Glottometrics 26
2013

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. 26 (2013), Lüdenscheid: RAM-Verlag, 2013. 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 26 (2013) **ISSN 2625-8226**

Contents

History of Quantitative Linguistics

Review

Glottometrics 26, 2013,1-8

Iranismen im Deutschen

Karl-Heinz Best

Abstract. Many authors examined the influence of loanwords in German (cf. Best 2001; Körner 2004, Ternes 2011). The present paper presents the development of Iranian borrowings and demonstrates that this process abides by the logistic law which in linguistics is known as Piotrowski Law.

Keywords: *loanwords, Persian, German, Piotrowski-law*

1. Iranismen

In diesem Beitrag geht es um die Entlehnung von Iranismen ins Deutsche. Es handelt sich dabei in fast allen Fällen um Wörter, die aus dem Persischen stammen oder zumindest über das Persische und in vielen Fällen über weitere Sprachen ins Deutsche gelangten. Dabei werden die Iranismen in einer Liste aufgeführt; die Liste enthält - soweit aus der Literatur zu entnehmen war – den Weg, den die jeweilige Entlehnung genommen hat und das Jahrhundert, in dem sie im Deutschen zuerst festgestellt wurde.

Die Untersuchung stützt sich auf gängige Wörterbücher, vor allem auf *Duden. Das große Fremdwörterbuch* (2007), sowie die etymologischen Wörterbücher *Duden. Das Herkunftswörterbuch* (2001), Kluge, *Etymologisches Wörterbuch der deutschen Sprache* (2002) und Pfeifer, *Etymologisches Wörterbuch des Deutschen* (1995). Bevorzugt wurden die Angaben von Kluge, da es das neueste der etymologischen Wörterbücher ist und sicher in Kenntnis seiner Vorläufer verfasst wurde.

Die Entlehnungswege wurden in der Regel für das jeweilige Wort insgesamt dargestellt, in etlichen Fällen handelt es sich wie z.B. bei „Angarien(recht)“ nur um Wortteile, in diesem Fall um die Konstituente „Angarien“; bei „Paprika“ fällt ein Teil der Entlehnungsgeschichte mit der von „Pfeffer“ zusammen.

Da die Angaben in den Wörterbüchern sich immer wieder unterscheiden, vor allem dann, wenn es um die Vollständigkeit der Entlehnungswege geht, aber auch in Hinblick auf die zeitlichen Angaben, kann die Zusammenstellung nicht ganz ohne Willkür erfolgen. Auch die Anzahl der Entlehnungen, ihre Schreibweise etc. ist natürlich von den Angaben der ausgewerteten Wörterbücher abhängig.

Spezialuntersuchungen zu den Iranismen im Deutschen scheint es bisher nicht zu geben. Die folgende Tabelle stellt die Entlehnungen ins Deutsche zusammen. Aufgenommen wurden Wörter, von denen mindestens eines der angegebenen Wörterbücher persische Herkunft annimmt oder wenigstens für möglich hält.

Tabelle 1
Entlehnungen von Iranismen im Deutschen

Entlehnung	Jhd.	Bedeutungshinweis	Entlehnungsweg
Achia, Atchia		Speise	hindi – pers.
Ajatollah, Ayatollah		schiitischer Ehrentitel	pers.
Akinakes		Kurzschwert	griech. – pers.
Angarien(recht)		Pflicht zu Transportleistungen	lat. – griech. – pers.

Angaroi		Eilboten	griech. – pers.
Anilin	18.	Farbstoff	frz. – arab. – portug. – pers.
Aranzini		überzuckerte Orangenschalen	it. – pers.
Arbuse		Wassermelone	russ. – pers.
Arsenik	15.	giftige Arsenverbindung	lat. – griech. – pers.
Aubergine	20.	Frucht	frz. – katalanisch – arab. – pers.
Awesta		Hl. Schrift der Parsen	pers.
Azur	17.	himmelblau	frz. – span. – arab. – pers.
Babist		Anhänger des Babismus, einer religiösen Bewegung	pers.
Babuschen/ Pampusche	18.	Hausschuhe	frz. – arab. – pers.
Bahai		Anhänger des Bahaismus	pers.
Bakschisch	17.	Geldgeschenk	... – pers.
Barbakane		Festungsanlage	frz./ it. – arab. oder pers.
Basar/ Bazar	16.	Markt	frz. – pers.
Bedeguar		Wucherung an Pflanze	frz. – pers.
Bendak		Kopfbedeckung	pers.
Bezoar		Magenstein	span. – arab. – pers.
Bombast	18.	Schwulst	engl. – frz. – lat. – griech. – pers.?
Bor	19.	chemisches Element, Kopfwort zu Borax	lat. – arab. – pers.
Borat		borsaures Salz	pers.
Borax		Natriumsalz der Borsäure	lat. – arab. – pers.
Bostan		Garten	türk. – pers.
Bronze	16.	Kupferlegierung	it. – arab.? – pers.
Bülbül		Nachtigall	pers. – arab.
Caravan		Wohnwagen	engl. – it. – lat. – pers.
Chan (Han)		Herberge, Karawanserei	pers./arab.
Csárda, Tscharda		Weinlokal	ungar. – türk. – pers.
Csardas, Csárdás	19.	ungarischer Tanz	ungar. – serbokroat. – türk. – pers.
Dakhma		Totenturm	pers.
Dalk		Mönchs-, Derwischkutte	pers.
Damad		Titel	pers.
Derwisch	16.	muslimischer Ordensangehöriger	türk. – pers.
Divan/ Diwan	17.	Sofa; Gedichtsammlung	frz. – it. – türk. – pers.
Doab		Zwischenstromland	pers. – sanskr.
Do-Gule		Teppichmuster	pers.
Dosar		Teppichmaß	pers.
Durbar		offizieller Empfang	engl. – hindi – pers.
Dutar		Laute	pers.
Echec		Schach	frz. – span. – pers.
Enderun		Frauenraum	türk. – pers.
Ferahan		Teppich	pers.

Ferman		Erlass	türk. – pers.
Frawaschi		Schutzgeist	awest.
Gaze	17.	leichtes Gewebe	ndl. – frz. – span. – arab.? – pers.?
Gazophylacium		Schatzkammer	lat. – griech. – pers.
Gerus		Teppich	pers.
Giaur		Ungläubiger, Nichtmoslem	türk. – pers.? – arab?
Hanum		Anrede für Frauen	türk. – pers.
Heris		Teppich	pers.
Hindu		Anhänger des Hinduismus	pers.
Hodscha		geistlicher Lehrer	türk. – pers.
Huri		Paradiesjungfrauen (Islam)	pers. – arab.
Ibrik		Wasserkanne	arab. – pers.?
Ilchan		mongolischer Herrschertitel	pers.
Jasmin	16.	Zierstrauch	frz. – span. – arab. – pers.
Juchten	17.	Juchtenleder	ndt. – russ. – turkotatar. ?– pers.?
Julep		Erfrischungsgetränkt	engl. – frz. – arab. / pers.
Kaftan	16.	Obergewand	türk. – arab. – pers.
kaki, khaki	20.	sandfarben	engl. – urdu – pers.
Kalian/ Kaliun		Wasserpfeife	pers.
Karawane	16.	Reisegruppe	it. – lat. – pers.
Karawanserei	17.	Unterkunft für Karawanen	pers.
Karmesin	15.	kräftiges Rot	it. – arab. – pers.?/altind.?
Kawir/Kewir		Salzwüste	pers.
Kelei		Teppich	pers.
Kelek		Floß	türk. – pers.
Kemantsche		Geige	pers.
Kenare		Teppich	pers.
Khedive		Fürstentitel	türk. / pers.
Kiosk	18.	Pavillon	frz. – türk. – pers.
Kulah		Mütze	türk. – pers.
Lack	16.	Schutzanstrich	it. – lat. – arab. / pers.– ind.
Lala		Erzieher	türk. – pers.
Lasur/Lazur	13.	Schicht aus durchsichtiger Farbe	lat. – arab./pers.?
Lila	18.	fliederfarben	frz. – span. – arab. – pers. – ind.
Limette		Zitrone	it. – pers.
Limonade	17.	Getränk	frz. – arab. – pers.
Limone	14.	Zitrone	frz. – arab. – pers.
Liwan		Raum, Moschee	arab. – pers.
Lumie		Zitrusfrucht	lat. – pers.
Magie	16.	Zauberkunst	lat. – griech. – pers.
Magier	18.	Zauberpriester	lat. – griech. – pers.
Magus	16.	Priesterstamm	pers.
Mahal		Teppich	pers.
Man		früheres pers. Gewicht	pers.

Manichäer		Anhänger des Manichäismus	lat. – nach pers. Religionsstifter „Mani“
Maral		Hirschart	pers.
Maristan		Anlage mit Hospital, Moschee, Stiftergrab	pers.
Markhor		Ziegenart	pers.
matt	13.	Schachausdruck	arab./pers.?
Mazdaznan		religiöse Heilsbewegung	pers.
Mir		Teppich	pers.
Miri		Teppichmuster	pers.
Mirza		Ehrentitel	pers.
Mogul		Herrscher	engl. – pers.
Moschus	17.	Sekret männl. Moschustiere	lat. – griech. – pers. – ind?
Mulla, Mullah		Ehrentitel, Titelträger	pers. – arab.
Mumie	16.	einbalsamierter Leichnam	it. – arab. – pers.
Muselman(n)	17.	Moslem	it., frz. – türk. – pers. – arab.
Namas, Namaz		Stundengebet	türk. – pers. – sanskr.
Naphtha	16.	Erdöldestillat	lat.? – griech. – pers.
Narde	9.	Pflanze/Duftstoff	lat. – griech. – arab.? – pers. – sanksr.
Nargileh		Wasserpfeife	pers.
Nauroz		Neujahrsfest	pers.
Nay		Blasinstrument	arab. – pers.
orange		goldgelb	frz. – pers.
Orange	17.	Frucht	frz. – span. – arab. – pers.
Padischah		Fürstentitel	pers.
Paprika	19.	Gemüse	ung. – serb. – lat. – griech. – pers. – ind.
Para		jugoslawische Währungseinheit	türk. – pers?
Paradies	8.	Garten Eden, Himmel	lat. – griech. – pers.
Parasange		altpers. Wegemaß	lat. – griech. – pers.
Parse		Anhänger des Zarathustra	pers.
Pascha	16.	herrischer Mann	türk. – pers.
Paschmina		Wollgewebe	pers.
Paschto, Paschtu		Sprache	pers.
Pehlewi		Mittelpersisch	pers.
Perborat		chem. Verbindung	lat./ pers.
Peri		feenhaftes Wesen	pers.
Perkal		Baumwollgewebe	frz. – pers.
Pfeffer	8.	Gewürz	lat. – griech. – pers. – ind.
Pilau/ Pilaw		Reiseintopf	pers. – türk.
Pistazie	16.	Baum, Frucht	lat. – griech. – pers.
Pomeranze	15.	Zitrusfrucht	2. Wortteil: it. – pers.
Poschti, Puschti		Teppich	pers.
Pul		afghanische Münze	pers.
Purim		Fest	hebr. – pers.
Pyjama	20.	Kleidungsstück	engl. – hindi – pers.
Ramasan		Ramadan	türk./ pers.

Regh		Maß für die Feinheit der Knüpfung	pers.
Reis	14.	Getreide	lat. – griech. – pers.? – ind. – semit?
Rhabarber	16.	Pflanze	It. – lat. – griech. – pers. ?
Rial		Währungseinheit	pers.; arab. – span.
Ribisel	15.	Johannisbeere	it., lat. – arab. – pers.
Roch, Rock		Riesenvogel im Märchen	arab. / pers.
Rochade	16.	Platztausch (Schach)	ndl. – frz. – span. – arab. – pers. – ind.
Saffian	17.	feines Ziegenleder	russ. – turksprachl. – pers.
Saki		Mundschenk (Dichtung)	arab. / pers.
Sandel(holz)	15.	Baumart	it.– arab. – pers. – ind. – drawida
Santalum		Sandelbaum	griech. – pers. – sanskr.
Saraband, Serabend		Teppich	pers.
Sarabanda, Sarabande		Tanz	frz./it. – span. – arab. – pers.
Sarafan		Überkleid	russ. – turkotatar. – pers.
Satrap		Statthalter	lat. – griech. – pers.
Schach	13.	Spiel	ndl. – frz. – arab. / pers. – ind.
Schah	20.	Kaiser von Persien	pers.
Schahbanu		Gemahlin des Schahs	pers.
Schah-in-Schah		König der Könige: Titel/ Träger des Titels	pers.
Schakal	17.	Tier	frz. – türk. – pers. – ind.
Schal	17.	Kleidungsstück	engl. – pers.
Schalwar		Frauenhose	pers./ türk.
Scharlach	12.	tiefes Rot, Infektionskrankheit	lat. – pers. – hebr.
Scheck	19.	Zahlungsmittel	engl. – ... – arab. – pers.?
Scheherazade/ Scheherezade		Märchenerzählerin	pers.
Schillum		Rohr zum Rauchen	engl. – hindi – pers.
Schiras		Teppich	pers.
Schorle(morle)	18.	Mischgetränk	slav. – pers.?
Seersucker		Baumwollgewebe	engl. – hindi – pers.
Sensal		Mahler	it. – arab. – pers.
Sepoy		Soldat	engl. – portug. – hindi – pers.
Serai		Wolltuch	pers.
Serail		Palast	frz. – it./ türk. – pers.
Sitar		Zupfinstrument	hindi – pers.
Softa		Student	türk. – pers.
Spahi		Krieger	frz. – türk. – pers.
Spinat	12.	Gemüse	span. – (lat.) – span. – arab. – pers.
Täbris		Teppich	pers.?
Taft	15.	Stoff	it. – pers.
Tambour	17.	Trommler	frz. – pers.

Tambur, Tanbur		Laute	arab. – pers?
Tamburin	13.	Schellentrommel	frz. – pers.
Tar		Laute	pers.
Tarbusch		arab. Bezeichnung für Fez	frz. – arab. – pers./ türk.
Tasse	16.	Gefäß	frz./ it. – arab. – pers.
Teppich	9.	Bodenbelag	lat. – griech. – pers?
Tiara	18.	Kopfbedeckung	lat. – griech. – pers.
Tiger	12.	Raubkatze	lat. – griech. – iran.
tigroid		tigerartig gestreift	griech. – pers.
Tinkal		Mineral	engl. – pers. – arab. – sanskr.
Toman		Rechnungseinheit	pers. – mongol.
Tschador/Tschadyr		langer Schleier	pers.
Turban	17.	Kopfbedeckung	it. – türk. – pers.
Turkbaff		Teppich	pers.
Wagireh		Probestück von Teppich	pers.
Zenana		Frauenbereich	hindi – pers.
Zendawesta		Awesta	pers.
Zerwanismus		Religion	pers.
Zitwer	11.	aromatisch duftendes Kraut	lat. – arab. – pers.
Zucker	13.	Süßmittel	it. – arab./ pers. – ind.
Zurna		Oboe	türk. – pers.

Es handelt sich um insgesamt 194 Wörter, teils in unterschiedlicher Schreibweise, von denen bei 68 das Jahrhundert der Entlehnung ins Deutsche angegeben werden kann.

2. Gesetzmäßigkeit der Entlehnungen

Wie alle derartigen Untersuchungen geht auch diese der Hypothese nach, dass die iranischen Entlehnungen ebenso wie andere Entlehnungen und sonstige Sprachwandel dem Piotrowski-Gesetz folgen (Altmann 1983, Altmann u.a. 1983; eine Vielzahl von Untersuchungen zu Entlehnungen und Sprachwandelprozessen sind auf der homepage des *Projekt Quantitative Linguistik*: http://wwwuser.gwdg.de/~kbest/ nachgewiesen):

$$(1) \qquad p = \frac{c}{1+ae^{-kt}}\,.$$

Um diese Annahme zu überprüfen, wird die Zahl der Wörter in Tabelle 1 nach Jahrhunderten aufgeschlüsselt in Tabelle 2 zusammengestellt; an die kumulierten Werte wird Formel (1) mit Hilfe der Software NLREG angepasst:

Tabelle 2
Zeitliche Entwicklung der iranischen Entlehnungen ins Deutsche

t	Jahrhundert	beobachtet	kumuliert	berechnet
1	8.	2	2	0.5622
2	9.	2	4	1.0264
3	10.	0	4	1.8647

4	11.	1	5	3.3566
5	12.	3	8	5.9463
6	13.	5	13	10.2507
7	14.	2	15	16.9154
8	15.	6	21	26.1802
9	16.	16	37	37.2996
10	17.	15	52	48.5151
11	18.	8	60	58.0087
12	19.	4	64	64.9237
13	20.	4	68	69.4284
a = 245.5773		b = 0.6083	c = 75.7018	D = 0.9877

Der Parameter *c* gibt an, gegen welchen Wert der Sprachwandel strebt; *t* steht für den jeweiligen Zeitabschnitt; *a* und *b* sind weitere Parameter. Der Determinationskoeffizient D = 0.9877, der höchstens den Wert 1.0000 erreichen kann, zeigt an, dass der beobachtete Prozess der Entlehnungen dem Piotrowski-Gesetz folgt. Abbildung 1 veranschaulicht dies, indem die beobachteten Werte (Punkte) kaum von den berechneten Werten (Linie) abweichen:

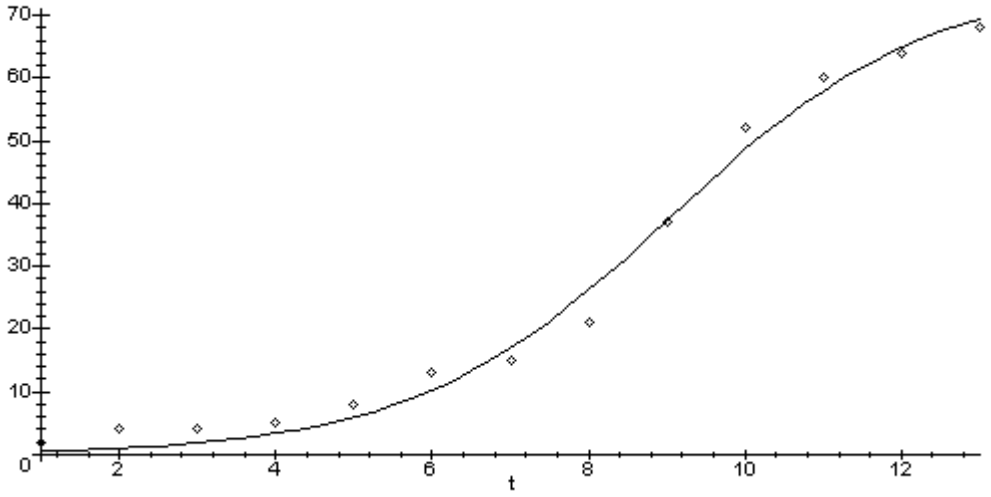


Abbildung 1. Zeitliche Entwicklung der iranischen Entlehnungen ins Deutsche

3. Zusammenfassung und Perspektive

Die Tabelle 1 zeigt, dass es im Deutschen eine große Anzahl von iranischen Entlehnungen gibt.

Die Hypothese, dass Sprachwandel beliebiger Art dem Piotrowski-Gesetz folgen, ließ sich erneut stützen.

Auch wenn sicher die eine oder andere Angabe sich als korrekturbedürftig erweisen wird, dürfte sich an dem Gesamtbild nicht allzuviel ändern, dass nämlich Entlehnungen aus dem Persischen über einen sehr langen Zeitraum erfolgt und ein erheblicher Teil der Entlehnungen in der Zeit vom 16. - 18. Jahrhundert ins Deutsche gelangt sind. Andere iranische Sprachen als mögliche Quellen von Entlehnungen konnten in dieser Untersuchung nicht berücksichtigt werden.

Literatur

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

Altmann, Gabriel, von Buttlar, H., Rott, W., & Strauß, U. (1983). A law of change in language. In: Brainerd, B. (ed.), *Historical linguistics* (S. 104-115). Bochum: Brockmeyer.

Best, Karl-Heinz (2001). Wo kommen die deutschen Fremdwörter her? *Göttinger Beiträge zur Sprachwissenschaft 5, 7-20.*

Duden. Das große Fremdwörterbuch. Herkunft und Bedeutung der Fremdwörter **(**^{4}2007). 4., aktualisierte Auflage. Mannheim/ Leipzig/ Wien/ Zürich: Dudenverlag.

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

Körner, Helle (2004). Zur Entwicklung des deutschen (Lehn-)Wortschatzes. *Glottometrics 7, 25-49.*

Pfeifer, Wolfgang [Ltg.] (^{2}1993/1995). *Etymologisches Wörterbuch des Deutschen.* München: dtv.

Ternes, Katarina (2011). Entwicklungen im deutschen Wortschatz. *Glottometrics 21, 25-53.*

Verwendete Software

MAPLE V Release 4. 1996. Berlin u.a.: Springer.

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

Glottometrics 26, 2013, 9 - 16

Sentence length and syntactic complexity in spoken and written English

Zhu Yujia[1]

Abstract: This study examines the sentence length and syntactic complexity in spoken and written English using the ICE corpus as the data source. The mean sentence length and the mean sentential syntactic complexity were computed and compared. The relationship between sentence length and the frequency of corresponding sentences, and that between the sentential syntactic complexity and the frequency of corresponding sentences can be described with the Wimmer-Altmann model.

Key words*: corpus, sentence length, sentential syntactic complexity*

1. Introduction

The nature and characteristics of speech and written language have been studied by many linguists. Bloomfield (1933) regarded writing as a recording system for speech, hence secondary to the latter. Chomsky (1965) considered writing as more representative of underlying grammar than speech, indicating that writing is more syntactically elaborate. Quirk (1985) described the difference between the two forms of language in terms of the medium through which they are transmitted, i.e., sound versus graphic; and as a result, spoken language employs more devices than written language, such as stress, rhythm, tempo etc, which are "impossible to represent with the relatively limited repertoire of conventional orthography (Quirk, 1985:26-27)." Biber et al (1999) noted that speech has a very high frequency of pronouns, deictic words, non-clausal or fragmentary components, and a low density of lexical words. Generally, these linguists approached the difference between the two forms of language mainly with the traditional methodology, with the notable exception of Biber, whose research in question was corpus based and data driven.

The present study focuses on sentence length (hereafter referred to as *SL*) and the sentential syntactic complexity (hereafter referred to as *SSC*) of spoken and written English within the theoretical frame and methodology of quantitative linguistics, intending to examine the relationship between *SL* and the frequency of sentences with corresponding *SL*; and *SSC* and the frequency of sentences with corresponding *SSC* in both spoken and written English, and searching for a formal model to describe such relationships mathematically.

The sentence is the largest stretch of language forming a syntactic construction (Biber et al. 1999), the highest-ranking unit of grammar (Quirk,1985), and an abstract grammatical element obtained from utterance in semantics (Saeed, 2000). Sentence length has attracted attention from quantitative linguists, who regard length of linguistic constructs as an important measurement revealing the inter-relationships among the different linguistic units of language. Köhler (1982), for instance, investigated the dependence of mean clause length on *SL*. Altmann (1988) concluded that *SL* depends on many different factors such as *SSC* and established a function of the probabilities of neighbouring length classes and the probability of the preceding length classes. Sigurd (2004) used a general function to describe the relation be-

[1] Address correspondence to: Zhu Yujia, School of Foreign Languages Dalian Maritime University. E-mail*: shamohaizhezhuyujia@126.com*

tween word length and frequency in English, Swedish and German. Some other researchers such as Best (2005), Fan et al (2010, 2012), Wang (2012) dug deeper into the sub-components of the sentence including words, phrases, morphemes and syllables. The present study differs from the above in that it studies the relationship between *SL* and the frequency of corresponding sentences; and that between *SSC* and the frequency of corresponding sentences in both spoken and written English. It is a quantitative contrastive study between the two forms of English, on which literature is few and far between. This study is corpus based, and the ICE-GB (the International Corpus of English, the British Component) was used as the data source.

2. Data and method

ICE-GB is made up of 500 2000-word text samples. Each sentence in ICE-GB is syntactically tagged. Table 1 displays the main syntactic labels of ICE-GB.

Table 1
Main syntactic labels of ICE-GB

A	adverbial	DTPE	predeterminer
ADJ	adjective	DTPO	determiner postmodifier
ADV	adverb	DTPR	determiner premodifier
AJHD	adjective phrase head	DTPS	postdeterminer
AJP	adjective phrase	ELE	element
AJPO	adjective phrase postmodifier	EMPTY	empty
AJPR	adjective phrase premodifier	EXOP	existential operator
ART	article	EXTHERE	existential there
AUX	auxiliary	FNPPO	floating NP postmodifier
AVB	auxiliary verb	FOC	focus
AVHD	adverb phrase head	FRM	formulaic expression
AVP	adverb phrase	GENF	genitive function
AVPO	adverb phrase postmodifier	GENM	genitive marker
AVPR	adverb phrase premodifier	IMPOP	imperative operator
CF	focus complement	INDET	indeterminate
CJ	conjoin	INTERJEC	interjection
CL	clause	INTOP	interrogative operator
CLEFTIT	cleft it	INVOP	inverted operator
CLOP	cleft operator	MVB	main verb
CO	object complement	N	noun
COAP	appositive connector	NADJ	nominal adjective
CONJUNC	conjunction	NONCL	nonclause
CONNEC	connective	NOOD	notional direct object
COOR	coordinator	NOSU	notional subject
CS	subject complement	NP	noun phrase
CT	transitive complement	NPHD	noun phrase head
DEFUNC	detached function	NPPO	noun phrase postmodifier
DISMK	discourse marker	NPPR	noun phrase premodifier
DISP	disparate	NUM	numeral
DT	determiner	OD	direct object
DTCE	central determiner	OI	indirect object
DTP	determiner phrase	OP	operator

P	prepositional	PUNC	punctuation
PARA	parataxis	PUNC	punctuation
PAUSE	pause	REACT	reaction signal
PAUSE	pause	SBHD	subordinator phrase head
PC	prepositional complement	SBMO	subordinator phrase modifier
PMOD	prepositional modifier	SU	subject
PP	prepositional phrase	SUB	subordinator
PREDEL	predicate element	SUBP	subordinator phrase
PREDGP	predicate group	TAGQ	tag question
PREP	preposition	TO	particle to
PROD	provisional direct object	TO	'to' infinitive marker
PROFM	proform	UNTAG	missing/unidentifiable items
PRON	pronoun	UNTAG	untag
PRSU	provisional subject	V	verb
PRTCL	particle	VB	verbal
PS	stranded preposition	VP	verb phrase
PU	parsing unit		

ICE-GB contains ICE-GB-S made up of 300 spoken texts totaling 697,985 words of speech and ICE-GB-W made up of 200 written texts totaling 424,194 words of written English. Both ICE-GB-S and ICE-GB-W were used as the data source for the present research.

The following is a syntactically parsed sentence with tags for syntactic functions and syntactic classes taken from file W1A-001.COR of ICE-GBS:

```
[<#22:1:B> <sent>]
PU,CL(main,cop,pres)
  SU,NP()
[<w>]
    NPHD,PRON(pers,sing) {I}
  VB,VP(cop,pres,semi,encl)
    OP,AUX(semi,pres,encl,disc1) {'m}
    A,AVP(excl)
[</w>]
      AVHD,ADV(excl) {just}
    OP,AUX(semi,pres,encl,disc1) {going to}
    MVB,V(cop,infin) {go}
  CS,AJP(prd)
    AJHD,ADJ(ge) {berserk}
  A,PP()
    P,PREP(ge) {for}
    PC,NP()
      DT,DTP()
        DTCE,ART(indef) {a}
      NPHD,N(com,sing) {whi`le}
```

The “clean” text of the above parsed sentence is as follows:

I’m just going to go berserk for a while.

SL was computed in number of words after the removal of the syntactic function and class labels; the *SL* of the above sentence is 10 (*I’m* is considered as two words). *SSC* was computed according to the number of immediate sentential syntactic functions such as *SU*,

subject; *VB*, verbal phrase etc. If there are 12 immediate syntactic function elements in a sentence, then its syntactic complexity is 12. In the above syntactically parsed sentence, there are four immediate syntactic function elements: *SU VB CS A PC*, therefore its *SSC* is four.

3. **Result and analyses**

3.1 Sentence length

The *SL* of ICE-GB-S and ICE-GB-W is displayed respectively in Table 2 and Table 3. There are altogether 59,516 sentences in ICE-GB-S. The mean *SL* in ICE-GB-S is 11.73. The range of *SL* in ICE-GB-S is from 1 to 121. However, short sentences whose length is less than 16 accounts for approximately 74.26% of the total. Sentences with length longer than 46 words only constitute 1.87% of the total sentences in ICE-GB-S.

Table 2
The distribution of sentence length in ICE-GB-S

SL	Number	SL	Number	SL	Number	SL	Number
1	4518	27	481	53	65	79	3
2	3669	28	445	54	42	80	8
3	5967	29	448	55	35	81	4
4	3897	30	396	56	44	82	6
5	3526	31	351	57	51	83	4
6	3415	32	382	58	31	84	7
7	3091	33	268	59	34	85	1
8	2856	34	282	60	25	86	3
9	2532	35	252	61	22	87	3
10	2365	36	243	62	15	88	3
11	2070	37	224	63	16	89	3
12	1822	38	182	64	20	90	3
13	1638	39	160	65	25	91	1
14	1470	40	149	66	16	93	1
15	1360	41	162	67	12	94	1
16	1166	42	121	68	17	95	3
17	1188	43	124	69	10	96	1
18	991	44	135	70	11	98	3
19	956	45	113	72	15	99	1
20	906	46	95	72	10	100	2
21	827	47	72	73	7	101	1
22	738	48	81	74	4	103	2
23	665	49	74	75	7	105	3
24	692	50	56	76	9	107	1
25	604	51	56	77	7	110	1
26	557	52	49	78	9	121	1

There are 23,937 sentences in ICE-GB-W. The mean *SL* is 17.72, much larger than that of ICE-GB-S, which is only 11.73. The percentage of short sentences whose length is shorter than 16 words in ICE-GB-W is 48.27%, much smaller than that of ICE-GB-S, which is 74.26%. All the statistics above indicate that sentences in written English are generally longer

than those in spoken English. Sentences with length longer than 46 words account for 2.87%, larger than that of of ICE-GB-S, which is 1.87%. However, unlike the long sentences in ICE-GB-W, those in ICE-GB-S are mainly caused by inserts, interjections, rephrasing, etc, which reflects the wordy and impreciseness of spoken language compared with written language.

Table 3
The distribution of sentence length in ICE-GB-W

SL	Number	SL	Number	SL	Number	SL	Number
1	772	24	601	47	64	70	2
2	890	25	527	48	51	71	3
3	884	26	497	49	31	72	4
4	759	27	470	50	54	73	4
5	732	28	446	51	40	74	1
6	695	29	420	52	39	75	1
7	746	30	333	53	43	76	4
8	770	31	336	54	35	77	3
9	775	32	322	55	30	78	4
10	760	33	263	56	29	79	3
11	782	34	255	57	17	81	1
12	739	35	221	58	12	84	2
13	772	36	221	59	18	86	1
14	743	37	185	60	17	88	1
15	736	38	182	61	16	89	1
16	805	39	145	62	11	91	1
17	787	40	133	63	12	94	1
18	751	41	104	64	13	98	1
19	691	42	105	65	6	99	1
20	705	43	105	66	6	101	1
21	666	44	93	67	9		
22	649	45	67	68	2		
23	609	46	87	69	6		

There is a certain relationship between *SL* and the frequency of sentences of corresponding length in both ICE-GB-S and ICE-GB-W. The number of sentences generally decreases along with the increase of the sentence length. The following equation was used by Nemcová and Serdelová (2005) to describe the relationship between the number of synonyms (*y*) of a word and the length of the word in syllables *x*:

$$y = ax^{b}e^{cx} + 1 \quad (1)$$

(1) is a special case of Wimmer & Altmann (2005). It also works in the relationship between *SL* and the frequency of sentences *NS* of corresponding length:

$$NS = a(SL)^{b}e^{c(SL)} + 1 \quad (2)$$

Figure 1 displays the model fit for both ICE-GB-S and ICE-GB-W. The left panel shows the relationship between *SL* and the frequency of sentences of corresponding length in ICE-GB-S and the right the relationship in ICE-GB-W. (2) yields a very good fit both for ICE-GB-S and

ICE-GB-W, with R^2 = 0.970, a = 5112.624, b = 0.156 and c = -0.112 for the former; and R^2 = 0.948, a = 0.438, b = 614.166 and c = -0.069 for the latter.

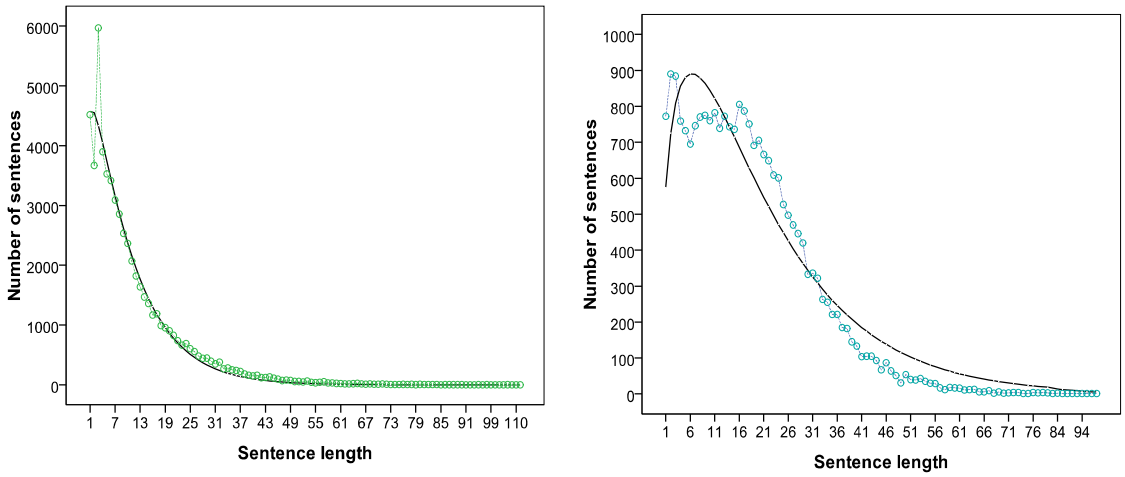


Figure 1. The relationship between *SL* and the frequency of sentences of corresponding length. Solid line: model fit, small circles: the observed value. Left panel: ICE-GB-S; right panel: ICE-GB-W

3.2. Sentential syntactic complexity

The *SSC* of both ICE-GB-S and ICE-GB-W is displayed in Table 4 and Table 5 respectively.

The range of *SSC* in ICE-GB-S is from 1 to 25. Sentences with *SSC* of one account for 20.24% of the total in ICE-GB-S. Sentences with *SSC* less than six constitute 85.15% of the total sentences in ICE-GB-S. Moreover, there are only five sentences with *SSC* of more than 20.

The mean *SSC* was calculated by adding up the *SSC* of each of the sentences totaling 212,472 and then divided by the total numbers of sentences in ICE-GB-S, which is 59,516 and the result is 3.57.

Table 4
The distribution of syntactic structural complexity in ICE-GB-S

SSC	Number	SSC	Number
1	12019	13	47
2	5839	14	32
3	13309	15	16
4	11552	16	12
5	7855	17	6
6	4160	18	7
7	2206	19	5
8	1150	21	1
9	603	22	1
10	312	23	1
11	157	24	1
12	101	25	1

The range of *SSC* in ICE-GB-W is from 1 to 24. Sentences with *SSC* of one account for only 10.63% of the total in ICE-GB-W, much smaller than that in ICE-GB-S, which is 20.24%. However, sentences with *SSC* less than six constitute about 84.80% of the total sentences in ICE-GB-W, which is similar to that in ICE-GB-S. The mean *SSC* in ICE-GB-W is 3.83, obtained from dividing the cumulative *SSC* of all the sentences in ICE-GB-W, 91,678, with the total number of sentences, 23,937.

Table 5
The distribution of syntactic structural complexity in ICE-GB-W

SSC	Number	SSC	Number
1	2545	11	34
2	1738	12	15
3	7060	13	8
4	5449	14	4
5	3505	15	3
6	1835	18	1
7	1009	19	1
8	437	20	1
9	202	22	1
10	86	24	1

Superficially, the mean *SSC* of ICE-GB-W does not seem much different from that of ICE-GB-S. However, a chi-square test using the cumulative *SSC* and the total number of sentences of ICE-GB-S and ICE-GB-W as the test data reveals that the difference in *SSC* of the two components of ICE-GB is significant, with a chi-square value of 66.612 and a *P* value of 0.000, indicating that the *SSC* of ICE-GB-W is significantly more complex than that of ICE-GB-S.

The relationship between the *SSC* and the frequency of sentences *NS* of corresponding *SSC* in both ICE-GB-S and ICE-GB-W can be also captured with (2), replacing sentence length *SL* with *SSC* we obtain (3):

$$NS = a(SSC)^{b} e^{c(SSC)} + 1 \quad (3)$$

Figure 2 displays the model fit for both ICE-GB-S and ICE-GB-W. The fit is good. For ICE-GB-S, R^2 = 0.896, a = 1226.855, b = 7.274 and c = -2.124; and for ICE-GB-W, R^2 = 0.854, a = 15740.343, b = 1.092 and c = -0.523.

4. Conclusion

This study examines *SL* and *SSC* and the frequency of the corresponding sentences between spoken and written English. The mean *SL* in written English is much larger than that in spoken English. On the other hand, the mean *SSC* in spoken English is significantly smaller than that in written English. The relationship between *SL* and the frequency of corresponding sentences and that between *SSC* and the frequency of corresponding sentences can be captured with the Wimmer-Altmann model.

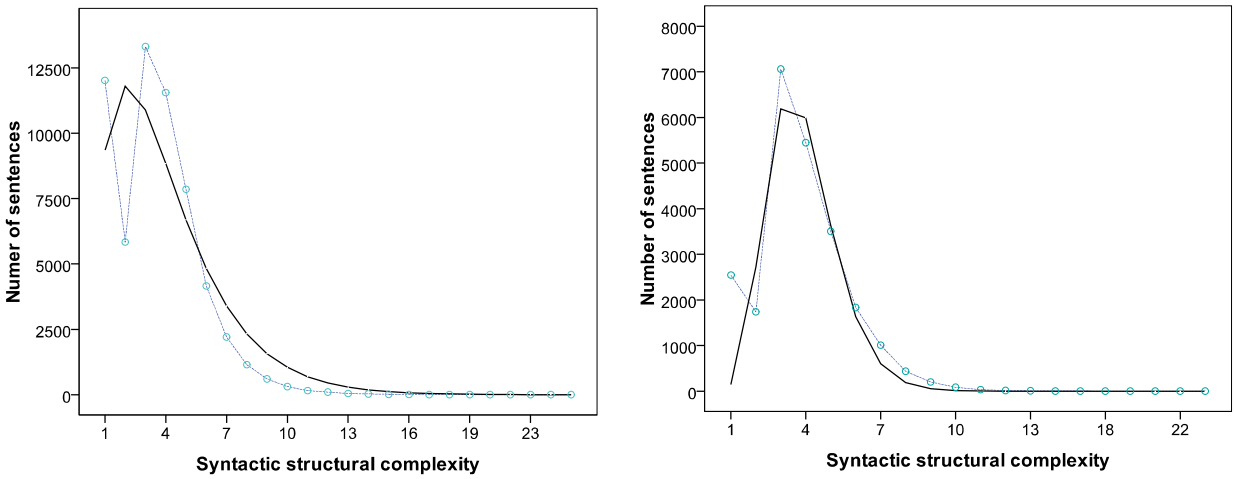


Figure 2. The relationship between *SSC* and the frequency of sentences of corresponding *SSC*. Solid line: model fit, small circles: the observed value. Left panel: ICE-GB-S; right panel: ICE-GB-W

References

Altmann, G. (1988). *Verteilungen der Satzlängen.* In: Schulz, K.-P. (ed.), *Glottometrika 9:147-169*. Bochum: Brockmeyer.

Best, K.-H. (2005). Satzläge. In: Köhler, R; Altmann, G; Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook.* Berlin, New York: de Gruyter, 298–304.

Biber, D. (1988). *Variation across Speech and Writing.* Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education Limited.

Bloomfield, L. (1933). *Language*. New York: Holt.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Fan, F., Grzybek, P., Altmann, G.. (2010). Dynamics of word length in sentence. *Glottometrics 20, 70-109.*

Köhler R. (1982). *Das Menzerathsche Gesetz auf Satzebene.* In: Lehfeldt, Werner; Strauss, Udo (eds.), *Glottometrika 4, 103–113.* Bochum: Brockmeyer,

Nemcová, E., Serdelová, K. (2005). On synonymy of Slovak. In: Altmann, G., Levickij, V. & Perebyinis, V. (eds.), *Problems of Quantitative Linguistics*: *194-209*. Chernivtsi: Ruta.

Quirk, R. (1985). *Comprehensive Grammar of the English Language.* London: Longman.

Saeed, J. (2000). *Semantics.* Beijing: Foreign Language Teaching and Research Press.

Sigurd, B. (2004). *Word Length, Sentence Length and Frequency – Zipf Revisited.* In: *Studia Linguistica 58(1), 37–52.* Blackwell Publishing Ltd.

Wang, H. (2012). Length and complexity of NPs in Written English. *Glottometrics 24, 79-87.*

Wimmer, G. & Altmann, G. (2005). Towards a unified derivation of some linguistic laws. In: Grzybek, P. (ed.). *Contributions to the science of language: Word length and related issues*: *93-117*. Boston: Kluver.

Glottometrics 26, 2013, 17-26

Kant‘s Terminology of Cognitive Capacities
A Quantitative Study on Lexicographic Polysemy in the “Critique of Pure Reason”

Thorsten Roelcke
Gabriel Altmann

Abstract. A relation between frequency and polysemy of terms for cognitive capacities in Kant’s “Critique of Pure Reason” is tested. The observed relation follows the Zipf-Alekseev function and is important in respect to the discussion of polysemy of terminology.

Keywords**:** *concept, Kant, frequency, polysemy, terminology*

1. Preliminary remarks

The “Critique of Pure Reason”, written in the 18th century by Immanuel Kant (“Kritik der reinen Vernunft”, 2nd ed. 1787/1968), is one of the most famous works in philosophy. The terminology of Kant has been taken as a prototype for scientific language of enlightenment (Roelcke 2005), and during the last centuries many dictionaries of the “Critique of Pure Reason” and Kant’s works in general have been published (Roelcke 1999; 2002). An important question of terminology research concerns the quantity of term’s meanings: From a normative point of view terms have to be monosemous (bi-uniqueness of form and meaning); in reality they are (and have to be) very polysemous. Kant’s terminology has been taken for a good example of a clear, distinct and last but not least monosemous one. In contrast to this conjecture a little dictionary (Roelcke 1989) shows that Kant’s terminology is a polysemous one: So we have to analyse the (natural or philological) regularities of the non-monosemous ideal (prototype) of philosophical terminology.

2. Terms and meanings in Kant’s “Critique of Pure Reason”

In Kant’s “Critique of Pure Reason” 52 terminus types denote each minimally one cognitive capacity; they appear as 2,654 terminus tokens in the philosophical text. Much of these terms show various single meanings, i. e. they are polysemous by philological interpretation (e. g. *Vernunft* has 23 different single meanings). As Figure 1 shows these meanings may be classified as follows: 1st transcendental (M1), 2nd (other) philosophical meanings (M2), 3rd (other) scientific (M3), and 4th other (non-scientific) meanings (M4).

In Table 1 we find a) the number of single meanings of each single term and meaning class (M1–4), b) the number of single meanings in total (MT), and c) the quantity of term tokens in total (TT); the number of term tokens of each single meaning has not been analyzed.

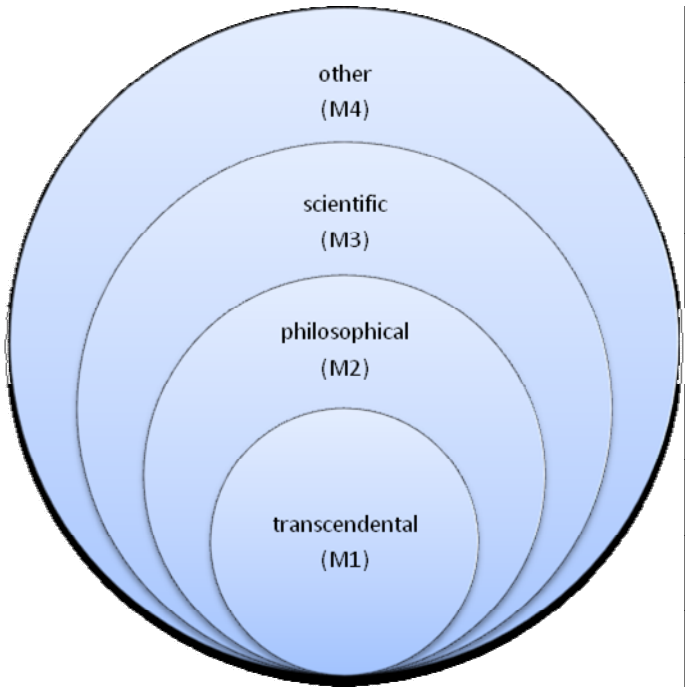


Figure 1
Meaning classes in the "Critique of Pure Reason" (Roelcke 1989).

Table 1
Terms and meanings (cf. Martin 1967; Roelcke 1989)

M1	=	class 1 – transcendental meanings
M2	=	class 2 – philosophical meanings
M3	=	class 3 – scientific meanings
M4	=	class 4 – non-scientific meanings
MT	=	meanings in total
TT	=	term tokens in total

The numbers on the right show the quantity of single meanings considering the four meaning classes (M1–M4), the quantity of single meanings in total (MT), and the quantity of term tokens in total (TT).

	Term	**M1**	**M2**	**M3**	**M4**	**MT**	**TT**
1	*Anschauungsvermögen*	1	0	0	0	1	4
2	*Denkungsvermögen*	0	1	0	0	1	1
3	*Dummheit*	0	0	0	1	1	1
4	*Einbildungskraft*	6	2	2	0	10	52
5	*Empfänglichkeit*	0	0	0	2	2	2
6	*Erkenntnißfähigkeit*	1	0	0	0	1	1
7	*Erkenntnißkraft*	1	1	1	0	3	20
8	*Erkenntnißquell*	4	0	0	1	5	2
9	*Erkenntnißvermögen*	2	0	0	0	2	20
10	*Fähigkeit*	1	0	0	1	2	2
11	*Fassungskraft*	0	0	0	1	1	1
12	*Gedächtniß*	0	0	0	1	1	1
13	*Gefühl*	0	0	0	3	3	5
14	*Gehirn*	0	0	0	1	1	2
15	*Geist*	0	0	0	2	2	4
16	*Geisteskraft*	0	0	0	1	1	1
17	*Gemüth*	3	0	0	0	3	45
18	*Gemüthskraft*	1	0	0	0	1	2

19	*Geschmack*	0	1	0	2	3	4
20	*Grundkraft*	1	0	0	0	1	11
21	*Grundquelle*	1	0	0	0	1	1
22	*Intelligenz*	1	1	0	0	2	28
23	*Klugheit*	0	0	0	1	1	3
24	*Kraft*	1	0	1	3	5	59
25	*Menschensinn*	0	0	0	2	2	2
26	*Menschenvernunft*	0	0	0	3	3	12
27	*Menschenverstand*	0	0	0	1	1	3
28	*Mutterwitz*	0	0	0	1	1	1
29	*Organ*	1	1	1	0	3	6
30	*Quell*	1	0	0	1	2	39
31	*Ratio*	0	1	0	0	1	10
32	*Receptivität*	3	0	0	0	3	18
33	*Scharfsinnigkeit*	0	0	0	2	2	3
34	*Sinn*	4	2	1	4	11	221
35	*äußerer Sinn*	2	0	0	0	2	20
36	*innerer Sinn*	2	0	0	0	2	68
37	*Sinnlichkeit*	3	2	0	0	5	167
38	*Spontaneität*	3	1	0	0	4	24
39	*Unterscheidungskraft*	0	0	0	1	1	2
40	*Unterscheidungsvermögen*	0	0	0	1	1	1
41	*Unvermögen*	0	0	0	1	1	8
42	*Urtheilskraft*	4	0	0	2	6	36
43	*Verbindungsvermögen*	1	0	0	0	1	1
44	*Vermögen*	1	0	0	3	4	106
45	*Vernunft*	12	9	0	2	23	1,047
46	*Vernunftvermögen*	2	2	0	1	5	7
47	*Verstand*	7	4	0	8	19	560
48	*Verstandesvermögen*	1	1	0	0	2	3
49	*Vorstellungsfähigkeit*	2	0	0	0	2	3
50	*Vorstellungskraft*	2	2	0	0	4	8
51	*Vorstellungsvermögen*	1	0	0	0	1	1
52	*Witz*	0	0	1	1	2	5
	Total:	**76**	**31**	**7**	**54**	**168**	**2,654**

The following hypotheses have to be tested:

(H1) The number of single meanings of term tokens in total (MT) abides by a certain rank-frequency distribution and displays a regular frequency spectrum.

(H2) The numbers of single meanings of term tokens in the four meaning classes (M1–M4) abide by a certain distribution (the numbers are too small for testing).

(H3) The number of term tokens in total (TT) abides by a certain frequency distribution.

(H4) The rank-frequency distribution of term tokens in total (TT) correlates with the rank-frequency distribution of meanings in total (MT).

Polysemy is the result of philological interpretation. Hence the question arises, whether the result of such an interpretation follows a certain function (H1) – and shows a somewhat "natural" distribution of linguistic entities (H3; H4). The distribution in the meaning types themselves (H2) seems to be secondary (the quantities are too small for testing).

3. Quantitative Analysis

The rise of polysemy is a kind of diversification process present both in formal and in semantic domains. If we consider a closed group of words, e.g. nouns in a dictionary, or a closed group of terms in a text, we can see that the number of meanings of individual terms is not uniformly distributed but follows a certain scheme which is well known in quantitative linguistics. In order to set up a model of this scheme, we start from the conjecture that the relative rate of change of the number of meanings in the individual word is proportional to the rate of change of its ranking in the given group, whereby we consider the rank as a continuous variable. This technique is analogous to the discrete approach but in this case we even know that the meaning of a word is rather a multidimensional continuous variable simplified for modelling purposes to ranks.

We can write our conjecture in the form

$$\frac{dy}{y-1} = \frac{s(x)}{h(x)}dx \tag{1}$$

where $s(x)$ is a function representing the diversification force of the speaker/author, and $h(x)$ is the unification force coming from the hearer who must care for communicative efficiency of words. An over- rich polysemy would be detrimental for the communication. Besides, we use here y-1 in the denominator on the left hand side because no word can have fewer than one meaning (except for proper names).

Now, the diversification force consists of a language- or communication constant, say a, and the direct force of the speaker influencing not directly the place of the word in the given system but merely the logarithm of its position. Hence one can define $s(x) = a + b \ln x$. The force of the hearer is simply a braking and controlling force defined as $h(x) = cx$. Inserting these conjectures in the above formula we obtain

$$\frac{dy}{y-1} = \frac{a + b\log x}{cx}dx \tag{2}$$

Solving this differential equation and setting a/c = B, $b/(2c) = C$, we obtain the function

$$y = 1 + Ax^{B+C\log x} \tag{3}$$

representing the Zipf-Alekseev function used for different purposes in linguistics, since different phenomena are constructed in this way. It can be remarked that not only the distribution of meanings to individual words of a group (column MT in Table 1) but also the frequency of these words in a text (column TT) abide by this regularity. If we apply formula (3) to our data, we obtain the fitting results as presented in Table 2.

Evidently, one can model this phenomenon also using a discrete probability distribution but it has some disadvantages: it is not better than the above approach in which the x

values may be considered discrete although there are no problems with the goodness-of-fit test. In a discrete probability distribution there are many frequencies whose values are zero but which must be taken into account; and some expected values are smaller than 1, a circumstance which is not realistic. The pooling of theoretical values is rather a subjective decision. However, using formula (3) we have all expected values > 1, and the determination coefficient is a sufficient sign of adequateness. Of course, the F-test and the t-tests for the parameters yield excellent results but here they are redundant.

Table 2

Fitting the Zipf-Alekseev function to the ranked sequence of MT and TT (Table 1)

Rank	MT	Formula (3)	TT	Formula (3)
1	23	23.65	1047	1052.90
2	19	16.41	560	515.65
3	11	12.23	221	283.32
4	10	9.61	167	170.93
5	6	7.85	106	110.40
6	5	6.60	68	75.06
7	5	5.67	59	53.14
8	5	4.96	52	38.87
9	5	4.40	45	29.22
10	4	3.95	39	22.48
11	4	3.59	36	17.65
12	4	3.29	28	14.11
13	3	3.04	24	11.47
14	3	2.83	20	9.46
15	3	2.65	20	7.91
16	3	2.49	20	6.70
17	3	2.36	18	5.74
18	3	2.24	12	4.97
19	3	2.14	11	4.35
20	2	2.05	10	3.85
21	2	1.97	8	3.43
22	2	1.89	8	3.09
23	2	1.83	7	2.81
24	2	1.77	6	2.57
25	2	1.72	5	2.37
26	2	1.67	5	2.20
27	2	1.63	4	2.05
28	2	1.59	4	1.93
29	2	1.55	4	1.82
30	2	1.52	3	1.73
31	2	1.49	3	1.65
32	2	1.46	3	1.58
33	1	1.44	3	1.52
34	1	1.41	3	1.46
35	1	1.39	2	1.42
36	1	1.37	2	1.38
37	1	1.35	2	1.34

38	1	1.34	2	1.31
39	1	1.32	2	1.28
40	1	1.30	2	1.25
41	1	1.29	2	1.23
42	1	1.28	1	1.21
43	1	1.26	1	1.19
44	1	1.25	1	1.18
45	1	1.24	1	1.17
46	1	1.23	1	1.15
47	1	1.22	1	1.14
48	1	1.21	1	1.13
49	1	1.20	1	1.12
50	1	1.20	1	1.11
51	1	1.19	1	1.10
52	1	1.18	1	1.09
	A = 22.6484 B = –0.4140 C = – 0.2044 R^2 = 0.98		A = 1051.8993 B = –0.7477 C = –0.4092 R^2 = 0.99	

Getting the spectrum of these distributions can be performed either theoretically (cf. Zörnig, Boroda 1992; Wimmer et al. 2003: 119 ff.) or practically by simply counting the number of ones, twos, etc. We simply state the spectra by evaluating the second and the fourth column in Table 2 and obtain the results in Table 3.

One can obtain different distributions for MT but all suffer from the fact that the majority of classes must be pooled and the number of degrees of freedom will be strongly reduced. In Table 3 we pooled all classes so that $NP_x > 5$. Much simpler is to model the spectrum as a function because here classes containing $f_x = 0$ can simply be eliminated. We simply use the power function (Zipf's law) with added 1 because we eliminated all zeroes and obtain the results in the last three columns of Table 3. Ignoring the added 1 the result would be slightly better, but theoretically worse because the last two theoretical values would be smaller than 1.

For testing Hypothesis H4, telling that the more meanings a word has, the more frequently it occurs in the text, a chi-square test for uniformity would be correct but unfortunately we have such a number of small frequencies that a chi-square would not be quite adequate. For this reason we simply compute the correlation of the two samples. Using standard formulae we obtain for the columns MT and TT the correlation coefficient $r = 0.91$ which for 50 degrees of freedom yields $t = 15.52$, a highly significant value. In order to find a function representing the increase of TT with increasing MT we may proceed in three ways: (1) reorder the values of MT and TT in Table 2 according to increasing MT; (2) for each x (from MT) take the means of frequencies y in TT; (3) to start from the frequencies and observe their influence to polysemy, i.e. to invert the dependence and use both approaches (1) and (2). Here we shall restrict ourselves to one direction.

Table 3
Frequency spectra of MT and TT

x	f_x	Geometric	x	f_x	$1 + ax^b$
1	20	19.4453	1	20	20.97
2	13	12.1738	2	13	9.59
3	7	7.6214	3	7	6.24
4	3	4.7714	4	3	4.70
5	4	2.9871	5	4	3.82
6	1	1.8701	6	1	3.26
7	0	1.1708	10	1	2.21
8	0	0.7330	11	1	2.08
9	0	0.4589	19	1	1.55
10	1	0.2873	23	1	1.44
11	1	0.1799			
12	0	0.1126			
13	0	0.0705			
14	0	0.0441			
15	0	0.0276			
16	0	0.0173			
17	0	0.0108			
18	0	0.0068			
19	1	0.0042			
20	0	0.0027			
21	0	0.0017			
22	0	0.0010			
23	1	0.0017			
	p = 0.3739; X^2 = 0.20; DF = 3, P = 0.98			a = 19.9659; b = –1.2169; R^2 = 0.94	

Again, we can use the above approach and for (1) consider the relative rate of change of mean frequency (y) as proportional to the relative rate of change of x. Here we obtain a simpler expression, namely

$$\frac{dy}{y-1} = \frac{b}{x}dx. \tag{4}$$

Solving this equation we obtain

$$y = 1 + ax^b. \tag{5}$$

We used again y-1 because frequencies cannot be smaller than 1. Fitting 5 to the above data, we obtain the function

$$y = 1 + 0.3911x^{2.5057}$$

yielding a determination coefficient $R^2 = 0.96$. Though the *y*-values have a great dispersion for a given *x*, the function displays a very good fitting.

In order to reduce the variance, we compute for each *x* (MT) the mean of all respective *y* (in TT) and obtain the course as presented in Table 4 (first two columns).

Table 4
The relationship of polysemy (*x*) to mean frequencies ($\bar{y}$)

x	$\bar{y}$	Formula (5)
1	2.80	1.20
2	15.31	2.30
3	15.71	4.93
4	46.00	9.62
5	58.75	16.86
6	36.00	27.09
10	52.00	106.28
11	221.00	137.58
19	560.00	608.65
23	1047.00	1024.92
a = 0.1955, b = 2.7311, $R^2 = 0.98$		

As can be seen (cf. fig. 2)[1], the dependence gets slightly stronger but the difference between the two determination coefficients is not relevant. In any case, it shows the interrelation between MT and TT, known from quantitative linguistics as the relationship between polysemy and frequency (cf. Köhler 1986, 2005).

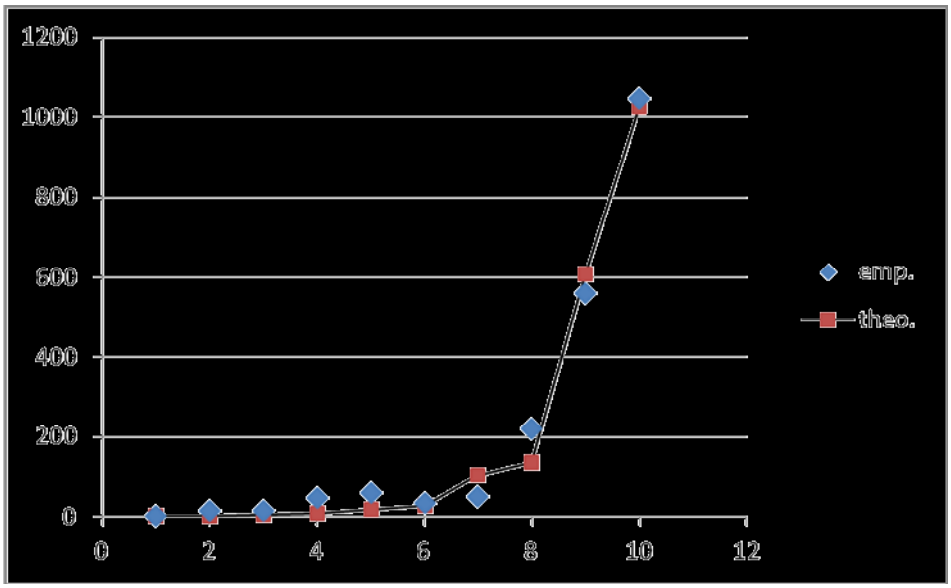


Figure 2
Polysemy and frequency of term in Kant's "Critique of Pure Reason"

[1] Thanks to Emmerich Kelih (Vienna) for drawing Figure 2.

4. Discussion

To return to the hypotheses above the conclusion is as follows: The numbers of term tokens in total (TT) as well as the numbers of single meanings in total (MT) follows a certain scheme, representing the Zipf-Alekseev function; both the relative rate of change of the frequency of terms and the relative rate of change of the number of their meanings are proportional to the rate of change of their ranking in the given group. Hence we can corroborate H1 and H3. In addition to this we note a certain relationship between polysemy and frequency with the result that H4 is corroborated, too. As the quantities of meanings of term tokens in the four meaning classes are too small for testing, we cannot accept or reject H2. As mentioned, the correlation between frequency and polysemy is well-known in quantitative linguistics. For at least three reasons the example of meanings in Kant's "Critique of Pure Reason" is interesting yet:

1. The (normative) ideal of bi-uniqueness of form and meaning in languages for specific purposes (LSP) has to be rejected, and we have to consider whether this ideal is someway "unnatural" for languages – including LSP. In our opinion this statement has or should have serious consequences for terminology work (e.g. that of ISO or DIN).
2. The terminology of Kant, traditionally taken as a prototype for clarity, distinctness and monosemy in scientific language of enlightenment, follows linguistic laws as well as other scientific languages. In respect to this the language of enlightenment is less culturally bound or influenced as often assumed.
3. Even the meaning of a word in total is rather a continuous variable simplified for modelling, especially for lexicographical purposes, the approach of several single meanings is an important method of philological interpretation. Obviously such human interpretation of language follows the same law as human language itself.

Kant's Terminology of Cognitive Capacities in the "Critique of Pure Reason" is a small example for lexicographic polysemy. But in respect to further studies concerning polysemy of terminology it is an inspiring one.

References

Kant, Immanuel (1787/1968). *Kritik der reinen Vernunft.* 2. Aufl. Akademie-Textausgabe. Unveränderter photomechanischer Abdruck des Textes der von der Preußischen Akademie der Wissenschaften 1902 begonnenen Ausgabe von Kants gesammelten Schriften. Bd. III. Berlin: de Gruyter.

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

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

Martin, Gottfried (Hrsg.) (1967*). Sachindex zu Kants Kritik der reinen Vernunft*. Bearbeitet von Dieter-Jürgen Löwisch. Berlin: de Gruyter.

Roelcke, Thorsten (1989). *Die Terminologie der Erkenntnisvermögen. Wörterbuch und lexikosemantische Untersuchung zu Kants „Kritik der reinen Vernunft"*. Tübingen: Niemeyer (Reihe Germanistische Linguistik 95).

Roelcke, Thorsten (1999). Die deutschsprachige Fachlexikographie der Philosophie in ihrem europäischen Umfeld: eine Übersicht. In: *Fachsprachen. Languages for Special Purposes. Ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft. An International Handbook of Special-Language and Terminology Re-*

search: 1995-2004. Hrsg. von Lothar Hoffmann, Hartwig Kalverkämper und Herbert Ernst Wiegand in Verbindung mit Christian Galinski und Werner Hüllen. Bd. 2. Berlin, New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft 14.2),

Roelcke, Thorsten (2002). Wörterbücher der Philosophie im Spannungsverhältnis zwischen philosophischem Diskurs und lexikographischer Struktur. *Lexicographica 18, 65-88.*

Roelcke, Thorsten (2005). Immanuel Kant. In: *Lexikologie. Lexicology. Ein internationales Handbuch zur Natur und Struktur von Wörtern und Wortschätzen. An international handbook on the nature and structure of words and vocabularies: 1532-1537.* Hrsg. von / Ed. by D. Alan Cruse, Franz Hundsnurscher, Michael Job, Peter Rolf Lutzeier. Bd. 2. Berlin-New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft),

Wimmer, G., Altmann, G., Hřebíček, L., Ondrejovič, S., Wimmerová., S. (2003*). Úvod do analýzy textov.* Bratislava: Veda.

Zörnig, Peter; Boroda, Moisei (1992). The Zipf-Mandelbrot law and the interdependencies between frequency structure and frequency distribution in coherent text. *Glottometrika 13, 205-218.*

Glottometrics 26, 2013, 27-37

Syntactic Complexity and Position in Hungarian

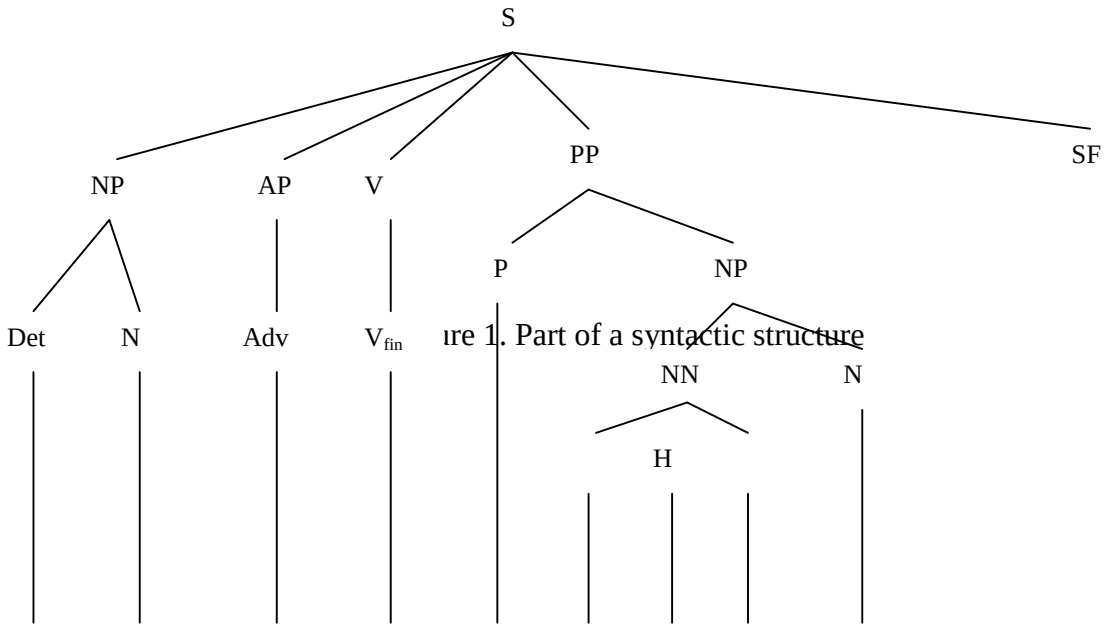
Reinhard Köhler, Sven Naumann, Trier

Abstract. Two properties of syntactic constructions are studied with respect to their frequency distributions in individual texts and in sub-corpora of the Hungarian "Szeged Treebank 2.0". Mathematical models of these distributions as presented in (Köhler/Altmann 2000) are tested on these data. It is found that the tests confirm the hypotheses in principle although some differences between Hungarian and previously tested languages were found. They might be explainable by differences in discourse organisation. Furthermore, differences in the distributions between corpus parts and between individual texts from different genres may indicate text type sensitivity.

Keywords: s*yntactic* ***c****omplexity, position in constituent, Hungarian, text type, discourse organisation, hyper-Pascal distribution, binomial distribution, negative hypergeometric distribution.*

1. Complexity

Complexity of syntactic constructions was defined in (Köhler 1999) as the number of immediate constituents. Fig. 1 illustrates how this property was operationalised: The top level node S is assigned complexity 5, the three constituents with the labels NP, PP, and NP have complexity 2, AP, V, and N complexity 1, and NN has 3.



ıre 1. Part of a syntactic structure

The first studies on frequency distribution of complexity were published in (Köhler/Altmann 2000). Theoretical considerations on the background of the synergetic-linguistic research framework yielded the hyper-Pascal distribution (cf. Wimmer/Altmann 1999) as a promising model of complexity distributions in texts:

$$P_x = \frac{\binom{k+x-1}{x}}{\binom{m+x-1}{x}} q^x P_0, \quad x = 0,1,2,...$$

with $P_0^{-1} = {}_2F_1(k,1;m;q)$ – the hypergeometric function – as normalising constant.

Empirical data from the English Susanne corpus and the German Negra corpus resulted in quite satisfying goodness-of-fit statistics ($C \approx 0.01$) and thus supported the hypothesis represented by the model. The present contribution reports on an analogous study on data from the Szeged Treebank 2.0, a syntactically analysed and annotated collection of Hungarian texts (http://www.inf.u-szeged.hu/projectdirs/hlt/; English version: http://www.inf.u-szeged. hu/ projectdirs/hlt/index_en.html). The Treebank was analysed on the basis of the same grammar type as the above-mentioned corpora, viz. a phrase structure grammar, which makes the three comparable. The Szeged Treebank differentiates the six text source types "business-news", "compositions" (a collection of essays written by pupils differentiated into age categories), "computer" (news in the field of computer technology), fiction, law, and newspapers. Individual studies were conducted on these six corpus parts.

The Hungarian data display an inhomogeneous situation: Of the 13 data sets in our study, only 4 give a good result with the hyper-Pascal distribution ($C < 0.008$); all these four data sets can be characterized as news texts. An example of one of the distributions with a very good fitting result is shown in Table 1 and Fig. 2.

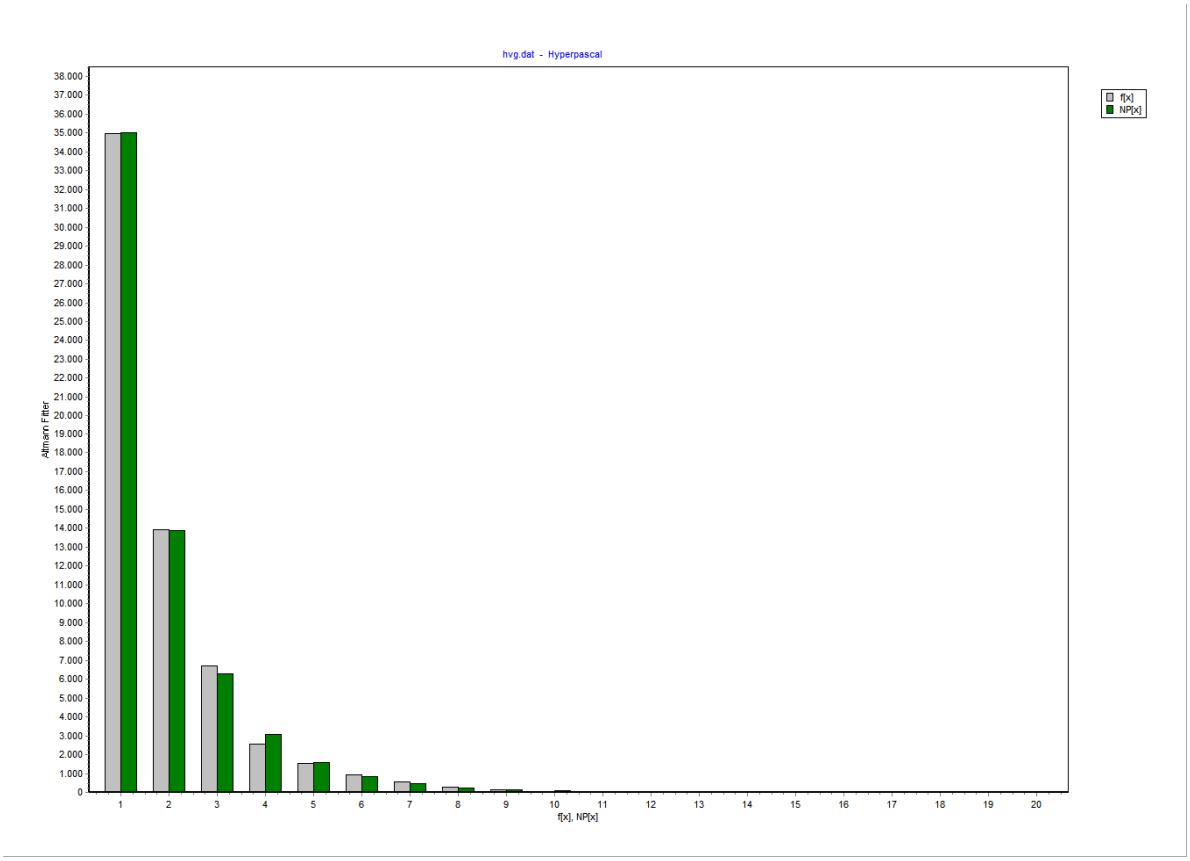


Figure 2. Graph of the results presented in Table 1

Table 1
Results of fitting the hyper-Pascal distribution to the data
from one of the news text sections in the Szeged Treebank 2.0

x[i]	**f[i]**	**NP[i]**
1	34984	34999.61
2	13916	13863.93
3	6684	6293.07
4	2559	3080.79
5	1520	1582.67
6	929	840.70
7	567	457.69
8	298	253.91
9	150	142.97
10	54	81.48
11	21	46.90
12	14	27.23
13	10	15.92
14	1	9.36
15	0	5.54
16	1	3.29
17	0	1.96
18	0	1.18
19	0	0.71
20	2	1.08
k = 1.9909, m = 3.2156, q = 0.6398		

X^2	**$P(X^2)$**	**DF**	**C**	**R^2**
210.1593	–	15	0.0034	0.9996

Four data sets give a good result with the Zipf-Mandelbrot distribution (C <= 0.090), and a single one a satisfying fit (C = 0.0140). The Waring distribution turned out as compatible with 6 of the data sets with a good C value (< 0.08) and with one at C = 0.0140. Remarkably, the data sets which give good results come from newspaper texts and computer news, and all these corpus parts share the good result. Exactly the same data sets are also compatible with the right truncated modified Zipf-Alekseev distribution.

Another interesting finding is that all the texts in our study are compatible with the modified negative binomial distribution except the two data sets from the “law” category, i.e. juridical texts. Fig. 3 and Table 2 show the results of fitting this distribution to one of the two data sets from the “computer news” section.

The negative hypergeometric distribution is also a good model of the complexity distributions of all the texts in the treebank, again with the exception of the juridical texts. These two data sets are best modelled by the right truncated modified Zipf-Alekseev distribution (cf. Table 3 and Fig. 4, which show the results for the file “szerzj”), whereas this model does not fit with the data from the “compositions” category.

Table 2
Results of fitting the Modified negative binomial distribution to the data
from one of the two “computer news” data sets (cwszt) in the Szeged Treebank 2.0

x[i]	**f[i]**	**NP[i]**
1	76725	76641.28
2	30524	30069.95
3	15239	14966.32
4	6092	7274.57
5	3515	3493.50
6	1843	1665.49
7	963	790.12
8	415	373.52
9	188	176.11
10	68	82.87
11	23	38.93
12	4	18.26
13	4	8.56
14	2	4.01
15	1	1.87
16	1	1.64

k = 1.1511; p = 0.5372; α = 0.8511				
X^2	$P(X^2)$	DF	C	R^2
290.8502	–	12	0.0021	0.9997

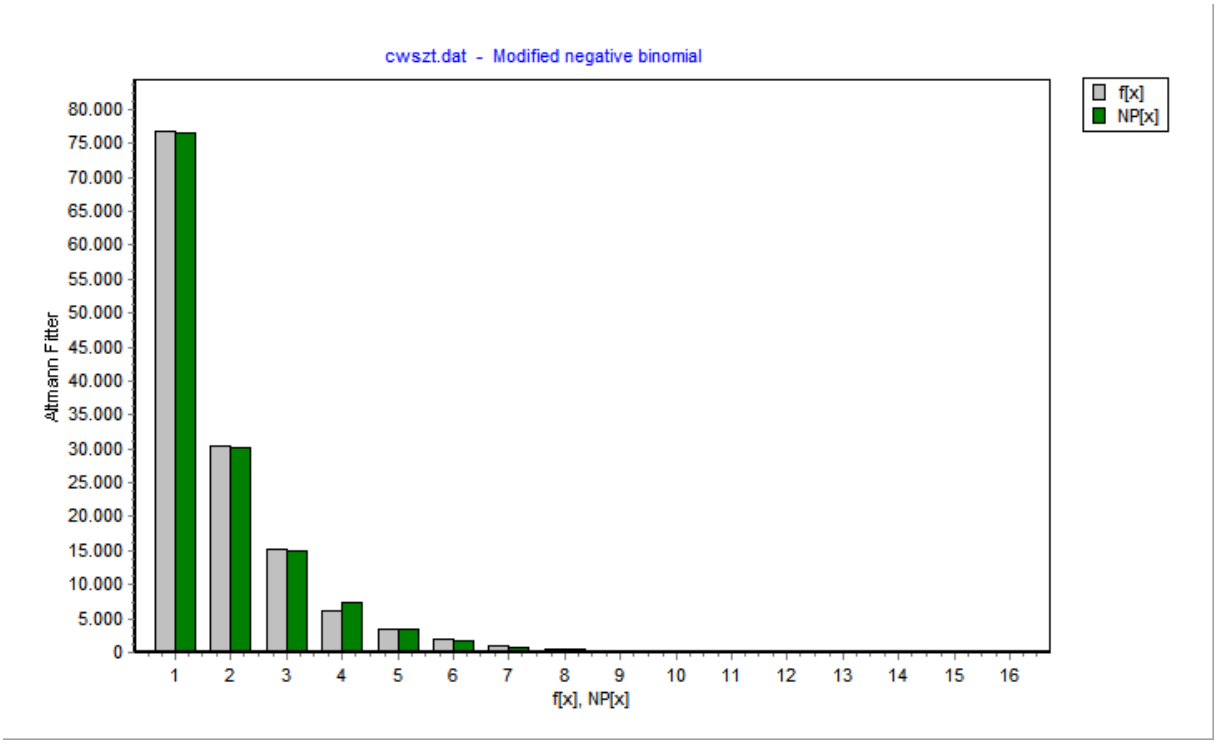


Figure 3. Graph of the results presented in Table 1

Table 3
Results of fitting the Right truncated modified Zipf-Alekseev distribution
to the data from one of the two "law" data sets (szerzj) in the Szeged Treebank 2.0

x[i]	**f[i]**	**NP[i]**
1	45682	45682.00
2	23882	24124.33
3	11861	10353.15
4	3144	4606.47
5	2066	2180.34
6	1211	1094.85
7	773	579.07
8	388	320.31
9	213	184.17
10	100	109.52
11	36	67.07
12	33	42.16
13	16	27.12
14	5	17.81
15	2	11.92
16	1	8.11
17	1	5.60

a = 0.2027; b = 1.0513; n = 17; γ = 0.5109

X^2	$P(X^2)$	DF	C	R^2
837.6989	–	12	0.0094	0.9981

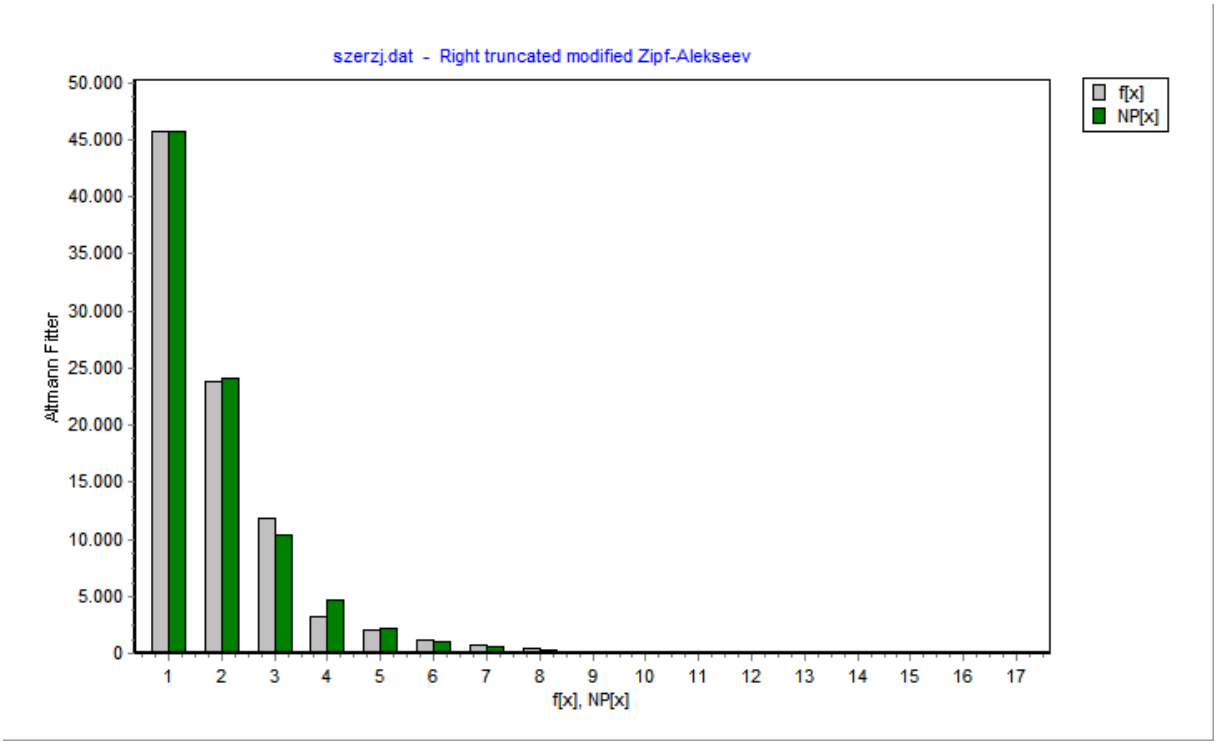


Figure 4. Graph of the results presented in Table 3

The results of the present study support the general hypothesis that complexity of syntactic constructions abide by a lawful distribution; the theoretical probability distributions which proved suitable to model the data from the Szeged Treebank 2.0, however, differ from the findings in (Köhler/Altmann 2000) for English and German. Currently, we have no evidence as to whether the difference is caused by properties of the Hungarian grammar, by the structure of the linguistic data, or still other factors. One possible factor could be excluded: corpus part vs. individual text. We studied one individual text from each of the corpus parts to find out whether their complexity distribution would differ from those of the corpus parts (i.e. text sorts). Five theoretical probability distributions (cf. Wimmer/Altmann 1999 for further details of the distributions) were fitted to the data:

(1) the hyper-Pascal d.

$$P_x = \frac{\binom{k+x-1}{x}}{\binom{m+x-1}{x}} q^x P_0, \quad x = 0,1,2,\ldots;$$

(2) the Zipf-Mandelbrot d.

$$P_x = \frac{(b+x)^{-a}}{F(n)}, \quad x = 1,2,3,\ldots,n$$

where $F(n) = \sum_{i=1}^{n} (b+i)^{-a}$;

(3) the Waring d.

$$P_x = \frac{b}{b+n} \frac{n^{(x)}}{(b+n+1)^{(x)}}, \quad x = 0,1,2,\ldots$$

where $x^{(n)} = x(x+1)(x+2)\ldots(x+n-1)$;

(4) the right truncated modified Zipf-Alekseev

$$P_x = \begin{cases} 1-\alpha, & x = 1 \\ \dfrac{\alpha\, x^{-(a+b\log_e x)}}{T}, & x = 2,3,4,\ldots,n \end{cases}$$

where $T = \sum_{j=2}^{n} j^{-(a+b\log_e j)}$;

(5) the modified negative binomial d.

$$P_x = \begin{cases} 1-\alpha+\alpha q^T, & x = 0 \\ \alpha\, p^x q^{T-hx} \left[\binom{T-(h-1)x-1}{x} + \sum_{k=1}^{h-1} \binom{T-h(x-1)+x-k-2}{x-1} q^{h-k} \right], & x = 1,2,\ldots,[T/h] \ \ (if\ \ T \geq h) \\ 1 - P(X \leq \lceil T/h \rceil), & x = [T/h]+1 \end{cases}$$

The results turned out as mixed as those on data from text collections:

Table 4
hyper-Pascal: four good results (two juridical texts and two newspaper texts)

Input data	X^2	$P(X^2)$	C	DF	k	m	q	N
computer-cwst-compl.dat	-	-	-	-	-	-	-	-
computer-win2000-compl.dat	-	-	-	-	-	-	-	-
fiction-pfred-compl.dat	168.42	0	0.2249	3	4.0689	35.5711	1.7152	749
fiction-utas-compl.dat	-	-	-	-	-	-	-	-
law-gazdtar-compl.dat	0.58	0.4449	0.0087	1	2.5879	1.1335	0.2819	67
law-szerzj-compl.dat	5.7	0.0579	0.0553	2	143.769	7.3967	0.0298	103
newspaper-hvg-compl.dat	26.06	0.0005	0.0196	7	12.5366	20.8999	0.7757	1328
newspaper-mh-compl.dat	0.76	0	0.019	0	0.3234	0.1811	0.3465	40
newspaper-np-compl.dat	2.29	0.5145	0.0111	3	1.4429	1.046	0.4016	206
newspaper-nv-compl.dat	7.32	0.0624	0.0383	3	0.5125	0.3591	0.4128	191

Table 5
Zipf-Mandelbrot: six with good fitting results and one acceptable one (i.e. three texts did not fit with the model: a juridical, a newspaper and a fiction text)

Input data	X^2	$P(X^2)$	C	DF	a	b	n	N
computer-cwst-compl.dat	8.44	0.2076	0.0112	6	3.9653	2.2278	10	752
computer-win2000-compl.dat	2.73	0.604	0.0042	4	7.3769	6.5331	8	655
fiction-pfred-compl.dat	7.75	0.1705	0.0103	5	4.532	1.9492	9	749
fiction-utas-compl.dat	8.13	0.0044	0.301	1	12	15.4286	5	27
law-gazdtar-compl.dat	3.71	0.0542	0.0553	1	4.263	4.2405	5	67
law-szerzj-compl.dat	14.19	0.0067	0.1378	4	5.2031	4.5531	9	103
newspaper-hvg-compl.dat	29.71	0.0005	0.0224	9	4.6199	3.2984	13	1328
newspaper-mh-compl.dat	4.51	0.1047	0.1129	2	2.0398	0.8224	6	40
newspaper-np-compl.dat	8.83	0.0317	0.0429	3	9.9886	10.6055	8	206
newspaper-nv-compl.dat	4.46	0.347	0.0234	4	5.8191	4.8367	9	191

Table 6
Waring: nine of ten texts yielded very good fitting results
(only one text from one of the four newspaper sub-corpora)

Input data	X^2	$P(X^2)$	C	DF	b	n	N
computer-cwst-compl.dat	11.55	0.1162	0.0154	7	5.2528	3.5058	752
computer-win2000-compl.dat	6.53	0.2578	0.01	5	8.4963	6.0719	655
fiction-pfred-compl.dat	10.6	0.1015	0.0142	6	4.0723	1.7155	749
fiction-utas-compl.dat	1.29	0.2564	0.0477	1	1.9033	0.5438	27
law-gazdtar-compl.dat	3.82	0.148	0.057	2	149904.067	134242.448	67
law-szerzj-compl.dat	9.87	0.0789	0.0959	5	289121.303	331284.827	103
newspaper-hvg-compl.dat	22.02	0.0088	0.0166	9	8.2439	6.9194	1328
newspaper-mh-compl.dat	2.37	0.4989	0.0593	3	12.4663	11.8429	40
newspaper-np-compl.dat	5.84	0.3216	0.0284	5	328241.681	299560.369	206
newspaper-nv-compl.dat	3.34	0.6483	0.0175	5	15.9491	13.9164	191

Table 7
Right truncated modified Zipf-Alekseev: five good and one acceptable
(bad results were obtained from a fiction, a juridical and two kinds of newspaper texts).

Input data	X^2	$P(X^2)$	C	DF	a	b	n	α	N
computer-cwst-compl.dat	5.44	0.3649	0.0072	5	0.3169	0.847	10	0.5997	752
computer-win2000-compl.dat	1.69	0.6383	0.0026	3	0.014	1.005	8	0.5832	655
fiction-pfred-compl.dat	4.7	0.3196	0.0063	4	0.7059	0.8268	9	0.7036	749
fiction-utas-compl.dat	-	-	-	-	-	-	-	-	-
law-gazdtar-compl.dat	-	-	-	-	-	-	-	-	-
law-szerzj-compl.dat	6.68	0.0355	0.0648	2	0.4125	1.0394	9	0.466	103
newspaper-hvg-compl.dat	20.28	0.005	0.0153	7	0.4232	0.8382	13	0.5437	1328
newspaper-mh-compl.dat	1.26	0	0.0316	0	2.0865	0.1427	6	0.475	40
newspaper-np-compl.dat	3.66	0.1604	0.0178	2	0.1251	1.1065	8	0.4806	206
newspaper-nv-compl.dat	2.59	0.4591	0.0136	3	1.6638	0.4006	9	0.534	191

Table 8
Modified negative binomial: all but a fiction and a newspaper text gave good results

Input data	X^2	$P(X^2)$	C	DF	k	p	α	N
computer-cwst-compl.dat	4.44	0.6169	0.0059	6	1.067	0.4956	0.7624	752
computer-win2000-compl.dat	3.91	0.4188	0.006	4	1.6173	0.5858	0.7232	655
fiction-pfred-compl.dat	3.61	0.6063	0.0048	5	0.6514	0.4835	0.7904	749
fiction-utas-compl.dat	0.65	0	0.0242	0	0.3737	0.3045	0.6457	27
law-gazdtar-compl.dat	0.33	0.5665	0.0049	1	8.3589	0.8912	0.8912	67
law-szerzj-compl.dat	5.04	0.0806	0.0489	2	2.6344	0.7233	0.9494	103
newspaper-hvg-compl.dat	23.74	0.0025	0.0179	8	0.7485	0.4353	0.9947	1328
newspaper-mh-compl.dat	2.18	0.14	0.0545	1	1.5162	0.5946	0.9865	40
newspaper-np-compl.dat	0.62	0.7337	0.003	2	3.7635	0.7794	0.8545	206
newspaper-nv-compl.dat	4.2	0.3795	0.022	4	0.8061	0.4779	1	191

The study showed thus that the individual properties of texts or text kinds (genres) seem to be reflected in the complexity distributions of syntactic constructions, a fact that might contribute to text categorization methods.

2. Position

Position of a syntactic constituent in the (immediate) mother construction is the second variable whose frequency distribution is studied in the present contribution. Fig. 1 can be used to illustrate how position is determined: The constituent with the label PP has position 4 in its mother construction with label S; NN has position 1 in its mother constituent PP, and the Adv node has position 1 in the AP phrase.

The same corpus parts as before were scrutinized with respect to their position distribution according to the definition given above. In this case, a satisfying result could not be achieved by means of models on the basis of probability distributions. This fact is likely to be due to the enormous amount of data which makes all kinds of goodness-of-fit tests fail which depend on the number of observations such as Chi-square and even its function C. Therefore, the fit of a continuous function instead of distributions was used as an alternative

method, which does not depend on degrees of freedom. It goes without saying that discrete as well as continuous models can be used to approximate empirical data since discreteness is a property of the model, not of reality. The data show an asymmetric unimodal shape, whence the function

$$y = ax^b e^{-cx}$$

was chosen. This function plays an important role in quantitative and synergetic linguistics where a considerable number of laws and hypotheses take this form. Table 9 shows the excellent results after fitting this function to the data.

Table 9
Results of fitting function (1) to the data from the Szeged Treebank 2.0

Sub-Corpus	R^2	**Parameter *a***	**Parameter *b***	**Parameter *c***
pupils-ess-8oelb	0.9830	15733.0743	7.7754	2.4076
pupils-ess-10elb	0.9849	18152.1967	7.8682	2.3929
pupils-ess-10erv	0.9906	16204.5326	7.2278	2.1610
computer-cwszt	0.9921	7681.1614	7.6485	2.0885
law-gazdtar	0.9987	15079.8964	5.7149	1.6062
newspaper-hvg	0.9956	3370.7919	7.0741	1.8902
newspaper-mh	0.9943	4139.4800	6.9083	1.9507
newsml	0.9927	15162.1250	7.1794	2.0045
newspaper-nv	0.9930	2136.3434	7.1711	2.0321
fiction-pfred	0.9856	20098.3985	6.1812	2.0784
law-szerzj	0.9980	7532.5500	5.5739	1.4999
fiction-utas	0.9912	11240.8648	6.4243	1.9247
computer-win2000	0.9926	4188.3745	7.1226	1.9573

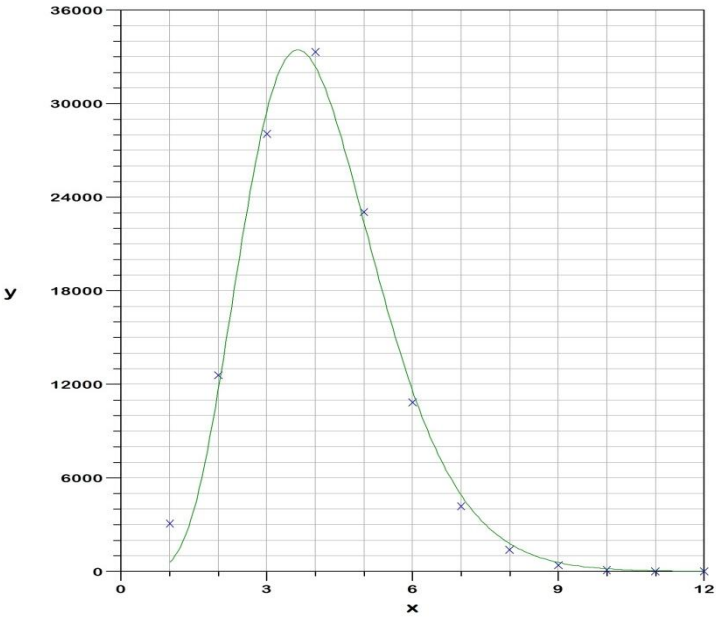


Figure 5. Plot of the fit of function (2) to the data from file “win2000”

In (Köhler/Altmann 2000), the Binomial distribution was theoretically derived for the distribution of the position of syntactic constructions in their mother constituent. This assumption was empirically supported by data from a German and an English corpus. Therefore, a set of individual texts was taken from the Szeged Treebank 2.0, and the Binomial distribution was fitted to these data. However, this distribution did not give a single acceptable fit to the Hungarian data. The derivation of the Binomial distribution was based on the (theoretically justified) assumption that complexity of syntactic constructions is distributed according to the hyper-Pascal distribution. The highest position value a constituent can take is that of the complexity of its mother constituent. Therefore, this parameter of the theoretical model is not free in the case of position and another parameter compactness, is a constant, which yields the Binomial distribution (cf. Köhler/Altmann 2000). The parameter *p* of the Binomial d. corresponds to the factor *E* in the original hyper-Pascal model and stands for fullness, i.e. the degree of informativeness of an expression. This parameter was assumed to be (more or less) constant within a text. This assumption may not be justified in Hungarian texts; the need for full information (or for a certain degree of redundancy) may vary within a text due to a yet unknown influence from the organisation of discourse. We take this possibility into account and consider the parameter *p* as a random variable representing the boundary conditions connected to the properties of the text types and the individual texts, instead of a constant. If this parameter of a Binomial distribution follows a beta distribution (a very flexible one) exactly the negative hypergeometric d. is obtained.

Fitting this distribution to our data from the individual texts yields very good results without any exception (cf. Table 10 and Fig. 6, which shows one of the fitting results).

Table 10
The results of fitting the negative hypergeometric distribution to the individual texts.

Input data	X^2	$P(X^2)$	C	DF	K	M	n	N
computer-cwst-pos.dat	3.28	0.6576	0.0044	5	5.211	0.7461	9	752
computer-win2000-pos.dat	0.7	0.9516	0.0011	4	3.8199	0.6557	7	655
fiction-pfred-pos.dat	3.92	0.4163	0.0052	4	4.9198	0.5238	8	749
fiction-utas-pos.dat	0.76	0.3837	0.0281	1	1.7226	0.4438	4	27
law-gazdtar-pos.dat	0	0.9598	0	1	2.7672	0.5885	4	67
law-szerzj-pos.dat	5.8	0.1219	0.0563	3	4.8927	0.7723	8	103
newspaper-hvg-pos.dat	14.01	0.1218	0.0106	9	7.1103	0.6536	16	1328
newspaper-mh-pos.dat	2.33	0.3113	0.0584	2	1.3574	0.4658	5	40
newspaper-np-pos.dat	0.21	0.8993	0.001	2	6.4656	0.9371	7	206
newspaper-nv-pos.dat	1.4	0.8438	0.0073	4	4.1057	0.6017	9	191

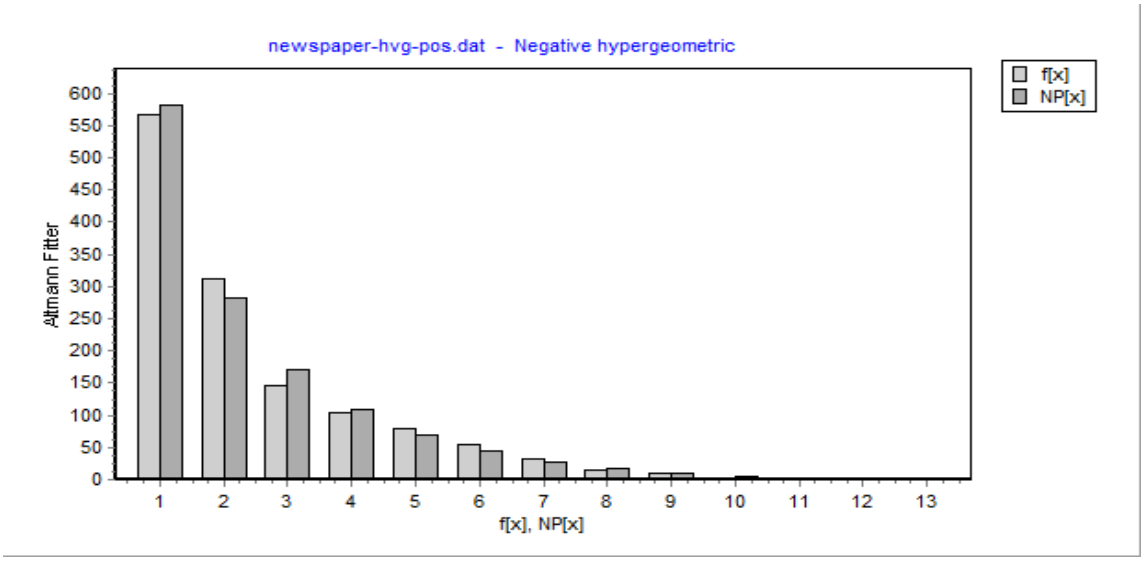


Figure 6. Graph of the position data from one of the newspaper texts and the theoretically expected values (negative hypergeometric distribution).

Conclusion

The study tested the mathematical models of the distributions of syntactic complexity and position in the mother constituent of syntactic constructions as developed and tested on data from German and English on sub-corpora and individual texts from the Szeged Treebank 2.0. The Hungarian data sets showed some differences from the previously investigated languages. In particular, the frequency distributions of position values do not follow the binomial distribution as in the other languages but can be modelled by a mixture of the binomial and the beta distributions. Thus, the assumptions made in (Köhler/Altmann 2000) are, in principle, supported.

An important question is open at this point: Which are the factors that cause the variation of the parameter *p* of the binomial d.? The first idea that differences in discourse organisation between Hungarian on the one side and German and English on the other side must be linguistically discussed and empirically investigated on data from other languages of different typological types.

A second question is rather theoretical: Is there a very general probability distribution capturing all data and taking parameter values specific for individual languages and text sort? Evidently, this complex question is a task for future research.

References

Köhler, Reinhard (1999). Syntactic Structures: Properties and Interrelations. *Journal of Quantitative Linguistics, 6/1, 46-57.*

Köhler, Reinhard; Altmann, Gabriel (2000). Probability Distributions of Syntactic Units and Properties. *Journal of Quantitative Linguistics, 7/3, 189-200.*

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

Glottometrics 26, 2013, 38-49

Entropy of a Zipfian Distributed Lexicon

L. C. Araujo[1], *T. Cristófaro-Silva*[2], *H. C. Yehia*[3]

Abstract. This article presents the calculation of the entropy of a system with Zipfian distribution. It shows that a communication system tends to present an exponent value close to, but greater than one. This choice both maximizes entropy and, at the same time, enables the retention of a feasible and growing lexicon. This result is in accordance with what is observed in natural languages and with the balance between the speaker and listener communication efforts. On the other hand, the entropy of the communicating source is very sensitive to the exponent value as well as the length of the observable data. Slight deviations on these parameters might lead to very different entropy measurements. A comparison of the estimation proposed with the entropy measure of written texts yields errors in the order of 0.3 bits and 0.05 bits for non-smoothed and smoothed distributions, respectively.

Keywords: *Zipf's law, entropy, communication*

1 Introduction

Statistical linguistics makes use of Zipf analysis, which is a statistical tool also used in several research fields, such as economics (Mandelbrot, 1963), gene expression (Furusawa and Kaneko, 2003), and chaotic dynamic systems (Nicolis et al., 1989). Zipf found a power-law relation for the rank frequency distribution of words in written texts[4] in natural languages (Zipf, 1949)

$$f_k \propto k^{-s} \quad (1)$$

where k is the rank of a word, f_k is the word token frequency with rank k and s is the exponent of the characterizing distribution. This empirical observation has become the most remarkable statement in quantitative linguistics. The observation of Zipf-like behavior is necessary in a natural text, as much as it is the necessary behavior of any source producing information content. This is because any randomly generated symbolic sequence will follow Zipf's law with an exponent between 1 and 2 (Miller, 1957; Li, 1992). The systematic organization of

[1] Universidade Federal de Sao Joao del-Rei - Campus Alto Paraopeba - Rod. MG 443, Km 7, 36420-000, Ouro Branco, MG, Brazil. E-mail: leolca@efsj.edu.br
[2] Universidade Federal de Minas Gerais - Faculdade de Letras - Av. Antonio Carlos 6627, 31270-901, Belo Horizonte, MG, Brazil. E-mail: thaiscristofarosilva@ufmg.br
[3] Universidade Federal de Minas Gerais - Departmento de Eletronica - Av. Antonio Carlos 6627, 31270-901, Belo Horizonte, MG, Brazil. E-mail: hani@ufmg.br
[4] Text is a term used in linguistics to refer to any written or spoken passage of whatever length, that does form a unified whole (Halliday, Hassan 1976).

language reflects the frequency usage of types. Studies have suggested that the frequency of usage is a key factor in the access of lexical items (Balota and Chumbley, 1984) and is also a driving force in language change (Bybee, 2002). Frequency plays an important role in understanding how human communication works. A useful example of Zipf analysis is given by Havlin (1995), who suggested a dissimilarity measure of two Zipf plots, from two different sources, which are smaller when the data comes from the same source and larger when it comes from different sources. This approach has been used to perform authorship attribution (Havlin, 1995).

Shannon (1948) proposed a mathematical way to deal with general communication systems and information transmission. The basic model considered by Shannon (1948) consists of i) an information source, which produces a message; ii) a transmitter, which operates on the message creating a signal suitable for transmission; iii) the channel, a mere medium where the signal is transmitted; iv) a receiver, which performs the inverse operation of the transmitter; and v) the destination, the person (or thing) for whom the message is intended. Communication is regarded as a sequence of random variables which are distributed according to the characteristics of the source. Entropy was defined as a measure of uncertainty in a random variable, what defines the expected value of the information content in a message. It is a measure of the average information produced by a source for the symbols produced in its output. Expressing entropy in bits gives us the average number of bits necessary to express each symbol produced by the source. Shannon (1948) also defined redundancy, what adds little, if any, information to a message, but helps overcome errors arriving on the information transmission process. On a language, we might regard redundancy as a measure of the restrictions imposed on that language due to its statistical structure, which might be, for example, an expression of physiological, perceptual and phonological constraints.

The entropy of English printed words was estimated by Shannon (1951) and Grignetti (1964) using a Zipfian distribution with a characteristic exponent $s = 1$. It is known that natural languages typically present $s \approx 1$ (Piotrovskii et al., 1994). Some types of human communication still present a greater exponent, for example, children's speech has been reported to present $s \approx 1.66$ and military combat text $s \approx 1.42$. Studies on animal communication also present Zipfian behaviour. For example, McCowan et al. (1999) present an exponent value of $s \approx 1.1$ and $s \approx 0.87$ for adult and infant dolphins, respectively. The value of the exponent s seems to be related to the possible existence of a wider lexicon. Larger values of s characterize systems still in formation whereas small values characterize a well-grounded system. In this paper, we present the estimation of the entropy of a system using an arbitrary Zipfian distribution and verify the effect of the characteristic exponent s on the entropy of the system. Some results are presented to compare the estimated entropy and the entropy found in written texts, as we consider *words* as the symbols produced by a Zipfian distributed source.

2 Entropy of the System

The entropy of a system using N symbols of probabilities P_k, where $k = 1$ to N , is given by

If we consider words as the symbols used by our system, the probabilities p_k might be estimated by counting the corpus tokens and dividing it by the sample size.

George Kingsley Zipf made important contributions on language statistics, by

performing word count experiments, from which he determined that there is a relationship between word frequency and rank: their product is roughly a constant (Zipf, 1949). The distribution of words in a text follows a power law:

$$p_k(s,N) = Ck^{-s} \tag{3}$$

where p_k stands for the probability of occurrence of the k-th most frequent word in the corpus; C is a normalizing constant, $C^{-1} = \sum_{n=1}^{N} n^{-s}$ which is the generalized harmonic number; k is the word rank; s is the slope, which characterizes the distribution; and N is the number of elements in the set. Zipf's law seems to hold in various languages (Zipf, 1949). "Investigations with English, Latin, Greek, Dakota, Plains Cree, Nootka (an Eskimo language), speech of children at various ages, and some schizophrenic speech have all been seen to follow this law"(Alexander et al., 1998).

Using the Zipfian value for the probabilities in Equation 2, we get

$$\begin{aligned} \bar{H} &= -\frac{1}{\ln 2}\sum_{k=1}^{N} Ck^{-s}\ln(Ck^{-s}) \\ &= \frac{sC}{\ln 2}\sum_{k=1}^{N}\frac{\ln k}{k^{s}} - \frac{\ln C}{\ln 2} \end{aligned} \tag{4}$$

the summation can be calculated following the steps proposed by Grignetti (1964). We are going to find a lower and an upper bound to the entropy in Equation 4 and, to achieve that, we need to obtain bounds such that

$$B_l \le \sum_{k=1}^{N} k^{-s}\ln k \le B_u \tag{5}$$

Figure 1 presents the function

$$f(x) = x^{-s}\ln x \tag{6}$$

for different values of s greater than one, which are usually found in human languages. From the first derivative of f,

$$f'(x) = x^{-s-1}(1 - s\ln x), \tag{7}$$

we conclude that f is a decreasing function for $x > e^{1/s}$, which can be verified in Figure 1.

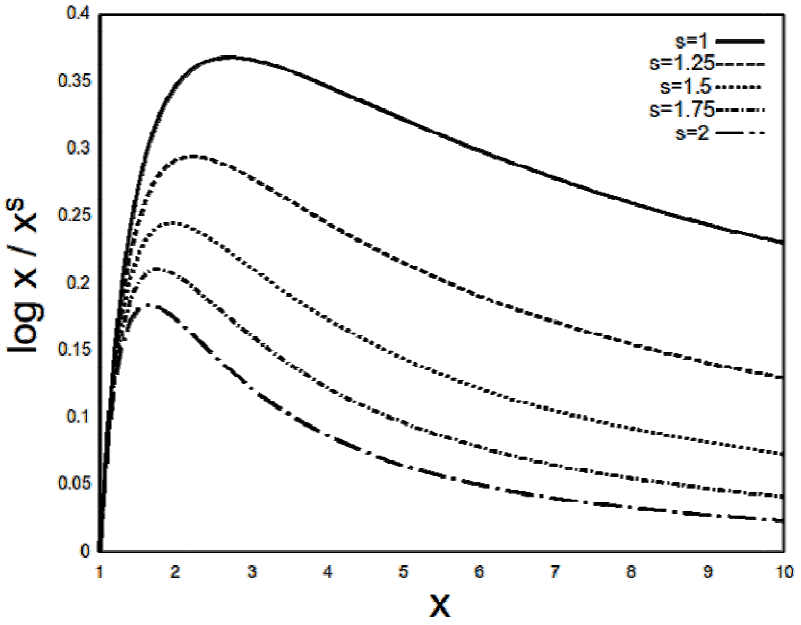


Figure 1: Function $f(x) = \ln x / x^s$ for different values of *s*

Particularly, for $s \geq 1$, the function *f* will be a decreasing function for $x > 3$ (for a general value of *s*, consider what is proposed in the next section, using q = 0). We might then approximate the summation using the Riemann sum approximation of an integral. The left Riemann sum S_l is an overestimate and the right Riemann sum S_r is an underestimate,

$$S_r \leq \int_a^b f(x)dx \leq S_l \tag{8}$$

Using Equation (8) we have

$$\sum_{n=4}^{N-1} \frac{\ln n}{n^s} \leq \int_3^{N-1} \frac{\ln x}{x^s} dx \leq \sum_{n=3}^{N-2} \frac{\ln n}{n^s} \tag{9}$$

and

$$\sum_{n=4}^{N} \frac{\ln n}{n^s} \leq \int_3^{N} \frac{\ln x}{x^s} dx \leq \sum_{n=3}^{N-1} \frac{\ln n}{n^s} \tag{10}$$

From equations (9) and (10) we conclude that

$$\int_3^N \frac{\ln x}{x^s} dx \leq \sum_{n=3}^{N-1} \frac{\ln n}{n^s} \leq \int_3^{N-1} \frac{\ln x}{x^s} dx + \frac{\ln 3}{3^s} \tag{11}$$

and, by adding the remaining terms (i.e. *n* = 1 and 2) to the summation, we get

$$\int_3^N \frac{\ln x}{x^s} dx + \frac{\ln 2}{2^s} \leq \sum_{n=1}^{N-1} \frac{\ln n}{n^s} \leq \int_3^{N-1} \frac{\ln x}{x^s} dx + \frac{\ln 3}{3^s} + \frac{\ln 2}{2^s} \tag{12}$$

The wanted bounds are given by adding the *N*-th term to the above equation,

$$B_l = \int_3^N \frac{\ln x}{x^s} dx + \frac{\ln 2}{2^s} + \frac{\ln N}{N^s} \leq \sum_{n=1}^{N} \frac{\ln n}{n^s} \leq \int_3^{N-1} \frac{\ln x}{x^s} dx + \frac{\ln 3}{3^s} + \frac{\ln 2}{2^s} + \frac{\ln N}{N^s} = B_u \tag{13}$$

Finally, the entropy bounds (H_l for the lower bound and H_u for the upper bound) are given using

equations (4) and (13)

$$\frac{sC}{\ln 2} B_l - \frac{\ln C}{\ln 2} = H_l \le \bar{H} \le H_u = \frac{sC}{\ln 2} B_u - \frac{\ln C}{\ln 2} \tag{14}$$

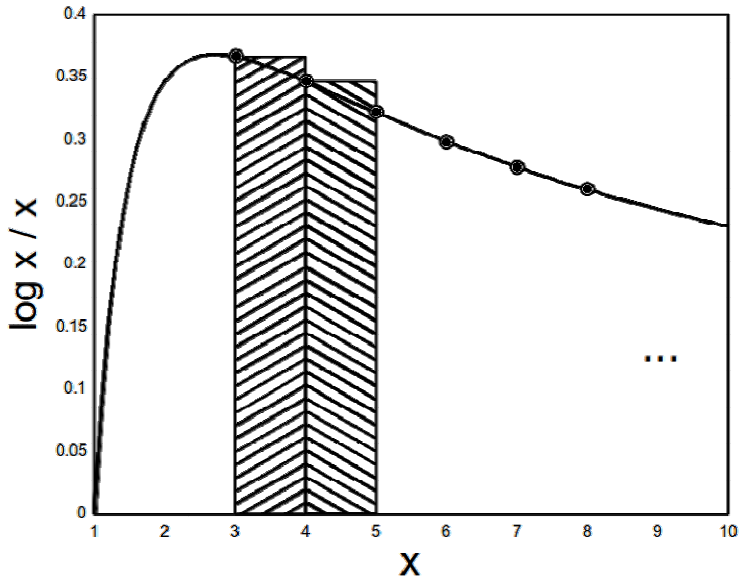


Figure 2: Left Riemann sum approximation of the integral.

The proposed approximation procedure is illustrated in Figure 2. The integral in Equation 13 is solved by parts, giving the following result, when s :

$$\int \frac{\ln x}{x^s} dx = \frac{x^{1-s}}{1-s}\left(\ln x - \frac{1}{1-s}\right), \tag{15}$$

where the integration constant is omitted, since it is irrelevant when evaluating the integral in an interval. When $s = 1$ the integral will result in

$$\int \frac{\ln x}{x} dx = \frac{(\ln x)^2}{2} \tag{16}$$

Using Equations 4, 12 and 15 (or 16) we are able to calculate the entropy bounds of a Zipfian distributed source for a given s and N. Figure 3 presents some results for different corpora. We observe that the entropy decreases with s and increases with N.

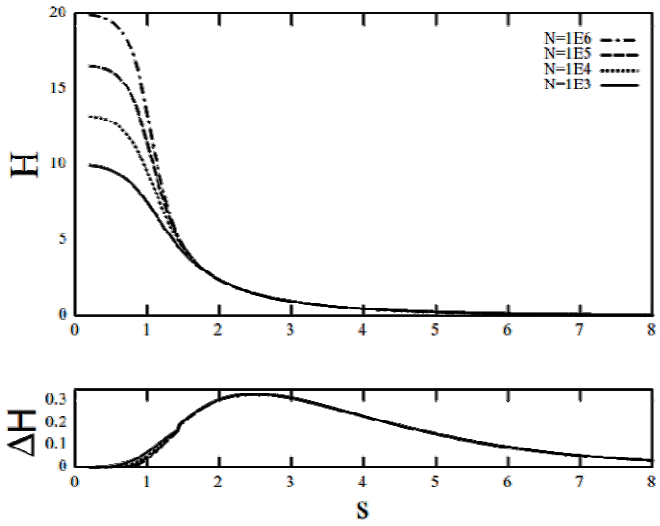


Figure 3: Entropy *H* (in bits) as a function of the Zipf exponent *s* and the number of types *N*. The upper plot presents the average Entropy estimated and the lower plot presents the difference between the upper and lower bounds of the entropy estimated.

The bounds for the entropy found here are rather tight (see *ΔH* in Figure 3), which leads to a good approximation as we consider the average of the bound values. The entropy is very sensitive to the parameter *s* in the vicinities of 1, where language communication is usually found. The size of the corpus also deeply influences the value of the entropy for small values of *s* The proposed method states a way to estimate the entropy of theoretical Zipfian distributed sources, given its characteristic exponent value *s* and the number of types *N* .

2.1 Entropy of Real Texts

In the previous section we presented a method to estimate the entropy of a Zipfian distributed source. In this section we present the results from some measurements of entropy for the following texts: Alice in Wonderland (Lewis Carroll); Hamlet (William Shakespeare); Macbeth (William Shakespeare); The Complete Works of William Shakespeare; and Ulysses (James Joyce).

Table 1 presents the results obtained from this comparative study using the Zipf exponent calculated through a maximum likelihood estimation and the number of types found in each text. Although the estimated entropy is close to the observed entropy, it is consistently smaller, for the reasons we argue below. The estimated Zipf's exponents are also biased, since there is a flattening in low rank.

Zipf's law is a model that coarsely approximates the distribution of words in language. Two important deviations of the model from real data are observed. On the low rank region, real data usually present a flattened pattern, which was accounted for by Mandelbrot (1965). This type of flattening results in a higher entropy, since it shapes a part of the distribution towards a uniform distribution. On the long tail (high rank), a staircase pattern is usually found, due to the undersampling of rare words. This second deviation from the ideal Zipfian model does not

produce a significant change on the entropy measure, therefore only the flattened pattern deviation will be considered here.

2.2 Zipf-Mandelbrot Entropy

In order to take into account the flattening observed on the low rank region of a Zipf plot, Mandelbrot (1965) introduced a modification of Zipf's law, adding a constant q to the rank k, resulting in Zipf-Mandelbrot's law

$$p_k(s,q,N) = C(k+q)^{-s} \tag{17}$$

where the new normalizing constant (a generalization of a harmonic number) is given by $C^{-1} = \sum_{n=1}^{N} (\mathrm{n}+\mathrm{q})^{-s}$.

Applying the same steps to this generalized formulation, the entropy will be given by

$$\bar{H} = \frac{sC}{\ln 2}\sum_{k=1}^{N}\frac{\ln(k+q)}{(k+q)^s} - \frac{\ln C}{\ln 2} \tag{18}$$

The new function that will be used by the Riemmann integral approximation is

$$f(x) = (x+q)^{-s}\ln(x+q), \tag{19}$$

which is decreasing for $x > e^{1/s} - q$. The constant q is a real value in the interval $[0;\infty)$. We shall then define an integer constant

$$\mathrm{K} = \max(\lceil e^{1/s} - q \rceil, 1), \tag{20}$$

which guarantees that the function $f(x)$ is decreasing for $x > K \geq 1$.

Using the left and right Riemann sum again, we find the inequalities below, which are respectively equivalent to equations 9 and 10:

$$\sum_{n=K+1}^{N-1}\frac{\ln(n+q)}{(n+q)^s} \leq \int_{K}^{N-1}\frac{\ln(x+q)}{(x+q)^s}dx \leq \sum_{n=K}^{N-2}\frac{\ln(n+q)}{(n+q)^s} \tag{21}$$

$$\sum_{n=K+1}^{N}\frac{\ln(n+q)}{(n+q)^s} \leq \int_{K}^{N}\frac{\ln(x+q)}{(x+q)^s}dx \leq \sum_{n=K}^{N-1}\frac{\ln(n+q)}{(n+q)^s} \tag{22}$$

From the above equations we conclude that

$$\int_K^N \frac{\ln(x+q)}{(x+q)^s}dx \leq \sum_{n=K}^{N-1} \frac{\ln(n+q)}{(n+q)^s} \leq \int_K^{N-1} \frac{\ln(x+q)}{(x+q)^s}dx + \frac{\ln(K+q)}{(K+q)^s} \tag{23}$$

which is equivalent to Equation 11. By adding the remaining terms we get the following boundaries

$$\begin{aligned} B_l &= \int_K^N \frac{\ln(x+q)}{(x+q)^s}dx + \sum_{n=1}^{K-1} \frac{\ln(n+q)}{(n+q)^s} + \frac{\ln(N+q)}{(N+q)^s} \\ &\leq \sum_{n=1}^{N} \frac{\ln(n+q)}{(n+q)^s} \\ &\leq \int_K^{N-1} \frac{\ln(x+q)}{(x+q)^s}dx + \sum_{n=1}^{K} \frac{\ln(n+q)}{(n+q)^s} + \frac{\ln(N+q)}{(N+q)^s} = B_u \end{aligned} \tag{24}$$

The integral in Equation 24 is solved by parts, giving the same results presented in equations 15 and 16, considering that we have $x + q$ instead of x. By adding the parameter q, the distribution suffers a flattening in the lower rank values and, consequently, the entropy of the source increases.

Table 1 also presents a comparison between the entropy estimates and the entropy found in real text data. We might observe that the usage of the Zipf-Mandelbrot model has improved the estimation of entropy. The improvement is more evident when smoothing is applied prior to the computation of the entropy.

2.3 Simple Good-Turing Smoothing

Many linguistic phenomena might be essentially regarded as infinite: words and sentences, for example. No matter how large a sample size is, it is always prudent to consider that many types have not appeared in that sample, thus they should not receive zero probability, since it is a matter of chance that they do not appear in the sample while others appear just once. Smoothing techniques reallocate the mass probability of types in order to provide a way of taking into account the probability of those types not observed in corpora. It adjusts the maximum likelihood estimates of probabilities in order to achieve a better estimate when there is insufficient data to accurately approximate them. The name *smoothing* is used because it tends to flatten the probabilities by lowering high probabilities and increasing the low ones (Chen and Goodman, 1998).

A particular technique that we use here is Good-Turing smoothing. It considers that unseen events together have a probability equal to the sum of the probabilities of all events that were observed only once, since they are equally rare. It is important to note that there is a relationship between Turing's smoothing formula and Zipf's law: both are shown to be instances of a

common class of re-estimation formula and Turing's formula "smooths the frequency estimates towards a geometric distribution. (...) Although the two equations are similar, Turing's formula shifts the frequency mass towards more frequent types[5]" (Samuelsson, 1996).

The Good-Turing method states an estimation f^* for the frequency of occurrence f based on the type count for a given frequency N_f and that given frequency plus one N_{f+1} :

$$f^* = (f+1)\frac{E[N_{f+1}]}{E[N_f]} \tag{25}$$

where $E[.]$ represents the expectation of a random variable. The estimation f^* is usually called the "adjusted number of observations", that represents how many words you are expected to see with a given frequency of occurrence. The probability of the unseen events will be approximated by $E[N_1]/N$. The value of N_1 is the largest value and the best estimate among all other N_f . For that reason, the value of N_1 is a good approximation of the value of $E[N_1]$.

One particular problem with the Good-Turing method is that, for a given f, N_{f+1} might not exist. Simple Good-Turing (SGT) (Gale, 1994) solves this problem by choosing $E[.]$ so that

$$E[N_{f+1}] = E[N_f]\left(\frac{f}{f+1}\right)\left(1-\frac{E[N_1]}{N}\right) \tag{26}$$

leading to

$$p_f^* = p_f\left(1-\frac{E[N_1]}{N}\right) \tag{27}$$

as the estimated probability for types with a given frequency f . This method was shown to be accurate in a Monte Carlo study using a predefined known model and by comparing the results with other smoothing techniques (Gale, 1994).

The results in Table 1 show the difference in the entropy measures for a given corpus with and without SGT smoothing. The estimated entropy, by the method proposed here, is significantly closer to the measured entropy when SGT smoothing is applied.

5 We have replaced *species* in the original text by *types* which is more appropriate in the context.

Table 1
Entropy of real texts (bits), with and without SGT smoothing, compared with the estimated entropy (bits) using the parameter *N* (number of types) found in the text, parameter *s* (Zipf exponent) found by a Maximum Likelihood Estimation (MLE) and the flattening parameter *q*, also found by MLE.

source	*N*	estimated parameters			entropy		estimated entropy	
		Zipf	Zipf-Mandelbrot		normal	sgt	Zipf	Zipf-Mandelbrot
		s	s	q				
Alice	3016	0.992	1.172	3.27	8.49	8.79	8.55	8.73
Hamlet	5447	0.991	1.087	1.64	9.04	9.08	9.09	9.13
Macbeth	4017	0.969	1.009	0.56	9.00	9.00	9.02	9.04
Shakespeare	29847	1.060	1.172	2.33	9.52	9.57	9.60	9.69
Ulysses	34391	1.025	1.085	1.18	10.19	10.25	10.22	10.25

3 Conclusion

The entropy of a system with Zipfian distributed symbols decreases with the characteristic exponent *s*. A value of *s* greater than one is a necessary condition for the convergence of the generalized harmonic number. In the limit ($N \to \infty$), it is regarded as the Riemann zeta function, which converges for real $s > 1$. An exponent *s* which satisfies this condition leads to a Zipfian distributed lexicon which will hold regardless of how big the lexicon is.

This limiting value of *s* close to one is also found by Cancho and Solé (2003) when they proposed "an energy function combining the effort for the hearer and the effort for the speaker". The minimization of this function leads to a Zipfian distribution where $s = 1$, which is consistent with what is found in human languages. An exponent s greater than 1 is necessary in order to guarantee a hypothetically growing lexicon without bounds. We might then expect a greater exponent when language acquisition is still in process and a smaller exponent, closer to one, when this learning period is consolidated. Rudimentary and severely restricted communication systems might experience an exponent smaller than one, since they are not expected to evolve and widen through time, and that choice increases the entropy of the source.

The maximum rank and the repertoire size are influenced by the length of the observation but, in practical aspects, it will always be limited due to a finite observation interval. The set of words observed in the sample will always lead to a finite lexicon. An infinite lexicon is only a hypothetical approximation, which is important to analyze under the assumption of the constantly growing underlying lexicon used in human communication. Figure 4 presents an adaptation from Mandelbrot (1953), where the entropy of a finite and an infinite lexicon are compared as functions of the Zipf exponent. From both figures 3 and 4, we might observe that the length of the sample is crucial in determining the entropy of the source. A simple truncation of the sample may lead to a severe distortion of the entropy estimate. It is also important to note that the

entropy estimate is much more sensitive for *s* in the vicinity of 1 , meaning that two sources with different characteristics might have similar values of their Zipf exponent, but present quite different entropy estimates. The proposed estimation for the entropy of a Zipfian distributed source presents consistent results with real data. As the estimated measure is very sensitive to the exponent *s*, a slight deviation from the true exponent might lead to a poor estimation of the information associated with the communication system.

The results in Table 1 show that there is a noticeable difference between the estimated entropy of Zipf's model and the real entropy of written texts. A better approach is given by considering the generalized Zipf-Mandelbrot law. Slight differences between the entropy measure and estimations are due to small deviations in the real data from the ideal model. If the Zipf or Zipf-Mandelbrot distributions are used straightforwardly on the estimation of the entropy of the source, attention must be paid to the possible deviation from the real entropy value, since what we have proposed here is a theoretical approximation of the entropy of a truly Zipfian distributed source.

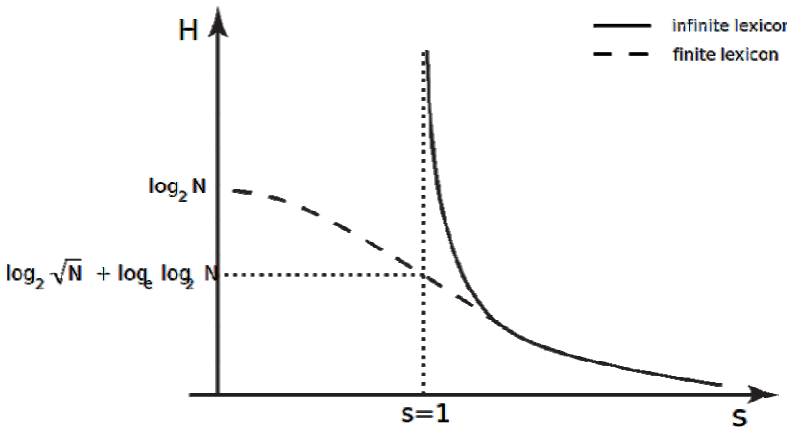


Figure 4: The entropy of a source with Zipfian distribution as a function of the characteristic exponent. Finite lexicon and infinite lexicon behaviour are compared (adapted from Mandelbrot (1953)).

Acknowledgments

This work has been supported by the Brazilian agencies CNPq and FAPEMIG.

References

Alexander, L., Johnson, R., Weiss, J. (1998). Exploring zipf's law. *Teaching Mathematics Applications 17 (4), 155–158.*

Balota, D. A., Chumbley, J.I. Jun. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of experimental*

psychology. Human perception and performance 10 (3), 340–357.

Bybee, J. (2002). Word Frequency and Context of Use in the Lexical Diffusion of Phonetically Conditioned Sound Change. *Language Variation and Change 14 (3), 261–290.*

Cancho, R. F., Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences of the United States of America 100 (3), 788–791.*

Chen, S. F., Goodman, J. (1998). *An empirical study of smoothing techniques for language modeling*. Tech. rep., Computer Science Group, Harvard University.

F**urusawa, C., Kaneko, K.** (2003). Zipf's law in gene expression. *Physical review letters 90 (8).*

Gale, W., Sampson, G. (1995). Good-Turing smoothing without tears. *Journal of Quantitative Linguistics 2, 217-237*

Grignetti, M. (1964). A note on the entropy of words in printed english. *Information and Control 7, 304–306.*

Halliday, M.A.K., Hasan, R. (1976). *Cohesion in English.* English language series. Longman.

Havlin, S. (1995). The distance between zipf plots. *Physica A: Statistical Mechanics and its Applications 216 (1), 148–150.*

Li, W. (1992). Random texts exhibit zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory 38 (6), 1842–1845.*

Mandelbrot, B. (1953). *Contribution à la théorie mathématique des jeux de communication.* Publications de l'Institut de Statistique de l'Université de París. Institut Henri Poincaré.

Mandelbrot, B. (1963). Oligopoly, mergers, and the paretian size distribution of firms. *External research note: Nc-246, IBM.*

Mandelbrot, B. (1965). Information theory and psycholinguistics. In: Wolman, B.B., Nagel, E.N. (Eds.), *Scientific Psychology: Principles and Approaches. Basic Books, 550–562.*

McCowan, B., Hanser, S. F., Doyle, L. R. (1999). Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires. *Animal Behaviour 57, 409–419.*

Miller, G. A. (1957). Some effects of intermittent silence. *The American Journal of Psychology 70 (2), 311–314.*

Nicolis, G., Nicolis, C., Nicolis, J.S. (1989). Chaotic dynamics, Markov partitions, and Zipf's law. *Journal of Statistical Physics 54 (54), 915–924.*

Piotrovskii, R.G., Pashkovskii, V.E., Piotrovskii, V.R. (1994). Psychiatric linguistics and automatic text processing. *Nauchno-Tekhnicheskaya Informatsiya, Seriya 2 28 (11), 21–25.*

Samuelsson, C. (1996). Relating Turing's formula and zipf's law. *CoRR cmp-lg/9606013.*

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal*

Shannon, C.E. (1951). Prediction and entropy of printed english. *Tech. Rep. 30, The Bell System Technical Journal.*

Zipf, G.K. (1949). *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology*. Hafner Pub. Co.

Glottometrics 26, 2013, 50-95

Some statistics for sequential text properties

Ioan-Iovitz Popescu, Bucharest
Peter Zörnig, Brasilia
Peter Grzybek, Graz
Sven Naumann, Trier
Gabriel Altmann, Lüdenscheid

Abstract. The aim oft he article is the measurement and the modelling of some sequential properties of word length, sentence length and word frequency by means of arc length, Hurst exponent and the distances between equal entities. Some of them were computed for various languages and their preliminary form has been shown.

Keywords: Arc length, Hurst exponent, distance, word length, sentence length, word frequency

1. Introduction

Written language moves in a one-dimensional space. The segmental entities make up a simple straight line. But if we begin to measure the properties of some entities quantitatively and replace them by the measured values, the straight line of written symbols changes and obtains a more or less fractal form. The given property need not be constant, it can begin to oscillate irregularly. In quantified form the text can be considered a time series whose properties can be scrutinized. Languages, individual texts, text sorts and properties may display differences with regard to their time series behaviour. The only condition is that the given property is variable (not constant like e.g. in monosyllabisms).

Time series have a number of properties which can be studied by a number of methods making up a whole discipline, therefore, we restrict ourselves here to the study of the smoothness/roughness of such series, i.e. to the variation of the values comparing subsequent neighbours.

There are a number of indicators measuring the smoothness of time series (cf. any text book on time series, e.g. Pandit, Wu 1983; Hamilton 1994; Brockwell, Davis 2010; Percival, Walden 2010; Kitagawa, Gersch 1996). We shall restrict ourselves to some of them used already previously in linguistics.

If the neighbouring values of some property in the time series differ strongly, the oscillation curve begins to cover more of the two-dimensional space than a simple straight line. Regular oscillation can easily be captured by Fourier analysis (or other methods) in such a way that the parameters are linguistically interpretable. But the more irregular the oscillation becomes, the more components must be added to the Fourier polynomial, and in that case the linguistic interpretation could become fuzzy.

Since in linguistic sequences there is no regular oscillation (except for those constructed artificially, e.g. rhythm in poetry), we can use the concept of smooth-

ness/roughness for our purposes and characterize texts, units, properties and in some cases also languages. Various properties can be found in Köhler, Altmann (2008: Chapter 9). In the present article we restrict ourselves to some elementary entities. As is known, the number of linguistic properties is infinite (cf. Altmann 2006) and linguistic entities are conceptual creations arising on the basis of the state of the art in linguistics.

Volatility and persistence of time series are most often measured using Hurst's exponent, Minkowski-sausage, Lyapunov coefficient and other indicators known from time-series research. Needless to say, all this can be used also in text analysis as initiated by L. Hřebíček (2000). The computation of the variance shows the variation of the values in the whole series, but does not yield an image of the neighbouring steps. Hence we begin with a slightly different approach considering the arc length between subsequent values of a property measured on respective units. This approach is naturally subdivided in several steps: first we choose a delimitable unit, e.g. syllable, morpheme, word, clause, sentence, verse, strophe, etc.; then we consider one of the enormous number of properties of these units. In the next step we analyse a text, replace the entities by the measured values and obtain a numerical sequence. Before we perform the two steps and evaluate the sequence, we state a linguistically substantiated hypothesis, define an indicator expressing the overall behaviour of the text, and propose a test for comparison of levels, units, properties, etc. Since our data/texts are usually very long, the tests can be performed asymptotically.

Our procedures are, nevertheless, merely experimental and inductive. We choose a property and scrutinize its sequential behaviour in order to obtain a first image. A theoretical substantiation can be added only after many languages and many texts will be analyzed.

2. Arc length

Here we shall define the extent of oscillation of values using the simple arc length between neighbouring entities defined as

$$L = \sum_{i=1}^{n-1} [(x_i - x_{i+1})^2 + 1]^{1/2} \tag{2.1}$$

which is the usual sum of Euclidean distances between the subsequent values. It has been used extensively in text analyses (cf. Popescu, Mačutek, Altmann 2009). It should not be used in cases where the x_i values vary in the interval [0;1] because the step itself would make up a much greater part of the distance than the values themselves. In those cases one can use e.g. the Manhattan metrics or some other variant of the Minkowski distance. In the next chapters we present some other methods.

Since this indicator depends on the text length, it has been relativized in different ways for different purposes, mostly in connection with word frequency data.

For the sake of easy manipulability one can consider the *mean arc length* defined as

$$\bar{L} = \frac{L}{n-1} \tag{2.2}$$

where n is the number of units in the sequence (n-1 is the number of individual arcs). Since the variance can be computed empirically, there is no problem with text comparisons, confidence intervals, etc. In order to illustrate the problem we consider the sequence of syllabic lengths of the verses in the poem *Der Erlkönig* (E) by Goethe and obtain:

$$E_1 = (8,7,8,8,9,6,6,6,7,7,7,6,8,5,6,6,7,6,6,8,9,5,8,7,8,9,9,6,6,7,7,7). \quad (2.3)$$

The elements represent the numbers of syllables in the individual verses, n = 32. The values do not vary strongly, because in the poem verse length is rather stereotype, predetermined, nevertheless, there is some oscillation within the interval [5,9]. Using the definition (2.1), we obtain the arc length

$$L = [(8-7)^2 + 1]^{1/2} + [(7-8)^2 + 1]^{1/2} + [(8-8)^2 + 1]^{1/2} + \ldots + [(7-7)^2 + 1]^{1/2} + + [(7-7)^2 + 1]^{1/2} = 50.6291$$

hence the mean (according to 2.2) is

$$\bar{L} = 50.6291/31 = 1.6332.$$

The variance of L_i from their mean is $\frac{1}{31}\sum_{i=1}^{31}(L_i - \bar{L})^2 = 21.3126$. According to the Central Limit Theorem we now obtain the variance of $\bar{L}$ as Var($\bar{L}$) = $Var(L)/n$ = 21.3126/31 = 0.6875, i.e. the variance of the mean is equal to the variance of individual arc lengths divided by n..

If all elements of the sequence are equal, we obtain $\bar{L}$ = 1. Here we see that only a small part of the sequential differences add something to the mean arc. Hence, the sequence is rather smooth.

Consider now the indicator

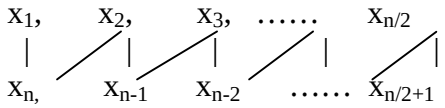
$$P = \frac{L - L_{min}}{1 + L_{mx} - L_{min}} \quad (2.4)$$

Where L_{max} is the maximal arc obtained under the given empirical conditions and L_{min} the minimal one. The minimal arc can be computed if we reorder the sequence in increasing (or decreasing) order. For our example it would be

$$E_2 = (5,5,6,6,6,6,6,6,6,6,6,6,7,7,7,7,7,7,7,7,7,8,8,8,8,8,8,8,8,9,9,9,9) \quad (2.5)$$

yielding $L_{min} = 27(1) + 4\sqrt{2} = 32.6569$.

In order to compute L_{max} one can proceed as follows: Order the numbers in non-decreasing order; if there are n elements (n being even), place the elements in the first row from x_1 to $x_{n/2}$, and the elements from the n^{th} to $n/2+1^{st}$, i.e. in reversed order in the second row. One obtains



(2.6)

Compute the arcs between the connected elements and add them to obtain a rough estimation of L_{max}. If the number of elements is odd, let the first row contain one element more, i.e. the last element in the first row is $x_{(n-1)/2+1}$ and the first element in the second row is again x_n. One computes also the distance between $x_{(n-1)/2+1}$ (which is the last element in the first row) and $x_{(n-1)/2}$ which is the last element in the second row.

This computation can easily be programmed and performed among others with Excel, for any text length. It is not necessary to check the individual permutations.

In our above example we obtain

5,5,6,6,6,6,6,6,6,6,6,6,7,7,7,7,
9,9,9,9,8,8,8,8,8,8,8,7,7,7,7,7,

Thus the maximum arc will be

$[(5-9)^2 + 1]^{1/2} + [(9-5)^2 + 1]^{1/2} + [(5-9)^2 + 1]^{1/2} + [(9-6)^2 + 1]^{1/2} + [(6-9)^2 + 1]^{1/2} + [(9-6)^2 + 1]^{1/2} + [(6-9)^2 + 1]^{1/2} + [(9-6)^2 + 1]^{1/2} + [(6-8)^2 + 1]^{1/2} + [(8-6)^2 + 1]^{1/2} + [(6-8)^2 + 1]^{1/2} + [(8-6)^2 + 1]^{1/2} + [(6-8)^2 + 1]^{1/2} + [(8-6)^2 + 1]^{1/2} + [(6-8)^2 + 1]^{1/2} + [(8-6)^2 + 1]^{1/2} + [(6-8)^2 + 1]^{1/2} + [(8-6)^2 + 1]^{1/2} + [(6-8)^2 + 1]^{1/2} + [(8-6)^2 + 1]^{1/2} + [(6-7)^2 + 1]^{1/2} + [(7-7)^2 + 1]^{1/2} + [(7-7)^2 + 1]^{1/2} + [(7-7)^2 + 1]^{1/2} + [(7-7)^2 + 1]^{1/2} + [(7-7)^2 + 1]^{1/2} + [(7-7)^2 + 1]^{1/2} + [(7-7)^2 + 1]^{1/2} + [(7-7)^2 + 1]^{1/2} = 3(4.1231) + 5(3.1623) + 14(2.2361) + 1.4142 + 8 = 68.8999.$

The above calculations can be summarized by the following formula:

$$L_{\max} = \begin{cases} \sum_{i=1}^{n/2} \sqrt{(x_i - x_{n+1-i})^2 + 1} + \sum_{i=1}^{n/2-1} \sqrt{(x_{i+1} - x_{n+1-i})^2 + 1} & \textit{for even } n \\ \sum_{i=1}^{(n-1)/2} \sqrt{(x_i - x_{n+1-i})^2 + 1} + \sum_{i=1}^{(n-1)/2} \sqrt{(x_{i+1} - x_{n+1-i})^2 + 1} & \textit{for odd } n \end{cases} \quad (2.7)$$

If e.g. n is even, the elements in the first and second sum correspond to the vertical and diagonal line segments in diagram (2.6), respectively.

The indicator P in (1.3) becomes P = (50.6291 – 32.6569)/(1 + 68.8999 – 32.6569) = 0.4826, hence we would consider the sequence as rather smooth.

Consider the behaviour of the indicator P. If all values of the sequence are equal, e.g. 2,2,2,2,2,…, then all values (L, L_{max} and L_{min}) are equal and we obtain 0/1 = 0. This sequence is extremely smooth. Here one sees why 1 has been inserted in the denominator: without 1 we would obtain 0/0 which is no definite value.

Now, take a maximally rough sequence in which there are only minimal and maximal values in regular succession, e.g. 1,10,1,10,1,10,…. In that case $L = L_{max}$ and $L_{max} - L_{min}$ is some increasing function of n, say = $k(n)$. Hence $P = k/(1+k)$. Taking the limit for $n \rightarrow \infty$ we obtain P = 1. Hence P is always between 0 and 1.

In the present article we shall consider sequences of word length, sentence length, and frequency, in different extent. The words will be replaced by their topical property and the smoothness/roughness of the sequence will be computed. Each computation will be accompanied by a hypothesis. Below, we add some further methods capturing the behaviour of the sequence.

2.1. Word length

Word length is the most frequent object of quantitative investigations because usually the data are readily available and one does not need determine the boundaries of syllables whose number represents the word length. Nevertheless, even here problems may arise e.g. with diphthongs, triphthongs. In some languages one counts also non-syllabic words or one considers them as clitics of the preceding or next word (cf. in Slavic languages the non-syllabic prepositions *s*, *z*, *v*, in Hungarian the conjunction *s* being the elliptic form of *és*, etc.). In strongly analytic languages the oscillation may be relatively small because words are not prolonged by affixation; as compared to this, in strongly synthetic languages, whether inflectional or agglutinative, words can attain a greater length and the irregular oscillation may be stronger. Hence we can preliminarily conjecture that the greater the word length roughness in text, the stronger is the synthetism of language.

Of course, the conjecture can be formulated in reverse order because it is not roughness which is the cause of synthetism but just conversely. We do not speak about causes in language but rather about links between properties, as is usual in dynamic systems. If we would consider synthetism as the independent variable, we would be forced to measure it in some way. Hence the above statement is merely a conjecture that can be tested.

In order to normalize P, and at the same time to perform the test for the significance of the smoothness/rougness of the data, we need its expectation and variance. We assume preliminarily that the expectation is 0.5, the mid of the range of P, because we have to do with very different entities. The variance is given as follows: for the given text L_{max} and L_{min} are some fixed values, hence

$$Var(P) = \frac{Var(L)}{(1 + L_{max} - L_{min})^2} \tag{2.8}$$

and the normalization yields

$$u = \frac{P - 0.5}{\sqrt{Var(P)}} \tag{2.9}$$

For illustration we compute (2.8) for the above mentioned verse length data and obtain

$P = 0.4826$,
$Var(L) = 21.3125$,
$Var(P) = 21.3125/(1 + 68.8999 - 32.6569)^2 = 0.01537$,
hence

$u = (0.4826 - 0.5)/\sqrt{0.01537} = -0.1404.$

We state that the text is smoother than expected ($u < 0$) but not significantly.

The respective values for 60 texts in 28 languages are presented in Table 2.2.

As can be seen, all word length sequences have a rather great roughness (viz. significantly positive u). One could order the texts according to u but there are great differences even within individual languages, referring rather to text size, text sort or stylistic differences, not to language type. Hence taking averages of P seems to be more adequate. However, at present, we cannot take averages in all languages because the number of texts in some of them is merely 1. Nevertheless, a preliminary order can be stated. Using Table 2.2 and taking the averages we obtain (in parenthesis is the number of analysed texts) the order presented in Table 2.1.

As can be seen, the conjecture concerning word length smoothness and synthetism can preliminarily be accepted: the smaller P, the stronger is the synthetism. However, further collecting of data is necessary. The result is very preliminary and shows merely the method. Nevertheless, one question remains open: why in some languages the regularity is greater than in other ones, i.e. what is the background mechanism cotrolling the *sequences* of word length?

Evidently, the languages are only partially ordered according to their degree of analytism/synthetism.

But even if the P-values seem to be quite similar, there may be significant differences between texts or languages. In order to state them, we simply perform the t-test for the difference of two means omitting languages in which only one text has been processed. The t-test with $n_1 + n_2 - 2$ degrees of freedom can be computed according to the formula

$$t = \frac{\bar{P}_1 - \bar{P}_2}{\hat{\sigma}_{\bar{P}_1 - \bar{P}_2}} \tag{2.10}$$

where

$$\hat{\sigma}_{\bar{P}_1 - \bar{P}_2} = \sqrt{\frac{\sum_{i=1}^{n_1}(P_{i1} - \bar{P}_1)^2 + \sum_{i=1}^{n_2}(P_{i2} - \bar{P}_2)^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{2.11}$$

and n_1, n_2 are the numbers of cases in the given language. The probability that t is greater or equal to the observed value can be found in the respective tables.

For the sake of illustration we perform the test for Slovak with $\bar{P}_{Sl}$ = 0.6135, sum of squared deviations = 0.000316, n = 2 and Sundanese with $\bar{P}_{Su}$ = 0.7097, sum of squared deviation = 0.000098, n = 2, yielding

$$t = \frac{|0.6135 - 0.7097|}{\sqrt{\frac{0.000316 + 0.000098}{2+2-2}\left(\frac{1}{2}+\frac{1}{2}\right)}} = 6.69$$

which is significant with 2 degrees of freedom (Sundanese has greater roughness than Slovak, the critical value is $t_{0.05,2} = 4.30$). The results will become more clear when the number of texts increases; in this form it is not quite true that Sundanese is more synthetic than Slovak.

Table 2.1
Oscillation of the word length arc in languages
(ordered by ascending mean of P)

Language	**mean P ascending**
Russian(1)	0,5907
Tamil(1)	0,5997
Slovenian(1)	0,6111
Slovak(2)	0,6135
Czech(5)	0,6137
Telugu(2)	0,6193
Malayalam(2)	0,6201
Latin(2)	0,6236
Indonesian(2)	0,6305
Serbian(2)	0,6444
Odia(2)	0,6475
Maninka(3)	0,6534
Hindi(2)	0,6582
German(5)	0,6675
Bulgarian(1)	0,6720
Hungarian(2)	0,6768
Welsh(2)	0,6777
Vai(3)	0,6799
Romanian(3)	0,6819
French(1)	0,6823
Macedonian(1)	0,7062
Italian(1)	0,7083
Sundanese(2)	0,7097
Bamana(4)	0,7247
Japanese(1)	0,7384
Kikongo(3)	0,7522
Akan(2)	0,7556
Tagalog(2)	0,7651

Table 2.2
Roughness in 60 texts from 28 languages

Language: Text	n	L	L_{min}	L_{max}	P	Var(L)	Var(P)	u
Akan: Agya Yaw Ne Akutu Kwaa	201	290,47	201,24	323,55	0,7236	0,2957	1,94E-05	50,71
Akan: Mma Nnsua Ade B□ne	143	218,62	143,66	237,84	0,7876	0,4536	5,01E-05	40,64
Bamana: Bamak□ sigicoya	1138	1739,97	1139,07	1966,32	0,7255	0,4910	7,16E-07	266,55
Bamana: Masadennin	2616	4057,15	2617,9	4559,65	0,7408	0,7347	1,95E-07	545,85
Bamana: Namak□r□ba halakilen	1406	1890,23	1406,66	2118,59	0,6783	0,2662	5,24E-07	246,35
Bamana: Sonsannin ani	2392	3615,73	2393,49	4013,4	0,7540	0,6195	2,36E-07	523,18
Bulgarian: Ostrovskij, Kak se kaljavaše stomanata (Chap. 1)	926	1644,85	927,07	1994,13	0,6720	0,5839	5,12E-07	240,47
Czech: Čulík, O čem jsou dnešní Spojené státy?	2003	3633,34	2006,14	4597,83	0,6276	0,8257	1,23E-07	364,10
Czech: Hvížďala, O předem zpackané prezidentské volbě	929	1615,07	930,49	2058,75	0,6062	0,6209	4,87E-07	252,23
Czech: Macháček, Slovenský dobrý příklad	340	599,02	341,07	751,56	0,6269	0,6063	3,58E-06	67,05
Czech: Spurný, Prekvapení v justici	288	499,95	289,07	632,99	0,6114	0,5822	4,89E-06	50,35
Czech: Švehla, Editorial, Voličův kalkul	288	482,14	291,23	610,38	0,5963	0,6727	6,56E-06	37,60
French: Dunkerque – La route des dunes (press)	1532	2558,39	1533,9	3034,36	0,6823	0,702	3,11E-07	326,74
German: Assads Familiendiktatur	1415	2587,27	1417,73	3228,75	0,6454	1,1336	3,45E-07	247,51
German: ATT0012 (press)	1148	2157,49	1150,31	2660,01	0,6667	1,2753	5,59E-07	223,00
German: Die Stadt des Schweigens	1567	2871,06	1569,73	3502,33	0,6730	1,1681	3,12E-07	309,52
German: Terror in Ost Timor	1398	2475,4	1400,31	2972,87	0,6832	0,9547	3,86E-07	295,07
German: Unter Hackern und Nobelpreisen	1363	2558,37	1365,31	3147,71	0,6690	1,2029	3,78E-07	274,77
Hindi: After the sanctionto love marriage (press)	1103	1648,69	1103,66	1895,7	0,6873	0,3134	4,98E-07	265,29
Hindi: The Anna Team on a cross-road (press)	860	1212,27	860,66	1418,53	0,6291	0,2214	7,09E-07	153,39
Hungarian: A nominalizmus Forradalma (press)	1314	2841,26	1316,31	3679,39	0,6451	1,3821	2,47E-07	291,68
Hungarian: Kunczekolbász (press)	458	1016,71	460,31	1244,57	0,7086	1,5538	2,52E-06	131,38
Indonesian: Pengurus PSM terbelah (press)	345	537,8	346,07	656,5	0,6156	0,3785	3,90E-06	58,54
Indonesian: Sekolah ditutup (press)	280	456,16	281,07	551,39	0,6453	0,463	6,29E-06	57,95

Italian: Il bosone di Higgs scoperto dal Cern (Internet)	2516	4974,2	2518,31	5984,39	0,7083	1,0795	8,98E-08	695,24
Japanese: Miki, Jinseiron Note	1805	3483,94	1809,06	4076,37	0,7384	1,1911	2,31E-07	495,45
Kikongo: Bimpa: Ma Ngo ya Ma Nsiese	823	1397,12	829,15	1621,03	0,7163	0,6927	1,10E-06	206,10
Kikongo: Lumumba speech	956	1557,76	957,07	1759,73	0,7474	0,4283	6,63E-07	303,88
Kikongo: Nkongo ye Kisi Kongo	768	1139,63	769,07	1235,41	0,7929	0,3269	1,50E-06	239,42
Latin: Cicero, In Catilinam I	1116	1948,25	1117,49	2470,53	0,6135	0,5908	3,22E-07	200,02
Latin: Cicero, In Catilinam II	3095	5632,4	3096,9	7097,23	0,6337	0,6637	4,15E-08	656,50
Macedonian: Ostrovskij, Kako se kaleše čelkiot (Chap. 1)	1123	2251,33	1124,07	2719,36	0,7062	0,8963	3,52E-07	347,63
Malayalam: Moralistic hooligans (press)	282	594,31	284,31	790,19	0,6116	1,3026	5,07E-06	49,56
Malayalam: No one should die (press)	288	668,42	290,31	890,86	0,6286	1,8581	5,13E-06	56,73
Maninka: Nko Doumbu Kende no.2 (press)	2076	3132,96	2077,07	3877,8	0,5860	0,4379	1,35E-07	234,27
Maninka: Nko Doumbu Kende no.7 (press)	1535	2394,57	1536,49	2814,04	0,6711	0,4746	2,90E-07	317,61
Maninka: Sìikán` (Constitution of Guinea, an excerpt)	1662	2950,33	1663,07	3492,93	0,7031	0,7119	2,12E-07	440,69
Odia: The Samaj, Bhuba-neshwar (28 June 2012), p. 4	348	549,27	349,49	662,51	0,6362	0,5059	5,13E-06	60,13
Odia: The Dharitri, Balasore (12th Feb, 2012), p. 10	630	1084,28	631,49	1317,72	0,6589	0,5642	1,19E-06	145,35
Romanian: Paler, excerpt from Aventuri solitare	891	1681,02	892,49	2008,51	0,7059	0,7763	6,22E-07	261,07
Romanian: Steinhardt, Jurnalul fericirii, Trei soluții	1511	2718,28	1512,49	3361,98	0,6516	0,8189	2,39E-07	310,02
Romanian: Popescu D.R., Vânătoarea regală	1006	1664,39	1007,07	1961,26	0,6882	0,5061	5,55E-07	252,63
Russian: Ostrovskij, Kak zakaljalas stal' (Chap. 1)	792	1319,67	793,49	1683,19	0,5907	0,5251	6,62E-07	111,55
Serbian: Ostrovskij, Kako se kalio čelik (Chap. 1,	994	1675,53	994,66	2050,33	0,6444	0,5376	4,81E-07	208,04
Slovak: Bachletová, Moja Dolná zem	872	1435,26	873,07	1770,07	0,6260	0,4523	5,61E-07	168,30
Slovak: Bachletová, Riadok v tlačive	924	1655,59	925,49	2139,51	0,6009	0,696	4,71E-07	146,94
Slovenian: Ostrovskij, Kako se je kalilo jeklo (Chap. 1)	977	1556,7	978,07	1923,95	0,6111	0,4304	4,80E-07	160,34
Sundanese: Agustusan (Online)	416	664,27	417,07	761,00	0,7167	0,438	3,68E-06	112,92
Sundanese: Aki Satimi (Online)	1283	2011,23	1283,66	2318,1	0,7027	0,3508	3,27E-07	354,31
Tagalog: Rosales, Kristal Na Tubig	1958	3794,27	1959,9	4309,66	0,7803	0,8705	1,58E-07	706,31
Tagalog: Hernandez, Limang Alas: Tatlong Santo	1738	3238,09	1739,9	3740,27	0,7486	0,7971	1,99E-07	557,24
Tagalog: Hernandez, Magpisan	1466	2838,63	1467,9	3255,29	0,7665	0,8667	2,71E-07	511,87
Tamil: Emu Bird Trading (press)	384	771,84	386,31	1028,18	0,5997	1,1346	2,75E-06	60,17

Telugu: Trailangaswamy (press)	295	616,92	297,31	810,27	0,6219	1,2091	4,58E-06	56,96
Telugu: Train Journey (press)	666	1299,12	668,73	1689,71	0,6168	1,1084	1,06E-06	113,41
Vai: Sa'bu Mu'a'	495	631,4	495,24	716,62	0,6123	0,1275	2,58E-06	69,93
Vai: Sherman, Mu ja vaa	3140	4079,9	3140,66	4579,51	0,6523	0,1571	7,58E-08	553,32
Vai: Vande	426	571,29	426,24	612,39	0,7750	0,1648	4,71E-06	126,80
Welsh: text 1 (gaenv)	985	1750,5	986,49	2094,86	0,6887	0,5934	4,82E-07	271,74
Welsh: text 2 (gasodl)	1002	1441,3	1002,66	1659,57	0,6667	0,2445	5,65E-07	221,82

The basic data were taken from Popescu et al. (2013) with the kind permission of E. Kelih, A. Rovenchak, A. Overbeck, H. Sanada, R. Smith, R. Čech, P. Mohanty and A. Wilson.

The results of pair-wise testing are presented in Table 2.3. As can be seen, there is a number of significant differences (grey). Though the number of used texts is very small, we can conjecture that not only texts may differ but also languages. A thorough investigation should be performed in such a way that first several texts of the same sort in one language must be analyzed, then the mean *P* could be used for comparison with another text sort in the same language, etc. This is a practically endless problem.

In our data (cf. Table 2.3), Tagalog seems to be an outlier among the other Indonesian languages (Malay, Sundanese), the group of Indo-European languages disintegrates, etc. But all these statements are merely conjectures that must be thoroughly tested. A classification of the results in Table 2.3 could be performed with the aid of many clustering methods but it is not our aim here.

Table 2.3
Testing the difference of mean $P(u)$ using the t-test

Language	Slo	Cze	Tel	Mal	Lat	Ind	Odi	Man	Hin	Ger	Hun	Wel	Vai	Rom	Sun	Bam	Kik	Aka
Czech 5	0,02																	
Telugu 2	0,46	0,55																
Malayalam 2	0,44	0,58	0,08															
Latin 2	0,63	0,87	0,41	0,27														
Indonesian 2	0,87	1,31	0,74	0,61	0,38													
Odia 2	2,01	2,88	2,43	1,94	1,58	0,91												
Maninka 3	0,87	1,49	0,76	0,73	0,65	0,49	0,13											
Hindi 2	1,41	2,42	1,33	1,26	1,13	0,85	0,34	0,10										
German 5	4,38	6,21	4,60	4,19	3,76	2,84	1,66	0,52	0,50									
Hungarian 2	1,85	3,22	1,80	1,73	1,60	1,32	0,87	0,46	0,43	0,47								
Welsh 2	3,85	5,49	5,18	4,15	3,64	2,56	1,91	0,53	0,63	0,86	0,03							
Vai 3	1,04	1,81	0,96	0,94	0,88	0,77	0,51	0,44	0,32	0,34	0,05	0,03						
Romanian 3	3,02	4,81	3,02	2,86	2,65	2,19	1,54	0,74	0,79	1,01	0,16	0,19	0,04					
Sundanese 2	6,69	8,90	12,12	8,15	7,03	4,82	4,66	1,24	1,72	3,83	1,01	2,46	0,47	1,30				
Bamana 4	4,28	6,91	4,24	4,13	3,96	3,57	3,00	2,03	2,18	3,55	1,52	1,83	0,99	1,81	0,60			
Kikongo 3	4,60	7,64	4,62	4,50	4,34	3,96	3,50	2,39	2,61	4,65	2,03	2,50	1,35	2,57	1,46	1,02		
Akan 2	4,14	7,20	4,25	4,10	3,94	3,55	3,19	2,00	2,25	4,44	1,75	2,30	1,12	2,34	1,40	0,98	0,09	
Tagalog 2	8,99	12,90	12,64	10,30	9,38	7,23	7,36	2,44	3,43	8,17	2,62	5,56	1,34	3,74	4,19	1,57	0,43	0,28

2.2. Sentence length (style)

Sentence length measured in terms of word numbers does not depend on language but on the communicative aim, on the style of the author, on text sort, on the age of the author, on the spontaneity of writing, etc. One can suppose that in text books for children the sentences are rather short; the same holds for poetry but not for prose where a sentence can consist of several hundreds of words. If no form restricts the writer - e.g. as in poetry, stage play, text-book, science, law, etc. - and (s)he writes spontaneously, the sentences can get longer. Punctuation did not exist from the beginning of writing, since it was introduced at a later time. Speech does not contain punctuation, sometimes one speaks a long time without any pause. This does not hold for stage plays where even monologues contain punctuation.

Thus we can set up several hypotheses concerning the roughness of sentence length and test them.

(1) The greater is the sentence length roughness, the more spontaneously the text has been written.

(2) A preliminary hypothesis concerning individual text sorts may be set up as follows: *text books for children* < *journalistic texts* < *poetry* < *stage play* < *law texts* < *scientific texts* < *prose.*

In the above ordering of text sorts, some of them are rather persistent than volatile, but the turning point is not yet known. Its finding is a matter of extensive testing.

Table 2.4 contains some computation of the indicator *P* in 15 texts of 9 languages. The order of languages does not correspond to the degree of their synthetism, hence *P* is a property of the given text. Texts in which *P* is maller than 0.5 have a certain „sentence rhythm".

A slightly more complex computation of sentence lengths may be performed in terms of clause numbers. However, the stating of the number of clauses cannot be made based on a general rule – which does not exist -, it is a problem of definition which may differ from language to language (and from linguist to linguist). Since we analyze only German data, we define clause as a construction containing a finite form of a verb, also auxiliary and modal. Hence the clause length of a sentence is simply the number of finite verbs in it. We used 20 German newspaper articles published in January 1999 in *Tageszeitung*, namely:

T1 Taz 22.01.1999: Kulturstadt Weimar: Winter mit fröhlicher Sonne.
T2 Taz 21.01.1999: Wer macht das Spiel?
T3 Taz 20.01.1999: Die Nerven behalten
T4 Taz 16.01.1999: Auf Elefanten Richtung Rhino!
T5 Taz 16.01.1999: Der Gefangene von Gaghan
T6 Taz 16.01.1999: Zeitschriften sind Originale
T7 Taz 09.01.1999: Die Friedhöfe an der Drina
T8 Taz 15.01.1999: Zwischen Finanzkrise und Handelskrieg
T9 Taz 15.01.1999: Das belgische Modell einer präventiven Ausländerfeindlichkeit
T10 Taz 15.01.1999: Zweierlei Recht im chilenischen Rechtsstaat
T11 Taz 15.01.1999: Frankreichs Front National entdeckt Osteuropa
T12 Taz 15.01.1999: Für einen neuen Stabilitätspakt
T13 Taz 15.01.1999: Kontrolle der Kapitalströme emerging markets - ein Gebot der Demokratie

Table 2.4
The arc of sentence length in 15 texts (in terms of word numbers)

Language	**Text**	**n**	**L**	**L_{max}**	**L_{min}**	**P**	**Var(L)**	**Var(P)**	**u**
Hungarian	A nominalizmus forradalma, press	63	703,42	982,87	96,73	0,6839	68,7281	8,73E-05	19.68
Indonesian	Pengurus, press	28	126,27	235,78	41,5	0,4341	13,481	3,54E-04	-3.51
Latin	Cicero, In Catilinam I	80	720,22	1225,11	101,36	0,5502	52,5753	4,16E-05	7.79
Romanian	Octavian Paler, Aventuri solitare (excerpt)	17	185,32	397,02	46,94	0,3942	89,365	7,25E-04	-3.93
Romanian	D.R. Popescu, Vânătoarea regală, Chapter 2	61	973,26	1367,43	157,99	0,6735	366,0799	2,50E-04	10.98
Romanian	N. Steinhardt, Jurnalul fericirii, Trei soluții	85	1260,5	1810,47	189,04	0,6604	446,615	1,70E-04	12.31
Russian	N. Ostrovskij, How the steel was tempered	76	605,21	840,72	108,33	0,6775	70,2603	1,31E-04	15.53
Slovak	Bachletová, Moja Dolna zem	92	617,45	892,5	118,38	0,6439	47,0381	7,83E-05	16.26
Slovak	Bachletová, Riadok v tlačive	78	641,52	836,31	92,12	0,7373	41,5035	7,47E-05	27.44
Slovenian	N. Ostrovskij, How the steel was tempered	84	754,91	1047,81	117,84	0,6843	79,5594	9,18E-05	19.24
Sundanese	Aki Satimi, press	147	673,99	1014,63	157,27	0,6020	10,5146	1,43E-05	27.00
Sundanese	Agustusan, press	53	209,97	280,28	57,38	0,6815	5,749	1,15E-04	16.95
Tagalog	Hernandez, Limang Alas, Tatlong Santo	104	894,34	1411,56	124,39	0,5977	44,6326	2,69E-05	18.84
Tagalog	Hernandez, Magpisan	111	965,66	1397,85	132,3	0,6580	48,1819	3,00E-05	28.83
Tagalog	Rosales, Kristal Na Tubig	139	1042,84	1717,3	171,91	0,5632	42,8283	1,79E-05	14.93

T14 Taz 13.01.1999: Für uns Serben wird hier kein Platz sein
T15 Taz 11.01.1999: Von Frust und Lust im samtenen Sweat-shop
T16 Taz 15.01.1999: Ethnische Definitionen als Machtpolitik
T17 Taz 15.01.1999: Bündnisse und Rivalitäten im Mittleren Afrika
T18 Taz 14.01.1999: Klick, klick, klick.
T19 Taz 11.01.1999: Wo Es war, soll Wir werden
T20 Taz 23.01.1999: Die Schlagerfamilie Ost.

For the German clause-variant (Table 2.5) we obtain a mean $P = 0.6086$. For the word-variant (Table 2.4) in a mixture of several languages it holds $P = 0.6161$. If these results turn out to be relatively constant, then we have found a very peculiar stability which may hold in the entire hierarchy (sentence – clause – word – syllable – sound duration). In order to state it, Sisyphean work is necessary. For the time being, we can state that this value tends to the value of the golden section minus 1 (0.6180)

Table 2.5
Arc of sentence lengths (in terms of clauses) in 20 German texts

Text	**n**	**L**	**L_{min}**	**L_{max}**	**P**	**Var(L)**	**Var(P)**	**u**
T1	148	226,9328	149,8929	294,5745	0,5288	0,5525	2,60E-05	5,65
T2	80	129,4372	81,4853	156,7301	0,6289	0.6184	1,06E-04	12,54
T3	112	188,1304	113,0711	220,2935	0,6936	0,6104	5,21E-05	26,81
T4	208	332,1408	209,4853	397,1550	0,6501	0,6508	1,83E-05	35,11
T5	246	361,5297	247,4853	437,9837	0,5955	0,4283	1,17E-05	27,95
T6	109	191,3951	110,8995	223,4260	0,7090	0,9889	7,67E-05	23,86
T7	107	176,4568	108,4853	223,7354	0,5847	0,7643	5,66E-05	11,26
T8	85	150,2687	86,0711	180,0518	0,6759	0,7251	8,04E-05	19,62
T9	97	148,8435	98,0711	188,5416	0,5551	0,4445	5,31E-05	7,56
T10	112	181,3166	113,0711	219,1969	0,6371	0,5075	4,42E-05	20,61
T11	95	137,2944	95,6569	176,1045	0,5112	0,2739	4,13E-05	1,75
T12	74	106,0632	74,6569	125,0160	0,6115	0,3458	1,31E-04	9,74
T13	120	170,7611	120,6569	213,9045	0,5316	0,2878	3,24E-05	5,56
T4	139	191,7779	140,0711	216,7953	0,6653	0,2736	4,53E-05	24,56
T15	105	165,7787	106,0711	201,4551	0,6195	0,4344	4,68E-05	17,47
T16	119	186,0022	120,8929	243,8176	0,5254	0,4684	3,05E-05	4,60
T17	110	163,2036	111,0711	204,6998	0,5509	0,4133	4,62E-05	7,49
T18	197	274,6528	197,6569	322,4205	0,6122	0,3032	1,92E-05	25,63
T19	79	131,3901	80,0711	163,9010	0,6050	0,5153	7,16E-05	12,40
T20	151	211,0678	152,0711	237,7794	0,6804	0,3759	5,00E-05	25,52

2.3. Frequency

The computation of frequency is simple, there is software available to perform this procedure. Having the frequencies of individual units (word forms or lemmas), we replace the respective words in the text by their frequency and compute the properties of the constructed sequence. We may start from two considerations: If there are many forms in language, then the number of seldom word-forms will be greater, many of them are placed in immediate neighbourhood, hence the arc will be smaller. This will be the case also with short texts. At the same time there are some few words (usually synsemantics) which occur quite frequently, hence the maximum arc will be longer and the minimum arc shorter. Hence P will be smaller.

Consider under these presuppositions 20 Slovak texts. Here we obtain the results as given in Table 2.6. The texts have different sizes which do not influence the value of P. The shortest text (n = 93) has P = 0.9319 while the longest text (n = 3704) has P = 0.8749. The texts were taken from http://quanta-textdata.uni-graz.at in 5 different text sorts as given in the following list.

Author	**Text sort**	**Text**	
1.anonymous	Agency (TASR)	Banskobystrický samosprávny kraj sa vzdáva zdravotníckych zariadení	03.11.2004
2.anonymous	Agency (TASR)	Bratislavskí taxikári zvyšujú poplatok za nástupenie o 100 %	09.11.2004
3.D. Dušek	Short story	1. máj 1977. In: *Kufor na sny*	1993
4.D. Dušek	Short story	Alfabet D.D.: *Kufor na sny*	1993
5.D. Hevier	Fairy tale	Kotrmelec a kotrmelec. In: *Futbal s papučou*	1989
6.D. Hevier	Fairy tale	Lyžicová Naháňačka. In: *Futbal s papučou*	1989
7.anonymous	Agency (TASR)	Čoskoro rozhodnú o investorovi pre SND	03.11.2004
8.Eugen Č.	Agency (TASR)	Ďakujem všetkým, ktorí nám dôverovali, verím v spravodlivosť	03.11.2004
9.B. Hochel	Short story	Kúpalisko	1997
10.B. Hochel	Short story	Muž, ktorému veľmi ukrivdili	1997
11.D. Hevier	Fairy tale	Krajina agord, Kap. 1, A	2001
12.D. Hevier	Fairy tale	Krajina agord, Kap. 1, B	2001
13.P. Holka	Short story	Kap. 1. *Normálny cvok*	1993
14.P. Holka	Short story	Kap. 2	1993
15.V. Šikula	Novel	Kap. 1. *Veterna ružica*	1995
16.V. Šikula	Novel	Kap. 2. *Veterna ružica*	1995
17.R. Sloboda	Novel	Kap. 1. *Pamäti*	1996
18.R. Sloboda	Novel	Kap. 2. *Pamäti*	1996
19.P. Vilikovský	Short story	Kap. 1. *Peší príbeh*	1992
20.P. Vilikovský	Short story	Kap. 2. *Peší príbeh*	1992

For all texts the value P deviates significantly from 0.5, and the value of u increases with the text size. The first three shortest texts are exceptions.

The mean values of P for the same text sorts are:

News = 0.7272 < *Fairy tale* = 0.7704 < *Short story* = 0.7939 < *Novel* = 0.82

As can be seen there is a clear hierarchy which can be left to literary scientists for interpretation. One could perform tests for the difference between text sorts but we are preliminarily content with this hierarchy.

The mean of all (n = 20) Slovak values is $\overline{P}_{Slk}$ = 0.7811, and the variance of the mean is $Var(\overline{P}_{Slk})$ = 0.0007125. The mean P is an indicator of written Slovak and since its empirical variance is known, it can be used for inter-language comparisons (s. below). The value of u increases with increasing n but there are some outliers which must be studied separately.

Table 2.6
The *P*-indicator in Slovak frequency data
(ordered according to increasing n)
(FT = fairy tale, NW = news, SS = short story, NO = novel)

Text sort	n	L	L_{max}	L_{min}	P	Var(L)	Var(P)	u
6. FT	95	145,74	160,28	95,24	0,7647	0,6134	1,41E-04	22,32
5. FT	123	288,48	386,64	125,82	0,6213	2,1312	3,11E-05	21,75
1. NW	229	456,51	670,11	230,49	0,5130	2,1451	1,10E-05	3,90
12. FT	267	983,65	1062,06	277,54	0,8989	21,7266	3,52E-05	67,23
11. FT	283	853,57	1033,72	286,14	0,7580	6,9087	1,23E-05	73,48
2. NW	293	686,41	812,69	294,9	0,7547	3,2351	1,20E-05	73,45
8. NW	351	991,92	1134,39	355,37	0,8161	6,8133	1,12E-05	94,45
19. SS	409	1443,77	1742,34	415,67	0,7744	11,2360	6,37E-06	108,67
20. SS	428	1731,6	2077,72	436,22	0,7887	15,8988	5,89E-06	118,91
7. NW	437	1329,85	1523,52	440,55	0,8204	8,4340	7,18E-06	119,59
3. SS	793	6877,33	7967,42	823,9	0,8473	142,7026	2,80E-06	207,70
13. SS	821	6405,93	8461,62	834,34	0,7304	60,5033	1,04E-06	225,94
14. SS	980	12081,5	14474,67	1014,47	0,8221	215,4821	1,19E-06	295,41
4. SS	1044	7281,84	8532,45	1061,9	0,8325	66,9222	1,20E-06	303,67
9. SS	1132	11036,96	13043,63	1165,91	0,8310	162,2079	1,15E-06	308,70
10. SS	1143	12302,14	16553,95	1183,01	0,7233	196,3127	8,31E-07	245,03
17. NO	1486	21620,58	24365,36	1541,63	0,8797	417,6505	8,02E-07	424,07
15. NO	1605	25912,97	33647,88	1649,83	0,7582	345,4501	3,37E-07	444,61
18. NO	3666	117620,74	134684,07	3795,34	0,8696	1902,8985	1,11E-07	1109,08
16. NO	5364	250592,66	322784,07	5529,3	0,7724	2830,8441	2,81E-08	1624,55

For Russian we used 20 texts as given below (taken from http://quanta-textdata.uni-graz.at) (here SP = Stage play, SS = Short story, NW = News, PL = Private letter, NO = Novel)

For these texts we obtain the results presented in Table 2.7. As can be seen, the value of u increases with increasing n. This is simply due to the fact that normality in

language is rather an exception. Linguistic data do not want to behave "normally", in the statistical sense of the word. Nevertheless, we use the tests as a first information.

	Author	**Text sort**	**Text**	**Year**
1	A.P. Čechov	SP	Svadba	1890
2	A.P. Čechov	SP	Čajka, 1. Act	1896
3	A.P. Čechov	SP	Čajka, 2. Act	1896
4	A.P. Čechov	SP	O vrede tabaka	1902
5	A.P. Čechov	SS	Chameleon	1884
6	A.P. Čechov	SS	Chirurgija	1884
7	L.N. Tolstoj	SS	Chozjain i rabotnik, Ch. 1	1895
8	L.N. Tolstoj	SS	Chozjain i rabotnik, Ch. 2	1895
9	E. Sal'nikova	NW	Lider ottepeli - žertva zastoja	Nezavisimaja gazeta, 01.11.2001
10	E. Sal'nikova	NW	Pereput'e zamyslov	Nezavisimaja gazeta, 12.05.2001
11	Olga Tropkina	NW	Podberezkin stal vtorym kandidatom	Nezavisimaja gazeta, 03.06.2000
12	Olga Tropkina	NW	Gotovjatsja novye zakony o vyborach	Nezavisimaja gazeta, 10.02.2001
13	L.N. Tolstoj	PL	to: A.A.Tolstaja	vom 19.9.1872
14	L.N. Tolstoj	PL	to: S.A.Tolstaja	vom 8.12.1884
15	L.N. Tolstoj	PL	to: S.A.Tolstaja	vom 28.02.1892
16	L.N. Tolstoj	PL	to: Nikolaj II	vom 16.01.1902
17	F.M. Dostoevskij	NO	Prestuplenie i nakazanie, Part I, Ch. 1	1886
18	F.M. Dostoevskij	NO	Prestuplenie i nakazanie, Part I, Ch. 2	1886
19	L.N. Tolstoj	NO	Anna Karenina, Book I, Ch. 1	1877
20	I.A. Gončarov	NO	Oblomov, Part I, Ch. 1	1858

It must be remarked that the newspaper texts were written in the present millennium. They are very short and represent a different text sort. In order to get a more thorough insight probably hundreds of texts must be analyzed and their size and origin must be taken into account. Our results show merely the method.

Considering again the mean P in individual text sorts:

Stage play (1,2,3,4) = 0.8176 < *Short story* (5,6,7,8) = 0.8234 < *Private letter* (13,14,15,16) = 0.8506 < *Novel* (17,18,19,20) = 0.8549 < *Comment/News* (9,10,11,12) = 0.8662

we obtain a first ordering of smoothness of frequencies in text sorts in Russian. The mean of all values is $\overline{P}_{Rus}$ = 0.8425 and the variance of P is $Var(P_{Rus})$ = 0.002224.

Table 2.7
The *P*-indicator in 20 Russian frequency data
(ordered according to ascending n)

Text	**n**	**L**	**L_{max}**	**L_{min}**	**P**	**Var(L)**	**Var(P)**	**u**
11	220	449,28	499,47	222,82	0,8156	3,2628	4,23E-05	48,52
09	251	694,16	718,85	255,0552	0,9447	7,681	3,56E-05	74,58
12	377	1295,03	1569,05	384,64	0,7680	13,1244	9,34E-06	87,69
19	702	6767,05	8039,61	726,75	0,8259	126,6025	2,37E-06	211,82
15	778	8590,44	10157,13	808,47	0,8323	164,2835	1,88E-06	242,42
10	785	5296,11	5629,94	807,2161	0,9306	99,4081	4,27E-06	208,32
13	879	12379,97	15273,9	912,3	0,7984	233,7948	1,13E-06	280,33
05	908	7326,17	8718,24	934,5	0,8211	104,086	1,72E-06	244,98
06	949	8623,27	10038,16	980,7	0,8437	155,2042	1,89E-06	249,90
14	1097	17306,11	18895,95	1158,63	0,9103	542,5134	1,72E-06	312,48
08	1453	21457,02	27407,91	1505,3372	0,7702	341,216	5,09E-07	378,95
07	1951	40905,22	47325,23	2040,91	0,8582	956,6879	4,67E-07	524,46
16	2175	47801,08	55137,98	2280,12	0,8612	1271,3319	4,55E-07	535,44
17	2595	76014,32	87041,45	2711,7	0,8692	1654,1425	2,33E-07	765,59
04	2656	54562,01	65429,5	2739,56	0,8266	813,3295	2,07E-07	718,01
03	2657	77408,03	92082,75	2743,56	0,8357	1167,4655	1,46E-07	877,84
01	2905	67799,73	85133,84	2979,7	0,7890	799,6347	1,18E-07	839,61
02	3389	101668,83	123370,02	3491,7	0,8190	1461,5572	1,02E-07	1000,19
20	3576	105276,39	125228,55	3680,21	0,8358	1262,087	8,54E-08	1149,06
18	5604	427782,29	480689,34	5946,08	0,8886	13758,4071	6,10E-08	1572,63

As can be seen, the mean *P* of Russian is slightly greater than that of Slovak. Using the given numbers we can test the difference by (2.2). First we multiply each *Var*(*P*) by *n* = 20 in order to obtain the sum of squares of deviations. For Slovak it is 0.01425 and for Russian it is 0.04448. Hence we obtain

$$t = \frac{|0.8425 - 0.7811|}{\sqrt{\frac{0.04448 + 0.01425}{20 + 20 - 2}\left(\frac{1}{20} + \frac{1}{20}\right)}} = 1.56$$

which is not significant, i.e. these two languages belong to the same frequency-roughness class.

For Slovenian, 20 texts were taken from http://quanta-textdata.uni-graz.at, too. They are as follows (*SN* – Short novel, *LN* – Letter novel, *OL* – Open letter, *Sc* = Science, *Hi* = History, *NS* = News).

	Autor	Text sort	Text	
1	Peter Kolšek	LN	Brina Švigelj - Visok hrib pravzaprav gora pri	1998
2	Peter Kolšek	LN	Brina Švigelj - Da od Zjutraj ko pri	1998
3	Matija Kočevar	SS	Izgubljene stvari	2001
4	Matija Kočevar	SS	Ko je vsega konec	2001
5	Milan Hladnik	Hi	Slovenska kmecka povest: In kakšen je bil konec	1990
6	Milan Hladnik	Hi	Slovenska kmecka povest: Iz česa vse je kmečka povest	1990
7	Boris Jež	NS	Kam z Jelinčičem!?	2002
8	Ivan Praprotnik	NS	Odhod z odra	2002
9	Grega Repovž	NS	Tako razlièna	2002
10	Ivan Cankar	SN	Hlapec jernej in njegova pravica, Ch. 1	1907
11	Ivan Cankar	SN	Zzodba o dveh mladih ljudeh, Ch. 1	1911
12	Fran Finžgar	SN	Strici, Ch. 1	1927
13	Fran Finžgar	SN	Strici, Ch. 2	1927
14	Nina Mazi	Sc	Tradicionalna medicina in WHO	1997
15	Dimitrij Zimsek	Sc	Klinična informatika	1997
16	Tone Škerlj	OL	Slovenski škofovski konferenci	13.03.2000
17	Tone Škerlj	OL	Mestni prostor in mestna uprava	27.01.1999
18	Tatjana Greif	OL	Odprto pismo Teološki fakulteti	2003
19	Marijan Pušavec	NO	Zbiralec nasmehov: Kral in angel	1991
20	Marijan Pušavec	NO	Zbiralec nasmehov: Zadnja škusnjava	1991

The results of computation are presented in Table 2.8. In Slovenian, the order of test sorts is

News = 0.7571 < *Open letter* = 0.7905 < *Short story* = 0.8077 < *Letter novel* = 0.8200 < *Novel* = 0.8358 < *Short novel* = 0.8540 < *Science* = 0.8629 < *History* = 0.8719.

Table 2.8
Computation of *P* in 20 Slovenian texts (ordered according to increasing *n*)

Text	**n**	**L**	**L_{max}**	**L_{min}**	**P**	**Var(L)**	**Var(P)**	**u**
T 2	225	504,82	584,06	228,37	0,7750	3,4991	2,75E-05	52,45
T 1	236	648,38	713,10	241,53	0,8609	7,6585	3,43E-05	61,63
T 16	296	1152,58	1386,18	306,13	0,7830	21,5403	1,84E-05	65,92
T 17	556	3878,35	4627,94	581,99	0,8145	103,1137	6,30E-06	125,35
T 18	563	2328,05	2843,10	574,37	0,7726	24,7419	4,80E-06	124,41
T 10	602	6980,76	8486,38	638,77	0,8080	225,9898	3,67E-06	160,83
T 9	625	4098,97	5261,43	647,16	0,7479	66,1066	3,10E-06	140,72
T 7	691	3798,75	4969,48	706,06	0,7252	33,6839	1,85E-06	165,49
T 8	775	8011,35	9835,30	810,81	0,7978	174,1371	2,14E-06	203,68

T 13	1023	26161,24	29862,88	1107,50	0,8712	1164,5566	1,41E-06	312,83
T 15	1199	9149,26	10763,63	1220,17	0,8308	81,3785	8,93E-07	349,95
T 14	1357	21099,12	23408,45	1418,94	0,8949	543,5198	1,12E-06	372,53
T 4	1578	31018,06	37409,64	1645,46	0,8213	599,4573	4,69E-07	469,29
T 20	1662	28755,15	34346,91	1702,42	0,8287	340,9218	3,20E-07	581,13
T 19	1679	26899,94	31593,85	1718,63	0,8429	361,7491	4,05E-07	538,56
T 12	1829	64957,88	73851,20	1960,11	0,8763	2489,1533	4,82E-07	542,21
T 3	2360	58583,59	73139,13	2439,42	0,7941	908,2377	1,82E-07	689,98
T 5	3304	103818,03	119803,24	3418,23	0,8626	1949,3819	1,44E-07	955,95
T 6	3554	115084,21	130131,18	3656,80	0,8810	1874,7870	1,17E-07	1112,96
T 11	4028	307502,60	356684,69	4325,71	0,8604	11578,7559	9,33E-08	1180,22

The mean P of all Slovenian texts is $\overline{P}$ = 0.8225, $Var(P)$ = 0.0021. Testing the difference according to (2.10) we obtain

t(Slovak, Russian) = 1.56
t(Slovak, Slovenian) = 3.37
t(Russian, Slovenian) = 0.003

hence only the difference between Slovak and Slovenian is significant.

3. Hurst exponent

Here we introduce another way of computing a characteristic of time series, namely the *Hurst exponent*. This measure of long time memory of time series was originally developed in hydrology and later applied to series in several other fields. It can also be applied in quantitative linguistics as shown by L. Hřebíček (2000). The Hurst exponent H takes usually values between 0 and 1, where H < 0.5 and H > 0.5 indicate volatility or the presence of a tendency, respectively. The most common method to determine H is by means of the rescaled range-statistic (*R/S*-statistic) which will be illustrated as follows.

Let us consider a time series $(x_1,x_2,\ldots,x_n)$. The *R/S* statistic consists of a sequence of values *RS(T)* which must be computed for T = 2, 3,...,n. For a given number T we define *RS(T)* as follows:

Calculate mean and standard deviation of the first T elements of the sequence by

$$\overline{x}_T = \sum_{i=1}^{T} x_i \qquad (3.1)$$

and

$$S_T = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(x_i - \overline{x}_T)^2} \quad . \qquad (3.2)$$

Calculate the sums of centralized values

$$Y_t = \sum_{i=1}^{t} (x_i - \bar{x}_T) \tag{3.3}$$

for $t = 1,...,T$ and compute

$$RS(T) = \frac{1}{S_T} \left[\max_{1 \le t \le T} Y_t - \min_{1 \le t \le T} Y_t \right] . \tag{3.4}$$

We consider again the sequence (1.2) concerning verse length in *Der Erlkönig:* (8,7,8,8,9,6,6,6,7,7,7,6,8,5,6,6,7,6,6,8,9,5,8,7,8,9,9,6,6,7,7,7) which has the length n = 32. Setting for example T = 10 we obtain from (3.1) and (3.2):

$$\bar{x}_{10} := \frac{1}{10}(8+7+8+8+9+6+6+6+7+7) = 7.2$$

and

$$S_{10} = \sqrt{\frac{1}{10}\sum_{i=1}^{10}(x_i - 7.2)^2} = \sqrt{0.96} .$$

The centralized values are

$(x_1 - 7.2, x_2 - 7.2, \ldots, x_{10} - 7.2) =$

$= (0.8, -0.2, 0.8, 0.8, 1.8, -1.2, -1.2, -1.2, -0.2, -0.2)$

and the corresponding partial sums of (3.3) are

$(Y_1, Y_2,...,Y_{10}) = (0.8, 0.6, 1.4, 2.2, 4.0, 2.8, 1.6, 0.4, 0.2, 0.0).$

Here we obtain $Y_{\max} = 4$ and $Y_{min} = 0$, hence

$$RS(10) = \frac{Y_{\max} - Y_{\min}}{S_{10}} = \frac{4-0}{\sqrt{0.96}} = 4.0825 .$$

Performing these computations for **a**ll values T = 2, 3,...,n = 32 (sequence length) yields the observed values *RS(T)* in the second column of Table 3.1. In applied sciences, one usually fits the function $(T/2)^H$ to such data, where H is the Hurst exponent. However, other possibilities have been tested, too.

The calculated values of the fitted power function are shown in the third column of Table 3.1. We obtain $RS(T) = (T/2)^{0.7981}$ with $R^2 = 0.96$, i.e. $H = 0.7981$. Since H is considerably larger than 0.5, the linguistic process can be considered as persistent. The results from Table 3.1 are plotted in Figure 3.1.

Table 3.1
Computation of *RS*(*T*) for verse length in *Der Erlkönig*

T	RS(T)	Computed		T	RS(T)	Computed
1				17	5.5817	5.518059
2	1.0000	1.000000		18	5.7540	5.775619
3	1.4142	1.382107		19	6.0276	6.030304
4	1.7321	1.738832		20	6.8070	6.282296
5	1.5811	2.077798		21	7.6449	6.531756
6	2.1213	2.403252		22	6.2201	6.778829
7	2.7217	2.717883		23	6.8343	7.023644
8	3.4412	3.023536		24	7.0259	7.266318
9	3.7741	3.321551		25	7.5842	7.506959
10	4.0825	3.612941		26	8.2256	7.745665
11	4.3708	3.898496		27	8.8205	7.982523
12	4.8054	4.178850		28	8.3816	8.217617
13	5.3533	4.454522		29	7.9786	8.451021
14	4.6771	4.725941		30	8.1011	8.682806
15	5.0196	4.993471		31	8.2218	8.913035
16	5.3405	5.257421		32	8.3408	9.141771

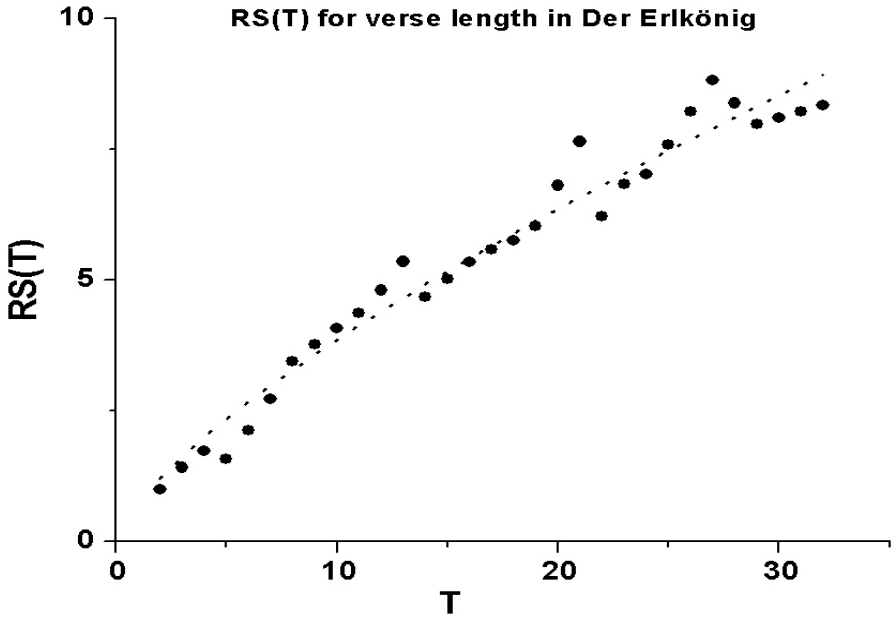


Figure 3.1. Plot of the *RS*(*T*) for verse length in *Der Erlkönig*

3.1. Word length

The computations of *H* for word length data are presented in Table 3.2.

Table 3.2
Computing the Hurst exponent for word length data

Language/Text	**n**	**H**	**R^2**	**P**
Akan: Agya Yaw Ne Akutu Kwaa	201	0,6776	0,9326	0,7296
Akan: Mma Nnsua Ade	143	0,6454	0,8296	0,7960
Bamana: Bamakɔ sigicogoya	1138	0,6439	0,9214	0,7264
Bamana: Masadennin	2616	0,6642	0,9087	0,7412
Bulgarian: Ostrovskij, Kak se kaljavaše …	926	0,7401	0,9598	0,6727
Czech: Čulík, O čem jsou dnešní Spojené státy?	2003	0,6276	0,8682	0,6279
Czech: Hvížďala, O předem zpackané prezidentské volbě	929	0,6878	0,8915	0,6068
Czech: Macháček, Slovenský dobrý příklad	340	0,7399	0,9201	0,6284
Czech: Spurný, Prekvapení v justici	288	0,5814	0,4141	0,6132
Czech: Švehla, Editorial, Voličův kalkul	288	0,7312	0,9323	0,5982
German: Assads Familiendiktatur	1415	0,6502	0,8449	0,6458
German: ATT0012 (Press)	1148	0,6350	0,9125	0,6671
German: Die Stadt des Schweigens	1567	0,6631	0,8192	0,6734
German: Terror in Ost Timor	1398	0,6719	0,8704	0,6837
German: Unter Hackern und Nobelpreisen	1363	0,6631	0,9108	0,6694
Hungarian: A Nominalizmus forradalma	1314	0,6545	0,7750	0,7095
Russian: Ostrovskij, Kak zakaljalas stal’	792	0,7233	0,9650	0,5914
Slovak: Bachletová, Moja dolná zem	872	0,6586	0,9444	0,6267
Slovak: Bachletová, Riadok v tlačive	924	0,6671	0,3282	0,6014

Evidently the Hurst exponent and the *P* indicator are not associated as shown in Figure 3.2. They represent different kinds of measurement.

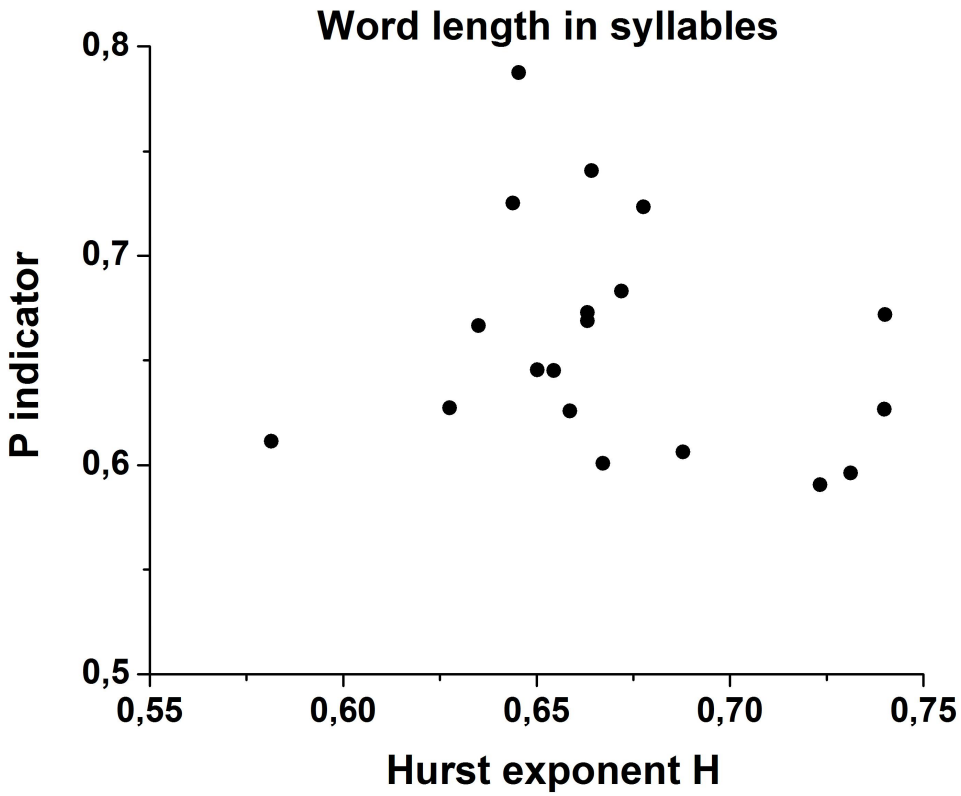


Figure 3.2. The independence of the indicators *H* and *P*

3.2. Frequencies

Let us now consider the form of the Hurst exponent in the frequency data. Besides "regular" results we obtain evident breaks in the text which can be interpreted as a loss of the internal rhythm, change of theme, a place where many corrections have been performed etc. Thus the breaks in the *RS*(*T*) curve can be interpreted textologically. For the texts in Slovak, Russian and Slovenian we obtain the results presented in Tables 3.3, 3.4 and 3.5.

Table 3.3
Hurst exponent in 20 Slovak texts

Text No	Sort	n	H	R^2
1	NW	229	fitting failed	
2	NW	293	0,6760	0,7625
3	SS	793	0,5426	0,1514
4	SS	1044	fitting failed	
5	FT	123	fitting failed	

6	FT	95	0,5597	0,6520
7	NW	437	0,5520	0,6239
8	NW	351	0,6822	0,9332
9	SS	1132	0,7171	0,9521
10	SS	1143	0,6223	0,9291
11	FT	283	0,7203	0,9353
12	FT	267	0,7079	0,9014
13	SS	821	0,6401	0,7272
14	SS	980	0,6320	0,8827
15	NO	1605	0,6409	0,8028
16	NO	5364	0,6428	0,9049
17	NO	1486	0,6101	0,9405
18	NO	3666	0,6473	0,8642
19	SS	409	0,6063	0,8371
20	SS	428	0,6429	0,6978

The mean is $\bar{H}$ = 0.6378 and the variance is *Var*(*H*) = 0.002715 (not taking into account the failed fittings, that is *n* = 17).

The ordering of text sorts is (according to their mean respective *H*) yields

Short story = 0.6262 < *Novel* = 0.6352 < *News* = 0.6367 < *Fairy tale* = 0.6626

which is, so to say, almost the reverse order as compared with *P* where we had *NS* < *FT* < *SS* < *NO*. More texts would surely lead to a better crystallisation of text sort.

As can be seen, one obtains different sequences. The first sort is monotonously increasing and can be well captured by the power function, for example text No 9 presented graphically in Figure 3.3.

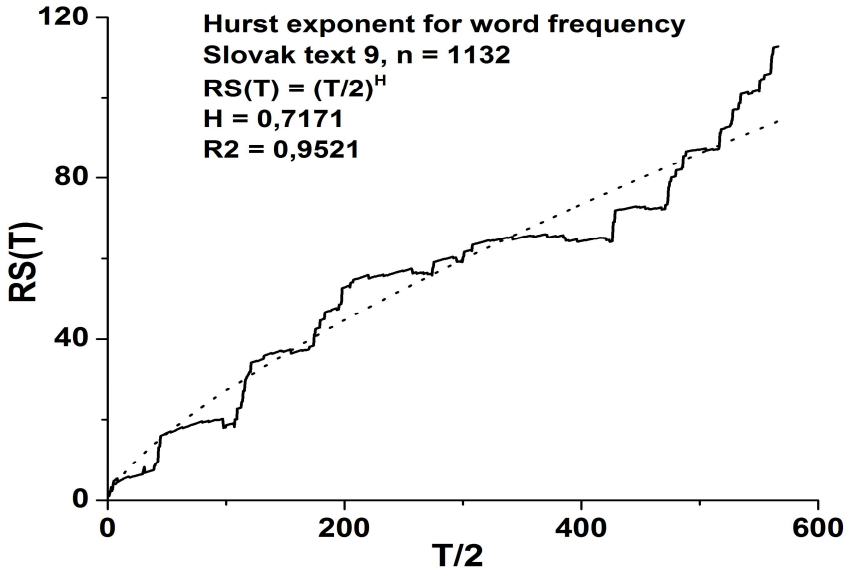


Figure 3.3. The RS(T) function for the Slovak text No. 9

A second alternative is represented by texts in which there is a turning point. Beginning from this point the *RS*(*T*) curve decreases (regularly or irregularly). This is the case e.g. in the Slovak text No 1 presented in Figure 3.4.

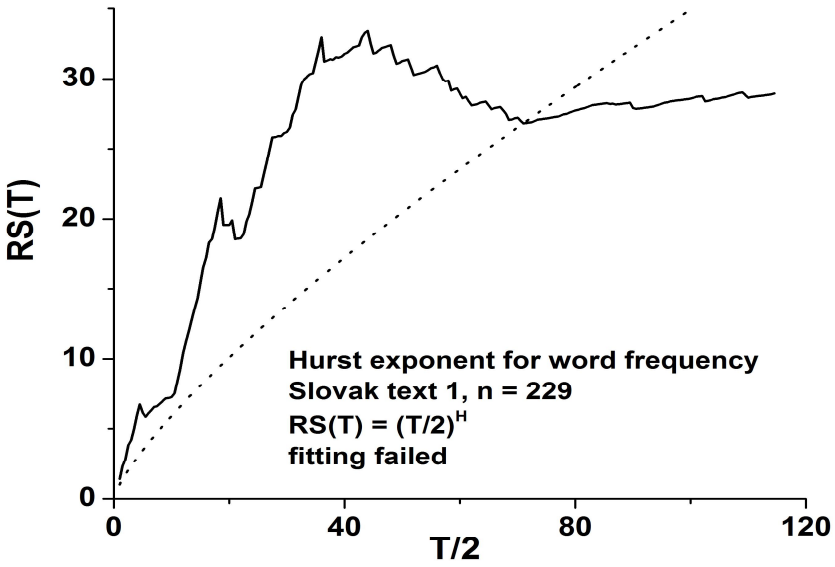


Figure 3.4. The RS(T) function in the Slovak text No 1

The third alternative represents a very irregular course of the curve testifying to the most probable pauses in writing or changes of the text. But one will surely find still other literary or textological "causes". An example for such a behaviour is the Slovak text no. 19, presented graphically in Figure 6.3. The general trend is conserved but at some places in the text there are positions which seem to represent a new begin of writing. Thus the Hurst exponent seems to be an interesting indicator of the dynamics of text generation.

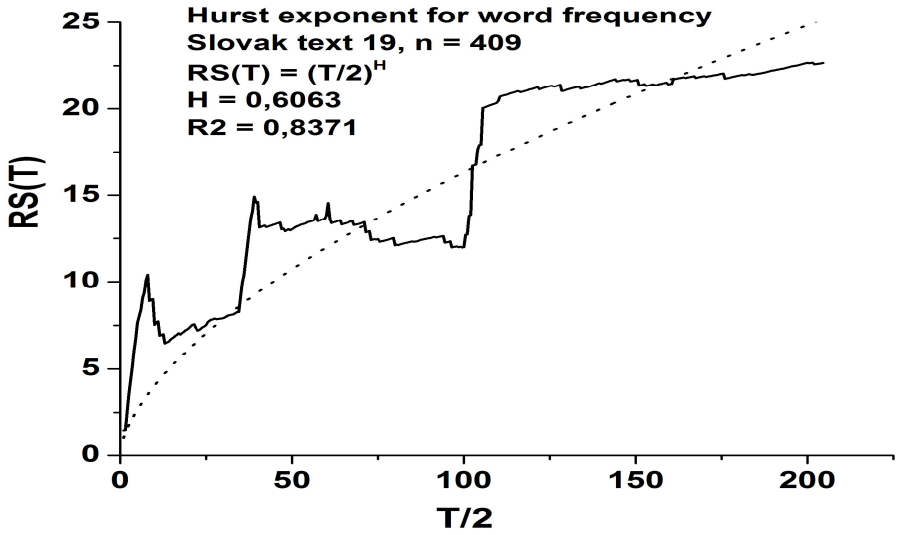


Figure 3.5. The RS(T) function in the Slovak text No 19

For the Russian texts we obtain the results presented in Table 3.4

Table 3.4
Hurst exponent in 20 Russian texts

Text	**Text sort**	**n**	**H**	**R^2**
1	SP	2905	0,5856	0,8990
2	SP	3389	0,6366	0,9152
3	SP	2657	0,6331	0,8391
4	SP	2656	0,8347	0,7816
5	SS	908	0,5254	0,5852
6	SS	949	0,6405	0,9342
7	SS	1951	0,5487	0,8764
8	SS	1453	0,7274	0,9539
9	NW	251	0,6811	0,9188
10	NW	785	fitting failed	
11	NW	220	0,6227	0,9315
12	NW	377	0,7099	0,9502
13	PL	879	0,6290	0,1603
14	PL	1097	0,5519	0,7721
15	PL	778	0,6519	0,9351
16	PL	2175	0,6853	0,8925
17	NO	2595	0,5820	0,8365
18	NO	5604	0,6157	0,9304
19	NO	702	0,6283	0,8378
20	NO	3576	0,6374	0,8842

In Russian we find another type of sequence which may, perhaps, display some partitions of the novel (cf. Figure 3.6).

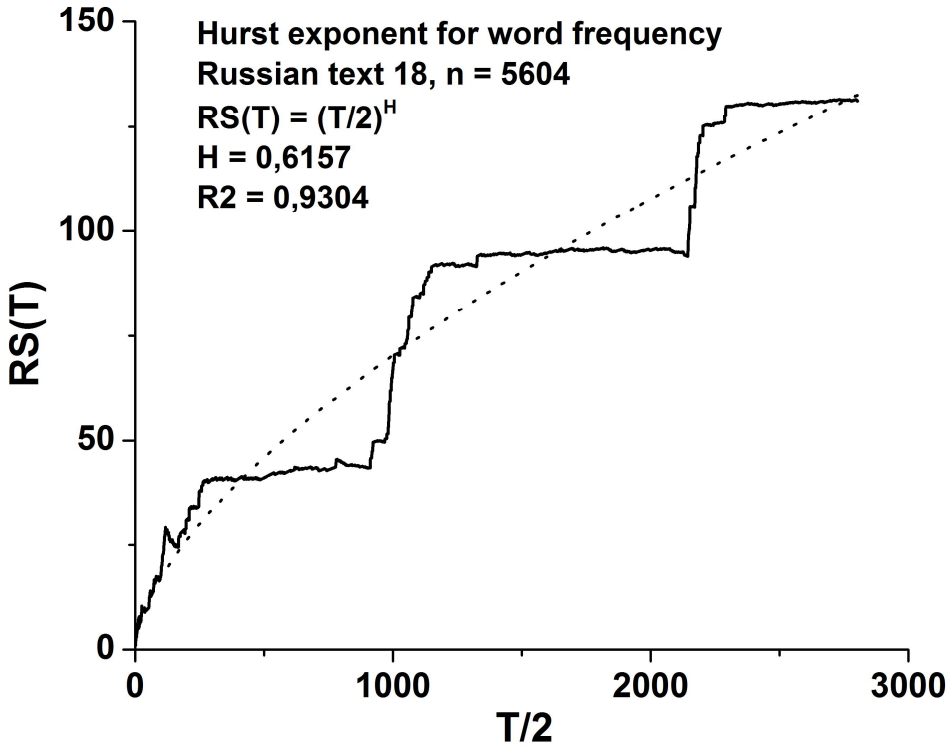


Figure 3.6. The RS(T) function in the Russian text No 18

The mean of *H* in Russian texts is $\bar{H}$ = 0.6383 and the variance is *Var*(*H*) = 0.00479. Here *n* = 19.

The order of text sorts according to their respective mean *H* is

Short story (0.6105) < *Novel* (0.6159) < *Private letter* (0.6295) < *News* (0.6712 < < *Stage play* (0.6725).

For the 20 Slovenian texts we obtain the results presented in Table 3.5. As can be seen, here the order of text sorts seems to be quite different to that based on *P*. Taking again the means of *H* from Table 3.5 we obtain

History (0.5551) < *Science* (0.5444) < *Letter novel* (0.6105) < *Open letter* (0.6120) < *Short novel* (0.6221) < *Short story* (0.6584) < *News* (0.6855) < *Novel* (0.6972).

The stability of this order must be tested using a number of other texts. A textological interpretation will be possible - perhaps - after analysing at least ten texts of each sort.

Table 3.5
Hurst exponent in 20 Slovenian texts

Text	**Text sort**	**n**	**H**	**R^2**
1	LN	236	0,6413	0,9431
2	LN	225	0,5797	0,9355
3	SS	2360	0,6584	0,9509
4	SS	1578	fitting failed	
5	Hi	3304	fitting failed	
6	Hi	3554	0,5551	0,5395
7	NS	691	fitting failed	
8	NS	775	0,6539	0,5147
9	NS	625	0,7171	0,9603
10	SN	602	0,6483	0,9503
11	SN	4028	0,6619	0,9005
12	SN	1829	0,6446	0,8254
13	SN	1023	0,5335	0,8379
14	Sc	1357	0,5335	0,8748
15	Sc	1199	0,5552	0,9134
16	OL	296	0,6235	0,8979
17	OL	556	0,5823	0,7711
18	OL	563	0,6301	0,8774
19	NO	1679	0,6491	0,9369
20	NO	1662	0,7452	0,9747

What is the relationship between P computed from the arc and *H* computed from normalized ranges? Are they correlated? As can easily be stated, there is no link between the two indicators. They characterize quite different properties of the text, though both capture the oscillation of the values. Even if we put together texts of the same sort in the three Slavic languages (there are only three such class up to now: SS, Nw, NO) we do not obtain any correlation. Hence the arc and the Hurst exponent are (preliminarily) independent of one another.

Computing the Hurst exponent for sentence lengths measured in terms of clause numbers we obtain for 20 German texts the results presented in Table 3.6.

Table 3.6
Hurst exponent in 20 German texts: sentence length in clauses

Text	**n**	**H**	**R^2**	**P**
T1	148	0,7244	0,7910	0,5288
T2	80	0,6343	0,7505	0,6289
T3	112	0,7296	0,7393	0,6936
T4	208	0,6296	0,8843	0,6501

T5	246	0,6195	0,3966	0,5955
T6	109	0,7178	0,2904	0,7090
T7	107	0,8157	0,8780	0,5847
T8	85	0,6428	0,5725	0,6759
T9	97	0,8059	0,8808	0,5551
T10	112	0,7220	0,9003	0,6371
T11	95	0,8444	0,9378	0,5112
T12	74	0,6700	0,8638	0,6115
T13	120	0,7778	0,5956	0,5316
T14	139	0,8545	0,9014	0,6653
T15	105	0,8703	0,7015	0,6195
T16	119	0,8766	0,9563	0,5254
T17	110	0,8377	0,9804	0,5509
T18	197	0,7487	0,9056	0,6122
T19	79	fitting failed		0,6050
T20	151	0,6929	0,5714	0,6804

The values of the determination coefficient are not always satisfactory but this is most probably a consequence of text corrections, our way of measurement or style. S can be shown in Figure 3.7., there is no relation between the Hurst exponent and the indicator P.

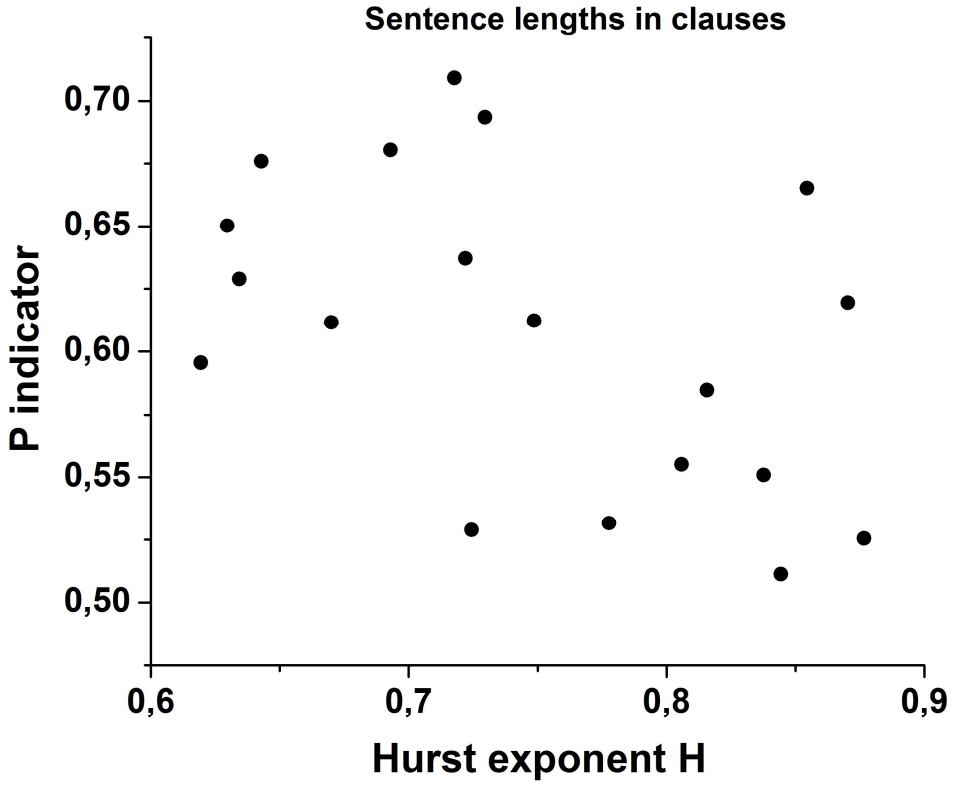


Figure 3.7. The independence of *H* and *P* indicators

4. Distances

In the analysis of a text we consider a sequence of numbers representing e.g. word length, sentence length or word frequency, we can measure the distance between equal numbers using different indicators. The simplest way is to consider the number of steps necessary to arrive at the next identical number. For example in (2.3) where we considered E_1 = (8,7,8,8,9,6,6,6,7,7,7,6,8,5,6,6,7,6,6,8,9,5,8,7,8,9,9,6,6,7,7,7) we begin with the first element, 8, and need 2 steps to meet the same number 8. The next 8 follows immediately after the sendond one, hence the distance is here 1.

If we count all distances and omit the distance of the last given element to the "next" (which would be infinity) we obtain the distribution of distances measured in this way. It is to be noted that this kind of measurement can be performed also with qualitative variables. In order to capture the course of f_x representing the number of distances x we may model it using the discrete or the continuous (c.f. Zörnig 2013) Zipf-Alekseev distribution. However, frequently the large distances have frequency 0 and in order to fit their distribution sequence one must pool many classes. Thus one uses mostly the continuous function for which the zero classes can be omitted. This is a "legal" procedure because the distance between identical elements can be measured in various (continuous) ways. Setting up a discrete or continuous model does not change the principal results because these concepts are chosen by the model builder, not by the reality. Thus we lean against the unified theory and conjecture that the relative rate of change of frequencies is proportional to the relative rate of change of distances, and the proportionality function is given as $g(x) = a + b \ln x$. We obtain the differential equation

$$\frac{df_x}{f_x} = \frac{a + b \ln x}{x} dx \tag{4.1}$$

The solution is

$$f_x = \frac{C}{x^{a + b \ln x}} \tag{4.2}$$

Which is an extended Zipf function. Would we consider (4.2) a discrete distribution, then C would be the normalizing constant. For our data concerning verse lengths we obtain the results presented in Table 4.1

Table 4.1
Distances between equally long verses in Erlkönig

x	1	2	3	4	5	6	7	8	9	16
$\mathbf{f_x}$	11	3	2	1	1	2	3	1	2	1
$\hat{f}_x$	10.92	3.37	2.11	1.66	1.47	1.37	1.33	1.32	1.33	1.68
a = 2.0384, b = 0.4924, C = 10.9218 , R^2 = 0.94										

As can be seen, the missing distances 10-15 are simply omitted. The formula expresses adequately the state of the affairs.

In the sequel we simply test the formula in order to obtain a background for such a procedure.

4.1. Word length

In order to test the adequateness of the above model, we fit it to the observed distances between words of identical length in texts from 28 languages. The parameter *C* is irrelevant because it merely expresses the frequency of the distance 1. The two other parameters could be used both for typological, text-sort analytic, stylistic and developmental analyses. Unfortunately the number of texts is too small in individual languages hence further research is necessary. But even at this preliminary stage one can observe e.g. differences between some Slavic languages for which the same text has been used. The interpretation must be postponed.

The fitting of the above model (4.2) to word length distances is presented in Table 4.2.

Table 4.2
Fitting the Zipf-Alekseev model to word length distances

Language/ Text	**Parameters**			
	a	**b**	**c**	$\mathbf{R^2}$
Akan: Agya Yaw Ne Akutu Kwaa	-0,3018	1,6105	88,1164	0,99
Akan: Mma Nnsua Ade Bɔne	-0,4506	1,2125	50,2644	0,99
Bamana: Bamakɔ sigicoya	0,5827	0,5406	462,4441	0,997
Bamana: Masadennin	1,1021	0,3529	1229,1857	0,998
Bamana: Namakɔrɔba halakilen	1,2538	0,2606	706,4107	0,9995
Bamana: Sonsannin ani	0,7449	0,565	1087,639	0,999
Bulgarian: Ostrovskij, Kak se kaljavaše stomanata (Chap. 1)	-0,6538	0,8604	204,1388	0,996
Czech: Čulík, O čem jsou dnešní Spojené státy?	0,1993	0,39	505,9607	0,998
Czech: Hvížďala, O předem zpackané prezidentské volbě	0,1272	0,4291	234,5507	0,99
Czech: Macháček, Slovenský dobrý příklad	0,0655	0,4914	85,9735	0,99
Czech: Spurný, Prekvapení v justici	0,2114	0,3764	72,787	0,98
Czech: Švehla, Editorial, Voličův kalkul	-0,0215	0,5184	72,0338	0,98
French: Dunkerque – La route des dunes (press)	-0,0558	0,767	493,4571	0,996
German: Assads Familiendiktatur	0,1938	0,4734	398,6965	0,998
German: ATT0012 (press)	0,1766	0,4816	320,2555	0,998
German: Die Stadt des Schweigens	0,1145	0,5835	466,9158	0,996
German: Terror in Ost Timor	-0,0027	0,6448	4.102.169	0,998
German: Unter Hackern und Nobelpreisen	0,248	0,4769	399,852	0,998

Hindi: After the sanctionto love marriage (press)	-0,2918	0,8888	346,947	0,999
Hindi: The Anna Team on a cross-road (press)	-0,3026	1,0229	305,7838	0,998
Hungarian: A nominalizmus Forradalma (press)	-0,3578	0,5469	232,5982	0,998
Hungarian: Kunczekolbász (press)	-0,4872	0,5948	77,0436	0,98
Indonesian: Pengurus PSM terbelah (press)	0,5951	0,279	109,2435	0,99
Indonesian: Sekolah ditutup (press)	-0,3865	0,8586	74,5227	0,97
Italian: Il bosone di Higgs scoperto dal Cern (Internet)	-0,4921	0,7455	544,6497	0,998
Japanese: Miki, Jinseiron Note	-1,1231	1,2207	381,4454	0,99
Kikongo: Bimpa: Ma Ngo ya Ma Nsiese	-0,1335	0,6449	219,7708	0,99
Kikongo: Lumumba speech	-0,9993	1,0809	222,1013	0,997
Kikongo: Nkongo ye Kisi Kongo	-1,2297	1,6979	233,6042	0,99
Latin: Cicero, In Catilinam I	0,1416	0,4203	283,3235	0,99
Latin: Cicero, In Catilinam II	-0,3045	0,5902	654,3742	0,999
Macedonian: Ostrovskij, Kako se kaleše čelkiot (Chap. 1)	-0,9083	0,9126	204,2451	0,996
Malayalam: Moralistic hooligans (press)	0,0922	0,3201	51,777	0,96
Malayalam: No one should die (press)	-0,0804	0,3471	43,3842	0,95
Maninka: Nko Doumbu Kende no.2 (press)	0,7791	0,3617	833,7428	0,999
Maninka: Nko Doumbu Kende no.7 (press)	0,4372	0,4898	550,7852	0,997
Maninka: Sìikán` (Constitution of Guinea, an excerpt)	-0,0746	0,6721	481,1838	0,996
Odia: The Samaj, Bhuba-neshwar (28 June 2012), p. 4	0,2968	0,4666	106,3381	0,98
Odia: The Dharitri, Balasore (12th Feb, 2012), p. 10	-0,1116	0,5341	146,6132	0,98
Romanian: Paler, excerpt from Aventuri solitare	-0,9371	1,1645	204,4162	0,98
Romanian: Steinhardt, Jurnalul fericirii, Trei soluții	-0,3386	0,8373	412,3248	0,99
Romanian: Popescu D.R., Vânătoarea regală	-0,3802	0,904	293,4566	0,995
Russian: Ostrovskij, Kak zakaljalas stal' (Chap. 1)	0,0277	0,5064	206,283	0,996
Serbian: Ostrovskij, Kako se kalio čelik (Chap. 1)	-0,0527	0,5345	255,1807	0,99
Slovak: Bachletová, Moja Dolná zem	-0,0647	0,5329	216,1887	0,998
Slovak: Bachletová, Riadok v tlačive	0,0388	0,4578	205,8707	0,99
Slovenian, Kak zakaljalas	0,1864	0,5209	299,9342	0,996
Sundanese: Agustusan (Online)	0,0622	0,5837	122,0413	0,99

Sundanese: Aki Satimi (Online)	-0,2768	0,7341	344,2042	0,998
Tagalog: Rosales, Kristal Na Tubig	-2,078	1,7239	303,0685	0,98
Tagalog: Hernandez, Limang Alas: Tatlong Santo	-1,0358	1,0807	349,0266	0,996
Tagalog: Hernandez, Magpisan	-1,4802	1,2429	236,0223	0,99
Tamil: Emu Bird Trading (press)	0,4878	0,2059	88,9384	0,97
Telugu: Trailangaswamy (press)	0,0935	0,3672	59,7176	0,98
Telugu: Train Journey (press)	0,1276	0,3614	141,1524	0,99
Vai: Sa'bu Mu'a'	0,7317	0,5156	232,4315	0,996
Vai: Sherman, Mu ja vaa	0,7717	0,5211	1494,9202	0,999
Vai: Vande	-0,515	1,5147	176,3501	0,99
Welsh: text 1 (gaenv)	-0,786	0,9916	226,8039	0,99
Welsh: text 2 (gasodl)	0,0522	0,7892	469,4913	0,996

4.2. Sentence length

We shall omit sentence length measured in terms of word numbers and concentrate on those measured in clause numbers. We have merely 20 German newspaper data. In all of them, the Zipf-Alekseev model can be fitted with very good results. The data are presented in Table 4.3. Again, the parameter *C* depends merely on the first value (or rather text size) and is not relevant.

Table 4.3
Fitting the Zipf-Alekseev function to distances between equal sentence lengths (in clauses) in 20 German newspaper texts.

T1			T2			T3		
x	f_x	$\hat{f}_x$	x	f_x	$\hat{f}_x$	x	f_x	$\hat{f}_x$
1	48	48.11	1	24	23.94	1	36	36.08
2	27	25.94	2	13	13.59	2	21	21.13
3	14	16.28	3	9	8.98	3	16	13.00
4	11	11.17	4	9	6.44	4	3	8.53
5	11	8.12	5	5	4.87	5	7	5.88
6	7	6.15	6	2	3.83	6	7	4.22
7	4	4.81	7	2	3.09	7	1	3.12
8	3	3.85	8	2	2.55	8	2	2.37
9	2	3.14	9	2	2.14	9	3	1.84
11	1	2.19	19	2	0.62	12	1	0.94
12	2	1.86	25	1	0.37	15	1	0.54
13	1	1.59	35	1	0.19	19	1	0.28
14	1	1.38	56	1	0.07	20	1	0.25

17	2	0.93				27	2	0.10
19	2	0.74				37	1	0.04
20	1	0.67				41	1	0.03
21	1	0.60				57	1	0.01
24	1	0.45				59	1	0.01
30	1	0.27						
31	1	0.25						
a = 0.7296, b = 0.2334 c = 48.1140, R^2 = 0.99			a = 0.6857, b = 0.1885 c = 23.9384, R^2 = 0.97			a = 0.5030, b = 0.3877 c = 36.0815, R^2 = 0.96		

T4			T5			T6		
x	f_x	$\hat{f}_x$	x	f_x	$\hat{f}_x$	x	f_x	$\hat{f}_x$
1	74	74.30	1	84	82.66	1	25	24.59
2	41	38.53	2	36	44.86	2	18	20.03
3	19	23.44	3	36	28.55	3	17	14.48
4	17	15.66	4	29	19.87	4	10	10.49
5	11	11.13	5	13	14.64	5	9	7.76
6	9	8.26	6	10	11.23	6	4	5.86
7	7	6.34	7	5	8.88	7	7	4.51
9	5	4.01	8	6	7.19	8	2	3.53
10	3	3.28	9	2	5.92	9	1	2.81
11	2	2.72	10	6	4.96	11	1	1.85
12	2	2.28	12	1	3.61	12	2	1.52
13	2	1.94	13	4	3.12	14	1	1.06
14	1	1.66	15	2	2.40	16	1	0.77
17	1	1.09	17	1	1.89	22	1	0.33
24	3	0.49	23	1	1.03	25	1	0.23
32	1	0.24	29	1	0.63	45	1	0.03
59	1	0.05	31	1	0.55			
67	1	0.03						
88	1	0.01						
a = 0.7717, b = 0.2534 C = 74.2955, R^2 = 0.99			a = 0.7354, b =0.2112 C = 82.6562, R^2 = 0.96			a = 0.02225, b = 0.4593 C = 24.5914, R^2 = 0.96		

T7			T8			T9		
x	f_x	$\hat{f}_x$	x	f_x	$\hat{f}_x$	x	f_x	$\hat{f}_x$
1	32	32.30	1	18	17.82	1	36	35.60
2	20	18.83	2	15	15.62	2	14	16.70
3	13	11.96	3	10	11.34	3	11	10.47
4	5	8.15	4	14	8.10	4	13	7.43
5	5	5.84	5	3	5.87	6	4	4.51
6	3	4.35	6	2	4.33	7	2	3.72
7	6	3.33	7	3	3.25	9	1	2.69
8	2	2.62	8	3	2.49	10	2	2.35
9	2	2.09	9	2	1.93	11	1	2.07

10	3	1.70	10	1	1.52	12	2	1.84
12	2	1.17	12	1	0.98	14	1	1.50
13	1	0.99	16	1	0.45	22	1	0.81
15	1	0.72	17	1	0.38	23	1	0.76
18	2	0.47	20	2	0.23	34	1	0.44
19	1	0.42	36	1	0.03	38	1	0.37
35	1	0.09	40	1	0.02			
36	1	0.08						
a = 0.5634, b = 0.3104 C = 32.3002, R^2 = 0.97			a = -0.1891, b = 0.5462 C = 17.8153, R^2 = 0.89			a = 1.0535, b = 0.0554 C = 35.6046, R^2 = 0.96		

T10			T11			T12		
x	f_x	$\hat{f}_x$	x	f_x	$\hat{f}_x$	x	f_x	$\hat{f}_x$
1	24	24.69	1	30	30.10	1	28	27.68
2	32	29.48	2	21	20.52	2	10	12.98
3	13	17.40	3	12	12.33	3	12	8.04
4	10	9.02	4	7	7.56	4	8	5.63
5	4	4.61	5	5	4.81	5	3	4.23
6	6	2.40	6	3	3.17	6	1	3.32
7	4	1.28	7	2	2.16	7	2	2.70
8	1	0.71	8	2	1.51	10	1	1.65
10	3	0.23	9	1	1.08	14	1	1.01
11	1	0.14	10	1	0.79	16	1	0.83
13	2	0.05	14	2	0.26	25	1	0.42
16	3	0.01	15	1	0.20	41	1	0.19
22	1	0.001	18	1	0.10			
27	1	0.0003	32	1	0.01			
32	1	0.00007	36	1	0.01			
a = -1.2381, b = 1.4171 C = 24.689, R^2 = 0.94			a = 0.1093, b = 0.6400 C = 30.1031, R^2 = 0.99			a = 1.0348, b = 0.0826 C = 27.6784, R^2 = 0.95		

T13			T14			T15		
x	f_x	$\hat{f}_x$	x	f_x	$\hat{f}_x$	x	f_x	$\hat{f}_x$
1	43	43.22	1	63	63.10	1	25	24.02
2	32	30.59	2	30	29.06	2	19	22.61
3	13	16.00	3	12	14.09	3	18	16.58
4	9	8.28	4	8	7.47	4	18	11.76
5	6	4.43	5	5	4.26	5	8	8.40
8	1	0.86	6	3	2.58	6	2	6.10
9	2	0.53	7	1	1.63	7	3	4.51
11	2	0.22	8	1	1.07	9	1	2.58
12	1	0.15	9	1	0.73	13	1	0.99
15	1	0.05	10	2	0.51	19	1	0.31
25	1	0.003	12	1	0.26	23	1	0.16
28	1	0.001	15	1	0.11	26	1	0.10

32	1	0.0005	20	1	0.03	77	1	0.0009
34	1	0.0003	22	1	0.02			
36	1	0.0002	29	1	0.006			
			37	1	0.002			
			39	1	0.001			
a = -0.1951, b = 1.0009 C = 43.2204, R^2 = 0.99			a = 0.6987, b = 0.6060 C = 63.0995, R^2 = 0.996			a = -0.3396, b = 0.6165 C = 24.0244, R^2 = 0.92		

T16			T17			T18		
x	f_x	$\hat{f}_x$	x	f_x	$\hat{f}_x$	x	f_x	$\hat{f}_x$
1	36	36.53	1	42	41.81	1	88	87.44
2	31	27.82	2	17	18.72	2	26	31.18
3	9	15.60	3	14	11.29	3	23	17.67
4	12	8.57	4	6	7.77	4	14	11.99
5	4	4.84	5	9	5.76	5	10	8.96
6	3	2.83	6	4	4.48	6	7	7.10
7	5	1.71	8	1	2.99	7	5	5.86
8	2	1.07	10	1	2.16	8	5	4.97
9	2	0.69	11	1	1.88	9	5	4.32
11	2	0.30	13	2	1.46	13	2	2.81
13	1	0.15	14	1	1.31	14	1	2.58
18	1	0.03	15	1	1.18	18	1	1.96
20	1	0.02	19	1	0.81	19	1	1.84
25	2	0.005	23	1	0.60	20	1	1.75
35	1	0.0006	26	1	0.49	21	1	1.66
43	1	0.0002	41	1	0.23	25	1	1.38
			44	1	0.21	34	1	1.01
a = -0.2598, b = 0.9417 C = 36.5342, R^2 = 0.95			a = 1.1044, b = 0.07922 C = 41.8129, R^2 = 0.98			a = 1.5421, b = -0.0786 C = 87.4357, R^2 = 0.99		

T19			T20		
x	f_x	$\hat{f}_x$	x	f_x	$\hat{f}_x$
1	19	19.15	1	69	68.82
2	16	15.84	2	21	22.43
3	12	10.95	3	11	12.37
4	8	7.51	4	13	8.34
5	1	5.25	5	9	6.24
6	2	3.75	6	6	4.97
7	7	2.74	7	1	4.13
9	3	1.55	8	2	3.54
10	1	1.19	9	2	3.10
13	1	0.59	10	2	2.76
28	1	0.05	11	2	2.49
29	1	0.04	15	2	1.82
36	1	0.02	20	1	1.39

			33	1	0.92
			43	1	0.76
			44	1	0.75
			54	1	0.66
a = -0.1265, b = 0.5785 C = 19.1548, R^2 = 0.91			a = 1.7125, b = -0.1371 C = 68.8191, R^2 = 0.99		

Here it is easier to study the relationship between the parameters *a* and *b* because we have 20 data sets taken from the same text sort (newspapers). If we order the two parameters according to increasing values of *a*, we obtain the results presented in Table 4.4, showing that $b = f(a)$ where *b* is a linear function of *a*. That means, the distances between identical sentence lengths are controlled by two mechanisms. The first controls the distances, the second guarantees an equilibrium between the parameters. The function expressed by the results in Table 4.4. is $b = 0.6448 - 0.5050a$ where $R^2 = 0.86$. Further data would yield still better results.

Table 4.4
The dependence of parameter *b* on parameter *a*
in German sentence length-distance relationship

a	**b**	**b = f(a)**
-1,2381	1,4171	1.2700
-0,3396	0,6165	0.8163
-0,2598	0,9417	0.7760
-0,1951	1,0009	0.7433
-0,1891	0,5462	0.7403
-0,1265	0,5785	0.7087
0,0223	0,4593	0.6335
0,1093	0,6400	0.5896
0,5030	0,3877	0.3908
0,5634	0,3104	0.3603
0,6857	0,1885	0.2985
0,6987	0,6060	0.2920
0,7296	0,2334	0.2764
0,7354	0,2112	0.2734
0,7717	0,2534	0.2551
1,0348	0,0826	0.1223
1,0535	0,0554	0.1128
1,1044	0,0792	0.0871
1,5421	-0,0786	-0.1339
1,7125	-0,1371	-0.2200

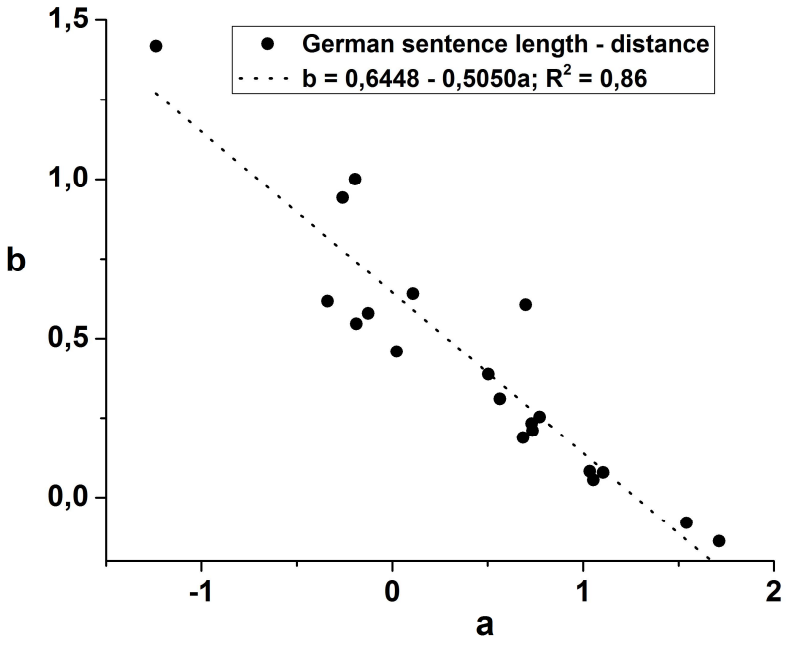


Figure 4.1. The dependence of parameter *b* on parameter *a* in Table 4.4

4.3. Frequencies

The distances between identical word frequencies present a rather different image. Frequency depends on thematic concentration, on vocabulary richness and other factors, and depends more on the given text than on frequencies which occur in corpuses. Hence one cannot expect a clearly expressed mutual dependence of parameters, quite the contrary. The frequencies may be very extreme, many of them are quite unique. Thus we expect a good fitting by the Zipf-Alekseev function but not a demonstrable relationship between the parameters.

The results of fitting (4.2) to the individual frequency vectors are presented in Table 4.5. All fits are very good, hence one can conjecture to have found the suitable function.

Table 4.5
Distances between equal frequencies fitted by the Zipf-Alekseev function (Slovenian)

Language/ Text	**Parameters**			
	a	**b**	**C**	**R^2**
T1	0,9781	0,0471	226,934	0,99
T2	0,1103	0,1871	377,1105	0,98

T3	0,8676	0,0972	182,6319	0,98
T4	0,8278	0,0761	139,5228	0,99
T5	1,0405	0,1604	208,0246	0,995
T6	0,7528	0,2848	149,2626	0,98
T7	0,8861	0,1138	230,508	0,99
T8	0,9338	0,2102	178,8466	0,99
T9	1,1223	0,0792	71,6603	0,99
T10	1,3713	0,0129	178,1989	0,99

Table 4.6
Ordered according to parameter *a*

a	**b**
0,1103	0,1871
0,7528	0,2848
0,8278	0,0761
0,8676	0,0972
0,8861	0,1138
0,9338	0,2102
0,9781	0,0471
1,0405	0,1604
1,1223	0,0792
1,3713	0,0129

Though it can be shown that parameter *b* decresases here with increasing *a*, the dependence is not linear but rather oscillating, hence no simple function can be used to captrure it. This need not hold for all languages, as can be shown below..

Table 4.7
Distances between equal frequencies fitted by the Zipf-Alekseev function (Slovak)

Text	**a**	**b**	**C**	**R^2**
T1	1,7279	-0,1064	99,9092	0,996
T2	1,1216	0,1879	112,8195	0,96
T3	0,8128	0,205	222,6503	0,99
T4	0,5194	0,2916	246,0265	0,99
T5	0,2425	0,253	24,1248	0,89
T6	2,1322	-0,2667	50,0155	0,998

T7	0,9093	0,2561	151,8341	0,99
T8	1,2987	0,0527	134,2203	0,995
T9	0,8963	0,2018	332,683	0,99
T10	0,8572	0,0655	234,9113	0,99
T11	1,4245	-0,006	108,6762	0,99
T12	1,3533	0,0232	103,4679	0,99
T13	0,3054	0,2708	144,8673	0,98
T14	0,7819	0,1503	234,3294	0,99
T15	0,3439	0,1664	204,4295	0,98
T16	0,0412	0,1821	368,2186	0,99
T17	0,4271	0,2894	312,9521	0,99
T18	0,3781	0,1754	489,6514	0,99
T19	0,4234	0,4662	121,6505	0,98
T20	0,211	0,5471	108,2932	0,96

If we consider the link between the parameters *a* and *b*, we may state that *b* can be expressed by a linear function of *a*, viz. $b = 0.3880 - 0.2686a$, $R^2 = 0.66$ The reordering according to increasing *a* is presented in Table 4.8. A graphical presentation can be found in Figure 4.2.

Table 4.8
The relationship between *a* and *b* in frequency distances in Slovak texts

a	**b**	**f(a)**
0.0412	0.1821	0,3769
0.2110	0.5471	0,3313
0.2425	0.2530	0,3229
0.3054	0.2708	0,3060
0.3439	0.1664	0,2956
0.3781	0.1754	0,2864
0.4234	0.4662	0,2743
0.4271	0.2894	0,2733
0.5194	0.2916	0,2485
0.7819	0.1503	0,1780
0.8128	0.205	0,1697
0.8572	0.0655	0,1578
0.8963	0.2018	0,1473
0.9093	0.2561	0,1438
1.1216	0.1879	0,0867
1.2987	0.0527	0,0392
1.3533	0.0232	0,0245
1.4245	-0.006	0,0054
1.7279	-0.1064	-0,0761
2.1322	-0.2667	-0,1847

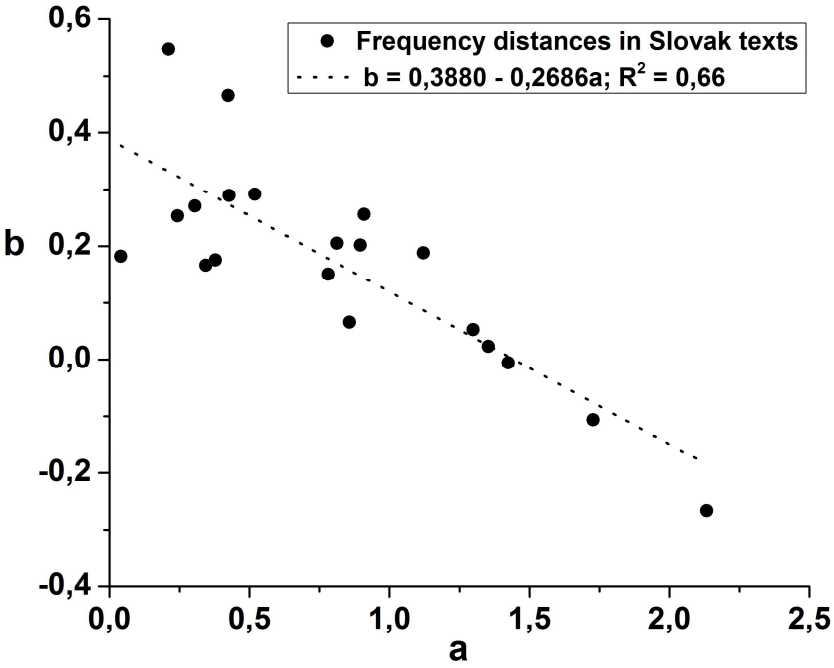


Figure 4.2. Dependence of *b* on *a* in frequency distances in Slovak texts

As could be expected, some of the texts display outliers whose character must be studied individually. It may be caused by the style of the author, by the theme, by the correction brought in after the text was ready, etc.

For the Russian data we obtain the results presented in Table 4.9.

Table 4.9
Distances between equal frequencies fitted by the Zipf-Alekseev function (Russian)

Text	a	b	c	R^2
T1	0,6464	0,0934	444,9898	0,99
T2	0,4194	0,1699	469,6606	0,99
T3	0,1792	0,2697	360,0277	0,98
T4	0,3580	0,2253	426,9983	0,99
T5	0,7201	0,1427	200,4294	0,97
T6	0,6948	0,2081	242,1476	0,99
T7	0,6315	0,1764	397,4242	0,99
T8	0,5626	0,1467	252,4396	0,99
T9	0,6437	0,9983	120,1178	0,99
T10	1,2262	0,1760	307,3182	0,99
T11	1,9545	-0,0997	112,4488	0,99
T12	1,5000	-0,0570	142.2132	0,99
T13	0,3863	0,2028	147,1140	0,98
T14	0,0583	0,4178	197,9598	0,97

T15	0,4529	0,2335	156,4488	0,97
T16	0,3969	0,1822	327,8049	0,99
T17	0,5885	0,1333	418.4794	0,99
T18	0,4989	0,1218	704,2576	0,99
T19	0,8346	0,1423	179,9826	0,99
T20	0,5189	0,1371	517,1111	0,99

Here merely one of the texts, namely T9 displays an outlier. Its cause could, perhaps, be found by analyzing the given text but this is not our aim. In any case, here *b* is linked with *a* in form $b = 0.2948 - 0.2043a$, $R^2 = 0.70$. The reordering of *a* and *b* values in presented in Table 4.10.

Table 4.10
Parameters *a* and *b* in Russian texts after reordering

Text	**a**	**b**	**b = f(a)**
T14	0,0583	0,4178	0,2829
T3	0,1792	0,2697	0,2582
T4	0,3580	0,2253	0,2217
T13	0,3863	0,2028	0,2159
T16	0,3969	0,1822	0,2137
T2	0,4194	0,1699	0,2091
T15	0,4529	0,2335	0,2023
T18	0,4989	0,1218	0,1929
T20	0,5189	0,1371	0,1888
T8	0,5626	0,1467	0,1799
T17	0,5885	0,1333	0,1746
T7	0,6315	0,1764	0,1658
T9	0,6437	0,9983	0,1633
T1	0,6464	0,0934	0,1627
T6	0,6948	0,2081	0,1529
T5	0,7201	0,1427	0,1477
T19	0,8346	0,1423	0,1243
T10	1,2262	0,1760	0,0443
T12	1,5000	-0,0570	-0,0117
T11	1,9545	-0,0997	-0,1045

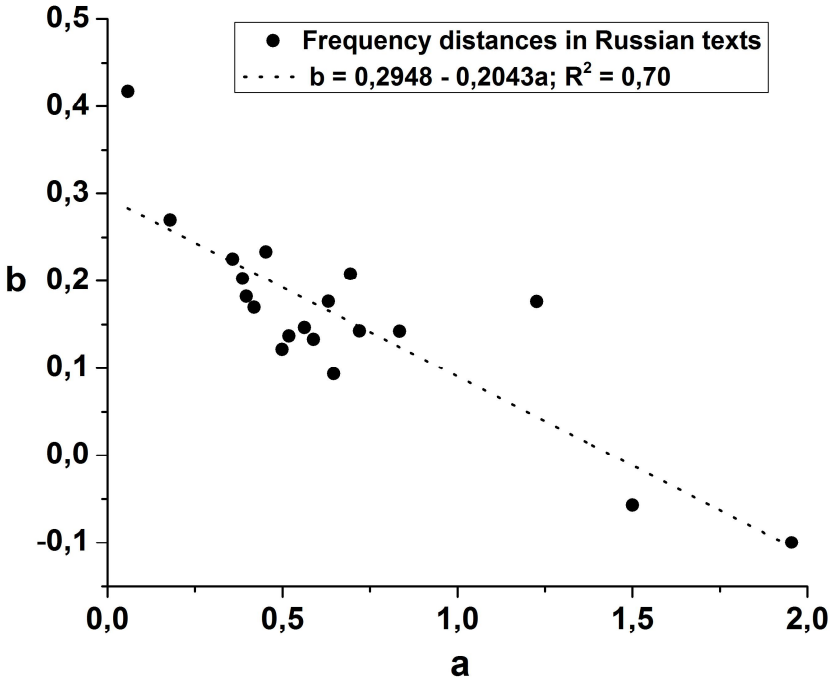


Figure 4.3. Relationship between a and b

Summary

Whatever property of text entites is scrutinized, i.e. measured and placed in the text replacing the original entity, the sequence itself may look chaotically forming rather a kind of fractal; nevertheless it conceals a number of stochastic regularities whose investigation is a task for the future. We merely selected some of them in order to show that they exist and abide by some regularities which after thorough testing may acquire the status of laws. Needless to say, a control cycle comprising all of them must be set up in order to see that they are linked with one another, and show the mathematical form of the link. Due to space restrictions, we merely showed the way directed to a deeper analysis. Arcs, Hurst exponents and distances have been analyzed in several languages for word length, sentence length and frequency representing only the surface of the text as a phenomenon. The next steps would be (1) the study of further languages in order to obtain a more extensive empirical basis, (2) the study of further properties taken from the infinite reservoir of language properties and (3) the linking of all these properties in control cycles in order to show that we have to do with a dynamic system. The study of these systems would be performed at a higher level using more complex mathematics.

References

Altmann, G. (2006). Fundamentals of quantitative linguistics. In: Genzor, J., Bucková, M. (eds.), *Favete linguis: 15-27*. Bratislava: Slovak Academic Press.

Brockwell, P.J., Davis, R.A. (2010). *Introduction to Time Series and Forecasting.* Berlin: Springer.

Çambel, A.B. (1993). *Applied chaos theory. A paradigm for complexity.* San Diego: Academic Press.

Hamilton, J.D. (1994). *Time series analysis.* Princeton University Press.

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

Kitagawa, G., Gersch W. (1996). Smoothness Priors Analysis of Time Series (Lecture Notes in Statistics 116), Springer, Berlin, Heidelberg, New York.

Köhler, R., Altmann, G. (2008, 2nd ed.). *Problems in Quantitative Linguistics, Vol. 1*. Lüdenscheid: RAM.

Mandelbrot, B.B. (1982). *The fractal geometry of nature.* New York: Freeman

Pandit, S.M., Wu, S.M. (1983). *Time series and system analysis with applications.* New York: Wiley.

Percival, D.B., Walden, A.T. (2010). *Wavelet Methods for Time Series Analysis.* Cambridge University Press.

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

Popescu, I.-I., Naumann, S., Kelih, E., Rovenchak, A., Overbeck, A., Sanada, H., Smith, R., Čech, R., Mohanty, P., Wilson, A., Altmann, G. (2013). Word length: aspects and languages. In: Köhler, R., Altmann, G. (eds.), *Issues in Quantitative Linguistics Vol. 3: 224-281*. Lüdenscheid: RAM-Verlag.

Zörnig, P. (2013). A continuous model for the distances between coextensive words in a text. *Glottometrics25, 54-68*.

History of Quantitative Linguistics

Since a historiography of quantitative linguistics does not exist as yet, we shall present in this column short statements on researchers, ideas and findings of the past – usually forgotten – in order to establish a tradition and to complete our knowledge of history. Contributions are welcome and should be sent to Peter Grzybek, peter.grzybek@uni-graz.at.

Historical Remarks on the Consonant-Vowel Proportion – From Cryptoanalysis to Linguistic Typology

The Concept of Phonological Stoichiometry (Francis Lieber, 1800-1872)

Peter Grzybek

Studies on the frequency of vowels and consonants in general, and on consonant-vowel proportions specifically, have a history which reaches back much longer than is usually assumed. Unfortunately, we know this history only most rudimentarily.

The beginning of such studies can be seen, it seems, in early cryptographical and cryptological literature (cf. Ycart 2013). Arab philosopher and mathematician Yaʿkūb ibn Ishāq Al Kindī (800-873), for example, in his manuscript *On Deciphering Cryptographic Messages*, which seems to be the oldest known description of cryptoanalysis by frequency analysis, clearly points out the different frequencies of vowels and consonants, and arrives at the conclusion that "the number of vowels in any language would be greater than non-vowels".[1]

Six centuries later, in the context of West European Renaissance, and ignorant of his Arab predecessor, Italian humanist Leon Battista Alberti (1404-1472) started his relevant ruminations in his *De Componendis Cifris* (ca. 1466-67); he arrived at the conclusion that if we take "one or two pages of poetry or prose and extract the vowels and consonants, listing them in separate series, vowels on one side and consonants on the other, you will no doubt find that there are numerous vowels"[2]. Moreover, what is even more important in our context, is Alberti's attempt to quantify the CV relation:

> From my calculations, it turns out that in the case of poetry, the number of consonants exceeds the number of vowels by no more than an octave[3] [non amplius quam ex octava], while in the case of prose the consonants do not usually exceed the vowels by a ratio greater than a sesquialtera [ferme ex proportione quam sesquitertiam]. If in fact we add up all the vowels on a page, let's say there are three hundred, the overall sum of the consonants will be four hundred.[4]

[1] Quoted after Ycart (2013: 1)
[2] Quoted after Ycart (2013: 1)
[3] There has been a debate as to the meaning of "an octave", but it seems reasonable to side with Ycart's (1012) argument and interpret it in terms of "one eighth".
[4] Quoted after Ycart (2013: 9)

The distinction of vowels and consonants on the basis of frequency became more or less common during the following centuries and, in fact, one of the standard procedures in crypto-analysis. There is no need to go into historiographic details here, concentrating rather on the vowel-consonant proportion specifically, which has been discussed to a much lesser degree. In this respect, Alberti was much more concrete as to numeric details than most of his followers, who nevertheless made concrete suggestions as to the CV proportion. One of them was David Arnold Conrad who, in his 1732[5] *Cryptographia denudata, sive Ars Decifrandi*, gave such an estimation of the CV relation: "The Vowels, generally five, are four times out-numbered by the Consonants, the Vowels must therefore recur most frequently" [Quoted after Ycart (2013: 6)]. Yet another estimation of this kind, containing a quantifying assertion about CV frequencies, can be found in Christian Breitenhaupt's 1737 *Ars Decifratoria*:

> The frequency of letters should be noted in general, since in any language vowels are more numerous than consonants. The reason for making these observations is obvious. Actually, for a given number of vowels, the corresponding number of consonants must be larger by five fourths; it cannot be otherwise, vowels being more frequent than consonants.[6]

Here is not the place to go into more details as to the history of studies on CV proportions. It seems that, from a historical point of view, and going beyond the narrow field of crypto-graphy, studies in this field have been epistemologically motivated by three major realms of interest:

1. *Genuinely linguistic*. The primary field of interest is, of course, linguistic: after all, it is a genuinely linguistic issue to define consonants and vowels, as well as other units; subsequent questions have mainly concentrated on typological issues, with regard to intra-lingual aspects (which factors have, within a given language, impact on the CV proportion?) as well as to inter-lingual and cross-linguistic aspects (is the CV proportion a possible characteristic for language typology?)
2. *Aesthetic and poetic*. In this respect, a leading question has been, is it possible to define phenomena like the euphony, or harmony, of a given language, or of individual texts in that language, on the basis of the CV relation?
3. *Pedagogic*. Here a major issue has been the question, if knowledge about the CV proportion can help in defining matters of text difficulty and understandability, or distinguish "easy-to-learn languages" from more (or less) "difficult" ones, and related questions.

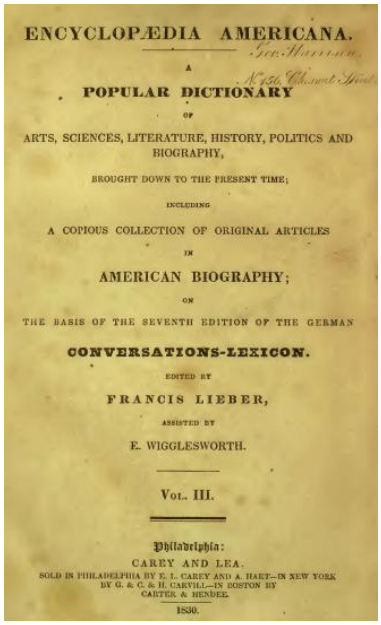
As a matter of fact, any answer to the second and to the third question must, in one way or another, start from an at least implicit assumption concerning the definition of the basic terms; in case the definitions are explicit, they depend, historically speaking, on the state of linguistic knowledge and, from a contemporary point of view, on the concrete linguistic theory chosen. In any case, the underlying definition is subsequently relevant, of course, for the frequencies of the distinguished items, as the basis for calculating the proportion between them. Quite naturally, genuinely linguistic approaches to the CV issue increased and have been prevailing in the in the 20th century, with the rise of linguistic theory. In this respect, mainly structuralist and typological approaches, as e.g., Isačenko's (1939/40), Krámský's (1946/48), or Skalička's (1966) seminal papers; these approaches have later been thoroughly reflected by Altmann and

[5] Conradus' texts were re-published 10 years later, in 1742, in the *Gentleman's Magazine*, in a series of articles.
[6] Quoted after Ycart (2013: 1)

Lehfeldt (1973) from a methodological point of view. And despite some justified critique, as brought forth e.g., by Kempgen (1991), one can only agree with Kelih (2010) stating that the basic idea about CV proportions in languages still today is of great relevance and timeliness.

Nevertheless, the historical interest in CV proportion is much older than these approaches, and it is just this historical background about which our knowledge is but fragmentary, from a historiographical point of view. Scholars, who were interested in related issues at the beginning of the 19th century, more often than not came from disciplines other than linguistics, the latter more or less in the *status nascendi*, rather than being an established branch of science. Among those scholars was, as has been shown elsewhere (cf. Grzybek 2006), Czech zoologist and mineralogist Svatopluk Presl, who not only presented one of the earliest Slavic letter statistics, but also considered to CV proportion to be an index at a language's euphony and degree of learning difficulty.

Among those early scholars interested in CV proportions also was the German-American Francis Lieber, whose linguistic contributions have almost been forgotten by today's linguistic audience.


ENCYCLOPÆDIA AMERICANA.

A

POPULAR DICTIONARY

OF

ARTS, SCIENCES, LITERATURE, HISTORY, POLITICS AND BIOGRAPHY,

BROUGHT DOWN TO THE PRESENT TIME;

INCLUDING

A COPIOUS COLLECTION OF ORIGINAL ARTICLES

IN

AMERICAN BIOGRAPHY;

ON

THE BASIS OF THE SEVENTH EDITION OF THE GERMAN

CONVERSATIONS-LEXICON.

EDITED BY

FRANCIS LIEBER,

ASSISTED BY

E. WIGGLESWORTH.

Vol. III.

Philadelphia:

CAREY AND LEA.

SOLD IN PHILADELPHIA BY E. L. CAREY AND A. HART.—IN NEW YORK BY G. & C. & H. CARVILL.—IN BOSTON BY CARTER & HENDEE.

1830.

Francis Lieber (March 18, 1800 – October 2, 1872), originally known by his German name Franz Lieber, was a German-American publicist, jurist, and political philosopher. Born in Berlin, he joined the Prussian Army during the Napoleonic Wars, and was wounded at Waterloo. Returning to Berlin after the Napoleonic wars, he attempted to pass the entrance exams to the University of Berlin; but being member of the Berliner Burschenschaft, a student fraternity, inspired by liberal and nationalistic ideas, which opposed the Prussian monarchy, he was denied admission. Moving to Jena he matriculated in 1820 to the University of Jena, and within a span of four months finished writing a dissertation in the field of mathematics. As the authorities caught up with him, he left Jena for Dresden to study topography, but as soon as the Greek Revolution of 1821 broke out, he volunteered his services. Lieber left Germany forever in 1825; for a short time he resided as a teacher in London, and in 1827 he embarked for the United States. During the next five years, during his residence in Boston (1827-32), he was occupied with the compilation of the 13-volume *Encyclopedia Americana: A popular dictionary of arts, sciences, literature, history, politics and biography, brought down to the present time; including a copious collection of original articles in American biography*. The encyclopedia was based on the 7th edition of the German *Brockhaus Conversations-Lexicon*, which had appeared in 1827 under the title *Allgemeine deutsche Real-*

Encyklopädie für die gebildeten Stände (Conversationslexikon), and which was, after *Dobson's Encyclopædia* (1789–1798), the first significant American encyclopedia.

In 1835, Lieber moved to Columbia, S.C., where he occupied the position of professor of political economy in the South Carolina College for twenty years; and here he produced his most important works: *A Manual of Political Ethics* (1838); *Legal and Political Hermeneutics* (1839); and *Civil Liberty and Self-Government* (1852). In 1856, he was called to Columbia College, New York, to take the chair of political economy, and in 1860 accepted the chair of political science in the Columbia Law School, giving up his chair of economics. He was the author of the *Lieber Code* during the American Civil War, also known as *Code for the Government of Armies in the Field* (1863), which laid the foundation for conventions governing the conduct of troops during wartime. Lieber died in New York, September 2, 1872.

Among Lieber's numerous works are a number on language, only three of them having been published, however. The most important in this respect are considered to be his 1837 article "On the Study of Foreign Languages", his 1850 contribution "On the Vocal Sounds of Laura Bridgeman", and his 1852 "Plan of Thought of the American Languages".

In the first of these articles, Lieber defended the teaching of the classical languages at schools; it is Lieber's discussion of the nature of Native American languages that had a lasting influence. In this article, as well as in the shorter one from 1852, Lieber praised native American languages, compared them favorably with the classical languages, and coined the term 'holophrastic' to describe their agglutinating or polysynthetic nature. Lieber explained this phenomenon in his 1837 and 1852 articles, discussed previous terms used to describe it (including agglutinative and polysynthetic), and explained why he considered his coinage 'holophrastic' to be a superior term, which indeed was used in works on Native American languages during the remainder of the nineteenth century.

Lieber's 1850 article is about the vocal sounds of Laura Bridgman, who is known as the first deaf-blind American child to successfully gain a significant education in the English language, some fifty years before the more famous Helen Keller. Bridgman was taught tactile finger spelling becoming completely fluent in it; she was not taught oral language and could only make a limited range of vocal sounds. She could communicate rapidly with anyone else who knew tactile finger spelling. Lieber's article was unique for its time.

In addition to the three published articles described above, and in addition to various unpublished articles, most probably many language-related articles in the *Encyclopedia Americana* were written by Lieber, although the contributions to the *Encyclopedia* were unsigned. One of these contributions to the third volume (1830) is on "Consonants". Among others, a number of calculations concerning the CV proportions of different languages are presented for comparative purposes, and it seems that these statistics, along with their interpretations, are among the earliest of this kind we know of.

According to Lieber (ibd., 450), the "various interesting relations of consonants to vowels, and of the sounds and letters in the different idioms, have not yet received any satisfactory investigation […]."

For Lieber, the study of euphony, or harmony, was a central concern in his cross-linguistic analyses of CV proportions: far away from saying that the euphony of a language depends entirely on this proportion (ibd. 450), Lieber was convinced of the fact that the "melodious sound or music of a language depends, in part, upon the proportion of the vowels to the consonants, a language becoming too hard if there are too many consonants" (ibd.).

In order to establish the CV relations of different languages, Lieber did not base his analyses on the paradigmatic level of inventories, but took passages from different texts, thus integrating the analysis of frequency of occurrence. Again, the author was well aware of a number of possible methodological problems. Attention was paid, among others, to text size: "The different passages were very similar in size, so that the number of syllables in each

would be very nearly the same." And although the author could not pay attention to further possibly intervening factors, he was at least well aware of the fact that the choice of different text types might result in possible differences: "To give anything like accuracy to such investigations, it is obvious that the results ought to be taken both from prose and poetry, also from many different writers, and the language of conversation" (ibd., 451).

For English, Italian, German, Portuguese and Spanish texts, three stanzas were taken from each of the following poems: the beginning of Lord Byron's *Childe Harold*, Torquato Tasso's *Gerusalemme Liberata*, Goethe's *Zueignung* (prefixed to his *Faust*) the *Luisiada* by Luís Vaz de Camões, and the Spanish epic poem *La Araucana by Alonso de Ercilla.* For French, he took 24 lines of the beginning of Racine's *Thébaïde*; for Greek (Ionic), 24 hexameters of the beginning of the *Odyssey*, and for the Attic dialect, the beginning of the *Anabasis*; for Latin, the 24 first hexameters of Ovid. In addition to these languages, Lieber offered data for Hawaiian (still termed Sandwich islands in the 18th century tradition of James Cook), Seneca Indian, Chahta Indian, Sanscrit, Malay, Persian, Hebrew, and common Arabic.

For some languages, Lieber reported separate results for what he termed 'orthographic proportion' vs. 'phonic proportion': those languages which he assumed to be characterized by an approximately 1:1 sound-letter relation, are counted by letters, all others by sounds (a sound possibly being represented by more than one letter).

Starting with an analysis of the *Odyssee*, a text in the Ionic dialect of Greek, Lieber found a CV proportion of 3:4, which he considered to be "a very melodious proportion" (ibd. 451). Comparing it to the results for the Attic dialect, for which he found a CV proportion of 1:1.006, he stated a difference of 0.327:

Ionic	=	3 : 4	=	1 : 1.333
Attic	=		=	1 : 1.006

				0.327

Similarly comparing Latin (with a CV relation of 6:5) to Italian (11:10), he found a difference of 10% between both languages. Table 1 contains the results for all languages as represented by Lieber, the data marked by an '*' being based on what Lieber termed 'phonetic proportions'; additionally, the last column contains the CV quotient based on the data given.

By way of a conclusion, Lieber arrived at the result that not only languages seem to be characterized by different CV proportions, but also do languages belonging to a common family seem to follow similar patterns. According to the author, it can easily be seen "that, in the languages of Latin origin, the proportion of consonants to vowels is much smaller than in the Teutonic idioms" (ibd., 452). But he was well aware of the pioneering state and limited reliability of his analyses and results. With due caution he frankly admitted that "the conclusions […] are rather to be regarded as indications of what might be learned from more thorough inquiries, than as facts from which general deductions can be safely drawn" (ibd., 451).

As a consequence, instead of jumping to hasty conclusions, Lieber (ibd., 452f.) suggested some kind of research program, including tasks as the following:

> "to compare the proportions of consonants to vowels, in such different families of languages; to show the proportions of the gutturals, labials, &c., of the different idioms; and, again, the proportion of these letters in the various families of languages, or according to the different parts of the earth to which they belong, as Asiatic, European, &c. languages, and many other calculations."

Table 1

Consonant-Vowel proportions for different languages (Lieber 1830)

		C	**V**
Sandwich islands		1	1.800
Greek	Ionic	1	1.333
	Attic	1	1.006
Portuguese		1.020	1
Common Arabic		1.080	1
Italian		1.100	1
Seneca Indians		1.180	1
Chata Indians		1.200	1
Sanscrit	*	1.200	1
Latin		1.200	1
Hebrew	*	1.200	1
Spanish		1.240	1
Persian	*	1.330	1
Malay	*	1.330	1
French	*	1.340	1
	orthographic	1.270	1
Dutch		1.500	1
English	*	1.510	1
	orthographic	1.520	1
Swedish		1.640	1
German	*	1.700	1
	orthographic	1.640	1

In his concluding remarks, But Lieber (ibd., 453), methodologically generalized and embedded his approach, referring to Duponceau's (1818) ruminations on English phonology:

> "We have no doubt that the more the science of languages is developed, the more obvious will be the necessity of the study of *phonology* […] the knowledge of the sounds produced by the human voice." And he was, on the one hand, a child of his time, but much ahead of his time, on the other, when he compared the contours of this field of phonology to be developed to contemporary approaches in chemistry, particularly stoichiometry[7]: "This branch of philology might be compared to the new department of *stœchiometry* in chemistry, which treats the proportions of the quantities of the elements in a state of neutralization or solution – a branch of science which everyday becomes more important […]".

[7] Stoichiometry is that branch of chemistry which deals with the relative quantities of reactants and products in chemical reactions; in this context, Lieber explicitly refers to relevant works of contemporary scholars such as Martin Heinrich Klaproth (1743-1817), Jöns Jakob Berzelius (1779-1848), and Johann Wolfgang Döbereiner (1870-1849).

With these remarks and his understanding of phonology, Lieber was much ahead of his time. Moreover, Lieber, with this short contribution, laid the foundations to make the calculation of CV proportions useful for issues of linguistic typology. This relates not only the cross-linguistic typology of different languages: with his remarks on making separate analyses for different kinds of texts (restricted, admittedly, to the rough juxtaposition of prose vs. poetry), this concerns intra-lingual specifics of text typology, as well. Both questions continue to play an important role till our days.

References

Duponceau, Peter S. (1818). English Phonology; Or, an Essay towards an Analysis and Description of the component sounds of the English Language. In: *Transactions of the American Philosophical Society, New Series, Vol. 1*; 228-264.

Grzybek, Peter (2006). A Very Early Slavic Letter Statistic and the Czech Journal 'Krok' (1841). Jan Svatopluk Presl (1791-1849). In: *Glottometrics*, 12; 88-91.

Harley, Lewis R. (1889). *Francis Lieber. His life and political philosophy*. New York: The Columbia University Press.

Isačenko, Aleksandr V. (1939/1940). Versuch einer Typologie der slavischen Sprachen. In: *Linguistica Slovaca*, 1/2; 64-76.

Kelih, Emmerich (2010a). Vokal- und Konsonantenanteil als sprachtypologisches Merkmal slawischer Literatursprachen. In: Fischer, Katrin B. et al. (eds.): *Beiträge der europäischen slavistischen Linguistik (Polyslav 13)*. München: Sagner; 70-77. [= Die Welt der Slaven Sammelbände; 40].

Kelih, Emmerich (2010b). „Wortlänge und Vokal-Konsonantenhäufigkeit: Evidenz aus slowenischen, makedonischen, tschechischen und russischen Paralleltexten." In: *Anzeiger für Slavische Philologie*, 36; 7-27.

Krámský, Jiří (1946-1948). Fonologické využití samohláskovych fonémat. In: *Linguistica Slovaca*, 4-6; 39-43.

Lieber, Francis (ed.) (1827-1832). *Encyclopædia Americana. A Popular Dictionary of Arts, Sciences, Literature, History, Politics and Biography, Brought down to the present time; including a copious collection of original articles in American biography; on the basis of the seventh edition of the German Conversations-Lexicon*. Philadelphia: Carey and Lea.

Lieber, Francis (1837). Consonants. In: *Encyclopædia Americana.* Philadelphia: Carey and Lea. Vol. III, 449-453.

Lieber, Francis (1837). On the Study of Foreign Languages, Especially of the Classic Tongues: A Letter to Hon. Albert Gallatin. In: *Southern Literary Messenger* III: 162-172. [Expanded version reprinted in: *The Miscellaneous Writings of Francis Lieber*, vol. 1. Philadelphia: J.B. Lippincott, 1881; 499-534.]

Lieber, Francis (1850). On the Vocal Sounds of Laura Bridgeman, the Blind Deaf Mute at Boston: Compared with the Elements of Phonetic Language. In: *Smithsonian Contributions to Knowledge* 2: 3-32. [Expanded version reprinted in: *The Miscellaneous Writings of Francis Lieber,* vol. 1. Philadelphia: J.B. Lippincott, 1881; 443-497.]

Lieber, Francis (1852). Plan of Thought of the American Languages. In: Schoolcraft, Henry (ed.), Historical and statistical i*nformation respecting the history, condition, and prospects of the Indian Tribes of the United States…* Vol. 2, Philadelphia: Lippincott, Grambo, 1851-1857; 46-349.

Mack, Charles R.; Lesesne, Henry H. (eds.) (2005). Francis Lieber and the Culture of the Mind: Fifteen Papers Devoted to the Life, Times, and Contributions of the Nineteenth-

century German-American Scholar, with an Excursus on Francis Lieber's Grave: Presented at the University of South Carolina's Bicentennial Year Symposium Held in Columbia, South Carolina, November 9-10, 2001. Columbia, S.C.: University of South Carolina Press.

Skalička, Vladimír (1966). Ein typologisches Konstrukt. In: *Travaux Linguistiques de Prague*, 2; 157-163.

Ycart, Bernard (2013). Letter counting: a stem cell for Cryptology, Quantitative Linguistics, and Statistics. In: *Historia Linguistica*, 40(3); 303-329. [= http://arxiv.org/abs/1211.6847]

Ycart, Bernard (2013). Alberti's letter counts. In: *Literary and Linguistic Computing*. [In print.– http://arxiv.org/abs/1210.7137]

http://archive.org/stream/encyclopaediaame03liebiala#page/n7/mode/2up

Review

Barry P. Scherr, James Bailey, Evgeny V. Kazartsev (eds.). *Formal Methods in Poetics: A Collection of Scholarly Works Dedicated to the Memory of Professor M.A. Krasnoperova*. RAM-Verlag, Lüdenscheid (Germany), 2011. 315 pp.

Reviewed by **Michael Wachtel,** Princeton University (USA)

The present collection of essays contains a range of approaches, all of which seek to apply strict methods of analysis (often statistical) to verse. Given the lengthy and distinguished history of such approaches in Russia, it is not surprising that most of the essays are devoted to Slavic poetry, with the exceptions usually authored by Slavs. The fact that the entire volume is written in English suggests that the editors (two of whom are American) wish to reach an audience beyond Eastern Europe. Certainly someone has put significant effort into English style; on the whole, the essays read exceedingly smoothly, and typos are few. However, the frequent use of probability theory and the esoteric national traditions involved (at least from the perspective of the Anglophone reader) may make the volume inaccessible to a broad readership. This would be a pity, because the quality of scholarship is often on a high level.

Empirical approaches to verse are sometimes taken to task for losing sight of aesthetic questions. Yet as many of the contributors to the volume show, there is no reason why the two cannot be combined in meaningful ways. And while the "scientific" tone may strike the un-initiated as dry, the advantage of formal approaches is that the authors lay out with unusual clarity the problem, the methodology, and the conclusions. As a result, even the reader who lacks the mathematical background to understand the details (for example, this reviewer) can usually follow the aims and accomplishments of the work.

Among the numerous "formal methods," some authors concentrate their analysis on a single text, filtered through a broad range of quantitative data. Several use comparative metrics to study translation. Still others cover the entire metrical repertoire of a given poet, or a significant subset of a poet's work. The collection is divided into two parts, the first devoted to Russian verse, the second to "European" verse. There would have been better ways to organize this material, e.g., the Klenin and Plungian contributions should have been adjacent rather than in different sections. However, rather than reorganize the volume now, my comments below treat the essays in the order they appear with the minor exception that I discuss two essays by the same author in the same paragraph.

Evgeny Kazartsev begins the volume by investigating the relationship of Lomonosov's iambs to those of Johann Christian Günther, a German poet whose works have long been recognized as having influenced Lomonosov. Based on a rigorous analysis of rhythmical patterns, Kazartsev concludes that Lomonsov used his German counterpart as a model not only in the earliest poems, but even later, where he seems to have followed Günther's lead in terms of the relationship of rhythm and genre. One question that Kazartsev does not address, however, is how to deal with German secondary stress. In his view, the lines: "Наследник имени и дел" and "Wer lehrt dich, tumme Tyraney?" are rhythmically identical. However, a German would surely read the lines differently, putting two stresses on "Tyraney" and thus making it a line without pyrrhics. Now, it may well be that Lomonosov spoke German without secondary stress (as Nabokov apparently spoke English), but this is an issue that at the very least needs to be considered.

In the next essay, Andrew Davis applies statistical analysis of punctuation to study the syntax of the Onegin stanza. His approach draws on work begun by Tomashevsky and Vinokur and continued most recently by Barry Scherr. While the importance of the fourth line as a syntactic border has long been recognized, Davis shows among other things how the rest

of the stanza's syntax is dependent on that fourth line. When the fourth line breaks neatly, there is a strong tendency for the eighth line to do so as well. However, when the fourth line lacks a syntactic break, the eighth becomes even less likely to close a segment. This explains why scholars who wish to see the Onegin stanza as a type of sonnet can easily find material to substantiate their claim, but it likewise shows why this claim is essentially false. Davis also discusses the syntax of later Onegin stanzas, showing, for example, that Lermontov's work in this form is stricter than Pushkin's, while Vikram Seth's novel *The Golden Gate* is still freer.

Sergei Andreev's essay draws on the methodology of M.A. Krasnoperova, memorialized in the volume's subtitle. This means that the approach is marked by higher mathematics, a challenge to the average student of poetics. Still, Andreev explains his approach for those who may not be able to follow the details. The basic idea is to investigate Tiutchev's lyric poems (in this case, those in iambic tetrameter quatrains) in terms of many formal features (rhythm, syntax, rhyme) and see how they correlate with each other. For example, even without statistical analysis one might assume that a stressed anacrusis tends to correlate with an unstressed first ictus, but it is hardly obvious how this might affect later ictuses or how it might vary from line to line. Andreev takes this data and divides it by periods, supporting a contention (made in his earlier work) that Tiutchev's later poetry differs from that of his early period. The essay offers a significant amount of information, though its significance will only become fully apparent in the context of a much larger data set (e.g., similar studies of other nineteenth-century Russian poets).

Emily Klenin's article concerns Fet's translation of Goethe's *Faust*. As a native speaker of German, Fet was especially sensitive to the rhythmic and metrical nuances of the original. He was the first Russian translator to retain the more unusual formal features, for example, the *Knittelvers* of Faust's opening monologue. (In Goethe's usage — which only vaguely recalled the medieval verse that goes by this name — *Knittelvers* was a somewhat irregular four-stress meter.) Given that Fet preserved this form in that famous opening soliloquy, the question arises as to why he did not do so in Faust's "Osterspaziergang" speech. The answer that Klenin suggests is that Fet — like some recent German metricians — did not read the "Osterspaziergang" passage as the same *Knittelvers* as the opening. Hence his rendering, which mixes binary and ternary meters, while not slavishly following the rhythm line by line, nonetheless creates an equivalent rhythmical texture.

Venera Kayumova's essay is devoted to a very important and rarely studied formal feature: hypermetrical stress. Using an enormous set of data (36,000) lines, she graphs the frequency of hypermetrical stresses in the iambic tetrameter of Russian poets over 140 years (1880–1920). Most incidences occur on the very first syllable of the line, and, statistically speaking, hypermetrical stressing decreases from the beginning to the end of the line. The most common type of hypermetrical stressing is in the form of a spondee (rather than that of a choriamb). Since the incidence of hypermetrical stress is far greater than theoretical predictions would lead us to expect, it can be safely assumed that hypermetrical stress is more a conscious decision on the part of the poet than an inevitable feature of Russian phonology.

In a staggeringly thorough analysis of Maksimilian Voloshin's poetry, Igor Karlovsky addresses the issue of free verse. Voloshin is often mentioned in this context, but Karlovsky shows that there are no finished poems by Voloshin that unambiguously meet the criteria for free verse. He suggests (logically enough) that Voloshin's lengthy stay in France influenced his untraditional — and relatively free — experiments with verse form. However, given Voloshin's famously poor knowledge of French language, it is probably no coincidence that he only began these experiments after he returned to Russia and was privy to Vyacheslav Ivanov's theories (and lectures) on poetics. Karlovsky not only surveys the variety of Voloshin's unusual forms (unrhymed variable meters, use of trochaic lines in iambic verse, etc.), but also connects them to genre, offering valuable lines of inquiry to future scholars.

Georgii Vasiutochkin has supplied two essays, both of which apply formal methods of poetic analysis to the interpretation of an individual text. The first, on Khodasevich's "Ne iambom li chetyrekhstopnym," begins with the well-established (by Andrei Belyi, then by Kirill Taranovsky) distinction between the iambic tetrameter of eighteenth-century poets and that of nineteenth-century poets. Placing Khodasevich's poem in this broad context, Vasiutochkin shows its rhythmic profile to be distant from Khodasevich's own earlier poetry and from that of the age of Pushkin, yet strikingly reminiscent of eighteenth-century practice. The closest fit is not Lomonosov's "Khotin Ode" (invoked in Khodasevich's poem), but Derzhavin. Some of Vasiutochkin's claims can be disputed; it is not clear why "Videnie murzy" should be singled out as the source rather than Derzhavin's verse more generally. Moreover, Vasiutochkin is apparently not aware that Khodasevich's poem is unfinished, a consideration that inevitably complicates a minute empirical analysis. Finally, one can argue that the sampling of a single poem is too small to be statistically significant. (And Vasiutochkin's rhythmical assumptions are sometimes debatable. To his credit, though, he shows precisely how he scans the poem.) Nonetheless, the statistics are so overwhelmingly revealing that even questioning a few stresses here and there would not change the overall picture. Vasiutochkin's other essay is similar in methodology; he seeks a Russian model for the trochaic hexameter of Joseph Brodsky's "Letters to a Roman Friend." His (unexpected) conclusion is that there was no specific model. Rather, he argues that Brodsky developed an odd rhythmic profile in an attempt to echo (albeit loosely) the metrics of antiquity, in particular Martial's Phalacian verse.

Barry Scherr's contribution concerns the sonnets of Arsenii Tarkovskii, a form that the poet turned to throughout his life and in this sense a microcosm of his poetry and poetics. Scherr shows how a supposedly "traditional" poet varies the sonnet form in terms of rhythm and even meter (with one anapestic sonnet), with the changes becoming more apparent as he grew older. He also shows the "border phenomena," poems that may be sonnets but cannot clearly be classified as such. Though the body of work is relatively small and some of the statistical evidence therefore insufficient, the essay is helpful both in understanding Tarkovskii's place in twentieth-century Russian verse as well as in understanding the sonnet more broadly.

Saule Abisheva's essay on the "classical" (i.e., syllabotonic) verse of David Samoilov is likewise an attempt to show how a traditional poet expresses his individuality. By studying all 9,000 lines of relevant verse, Abisheva can be more precise than previous analyses (of such authorities as Pavel Rudnev and M.L. Gasparov), since these were based on sampling. In many respects, however, the results still show that this verse is typical in its rhythmical features or, in the author's words, "an inconspicuous striving for heterogeneity" (p. 158). Still, there are some oddities (e.g. unstressed ictuses in ternary verse) that stand out and deserve attention.

Svetlana Efimova's work on the metrical and stanzaic forms of Konstantin Vasiliev (1955–2001) is similar in approach to Abisheva's essay. Vasiliev was obscure during his lifetime, but has apparently become quite celebrated since his death. Her brief discussion, which suggests a periodisation of his work as well as a delineation of the distinctive (as well as typical) features of his poetry, is followed by a series of charts and graphs twice its length. The main problem of the work of Efimova (and Abisheva) is that ultimately a computer could produce it (see comments below on Ibrahim and Plecháč). On the one hand, there is something reassuring about an essay in which every statement can be empirically verified. On the other hand, there is something troubling about this very phenomenon. M.L. Gasparov edited a compendium of the metrical repertoire of nineteenth-century poets, and the volume can be useful. However, it does not pretend to be poetic analysis, but merely to serve as raw material for future studies. The work of Abisheva and Efimova presumably also belong to this

genre, yet many other poets would require this same detailed analysis before the results can be appreciated.

These essays conclude the first part of the book ("Russian Verse"). The second part ("European Verse") contains a number of exceptionally fine contributions. In the opening article, Vladimir A. Plungian discusses a single brief poem by Robert Louis Stevenson and its "classic" Russian translation by Andrei Sergeev. Plungian explains that the translator's equimetrical solution masks a profound problem. By translating an English poem in "dol'nik" into a Russian poem in the same form, the translator cannot but create an anachronism, since "dol'nik" first came into widespread use in Russia in the early twentieth century. Thus Stevenson's simple poem sounds to a Russian ear like the work of Gippius or Blok. Plungian also looks at small semantic differences and shows how these subtly change the tone of the poem and align it with the same literary-historical problems as the meter. The essay combines synchronic and diachronic analysis in a convincing and accessible way.

Reuven Tsur devotes his article to the question of performance, to the way actual readers realize the written text. Tsur makes a number of superlative points on the interplay of syntax and prosody, with concise and insightful discussion of formal devices (e.g., caesura, enjambment). These observations are then applied to specific readings of Milton's sonnet "On his Blindness." Because Milton's complicated syntax is often at odds with the formal boundaries of the sonnet, the poem serves as a particularly good example of the difficulties facing a reader. Tsur uses audio processors to study readings that can be found on the internet. His point is not to argue for a single correct reading (though he does find fault with some aspects of each reader), but rather to show how each reader makes use of the tension between syntax and versification. Such an approach does not particularly help in interpretation (those hoping to understand what Milton is actually saying will not find much guidance here), but it does nonetheless contribute a great deal to an understanding of how poetry communicates in the broadest sense.

Alfred Behrmann's essay is one of the most accessible in the volume and also one of the finest. Using a few well-chosen passages from Shakespeare, he investigates the role of paradox, in this case the use of prose versus poetry, a subject far less obvious than is often assumed. In a wonderful discussion of the famous scene after Julius Caesar's death, Behrmann shows how the "poetic" Brutus uses prose, while the prosaic Mark Antony speaks in verse, all the while insisting that he lacks the skill to do so. (It might be added, however, that Brutus' prose has strongly poetic elements, e.g. the repeated "If any, speak, for him I have offended," which scans as an impeccable iambic pentameter line.) Likewise, in "Love's Labour's Lost," a veritable treasure trove of poetic forms, the lowliest characters display no less poetic resourcefulness than the learned ones.

In the tradition of the Abisheva and Efimova essays, Vadim Andreev's article consists largely of a catalogue of the work a single poet. In this case, however, the poet is American (Edgar Allen Poe), the verse is limited to iambs, and the purpose is not simply descriptive, but diachronic. In other words, the author is interested in delineating changes that occur in the poet's work over time. Especially impressive is the number of verse features that Andreev has charted; 34 in all, among them not just rhythm, but also morphology and syntax. Since Andreev includes very few methodological examples (i.e., how these characteristics are defined), a lot has to be taken on faith. Potential problems are legion; for example, though the English iamb tolerates stress reversal and is thus much freer than the Russian, Andreev's rhythmic analysis seems to have been taken wholesale from that traditionally used to study Russian verse. In any case, without seeing the evidence, it is difficult to evaluate Andreev's conclusions. On the basis of stylistic markers, he claims to have delineated a system so refined that it would allow him to give approximate dates to works by Poe that have heretofore lacked reliable dates of composition.

The work of Viktor Levitskii and Olga Naidesh is devoted to the "phonosemantic" qualities of verse. This term implies something that poets have argued about for centuries, but which verse theorists have been hesitant to accept: namely, that specific sounds have specific semantic associations consistent from poet to poet. The study is based on German-language poetry, but the authors suggest that the same results would be found in other languages as well. The first study traces the appearance of the "r" sound versus the "l" sound. According to the authors, the former appears with much greater frequency (by a factor of ten) in poems in a "minor" key, while the latter appears with much greater frequency in poems of a "major" key. (The major and minor distinctions are "based on the expert opinion of literary scholars" [p. 249].) The authors then build on their initial discovery by investigating consonant clusters that include either the "l" or the "r" phoneme. Even if one accepts their results as correct, the approach suffers from oversimplification. Poetic semantics is reduced to questions of happy versus sad, and even the relative prominence of the phonemes in question is not taken into account. Surely those in stressed syllables and rhyme position should count more heavily than others.

The essay of Mikhail Lotman and Maria-Kristiina Lotman, devoted to the Estonian trochaic tetrameter, would seem on first glance to be of only local interest. However, the essay is outstanding, with potential application to syllabo-tonic verse of numerous national traditions. According to the authors, Estonian verse is remarkable for the tenacity with which it clings to the metrical scheme. Since rhythmic variation is so rare, the authors seek out factors that might allow some discrimination between historical periods and even individual authors. They propose a two-pronged approach. First, they consider the stressing not on the ictuses (as these are virtually always stressed), but on the weak syllables. Second, they introduce the notion of various degrees of stress, ranging from none whatsoever to phrasal stress (the most important syllable in an entire phrase). Admittedly, such an approach introduces an element of judgment and thus departs from the strictly empirical data sought by many formal theorists. However, it has the enormous advantage of reflecting the verse as it is actually recited (an issue that scholars of Russian verse tend to ignore). Most important: when the authors compare their results to that from randomly occurring trochaic passages in Estonian prose, they discover significant differences. And poets of different periods display different proclivities. Hence the authors conclude that Estonian poets indeed are using the rhythmic qualities of their language for aesthetic purposes.

The article by Robert Ibrahim and Petr Plecháč is rather technical, but the main point is clear and significant. Computer studies of verse have up until now been based on organizing information that is painstakingly gathered by hand. For example, the magnificent "natsional'nyi korpus russkogo iazyka" (http://www.ruscorpora.ru) includes basic metrical information on thousands of poems, but the computer has not compiled this data. It is simply a convenient storehouse, a means for allowing a researcher quick access to other scholars' grunt work. Ibrahim and Plecháč are attempting something quite different; they are hoping to have a computer gather the data, not simply organize it. In other words, they are teaching the computer to scan Czech verse. If Czech poetry (or any national poetic tradition) were entirely accessible in this form, it would allow a far more detailed picture than is presently available.

The final contribution to the volume, by Ján Mačutek, exceeds the competence of this reviewer. The essay concerns the rhythmic patterns in five Slovak poems, and the results are apparently language-specific, i.e., they cannot necessarily be transferred to other national traditions. The author, a mathematician, makes demands that few humanists will be able to follow, e.g., his discussion of the "1-shifted right truncated negative binomial distribution." I must leave it to others to determine how successfully this methodology explains the complexities of poetic rhythm.

Taken as a whole, the volume shows that verse theory continues to thrive, even after the deaths of such luminaries as M.L. Gasparov and M.A. Krasnoperova (whose complete bibliography is included [pp. v–xi] after a brief opening appreciation of her work [pp. iii–iv]). It is encouraging to see the variety of approaches in this volume, many of which point in directions well worth further exploration.