Glottometrics 27
2014

RAM-Verlag

ISSN 2625-8226

Glottometrics

Glottometrics ist eine unregelmäßig erscheinende Zeitdchrift (2-3 Ausgaben pro Jahr) für die quantitative Erforschung von Sprache und Text.
Beiträge in Deutsch oder Englisch sollten an einen der Herausgeber in einem gängigen Textverarbeitungssystem (vorrangig WORD) geschickt werden.
Glottometrics kann aus dem **Internet** heruntergeladen werden (**Open Access**), auf **CD-ROM** (PDF-Format) oder als **Druckversion** bestellt werden.

Glottometrics is a scientific journal for the quantitative research on language and text published at irregular intervals (2-3 times a year).
Contributions in English or German written with a common text processing system (preferably WORD) should be sent to one of the editors.
Glottometrics can be downloaded from the **Internet (Open Access)**, obtained on **CD-ROM** (as PDF-file) or in form of **printed copies.**

Herausgeber – Editors

Bestellungen der CD-ROM oder der gedruckten Form sind zu richten an

Orders for CD-ROM or printed copies to RAM-Verlag RAM-Verlag@t-online.de

Herunterladen/ Downloading: https://www.ram-verlag.eu/journals-e-journals/glottometrics/

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Glottometrics. 27 (2014), Lüdenscheid: RAM-Verlag, 2014. Erscheint unregelmäßig.
Diese elektronische Ressource ist im Internet (Open Access) unter der Adresse https://www.ram-verlag.eu/journals-e-journals/glottometrics/ verfügbar.
Bibliographische Deskription nach 27 (2014) **ISSN 2625-8226**

Contents

Articles

Glottometrics 27, 2014, 1-9

Four reasons for a revision of the Transitivity Hypothesis

Radek Čech, Ostrava

Abstract. Since the Transitivity Hypothesis was introduced thirty four years ago, it has become one of the most influential approaches to the functioning of transitivity in natural language. Despite the huge impact of the approach, at least within functional linguistics, some fundamental theoretical and methodological problems still remain unsolved; this seriously undermines the entire approach. The aim of this study is to analyze the four most crucial shortcomings of the approach and to propose solutions. Specifically, the study focuses on (1) the consequences of the absence of a sound theoretical foundation, (2) the ambiguity of the Hypothesis, (3) methodological deficiencies, and (4) the dubious validity of the Transitivity Hypothesis with regard to its universality. This study also takes into account later modifications of the Transitivity Hypothesis, particularly the frequency-based approach which has been advanced by the authors of the Transitivity Hypothesis.

Keywords: Transitivity, probability, definitions

1. Introduction

The Transitivity Hypothesis (hereinafter TH) was proposed thirty four years ago by P. Hopper and S. Thompson (1980). Since its publication, Hopper and Thompson's paper has been considered a seminal contribution to the research into the functioning of transitivity in language, and it has been cited in the majority of studies focusing on transitivity – at least those taking a functional linguistic approach. By way of illustration, the Web of Science database reflects the huge impact of Hopper and Thompson's paper – it is the second most cited article (with 756 citations; an average 32.9 citations per year) which has ever been published in *Language*, the Journal of the Linguistic Society of America. The impact of Hopper and Thompson's approach to transitivity is indisputable. Moreover, the authors formulated their view on the functioning of transitivity in the form of an empirically testable hypothesis; this has significantly increased their ideas' attraction to researchers. In summary, the TH represents a highly heuristic view of the one of the most fundamental properties of language, and the form of the TH enables us to characterize it as an empirical scientific approach.

However, closer observation of the TH reveals some fundamental problems, both theoretical and methodological. Surprisingly enough, among the large number of studies referring to the TH, only a tiny minority of them (e.g. Tsunoda 1985, Olsen – MacFarland 1996, LaPolla et al. 2011) have focused on critical analysis of the theoretical and methodological foundations of the TH. The majority of studies take the TH for granted, or merely propose slight modifications to it. The aim of this article is to show that fundamental problems seriously undermine the TH and that if the heuristic value of the TH is not to be lost, these fundamental problems must be solved. The present article offers a critique of the TH while also proposing some solutions to the challenges identified.

2. The main characteristics of the Transitivity Hypothesis

According to Hopper and Thompson (1980), transitivity represents a crucial relationship in language which has a number of universally predictable consequences in grammar. Importantly, transitivity is not viewed in a traditional sense – according to which the presence (or absence) of the object in the sentence is the only parameter distinguishing between transitive (or intransitive) clauses. Instead, Transitivity[1] is regarded as a continuum: it is a matter of the grammar (and semantics) of the entire clause and it "can be broken into its component parts (...), they allow clauses to be characterized as MORE or LESS Transitive: the more features a clause has in the 'high' column 1A–J, the more Transitive it is" (p. 253); see Table 1.

Table 1
Transitivity parameters (Hopper – Thmopson 1980, p. 252)

	Parameter	**High Transitivity feature**	**Low Transitivity feature**
A	PARTICIPANTS	2 or more	1
B	KINESIS	action	non-action
C	ASPECT	telic	atelic
D	PUNCTUALITY	punctual	non-punctual
E	VOLITIONALITY	volitional	non-volitional
F	AFFIRMATION	affirmative	negative
G	MODE	realis	irrealis
H	AGENCY	Agent high in potency	Agent low in potency
I	AFFECTEDNESS of Object	Object totally affected	Object not affected
J	INDIVIDUATION of Object	Object highly individuated	Object non-individuated

The value of Transitivity in a sentence is determined by the presence of high Transitivity features, so the sentence

(1) *Susan left*

is more Transitive than the sentence

(2) *Jerry likes beer*

because sentence (1) has more high-Transitivity features (Kinesis: action; Aspect:

1 The authors use the term Transitivity (or Transitive) with a capital T to designate this specific understanding of the notion.

telic; Punctuality: punctual; Volitionality: volitional) than sentence (2) (Participants: two) (ibid. p. 254).

The most important aspect of the TH, in my opinion, lies in its prediction hypothesizing the relationships between the components: “If two clauses (a) and (b) in a language differ in that (a) is higher in Transitivity according to any features 1A-J, then, if concomitant grammatical or semantic difference appears elsewhere in the clause, that difference will also show (a) to be higher in Transitivity” (ibid, p. 255). Component features should co-vary extensively and systematically, so “whenever two values of the transitivity components are necessarily present (...) they will agree in being either both high or both low in value” (ibid., p. 254). In summary, Transitivity causes a very wide range of correlations in the grammar of language.

3. Reasons for the revision of the Transitivity Hypothesis

3.1 The origin of Transitivity – a proper theory is needed

Let us try to examine Transitivity from a more global point of view. It has been shown in Section 2 that according to the TH, Transitivity controls relationships among very different grammatical and semantic facets of language. Consequently, Transitivity should be viewed as a kind of linguistic ‘supra-category’, and it is necessary to answer the question of the origin of this important property of language.

Hopper and Thompson, at the beginning of their study, promise to present a “broad theory of Transitivity” (1980, p. 251). First, they state that Transitivity “involves a different facet of the effectiveness or intensity with which the action is transferred from the participant to another” (p. 252). The article then gives plenty of examples which are intended to corroborate the TH. Next, the authors articulate the need to find some underlying unitary principle which enables the TH to be explained; however, the authors admit that a superordinate semantic principle including all Transitivity components has not been discovered, and turn their focus to pragmatics.

Generally, the authors assume that a “linguistic universal originates in a general pragmatic function, and that the universal is not explained until this function has been isolated and related to this universal“ (p. 280). Consequently, since Transitivity is viewed as being a universal property of language, it should be connected to some communicative function.

In particular, the authors relate Transitivity to text properties. Accorrding to them, any text consists of both a more relevant part, referred to as the *foreground*, and a less relevant part, the *background*. The foreground supplies the main points of the discourse and crucially contributes to the speaker’s communicative goal, while the background merely assists, amplifies, or comments on it (cf. ibid p. 280). In languages like English, which do not express foregrounding by a single morphosyntactic feature, the foreground manifests itself by a cluster of properties. According to the authors, this cluster is precisely that set of properties which characterize high Transitivity (cf. ibid p. 284). Further, foregrounding is marked on a probabilistic basis, so “the likelihood that a clause will receive a foregrounded interpretation is proportional to the height of the scale of Transitivity. From the performer’s point of view, the decision to foreground a clause will be reflected in the decision to encode more (rather than fewer) Transitivity features in the clause" (ibid. p. 284). In summary, Transitivity can be viewed as a discourse-motivated mechanism which governs the behaviour of particular Transitivity features.

However, does this kind of explanation really represent the promised "broad theory of Transitivity"? Even if one sets aside the methodological problems (see Section 3.3, 3.4), some fundamental questions arise: Is the TH proposed in relationship to other hypotheses? Why were the particular parameters chosen? What is the relationship between particular parameters and discourse characteristics (foreground vs. background)? Why should some features manifest foregrounding (or back-grounding) and others not? For example, why should an affirmation be more effective at achieving the speaker's communicative goals than a negation? What are the relationships among particular Transitivity parameters? Are they uniform? Or do they constitute a hierarchy?

Without answers to questions of this kind, the TH is not much more than a statement concerning some correlative relationships within language. However, one shoud bear in mind that "[i]n any data, some correlations can be found if all you are looking for is correlations!" (Fraassen 2002, p. 159). To summarize, a description of correlations is no theory; moreover, the mere presence of correlation does not guarantee that the correlation is a manifestation of the theory (or better, the manifestation of a law which is derived from the theory).

3.2 Ambiguity of the hypothesis

At first sight, the TH is set forth with crystal clarity: "If two clauses (a) and (b) in a language differ in that (a) is higher in Transitivity according to any features 1A-J, then, if concomitant grammatical or semantic difference appears elsewhere in the clause, that difference will also show (a) to be higher in Transitivity.

The converse of this hypothesis, that there is a similar correlation among low-Transitivity features, is implicit. (…) The Transitivity Hypothesis also predicts that the opposite type of correlation will not be found, where a high-Transitivity feature systematically co-varies with low-Transitivity feature in the same clause" (p. 255).

However, even a cursory glance at Table 1 reveals unsustainable consequences of the TH. Specifically, if no co-variation between particular low-Transitivity and high-Transitivity features is predicted, it should not be possible, for example, to use an atelic verb predicate in a two-participant sentence or a punctual verb in a negative sentence. The prediction given by the TH evidently contradicts the user's common language experience. For example, the sentence

(3) *Peter did not kick the ball*,

containing the negative punctual verb, is undoubtedly well-formed and commonly used in English.[2]

2 The Google search engine finds approximately 66 000 instances of the string "did not kick the ball" [25th February 2014].

In order for the TH to remain meaningful, it is necessary to view the correlative relationships between particular parameters not in the strict sense, but probabilistically. In fact, this approach is implicitly adopted by the authors of the TH; besides the examples which fit the original strict formulation of the TH, some examples formulated as tendencies are also used for corroboration of the hypothesis. For example, it is stated that "an animate O [object] is *more conducive* to the selection of the accusative than an inanimate O [object]; a singular O [object] is *more likely* to be (and is *more acceptable*) in the accusative than a plural O [object]" (Hopper – Thompson 1980, p. 279) [my italics]. Moreover, if the authors claim that Transitivity should be higher in the foreground than in the background, the probabilistic approach is anticipated; particularly, in the foreground *more* high Transitivity features should appear in the sentence than in the background, which means that in the foreground there should be a *higher* correlation between high Transitivity features than in the background.

In the light of these facts, it is hard to comprehend why the authors did not originally formulate the TH probabilistically. The original 'strict' form of the hypothesis is ambiguous, which seriously confuses the whole approach.

3.3 A frequency-based approach to the Transitivity Hypothesis – a proper methodology is needed

A frequency-based approach to the TH is explicitly adopted in Thompson and Hopper's later work (2001) focusing on the relationship between language form, namely conversation, and Transitivity. However, Čech and Pajas (2009) revealed some fundamental deficiencies of their approach; first, the prediction concerning the relationship between language form and Transitivity presented in Thompson and Hopper's (2001) paper lacks the form of an empirically testable hypothesis. For example, it is stated that Transitivity is low in conversation, and consequently the majority of clauses turn out to have one participant. The presented results seem to confirm the prediction: 73% of one-participant clauses and 27% of two or more-participant clauses were detected in the observed dataset. Nevertheless, what does it actually mean when one says that something is 'low' or 'high' without an explicit scale factor? In other words, what percentage of one-participant clauses is 'enough' to say that Transitivity is low? Moreover, the authors did not explicitly formulate the claim that Transitivity is low *in comparison* to written language (or a particular genre), although this is probably assumed implicitly. However, without a clearly formulated hypothesis, e.g. *the ratio of one-participant clauses, in comparison to two or more-participant clauses, is higher in conversation than in written language*, neither the statement concerning the relationship between conversation and Transitivity, nor the presented empirical findings, have any scientific validity.

Next, the differences among distributions are interpreted without any statistical test. As Altmann and Lehfeldt (2004) pointed out, this represents "a disease of the frequentism that could be called a children's illness if it had not have lasted already for such a long time" (p. 278).

Last but not least, one of the most important deficiencies of the TH lies in the vagueness of its definition of particular Parameters. In the majority of cases, it is assumed that notions such as *negation, punctuality, affectedness* etc. are not problematic; consequently, these notions are defined superficially, despite the fact that it is well-known in linguistics that even relatively well-established notions are not unequivocal (cf. Brown, 2005). However, without clear definitions, at least operational ones, the analyses are obscure, and obviously a different comprehension of the notions will bring different results.

3.4 (Non-)universality of the Transitivity Hypothesis

The crucial importance of the TH is dependent on its universal validity. To emphasize this aspect of the TH, Hopper and Thompson claim at the very beginning of their article that Transitivity has “a number of universally predictable consequences in grammar” (p. 251). However, although the TH is indeed originally formulated universally, without any restrictions – cf. “whenever two values of the Transitivity components are necessarily present (...) they will agree in being either both high or both low in value” (p. 254) – the first constraint on its universal validity is posited by the authors. Transitivity is viewed by them as a discourse property, which means that it should reflect a distinction between foregrounded and backgrounded discourse. Consequently, if one thinks of a language which obligatorily expresses for example both an object and aspect, the higher correlation between these two parameters should appear in the foreground rather than in the background. So the prediction can be viewed as universal, but only in the case of the foreground. Not surprisingly, Hopper and Thompson emphasize this aspect of the approach in the conclusion of their article: “Semantic and grammatical properties which are irrelevant to foregrounding are also irrelevant to Transitivity” (p. 294). However, it is unclear why this constraint was not incorporated into the original hypothesis and why the authors have not predicted that ‘whenever two values of the Transitivity components are necessarily present *in the foreground* they will agree in being either both high or both low in value’. In my view, such a formulation would significantly clarify the approach.[3]

Another restriction of the TH is presented by the authors in their study focusing on the relationship between Transitivity and conversation (Thompson – Hopper 2001). It is stated that conversation is low in Transitivity; this is illustrated by the character of two-participant clauses. More concretely, the observation of conversation has revealed strong correlation between two-participant clauses (which manifest a high-Transitivity feature) and low-Transitivity features, such as Non-action, Atelic, Non-punctual and so on (see Table 2).

3 The relationship between the universal status of Transitivity and discourse properties is emphasized in Hopper and Thompson’s later works, cf. “a cross-linguistic function of ‘Transitivity’ is of a central importance in universal grammar, and at the same time is derived from discourse salience of prototypically transitive clauses” (Hopper – Thompson 1984, p. 707).

Table 2
The ratio of low Transitivity features in two-participant clauses in conversation (based on Thompson & Hopper 2001, p. 37).

Kinesis: Non-action	86%
Aspect: Atelic	86%
Punctuality: Non-punctual	98%
Affectedness: Non-affected Object	84%
Mode: Non-irrealis	70%
Individuation: Non-individuated Object	55%
Volitionality: Non-volitional Agent	50%
Agency: Potent Agent	97%

However, the results in Table 2 indicate co-variation of opposite features, which is in direct contradiction with the prediction of the TH (see Section 2). This means that the TH is not valid for conversation, and its universality is radically restricted to just one part of discourse – the foreground – in one form, i.e. written, of language. Moreover, no clear criteria for distinguishing the foreground and background are put forth.

In summary, the TH is presented as a language universal 1) with highly restricted validity and 2) without a methodology enabling researchers to test its validity empirically, because of the absence of interpersonally observable criteria for the delimitation of the foreground.

4. Conclusion and proposals

Although Hopper – Thompson's approach to Transitivity has opened up an interesting way of viewing a very important aspect of the functioning of language, fundamental theoretical and methodological deficiencies undermine the entire approach. However, in my opinion these deficiencies are solvable. The proposals for solutions are as follows:

1. The TH should be implemented into a theory of language. This would clarify both the general status of Transitivity and the character of predicted relationships between particular parameters. In other words, both Transitivity, as a property of language, and the TH should be derived from more general principles which rule linguistic behaviour.
2. The TH should be formulated probabilistically. A probabilistically formulated hypothesis reflects the true intention of the authors, and – more importantly – it enables results to be tested empirically by using common statistical methods.
3. The features of parameters should be quantified.
4. The vagueness should be removed from the definitions of particular parameters. This would make it possible to quantify unambiguously the features of parameters, and consequently would provide a high level of validity (and comparability) of results. In practice, it means that the definitions must be unequivocal.

5. The majority of parameters are defined dichotomically, despite being far more complex in nature. For example, parameter A (number of participants) only distinguishes between one-participant and two or more-participant sentences, although there are obvious differences in the linguistic behaviour of participants which are represented by a direct object, indirect object, prepositional object, and adverbial. It therefore seems more reasonable to define, if possible, the features of parameters as a scale. Dichotomy of properties is a heredity having its origin in structuralism.
6. The results should be interpreted using common statistical methods. The first step would be the translations of conjectures into the language of statistics.
7. The relationship between Transitivity and discourse should be reconsidered; either a clear definition of the foreground must be given (with a method for distinguishing between the foreground and background), or Transitivity has to be redefined in genuinely universal terms, i.e. without restrictions as to discourse type (or language form).

If implemented, these proposals would bring Hopper and Thompson's approach into the field of empirical/experimental science – which seems to be in accordance with the linguistic stance taken by the authors themselves (cf. Hopper 1987, Bybee & Hopper 2001).

Acknowledgement

R. Čech was supported by GAČR (Czech Science Foundation) No. 405/08/P157 – Components of transitivity analysis of Czech sentences (emergent grammar approach).

References

Altmann, G. , Lehfeldt, W. (2004). Book review. (Bybee, J. - Hopper, P. (eds.) (2001): *Frequency and the Emergence of Linguistic Structure*. Amsterdam, Philadelphia: John Benjamins). *Journal of Quantitative Linguistics 11, 275–304.*

Brown, K. (ed.) (2006). *The Encyclopedia of Language and Linguistics*. Oxford: Pergamon.

Bybee, J., Hopper, P. (2001). Introduction to frequency and the emergence of linguistic structure. In Bybee, J., Hopper, P. (eds.), *Frequency and The Emergence of Linguistic Structur*e: *1-24*. Amsterdam/Philadelphia: John Benjamins.

Čech, R., Pajas, P. (2009). Pitfalls of the Transitivity Hypothesis: Transitivity in conversation and written language in Czech. *Glottotheory 2, 2009, 41-49.*

Fraassen, B.C. (2002). *Empirical Stance*. New Haven & London: Yale University Press.

Hopper, P. (1987). Emergent grammar. In: *Proceedings of the thirteenth annual meeting of the Berkley Linguistics Society: 139-157*. Berkley: Berkley Linguistics Society, 1987.

Hopper, P., Thompson, S. (1980). Transitivity in Grammar and Discourse. *Language 56, 251-299.*

Hopper, P., Thompson, S. (1984). The Discourse Basis for Lexical Categories in Universal Grammar. *Language 60, 703-752.*

LaPolla, R. J., Kratochvil, F., Coupe, A. R. (2011) On Transitivity. *Studies in Language 35 (3), 469-491*.

Olsen, M. B., Macfarland, T. (1996). Where's Transitivity? Paper presented at the *Seventh Annual Formal Linguistic Society of Mid-America conference*, May 17–19, 1996, The Ohio State University.

Thompson, S., Hopper, P. (2001). Transitivity, clause structure, and argument structure: Evidence from conversation. In: Bybee, J., Hopper, P. (eds.): *Frequency and the Emergence of Linguistic Structure: 27–56*. Amsterdam, Philadelphia: John Benjamins,

Tsunoda, T. (1985). Remarks on Transitivity. *Journal of Linguistics 21, 385-396.*

Glottometrics 27, 2014, 10-17

Hebraismen im Deutschen

Karl-Heinz Best

Abstract. The present paper presents the development of Hebraic borrowings in German and demonstrates that this process abides by the logistic law which in linguistics is known as Piotrowski Law.

Keywords: Borrowing, Hebrew, Piotrowski Law.

Vorbemerkung

Zwei Ziele werden mit diesem Beitrag verfolgt:

1. Es sollen die in der deutschen Gemeinsprache vorkommenden Hebraismen erfasst werden. Datenquelle sind entsprechend allgemeine Wörterbücher des Deutschen, keine Wörterbücher mit spezieller Fachterminologie.
2. Es soll ein weiteres Mal überprüft werden, ob die Übernahme der noch heute gebräuchlichen Hebraismen über die Jahrhunderte hinweg in Übereinstimmung mit dem Piotrowski-Gesetz (Altmann 1983) verläuft und wie sich dieser Trend darstellt.

Hebraismen sind gelegentlich erfasst worden (z.B. Kreuzer 2001), aber nicht in der für diesen Beitrag erforderlichen Form. Deshalb wurde dieser Wortschatz mit den erforderlichen Informationen hier erneut zusammengestellt.

Vorgehen

Die vorliegende Untersuchung knüpft eng an die zu den Jiddismen im Deutschen an, sowohl methodisch als auch inhaltlich. Als Hebraismen werden alle Wörter definiert, die aus dem Hebräischen oder auch über das Hebräische ins Deutsche gekommen sind, auch wenn ihr letzter Ursprung auf eine andere Sprache zurückgeht. Viele dieser Entlehnungen haben das Deutsche über das Jiddische erreicht; die Daten dieser Wörter wurden der entsprechenden Untersuchung (Best 2006) entnommen und nicht neu bearbeitet.

Als Hebraismen wurden diejenigen Wörter aufgenommen, die in Duden (^{2}1999) als solche ausgewiesen sind. Die Datierung erfolgt primär nach Kluge (^{24}2002), wo möglich. Wo im Duden „gaunerspr." als Entlehnungsstation steht, findet man bei Kluge oft „rotwelsch". Diese beiden Zuweisungen werden in der Literatur offenbar nicht systematisch unterschieden. Hier wurde nach Kluge „rotwelsch" eingefügt, wo er diese Angabe hat. Beide, (rotwelsch) und (gaunerspr.), werden in Klammern gesetzt, da sie keine eigenen Sprachen sind, sondern nur Sondersprachen des Deutschen.

Kluge (^{24}2002) wird auch bei den Angaben zur Entlehnungsgeschichte vertraut, da dieses Wörterbuch bei der Untersuchung zu den Jiddismen die zuletzt erfolgte

Bearbeitung eines etymologischen Wörterbuchs war. Sie werden um einige Angaben aus *Duden Herkunftswörterbuch* (2001) und Pfeifer (²1993/1995) ergänzt.

Übersicht über die Hebraismen im Deutschen

Die folgende Tabelle stellt die Hebraismen zusammen. (Die Bedeutungshinweise dienen lediglich der groben Orientierung. Außerdem wird angegeben, in welchem Jahrhundert und auf welchem Weg ein Hebraismus im Deutschen erscheint. Fragezeichen zeigen unsichere Zuordnungen an.

Tabelle 1
Hebraismen im Deutschen

Entlehnung	Jhd.	Bedeutungshinweis	Entlehnungsweg
acheln	16.	essen	(rotwelsch) -jidd. - hebr.
Adonai		mein Herr, Name Gottes im AT	hebr.
[2]Agora	20.	Untereinheit des Schekel	hebr.
amen	8.	Gebetsformel	lat. - griech. - hebr.
Ariel		Name	hebr.
Baal		semit. Wetter- und Himmelgott	hebr.
Bafel	19.	schlechte Ware; Gerede	jidd.? - hebr.?
baldowern	19.	auskundschaften	(rotwelsch) - jidd. - hebr.
Balsam	11.	Linderungsmittel	lat. - griech. - hebr.
Barches		weißes Festtagsbrot	jidd. - hebr.
[1]Bar-Mizwa		Jude nach Vollendung des 13. Lebensjahres	hebr.
[2]Bar-Mizwa		Feier zur Initiation von [1]Bar-Mizwa	hebr.
Bat-Mizwa		Jüdin nach Vollendung des 13. Lebensjahres	hebr.
Beelzebub	8.	oberster Teufel	hebr.
Behemot(h)		Tier	hebr.
Beisel, Beisl, Beiz(e)	20.	einfaches Gasthaus	(rotwelsch) - jidd. - hebr.
Belial, Beliar		Teufel	hebr.
Ben		Teil von Eigennamen	hebr./arab.
Beschores		unredlicher Gewinn	jidd.- hebr.
betucht	17.	wohlhabend	jidd. - hebr.
bigott	18.	übertrieben fromm	frz. - jidd.?
Bisam	9.	Moschus	mittellat. - hebr.
Chanukka		Fest	hebr.
Cherub, Kerub		Engel	hebr.
Chuzpe	20.	Dreistigkeit	jidd. - hebr.
Daffke	20.	aus Daffke: nun gerade	(rotwelsch) - jidd. - hebr.
Dalles	18.	Armut; Erkältung	jidd. - hebr.
dibbern	15.	leise miteinander sprechen	(rotwelsch) - jidd. - hebr.
Eden		Paradies	hebr.
Elohim		Gott	hebr.
Essener		Name	hebr.?

Ezzes, Eizes	19.	Tipps	(rotwelsch) - jidd. - hebr.
Ganeff	19.	Ganove	(rotwelsch) - jidd. - hebr.
Ganove	20.	Verbrecher	(rotwelsch) - jidd. - hebr.
Gauner	16.	Spitzbube	(rotwelsch) - jidd. - hebr.?
Gehenna		Tal Hinnoms	kirchenlat. - griech. - hebr.
Geseier, Geseire	19.	unnützes Gerede	(rotwelsch) - jidd. - hebr.
Goi	18.	Nichtjude	jidd. – hebr.
Golem		Sagenfigur	hebr.
Golgatha		Schädelstätte	kirchenlat. - griech. - hebr.
Großkotz		Wichtigtuer	jidd. - hebr.?
Hagana	20.	militärische Organisation	hebr.
hallelujah	14.	Interjektion	kirchenlat. - hebr.
hosianna		Interjektion	kirchenlat. - griech. - hebr.
Ischariot		Name	hebr.?
Ische	18.	Mädchen	jidd. - hebr.
Kabale	17.	Intrige	frz. - hebr.
Kabbala		Geheimlehre	hebr.
Kaddisch		jüdisches Gebet	jidd. - aram. - hebr.
Kaff	19.	elendes Nest	(rotwelsch) - jidd. - hebr.
Kaffer	18.	Dummkopf	(rotwelsch) - jidd. - hebr.
Kafiller		Schinder, Abdecker	(gaunerspr.) - jidd. - hebr.
Kalle	18.	Braut, Geliebte, Prostituierte	(rotwelsch) - jidd. - hebr.
kapores	18.	kaputt	(rotwelsch) - jidd. - hebr.
Karäer		Anhänger einer Sekte	hebr.
Kassiber	19.	heimliches Schreiben	(rotwelsch) - jidd. - hebr.
Katzoff, Katzuff	18.	Fleischer	(gaunerspr.) - jidd. - hebr.
Kibbuz	20.	ländliches Kollektiv	hebr.
Klezmer	20.	jüdische Instrumentalmusik	amerik. - jidd. - hebr.
Kluft	18.	Kleidung	(rotwelsch) - jidd. - hebr.
Knast	19.	Haftstrafe	(rotwelsch) - jidd. - hebr.
Knesset(h)		Parlament	hebr.
kochem	19.	klug	(gaunerspr.) - jidd. - hebr.
Kohl	18.	Geschwätz	jidd. – hebr.?
koscher	18.	den jüdischen Speisegesetzen gemäß	jidd. - hebr.
Leviathan	17.	Staatssymbol (bei Hobbes)	hebr.
Likud(block)	20.	Parteienbund	hebr.
machulle	19.	pleite, ermüdet	(rotwelsch) - jidd. - hebr.
Macke	20.	Tick	jidd. - hebr.
Makkabi		Name	hebr.
Maloche	18.	schwere Arbeit	(rotwelsch) - jidd. - hebr.
Manna	14.	Nahrung	spätlat. - griech. - hebr.
Mapai		Parteienname	hebr.
Massel	20.	unerwartetes Glück	jidd. - hebr.
Massora		Textkritik	hebr.
Massoret		Schriftgelehrter	hebr.
Matze, Mazze, Matzen, Mazzen	15.	ungesäuertes Fladenbrot	jidd. - hebr.
mauern	20.	defensiv spielen	(rotwelsch)? - jidd.? - hebr.

Mauschel		(armer) Jude	jidd. - hebr.
mauscheln	17.	betrügen, undeutlich sprechen	jidd. - hebr.
Menora		Leuchter	hebr.
meschugge	19.	verrückt	(rotwelsch) - jidd. - hebr.
Messias	18.	Heilsbringer	kirchenlat. - griech. - hebr.
mies	19.	schlecht, hinterhältig	(rotwelsch) - jidd. - hebr.
Mikwe		Tauchbad	hebr.
Mischna		Rechtssammlung	hebr.
Mischpoche, Mischpoke, Muschpoke	20.	Familie, Gesellschaft, Bande	(rotwelsch) - jidd. - hebr.
Misrach		Himmelsrichtung	hebr.
Misrachi		zionistische Organisation	hebr.
Mitzwa		gute Tat	jidd. - hebr.
molum	18.	angetrunken	(rotwelsch) - jidd. - hebr.
Moos	18.	Kleingeld	(rotwelsch) - jidd. - hebr.
mosern	18.	nörgeln	(rotwelsch) - jidd. - hebr.
Naute		ein Konfekt	jidd. - hebr.?
Nimrod		Jäger	hebr.
Ophir		sagenhaftes Land	lat. - griech. - hebr.
Paschalis		Name	hebr.
Peies		lange Schläfenlocke	jidd. - hebr.
Pessach		Passah	jidd. - hebr.
Pharisäer	18.	Heuchler	spätlat. - griech. - hebr.
Platte		die Platte putzen: fliehen	(gaunerspr.)? - jidd. - hebr.
Pleite	19.	Bankrott	(rotwelsch) - jidd. - hebr.
Purim		Fest	hebr. - pers.
Rabbi	16.	Schriftgelehrter	kirchenlat. - griech. - hebr.
Rabbiner		Schriftgelehrter	kirchenlat. - griech. - hebr.
Rebbes		Reibach	jidd. - hebr.
Reibach, Rebbach, Rewach	19.	unverhältnismäßiger Gewinn	(rotwelsch) - jidd. - hebr.
Rochus	19.	Zorn, Wut	(rotwelsch) - jidd. - hebr.
Sabbat	13.	Ruhetag	lat. - griech. - hebr.
Sabre	20.	eingeborener Jude	hebr.
Sadduzäer		Person eines Priesteradels	lat. - griech. - hebr.
Samiel		Name des Satans	griech. - hebr.
Samstag	9.	Samstag	lat. - griech. - hebr.
Sanhedrin		Ratsversammlung	hebr.
Satan (in Zusammensetzung)	8.	Satan	kirchenlat./griech. - hebr.
Schabbes	18.	Sabbat	jidd. - hebr.
Schacher	19.	gewinnorientierter Handel	hebr.
schachern	17.	Handel treiben	(rotwelsch) - jidd. - hebr.
schächten	17.	schlachten	jidd. - hebr.
Schadchen	19.	Heiratsvermittler	hebr.
schäkern	18.	scherzen, flirten	jidd.? - hebr.
Schammes		Diener in Synagoge, Assistent	jidd. - hebr.
Schamott		wertloses Zeug	jidd. - hebr.

Schaude, Schode, Schaute, Schote	16.	Narr	(gaunerspr.) - jidd. - hebr.
Schekel	20.	Währungseinheit	hebr.
Schibboleth		Erkennungszeichen	hebr.
schicker	19.	(leicht) betrunken	(rotwelsch)/jidd. - hebr.
Schickse	18.	leichtlebige Frau, Jüdin	(rotwelsch)/jidd. - hebr.
Schlemihl	19.	Pechvogel, Schlitzohr	jidd. - hebr.?
Schmiere	18.	Wache, Polizei	(rotwelsch) - jidd. - hebr.
Schmu	18.	unredlicher Gewinn, Schwindel	(rotwelsch) - jidd. - hebr.?
Schmus	18.	Getue, Geschwätz	(rotwelsch) - jidd. - hebr.
schmusen	18.	kosen	(rotwelsch) - jidd. - hebr.
Schoah, Shoah, Shoa	20.	Holocaust	hebr.
schofel	18.	schäbig, kleinlich	(rotwelsch) - jidd. - hebr.
Sekel		Gewichtseinheit	lat. - griech. - hebr.
Seraph		Engel	lat. - hebr.
Sore	18.	Diebesgut	(rotwelsch) - jidd. - hebr.
stiekum	20.	heimlich	(rotwelsch) - jidd. - hebr.
Stuss	18.	Unsinn	(rotwelsch) - jidd. - hebr.
Tacheles	20.	Tacheles reden: Klartext reden	jidd. - hebr.
taff		robust, hart	jidd. - hebr.
Talmud		Gesetzessammlung	hebr.
Tefilla		jüdisches Gebet, -sbuch	hebr.
Thora		mosaisches Gesetz	hebr.
Tinnef	19.	wertloses Zeug, Unsinn	(rotwelsch) - jidd. - hebr.
Tohuwabohu	19.	Chaos	hebr.
Tokus		Hintern	jidd. - hebr.
treife		nicht koscher	jidd. - hebr.
türmen	19.	davonlaufen	(gaunerspr.)? - hebr.
verknacken	19.	bestrafen	jidd. - hebr.
Zimt	11.	Gewürz	lat. - griech. - hebr. - malay.
Zion, Sion		Tempelberg	hebr.
zocken	19.	Glücksspiele machen	(rotwelsch) - jidd. - hebr.
Zoff	20.	Streit	(rotwelsch) - jidd. - hebr.
Zores	19.	Ärger, Wirrwarr	(rotwelsch) - jidd. - hebr.
Zosse, Zossen	18.	(altes) Pferd	(rotwelsch) - jidd. - hebr.

Insgesamt wurden 157 Hebraismen erfasst, von denen 94 aufgrund der Angaben bei Kluge (^{24}2002) und in den anderen etymologischen Wörterbüchern datiert werden können.

Verlauf des Entlehnungsprozesses

Die folgende Tabelle gibt Auskunft darüber, in welchem Jahrhundert wie viele Hebraismen das Deutsche erreichten. Diese Daten werden zusätzlich kumuliert aufgeführt. An diese kumulierten Werte wird das Modell des unvollständigen Sprachwandels in der Form

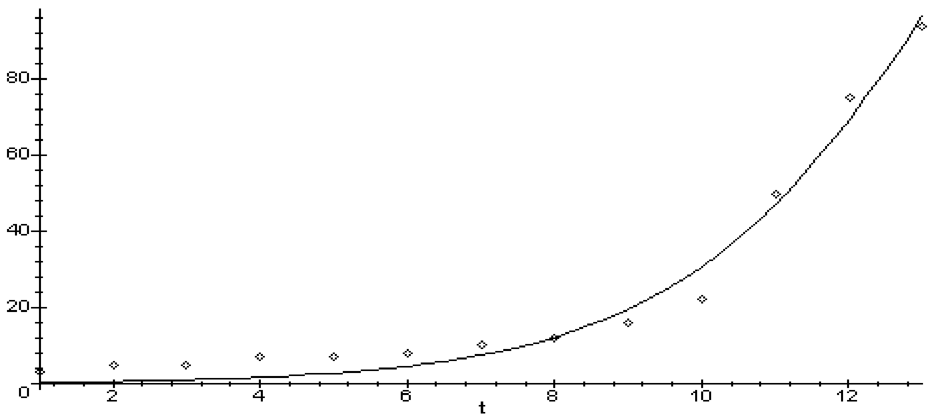
(1) $$p = \frac{c}{1 + ae^{-bt}}$$

angepasst, um zu sehen, ob der Gesamtprozess gesetzmäßig verläuft. Das Ergebnis findet sich in der folgenden Tabelle 2:

Tabelle 2
Entwicklung der Hebraismen im Deutschen

Jhd.	t	beobachtet	kumuliert	berechnet
8.	1	3	3	0.36
9.	2	2	5	0.60
10.	3	0	5	1.00
11.	4	2	7	1.66
12.	5	0	7	2.74
13.	6	1	8	4.53
14.	7	2	10	7.44
15.	8	2	12	12.12
16.	9	4	16	19.51
17.	10	6	22	30.78
18.	11	28	50	47.15
19.	12	25	75	69.31
20.	13	19	94	96.60
a = 1090.5841 *b* = 0.5091 *c* = 237.3935 *D* = 0.9769				

Legende zur Tabelle 2: *a*, *b* und *c* sind die Parameter des Modells; *c* gibt den Zielwert an, auf den nach der Berechnung der Prozess hinausläuft. *D* ist der Determinationskoeffizient, der höchstens den Wert 1 erreichen kann. Das Ergebnis ist hervorragend, wie der Testwert *D* = 0.9769 und die folgende Graphik (Abb. 1) zeigen. Parameter *c* ist mit Vorsicht zu interpretieren, da der Prozess der Entlehnungen noch nicht erkennbar den Wendepunkt überschritten hat (Best 2009) und damit der weitere Verlauf sehr unterschiedlich sein kann.



Graphik zu Tabelle 2: Entwicklung der Hebraismen im Deutschen

Schlussbemerkungen

Die Untersuchung hat ergeben, dass in der Gemeinsprache mit rund 150 mehr oder weniger geläufigen Hebraismen zu rechnen ist. Ihre Entlehnung ins Deutsche kann vom 8. Jahrhundert an beobachtet werden und hält auch im 20. Jahrhundert noch an, wobei das große Interesse in Deutschland am neu gegründeten Staat Israel eine bedeutsame Rolle spielt.

Der Prozess der Einbürgerung von Entlehnungen aus dem Hebräischen folgt dem Piotrowski-Gesetz mit sehr guter Übereinstimmung, so wie viele andere Entlehnungsprozesse auch (Ternes 2011).

Es ist zu beachten, dass der Verlauf der Entlehnungen noch deutlich komplizierter ist, als hier dargestellt, da nur die heute noch gebräuchlichen Hebraismen erfasst wurden. Es ist aber damit zu rechnen, dass in den vergangenen Jahrhunderten auch Hebraismen übernommen wurden, die dann wieder außer Gebrauch gerieten, so dass sie mit dem hier angewendeten Verfahren nicht erfasst werden konnten.

Jenseits der Grenzen der Gemeinsprache finden sich weitere Hebraismen. Hiermit sei beispielhaft auf Scheer-Nahor (1998/99) für Hebraismen im Badischen Wörterbuch und Matras (1996) für ihr Vorkommen in der Sondersprache der Viehhändler verwiesen.

Literatur

Altmann, Gabriel (1983). Das Piotrowski-Gesetz und seine Verallgemeinerungen. In: Karl-Heinz Best und Jörg Kohlhase (Hrsg.), *Exakte Sprachwandelforschung: 54–90*. Göttingen: edition herodot.

Best, Karl-Heinz (2006). Quantitative Untersuchungen zu den Jiddismen im Deutschen. *Jiddistik Mitteilungen 36, 1-14.*

Best, Karl-Heinz (2009). Sind Prognosen in der Linguistik möglich? In: Tilo Weber und Gerd Antos (Hrsgs), *Typen von Wissen. Begriffliche Unterscheidung und Ausprägungen in der Praxis des Wissenstransfers: 164-175*. Frankfurt/M.: Lang.

Duden. Herkunftswörterbuch (2001). 3., völlig neu bearbeitete und erweiterte Auflage. Mannheim/ Wien/ Zürich: Dudenverlag.

Duden. Das große Wörterbuch der deutschen Sprache in 10 Bänden. (²1999). 3., völlig neu bearbeitete und erweiterte Auflage. Mannheim/ Leipzig/ Wien/ Zürich: Dudenverlag.

Kluge. Etymologisches Wörterbuch der deutschen Sprache. (^{24}2002). Bearb. v. Elmar Seebold. 24., durchgesehene und erweiterte Auflage. Berlin/ New York: de Gruyter.

Kreuzer, Siegfried (2001). Von Ave bis Zores. Hebräische und semitische Wörter in unserer Sprache. *Zeitschrift für Literaturwissenschaft und Linguistik (LiLi) 121, 98-114.* (https://www.google.de/search?q=%22Hebr%C3%A4ische+W%C3%B6rter+im+Deutschen+%22&ie=utf-8&oe=utf-8&rls=org.mozilla:de:official&client=firefox-a&gws_rd=cr&ei=QM-UUu2_FYjJygOXrYDYDg)

Matras, Yaron (1996). Sondersprachliche Hebraismen: Zum semantischen Wandel in der hebräischen Komponente der südwestdeutschen Viehhändlersprache. In: Klaus Siewert (Hrsg.), *Rotwelsch-Dialekte: Symposion Münster, 10. – 12. März 1995: 43-58.* Wiesbaden: Harrassowitz.
Pfeifer, Wolfgang [Ltg.] (²1993/1995). *Etymologisches Wörterbuch des Deutschen.* München: dtv.
Scheer-Nahor, Friedel (1998/99). *Hebraismen im Badischen Wörterbuch.* Freiburg, Magisterarbeit. (Wortliste unter http://www.scheer-nahor.de/wortlist.pdf.)
Ternes, Katharina (2011). Entwicklungen im deutschen Wortschatz. *Glottometrics 21, 25-53.*

Software

NLREG. Nonlinear Regression Analysis Program. Ph. H. Sherrod. Copyright (c) 1991–2001.

Glottometrics 27, 2014, 18-53

Some aspects of Slavic phonemics and graphemics

Emmerich Kelih, Vienna
Ioan-Iovitz Popescu, Bucharest
Gabriel Altmann, Lüdenscheid

Abstract. For the lowest linguistic level (phonemes, graphemes) some indicators of the given systems are presented and compared in 12 Slavic languages. In some cases, the divergence of the family can be shown.

Keywords: *lambda, repeat rate, entropy, ranking, Ord's criteria, Gini's coefficient, Pearson's excess, stratification, comparisons, control cycle.*

The lowest level of language is considered here as that concerning phonemes, letters (in languages using an alphabet) and graphemes. Letters and graphemes need not coincide as is well known from European languages (e.g. English). For all types of entities the frequencies can be computed, ranked in the usual way and the rank-frequency distribution can be characterized either as a distribution (usually some type of Zipfian distribution) with all its properties, or by means of some indicators expressing some further properties. Usually one computes the entropy and/or the repeat rate expressing the degree of non-uniformity of the occurrence of letters/graphemes/phonemes. Here we shall apply several indicators and compare or order Slavic languages.

As is well known, neither sounds, phonemes, letters or graphemes (practically nothing in language) are distributed uniformly because there is a need for redundancy which causes a certain excess in the rank-frequency distribution. But graphemes/letters may have slightly different properties because they are secondary constructions. The counting of sounds is unproductive because sounds have only intervals of measurable properties which may be different for each speaker. Problematic is also the computation of letter frequencies in English because written English uses today rather a hieroglyphic script whose components (motifs) are made of Latin letters. Nevertheless, they can be identified unequivocally.

Here we begin with considering the properties of the distribution of graphemes and phonemes in Slavic languages, a problem known from many publications. It is not our aim to propose a new distribution model, because there are a number of them. In order to allow further investigations we present the data in the Appendix. They are taken from Chapter 1 of the novel *Kak zakaljalas stal'* (How the steel was tempered) by Ostrovskij written in Russian and translated into all Slavic languages (cf. Kelih 2009a, 2009b). For the analysis of the grapheme and phoneme frequencies the word types in the above mentioned texts have been used.

The results can be used for description, typology, areal and historical study, etc. This restriction warrants homogeneity of data that cannot be attained using a corpus. In a study of this nature, the text-sort should be the same in all languages under study, with almost similar text size, etc. So if there are particular influences in text, here only some background laws may be presented. If there are some links between the properties, then outliers signaling a disturbance will be shown. The "disturbance" may mean either an innovation which is an element of self-organization, or, more frequently, especially with graphemes, it is a kind of retardation causing a disharmony with the phonetic development. Classical examples are English and French. The written language cannot easily leave a strong attractor, a circumstance causing ever greater difficulties and efforts in learning to write.

Lambda

Let us begin with the lambda-indicator proposed and used previously already for higher units (lemmas, words, etc.), cf. Popescu, Čech, Altmann (2011); Popescu, Zörnig, Altmann (2013); Popescu, Mačutek, Altmann (2009, 2010).

The lambda indicator is a function of the arc length between the neighboring ordered (ranked) frequencies. The components of the arc are defined as

$$L_r = \sqrt{[f(r) - f(r+1)]^2 + 1}, \qquad (1)$$

viz. as the Euclidean distances between the neighboring frequencies, and their sum is the arc

$$L = \sum L_r = \sum_{r=1}^{V-1} \sqrt{(f(r) - f(r+1))^2 + 1} \qquad (2)$$

where *V* is the inventory of entities (= greatest rank). Since *L* depends strongly on text size, in the literature it is relativized in different ways: either dividing it by its maximum or simply by *N*, the text size. However, there still remains a trace of dependence which can be partially removed by defining rather

$$\Lambda = \frac{L}{N} log_{10} N\,. \qquad (3)$$

Other modifications concerning word frequencies are used, too.

The rank-frequency distribution of graphemes/phonemes is defined as a pair <*r*, *f*(*r*)> where *r* is the rank and *f*(*r*) is the frequency at rank *r*. It is irrelevant whether one uses a corpus or a dictionary. The choice merely modifies the result.

For the sake of illustration let us consider the phonemes in the Slovene version of the first chapter of the novel *Kak zakaljalas stal'* by Ostrovskij (see Appendix):

[1361, 1103, 1100, 1038, 839, 718, 660, 582, 531, 516, 467, 445, 388, 381, 381, 264, 263, 239, 233, 180, 179, 174, 103, 92, 86, 60, 24, 17].

The arc can be computed as

$$L = [(1361 – 1103)^2 + 1]^{1/2} + [(1103 – 1100)^2 + 1]^{1/2} + \ldots + [(24 – 17)^2 + 1]^{1/2} = = 1346.6339$$

Since $N = 12424$, using (3) we obtain $\Lambda = 1346.6639(\log_{10}12424)/12424 = 0.4438$.

First we test the hypothesis that *the smaller the inventory of graphemes/phonemes, the greater is the lambda-indicator*. The hypothesis follows from the requirement of language carriers to create sufficient redundancy. In small inventories, this can be done by emphasizing some phonemes, viz. rendering their frequencies higher than it is usual with elements of large inventories (e.g. that of words). Thereby lambda increases. Needless to say, the dependence cannot be quite smooth because every language has its dynamic history, borrowing from other languages, trends, different text-sorts, etc. Besides, inventory and redundancy are elements of the synergetic control cycle (cf. Köhler 2005) that must be held in equilibrium.

As is well known, phoneme/grapheme and letter frequencies are formed differently. A full 1:1 correspondence, phoneme = letter, is rather an exception. Though steps in still deeper levels are possible, e.g. in sounds, distinctive features and muscle effort of sounds, graphical motifs of letters or (iconic, symbolic) signs, we restrict ourselves to those for which there are many available data. Let us begin with phoneme frequencies of the above mentioned 12 Slavic languages.

Consider first the phonemes in 12 Slavic languages. In Table 1 they are ordered according to increasing inventory *V* of phonemes. The inventory is defined by actually occurring phonemes in the text, thus in some cases (e.g. Slovene phoneme inventory consists of 29 phonemes, whereas in the text only 28 of them are realized) differences between systemic inventory size and the observed units are obtainable. A detailed discussion of the problems of the determination of the grapheme inventory size for the Slavic languages can be found in Kelih (2013: 57-61). For the performed analysis the same principles are applied.

Table 1
The Λ indicator of phoneme frequencies in 12 Slavic languages
(First chapter of the novel by Ostrovskij)

Language	N	V	L	Var(L)	Λ	Var(Λ)
Slovene	12424	28	1346.6339	3859.04580360	0.4438	0.00041909
Serbian	11529	31	1384.4265	6994.31412495	0.4877	0.00086815
Croatian	11792	31	1425.2402	6395.67843232	0.4921	0.00076250
Macedonian	10698	32	1447.0694	6788.04312822	0.5450	0.00096294
Ukrainian	12581	36	1252.2998	2716.63888167	0.4081	0.00028848
Upper-Sorbian	12609	37	1173.6440	2882.70773201	0.3817	0.00030490
Czech	11070	40	978.9233	1375.74842633	0.3576	0.00018361
Bulgarian	11219	42	1648.8832	13353.23548129	0.5952	0.00174012
Russian	13068	42	1432.8106	4906.15344372	0.4513	0.00048676
Polish	12697	42	1211.0696	3630.87709420	0.3914	0.00037928

Belorussian	12950	43	2514.5365	45516.09520476	0.7985	0.00458975
Slovak	11857	46	1069.7772	1577.28112948	0.3676	0.00018621

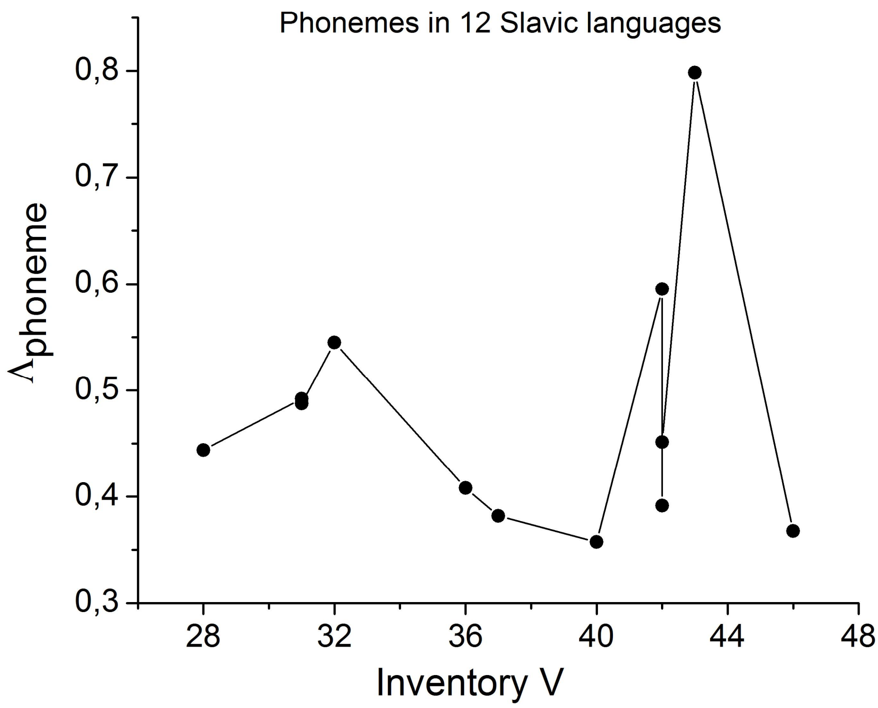


Figure 1. <V,Λ> for phonemes in Slavic languages

As can be seen in Figure 1, Λ does not depend of V. Belorussian is clearly an outlier[1] but Bulgarian and Macedonian also display a diverging trend. Comparing the greatest Λ = 0.7985 in Belorussian with the smallest one (Λ = 0.3817) in Upper-Sorbian using the asymptotic normal test and computing the variances directly from the empirical data we apply

$$u = \frac{|\Lambda_1 - \Lambda_2|}{\sqrt{Var(\Lambda_1) + Var(\Lambda_2)}}, \tag{4}$$

yielding in our case

$$u = |0.7985 - 0.3817|/(0.00458975 + 0.00030490)^{1/2} = 5.96,$$

[1] The outstanding behaviour of Belorussian has already been noticed by Kelih (2012). Belorussian is – in comparison to other Eastern Slavic languages – known for its mainly phonetically determined orthography, whereas for instance Russian and Ukrainian are governed by phonemic and morphologic orthographic principles. Some further explanations are given at the end of the paper.

which is significant and shows that the Slavic languages diverge in their use of phonemes. If we compare Upper-Sorbian with Bulgarian (Λ = 0.5925) and use (4), we obtain

$$u = |0.5952 - 0.3817|/(0.00174012 + 0.00030490)^{1/2} = 4.72,$$

which is significant, too, and shows the phonemic disintegration of this family.

If we want to compare two languages, we may take the mean of all lambdas in one language and compute their variance directly from the data. One can, of course, pool the different data to obtain a common variance, one can compute the degrees of freedom in a special way, one can use a slightly more exact test using theoretical variances (cf. Zörnig 2014), but we make the computation as simple as possible.

Table 2 contains the results based on grapheme frequencies in Slavic languages using the data in Appendix.

Table 2
Lambda for graphemes in 12 Slavic languages

Language	**N**	**V**	**L**	**Var(L)**	**Λ**	**Var(Λ)**
Slovene	12424	25	1431.7879	6375.15815015	0.4718	0.00069234
Serbian	11529	30	1383.6983	5373.52083866	0.4875	0.00066698
Croatian	11792	30	1424.1910	5253.28328107	0.4918	0.00062630
Bulgarian	11063	30	1408.4964	8657.99874314	0.5148	0.00115682
Macedonian	10700	31	1448.6484	7378.21575330	0.5455	0.00104631
Russian	13081	33	1356.6461	3724.47519235	0.4269	0.00036887
Ukrainian	12545	33	1145.2940	1761.75563071	0.3742	0.00018804
Belorussian	12982	33	2031.1878	55758.16553760	0.6436	0.00559777
Czech	10983	40	919.9590	1067.43074920	0.3385	0.00014448
Slovak	12057	42	1080.7223	1327.70384251	0.3658	0.00015213
Polish	13635	32	1197.1586	2504.10350641	0.3630	0.00023026
Upper-Sorbian	13002	34	1173.0818	2813.16779987	0.3712	0.00028165

Here, again, Belorussian is an outlier. Without it we obtain a decreasing trend as shown in Figure 2. Comparing again the greatest (Macedonian = 0.5455) and the smallest (Slovak = 0.3658) Λ we obtain

$$u = |0.5455 - 0.3558|/(0.00104631 + 0.00015213)^{1/2} = 5.48$$

a highly significant difference.

Here at least a slight dependence of Λ on inventory V can be traced down, but it can be captured only with a polynomial function, even if one omits Belorussian. In this sense, the Slavic family diverges, too.

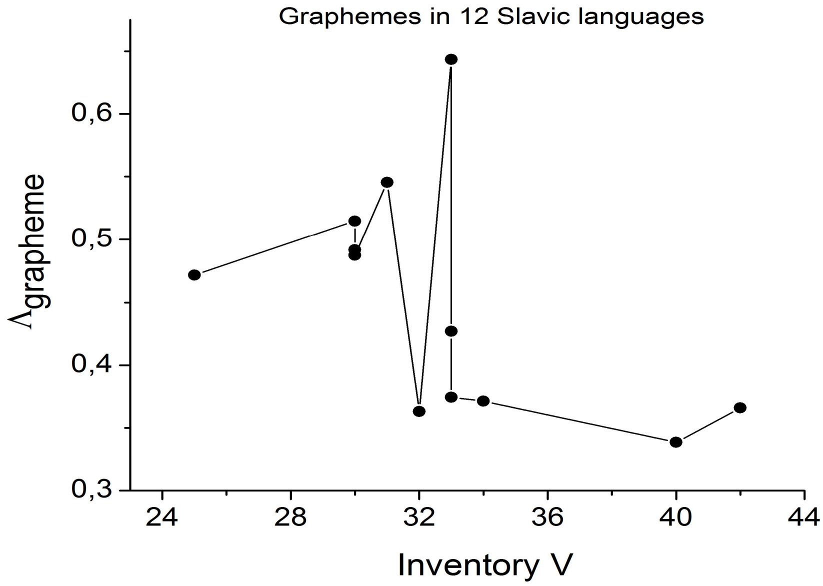


Figure 2. <V, Λ> for graphemes in Slavic languages

As a next problem we compare the Lambdas of the phonemic and the graphemic frequencies. If there is no divergence between the two levels, then the values will be quite near to one another. In Figure 3 the individual values of *Λ* can be seen. In most cases the *Λ* of phonemes is greater than that of graphemes. For the lowest level of language it preliminarily holds that *the level having the greater inventory has smaller lambdas.*

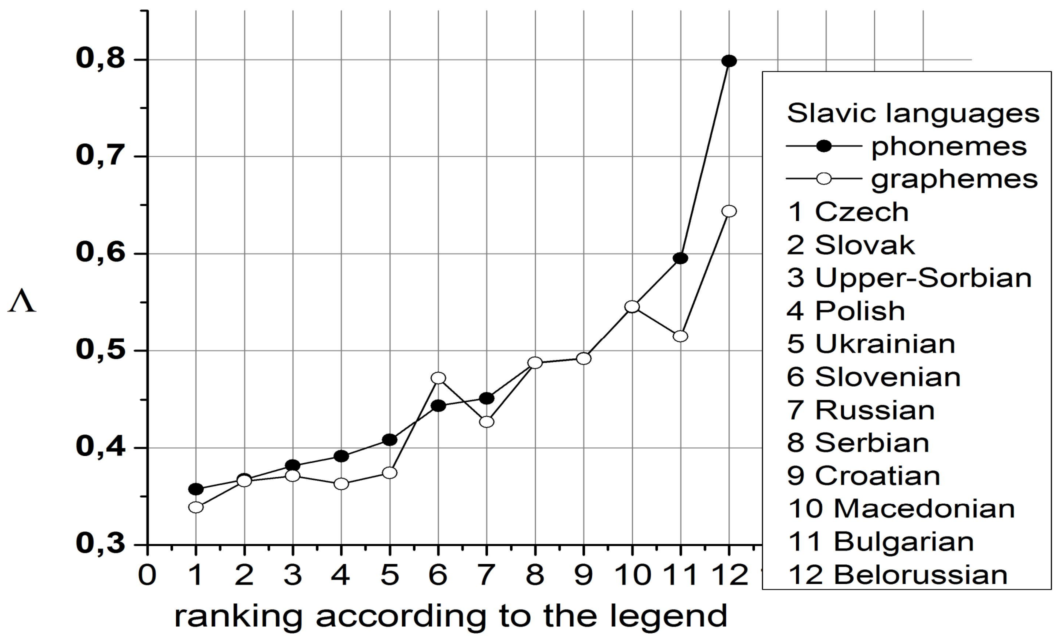


Figure 3. Lambdas of phoneme and grapheme frequencies in 12 Slavic languages

If we look at the relationship between the Lambdas of phonemes and graphemes (in the same language), we can simply state that there is a strong correlation as can be seen in Figure 4. It is not linear but if we omit Belorussian as an outlier (see Figure 5), it can be made linear.

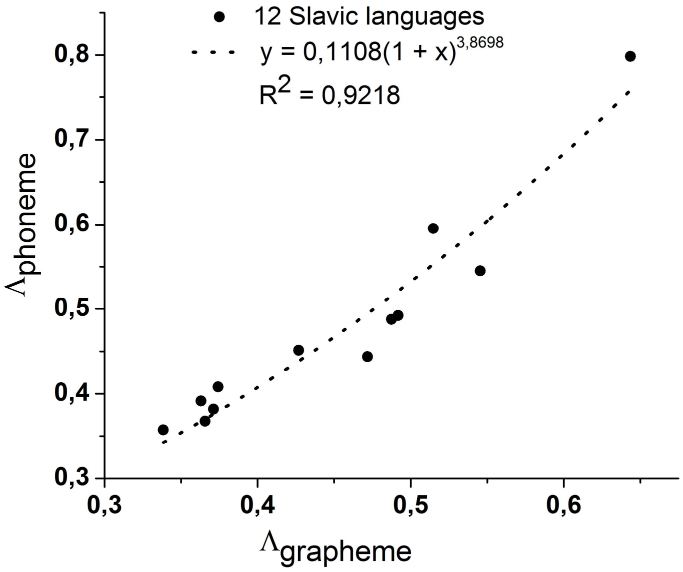


Figure 4. Non-linear relationship between the Lambdas of graphemes and phonemes

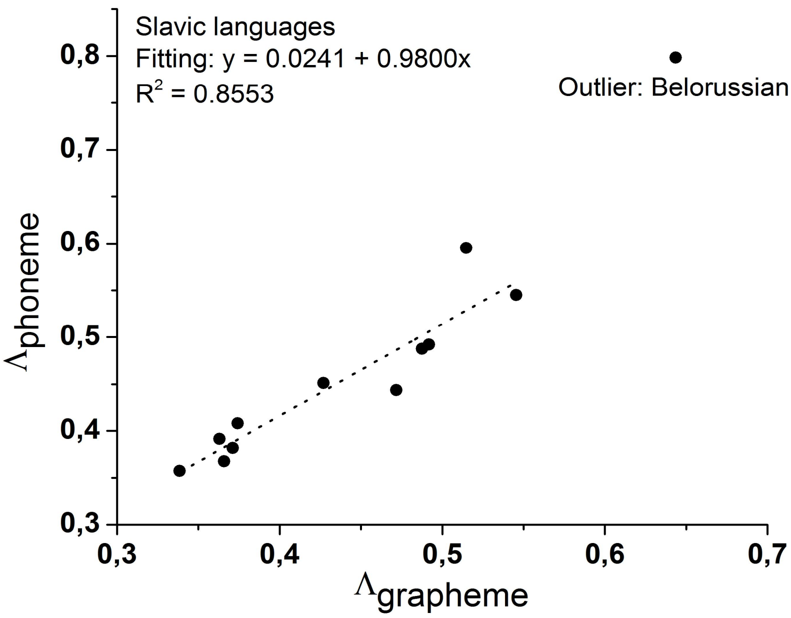


Figure 5. Linear relationship between the Lambdas of graphemes and phonemes

The non-linearity is more probable if we consider languages like English or French. Nevertheless, further languages are necessary to obtain a more sophisticated answer.

Repeat rate

The repeat rate is a measure of concentration. The more the frequencies are concentrated on a small number of entities, the greater it is. Hence, at the same time, it shows the deviation of the distribution from uniformity. This indicator is defined as

$$RR = \frac{1}{N^2}\sum_{r=1}^{V} f(r)^2 \qquad (5)$$

and it moves in the interval $<1/V; 1>$. Usually, one relativizes it as

$$RR_{rel} = \frac{1 - RR}{1 - 1/V}. \qquad (6)$$

It has been used frequently in word and grapheme frequency studies. Its variance is defined as

$$Var(RR) = \frac{4}{N}\left(\sum_{r=1}^{V} p_r^3 - RR^2\right), \qquad (7)$$

where N is the sample size and $p_r = f(r)/N$ (cf. e.g. Altmann, 1988; Popescu et al. 2009).

For the 12 Slavic languages we obtain the results presented in Table 3.

Table 3
Repeat Rate in 12 Slavic languages

Language	Phonemes			Graphemes		
	V	**RR**	**Var(RR)**	**V**	**RR**	**Var(RR)**
Slovene	28	0.0592	0.00000255	25	0.0617	0.00000278
Serbian	31	0.0598	0.00000291	30	0.0606	0.00000297
Croatian	31	0.0594	0.00000281	30	0.0602	0.00000286
Bulgarian	42	0.0634	0.00000349	30	0.0600	0.00000304
Macedonian	32	0.0654	0.00000378	31	0.0668	0.00000387
Russian	42	0.0513	0.00000200	33	0.0514	0.00000183
Ukrainian	36	0.0504	0.00000188	33	0.0491	0.00000174
Belorussian	43	0.0664	0.00000408	33	0.0540	0.00000244
Czech	40	0.0436	0.00000159	40	0.0439	0.00000157
Slovak	46	0.0438	0.00000152	42	0.0485	0.00000179
Polish	42	0.0446	0.00000149	32	0.0487	0.00000155
Upper-Sorbian	37	0.0465	0.00000162	34	0.0469	0.00000156

It can be shown that for phonemes there is no dependence of *RR* on the inventory; with graphemes, there is a slight linear decrease of *RR* with increasing *V*. However, if we omit some outliers, we obtain, as can be expected, “better” results. In some Slavic languages (Bulgarian, Macedonian, Belorussian) there is a slight disequilibrium which is compensated by some other properties. The same holds for the graphemes where a (decreasing) dependence is overt but there are several outliers, too.

We can conjecture that graphemes represent a set whose frequencies are more influenced by the inventory than that of phonemes. This conjecture can, of course, be tested only if many languages are considered. However, it can be shown that the differences between the greatest and the smallest *RR* both in phonemic and graphemic view respectively are significant. The Slavic languages are in this respect divergent.

The relationship between *RR* and Λ can be computed using the above tables. Omitting Belorussian we obtain for graphemes a simple linear relationship $RR = 0.0103 + 0.1020\Lambda$ with $R^2 = 0.94$; including Belorussian, the linear relationship reduces to $R^2 = 0.52$ but the *F*-test is still significant. One can capture the complete data with the Lorentzian function yielding $RR = a/[1 + ((\Lambda-b)/c)^2]$, yielding $a = 0.0625$, $b = 0.5356$, $c = 0.3000$ and $R^2 = 0.91$. In any case we see that the two indicators are not independent. The plotting of the dependence is shown in Figure 6.

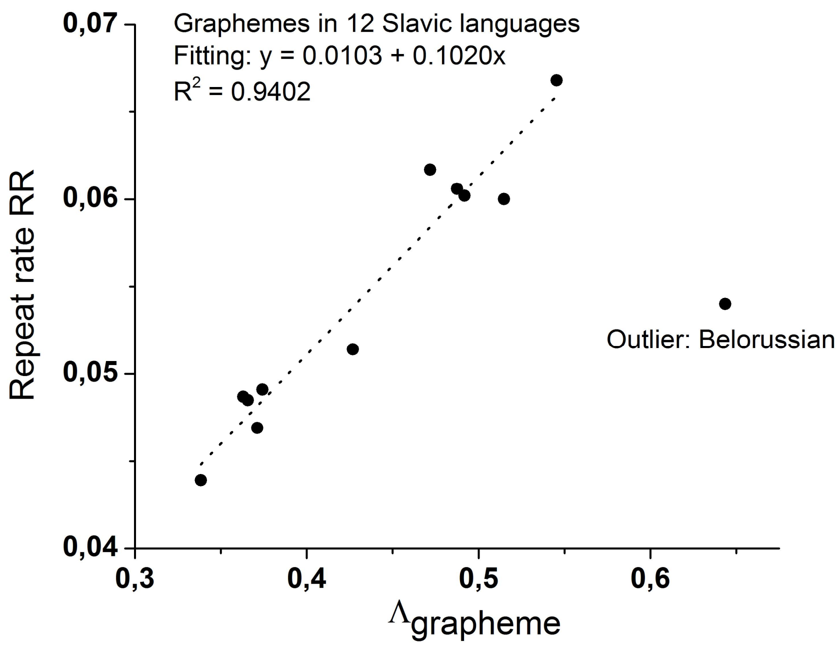


Figure 6. The link between *RR* and Lambda for graphemes

For phonemes we obtain $RR = 0.0086 + 0.1002\Lambda$ with $R^2 = 0.85$. If we insert also Belorussian, the *F*-test remains significant and the determination coefficient reduces to $R^2 = 0.52$. Using the Lorentzian function, we obtain $a = 0.0705$, $b = 0.6844$, $c = -0.4244$ and $R^2 = 0.90$. Still better fitting can be obtained, e.g. using the Zipf-Alekseev function. Again, *RR* and the Lambda for phonemes are not independent. The plotting is presented in Figure 7.

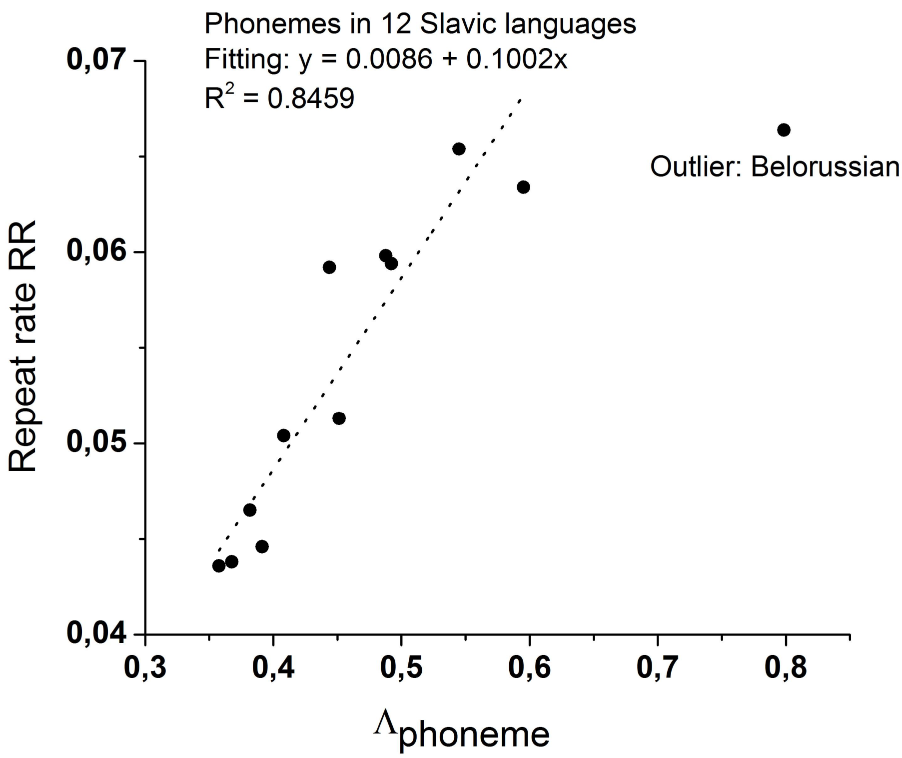


Figure 7. The link between *RR* and Lambda for phonemes

Entropy

For our purposes, entropy is also an indicator of uncertainty, here non-uniformity. Hypothetically, the greater the repeat rate, the smaller the entropy, and at the same time, the greater the lambda indicator the smaller the entropy. Thus one can characterize a distribution's non-uniformity and compare it with other samples/languages using the entropy. Usually, non-uniformity is tested simply by the chi-square test but since this increases with the increase of the sample size, it is not quite reliable. Besides, we do not want to test the uniformity but characterize the non-uniformity.

Entropy is defined in many forms (cf. Esteban, Morales. 1995); here we use the Shannon version given as

$$H = -\sum_{r=1}^{V} p_r \log_2 p_r \,, \tag{8}$$

where again, $p_r = f(r)/N$. The variance of the Shannon-entropy is

$$Var(H) = \frac{1}{N}\left(\sum_{r=1}^{V} p_r \log_2{}^2 p_r - H^2\right). \tag{9}$$

For the phoneme/grapheme level in 12 Slavic languages we obtain the results presented in Table 4. Usually one computes the relative entropy defined as

$$H_{rel} = \frac{H}{H_0} = \frac{H}{\log_2 V} \tag{10}$$

and

$$Var(H_{rel}) = \frac{Var(H)}{(\log_2 V)^2}, \tag{11}$$

but for our purposes the raw value of *H* is sufficient because we have the same text in all languages.

Table 4
Entropies of phonemic and graphemic systems in 12 Slavic languages

Language	Phonemes			Graphemes		
	V	H	Var(H)	V	H	Var(H)
Slovene	28	4.3430	0.00147759	25	4.2689	0.00006941
Serbian	31	4.3941	0.00010422	30	4.3544	0.00009838
Croatian	31	4.4005	0.00010024	30	4.3603	0.00009438
Bulgarian	42	4.3811	0.00013893	30	4.3626	0.00009833
Macedonian	32	4.2937	0.00012620	31	4.2232	0.00011307
Russian	42	4.7196	0.00011454	33	4.5401	0.00007227
Ukrainian	36	4.6095	0.00008523	33	4.5914	0.00007175
Belorussian	43	4.5924	0.00015749	33	4.6085	0.00008595
Czech	40	4.8130	0.00009462	40	4.7408	0.00007990
Slovak	46	4.8303	0.00010020	42	4.6777	0.00010249
Polish	42	4.8198	0.00009060	32	4.5821	0.00006027
Upper-Sorbian	37	4.7369	0.00008389	34	4.6802	0.00006934

It can be shown that the entropies of phonemes and graphemes are linked with *V* in a way that could still be expressed linearly. Even here, we must reckon with outliers. Thus concerning phonemes, Bulgarian is an outlier and omitting it we obtain the relationship $H = 3.4834 + 0.0300V$ with $R^2 = 0.76$ and a significant *F*-test. For graphemes, the outlier is Macedonian. Omitting it we obtain $H = 3.5611 + 0.0293V$ with $R^2 = 0.74$ and a significant *F*-test.

Since Λ, entropy and repeat rate express the degree of non-uniformity, they may display some common trend. Taking the individual values from the above Tables 1, 2, 3 and 4 and ordering them according to respective languages, we obtain the following regressions for *H*: $H = f(\Lambda)$, $H = f(RR)$ which in positive case, hold also in the opposite direction. For the link between Lambda and *H* of *graphemes* we obtain H = 5.5079 –

2.3584*Λ* with R^2 = 0.89 but with omitting the Belorussian outlier as can be seen in Figure 8.

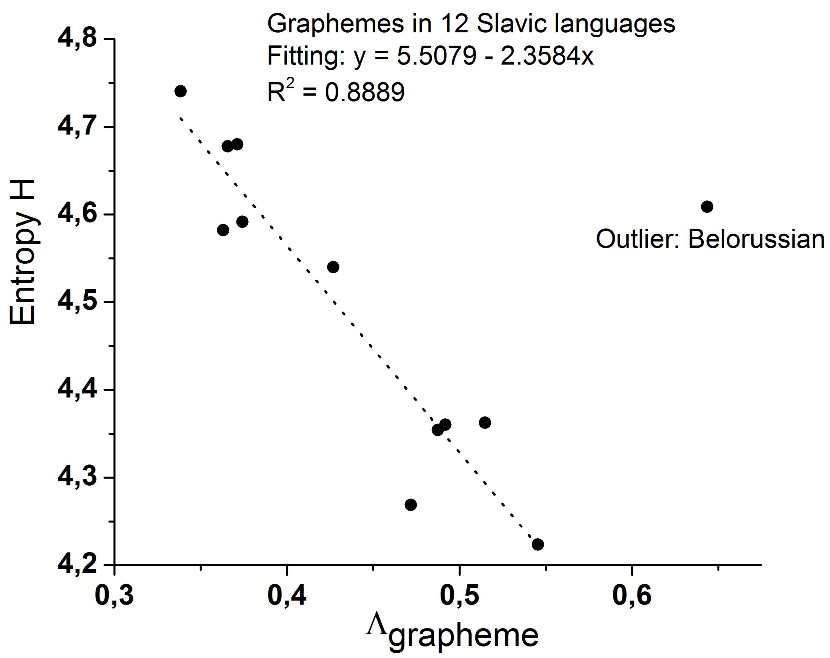


Figure 8. The link between *H* and Lambda for graphemes

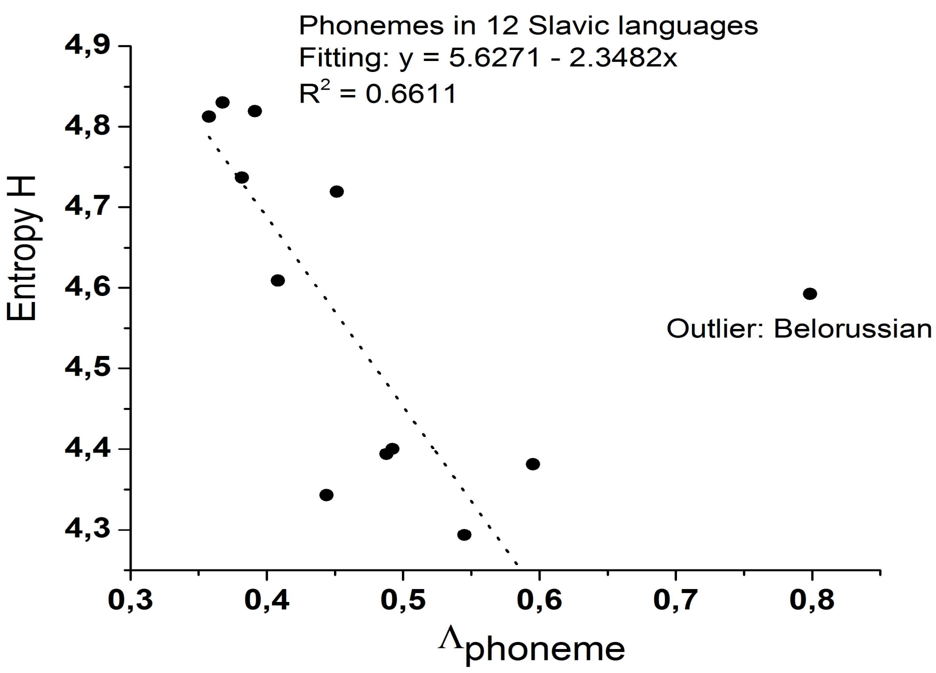


Figure 9. The link between *H* and Lambda for phonemes

The relationship between *H* and Lambda for phonemes is linear, too, of course omitting Belorussian. We obtain *H* = 5.6271 – 2.3482*Λ* with R^2 = 0.66. The result is presented graphically in Figure 9.

The link between *H* and repeat rate is as follows: For phonemes we obtain (omitting Belorussian) the results presented in Figure 10.

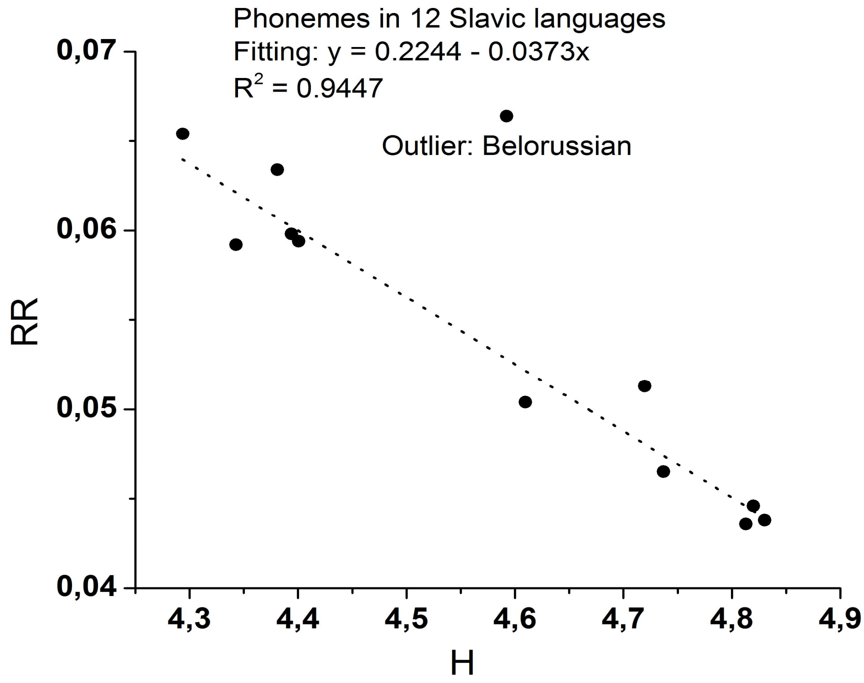


Figure 10. The link between *RR* and *H* for phonemes

And for graphemes (omitting Belorussian) in Figure 11. Both relationships are linear, the formulas of the straight lines are in the respective Figures.

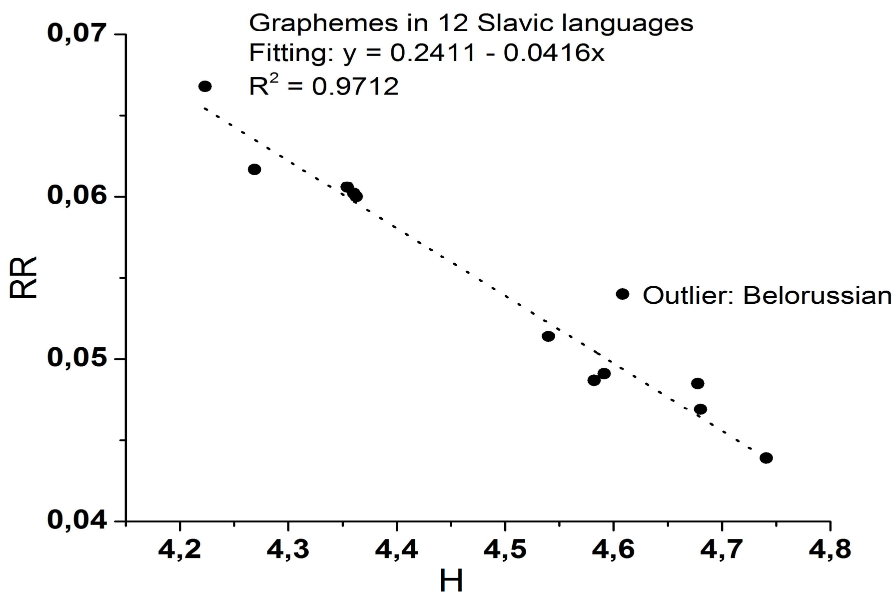


Figure 11. The link between *RR* and *H* for graphemes

Thus we obtain the control cycle in which all properties (*V, H, RR, Λ*) are linked with one another, even if we were forced to omit the outliers. Of course, even with the inclusion of outliers we would find a function with more parameters but this is *cura posterior*. In any case, the relationships must be analyzed in further languages and afterwards it will be easier to say why a certain language is an outlier.

Frequency distribution

For modeling the rank-frequency distribution of phonemes and graphemes a relatively great number of theoretical distributions have been proposed. The most frequently applied ones are Zipf *d*., geometric *d*., Good *d*., Zipf-Mandelbrot *d*., Yule *d*., Altmann's sequence, but there is none that would hold for all cases. This is perhaps caused by two circumstances: (1) There are a number of boundary conditions associated with every language; the proposed distributions or functions do not have parameters capturing this local deviation; or one did not find a general distribution. (2) As known, frequency distributions of linguistic entities represent stratified populations. Stratification can be revealed (cf. Popescu, Altmann, Köhler 2010) but it does not lead to a distribution, the two views are independent. Here we shall try to find a distribution or function common to all Slavic languages and consider the properties of the empirical distributions. We start from the unified theory (cf. Wimmer, Altmann 2005) and conjecture a very simple relationship that can be expressed in form of a differential equation

$$df(r) = -\frac{b}{r+c}dr \qquad (12)$$

i.e. the change of frequency is inversely proportional to the change of the rank. It is not necessary to involve further parameters. The solution of (12) is

$$f(r) = a - b*\log(r + c) \qquad (13)$$

Parameter *a* depends evidently on the value of the first rank, hence it is irrelevant (it is the integration constant). Parameter *c* is a modifying parameter controlling the decrease (it is a slight displacement of the rank scale). The main parameter is here *b* which depicts constancy of the decrease by ranks. As usual, it is extreme in Belorussian, the greatest value is in Polish. In Table 5 the values of the function are presented for graphemes, in Table 6 those for phonemes. As can be seen, the determination coefficient is very high in all cases.

Table 5
Fitting function (14) to the ranked sequence of graphemes
(ordered according to parameter *b*)

Language	**a**	**b**	**c**	**R^2**
Belorussian	881.1225	211.4507	-0.99574	0.9449
Upper-Serbian	1372.2791	372.3313	0.47607	0.9804
Slovak	1369.7942	375.8037	0.84486	0.9872

Bulgarian	1303.1611	378.7842	-0.15809	0.9830
Czech	1417.2351	380.8908	3.10282	0.9890
Macedonian	1373.8308	411.3184	-0.13608	0.9881
Serbian	1408.1506	411.4441	-0.00132	0.9778
Croatian	1418.8751	413.3876	-0.05303	0.9783
Ukrainian	1582.4807	438.8682	1.61928	0.9823
Russian	1631.0497	458.0895	1.11733	0.9862
Slovene	1756.7020	517.5420	0.85872	0.9818
Polish	2022.5041	558.5876	3.49243	0.9855

Table 6
Fitting function (14) to the ranked sequence of phonemes
(ordered according to parameter *b*)

Language	**a**	**b**	**c**	$\mathbf{R^2}$
Belorussian	969.8745	257.2112	-0.99749	0.9769
Czech	1098.8524	293.7206	0.40911	0.9929
Slovak	1128.9136	300.4645	0.09619	0.9911
Polish	1150.9590	305.7587	-0.26212	0.9658
Bulgarian	1138.7241	322.7730	-0.78186	0.9796
Upper-Serbian	1221.7867	328.8419	-0.04801	0.9614
Russian	1225.7875	334.3230	-0.57187	0.9548
Ukrainian	1353.7021	375.3046	0.14973	0.9776
Macedonian	1295.4153	383.8657	-0.32835	0.9858
Serbian	1350.5137	391.5668	-0.14831	0.9714
Croatian	1362.2198	393.8406	-0.19290	0.9727
Slovene	1817.4854	532.2674	1.35312	0.9895

It has to be remarked that modeling with the aid of a distribution or a function are merely two tentative approaches approximating some real phenomenon. They do not express “truth” but our concept formation. To work with a function (sequence), i.e. without normalization, is simpler than with a distribution in which one must frequently consider also classes with zero frequency and test with a chi-square which is not appropriate for great sample sizes. It is misleading especially in classes with small frequencies.

Ord’s criteria

J.K. Ord (1972) proposed an indicator based on the first three moments of the distribution ascribing the data a place in Cartesian coordinates. It has been frequently used especially in text analysis. The indicators are

$$I = \frac{m_2}{m_1'}, \qquad S = \frac{m_3}{m_2}, \qquad (14)$$

where m_1‘ is the mean and m_2, m_3 are the second and third central moments. If we compute the moments, we obtain the results presented in Table 7 and displayed graphically in Figures 12a and 12b. If there is some order in the data, then the points are placed in a small domain or directly on a straight line.

Table 7
Ord’s criteria for phonemes and graphemes.

Phonemes

Language	V	m_1‘	m_2	m_3	I	S
Belorussian	43	10.3647	86.6890	717.6720	8.3639	8.2787
Bulgarian	42	8.3789	49.7599	401.6300	5.9387	8.0714
Croatian	31	8.6251	47.4543	299.0798	5.5019	6.3025
Czech	40	11.2977	81.3224	647.3048	7.1981	7.9597
Macedonian	32	7.9354	43.0041	292.7511	5.4193	6.8075
Polish	42	11.4775	86.6887	664.0608	7.5529	7.6603
Russian	42	10.6448	85.5212	787.8290	8.0341	9.2121
Serbian	31	8.5726	47.1407	301.1275	5.4990	6.3878
Slovak	46	11.3648	84.3103	700.4570	7.4186	8.3081
Slovene	28	8.3169	40.5180	220.9837	4.8718	5.4540
Ukrainian	36	9.9399	61.2163	388.0371	6.1586	6.3388
Upper-Sorbian	37	10.8865	74.8883	514.0581	6.8790	6.8643

Graphemes

Language	V	m_1‘	m_2	m_3	I	S
Belorussian	33	10.4865	64.2783	311.0694	6.1296	4.8394
Bulgarian	30	8.4206	44.1674	258.1464	5.2452	5.8447
Croatian	30	8.4163	43.6911	248.5775	5.1912	5.6894
Czech	40	10.9044	67.1346	426.5093	6.1567	6.3530
Macedonian	31	7.6230	36.5160	207.6706	4.7902	5.6871
Polish	32	9.8516	53.1699	288.2328	5.3971	5.4210
Russian	33	9.4860	52.2423	316.3570	5.5073	6.0556
Serbian	30	8.3651	43.4747	252.8360	5.1972	5.8157
Slovak	42	10.1885	67.5630	575.3327	6.6313	8.5155
Slovene	25	8.0223	37.0356	178.7085	4.6166	4.8253
Ukrainian	33	9.8677	55.2429	316.3718	5.5984	5.7269
Upper-Sorbian	34	10.5212	67.1973	425.3698	6.3869	6.3302

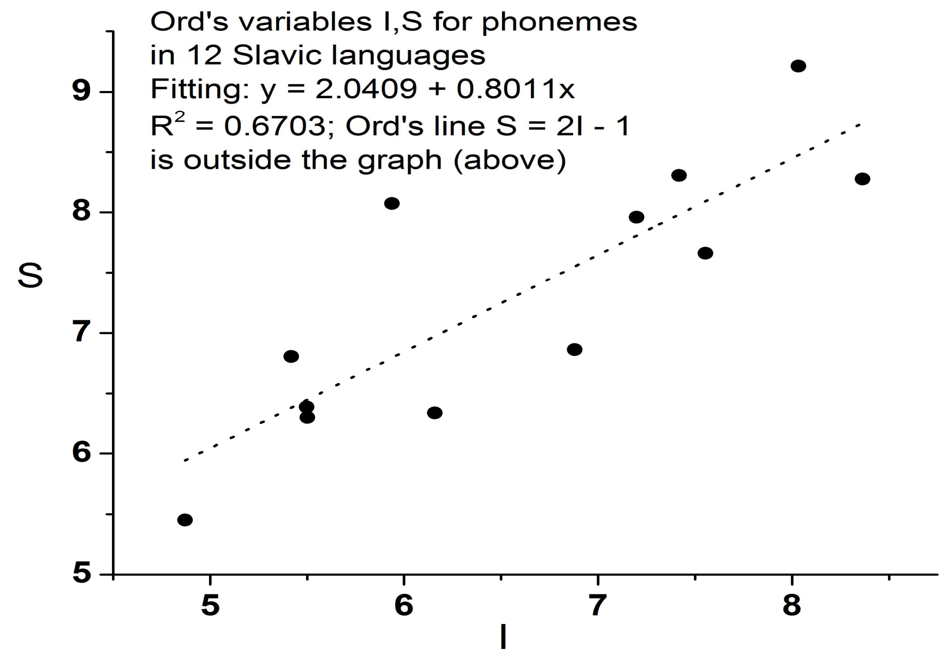


Figure 12a. <I,S> for phonemes

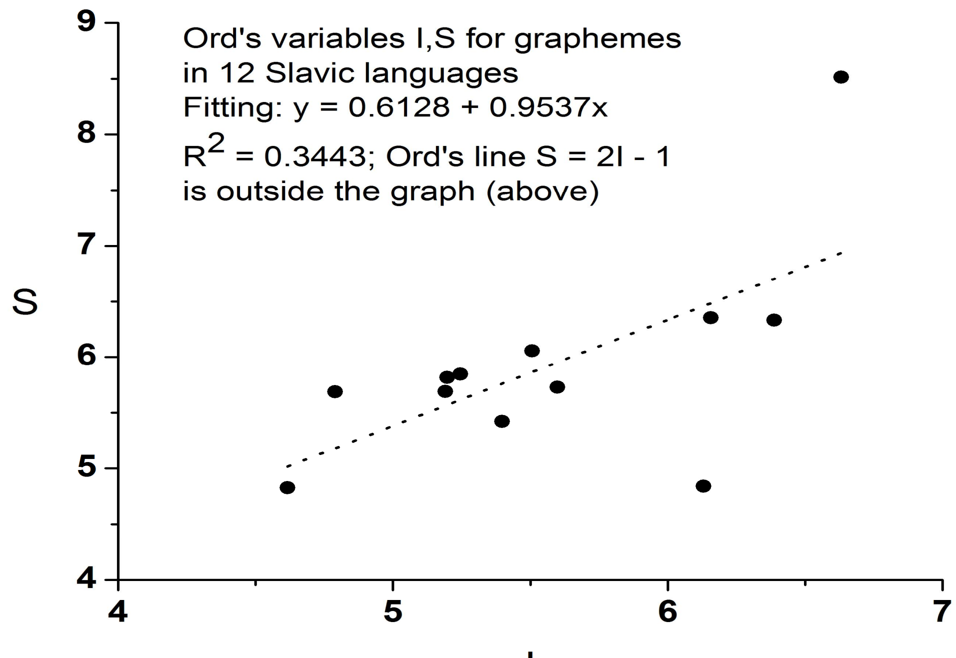


Figure 12b <I,S> for graphemes

Since all S values are placed below the line $S = 2I - 1$ determining the upper boundary of the negative hypergeometric distribution, we can state at least the domain where most probably the phoneme and grapheme frequencies of all languages are placed. Testing the adequacy of the negative hypergeometric distribution using the chi-square test is, of course, irrelevant because the sample sizes are enormous.

Whatever indicator has been computed, one could observe the existence of outliers which are easily seen in the Figures. This fact merely shows that there are some local boundary conditions which should be taken into account. Adding more languages, perhaps we succeed in finding the force influencing the given deviation (cf. Köhler 2005).

The excess

Rank-frequency distributions or ordered sets have a number of other properties which can be used for comparisons. All previous indicators of Lambda, entropy and repeat rate can be considered at the same time as indicators of the excess of the distribution. Usually the greater the excess or kurtosis is, the greater is Lambda and repeat rate and the smaller is entropy. But there is a classical Person's coefficient of excess defined as

$$\beta_2 = \frac{m_4}{m_2^2}, \qquad (15)$$

Or comparing it with the normal distribution, one subtracts 3 from (15). Since ranked frequencies decrease monotonously, we will compute (15) for all Slavic languages.

Table 8a
Person's excess of graphemes of Chap.1 of Ostrovskij's novel KZS

Language (alphabetically)	**N**	**V**	**m_2**	**m_4**	**β_2**
Belorussian	12982	33	64.2783	9810.7847	2.3745
Bulgarian	11063	30	44.1674	5589.3597	2.8652
Croatian	11792	30	43.6911	5603.0029	2.9352
Czech	10983	40	67.1346	12354.8942	2.7412
Macedonian	10700	31	36.5160	4126.5923	3.0947
Polish	13635	32	53.1699	7567.9362	2.6770
Russian	13081	33	52.2423	7873.4229	2.8848
Serbian	11529	30	43.4747	5638.4235	2.9832
Slovak	12057	42	67.5630	15785.3389	3.4581
Slovene	12424	25	37.0356	3633.4449	2.6490
Ukrainian	12545	33	55.2429	8384.3189	2.7474
Upper-Sorbian	13002	34	67.1973	11754.1049	2.6031

Table 8b
Person's excess of phonemes of Chap.1 of Ostrovskij's novel KZS

Language (alphabetically)	**N**	**V**	**m_2**	**m_4**	**β_2**
Belorussian	12950	43	86.6890	21181.2383	2.8185
Bulgarian	11219	42	49.7599	9840.4944	3.9743
Croatian	11792	31	47.4543	6983.2629	3.1010
Czech	11070	40	81.3224	19349.4407	2.9258
Macedonian	10698	32	43.0041	6095.5609	3.2960
Polish	12697	42	86.6887	20497.8580	2.7276
Russian	13068	42	85.5212	22828.0356	3.1212
Serbian	11529	31	47.1407	6959.0793	3.1315
Slovak	11857	46	84.3103	21609.6449	3.0401
Slovene	12424	28	40.5180	4706.1317	2.8666
Ukrainian	12581	36	61.2163	10377.8929	2.7693
Upper-Sorbian	12609	37	74.8883	15183.5897	2.7074

Gini's coefficient

Gini's coefficient is, as a matter of fact, the space between the Lorenz curve and the diagonal of the Cartesian coordinate system. A simple computation can be performed using the formula

$$G = \frac{1}{V}\left(V + 1 - 2m_1'\right), \qquad (16)$$

which, for large *V*, can be simplified in

$$G = 1 - \frac{2m_1'}{V}. \qquad (17)$$

The formula can be used for different purposes (cf. Popescu et al. 2009: 54 ff.), here we may measure with it the divergence from the uniformity of frequencies. The greater is *G*, the greater is the divergence of the frequencies. As a matter of fact, the greater is *G*, the smaller is the entropy and the greater is the repeat rate. Hence *G* can be used alternatively. For comparative purposes one can use the variance of *G* which is simply

$$Var(G) = \frac{4\sigma^2}{V^2 N}, \qquad (18)$$

where σ^2 is the variance of the independent variable (rank), *N* is the sample size and *V* is the inventory size.

For the Slavic languages we obtain the results presented in Table 9.

Table 9
Gini's coefficient for phoneme and grapheme frequencies in Slavic languages

Language	Phonemes				Graphemes			
	N	V	G	Var(G)	N	V	G	Var(G)
Belorussian	12950	43	0.5412	0.00056	12982	33	0.3948	0.00095
Bulgarian	11219	42	0.6248	0.00068	11063	30	0.4720	0.00135
Croatian	11792	31	0.4758	0.00118	11792	30	0.4722	0.00126
Czech	11070	40	0.4601	0.00076	10983	40	0.4798	0.00076
Macedonian	10698	32	0.5353	0.00122	10700	31	0.5405	0.00130
Polish	12697	42	0.4773	0.00060	13635	32	0.4155	0.00096
Russian	13068	42	0.5169	0.00058	13081	33	0.4554	0.00094
Serbian	11529	31	0.4792	0.00121	11529	30	0.4757	0.00129
Slovak	11857	46	0.5276	0.00053	12057	42	0.5386	0.00063
Slovene	12424	28	0.4417	0.00138	12424	25	0.3982	0.00173
Ukrainian	12581	36	0.4756	0.00082	12545	33	0.4323	0.00098
Upper-Sorbian	12609	37	0.4386	0.00078	13002	34	0.4105	0.00089

Control cycle

If one investigates the primary language – as opposed to writing which is secondary – one expects to find properties linked to control cycles similar to those developed by R. Köhler (1986, 2005). Some of the links have been shown above but outliers disturb their exact form. However, if we expect a perfect self-regulation, its disturbance can show us the outliers, i.e. those languages which develop in some other direction and temporarily abandon the perfect equilibrium. Sometimes the "cause" may be found directly but in most cases the history of the language should be investigated, especially artificial interventions like script creation or borrowing (cf. e.g. Chinese → Japanese, Korean; or Greek/Latin → Slavic languages; or hieroglyphs → hieratic script), or conservativism: letting spoken language develop without adapting the written form (English, French,…), but also the borrowing of words which can strongly change the frequency of phonemes/graphemes. There are cases in which there is no possibility to adapt the writing, e.g. in Slovak, the preposition "s" (with) is pronounced as [s] in front of voiceless consonants, and [z] in front of voiced consonants and vowels, while the preposition "z" (from) also has two pronunciations: [s] and [z] according to the following sound. In such cases a disequilibrium may develop. Hence our aim can be merely the finding the control cycle and simultaneously the outliers for each link separately. The summary of results is presented in Tables 10a, b.

Table 10a
Indicators of phonemes

Language	**N**	**V**	**Λ**	**RR**	**H**	**G**	**I**	**S**	**β_2**	**b**
Belo-russian	12982	33	0.6436	0.0540	4.6085	0.3948	6.1296	4.8394	2.3745	211.45
Bulgarian	11063	30	0.5148	0.0600	4.3626	0.4720	5.2452	5.8447	2.8652	378.78
Croatian	11792	30	0.4918	0.0602	4.3603	0.4722	5.1912	5.6894	2.9352	413.39
Czech	10983	40	0.3385	0.0439	4.7408	0.4798	6.1567	6.3530	2.7412	380.89
Maced-onian	10700	31	0.5455	0.0668	4.2232	0.5405	4.7902	5.6871	3.0947	411.32
Polish	13635	32	0.3630	0.0487	4.5821	0.4155	5.3971	5.4210	2.6770	558.59
Russian	13081	33	0.4269	0.0514	4.5401	0.4554	5.5073	6.0556	2.8848	458.09
Serbian	11529	30	0.4875	0.0606	4.3544	0.4757	5.1972	5.8157	2.9832	411.44
Slovak	12057	42	0.3658	0.0485	4.6777	0.5386	6.6313	8.5155	3.4581	375.80
Slovene	12424	25	0.4718	0.0617	4.2689	0.3982	4.6166	4.8253	2.6490	517.54
Ukrainian	12545	33	0.3742	0.0491	4.5914	0.4323	5.5984	5.7269	2.7474	438.87
Upper-Sorbian	13002	34	0.3712	0.0469	4.6802	0.4105	6.3869	6.3302	2.6031	372.33

Table 10b
Indicators of graphemes

Language	**N**	**V**	**Λ**	**RR**	**H**	**G**	**I**	**S**	**β_2**	**b**
Belo-russian	12982	33	0.6436	0.0540	4.6085	0.3948	6.1296	4.8394	2.3745	211.45
Bulgarian	11063	30	0.5148	0.0600	4.3626	0.4720	5.2452	5.8447	2.8652	378.78
Croatian	11792	30	0.4918	0.0602	4.3603	0.4722	5.1912	5.6894	2.9352	413.39
Czech	10983	40	0.3385	0.0439	4.7408	0.4798	6.1567	6.3530	2.7412	380.89
Maced-onian	10700	31	0.5455	0.0668	4.2232	0.5405	4.7902	5.6871	3.0947	411.32
Polish	13635	32	0.3630	0.0487	4.5821	0.4155	5.3971	5.4210	2.6770	558.59
Russian	13081	33	0.4269	0.0514	4.5401	0.4554	5.5073	6.0556	2.8848	458.09
Serbian	11529	30	0.4875	0.0606	4.3544	0.4757	5.1972	5.8157	2.9832	411.44
Slovak	12057	42	0.3658	0.0485	4.6777	0.5386	6.6313	8.5155	3.4581	375.80
Slovene	12424	25	0.4718	0.0617	4.2689	0.3982	4.6166	4.8253	2.6490	517.54
Ukrainian	12545	33	0.3742	0.0491	4.5914	0.4323	5.5984	5.7269	2.7474	438.87
Upper-Sorbian	13002	34	0.3712	0.0469	4.6802	0.4105	6.3869	6.3302	2.6031	372.33

Now the dependences between individual indicators can be computed as given in Table 11. The formulas are not derived from a theoretical background but simply fitted iteratively.

Tables 11a and 11b display the situation in the Slavic family quite colorfully. The following consequences can be drawn:

(1) At the lowest level (phonemic/graphemic), the family disintegrates. There is no interrelation having the same form in all languages. Each interrelation displays one or more outliers. This simply means that the attractors develop, change their place in the system and new interrelations arise.

(2) Though we believe in the law-like character of the control cycle, the transitivity of the formulas is not given. That means, one cannot replace a symbol in an equation by the respective whole equation of another relation in order to obtain the complete control cycle. We proceeded empirically as follows: Comparing some two properties in the whole family, we omitted stepwise all languages which proved to be outliers, until we obtained a sufficient determination coefficient. Thus the interrelations have different weight within the Slavic family. In some cases, we were forced to omit maximally seven languages. If we display the relations among indicators graphically, then each relation obtains a weight which is equal to the number of languages obeying it.

(3) Here we do not furnish a theoretical substantiation to the formulas in Table 11. The number of indicators is too great and one would be forced to take into account not only phonemic/graphic criteria in order to find the boundary conditions. We are persuaded that the knowledge of all boundary conditions would yield much more coherent results, hence the present one should merely indicate the way for future research.

(4) As far as it was possible, we found a very simple formula, usually a linear relation which is more frequently represented with graphemes. All relationships can be derived from the unified theory but for the time being, we cannot predict their validity in other language families where the boundary conditions may be different. It would be helpful to have a similar analysis concerning other language families or simply individual languages. Traditionalism and simultaneous borrowing can disturb a number of equilibria, hence the study of boundary conditions in a group like the Romance languages would be rather an adventure.

(5) In Figure 13a and 13b all links classified according to their form are shown. As can be seen, the graphemic links develop towards linearity. The weight of the edge marked with a number represents the number of languages that did not display a deviation from the given relationship. Since there were 12 languages, one can see the strength of individual links. The simplification of the graphemic stratum shows that there is conscious constructive thinking in its formation – which mostly does not exist in spoken language which is full of spontaneous errors, imitations, tendencies, deliberate deviations, etc. A part of the boundary conditions could be discovered (age, gender, social stratum, education, dialect, etc.) but this presupposes enormous work for every language.

Table 11a
Outliers and the interrelation between indicators: phonemes

Relation	Formula	R^2	Outliers	Weight = 12 - Outliers
Λ - RR	RR = 0.0086 + 0.1002Λ	0.85	Bel	11
Λ - H	H = 5.9119 – 2.98512Λ	0.85	Bel Bul Sln	9
Λ – G	G = 0.46019 + 21.95Λ^9.42792	0.82	Bel Slk	10
Λ – I	I = 3.06489Λ^(–0.85909)	0.82	Bel Bul Rus Sln	8
Λ – S	S = 19.63321 – 32.00327Λ	0.81	Bel Bul Mac Rus Ser Cro	6
Λ – β_2	β_2 = 0.06041|Λ+0.72973|^11.10837 + 2.54684	0.91	Bel Cze Slk	9
Λ – b	b = 255 + 132exp(–0.5((Λ–0.5)/0.09)^2)	0.75	Sln	11
RR – H	H = 5.93464 – 25.43339RR	0.94	Bel	11
RR – G	G = 0.46927 + 0.00000494453*exp((RR–0.03166)/0.0036)	0.91	Bul Rus Slk Sln Sor	7
RR – I	I = 14.05066(1 + RR)^(–15.39951)	0.86	Bel Sln Rus	9
RR – S	S = 0.1098RR^(–1.36681)	0.91	Bel Rus Cro Ser Bul Mac	6
RR – β_2	β_2 = 2.47643 + 11.62083RR	0.69	Bel Bul Sln Pol Sor Ukr	6
RR – b	b = –1420.33421+1828.37844exp(–0.5((RR–0.05452)/0.02943)^2)	0.92	Mac Rus Sln	9
H – G	G = 0.90698 – 0.09377H	0.38	Bel Bul Slk Sln Rus	7
H – I	I = –12.58463 + 4.12706H	0.91	Bel Rus	10
H – S	S =exp(55.60953 – 23.88323H + 2.65056H^2)	0.94	Bel Bul Rus Sln	8
H – β_2	β_2 = 2.63681+ 157486000*0.01134^H	0.97	Bul Cze Rus Slk Sln	7
H – b	b = 389.58412 – 89.93838exp(–0.5((H – 4.83338)/0.11409)^2)	0.98	Bel Bul Sln	9
G – I	I = 0.60436 + 14.38158G	0.99	Bul Cro Mac Ser Slk Sln Ukr	5
G – S	S = 7.78482|G – 0.4417|^0.72965 + 5.6365	0.92	Bel Cze Pol Rus Slk Sor	6
G – β_2	β_2 = 0.25891 + 5.74083G	0.86	Bel Pol Ukr	9
G – b	b = 1541.81678 – 2373.55103G	0.90	Bul Cze Mac Pol Sor	7

I – S	S = 1.51531 + 0.86293I	0.81	Bul	11
I – β_2	β_2 = 2.87907 + 0.26084*sin(Π*(I – 0.27597)/1.14778)	0.95	Bul Mac Pol	9
I – b	b = 270108,03127exp(–I/0,68857) + 298,91455	0,87	(no outliers)	12
S – β_2	β_2 = 3,0068 + 0,28911*sin(Π*(x – 2,83204)/1,73645)	0,71	Bul Pol	10
S – b	b = 73529,72493exp(–S/0,95529) + 290,45677	0,87	(no outliers)	12
β_2 – b	b = 360,34289 + 69,27894*sin(Π*(β_2 – 0,66999)/0,63747)	0,91	Bel Cro Ser Sln Ukr	7

Table 11b
Outliers and the interrelation between indicators: graphemes

Relation	Formula	R^2	Outliers	Weight = 12 - Outliers
Λ - RR	RR = 0.0103 + 0.1020Λ	0.94	Bel	11
Λ - H	H = 5.5079 – 2.3584Λ	0.89	Bel	11
Λ – G	G = 0.21992 + 0.5356Λ	0.83	Bel Cze Slk Sln	8
Λ – I	I = 3.66045Λ^(–0.49105)	0.79	Bel Sln Pol Slk	8
Λ – S	S = 7.3938|Λ–0.70604|^0.15205	0.94	Slk Sln Pol Ukr	8
Λ – β_2	β_2 = 2.05116 + 1.80696Λ	0.74	Bel Slk Sln	9
Λ – b	b = –1643040 + 1643480exp(–0.5((Λ–0.44496)/12.19973)^2)	0.80	Pol Sln	10
RR – H	H = 5.76146 – 23.40983RR	0.97	Bel	11
RR – G	G = 0.40562 + 1530260RR^6.01079	0.91	Cze Pol Sln Mac	8
RR – I	I = 11.11697 – 96.34061RR	0.93	Cze Pol Ukr Rus Sln	7
RR – S	S = 7.7868 – 32.6627RR	0.92	Bel Slk Pol Ukr Sln	7
RR – β_2	β_2 = 1.86587 + 17.92233RR	0.77	Bel Slk Sln	9
RR – b	b = 298.09804 + 1698.821RR	0.53	Bel Sln Pol Ukr Rus	7
H – G	G = 1.656 – 0.26884H	0.86	Cze Slk Sln	9
H – I	I = –9.13503 + 3.26852H	0.81	(no outliers)	12
H – S	S = –0.4679 + 1.4411H	0.94	Bel Slk Sln Pol Ukr	7
H – β_2	β_2 = 6.44078 – 0.80215H	0.76	Bel Slk Sln	9
H – b	b = 409.14238 + 44.16093*sin(Π*(H – 0.34597)/0.13818)	0.83	Bel Pol Sln	9
G – I	I = 10.55463 – 11.27104G	0.97	Cze Mac Pol Slk Sln Sor	6
G – S	S = –2020.7638+2027.16906exp(–0.5((G–0.50236)/2.70374)^2)	0.98	Bul Cro Mac Ser Slk Sor	6
G – β_2	β_2 = 1.1838 + 3.63074G	0.91	Bel Cze Slk	9
G – b	b = 587.49349 – 208.024exp(–0.5((G–0.51058)/0.07836)^2)	0.69	Bel Sor Pol	9

I – S	S = 3.35663 + 0.4751I	0.91	Bel Pol Slk Sln Ukr	7
I – β_2	β_2 = 4.37111 – 0.27486I	0.87	Bel Pol Slk Sln	8
I – b	b = 2333.31884*I^(–0.97655)	0.95	Bel Bul Cro Mac Pol Ser	6
S – β2	β_2 = 1.56292 + 0.22334S	0.89	Cze Mac Sor	9
S – b	b = 130389.8264exp(–S/0.71482) + 366.76798	0.86	Bel Pol Rus	9
β_2 – b	b = –6409480 + 6409900exp(–0.5((β2–3.08851)/92.98925)^2)	0.85	Pol Rus Sln Ukr	8

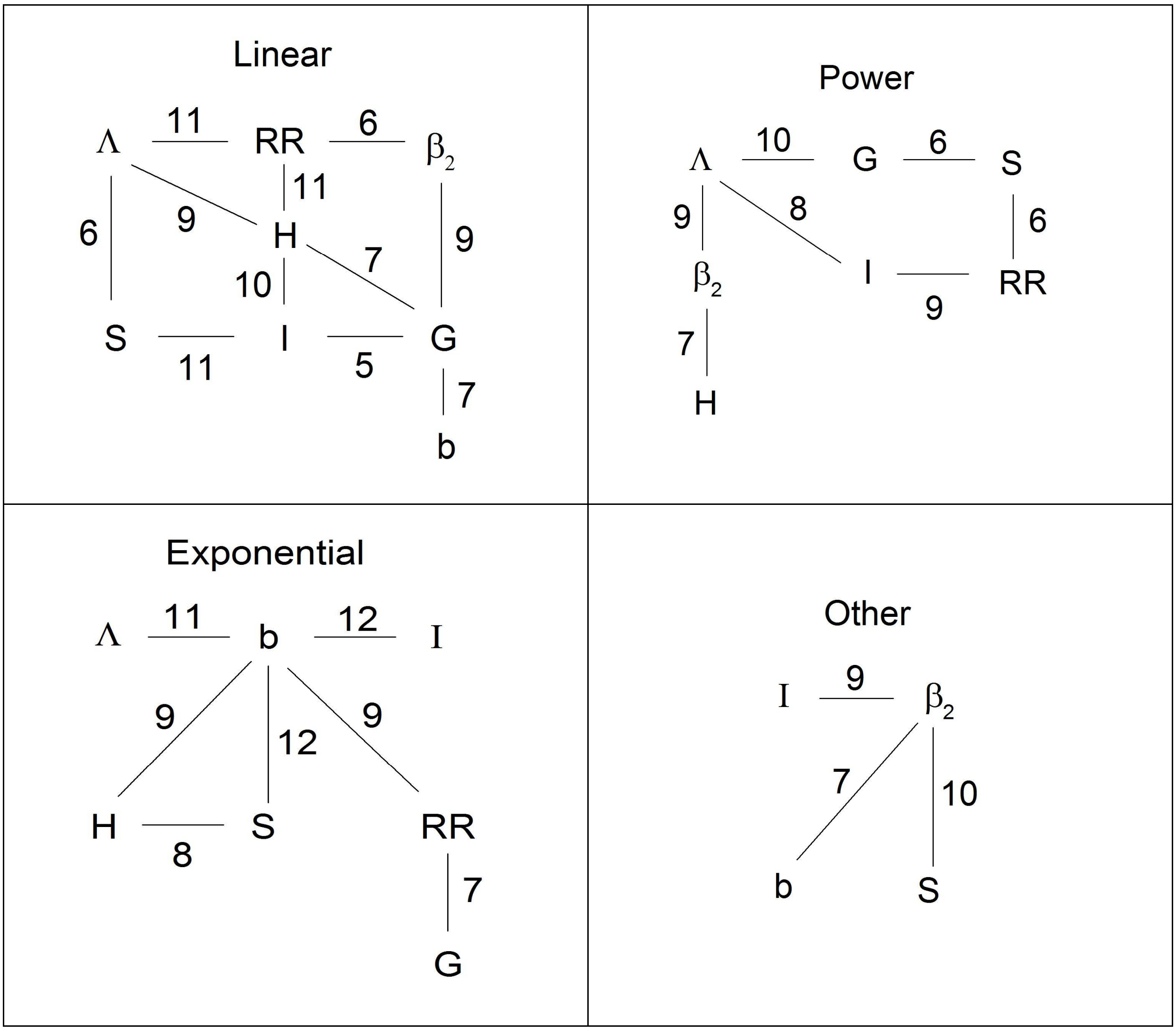


Figure 13a. Control cycle of phonemes

(6) In the next table one can see that language families usually diversify, and individual links become weaker. The causes mentioned in (5) are more or less relevant in individual languages. We obtain the results presented in Table 12 in which the degree of deviation from the family is shown. The phonemic divergence is rather "natural" and is a result of self-regulation, while the graphemic deviance is culturally conditioned and results rather from self-organization.

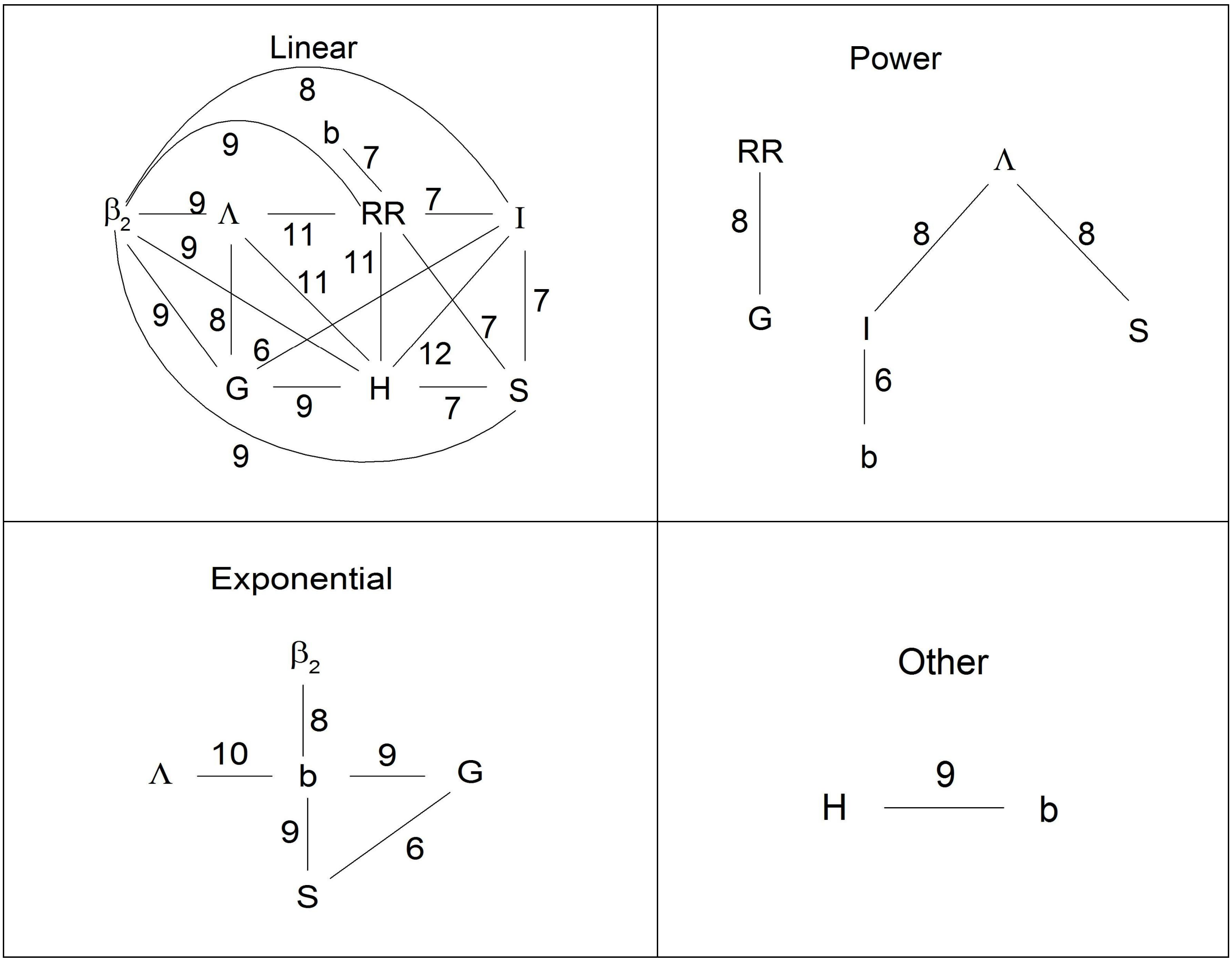


Figure 13b. Control cycle of graphemes

Table 12 (Outliers)
Number of outlier occurrences out of 28 possible cases

Phonemes			Graphemes	
Belorussian	17		Belorussian	18
Bulgarian	15		Slovene	18
Slovene	13		Polish	16
Russian	11		Slovak	14
Slovak	7		Czech	7
Macedonian	6		Ukrainian	7
Polish	6		Macedonian	5
Serbian	4		Russian	4
Croatian	4		Upper-Sorbian	4
Czech	4		Bulgarian	2
Ukrainian	4		Croatian	2
Upper-Sorbian	4		Serbian	2

From the linguistic point of view Table 12 displays a rather interesting behavior of the analysed Slavic languages. Since no general discussion of Slavic phonology and graphematics is possible here, we refer to Belorussian, Slovene, Bulgarian and Polish only. Generally the analysis of Slavic grapheme and phoneme frequencies shows the "individual" and "autonomous" organization of these two levels. In this sense one has to interpret the behavior as an empiric justification to differentiate graphemics strongly from phonology. Whereas graphemes can be understood as basic constituents of written language, phonemes are inherently basic units of the spoken language. In any case the level of linguistic abstractness is in phonology much higher than in graphemics. The results of the control cycle for Slavic languages show that any of the stated levels is connected with some particular problems and "disturbances" and imbalance. Coming back to the analyses of Slavic languages the results are particularly quite surprising, although, explainable. At least some hints and general problems of the Slavic languages, which quite often occur as outliers in respect to their frequency behavior, can be given. Generally one would expect a rather similar and homogenous picture for phonemes and graphemes, that could be explained due to the generally rather narrow grapheme – phoneme correspondence of the Slavic languages.

In case of Belorussian it is obvious that the grapheme as well as the phoneme levels seems to be disturbed from a synergetic point of view. This can be explained by the combination of rather different orthographic principles. On one hand the leading orthographic principle of Belorussian is phonetically determined, neither phonemes nor graphemes, but sounds are encoded. Additionally Belorussian utilizes a rather economic marking of palatalization (with special signs which historically were used for the marking of a vowel, but which lost their function). This kind of marking of the palatalization also occurs in other Eastern Slavic scripts, traditionally treated as scripts, which are based on morphophonemic principles. Due to a leading phonetically determined script the analysis of the grapheme level is rather complicated, since reduced vowels are encoded as sounds. This leads to further problems in determining the phoneme frequency based on a written text. Generally this mixture of different factors in regard to phonology and the grapheme level are a first attempt to explain the observed complications. Slovene shows in both cases (phoneme and grapheme level) a rather specific behavior and occurs as an outlier too; in respect to the grapheme level this language has clearly less graphemes, since some phonemes (semivowel, and some phonologically relevant long open vowels) are not expressed by special graphical signs. In this respect the Slovene grapheme inventory is underspecified. Regarding the phoneme level Slovene is known as a language with a rather complex interrelation of vowel quantity, openness of vowels and pitch accent. Currently the Slovene phonological and especially prosodic system is in transition, pitch accent is already partly lost and there is ongoing discussion about it (cf. Kelih 2013b for details). In Bulgarian – a language which quite often occurs as outlier on the phonological level –the unequal extent of palatalization has to be mentioned. Since Bulgarian standard is based on the Eastern vernaculars, which indeed have phonologically relevant palatalized phonemes. However, all of them are regarding their position rather restricted within word forms which again causes a rather low frequency of this phonemes at the text level. In this respect the phonological systems show a significant under-exploitation of these particular phonemes. Polish, a language which attracts attention on the graphemic level, is well known for its extensive use of digraphs, which in our analysis has not been analyzed as combined units, but separate units. Before a reliable linguistic diagnostics can be given in this respect, different approaches for the determination of the grapheme inventory size have to be applied.

In any case, summarizing some possible influence factors and boundary conditions, one has to state, that based on the performed analysis clearly more in-depth studies of Slavic phonology and scripts are required.

Other kinds of diversification, e.g. in the vocabulary or grammar, should be stated in the same way and compared with phonemic/graphemic images. One could acquire a scale of diversification or a scaled distance within the family. If one has the same text, one can for each sentence state how many words/word-forms/morphemes are genetically related and set up an indicator of divergence. However, this is a task for the future and needs rather a team work because it is not only synchronic analysis but requires a good etymological knowledge.

Stratification

It has been shown in different publications (cf. Popescu, Altmann, Köhler 2010; Popescu, Čech, Altmann 2011a; Altmann, Popescu, Zotta 2013) that classes of linguistic entities are not monolithic; they always display some stratification. This is caused both by steady diversification of linguistic entities and by their different nature – leaving aside the boundary conditions associated with every linguistic phenomenon. That means, any classification of linguistic entities puts together different strata. Though we cannot say with certainty which entities belong to individual strata, we can at least detect their number. In the domain of phonemes we can conjecture that vowels and consonants abide by different laws, but this has never been showed. The same holds for parts-of-speech classes or any other classification.

The formula revealing the number of strata has been defined as

$$f(r) = c + a_1 * \exp(-b_1/r) + a_2 * \exp(-b_2/r) + a_3 * \exp(-b_3/x) + \ldots \qquad (19)$$

where a_i is amplitude, b_i exponent (decay constant), and c additive fitting constant or offset (its value is practically unity only for distributions containing hapax legomena such as word rank frequencies). If two exponents are equal (or almost equal), one may eliminate one of the components of (15) and obtain the number of strata as the number of quite different exponents b. In our case it can be shown that phoneme and grapheme frequencies in Slavic languages mostly consist of two strata as shown in Table 13.

Table 13
Stratification of phonemes and graphemes
with fitting offset *c* free

Phonemes

Language	*c*	a_1	b_1	a_2	b_2	a_3	b_3	R^2
Slovene	-6663.76	937.36	4.99	3633.80	327.53	3633.80	332.19	0.9890
Serbian	-588.38	558.94	2.53	518.97	2.54	1311.85	38.38	0.9717
Croatian	-779.56	1138.48	2.47	751.34	46.60	751.53	46.73	0.9737
Bulgarian	-42.07	9392680.00	0.10	302.14	6.62	952.18	11.95	0.9900
Macedonian	-7593.29	2796920000.00	0.06	1133.33	5.45	7923.85	725.69	0.9902
Russian	-746.39	1467.08	2.96	631.86	78.24	630.67	78.25	0.9749
Ukrainian	-5294.88	1009.94	3.10	3189.17	315.35	2713.08	319.01	0.9833
Belorussian	-540.30	3001.44	0.89	3002.22	0.89	1120.00	55.78	0.9954
Czech	-3569.52	546.98	2.56	605.12	20.53	3629.20	1029.05	0.9933
Slovak	-189.00	681.38	2.83	420.31	28.72	420.31	28.72	0.9925
Polish	-5445.27	1032.90	3.19	2977.23	451.72	2977.23	455.02	0.9730
Upper-Sorbian	-4664.31	1002.54	3.01	2915.60	327.30	2310.26	334.57	0.9711

Graphemes

	c	a_1	b_1	a_2	b_2	a_3	b_3	R^2
Slovene	-5159.55	976.75	4.08	2986.39	219.86	2857.81	226.28	0.9798
Serbian	-1243.75	1073.53	2.74	701.62	65.99	1238.08	66.01	0.9787
Croatian	-972.70	594.72	2.50	502.05	2.53	1715.41	51.99	0.9800
Bulgarian	-10990.28	3947980000.00	0.06	821.76	6.35	11407.70	801.36	0.9882
Macedonian	-588.35	25853800.00	0.09	937.03	5.39	1066.43	48.28	0.9895
Russian	-261.94	1243.44	0.80	665.28	20.39	665.20	20.39	0.9932
Ukrainian	-854.95	645.91	2.49	824.75	50.12	824.27	50.12	0.9861
Belorussian	-7131.12	197969.33	0.20	3921.88	350.50	3924.88	361.44	0.9974
Czech	-255.35	323.81	2.14	486.12	27.79	505.54	27.81	0.9892
Slovak	-71.65	306.21	2.44	527.92	15.12	526.23	15.12	0.9900
Polish	-2170.77	562.82	3.58	1491.94	100.34	1492.06	100.35	0.9862
Upper-Sorbian	-4621.73	858.39	3.56	2613.71	287.61	2613.71	296.24	0.9839

The gray cells have very close *b*-exponents, hence belong to a single stratum. Thus in these examples there are only two strata. Exceptions are, however, Bulgarian, Macedonian, and Czech (for phonemes only) with three strata.

A somewhat simpler strata landscape, yet with a quite high determination coefficient, we get by truncating rightwards the fitting at the fixed offset value $c = 1$, that is at the minimum unity frequency, as shown in Table 14. This time only the Macedonian (graphemes) distribution is resolved as a tri strata superposition, the rest remaining bistratal or monostratal.

However, this way of computing stratification shows that there are great differences between the parameter a_1. It seems that outliers (Bulgarian, Macedonian, Belorussian) display a very great first parameter, which can be interpreted only as a slow drifting away from the equilibrium and the Slavic family – at least in this sense.

Table 14
Stratification of phonemes and graphemes
with fitting offset fixed $c = 1$

Phonemes

Language	c	a_1	b_1	a_2	b_2	a_3	b_3	R^2
Slovene	1	486.18	9.50	486.18	9.50	486.18	9.50	0.9835
Serbian	1	392.01	1.85	353.65	1.85	1066.04	11.08	0.9645
Croatian	1	806.69	1.79	535.99	11.29	535.99	11.29	0.9654
Bulgarian	1	628276.22	0.14	617.89	9.39	617.93	9.42	0.9885
Macedonian	1	6743050.00	0.10	249.09	3.90	1159.18	9.09	0.9879
Russian	1	615.80	2.44	624.04	2.44	769.46	15.41	0.9701
Ukrainian	1	637.71	2.16	489.00	13.25	488.96	13.25	0.9691
Belorussian	1	3204.93	0.81	3209.07	0.81	708.86	16.57	0.9888
Czech	1	413.21	1.69	155.51	14.14	672.33	14.14	0.9880
Slovak	1	488.78	1.88	385.51	14.03	485.06	14.03	0.9847
Polish	1	772.41	2.33	395.91	15.99	395.84	15.99	0.9616
Upper-Sorbian	1	744.47	2.42	405.15	15.87	404.98	15.87	0.9582

Graphemes

Language	c	a_1	b_1	a_2	b_2	a_3	b_3	R^2
Slovene	1	319.53	2.13	670.94	10.22	669.02	10.22	0.9762
Serbian	1	663.43	1.83	558.95	10.71	559.15	10.72	0.9693
Croatian	1	728.00	1.77	534.18	10.93	587.75	10.94	0.9691
Bulgarian	1	632134.21	0.13	681.42	9.92	522.81	9.95	9.9163
Macedonian	1	420547.67	0.13	152.61	8.50	1204.44	8.50	0.9855
Russian	1	436.94	11.09	436.94	11.09	436.94	11.09	0.9692
Ukrainian	1	396.58	11.78	396.58	11.78	396.58	11.78	0.9639
Belorussian	1	744440.79	0.16	409.76	17.59	409.66	17.60	0.9736
Czech	1	309.32	13.12	309.32	13.12	309.32	13.12	0.9728
Slovak	1	392.16	11.07	392.16	11.07	392.16	11.07	0.9861
Polish	1	421.45	12.23	421.45	12.23	421.45	12.23	0.9706
Upper-Sorbian	1	520.55	2.44	472.52	14.58	473.03	14.58	0.9755

To conclude, it would be premature to identify the strata. To this end not only the quantitative aspect (rank-frequency) but also the qualitative nature of phonemes/ graphemes in "all" languages should be analyzed. Besides, the phenomenon of stratification exists at all levels of language hence a theoretical solution is rather a task for the whole century.

Conclusion

Studying a family of languages on the lowest level one can ascertain the state of its disintegration. This may be quite different on "higher" levels where not only the lower ones interact but also the cultural development exerts a strong influence. The attractors which are active at the time of unity of the given languages are abandoned and new ones are sought. But they need not be the same in all languages of the group. A drastic example is that of the Indo-European family. Perhaps the most drastic example is English where one must always ask: which English? In Chinese, the inhabitants of Canton and Beijing make themselves understood by writing the signs on the hand: a possible future image of English. Regarding the situation of Slavic languages maybe it is not so drastic, but at least a quite remarkable diversification within one closely related language family has been noticed.

References

Altmann, G. (1988). *Wiederholungen in Texten*. Bochum: Brockmeyer.

Altmann, G., Popescu, I.-I., Zotta, D. (2013). Stratification in texts. *Glottometrics 25, 85-93.*

Esteban, M.D., Morales, D. (1995). A summary of entropy statistics. *Kybernetica 31(4), 337-346.*

Kelih, E. (2009a). Slawisches Parallel-Textkorpus: Projektvorstellung von "Kak zakaljalas' stal' (KZS)". In: E. Kelih, V.V. Levickij, G. Altmann (eds.), *Methods of Text Analysis. Metody analizu tekstu: 106-124.* Černivci: ČNU.

Kelih, E. (2009b). Preliminary analysis of a Slavic parallel corpus. In: J. Levická, R. Garabík (eds.), *NLP, Corpus Linguistics, Corpus Based Grammar Research. Fifth International Conference Smolenice, Slovakia, 25-27 November 2009. Proceedings: 175-183*. Bratislava: Tribun

Kelih, E. (2011). Ein empirischer Regelkreis: Graphemhäufigkeiten in slawischen Sprachen. *Glottotheory* 3(2), 23–34.

Kelih, E. (2013a). Grapheme inventory size and repeat rate in Slavic language. *Glottotheory* 4(1), 56–71.

Kelih, E. (2013b). Silben- und akzentzählende Sprachen: Das Slowenische in typologi scher Sichtweise, in: *Wiener Slavistisches Jahrbuch 58, 188-211*.

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 760-774.* Berlin: de Gruyter.

Ord, J.K. (1972). *Families of frequency distributions.* London: Griffin.

Popescu, I.-I. et al. (2009). *Word frequency studies*. Berlin-New York: Mouton de Gruyter.

Popescu, I.-I., Altmann, G., Köhler, R. (2010). Zipf's law – another view. *Quality and Quantity 44(4), 713-731.*

Popescu, I.-I., Čech, R., Altmann, G. (2011). *The Lambda-structure of texts.* Lüdenscheid: RAM-Verlag.

Popescu, I.-I., Čech, R., Altmann, G. (2011a). On stratification in poetry. *Glottometrics 21, 54-59.*

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

Popescu, I.-I., Mačutek, J., Altmann, G. (2010). Word forms, style and typology. *Glottotheory 3(1), 89-96.*

Popescu, I.-I., Zörnig, P., Altmann,G. (2013). Arc length, vocabulary richness and text size. *Glottometrics 25, 43 – 53.*

Wimmer, G., Altmann, G. (2005). Unified derivation of some lingruistic laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 791-807.* Berlin: de Gruyter.

Appendix

Phoneme and grapheme frequencies in Slavic languages based on the translation of Chapter 1 of the novel *Kak zakaljalas stal'* by N. Ostrovskij

Graphemes

Slovene	1447, 1124, 1103, 1098, 839, 718, 660, 560, 523, 502, 463, 445, 403, 388, 381, 272, 263, 239, 237, 194, 174, 168, 103, 103, 17
Serbian	1389, 1142, 1038, 947, 639, 545, 545, 499, 487, 458, 457, 419, 409, 394, 384, 236, 235, 208, 201, 169, 165, 120, 108, 85, 70, 62, 58, 36, 16, 8
Croatian	1427, 1144, 1098, 900, 661, 559, 547, 537, 484, 465, 463, 405, 401, 399, 398, 350, 241, 204, 203, 171, 167, 123, 86, 84, 84, 73, 58, 39, 15, 6
Bulgarian	1410, 919, 905, 895, 765, 635, 597, 470, 454, 433, 405, 394, 354, 273, 271, 251, 232, 204, 203, 199, 185, 180, 129, 102, 58, 58, 50, 17, 11, 4
Macedonian	1447, 1023, 999, 875, 815, 618, 608, 450, 449, 377, 369, 356, 355, 279, 258, 258, 225, 188, 172, 169, 140, 75, 70, 41, 25, 24, 14, 9, 6, 4, 2
Russian	1356, 1091, 843, 834, 767, 747, 704, 653, 638, 593, 491, 458, 447, 419, 383, 298, 297, 264, 227, 227, 218, 195, 189, 164, 157, 117, 101, 59, 57, 53, 28, 4, 2
Ukrainian	1158, 1053, 968, 813, 651, 597, 580, 575, 547, 536, 490, 467, 464, 439, 433, 335, 285, 278, 257, 216, 206, 201, 185, 179, 163, 136, 94, 87, 47, 44, 23, 23, 15

Belorussian	2036, 681, 638, 607, 599, 593, 548, 531, 517, 499, 477, 446, 423, 412, 398, 393, 390, 351, 344, 333, 259, 242, 214, 195, 188, 162, 141, 105, 99, 78, 49, 28, 6
Czech	915, 753, 731, 711, 602, 533, 502, 482, 441, 434, 429, 406, 403, 380, 355, 318, 241, 229, 229, 223, 210, 173, 171, 170, 159, 142, 131, 126, 103, 100, 90, 35, 17, 9, 9, 6, 6, 5, 3, 1
Slovak	1078, 1064, 916, 748, 686, 616, 609, 575, 541, 476, 472, 438, 411, 390, 329, 261, 219, 200, 198, 198, 189, 178, 171, 151, 142, 124, 119, 112, 103, 81, 61, 59, 31, 26, 15, 14, 14, 12, 11, 11, 4, 4
Polish	1210, 1041, 986, 928, 889, 674, 674, 567, 564, 562, 560, 543, 469, 447, 441, 408, 378, 343, 256, 233, 226, 224, 217, 177, 176, 106, 102, 95, 77, 27, 19, 16
Upper-Sorbian	1181, 1100, 1047, 749, 714, 636, 583, 514, 493, 474, 462, 458, 445, 400, 373, 317, 307, 300, 253, 252, 250, 219, 214, 211, 179, 158, 145, 139, 137, 133, 104, 30, 15, 10

Phonemes

Slovene	1361, 1103, 1100, 1038, 839, 718, 660, 582, 531, 516, 467, 445, 388, 381, 381, 264, 263, 239, 233, 180, 179, 174, 103, 92, 86, 60, 24, 17
Serbian	1389, 1142, 1038, 947, 554, 545, 545, 501, 487, 458, 433, 419, 407, 394, 384, 236, 235, 208, 201, 169, 165, 120, 108, 85, 85, 70, 62, 58, 40, 36, 8
Croatian	1427, 1144, 1098, 900, 577, 559, 547, 541, 484, 463, 437, 405, 399, 398, 397, 350, 241, 204, 203, 171, 167, 123, 86, 84, 84, 84, 73, 58, 43, 39, 6
Bulgarian	1642, 919, 905, 895, 823, 610, 578, 484, 427, 408, 402, 343, 329, 269, 264, 251, 235, 214, 199, 174, 168, 157, 127, 90, 55, 51, 48, 25, 25, 19, 13, 12, 9, 8, 8, 7, 7, 7, 5, 4, 2, 1
Macedonian	1447, 1023, 999, 875, 827, 618, 523, 453, 431, 380, 357, 343, 279, 259, 258, 221, 194, 185, 181, 172, 168, 138, 85, 69, 66, 44, 41, 24, 14, 13, 9, 2
Russian	1431, 1399, 1108, 815, 588, 500, 488, 483, 472, 444, 437, 421, 386, 379, 294, 276, 267, 266, 230, 202, 201, 199, 189, 181, 172, 153, 152, 132, 128, 112, 97, 89, 86, 67, 59, 57, 53, 23, 13, 12, 5, 2
Ukrainian	1253, 1158, 1053, 813, 644, 626, 500, 498, 477, 467, 461, 435, 398, 395, 384, 378, 335, 280, 256, 216, 206, 202, 197, 185, 161, 141, 133, 88, 69, 51, 49, 23, 22, 18, 6, 3
Belorussian	2510, 1245, 622, 593, 491, 479, 468, 434, 433, 398, 394, 394, 389, 304, 296, 263, 259, 254, 252, 249, 238, 218, 210, 206, 202, 178, 144, 139, 108, 102, 101, 86, 79, 75, 65, 19, 17, 11, 10, 8, 4, 2, 1
Czech	976, 915, 711, 675, 582, 497, 490, 466, 451, 440, 392, 355, 341, 321, 318, 308, 241, 237, 229, 204, 203, 171, 164, 162, 159, 133, 133, 131, 108, 93, 90, 83, 68, 63, 56, 41, 39, 20, 3, 1
Slovak	1064, 1003, 791, 746, 575, 548, 496, 489, 469, 450, 425, 390, 354, 329, 293, 283, 267, 266, 248, 227, 219, 197, 192, 190, 179, 171, 159, 125, 120, 103, 99, 75, 69, 61, 61, 26, 17, 17, 15, 14, 11, 10, 5, 4, 4, 1
Polish	1210, 1023, 986, 889, 543, 481, 475, 459, 448, 438, 430, 420, 406, 373, 332, 320, 298, 266, 256, 233, 226, 225, 209, 186, 181, 177, 176,

	172, 167, 158, 148, 95, 79, 46, 38, 33, 31, 27, 24, 5, 5, 3
Upper-Sorbian	1181, 1100, 1047, 805, 583, 506, 474, 462, 458, 453, 447, 391, 383, 377, 373, 371, 307, 297, 291, 286, 252, 240, 211, 179, 145, 140, 130, 129, 128, 109, 93, 81, 69, 64, 22, 15, 10

Glottometrics 27, 2014, 54-88

The lambda structure of language levels

Ioan-Iovitz Popescu
Gabriel Altmann

Abstract. The aim of the article is to present a survey of the computation of the indicator lambda for units of different levels and study the dependence on the inventory size and on the abstractness of the given level. Hence two hypotheses are tested.

Keywords: lambda, language levels, phonemics, graphemics, lexicon, text, hreb, cases, dependence, affixes

Introduction

The "lambda-structure" of texts has been thoroughly studied only for the level of words (cf. Popescu, Čech, Altmann 2011; Popescu, Zörnig, Altmann 2013; Popescu, Mačutek, Altmann 2009, 2010). In the present study we want to make a survey of its forms on different levels of language.

The lambda indicator is a function of the arc length between the neighboring ordered (ranked) frequencies. The components of the arc are defined as

$$L_r = \sqrt{[(f(r) - f(r+1)]^2 + 1} \qquad (1)$$

i.e. as the Euclidean distances between the frequencies *f*, and their sum is the arc

$$L = \sum L_r = \sum_{r=1}^{V-1} \sqrt{(f(r) - f(r+1))^2 + 1} \qquad (2)$$

where *V* is the inventory of entities. Since *L* depends strongly on text size, in the literature it was relativized in different ways: either dividing it by its maximum or simply by *N*. However, there still remained a trace of dependence which could be partially removed by defining

$$\Lambda = \frac{L}{N} log_{10} N\,. \qquad (3)$$

Other modifications concerning word frequencies are used, too.

Here, our aim is to study the behavior of lambda at different levels of language. We can state two hypotheses: (1) the higher the level, the greater becomes the lambda; e.g. the lambda of word frequencies is greater than that of phonemes. (2) The frequencies of basic forms of entities have always a smaller lambda than the frequencies of allo-forms: phonetic, morphological, syntactic, semantic variants. For example, the lambda of word forms in a text is always greater than that of lemmas. This boils down to the hypothesis that the smaller the inventory of entities, the greater is lambda. However, the morphological tendencies in the language (analytism, synthetism) play a basic role and should not be intermixed. If we do it, we must reckon with outliers and other irregularities.

The first hypothesis concerns merely the main levels: phonemics, morphology, syntax, semantics; however, there may be great differences between languages, text-sorts, styles and many factors can influence the results (age, gender, education, religion, etc.). Using lambda, we merely want to characterize texts or languages.

The second hypothesis can be developed in different directions of the hierarchies. If we define parts of speech, we obtain a special lambda; if we now take one of the POS and set up the frequencies of its lemmas, we obtain a greater lambda; further, if we take one lemma and study the frequency of its meaning variants, morphological forms, dialectal variants, syntactic functions, etc., separately, we obtain always a greater lambda. This is caused by the fact, that the deeper is the position in the hierarchy, the more concentrated are the frequencies on the main representative, and the arc components between the first and second rank increases. The concentration means at the same time that the *repeat rate* increases with the level in the hierarchy and the entropy decreases.

Lambda is also one of the possibilities of measuring the excess of the rank-frequency distribution. Other possibilities were proposed by K. Pearson, e.g. the ratio of the fourth and the second central moments.

In the sequel we shall scrutinize different entities and perform also some comparisons.

Phonemes and letters

As is well known, phoneme/sound/letter frequencies are formed differently because there must be a certain amount of redundancy, while this is not necessary e.g. with word frequencies in text of the same text-sort. In general, we consider the frequencies of a set ordered in non-increasing order, practically in its rank order. Of course, the same procedure can be applied also to the distribution of any measurable property playing the role of the independent variable. Continuous variables may be pooled to groups.

Whatever unit we choose, we first consider a single text and compute the frequency of units of the given set, e.g. letters, syllables, morphemes, words, word length, clause length, rhythmic patterns, semantic classes, etc. Each of the distributions yields a lambda value which can be compared in languages, evolution, text-sorts, but at the same time, the levels of language may be compared.

Let us consider the phonemes in the poem *Lacul* by the Romanian poet M. Eminescu. Ordering them we obtain the series

[37,33,32,29,27,27,26,25,24,21,18,15,14,13,11,9,8,7,7,7,6,6,3,2,2,2,1,1,1].

The arc length can be computed according to (2) as

$$L = [(37 – 33)^2 + 1]^{1/2} + [(33 – 32)^2 + 1]^{1/2} + \ldots + [(1 – 1)^2 + 1]^{1/2} = 50.1990.$$

Since N = 414 and log(N) = log(414) = 2.6170, we obtain

$$\Lambda(\text{Romanian phonemes/Lacul}) = (50.1990/414)2.6170 = 0.3173.$$

Let us begin with letter frequencies. This can be considered the lowest level of language because it contains only secondary symbols abstracted from the primary level of sounds or phonemes. Though steps in deeper levels are possible, e.g. distinctive features and muscle effort of sounds, or graphical motifs of letters or (iconic, symbolic) signs, we restrict ourselves to those for which there are many available data.

Consider first the letter frequencies in 12 English novels as presented in Table 1. For completeness we add also the variance of Lambda computed as $Var(\Lambda) = (\log_{10}N/N)^2 Var(L)$, while $Var(L)$ is computed directly from the L_r-values. The mean Λ of these texts is 0.6974.

Table 1
Letter frequencies in 12 English novels
(e-texts from http://www.gutenberg.org/browse/scores/top)

Author: Text	N	V	L	Var(L)	Λ	Var(Λ)
Charles Dickens: David Copperfield	1503528	26	181177,0444	95269821,7918	0,7443	0,00160806
Charles Dickens: Great Expectations	761751	26	91566,0117	21414377,7815	0,7070	0,00127674
Charles Dickens: A Christmas Carol	121498	26	14828,5788	715514,2944	0,6206	0,00125311
James Joyce: Ulysses	1182311	26	140388,0123	73998496,9439	0,7211	0,00195221
Conan Doyle: Sherlock Holmes	431143	26	52886,0364	8328582,7923	0,6912	0,00142252
Mark Twain: Huckleberry Finn	421468	26	46966,1942	3347373,2108	0,6268	0,00059619
John Milton: Paradise Lost	356888	26	42552,2197	5979153,3798	0,6620	0,00144730
H.G. Wells: The War of the Worlds	266023	26	33293,1516	3034550,0077	0,6789	0,00126195
Jonathan Swift: Gulliver's Travels	454690	26	57933,0300	12271525,1365	0,7209	0,00189998
Emily Bronte: Wuthering Heights	497933	26	63575,0361	17170836,2165	0,7274	0,00224785
Charlotte Bronte: Jane Eyre	787557	26	100285,0239	43380170,3114	0,7508	0,00243155
Bram Stoker: Dracula	638106	26	78959,2011	20236956,4427	0,7183	0,00167474

Table 2
Russian letter frequencies
(According to Grzybek, Kelih 2003).

Text	N	V	L	Var(L)	Λ	Var(Λ)
Ol´chin 1907	22304	29	2410,9721	18622,0180	0,4700	0,00070781
Proskurin 1933	999202	33	109689,0384	22115856,4027	0,6586	0,00079735
Kalinina 1968	100000	31	10954,3592	268918,7303	0,5477	0,00067230
Grigor´ev 1980a	50000	32	5662,5558	78115,7019	0,5322	0,00068993
Grigor´ev 1980b	99986	32	11388,637	312341,6696	0,5695	0,00078105
Dietze 1982	429257	32	44016,0314	2295153,243	0,5776	0,00039520

If we compare two authors, e.g. *Jane Eyre* by Charlotte Bronte (highest English Λ) with *A Christmas Carol* by Charles Dickens (smallest Λ), we can use the asymptotic normal test defined as

$$u = \frac{|\Lambda_1 - \Lambda_2|}{\sqrt{Var(\Lambda_1) + Var(\Lambda_2)}}, \quad (4)$$

yielding in our case

$$u = [0.7274 - 0.6206|/(0,00243155 + 0,00125311)^{1/2} = 1.76$$

which is not significant, hence all English texts display a rather constant lambda.

If we want to compare two languages, we may take the mean of all lambdas in one language and compute their variance directly from the data. One can, of course, pool the different data to obtain a common variance, one can compute the degrees of freedom in a special way, but we make the computation as simple as possible. The mean of English lambdas is $\bar{\Lambda}_{English}$ = 0.6974, Var($\bar{\Lambda}_{English}$) = 0.00014; the same values for Russian texts are $\bar{\Lambda}_{Russian}$ = 0.5595, Var($\bar{\Lambda}_{Russian}$) = 0.000532. Using the t-test with $n_E + n_R - 2$ degress of freedom, we obtain

$$t = |0.6974 - 0.5595|/(0.00014 + 0.000532)^{1/2} = 5.32,$$

showing that the difference between English and Russian is significant. Hence mean Λ can be used at least for the ordering of alphabetic languages.

Table 3 contains lambdas concerning mixed samples from different languages. The main source is the collection presented on the Internet: http://www.cryptogram.org/cdb/words/frequency.html, the other sources are shown under the table.

The values of lambda are ordered. As can be seen, there is only one remarkable fact: all Austronesian languages have a very high lambda though neither *N* nor *V* differ drastically from those in other languages. The Slavic languages are rather in the first part of the table. Hence further investigations using much more extensive data could perhaps be used for genetic or typological classification.

Table 3
Lambda of letters in mixed samples of a language

Language	N	V	L	Var(L)	Λ ascending	Var(Λ)
Polish	6841	32	651.7516	985.2137663	0.3654	0.00030963
Czech	8075	37	762.4237	955.5444859	0.3689	0.00022371
Swedish	4894	26	507.8443	869.8651985	0.3829	0.00049442
Serbian	8624	29	875.3375	2561.7835	0.3995	0.00053354
Albanian	4590	34	501.1058	364.8967334	0.3998	0.00023224
Estonian	5011	24	567.5124	807.361043	0.4190	0.00044015
Greek(modern)	6351	25	756.3991	1976.03709	0.4529	0.00070848
Kurdish	7199	31	862.1383	3695.653153	0.4619	0.00106098
Maltese	8680	32	1058.9411	2435.740619	0.4805	0.00050148
Hungarian	9620	37	1180.4987	3201.957793	0.4888	0.0005489
Guarani	7482	36	949.3526	3314.071074	0.4916	0.00088848
Slovak	148935	42	14190.3149	337489.7164	0.4929	0.00040715
Italian	4882	22	653.9277	2379.553599	0.4941	0.00135838
Latin	8281	21	1075.0203	3670.871726	0.5086	0.0008218
Macedonian	8662	29	1145.3366	2515.746183	0.5207	0.00051987
Mazateco	6624	28	910.4608	2498.752344	0.5252	0.0008315
Sardinian	7565	27	1035.9852	2818.942608	0.5312	0.00074108
German	96365	46	10275.3015	268247.4621	0.5314	0.00071753
Gascon	12259	34	1604.0673	8111.541627	0.5350	0.00090222
Scottish Gaelic	1393	30	244.4406	358.5576369	0.5517	0.00182645
Slovenian	313735	25	31539.3156	2684118.9204	0.5526	0.00082387
Portuguese	4283	33	652.7371	1137.240213	0.5535	0.00081769
French	9625	31	1356.8022	8287.009086	0.5615	0.0014194
Gagauz	9121	32	1317.1583	8104.909776	0.5719	0.00152779
Walloon	18325	35	2229.6686	18002.47581	0.5754	0.00119889
German	5732	27	890.3923	3462.070159	0.5838	0.00148836
English (Fry)	492745	44	51496.0935	7716748.5117	0.5949	0.00102995
Lithuanian	8845	31	1346.9639	7368.963385	0.6010	0.00146716
Chechewa	8710	26	1331.2414	8559.118878	0.6022	0.00175142
Finnish	5339	21	869.7351	5402.738988	0.6072	0.00263342
Spanish	5275	27	861.5076	3160.600933	0.6079	0.00157373
Romanian	6268	26	1029.489	4500.045635	0.6237	0.00165147
Huasteco	8276	29	1346.391	9323.565194	0.6374	0.00208944
Georgian	13000	33	2056.685	15942.6105	0.6509	0.00159700
Greek (classic)	2517	25	485.6012	2095.59317	0.6561	0.00382582
Danish	9719	25	1602.2973	22357.29607	0.6574	0.0037636
Chuuk	8893	22	1533.5387	8357.267008	0.6810	0.00164798
Chayahuita	9089	25	1596.1847	12347.06235	0.6952	0.00234205
Inuktikut	15183	18	2567.2348	30294.75474	0.7070	0.00229767
Frisian	14332	32	2501.7733	45017.56598	0.7255	0.00378603
Finnish	2491208	27	296513.1142	176316033.7259	0.7613	0.00116237
Dutch	4135	26	886.897	7210.064637	0.7757	0.00551517
Kikongo	9339	21	1841.4972	32632.8108	0.7829	0.005898

Sea Dayak	19999	21	3774.12238	503921.1939	0.8117	0.02330701
Indonesian 1	188439	29	30130.1756	4920041.477	0.8435	0.00385567
Indonesian	10106	24	2294.6436	84442.86327	0.9093	0.0132592
Javanese	11505	24	2638.544	112538.0939	0.9313	0.01402065
Malay	10457	25	2506.1523	79056.34904	0.9633	0.0116801
Fijian	8604	21	2112.921	83936.48658	0.9663	0.01755389
Hawaiian 1	7985	13	2015.3751	117466.8135	0.9849	0.0280544
Malagasy	10324	32	2540.6878	44013.83794	0.9955	0.00675767
Maori	10950	16	2750.4393	114162.2309	1.0146	0.01553571
Indonesian 2	92853	29	20916.5655	6645150.6361	1.1191	0.0190231
Hawaiian 2	19458	13	5225.2824	642691.3604	1.1518	0.03122756
Tagalog	10154	23	3277.7647	295790.1002	1.2934	0.04605421

Slovenian, Slovak, Serbian: Grzybek, Kelih (personal communication); Finnish: Pääkkönen (1994); Indonesian 1: Altmann (2005: dictionary); Indonesian 2 (text), Georgian, Hawaiian and Sea Dayak: Altmann, Lehfeldt (1980); English: Fry (1947) (Internet); German: Meyer (1967), Best (2004/2005);

By modifying lambda according to *N* we obtain independence of lambda on *N*: as a matter of fact, a strongly oscillating horizontal line, as can be seen in Figure 1.

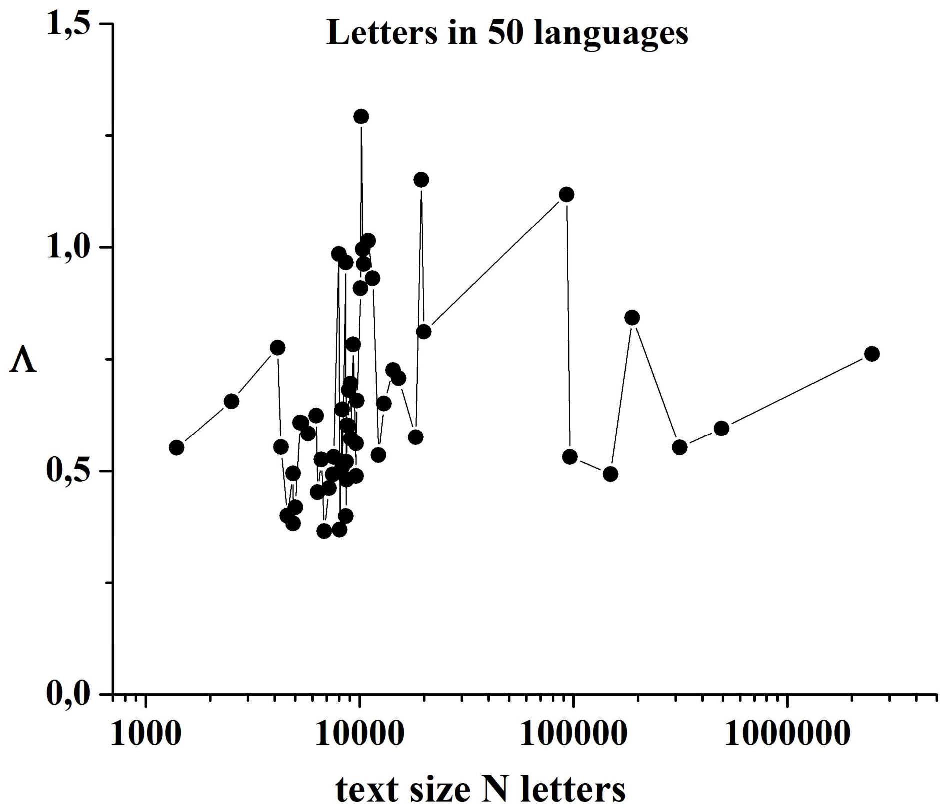


Figure 1. <N,Λ> for the languages in Table 3

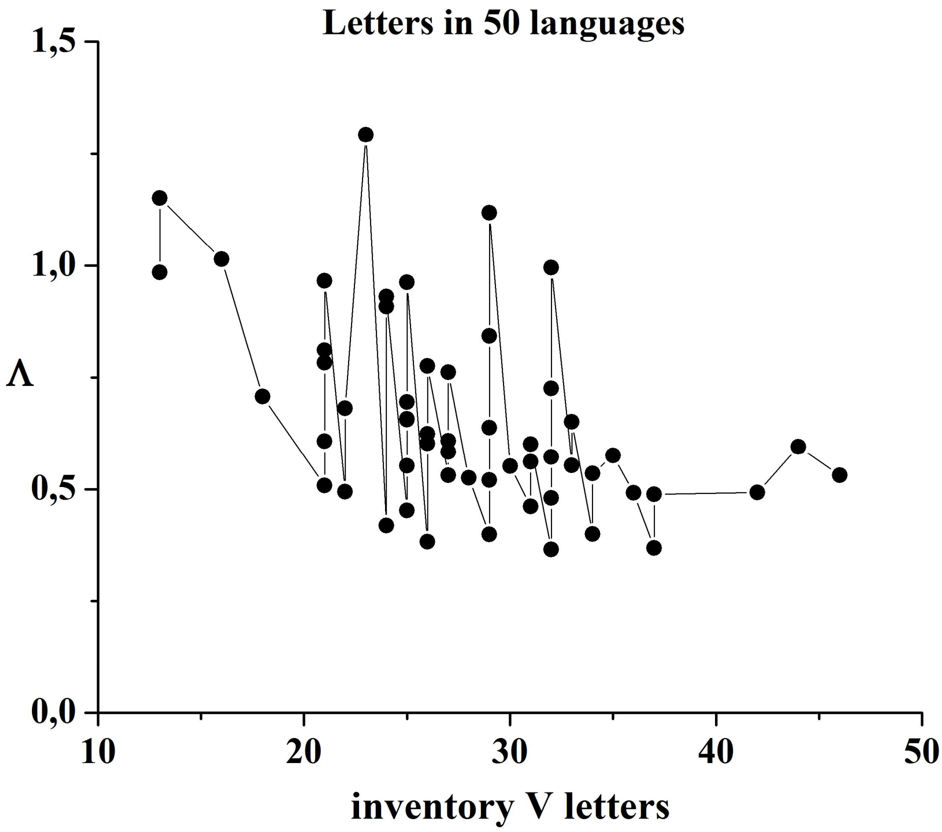


Figure 2. Dependence of lambda on the inventory *V* (letters) (Table 3)

However, there is a clear descending dependence of lambda on the inventory as can be seen in Figure 2. The greater the inventory, the smaller lambda, hence the second hypothesis is corroborated. The “cause” is simple: the more letters there are in the inventory, the more even is the rank-frequency distribution. Of course, one finds outliers but we suppose that adding further data, the decreasing trend would be strengthened.

For lambda values of phonemes in Slavic languages cf. Kelih, Popescu, Altmann (2014).

Closed classes of higher levels

Cases

Case is a linguistic category with quite different interpretations. The concept itself is a heredity from Latin grammar but does not have the same form in every language. In some of them it is expressed by inflexion, in other ones by a preposition, postposition suffix or particle and still in other ones by word order. Since for a given language the inventory is always the same, e.g. in German there are 4 cases, the inventory cannot influence the lambda. If we order the lambda values according to *N*, we do not obtain a monotonous increase, thus the only “causes” of the differences can be either ran-

domness or text-sort or style. Since texts No. 1—10 are newspaper texts and No. 11—20 sagas, we compare the means yielding: Texts
, though in Russian, we obtain for the difference of extreme values of lambda (Text 1—10: 0.7596, texts 11—20: 0.7643, i.e. the only "cause" is randomness. We can state that the use of cases in German is a purely grammatical matter. The interval of lambda in German is relatively large: (0.58; 1.05).

In Slovenian, there are 6 cases and the lambda interval is (0.56; 0.75), i.e. the upper boundary is much lower than that in German.

Slovak contains 6 cases and in some cases also the seventh one (vocative), and the interval is (0.46; 0.99).

The Russian has the largest interval: (0.33; 1.03) with 6 cases.

If we consider the case as a representative of grammatical phenomena, we can conjecture that for a language lambda does not display significant results 7 and 10) a t = 2.44 with 8 DF which is slightly greater than the critical value at α = 0.05.

Table 3a
Rank-frequencies of German cases
(Popescu, Kelih, Best, Altmann 2009)

Text	**Data**	**N**	**V**	**L**	**Var(L)**	**Λ**	**Var(Λ)**
5	30,28,27,4;	89	4	26.672	149.9339768	0.5842	0.07193124
3	48,44,27,10;	129	4	38.1819	55.52402752	0.6247	0.01486309
15	32,26,23,3;	84	4	29.2700	81.21094347	0.6705	0.04261791
2	50,47,33,7;	137	4	43.2172	130.71270041	0.6740	0.03179597
18	46,45,28,6;	125	4	40.4663	115.579552	0.6788	0.03252524
19	43,39,34,4;	120	4	39.2388	215.3862635	0.6799	0.06466067
8	32,21,20,4;	77	4	28.4908	55.21244172	0.6980	0.03314094
9	51,39,34,7;	131	4	44.1591	125.4952613	0.7137	0.03278217
13	48,45,35,3;	131	4	45.2278	227.0747361	0.7310	0.05931701
4	40,32,24,3;	99	4	37.1483	56.00049164	0.7488	0.02275536
6	73,49,42,13;	177	4	60.1091	132.3154478	0.7634	0.02134255
20	64,61,38,6;	169	4	58.1996	217.9672223	0.7672	0.03787948
1	45,40,11,6;	102	4	39.2153	190.6936974	0.7722	0.07394741
16	46,34,28,3;	111	4	43.1443	93.76085657	0.7950	0.03183460
10	49,35,27,4;	115	4	45.1197	56.70278183	0.8085	0.01820695
17	56,46,38,2;	142	4	54.126	243.2290007	0.8204	0.05587784
14	43,29,24,2;	98	4	41.1574	71.67801728	0.8363	0.02959204
12	48,37,20,1;	106	4	47.101	17.24859373	0.8999	0.00629685
7	58,32,26,4;	120	4	54.1247	111.252777	0.9378	0.03339897
11	78,53,28,2;	161	4	76.0592	0.3328213	1.0425	0.00006253

Table 3b
Rank-frequencies of Slovenian cases

Text	Data	N	V	L	Var(L)	Λ	Var(Λ)
1	64,54,52,21,20,7;	218	6	57.7547	143.2198076	0.6195	0.016480
2	89,65,49,39,23,13;	278	6	76.183	33.05741413	0.6698	0.002555
3	86,78,45,29,27,7;	272	6	79.3697	139.5227131	0.7104	0.011178
4	81,64,42,26,17,15;	245	6	66.3748	59.46946505	0.6473	0.005655
5	26,16,14,11,9,1;	77	6	25.7465	13.35576592	0.6308	0.008017
6	82,77,52,31,27,16;	285	6	66.3113	88.39074567	0.5712	0.006558
7	78,34,28,23,20,12;	195	6	66.4177	298.1845917	0.7800	0.041124
8	43,34,28,15,14,5;	139	6	38.6462	18.57374995	0.5958	0.004415
9	67,59,42,23,16,14;	221	6	53.4251	50.28805474	0.5667	0.005659
10	133,97,91,48,28,16;	413	6	117.1749	246.0026789	0.7422	0.009870

Table 3c
Rank-frequencies of Slovak cases

Text	Data	N	V	L	Var(L)	Λ	Var(Λ)
1	26,18,16,13,8,2;	83	6	24.6424	5.38764202	0.5698	0.00288023
2	14,13,5,4,2,2,2;	42	7	15.1268	7.57271161	0.5846	0.01131158
3	53,52,47,46,42,6;	246	6	48.0644	220.49048771	0.4671	0.02082837
4	36,31,30,25,10,5,4;	141	7	33.1588	24.94983911	0.5054	0.00579683
5	67,50,44,28,15,15,2;	221	7	66.2202	38.82960059	0.7025	0.00436957
6	39,36,22,12,10,4,2;	125	7	37.8027	23.36514381	0.6342	0.00657519
7	27,22,10,7,4,2,	72	6	25.7012	15.97231983	0.6630	0.01062875
8	28,20,10,9,7,3;	77	6	25.8855	13.99699133	0.6342	0.00840161
9	163,105,72,43,24,22,2;	431	7	161.3284	345.43873655	0.9861	0.01290641

Table 3d
Rank-frequency of Russian cases

Text	Data	N	V	L	Var(L)	Λ	Var(Λ)
1	134,74,43,24,19,15;	309	6	119.2729	530.69900097	0.9611	0.03446002
2	64,25,19,8,7,6;	129	6	58.9694	247.38067552	0.9648	0.06622073
3	36,31,28,27,6,3;	131	6	33.8616	65.16965497	0.5473	0.01702377
4	54,28,21,14,4,2;	123	6	52.4473	83.21402101	0.8911	0.02402367
5	35,15,15,12,7,4;	88	6	32.4486	59.3545502	0.7170	0.02897990
6	66,35,21,11,11,6;	150	6	61.2007	134.4737844	0.8879	0.02830149
7	36,28,28,22,17,15;	146	6	22.4801	8.23223773	0.3333	0.00180912
8	47,33,23,18,16,10;	147	6	37.5034	21.17477009	0.5529	0.00460287
9	83,31,21,15,12,12;	174	6	72.3045	452.1027449	0.9310	0.07496312
10	51,17,10,9,7,2;	96	6	49.8351	185.8232821	1.0290	0.07922893

In Hungarian, Vincze (2013) found 24 cases, a quite normal state for a strongly agglutinating language, but since text-sorts do not differ, we present merely the total. The results are presented in Table 3e.

Table 3e
Rank-frequencies of Hungarian "cases"

	N	**V**	**L**	**Var(L)**	**Λ**	**Var(Λ)**
Total	381082	24	206705,3251	1065504666,8923	3,0272	0,22853115

For ordering the languages it is sufficient to consider the means of lambda. We obtain the comparative result in Table 3f

Table 3f
Mean lambda for case frequencies in 5 languages

Language	**V**	$\bar{\Lambda}$
Slovak	6	0.6385
Slovenian	6	0.6534
German	4	0.7623
Russian	6	0.7815
Hungarian	24	3.0272

In the grammatical domain one expects higher lambdas than in the phonemic one.

Parts-of-speech

In order to obtain a more systematic survey we analyzed the distribution of parts-of-speech in 60 End-of-Year speeches of Italian presidents (cf. Tuzzi, Popescu, Altmann 2010). The survey is at the same time an image of historical change of lambda. The results are presented in Table 4. One can see that mechanical ascription yields 8 to 11 parts-of-speech and is a grammatical convention. The interval of lambda is <0.4866; 0.8730> and it increases irregularly in the course of time as can be seen in Figure 3. There are some outliers that can be ascribed to an individual president, e.g. Saragat, and Cossiga 1991. However, if we sort the data according to size *N* or inventory *V*, we obtain a slightly increasing lambda, as can be seen in Figures 4 and 5.

The values of lambda for different languages can be found in Table 5.

Table 4
POS in Italian

Text	**N**	**V**	**L**	**Var(L)**	**Λ**	**Var(Λ)**
1949Einaudi	194	9	41,2585	14,17393027	0,4866	0,00197117
1950Einaudi	150	9	42,9954	22,27494143	0,6237	0,00468801

1951Einaudi	230	8	40,0384	16,49819915	0,4111	0,00173956
1952Einaudi	179	8	40,5178	26,24534092	0,5099	0,00415730
1953Einaudi	190	9	46,8524	7,94381255	0,5619	0,00114266
1954Einaudi	260	9	57,1609	31,08272589	0,5309	0,00268162
1955Gronchi	388	9	83,1137	57,21616227	0,5546	0,00254720
1956Gronchi	665	8	154,5905	371,49438891	0,6562	0,00669385
1957Gronchi	1130	10	262,6558	550,20728725	0,7097	0,00401648
1958Gronchi	886	10	200,5435	276,79612126	0,6671	0,00306324
1959Gronchi	697	9	180,6438	308,71126865	0,7369	0,00513703
1960Gronchi	804	10	194,6857	258,45099584	0,7035	0,00337470
1961Gronchi	1252	9	302,2064	554,55856374	0,7477	0,00339462
1962Segni	738	8	167,2172	224,24762187	0,6498	0,00338681
1963Segni	1057	10	249,2705	230,25302964	0,7132	0,00188469
1964Saragat	465	9	99,8500	57,24962571	0,5728	0,00188391
1965Saragat	1053	10	264,8750	715,44768984	0,7603	0,00589431
1966Saragat	1199	10	322,3163	629,98882412	0,8277	0,00415397
1967Saragat	1056	11	261,6730	329,74709879	0,7493	0,00270346
1968Saragat	1174	10	302,2558	528,25572658	0,7903	0,00361153
1969Saragat	1584	11	392,7435	1272,13907554	0,7934	0,00519108
1970Saragat	1929	11	488,7010	1800,34850340	0,8323	0,00522217
1971Leone	262	10	69,1937	37,87815904	0,6387	0,00322706
1972Leone	767	10	180,3891	172,80258631	0,6785	0,00244450
1973Leone	1250	10	298,2023	713,18610058	0,7388	0,00437764
1974Leone	801	10	197,4668	454,55377979	0,7158	0,00597315
1975Leone	1328	9	310,2143	716,41252858	0,7296	0,00396247
1976Leone	1366	10	320,7748	1113,00720642	0,7363	0,00586403
1977Leone	1604	10	356,6583	1701,63651366	0,7127	0,00679469
1978Pertini	1493	10	322,2778	853,83396929	0,6851	0,00385908
1979Pertini	2302	11	498,1817	1407,16706369	0,7276	0,00300163
1980Pertini	1360	11	314,7568	670,75695958	0,7252	0,00356088
1981Pertini	2818	11	571,2389	4401,84609540	0,6993	0,00659745
1982Pertini	2487	11	507,1897	2232,09597846	0,6925	0,00416115
1983Pertini	3748	11	783,1100	5158,54184463	0,7467	0,00469017
1984Pertini	1340	10	285,4930	433,96909317	0,6662	0,00236338
1985Cossiga	2359	11	610,5849	3389,84486348	0,8730	0,00692926
1986Cossiga	1349	10	321,3436	982,93450447	0,7456	0,00529167
1987Cossiga	2091	10	491,1875	1418,72047080	0,7800	0,00357732
1988Cossiga	2385	10	552,1696	1999,27378437	0,7819	0,00400944
1989Cossiga	1912	10	435,5390	1066,35862602	0,7475	0,00314101
1990Cossiga	3347	10	782,1227	2999,82154981	0,8236	0,00332673
1991Cossiga	418	10	93,7939	62,56546733	0,5882	0,00246022
1992Scalfaro	2772	11	654,9161	3979,83273437	0,8134	0,00613904
1993Scalfaro	2941	11	683,2213	4297,09617228	0,8058	0,00597678
1994Scalfaro	3605	11	865,2990	7940,75065641	0,8538	0,00773029

1995Scalfaro	4228	11	991,1198	7812,35079528	0,8500	0,00574645
1996Scalfaro	2085	10	524,3185	4258,18218976	0,8347	0,01079082
1997Scalfaro	5015	11	1103,1102	13352,86524480	0,8139	0,00726942
1998Scalfaro	3995	11	967,2921	7935,06571490	0,8720	0,00644893
1999Ciampi	1941	11	503,2617	2353,30726201	0,8525	0,00675302
2000Ciampi	1844	10	420,2012	1244,52718176	0,7442	0,00390347
2001Ciampi	2097	11	547,3533	2544,15964743	0,8670	0,00638325
2002Ciampi	2129	10	549,3121	2958,36391236	0,8587	0,00722957
2003Ciampi	1565	11	407,8606	1397,32989671	0,8325	0,00582210
2004Ciampi	1807	10	447,3706	1982,52187369	0,8063	0,00644060
2005Ciampi	1193	11	289,4386	375,28136434	0,7464	0,00249591
2006Napolitano	2204	11	501,3999	2683,86806855	0,7606	0,00617541
2007Napolitano	1794	11	416,4698	1426,80978341	0,7554	0,00469363
2008Napolitano	1713	11	408,8353	1223,26334927	0,7718	0,00435933

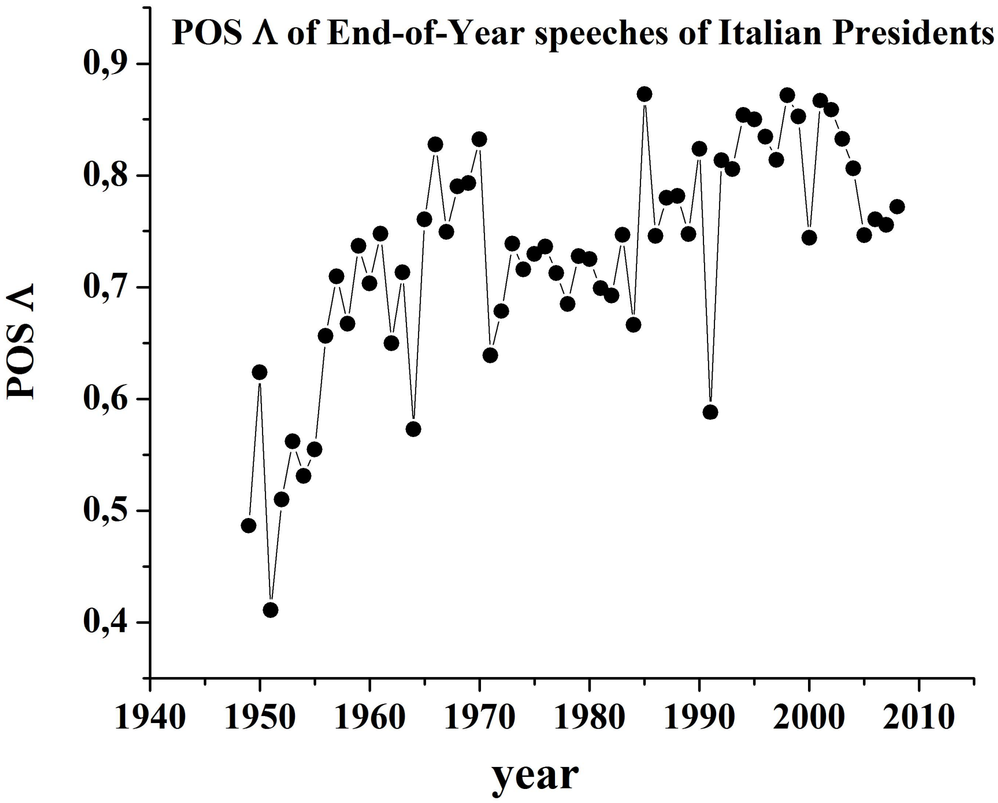


Figure 3. Lambda of POS of End-of-Year speeches of Italian presidents

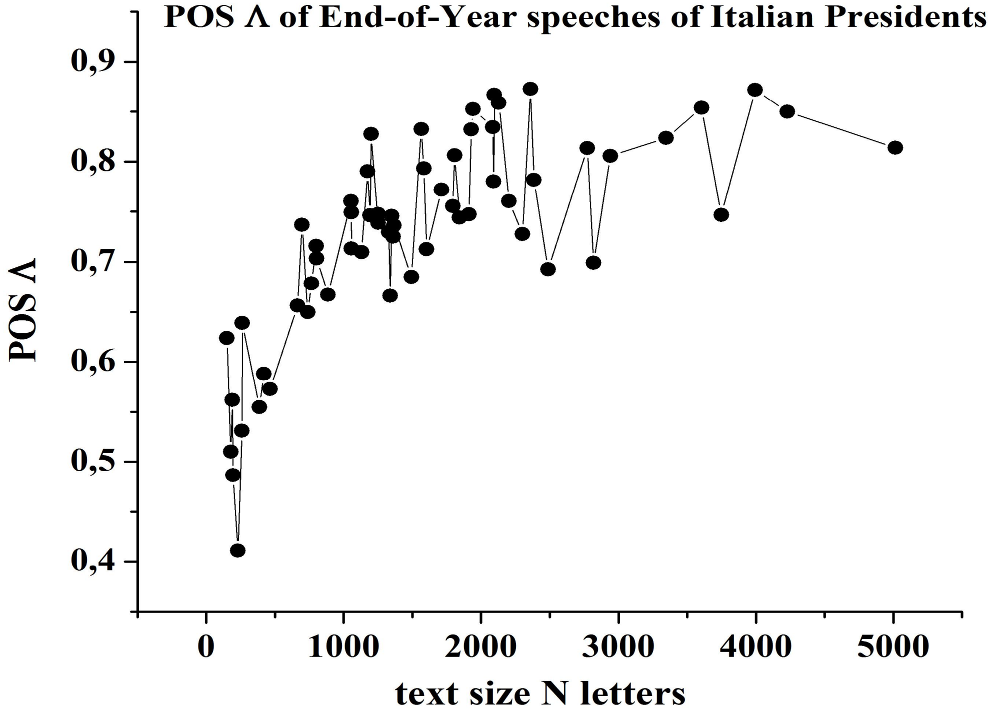


Figure 4. POS in Italian ordered according to *N*

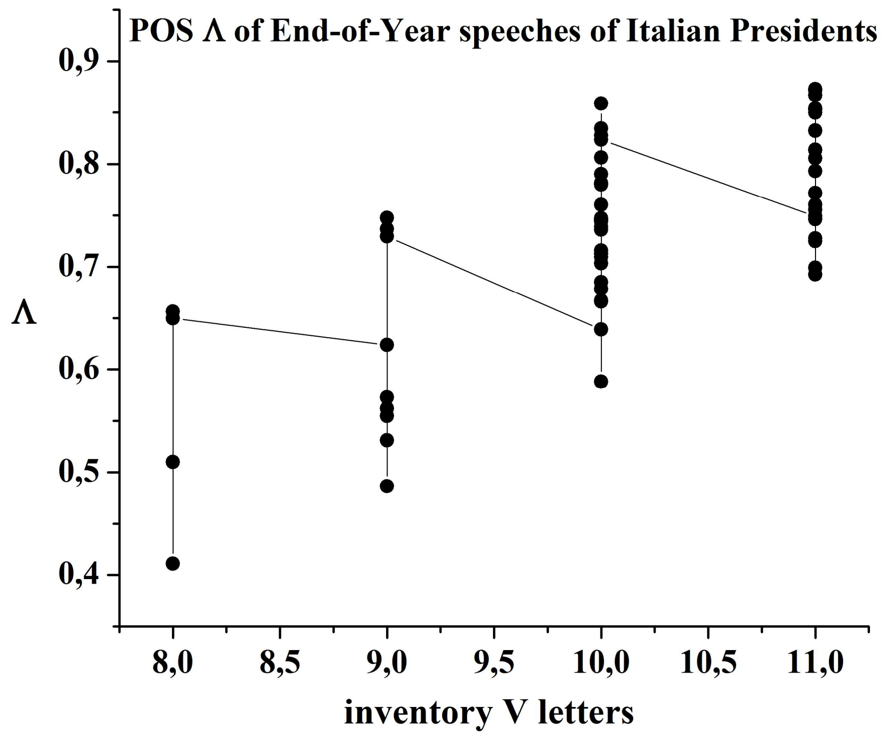


Figure 5. Pos in Italian ordered according to *V*

Table 5
POS in different languages

Text	POS frequency	N	V	L	Var(L)	Λ	Var(Λ)
Chinese	247, 228, 140,133, 107, 81, 55, 27	1018	8	220.18	674.58	0.6505	0.005889
German 1	192, 161, 153,112, 111, 104, 97,70	1000	8	122.67	231.91	0.3680	0.002087
German 2	2032, 1939, 1532, 1338, 1179, 974, 914, 761	10669	8	1271.03	12578.19	0.4799	0.001793
German SMS	2815, 2550, 2416, 1606, 1459, 767, 541, 175	12329	8	2640.01	72354.22	0.8760	0.007966
Latin	347, 173, 142, 98, 93, 59, 40, 39, 9	1000	9	338.60	3041.81	1.0158	0.027376
Polish	144188, 79995, 71988, 56812, 33605, 31833, 21428, 18757, 8076, 650	467332	10	143538.00	369032109.86	1.7414	0.054315
Portuguese	2586, 1607, 949, 819, 776, 680, 478, 440, 352	8687	9	2234.04	120785.91	1.0130	0.024832
Portuguese (Brazilian)	2930, 2265, 1743, 1708, 1602, 1040, 936, 394	12618	8	2536.03	71470.77	0.8242	0.007550

(Portuguese-Brasilian Port.: 1.05, DF = 13; German 2-German SMS = 4.01, DF = 12; Chinese-Latin = 2.00, DF = 13; Polish-Latin = 2.54, DF = 15)

For Hungarian, Vincze (2013) prepared a count of POS for six different text-sorts and distinguished 14 parts-of-speech. As can be seen in Table 6, the lambdas are higher than in other languages. This can be ascribed both to the great *N* and *V* but most probably Hungarian as a strongly synthetic language produces these results. None of the above languages attains such high values. The results lean against the Szeged Treebank.

Table 6
POS in six Hungarian text sorts (Vincze 2013)

Text sort	N	V	L	Var(L)	Λ	Var(Λ)
Composition	279329	14	58552.0087	28960087.0393	1.1416	0.01100884
Literature	186531	14	44697.0159	18535023.6688	1.2630	0.01479913
Law	221491	14	78540.0416	110659865.5069	1.8954	0.06445124
Newspaper	187276	14	61868.0111	85230872.4526	1.7418	0.06755584
Newsml	200084	14	79586.1674	164184826.9060	2.1086	0.11525472
Computer	179732	14	74515.0373	124331855.3950	2.1785	0.10627109

Dependency relations

V. Vincze (2013) prepared also counts concerning the dependence relations between the verb and its arguments, i.e. mostly verb valence. She found maximally 25 different cases and subdivided the texts into text sorts. The results are presented in Table 7. A comparison with other languages is, preliminarily, not possible.

Table 7
Dependency relation in Hungarian
(Vincze 2013)

Text sort	**N**	**V**	**L**	**Var(L)**	**Λ**	**Var(Λ)**
Composition	284436	25	46941.1026	10545125.6276	0.9001	0.00387713
Literature	189731	24	35177.1050	9541790.7956	0.9786	0.00738439
Law	224218	25	80590.2962	98762838.5430	1.9232	0.05624307
Newspaper	190404	25	51745.1595	29921166.2296	1.4348	0.02300599
Newsml	201523	25	66869.2328	60777256.6315	1.7601	0.04210678
Computer	184605	25	49168.1296	23438122.0080	1.4026	0.01907382

Semantics

Meaning diversification

Almost every word in non-scientific literature has several meanings. The meanings can be found in monolingual dictionaries or in WORDNET, if it exists for the given language. However, the individual meanings do not occur with the same frequency; the main meaning is usually very conspicuous. If one orders the meanings, one can see it at once. For example, in the English WORDNET the word *belly* has V = 6 meanings, and their frequency sum in the given data is N = 14. The frequencies are given as: 8, 2, 1, 1, 1, 1. The values of lambda for some words are presented in Table 8 in descending order of lambda. The last word shows that no diversification yields lambda = 0. In order to show the prevalence of the main meaning we added the column *f*(1).

Table 8
Meaning diversification of English words
(Fan, Popescu, Altmann, 2008).

Word	**N**	**V**	**L**	**f(1)**	**Λ**
Year	865	4	831.15	832	2.8221
Walk	1208	17	1099.55	1092	2.8054
Cut	2138	71	1728.05	1672	2.6915

Blood	677	6	637.35	637	2.6648
Say	3547	12	2593.96	2593	2.5961
Name	847	15	703.70	698	2.4325
Woman	587	4	475.24	480	2.2415
Eye	291	6	264.50	264	2.2395
Child	823	4	622.09	625	2.2037
Water	1026	10	747.84	744	2.1948
Night	1041	8	735.80	736	2.1328
Man	2283	13	1441.57	1437	2.1207
Die	160	14	152.10	142	2.0953
Hand	265	16	225.68	216	2.0637
Tree	113	7	111.00	107	2.0168
Kill	121	17	116.24	103	2.0009
Eat	680	6	478.22	479	1.9920
Hear	356	5	274.21	275	1.9652
Mother	107	7	103.42	100	1.9615
Right	1032	35	670.49	649	1.9580
Road	99	4	95.42	95	1.9235
New	1648	12	982.99	980	1.9188
Know	968	12	597.17	593	1.8420
Fire	1017	17	620.87	616	1.8359
Head	337	42	241.75	208	1.8132
Foot	1282	14	745.23	740	1.8066
Husband	71	2	69.01	70	1.7994
Animal	69	3	67.01	67	1.7858
Hair	64	6	62.01	59	1.7500
Leg	90	9	79.52	75	1.7267
Father	86	9	76.66	72	1.7243
Good	303	27	207.53	190	1.6996
Sit	187	8	136.25	134	1.6553
Bad	72	17	63.84	51	1.6468
Grass	50	10	48.01	41	1.6314
Sky	50	2	48.01	49	1.6314
Black	91	23	74.25	56	1.5985
Knee	55	3	50.25	51	1.5899
See	1227	25	626.97	617	1.5783
Dog	50	8	46.43	42	1.5776
Smoke	91	10	73.23	68	1.5766
Day	1314	10	648.65	648	1.5395
Ice	40	10	38.02	31	1.5226
Stone	629	16	338.61	330	1.5066
Laugh	83	4	65.04	65	1.5039
Tooth	43	5	39.43	38	1.4978

Moon	38	9	36.02	30	1.4974
Neck	38	5	36.02	34	1.4973
White	117	25	84.33	65	1.4907
Nose	45	14	40.25	30	1.4789
Bird	36	6	34.02	31	1.4706
Old	1066	9	516.96	515	1.4683
Sea	43	4	38.25	38	1.4530
Forest	39	3	35.43	36	1.4454
Push	88	15	65.31	56	1.4431
Breathe	33	9	31.02	25	1.4274
Wind	47	15	39.85	29	1.4176
Stand	330	24	183.89	169	1.4034
Sun	65	7	50.06	47	1.3961
Mouth	74	11	54.90	49	1.3867
Sleep	85	6	60.04	58	1.3628
Pull	90	24	62.24	44	1.3515
Green	44	14	36.12	26	1.3493
Live	264	19	144.63	133	1.3267
Salt	39	10	32.15	26	1.3115
Think	602	14	283.38	277	1.3085
Warm	57	13	42.35	34	1.3047
Leaf	25	6	23.03	20	1.2876
Fat	34	10	28.44	22	1.2810
Cold	75	16	51.20	40	1.2801
Wet	34	9	28.19	23	1.2697
Dust	64	7	44.30	42	1.2501
Fear	127	8	75.29	73	1.2473
Egg	23	5	21.03	19	1.2450
Rub	23	5	21.03	19	1.2450
Blow	72	29	47.92	25	1.2362
Throw	110	20	66.14	53	1.2274
Snake	29	8	24.27	20	1.2236
Ear	51	5	36.53	36	1.2231
Flower	41	4	31.09	31	1.2231
Turn	2091	38	765.49	744	1.2155
Bite	25	13	21.46	12	1.2002
Dirty	25	13	21.46	12	1.2002
Round	48	26	34.06	13	1.1930
Earth	105	9	61.28	57	1.1796
Fly	74	20	46.66	33	1.1787
Near	80	9	48.13	44	1.1450
Yellow	42	8	29.52	26	1.1411
Float	33	15	24.70	14	1.1366

Spit	18	8	16.05	11	1.1193
Mountain	18	2	16.03	17	1.1179
Wipe	18	2	16.03	17	1.1179
Cloud	51	13	33.10	24	1.1081
Root	29	15	21.89	11	1.1039
Seed	28	13	21.19	12	1.0954
Red	79	8	45.20	43	1.0858
Fall	169	44	81.90	46	1.0796
Hold	3906	45	1154.76	1134	1.0619
Hunt	19	15	15.65	4	1.0533
Come	792	22	286.15	275	1.0473
Tongue	27	10	19.70	14	1.0444
Sing	86	5	46.30	46	1.0414
Dry	59	19	34.53	20	1.0363
Fight	729	9	263.86	268	1.0361
Full	94	13	49.02	42	1.0290
Liver	15	5	13.05	11	1.0232
Lie	208	10	91.24	89	1.0168
Squeeze	34	17	22.52	10	1.0144
Dull	30	19	20.58	5	1.0131
Heart	88	10	45.76	42	1.0111
Stick	48	26	28.40	7	0.9949
Fog	25	4	17.69	18	0.9890
Swim	14	3	12.05	12	0.9861
Scratch	29	14	19.49	9	0.9826
Tail	17	11	13.54	6	0.9798
Fish	22	6	15.87	14	0.9686
Horn	19	11	14.36	7	0.9664
Dig	23	11	16.25	9	0.9619
Split	31	19	19.66	5	0.9457
Count	52	11	28.44	23	0.9384
Skin	28	11	17.89	11	0.9247
Far	155	10	64.79	62	0.9156
Tie	47	18	25.72	13	0.9150
Bark	12	9	10.16	4	0.9139
Star	25	12	16.34	8	0.9134
Smooth	30	12	18.48	11	0.9100
Freeze	26	14	16.71	7	0.9093
Feather	12	7	10.10	6	0.9082
Drink	74	10	35.82	32	0.9049
Rain	44	4	24.23	25	0.9049
Back	302	28	109.19	92	0.8967
Left	485	24	161.43	151	0.8939

Bone	17	6	12.31	10	0.8908
Hit	1627	24	449.55	440	0.8873
Correct	40	12	22.05	15	0.8830
Wash	56	21	27.76	13	0.8667
Fruit	16	5	11.48	10	0.8637
Straight	59	21	28.69	14	0.8611
Belly	14	6	10.50	8	0.8593
Sand	14	4	10.48	10	0.8577
Play	331	52	109.34	70	0.8324
Burn	55	20	26.09	11	0.8256
Breast	12	6	8.54	6	0.7678
Give	805	42	206.70	181	0.7461
Sharp	45	15	19.15	9	0.7036
Flow	66	14	25.12	18	0.6924
Rope	8	4	6.12	5	0.6912
Swell	24	11	11.66	5	0.6704
Suck	10	6	6.65	4	0.6650
Stab	6	6	5.00	1	0.6485
Wing	31	10	13.23	8	0.6363
Snow	37	6	14.71	13	0.6233
Guts	8	6	5.41	2	0.6112
Smell	46	8	15.99	14	0.5781
Worm	8	5	4.83	3	0.5451
Meat	6	3	4.16	4	0.5398
Sew	6	2	4.12	5	0.5343
Ash	5	4	3.41	2	0.4773
Louse	5	4	3.41	2	0.4773
Lake	5	3	3.24	3	0.4524
Vomit	4	4	3.00	1	0.4515
Rotten	3	3	2.00	1	0.3181

It is noteworthy to compare Λ with its maximum value Λ_{max} corresponding to the longest possible arc length $L_{max} = f(1) + V - 2 = N - 1$ (cf. Popescu, Lupea, Tatar, Altmann 2014: Ch. 3.2.3. The lambda indicator). Introducing this expression in Eq. (3) we obtain an approximate Λ_{max} as

$$\Lambda_{max} = (1 - 1/N)\mathrm{Log}_{10}N \approx \mathrm{Log}_{10}N \text{ (for } N >> 1)$$

The positions of individual words in their relation to maximum Lambda, Λ_{max}, are displayed in Figure 6. As can be seen, the greater is N, the greater is the dispersion of values. A linear increase cannot be stated. The dots lie in the area between the straight lines $\Lambda = \mathrm{Log}_{10}(N-1)$ and $\Lambda = 0.2568\mathrm{Log}_{10}N$.

The same holds true for the relationship between Lambda and *V* which yields a horizontal, strongly oscillating line

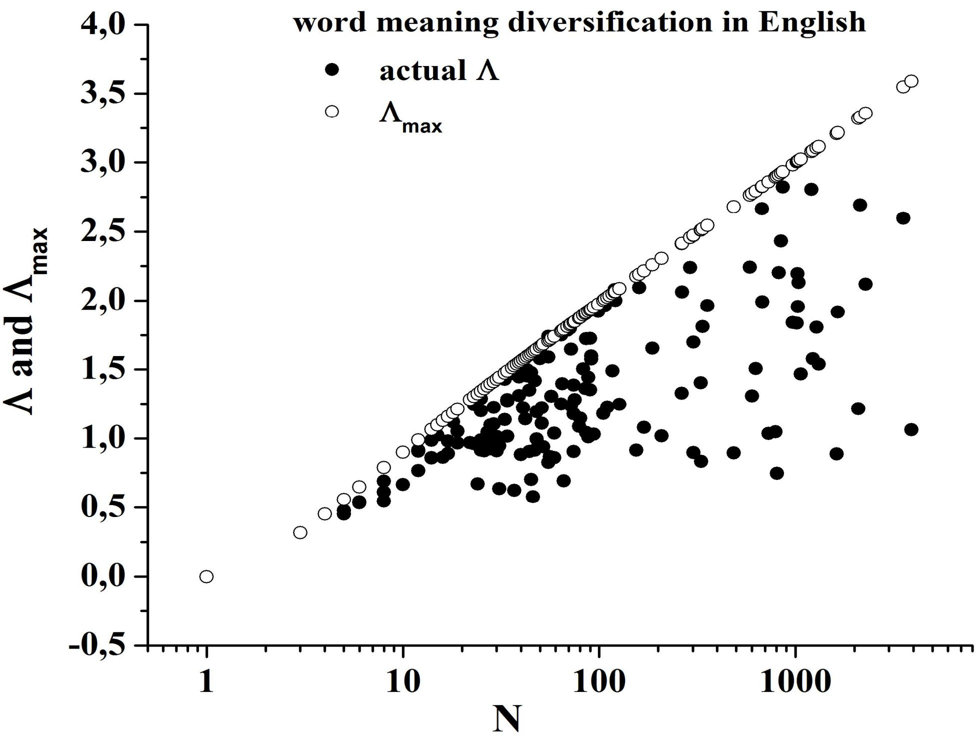


Figure 6. Lambdas of meaning diversifications in English

Word associations

Associations are representatives of connotative meaning. There are no two people having the same battery of associations of any word. Associations depend on personal history. A great part of them is the contents of a conversation in which person A says to person B matter that is not known to B. Hence learned or experienced matter whereby the main meaning remains unchanged. In association dictionaries one can observe that some words are heavily loaded with associations, e.g. *music* or *father*. Now, asking many persons for their associations concerning a word, we obtain a frequency distribution of connotations for which lambda can be computed in order to see the place of associations in the hierarchy of language levels.

In Table 9 one finds the associations of French based on Thérouanne, Denhière (2004) with computed lambda. The interval of lambda is <0.6382; 1.7504>. The ordering according to inventory does not yield any trend.

Table 9
French word associations
(Thérouanne, Denhière 2004)

Word	**Inventory**	**f(1)**	**L**	**Λ**
accès	42	17	52.06	1.0412
accolade	49	16	58.68	1.1736
adresse	36	22	50.96	1.0192
affection	27	30	49.67	0.9934
air	33	19	45.08	0.9016
ampoule	16	65	75.87	1.5174
arête	20	71	86.29	1.7258
artifice	19	65	78.41	1.5682
aube	25	26	44.58	0.8916
aval	17	75	87.52	1.7504
avocat	38	13	44.11	0.8822
baie	51	10	54.65	1.0930
baleine	33	13	38.71	0.7742
bâtiment	33	38	64.80	1.2960
bide	20	66	80.35	1.6070
bidet	19	40	52.24	1.0448
bière	33	35	62.13	1.2426
bise	25	40	57.96	1.1592
blaireau	45	25	63.92	1.2784
bouc	23	40	56.03	1.1206
bourdon	23	40	56.62	1.1324
bretelle	28	26	47.46	0.9492
cachet	34	22	48.61	0.9722
cafard	30	30	53.34	1.0668
calcul	25	25	42.47	0.8494
canapé	25	29	47.75	0.9550
cancer	24	57	75.32	1.5064
canne	26	23	41.12	0.8224
carrière	37	14	43.26	0.8652
case	44	13	50.41	1.0082
chausson	25	39	57.54	1.1508
chemise	26	29	47.67	0.9534
chenille	30	49	73.34	1.4668
cheville	35	33	61.49	1.2298
cliché	22	62	78.89	1.5778
comté	21	71	87.07	1.7414
conception	43	11	47.47	0.9494
cor	27	36	57.08	1.1416
couette	24	35	51.84	1.0368
cousin	31	29	53.47	1.0694
dauphin	31	16	40.28	0.8056
dé	20	58	72.42	1.4484
déduction	36	30	59.08	1.1816

devise	58	14	65.88	1.3176
discipline	37	31	61.33	1.2266
dossier	56	9	59.13	1.1826
éclair	22	29	43.45	0.8690
élan	42	25	60.81	1.2162
ellipse	39	12	43.78	0.8756
entretien	36	21	50.07	1.0014
esquimau	13	50	56.96	1.1392
essai	45	23	61.19	1.2238
étalon	17	71	83.41	1.6682
étiquette	43	19	54.77	1.0954
expiration	18	54	66.69	1.3338
exposant	43	10	47.19	0.9438
facture	29	13	35.00	0.7000
faculté	35	27	55.36	1.1072
farce	30	29	52.57	1.0514
feuille	25	32	51.10	1.0220
filature	43	16	51.57	1.0314
fléau	48	21	62.35	1.2470
flûte	22	35	50.83	1.0166
forfait	30	29	53.21	1.0642
four	18	25	34.98	0.6996
fraise	20	37	49.75	0.9950
fronde	38	17	47.72	0.9544
fugue	42	34	70.38	1.4076
garrot	30	36	59.31	1.1862
gratin	26	22	40.51	0.8102
gravité	58	11	62.88	1.2576
grenade	23	30	45.69	0.9138
grève	45	21	59.93	1.1986
grue	38	24	55.35	1.1070
héroïne	37	23	52.44	1.0488
identité	22	38	53.03	1.0606
imposition	37	23	53.02	1.0604
index	29	32	55.05	1.1010
induction	45	44	84.08	1.6816
iris	18	59	71.87	1.4374
lama	29	20	40.71	0.8142
latitude	24	68	87.07	1.7414
légende	21	31	44.66	0.8932
lentille	31	23	46.92	0.9384
lettre	33	17	42.34	0.8468
lézarde	39	14	45.53	0.9106
licence	28	35	55.82	1.1164
livre	34	19	45.03	0.9006
maîtresse	23	44	60.38	1.2076
majorité	47	17	56.69	1.1338
manège	25	31	49.19	0.9838

maquereau	14	64	72.02	1.4404
marabout	34	19	45.65	0.9130
melon	27	34	54.37	1.0874
mémoire	50	19	63.38	1.2676
milieu	34	51	79.42	1.5884
morse	37	15	44.18	0.8836
mortier	48	18	58.95	1.1790
motif	32	29	54.01	1.0802
mousse	37	20	50.46	1.0092
mule	28	43	65.30	1.3060
mutation	29	23	45.62	0.9124
mystère	51	12	56.40	1.1280
navet	24	45	62.14	1.2428
note	38	15	46.38	0.9276
oeillet	23	54	70.44	1.4088
orbite	28	27	47.84	0.9568
page	20	44	57.98	1.1596
palais	34	27	53.75	1.0750
parabole	28	23	44.26	0.8852
parquet	23	47	63.73	1.2746
partition	23	59	77.17	1.5434
patron	34	25	52.93	1.0586
pêche	23	40	56.28	1.1256
pensée	49	12	54.40	1.0880
pépin	15	24	31.92	0.6384
perception	31	23	46.44	0.9288
perche	36	25	53.75	1.0750
pétrin	32	37	62.32	1.2464
pieu	36	37	67.34	1.3468
pignon	38	42	75.00	1.5000
platine	17	30	38.55	0.7710
plongeur	32	19	43.53	0.8706
police	52	15	60.96	1.2192
polo	20	20	31.91	0.6382
pouce	12	55	60.96	1.2192
profession	29	35	56.82	1.1364
puce	33	17	43.27	0.8654
punaise	43	32	69.42	1.3884
pupille	19	56	69.12	1.3824
quarantaine	41	32	66.74	1.3348
radiation	48	17	58.67	1.1734
rame	24	43	60.40	1.2080
rate	33	33	60.03	1.2006
recette	18	65	77.65	1.5530
réflexion	36	37	66.73	1.3346
remise	35	19	46.82	0.9364
réplique	39	25	57.05	1.1410
révolution	43	18	54.66	1.0932

secrétaire	38	26	57.26	1.1452
sinus	21	53	68.88	1.3776
sirène	34	17	44.87	0.8974
sol	38	41	73.09	1.4618
solution	19	44	56.39	1.1278
somme	25	37	55.05	1.1010
souci	33	33	59.73	1.1946
soupir	43	16	52.03	1.0406
spectre	24	51	68.30	1.3660
stade	26	41	60.73	1.2146
tare	55	17	65.53	1.3106
timbale	33	22	48.47	0.9694
timbre	16	57	66.99	1.3398
trafic	26	32	50.86	1.0172
trapèze	24	49	67.08	1.3416
treillis	18	30	41.44	0.8288
trombone	31	14	38.33	0.7666
truffe	24	30	46.71	0.9342
tuteur	41	14	47.63	0.9526
vase	22	66	82.31	1.6462
vecteur	37	26	55.84	1.1168
vedette	22	58	74.17	1.4834
vol	34	25	52.92	1.0584

Colors

One can study a special class of words belonging to a semantic or grammatical set. One can state that the individual elements do not occur with the same frequency. Ranked according to their occurrence they display a special lambda. Here we chose color names whose number and individual frequencies are different in different languages. Pawlowski (1999) collected some counts which are evaluated in Table 10

Table 10
Lambdas of color names

Language	N	V	L	Var(L)	Λ	Var(Λ)
Czech	2500	12	601.9840	3922.1928	0.8182	0.007246
English	1358	12	358.8057	2136.7226	0.8278	0.011372
French (Juilland)	460	8	129.3329	131.0712	0.7487	0.004392
French (Engwall)	1238	10	289.4484	1033.1335	0.7231	0.006448
Italian	706	10	144.4123	91.9735	0.5827	0.001498
Polish	391	11	92.6950	93.9738	0.6145	0.004130
Romanian	564	7	147.3701	538.6688	0.7189	0.012818
Russian	2278	12	457.5393	1657.2892	0.6744	0.003600
Slovak	2026	12	467.2186	2019.4162	0.7625	0.005379
Spanish	486	10	139.5078	177.5634	0.7712	0.005426
Ukrainian	1315	11	306.3862	842.5280	0.7267	0.004740

As can be seen, there is no similarity within a genetic family. Both the numbers of colors (*V*) found by Pawlowski (1999) are different, and the lambdas move in the small interval (0.58; 0.83). Nevertheless, we may conjecture that different semantic classes may move in different intervals.

Words

For testing the second hypothesis conjecturing that the variants have a greater lambda than the basic forms (e.g. word forms vs. lemmas) we may use the count of words and lemmas in 60 End-of-Year speeches of Italian presidents. The data are presented in Tables 11 and 12.

Table 11
Word-forms lambda in End-of-Year speeches of Italian presidents

Text	**N**	**Inventory**	**L**	**Word Λ**	**Var(Λ)**
1949Einaudi	194	140	143.5432	1.6928	0.001487
1950Einaudi	150	105	108.7800	1.5781	0.003510
1951Einaudi	230	169	172.2333	1.7686	0.001979
1952Einaudi	179	145	146.8929	1.8488	0.001530
1953Einaudi	190	143	145.8191	1.7489	0.002234
1954Einaudi	260	181	186.2913	1.7303	0.002064
1955Gronchi	388	248	255.3558	1.7038	0.001467
1956Gronchi	665	374	392.7529	1.6672	0.000765
1957Gronchi	1130	549	599.3767	1.6194	0.000731
1958Gronchi	999	460	488.0442	1.6236	0.000978
1959Gronchi	697	388	409.8201	1.6718	0.000776
1960Gronchi	804	434	462.2283	1.6703	0.000865
1961Gronchi	1252	622	674.0538	1.6677	0.000608
1962Segni	738	381	404.0091	1.5701	0.000629
1963Segni	1057	527	559.5185	1.6008	0.001093
1964Saragat	465	278	289.0321	1.6580	0.001090
1965Saragat	1053	510	547.7775	1.5736	0.001012
1966Saragat	1199	597	624.7671	1.6031	0.000875
1967Saragat	1056	526	562.9810	1.6120	0.000941
1968Saragat	1174	562	602.8260	1.5774	0.000804
1969Saragat	1584	692	759.8210	1.5357	0.000422
1970Saragat	1929	812	877.5755	1.4946	0.000932
1971Leone	262	168	173.0226	1.5970	0.001280
1972Leone	767	394	414.7079	1.5598	0.001091
1973Leone	1250	616	669.2188	1.6580	0.000680
1974Leone	801	426	445.7840	1.6160	0.000799

1975Leone	1328	632	678.9746	1.5968	0.000665
1976Leone	1366	649	685.1578	1.5727	0.000484
1977Leone	1604	717	780.7230	1.5601	0.000499
1978Pertini	1493	603	639.4469	1.3602	0.000791
1979Pertini	2302	800	848.3508	1.2348	0.000688
1980Pertini	1360	535	567.9546	1.3086	0.001089
1981Pertini	2818	911	983.9384	1.2042	0.000525
1982Pertini	2487	854	921.7382	1.2590	0.000481
1983Pertini	3748	1149	1236.6461	1.1797	0.000318
1984Pertini	1340	514	539.1823	1.2583	0.000695
1985Cossiga	2359	859	955.7467	1.3665	0.000532
1986Cossiga	1349	561	610.0912	1.4165	0.000915
1987Cossiga	2091	904	993.7626	1.5774	0.000438
1988Cossiga	2385	875	976.9096	1.3839	0.000543
1989Cossiga	1912	778	842.2127	1.4455	0.000594
1990Cossiga	3327	1222	1351.7941	1.4243	0.000372
1991Cossiga	418	241	254.7695	1.5976	0.002019
1992Scalfaro	2772	978	1072.8016	1.3316	0.000464
1993Scalfaro	2941	1074	1179.3043	1.3904	0.000410
1994Scalfaro	3605	1190	1333.2622	1.3152	0.000268
1995Scalfaro	4145	1341	1492.5157	1.2787	0.000346
1996Scalfaro	2085	866	934.0381	1.4869	0.000594
1997Scalfaro	4909	1405	1538.4429	1.1357	0.000273
1998Scalfaro	3995	1175	1281.1874	1.1550	0.000332
1999Ciampi	1941	831	877.3226	1.4862	0.000439
2000Ciampi	1844	822	871.2039	1.5429	0.000540
2001Ciampi	2097	898	965.5417	1.5288	0.000422
2002Ciampi	2129	909	984.9410	1.5397	0.000517
2003Ciampi	1565	718	763.4969	1.5585	0.000816
2004Ciampi	1807	812	869.7050	1.5676	0.000527
2005Ciampi	1193	538	576.2236	1.4860	0.000687
2006Napolitano	2204	929	1033.5266	1.5677	0.000590
2007Napolitano	1794	793	874.5688	1.5878	0.000476
2008Napolitano	1713	775	831.2543	1.5692	0.000687

Table 12
Lambda of lemmas in the End-of-Year speeches of Italian presidents

Text	N	Inventory	L	Lemma Λ	Var(Λ)
1949Einaudi	194	119	127.1045	1.4989	0.00002269
1950Einaudi	150	91	98.1783	1.4243	0.00003760
1951Einaudi	230	150	160.0910	1.6439	0.00002493
1952Einaudi	179	123	126.9508	1.5978	0.00000641
1953Einaudi	190	120	129.5522	1.5538	0.00003896
1954Einaudi	260	154	170.2779	1.5816	0.00005704
1955Gronchi	388	206	220.3942	1.4705	0.00001092
1956Gronchi	665	321	367.2689	1.5590	0.00004024
1957Gronchi	1130	461	558.7041	1.5095	0.00004519
1958Gronchi	999	380	477.2652	1.4330	0.00006120
1959Gronchi	697	332	388.3051	1.5840	0.00003966
1960Gronchi	804	375	429.7412	1.5529	0.00002647
1961Gronchi	1252	512	628.0524	1.5539	0.00004494
1962Segni	738	329	390.0507	1.5158	0.00004005
1963Segni	1057	449	523.7339	1.4984	0.00003045
1964Saragat	465	226	252.4964	1.4484	0.00002331
1965Saragat	1053	429	508.8958	1.4607	0.00002620
1966Saragat	1199	501	610.7240	1.5682	0.00005171
1967Saragat	1056	459	538.5981	1.5422	0.00004741
1968Saragat	1174	468	569.2860	1.4885	0.00004199
1969Saragat	1584	573	702.3905	1.4189	0.00004379
1970Saragat	1929	672	846.7585	1.4421	0.00003780
1971Leone	262	141	158.1689	1.4599	0.00012093
1972Leone	767	328	379.7273	1.4282	0.00002673
1973Leone	1250	503	598.5375	1.4829	0.00002888
1974Leone	801	347	397.6459	1.4415	0.00002247
1975Leone	1328	530	640.0503	1.5053	0.00004524
1976Leone	1366	532	617.9603	1.4184	0.00001416
1977Leone	1604	581	685.0994	1.3690	0.00001682
1978Pertini	1493	481	555.4949	1.1810	0.00000811
1979Pertini	2302	625	738.0466	1.0779	0.00000662
1980Pertini	1360	426	490.7566	1.1307	0.00001229
1981Pertini	2818	698	842.6040	1.0316	0.00000743
1982Pertini	2487	668	833.2466	1.1377	0.00001527
1983Pertini	3748	884	1139.1061	1.0862	0.00001783
1984Pertini	1340	398	485.0885	1.1320	0.00002507
1985Cossiga	2359	701	876.4488	1.2531	0.00001401

1986Cossiga	1349	474	584.2881	1.3557	0.00004104
1987Cossiga	2091	756	929.6268	1.4762	0.00001944
1988Cossiga	2385	711	909.2989	1.2877	0.00002677
1989Cossiga	1912	650	796.8741	1.3676	0.00001871
1990Cossiga	3327	956	1230.0395	1.3022	0.00001585
1991Cossiga	418	207	236.8690	1.4853	0.00006803
1992Scalfaro	2772	783	952.6258	1.1832	0.00001194
1993Scalfaro	2941	861	1055.3656	1.2447	0.00001259
1994Scalfaro	3605	926	1195.7734	1.1798	0.00001734
1995Scalfaro	4145	956	1225.1724	1.0693	0.00001290
1996Scalfaro	2085	701	837.2508	1.3328	0.00002275
1997Scalfaro	4909	956	1227.4088	0.9229	0.00001083
1998Scalfaro	3995	916	1193.9417	1.0763	0.00001626
1999Ciampi	1941	656	812.2236	1.3759	0.00004503
2000Ciampi	1844	670	777.9629	1.3778	0.00001799
2001Ciampi	2097	726	876.5757	1.3885	0.00003540
2002Ciampi	2129	747	897.7441	1.4034	0.00002740
2003Ciampi	1565	575	675.0750	1.3780	0.00002568
2004Ciampi	1807	652	702.8227	1.2668	0.00002307
2005Ciampi	1193	438	525.9108	1.3563	0.00003981
2006Napolitano	2204	760	871.9451	1.3226	0.00001049
2007Napolitano	1794	661	783.4723	1.4210	0.00001268
2008Napolitano	1713	618	720.3159	1.3598	0.00000876

Comparing the individual speeches, one can see that at this high level word forms have a greater lambda because some forms have priority over other ones.

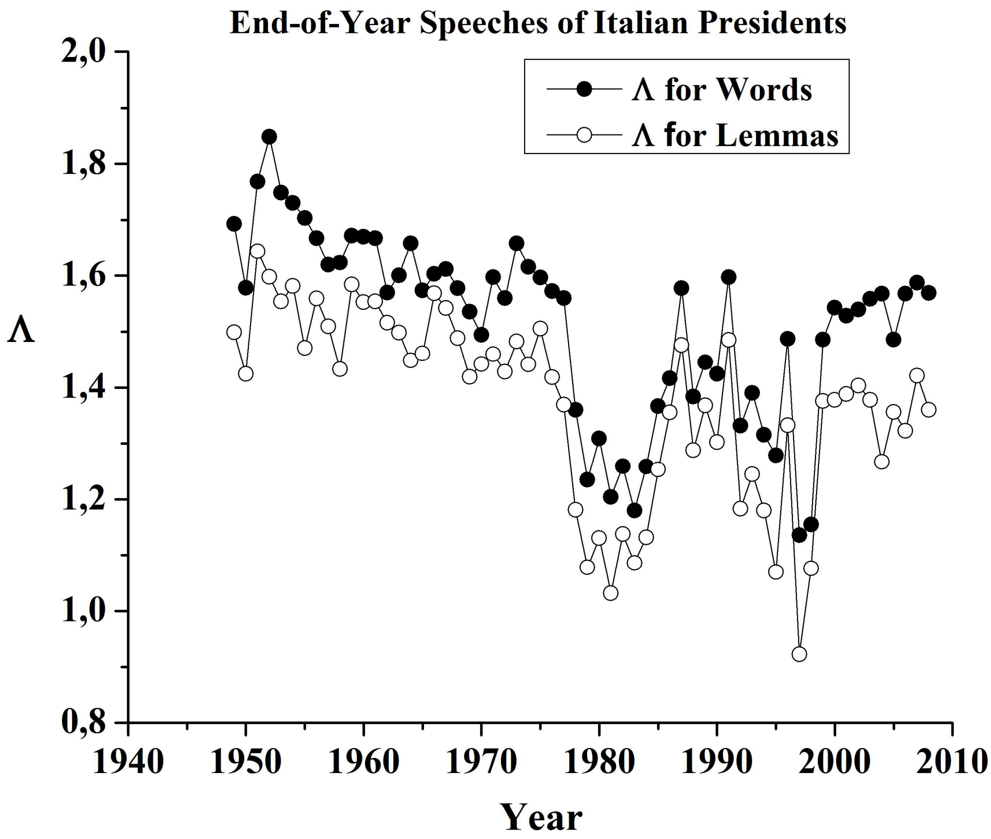


Figure 7. The course of Lambda in the End-of-Year Speeches of Italian Presidents

Suprasentence

Hrebs

Originally, hrebs were considered aggregates of sentences associated by a word, a synonym or a reference (cf. Hřebíček 1997), today one may define also word-hrebs, morpheme-hrebs or phrase-hrebs, etc., without recourse to the underlying sentence. Hence elements of a hreb may occur in the same or in different sentences. The individual hrebs as wholes have, of course, different frequencies of occurrence in the text and the individual elements of specific hrebs, too. Hence we obtain a distribution for the text and individual distributions for each individual hreb. Hrebs have all properties of linguistic entities and may be studied separately. They may have intersections, i.e. an entity can be element of several hrebs at the same time – a usual phenomenon with pronouns or even personal endings of verbs. Hrebs are useful for measuring e.g. the thematic concentration of texts and other text properties.

Here we restrict ourselves to word-hrebs in some Romanian poems written by M. Eminescu. The results are presented in Table 13. The parenthesis with a number in the second column signifies the number of 1 in the rank-ordered distribution. As can be seen, the value of lambda is in all cases greater than 1.

Table 13
Hrebs in Romanian poems by M. Eminescu

Text	**Frequencies**	**N**	**L**	**Λ**
Lacul	9,6,4,3,3,2,2,2,2,2,2,2,(39)1	78	54.6410	1.3255
Dintre sute de catarge	9,9,4,3,3,3,2, (19)1	52	30.3417	1.0013
La mijloc de codru	6,3,2,(22)1	33	26.9907	1.2420
Pe lângă plopii fără soț	19,16,5,4,3,3,3,3,3,3,2,2,2,2,2,2,2, 2,2,2,2,2,2,2,(58)1	148	94.8645	1.3911
Peste vârfuri	5,4,2,2,2,2,2,(19)1	38	27.0645	1.1252
Somnoroase păsărele	5,4,3,3,3,2,2,2,2,(26)1	52	35.6569	1.1767
Atât de fragedă	23,15,4,4,3,3,3,3,2,2,2,2,2,2,(64)1	134	95.3503	1.5136
La steaua	7,4,4,4,4,3,3,2,2,2,(20)1	55	32.4049	1.0254
Trecut-au anii…	10,7,4,3,3,3,3,3,2,2,2,(24)1	66	39.5672	1.0908
Ce te legeni?	11,10,4,3,3,2,2,2,2,2,2,2,(21)1	66	38.7396	1.0680
Mai am un singur dor	17,6,5,3,3,3,3,2,2,2,2,2,2,2,2,2,2, 2,2,35(1)	101	66.5241	1.3202

Comparison

For the sake of comparison, we collect all results in Table 14 but use merely the averages of the data in the previous tables. Thereby some shifts may occur. We consider the following levels:

1. Phonemic-graphemic
2. Closed classes: cases, parts-of speech, colors
3. Syntactic relations: dependence
4. Meaning diversification: word meanings, associations
5. Lexicon: word forms, lemmas
6. Suprasentence units: hrebs
7. Complex grammatico-semantic units: Hungarian affixes.

This is, of course, a very elementary scaling, not having sufficient empirical data

Table 14
Mean lambdas and inventories

Entity	Language	Average V	Average Λ
Letters	English	26	0.6974
	Russian	31.50	0.5593
	Different languages	27.87	0.6573
	12 Slavic languages	32.75	0.4496
Phonemes	12 Slavic languages	37.5	0.4767
Cases	German	4	0.7624

	Slovene	6	0.6534
	Slovak	6.56	0.6285
	Russian	6	0.7815
Parts-of-speech	Italian	10.08	0.7277
	Different languages	8.5	0.8711
Categories: Colors	Different languages	10.72	0.7244
Dependence relations	Hungarian	24.83	1.3999
Meaning diversification	English words	12.57	1.2879
	French word associations	31.33	1.1366
Word forms	Italian	655.08	1.5153
Lemmas	Italian	527.48	1.3736
Hrebs	Romanian	74.82	1.2073
Grammatico-semantic units	Hungarian affixes	24	3.0272

Table 14 shows a slight increase of lambda with increasing level as can be seen in Figure
8

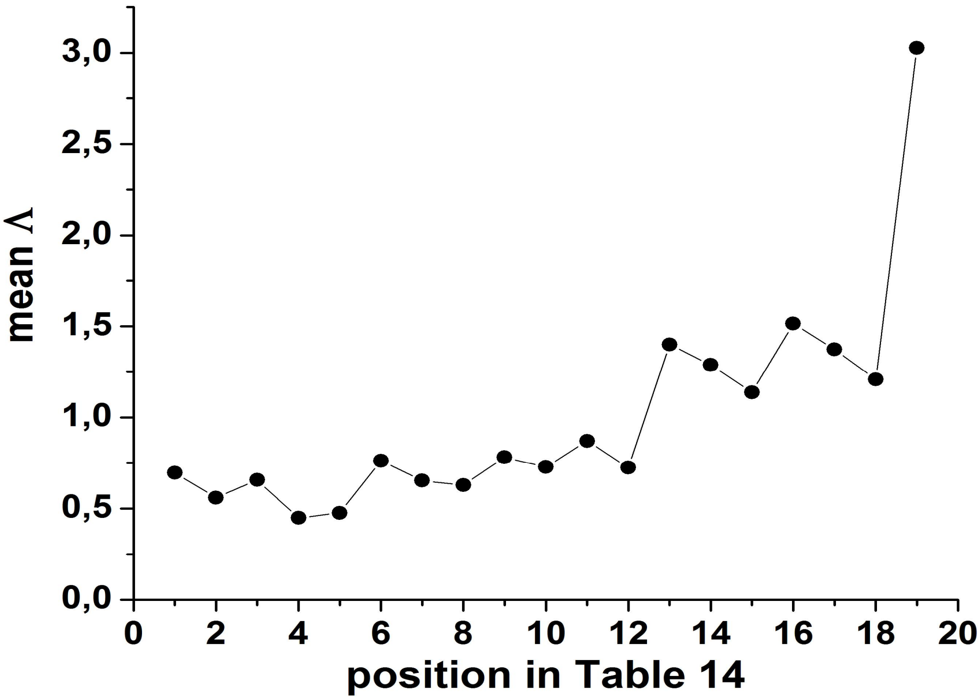


Figure 8. Increase of lambda with increasing level

If we take averages of individual levels, we obtain the results presented in Table 15 and Figure 9.

Table 15
Mean lambda for individual levels

Level	**Lambda**
Phonemes/graphemes	0,5681
Closed classes: Cases, POS, Colors	0,7356
Syntactic relations: Dependence	1,3999
Meaning diversification: word meanings, associations	1,2123
Lexicon: Word forms and lemmas	1,4445
Suprasentence units: Hrebs	1,2073
Grammatico-semantic units: Hungarian affices	3,0272

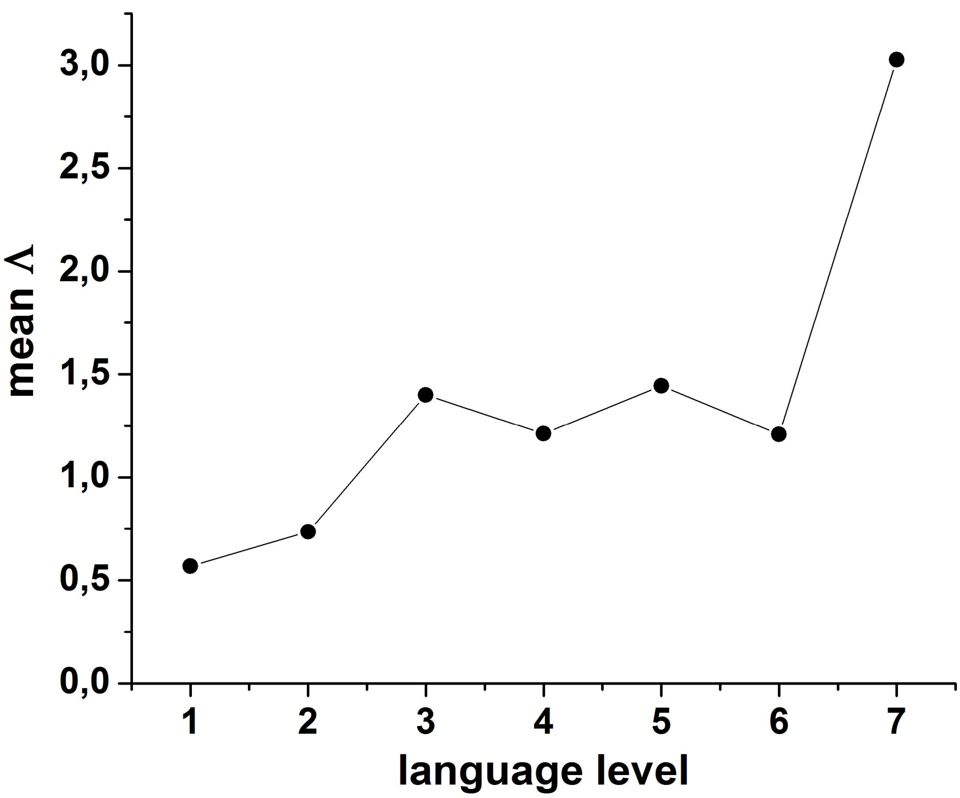


Figure 9. Increase of lambda according to levels

We can conclude that the inventory of entities is a factor influencing the lambda only on the same level of language while the abstractness of the level exerts a stronger influence on lambda regardless of the size of the inventory. Nevertheless, this conjecture may change if many other phenomena and languages will be analyzed.

References

Altmann, G. (1993). Phoneme counts: Marginal remarks on the Pääkkönen article. In: Altmann, G. (Ed.*), Glottometrika 14: 54-68*. Trier: Wissenschaftlicher Verlag Trier.

Altmann, G., Lehfeldt, W. (1980). *Einführung in die quantitative Phonologie*. Bochum: Brockmeyer.

Best, K.-H.(2004/2005). Laut- und Phonemhäufigkeiten im Deutschen. *Goettinger Beiträge zur Sprachwissenschaft 10/11, 21-32.*

Fan, F., Popescu, I.-I., Altmann, G. (2008). Arc length and meaning diversification in English. *Glottometrics 17, 79-86.*

Fry, D.B. (1947). The frequency of occurrence of speech sounds in Southern English. *Archives néerlandaises de phonétique expérimentale 20, 103-106.*

Grzybek, P., Kelih, E. (2003). Graphemhäufigkeiten (am Beispiel des Russischen). Teil I. Methodologische Vor-Bemerkungen und Anmerkungen zur Geschichte der Erforschung von Graphemhäufigkeiten im Russischen. *Anzeiger für Slavische Philologie 31, 131-162.*

Hřebíček, L. (1997). *Lectures on Text Theory*. Prague: Oriental Institute.

Kelih, E., **Popescu, I.-I., Altmann, G.** (2014). Some aspects of Slavic phonemics and graphemics. *Glottometrics 27, 18-53.*

Meier, H. (1967^2). *Deutsche Sprachstatistik*. Hildesheim: Olms.

Nemcová, E., Popescu, I.-I., Altmann, G. (2010). Word associations in French. In: Berndt, A., Böcker, J. (eds.), *Sprachlehrforschung: Theorie und Empirie: 223-237*. Frankfurt: Lang

Pääkkönen, M. (1993). Graphemes and context. In: Altmann, G. (ed.), *Glottometrika 14, 1-53.* Bochum: Brockmeyer.

Pawlowski, A. (1999). The quantitative approach to cultural anthropology: Application of linguistic corpora in the analysis of basic colour terms. *Journal of Quantitative Linguistics 6(3), 222-234.*

Popescu, I.-I., Čech, R., Altmann, G. (2011). *The Lambda-structure of Texts.* Lüdenscheid: RAM-Verlag.

Popescu, I.-I., Kelih, E., Best, K.-H., Altmann, G. (2009). Diversification of the case. *Glottometrics 18, 32-39.*

Popescu I.-I., Lupea, M., Tatar, D., Altmann, G. (2014). *Quantitative Analysis of Poetry*. Mouton de Gruyter (in print).

Popescu, I.-I., Mačutek, J., Altmann, G. (2009). *Aspects of Word Frequencies.* Lüdenscheid: RAM-Verlag.

Popescu, I.-I., Mačutek, J., Altmann, G. (2010). Word forms, style and typology. *Glottotheory 3(1), 89-96.*

Popescu, I.-I., Zörnig,P., Altmann,G. (2013). Arc length, vocabulary richness and text size. *Glottometrics 25, 43 – 53.*

Thérouanne, P., Denhière, G. 2004). Normes d'associantion libre et fréquence relatives des acceptions pour 162 mot homonyms. *L'Année Psychologique 104, 537-595.*

Vincze, V. (2013). Domain differences in the distribution of parts of speech and dependency relations in Hungarian. *Journal of Quantitative Linguistics 20(4), 314-448.*

Source: LETTER FREQUENCY STATISTICS
http://www.cryptogram.org/cdb/words/frequency.html

Texts

German

Text 01: „Nicht blind genug" heißt Startverbot, ET (= Eichsfelder Tageblatt), 9.9.2008, S. 28, Sparte: „Sport" .
Text 02: Es bleibt dabei: Mit links ist gut, ET, 9.9.08, S. 29, „Sport".
Text 03: Serena Williams ist wieder am richtigen Platz. ET, 9.9.08, S. 29, „Sport".
Text 04: Teuber: Hoffnungsträger und Vorbild zugleich. ET, 9.9.08, S. 28, „Sport".
Text 05: Eiskalte Gieboldehäuser besiegen Pferdeberg. ET, 17.9.08, S. 27, „Sport".
Text 06. Über Peking „kann man nur in Superlativen sprechen". ET, 17.9.08, S. 28, „Sport".
Text 07. Verletzter Czyz siegt für kranken Vater. ET, 17.9.08, S. 28, „Sport".
Text 08: Werder enttäuscht Bremer Fans. ET, 17.9.08, S. 29, „Sport",
Text 09. Schröder wartet zwei Stunden auf Gold. ET, 15.9.08, S. 20, „Sport".
Text 10. Bötzel fehlt noch immer die Medaille. ET, 15.9.08, S. 20, „Sport".

Deutsche Sagen. Hrsg. von den Brüdern Grimm. Berlin: Rütten & Loening 1984.
Text 11: Die drei Bergleute im Kuttenberg. S. 35f.
Text 12: Die Springwurzel. S. 41f.
Text 13: Die Schlangenjungfrau. S. 44f.
Text 14: Des kleinen Volks Hochzeitsfest. S. 58f.
Text 15: Zwerge leihen Brot. S. 61
Text 16: Das Bergmännlein beim Tanz. S. 65f.
Text 17: Der Wassermann. S. 73f.
Text 18: Die Elbjungfer und das Saalweiblein. S. 82f.
Text 19: Der Alraun. S. 120f. (1 lat. Zitat ausgelassen)
Text 20: Das Vogelnest. S. 124f. (1 lat. Wort ausgelassen)

Slovenian

Text 1-8: Cankar, Ivan (1898 – 1902): Private letters to Ana Lušinova. Ljubljana: DZS.
Text 9: Prežihov, Voranc (1940): Samorastniki. Chapter 1. (Novel). Ljubljana: Naša založba.
Text 10: Prežihov, Voranc (1940): Samorastniki. Chapter 2. Ljubljana. (Novel) Ljubljana: Naša založba.

Russian

All texts are from http://lib.ru/LITRA/CHEHOW/ (October 10, 2008)

Čechov, A.P.
Text 01: Chameleon. (1884).

Text 02: Ušla. (1983)
Text 03: Sovremennye molitvy. (1883).
Text 04: Sovet. (1883).
Text 05. Idillija. (1884)
Text 06: Na gvozde. (1883)
Text 07: Po-Amerikanski. (18ß0)
Text 08: Radost'. (1883)
Text 09: Rjažennye. (1883)
*Text 10: Temnuju noč'ju. (1883)

Slovak
All texts are from http://zlatyfond.sme.sk (October 1, 2008)

Text 01: Ján Stacho, Apokryfy: Noc
Text 02: Rudolf Dilong: Nevolaj, nevolaj: Minieme sa.
Text 03: Ján Ondruš, Korenie: Chodec po povraze
Text 04: Ján Kovalik Ústiansky, Z pút k slobode: Bratom za Oceánom
Text 05: Anton Prídavok, Lámané drieky
Text 06: Jozef Gregor Tajovský, Zajac
Text 07: Pavol Ušák Oliva, Čierne kvietie: Hviezdy a smútok
Text 08: Lýdia Vadkerti-Gavorníková, Trvanie: Leto
Text 09: Janko Kráľ, Šahy. 1849

Bibliography: Motifs

The linguistic motif, a new unit for sequential analyses, was introduced because in quantitative linguistics no adequate means were available for investigations in the syntagmatic dimension. Almost all studies were devoted to paradigmatic phenomena; models such as probability distributions and functions were (and are) predominant, which ignore the sequential organisation of the units in the text. Those methods which have been used to improve this situation did not prove to be appropriate for language and text or could achieve only part of their aims (cf. Köhler, Naumann 2010).

The construction of this unit, the motif (originally called segment or sequence, cf. Köhler 2006, 2008a,b; Köhler/Naumann 2008, 2009, 2010;) was inspired by the so-called F-motiv for musical "texts" (Boroda 1982). Boroda was in search for a unit which could replace the word as used in linguistics for frequency studies in musical pieces. Units common in musicology were not usable for his purpose, and so he defined the "F-Motiv" with respect to the duration of the notes of a musical piece.

For the purposes of linguistics, a much more general unit is needed; even the original definition,

> *the longest continuous sequence of equal or increasing values representing a quantitative property of a linguistic unit.*

was already generalised several times (cf. Beliankou, Köhler 2013). One of the aims of the most recent generalisations was to enable the researcher to form motifs from non-numerical data.

Meanwhile, a number of studies have been performed on the basis of various versions of motifs, e.g. frequency, polysemy, length and other motifs.

Beliankou, A., Köhler, R., Naumann, S. (2013). Quantitative properties of argumentation motifs. In: Obradović, I.. Kelih, E., Köhler, R. (eds.). *Methods and Applications of Quantitative Linguistics. Selected Papers of the 8th International Conference on Quantitative Linguisttics (QUALICO) in Belgrade, Serbia, April 26-29, 2012: 35-43*. Belgrade: Academic Mind**.**

Boroda, M.G. (1982). Häufigkeitsstrukturen musikalischer Texte. In: Orlov, J.K., Boroda, M.G., Nadarejšvili, I.Š. (eds.), *Sprache, Text, Kunst. Quantitative Analysen: 231-262.* Bochum: Brockmyer.

Čech, R., Altmann, G. (2011). *Problems in Quantitative Linguistics Vol. 3.* Lüdenscheid: RAM-Verlag.

Köhler, R. (2006). The frequency distribution of the lengths of length sequences. In: Genzor, J., Bucková, M. (eds.), *Favete linguis. Studies in honour of Viktor Krupa: 142-152*. Bratislava: Academic Press.

Köhler, R. (2008). Word length in text. A study in the syntagmatic dimension. In: Mislovičová, S. (ed.), *Jazyk a jazykoveda v pohybe: 416-421*. Bratislava: VEDA

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

Köhler, Reinhard (2014, to appear). Linguistic motifs. In: Mačutek, Ján., Mikros, Georgios (eds.). *Sequential Analysis*. Berlin, New York: de Gruyter.

Köhler, R., Altmann, G. (2009). *Problems in Quantitative Linguistics Vol. 2*. Lüdenscheid: RAM-Verlag.

Köhler, R., Naumann, S. (2008). Quantitative text analysis using L-, F- and T-segments. In: Preisach, B., Schidt-Thieme, D. (eds.), Data Analysis, Machine Learning and Applications. *Proceedings of the Jahrestagung der Deutschen Gesellschaft für Klassifikation 2007 in Freiburg: 637-646.* Berlin-Heidelberg: Springer.

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

Popescu, I.-I., Zörnig, P., Grzybek, P., Naumann, S., Altmann, G. (2013). Some statistics for sequential text properties. *Glottometrics 26, 50-94.*

Mačutek, Ján (2009): Motif richness. In: Köhler, Reinhard (ed.), *Issues in Quantitative Linguistics: 51-60.* Lüdenscheid: RAM-Verlag.

Sanada, H. (2010). Distribution of motifs in Japanese texts. In: Grzybek, P., Kelih, E., Mačutek, J. (eds.). *Text and Language. Structures – Functions – Interrelations – Quantitative Perspectives: 183-193.* Wien: Praesens.

Reinhard Köhler

Announcement

Quantitative Index Text Analyser (QUITA)

Miroslav Kubát, Vladimír Maltach, Radek Čech
(Palacký University, Olomouc)

New software for a quantitative text analysis has been developed at Palacký University in Olomouc, the Czech Republic. *Quantitative Index Text Analyser* (QUITA) covers the most common indicators, especially those connected with frequency structure of a text. In addition to computing results of the indicators, QUITA provides also statistical testing and graphical visualization of obtained data.

QUITA is a versatile tool with many uses designed for researchers from various disciplines (linguistics, criticism, history, sociology, psychology, politics, biology, etc.). The program enables basic text processing functions like creating word lists, text lemmatizing or creating n-grams. The program also provides more advanced tools such as a random text creator or a binary file translator. However, the main part of the software is an indicator computing. Although the authors focused mainly on the indicators connected to frequency structure of a text (e.g. *h*-point, entropy, repeat rate, adjusted modulus, Gini's coefficient, lambda), there are also several other characteristics such as thematic concentration, activity & descriptivity or writer's view.

The main purpose of QUITA is to provide user-friendly tool of quantitative text analysis for researchers (especially from the humanities) without deeper knowledge of quantitative linguistics, statistics and programming. Apart from generating results, QUITA also enables a simple statistical comparison and creating charts. There is no need to use any additional software such as spreadsheet applications or special statistical programs. In sum, QUITA is the program that combines all important parts of any quantitative research: obtaining results, statistical testing and graphical visualization.

In order to compare texts for authorship attribution, genre analysis or another purpose, the differences between obtained resulting values of several indicators can be statistically tested. QUITA provides not only statistical testing among particular texts but also among groups of texts. For creating graphs of obtained data, there is a special tool "Chart Wizard" which offers wide range of chart types and editing options. All results can be copied via clipboard or saved directly as CSV file. The charts can be saved as image files.

QUITA is a tool with wide range of application, from stylometry to DNA analysis. Although almost all indicators in the software were proposed as features for common linguistic research (e.g. authorship attribution, genre or thematic analysis), possibilities are practically endless. Biologists can use one of available tokenizers (DNA Triplet Tokenizer, DNA Nucleotide Tokenizer) to handle with DNA as a text and apply the indicators, for instance. There is also an option to use different units other then words or lemmas such as characters, n-grams, etc. It should be noted that the software is designed as multilingual tool; QUITA therefore works with almost all scripts and includes several tokenizers and lemmatizers. Nevertheless, especially the number of lemmatizers is still limited but it should be significantly extended in a next version of the software.

Since QUITA aims to help as many researchers as possible, the program will be distributed as freeware. Thus everybody can use QUITA without any restrictions. The software can be downloaded on the website http://oltk.upol.cz/software.

The software was developed as a student project at the Department of General Linguistics at Palacký University in Olomouc, the Czech Republic. The team consists of two students (Vladimír Matlach, Miroslav Kubát) and their supervisor Radek Čech. The indicators included in QUITA were mostly selected in accordance with following books: *Word frequency studies* (Popescu et al. 2009), *Aspects of Word Frequencies* (Popescu et al. 2009) and *Metody kvantitativní analýzy (nejen) básnických textů* (Čech et al. 2013).

Acknowledgement

QUITA (Quantitative Index Text Analyser) was supported by the student project IGA (no. FF_2013_031) of the Palacký University, Olomouc, Czech Republic.

References

Čech, R., Popescu, I. I., Altmann, G. (2013). *Metody kvantitativní analýzy (nejen) básnických textů*. Olomouc: Univerzita Palackého v Olomouci. (in press)

Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B.D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., Vidya, M.N. (2009). *Word frequency studies*. Berlin-New York: Mouton de Gruyter.

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

Books received

David, J., Čech, R., Davidová Glogarová, J., Radková, L., Šústková, H. (2013). *Slovo a text v historickém kontextu. Perspektivy historickosémantické analýzy jazyka.* Brno: Host. Pp. 324.

Janda, Laura A. (ed.) (2013). *Cognitive Linguistics: The Quantitative Turn. The Essential Rreader.* Berlin-Boston: Walter de Gruyter. Pp. 321

Pickl, Simon (2013). *Probabilistische Geolinguistik. Geostatistische Analysen lexikalischer Variation in Bayerisch-Schwaben.* Stuttgart: Franz Steiner Verlag.