard to the hearer), and a stabilizing control of the language itself. Putting all this together we obtain the difference equation

$$P_x = \frac{k+x-1}{m+x-1} qP_{x-1}, \tag{7}$$

where k represents the "force" of the speaker/writer, m that of the hearer/reader, and q is the language constant. Solving this equation we obtain the hyper-Pascal distribution

$$P_x = \frac{\binom{k+x-1}{x}}{\binom{m+x-1}{x}} q^x C, \quad x = 0,1,2,... \tag{8}$$

where C is the normalizing constant. Since there are no motifs of length $x = 0$, we must displace the distribution one step to the right and obtain

$$P_x = \frac{\binom{k+x-2}{x-1}}{\binom{m+x-2}{x-1}} q^{x-1} C, \quad x = 1,2,... \tag{9}$$

Table 4: Motif lengths in *Čas pre nádych vône*

Length	Frequency	Theoretical
1	19	21.40
2	61	55.71
3	19	22.83
4	10	8.79
5	5	5.26
k = 0.1664, m = 0.0230, q = 0.3595, X^2 = 1.60, DF = 1, P = 0.21		

where C can be expressed by the hypergeometric function $C = ({}_2F_1(k,1;m;q))^{-1}$. If we compute (9), we obtain the values in the third column of Table 4.

This result is, of course, preliminary, but it shows the cooperation of forces in forming the morphological complexity of a text. For different texts and languages we obtain different parameters. The computation presupposes sufficiently long texts in which there are at least five different motif-length classes. If this is not given, one must take a special or limiting case of the hyper-Pascal distribution (cf. Wimmer, Altmann 1999). In our case, we restrict ourselves to the Poisson distribution which is a limiting case of hyper-Pascal, when $m = 1$, $k \to \infty$, $q \to 0$, $kq \to a$. We obtain (in 1-displaced case)

$$P_x = \frac{a^{x-1}e^{-a}}{(x-1)!}, \quad x = 1,2,3,... \tag{10}$$

Some results are presented in Table 5.

Since complexity motifs present an expression of grammatical sequencing of words, the result shows that even on this very abstract level laws are active.

In Slovak, the lengths of motifs are relatively small. It may be tested whether this property (viz. motif length) is linked with the degree of synthetism which must, of course, be measured in a different way.

6 Distances

A writer has a special technique of using words depending on his/her age, education, gender, living space, reader circle, etc. This boils down to the fact that his complexity degrees follow some particular pattern, i.e. special complexities are repeated in some distances. The grammar of the language forces him to use very simple words lying between the more complex ones, thus even if we cannot per-

Table 5a: Complexity motifs in the poems by E. Bachletová

x	*Aby spriesvitnela*		*Iba neha*		*Moje určenie*	
	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	10	10.84	2	2.31	9	12.83
2	14	11.03	27	26.74	26	19.36
3	3	5.62	12	12.66	11	14.60
4	2	1.90|	7	5.73	10	7.34
5	1	0.60|	3	2.55	2	3.87
6			0	1.13|		
7			1	0.88|		
	a = 1.0177, X^2 = 2.18 DF = 2, P = 0.37		k = 0.1024, m = 0.0038, q = 0.4311, X^2 = 0.95 DF = 2, P = 0.62		a = 1.5088, X^2 = 6.18 DF = 3, P = 0.10	

Table 5b

x	*Hladanie odpovedi*		*Dielo Stvoriteľa*		*Sila ľudského ducha*	
	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	8	8.27	9	9.70		
2	13	12.86	18	17.87		
3	6	5.61	12	10.74		
4	2	2.12	5	5.57		
5	1	1.15	3	2.73		
6			1	1.29		
7			–	0.60|		
8			–	0.27|		
9			1	0.23|		
	k = 0.5576, m = 0.1117 q = 0.3115, X^2 = 0.86 DF = 1, P = 0.80		k = 0.6152, m = 0.1419 q = 0.4250, X^2 = 0.36 DF = 3, P = 0.95			

ceive a regular oscillation of degrees, we may look for a distance pattern. The problem of distances between equal entities has been scrutinized expecially with frequencies and adnominal modificators, and it can be used also for studying the patterns of parts-of-speech or any other kind of linguistic units.

Table 6: Distances between words of equal complexity in the text *Čas pre nádych vône*

Distance	Frequency	Theoretical
1	45	45.71
2	58	55.81
3	33	35.77
4	21	20.48
5	8	11.72
6	10	6.99
7	7	4.42
8	6	3.01
9	5	2.21
10	1	1.74
11	3	1.47
12	3	1.30
13	2	1.20
14	3	1.13
15	2	1.09
17	4	1.04
18	1	1.03
19	2	1.02
20	1	1.01
21	1	1.01
23	3	1.01
24	3	1.00
27	1	1.00
29	2	1.00
30	1	1.00
33	1	1.00
34	1	1.00
38	1	1.00
39	1	1.00
40	1	1.00
42	1	1.00
51	1	1.00
58	1	1.00
68	1	1.00
75	1	1.00
77	1	1.00
86	1	1.00
104	1	1.00

Let us define the distance between two equal entites as the number of steps necessary to attain the next equal entity from the previous, i.e. in the sequence *a,b,b,b,a,* the second *a* lies 4 steps behind the first. We study the number (y) of steps of length $x = 1, 2, 3, \ldots$ in text and approach the problem using continuous functions. We start again from the conjecture that the relative rate of change of distance classes is some function of language forces (requirements), i.e. $y'/y = g(x)$. We define $g(x) = b/x$ where b is some language constant, but since the author always effects this constant, we set a modifying function –c ln x and obtain

$$\frac{y'}{y} = \frac{b - c \ln x}{x}. \tag{11}$$

Since we do not consider zero distances, we may write y as $y - 1$. Solving (11) and renaming the parameters we obtain

$$\sin y = 1 + ax^{(B - C \ln x)}.$$

Fitting this function to the number of distances in *Čas pre nadych vone* we obtain the results presented in Table 6. The resulting formula is

$$y = 1 + 44.7054x^{(1.1022 - 1.2888 \ln x)}, \quad R^2 = 0.98$$

Modelling discrete sequences by continuous functions and ignoring the normalisation necessary for every distribution is usual in sciences. "Discrete" and "continuous" are only our conceptual constructions by means of which we want to capture some real pattern. Using a function we need not use each value of the variable, and need not pool classes as is usual with testing by means of the chi-square criterion. The determination coefficient is a sufficient warrant for the goodness-of-fit.

In strongly analytic languages the distances between equal complexities will be very short because here mostly roots appear in the text; and there will be some few great distances between higher complexities. Since it depends on the character of the language, an indicator of distances could be used also in typology.

References

Anderson, S. R. (2012): Dimensions of morphological complexity. In: *Understanding and measuring morphological complexity*. Ed. By M. Baerman, D. Brown, G. G. Corbett, (eds.).

Bane, M. (2008): Quantifying and measuring morphological complexity. *Proceedings of the 26th West Coast Conference on Formal Linguistics.* Ed. by Ch. B. Chang, H. J. Haynie. Somerville, MA 2008.

Beliankou, A.; Köhler, R.; Naumann, S. (2013): Quantitative Properties of Argumentation Motifs. In: *Methods and Applications of Quantitative Linguistics. Selected papers of the VIIIth International Conference on Quantitative Linguistics (QUALICO) in Belgrade, Serbia, April 16–19, 2012.* Ed. by I. Obradović, E. Kelih, R. Köhler, pp. 35–43.

Bickel, B.; Nichols, J. (2005): Inflectional synthesis of the verb. In: *The World Atlas of Language Structures: 94–97,* Ed. by M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie. Oxford: Oxford University Press, 94–97.

Fan, F., Altmann, G. (2007): Measuring the cohesion of compounds. In: *Problems of typological and quantitative lexicology*. Ed. by V. Kaliuščenko, R. Köhler, V. V. Levickij. Černovcy: RUTA, 190–209.

Gell-Mann, M. (1995): What is complexity? In: *Complexity* 1, 16–19.

Goldsmith, J. A. (2001): Unsupervised learning of the morphology of a natural language. In: *Computational Linguistics* 27, 153–198.

Goldsmith, J. (2006): An algorithm for the unsupervised learning of morphology. In: *Natural Language Engineering* 12.3, 1–19.

Hřebíček, L. (2000): *Variation in sequences*. Prague: Oriental Institute.

Juola, P. (1998): Measuring linguistic complexity: The morphological tier. In: *Journal of Quantitative Linguistics* 5, 206–213.

Juola, P. (2007): Assessing linguistic complexity. In: *Language complexity: Typology, contact, change*. Ed. by M. Miestamo, K. Sinnemäki, F. Karlsson. Amsterdam: John Benjamins, 89–108.

Köhler, R. (2006): *The frequency distribution of the lengths of length sequences.* In: *Favete linguis. Studies in honour of Victor.* Ed. by J. Krupa Genzor, M. Bucková. Bratislava: Slovak Academic Press, 145–152.

Köhler, R. (2008a): Word length in text. A study in the syntagmatic dimension. In: *Jazyk a jazykoveda v prohybe.* Ed. by S. Mislovičová. Bratislava: VEDA Vydavatel'stvo SAV, 416–421.

Köhler, R. (2008b): Sequences of linguistic quantities. Report on a new unit of investigation. In: *Glottotheory* 1(1). 115–119.

Köhler, R.; Naumann, S. (2008): Quantitative text analysis using L-, F- and T-segments. In: *Data Analysis, Machine Learning and Applications*. Ed. by B. Preisach, D. Schmidt-Thieme. Berlin, Heidelberg: Springer, 637–646.

Köhler, R.; Naumann, S. (2009): A contribution to quantitative studies on the sentence level. In: *Issues in Quantitative Linguistics*. Ed. by R. Köhler. Lüdenscheid: RAM-Verlag, 34–57.

Köhler, R.; Naumann, S. (2010): A syntagmatic approach to automatic text classification. Statistical properties of F- and L-motifs as text characteristics. In: *Text and Language.* Ed. by P. Grzybek, E. Kelih, J. Mačutek, J Wien: Praesens, 81–89.

Lempel, A.; Ziv, J. (1976): On the complexity of finite sequences. In: *IEEE Transactions in Information Theory* 22, 75–81.

Mačutek, J. (2009): Motif richness. In: *Issues in Quantitative Linguistics*. Ed. by R. Köhler. Lüdenscheid: RAM-Verlag, 51–60.

McWhorter, J. (2001): The world's simplest grammars are creole grammars. In: *Linguistic Typology* 5, 125–66.

Nichols, J. (2007): The distribution of complexity in the world's languages. *81st Annual Meeting of the Linguistic Society of America.*

Nichols, J. (2010): Linguistic complexity: a comprehensive definition and survey. In: *Language complexity as an evolving variable*. Ed. by G. B. Sampson, D. Gil, P. Trudgill. Oxford: Oxford University Press, 109–124.

Pellegrino, F.; Coupé, C.; Marsico, E. (2007): An information theory-based approach to the balance of complexity between phonetics, phonology, and morphosyntax. In: *81st Annual Meeting of the Linguistic Society of America*.

Popescu, I.-I.; Mačutek, J.; Altmann, G. (2009): *Aspects of word frequencies*. Lüdenscheid: RAM-Verlag.

Roelcke, Th. (2011): *Typologische Variation im Deutschen. Grundlagen – Modelle – Tendenzen*. Berlin: Schmidt.

Sanada, H. (2010): Distribution of motifs in Japanese texts. In: *Text and Language*. Ed. by P. Grzybek, E. Kelih, J. Mačutek. Wien: Praesens, 183–194.

Shosted, R. (2006): Correlating complexity: A typological approach. In: *Linguistic Typology* 10, 1–40.

Siegel, J. (2004): Morphological simplicity in pidgins and creoles. In: *Journal of Pidgin and Creole Languages* 19, 139–162.

Vulanović, R. (2007): On measuring language complexity as relative to the conveyed linguistic information. In: *SKY Journal of Linguistics* 20, 399–427.

Wimmer, G.; Altmann, G. (1999): *Thesaurus of univariate discrete probability distributions*. Essen: Stamm.

Wimmer, G.; Altmann, G. (2005): Unified derivation of some linguistic laws. In: *Quantitative Linguistics. An International Handbook*. Ed. by R. Köhler, G. Altmann, R. G. Piotrowski. Berlin: de Gruyter, 791–807.

Ziegler, A.; Altmann, G. (2002): *Denotative Textanalyse*. Wien: Praesens.