universität wien

Dr. Emmerich Kelih

Institut für Slawistik

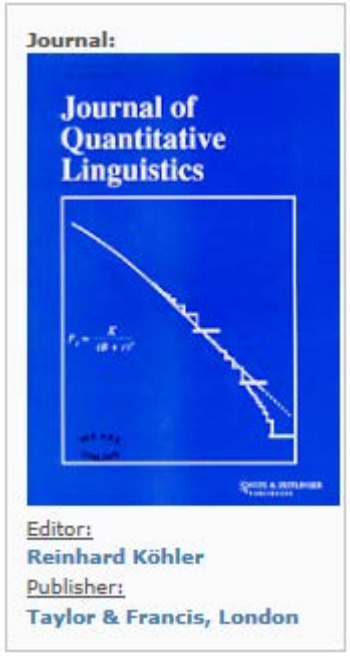
Morphosyntactic encoding strategies - A quantitative perspective

22.08. 2013

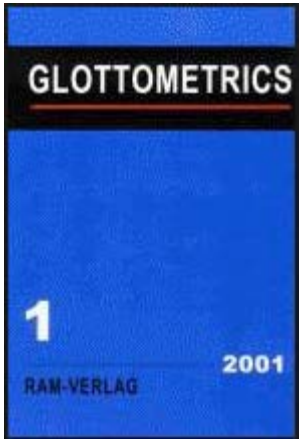
IQLA
International Quantitative Linguistics Association

Quantitative Linguistic as integral part of General Linguistics

since 1994



since 2001



since 2008

IQLA

International Quantitative Linguistics Association

Qualico 2014 (May, 29th – June, 1st in Olomouc/CZ)

QUALICO 2014

20th anniversary of IQLA and Journal of Quantitative Linguistics (JQL)

Olomouc (Czech Republic), May 29 - June 1, 2014.

IMPORTANT DATES

Abstract Submission Deadline	December 15, 2013
Notification of Abstract Acceptance	February 15, 2014
Conference	May 29 - June 1, 2014
Full Paper Submission Deadline	September 30, 2014
Notification of Full Paper Acceptance	November 30, 2014

+ presentation of recent projects (Poster-Session)

Monographs, Bibliography, Book series etc.

Quantitative Linguistics - Brockmeyer (Bochum) 1978-1992

***Quantitative Linguistics* - de Gruyter** (Berlin u.a.)

Studies in Quantitative Linguistics (RAM-Verlag, Lüdenscheid)

Bibliography:

Köhler, R. (1995): ***Bibliography of Quantitative Linguistics.*** Amsterdam/Philadelphia: Benjamins.

Köhler, R.; Altmann, G.; Piotrowski, R.G. (eds.) (2005): ***Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook.*** Berlin, New York: de Gruyter. [= Handbücher zur Sprach- und Kommunikationswissenschaft; 27]

in preparation:
Köhler, R.; Grzybek, P.; Naumann, S. (eds.) (2013ff): *Quantitative und Formale Linguistik.* Berlin u.a.: de Gruyter. [= Wörterbücher zur Sprach- und Kommunikationswissenschaft; 9] [in German, an English version is planed]

Aims and methods of QL:

Analysis and exploration of regularities, tendencies and statistical laws of language and texts

(1) **distributional laws (Zipf's law)**

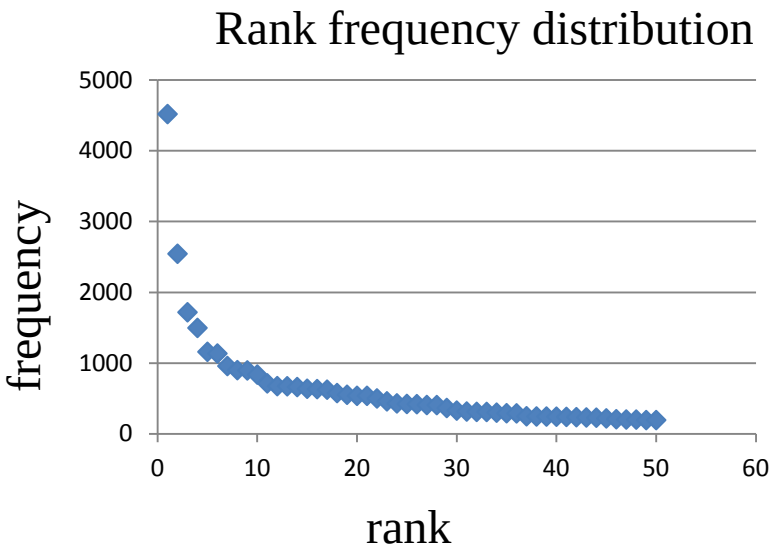
(2) **functional laws (Menzerath's law)**

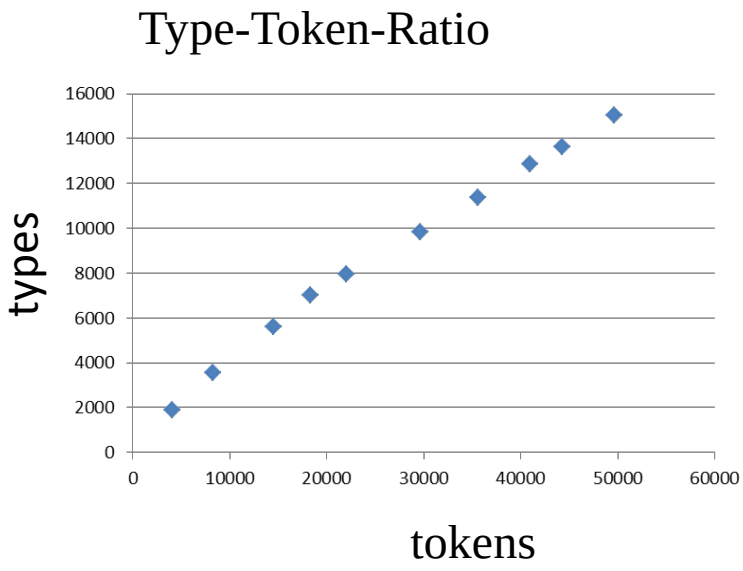
(3) **developmental laws** (**Piotrovskij law:** diachronic development of linguistic entities)

- QL goes beyond description of linguistic phenomena
- QL tries to find stochastic and mathematical models of linguistic entities and phenomena
- QL emphasize and explores the linguistic relevance of frequency/length of linguistic entities generally in **language systems** and **texts**

Todays talk:

1. Introduction to the quantitative analysis of the **lexical structure of texts**





2. Relevance of the **Type-Token-Ratio** for crosslinguistic comparision and measurement of analytism/synthetism

Overview:

1. Morphosyntactic encoding strategies
2. Analytic vs. synthetic languages: An “old school” approach in typology?

3. Possibilities of a quantitative analysis:

3.1. Type-Token-Ratio in Parallel Texts

3.2. Zipf’s law and morphosyntactic properties

3.3. Empirical results (Slavic languages, English, and Chinese)

Language Typology: „Traditional“ morphological classification:

isolating, agglutinative , polysynthetic und fusional languages …

↓

analytic vs. synthethic languages

Analytic:
grammatical information is encoded within one „autonomous“ word form (Hinrichs 2000a)

or:
Redistribution of lexical and grammatical information from the morphological (infusional) level to the syntactical level

Example: Future tense I

‘we will see‘

Russian: Synthetic

po-smotr-im
PRE.ASP.FUT.-see-1PL

Serbian: Analytic

ће-мо	*виђе-ти*
ће-mo	*vide-ti*
AUX.FUT-1PL.PRES	see-Inf.

Areal linguistic typology for Slavic languages (Gvozdanović 2009)

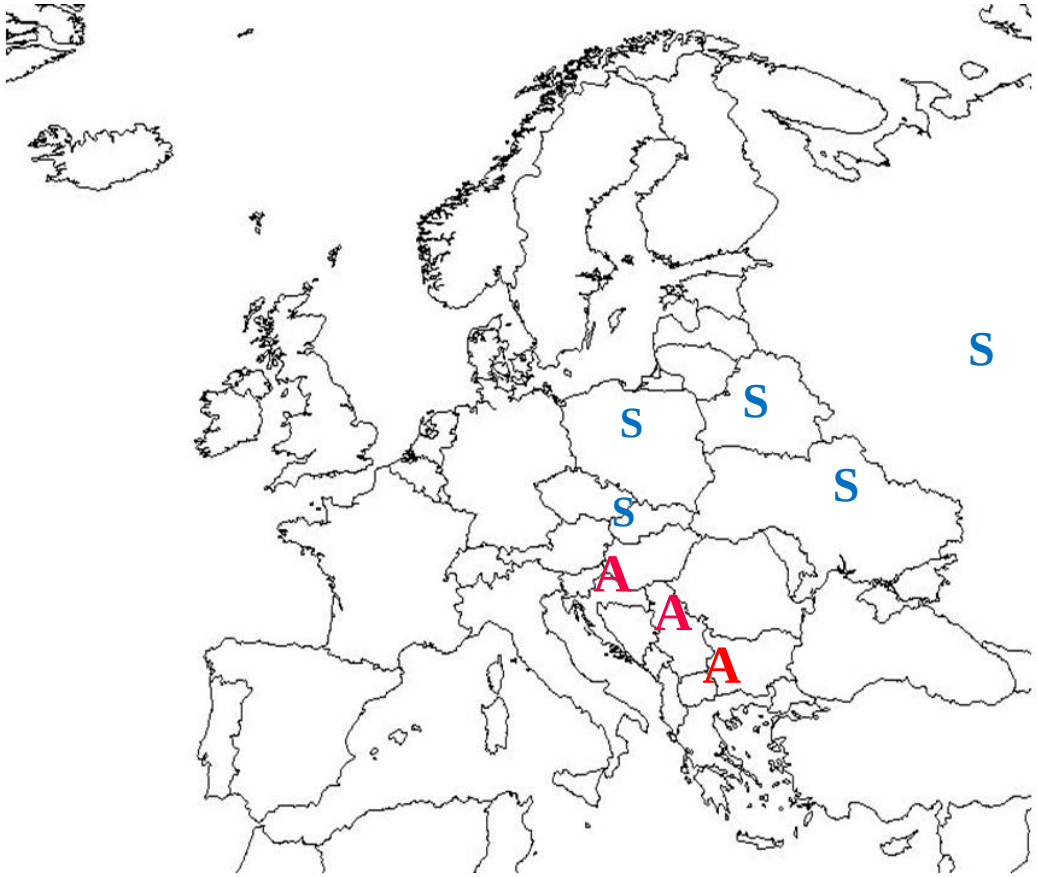
A. Southern areal: Analytic

- South-eastern areal (Bulgarian and Macedonian)
- Serbian as transitional area
- southwestern areal (Slovene, Croatian)

B. **nothern area: Synthetic**

-West Slavic languages

- East Slavic languages



traditional topoi of Slavic and General Linguistics

Synthetic	Transitional area (B/C/S)	Analytic
Russian (East Slavic) Czech (West Slavic)		Bulgarian Macedonian

What about English?

What about Chinese?

Open questions and problems?

- are there „totally“ analytic/synthetic languages?
- Is Russian more synthetic than Czech?
- Is Bulgarian more analytic than English?
- Is Chinese more analytic than English?

======

- which language level should be analysed? (system vs. Text)
- which characteristics are take for the classification?
- what about the degree of (A) and (S) **within** one language?

Example 1: Comparision in Slovene:

I. Synthetic (for frequently used adjectives)

lep-Ø	*lep-š-i*	*naj-lep-š-i*
nice-M.NOM	nice-COMP-M.DEF	SUPERLAT-nicest- COMP-M.DEF
'nice'	'nicer'	'nicest'

bogat-Ø	*bogat-ejš-i*	*naj-bogat-ejš-i*
rich-M.NOM	rich-COMP-M.DEF.	SUPERLAT-rich- COMP-M.DEF.
'richer'	'richer'	'richest'

II. Analytic (usually for polysylabic words, colours and endings with -en, -av, -ast, -a)

zaželen-Ø	*bolj zaželen-Ø*	*najbolj zaželen-Ø*
favored-M.Nom	COMP favored-M.NOM.	SUPERLAT -favored-M.NOM.
'favored'	'more favored'	'most favored'

muhast-Ø	*bolj muhast-Ø*	*najbolj muhast-Ø*
bothersome-M.Nom	COMP bothersome-M.NOM	SUPERLAT-bothersome-M.NOM.
'bothersome'	'more bothersome'	'most bothersome'

Example 2: Future tense in Serbian

I. Analytic: enclitic form (“ short form”) of хтети/hteti („will“) + Infinitive

ми	*ће-мо*	*виђе-ти*
mi	*će-mo*	*vide-ti*
we	AUX.FUT-1PL.PRES.	see-Inf.
'we will see'		

II. Synthetic: Infinitive – ti + enclitic form of хтети/hteti („will“)

виђе-ће-мо	*raђu-ће-me*
vide-će-mo	*radi-će-te*
see.INF-AUX-1PL	work.INF.-AUX-2PL
'we will see'	'you will see'

but: Infinitives with –ći →: доћи ћу/doći ću (orthographical convention!)

III. Analytic: Enclitic form хтети/hteti („will“) + da-construction + Verb

то	*ћемо*	*ђа*	*видимо*
to	će-mo	da	vid-imo
this/that	will-1Pl	da-conjunction	see-1Pl
'that we will see'			

Example 2a: Future tensen in Croatian

I. Analytic: Infinitive –i + enclitic form of hteti

videt	*će-mo*	*radit*	*će-te*
see.INF	AUX-1PL	work.INF	AUX-2Pl
'we will see'		'you will work'	

but: -ti at the infinitve, if immeadetly before the auxiliary verb

Intermediate Results & Desiderata

- different encoding strategies in Serbian and Croatian
- morphosyntactical and phonological determination
- da-construction is considered to be typically Serbian
- no systematic different corpus-based analyses of different approaches
- role of orthographical conventions?
- role of the normative grammar?

Example (3): Prepositional phrases in Russian

I. Analytic

učeb-nik–Ø	*po*	*matematik-e*
schoolbook-NOM.SG.	about.PREP.	mathematics-DAT.SG.
'maths school book'		

plan–Ø	*po*	*vypusk-u*
plan-Nom.Sg.	about.PREP.	production-DAT.SG.
'production plan'		

II. Synthetic

učeb-nik–Ø	*matematik-i*
schoolbook-NOM.SG.	mathematics-GEN.SG.
'maths school book'	

plan–Ø	*vypusk-a*
plan-Nom.Sg.	production-Gen.Sg.
'production plan'	

In between résumé

Within one language one particular grammatical categories can be expressed

1. both analytic and synthetic (!)
2. usage of one these strategies can be determined by stylistic factors
3. usage can be text type specific
4. usage can be determined by phonological or morphosyntactical factors
5. areal and geographical factors can play a role

Further questions and remarks (1):

1. How a language can be characterised as analytic or synthetic?

→ Analysis of **one selected grammatical features** (tense system, expression of modality, comparison … …)

→ Analysis of texts (as conglomerate of different realised encoding strategies), particularly **parallel texts are relevant** for crosslinguistic purposes

Further questions and remarks (2)

How we can determine the degree of analytism/synthetism quantitatively?

Many suggestions and ideas:

1. Greenberg-Index (1960): number of morphemes per word form

   = word length in the number of morphemes

2. number of roots in relation to all morphemes

3. number of words with inflection (see Altmann/Lehfeldt 1973)

4. number of monosyllables in one language

5. frequency of word forms in parallel texts

6. frequency of prepositions and auxiliaries

7. frequency of Hapax Legomena

8. selected parameters from Zipf's law

… …

9. **Type-Token-Ratio** in **parallel texts**

Analysis of parallel texts is a standard tool in

- language Typology,
- crosslinguistic Research
- Quantitative Linguistics
- synergetic linguistics
- …

Some basic problems of parallel texts:

- Simplification (syntactical level (e.g. shorter sentence length in the target text than in the source text), lexical level (smaller number of tokens and types)

- translators' strategy of explicitation (e.g. the process of rendering information that is only implicit in the source text explicit in the target text)

- Missing/or not translated parts (must be checked in the process of alignment)

→nevertheless parallel texts are a reliable recourse for text-based research - a-priori low inter-lingual heterogeneity !

Analysis of the Type-Token-Ratio in parallel texts

-“whole” texts are required

- building up of a parallel text corpus (QUANTA-project, Graz, Vienna)

- includes translations of Kak zakaljalas’ stal’/How the steel was tempered” (KZS) by Nikolaj Ostrovskij and Master i Margarita by M. Bulgakov

- KZS: Socialist realism novel from 1932-1934

- Author: N.A. Ostrovskij

- 10 chapters (scanned, OCR, plain text) for **12 Slavic Languages**: Belarussian, Ukrainian, Russian, Czech, Polish, Slovak, Upper-Sorbian, Bulgarian, Macedonian, Croatian (ijekavian), Serbian (ekavian), Slovenian translations, **English** and **Chinese**

- simple text pre-processing of the texts

- no tagging and no annotation up to now !

- some tentative quantitative studies based on KZS has been performed, among them of the **Type-Token-Ratio**

What is the Linguistic Meaning of the Type-Token-Ratio ?

- vocabulary richness?

- information flow in a text?

→ TTR in parallel texts is an indicator for analytic/synthetic languages
→ TTR can be interpreted as “morphological richness”

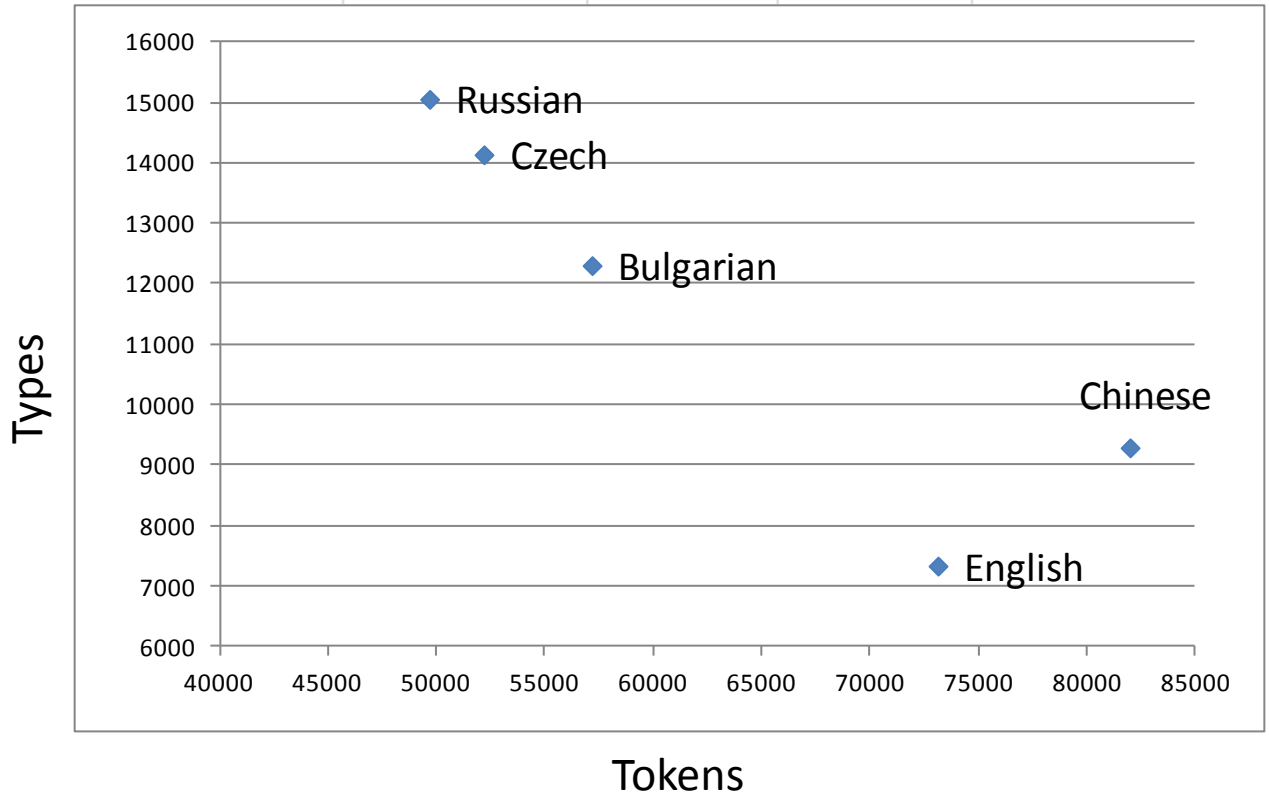
Inflected languages: more types (due to inflection), hence less times the same/identical form appears in a text.

Analytic languages: less word form types in a text, but the same forms are repeated very often.

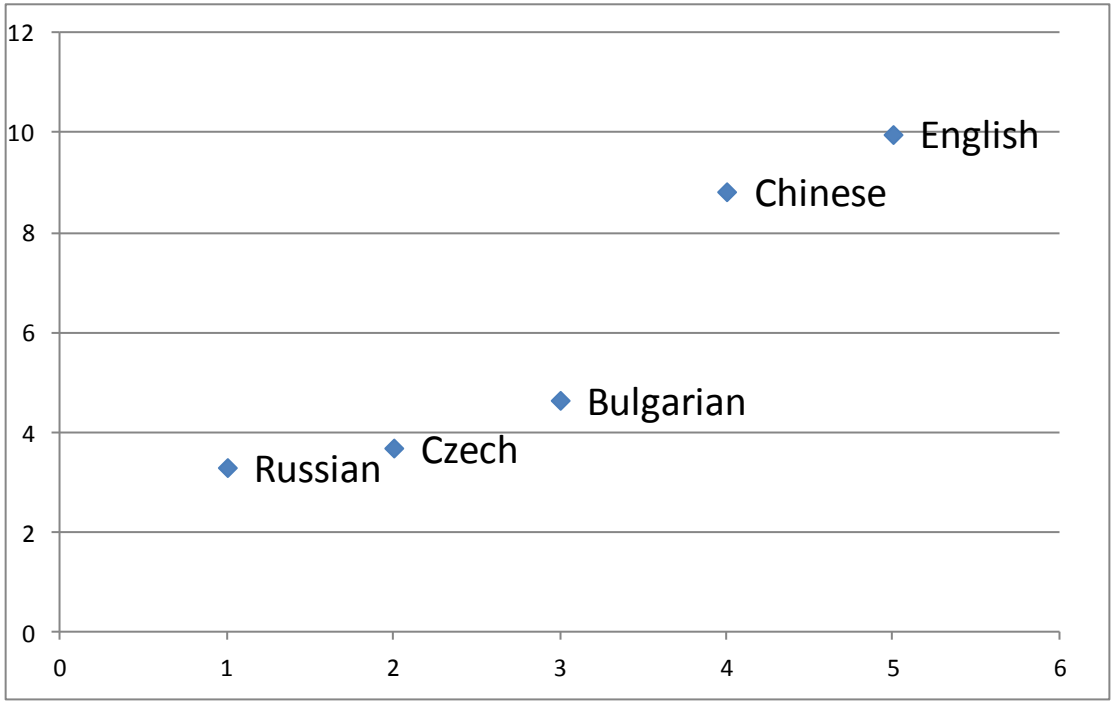
Synthetic languages: more grammatical information within one word form, hence more different word form types are requiered, but the frequency of this types in a text is lower.

First results: Number of Tokens and Types

Language	Tokens	Types
Russian	49672	15053
Czech	52180	14136
Bulgarian	57165	12303
English	73123	7333
Chinese	81982	9289



Type-Token-Ratio = TOKENS/TYPES



Language	TTR
Russian	3,30
Czech	3,69
Bulgarian	4,65
Chinese	8,83
English	9,97

→High TTR for analytic languages, and a low TTR for synthetic languages!

What about the factor text length?

TTR in cumulated chapter of KZS (How the steel was tempered)

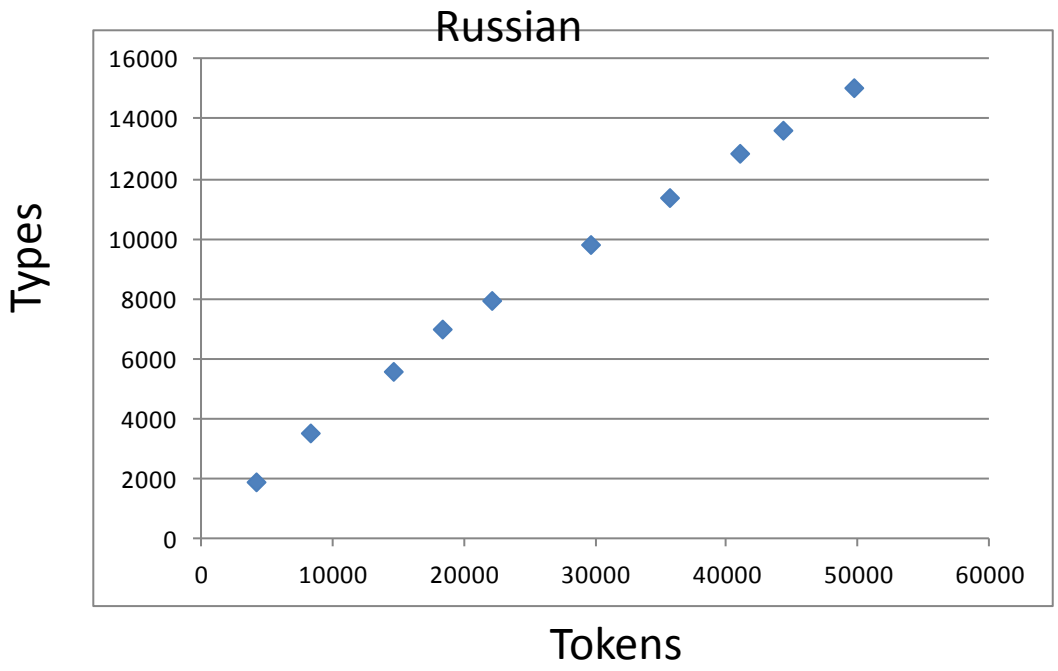
Chapter 1, chapter 1+2, chapter 1+2+3 … …

Modelling the TTR in cumulated chapters:

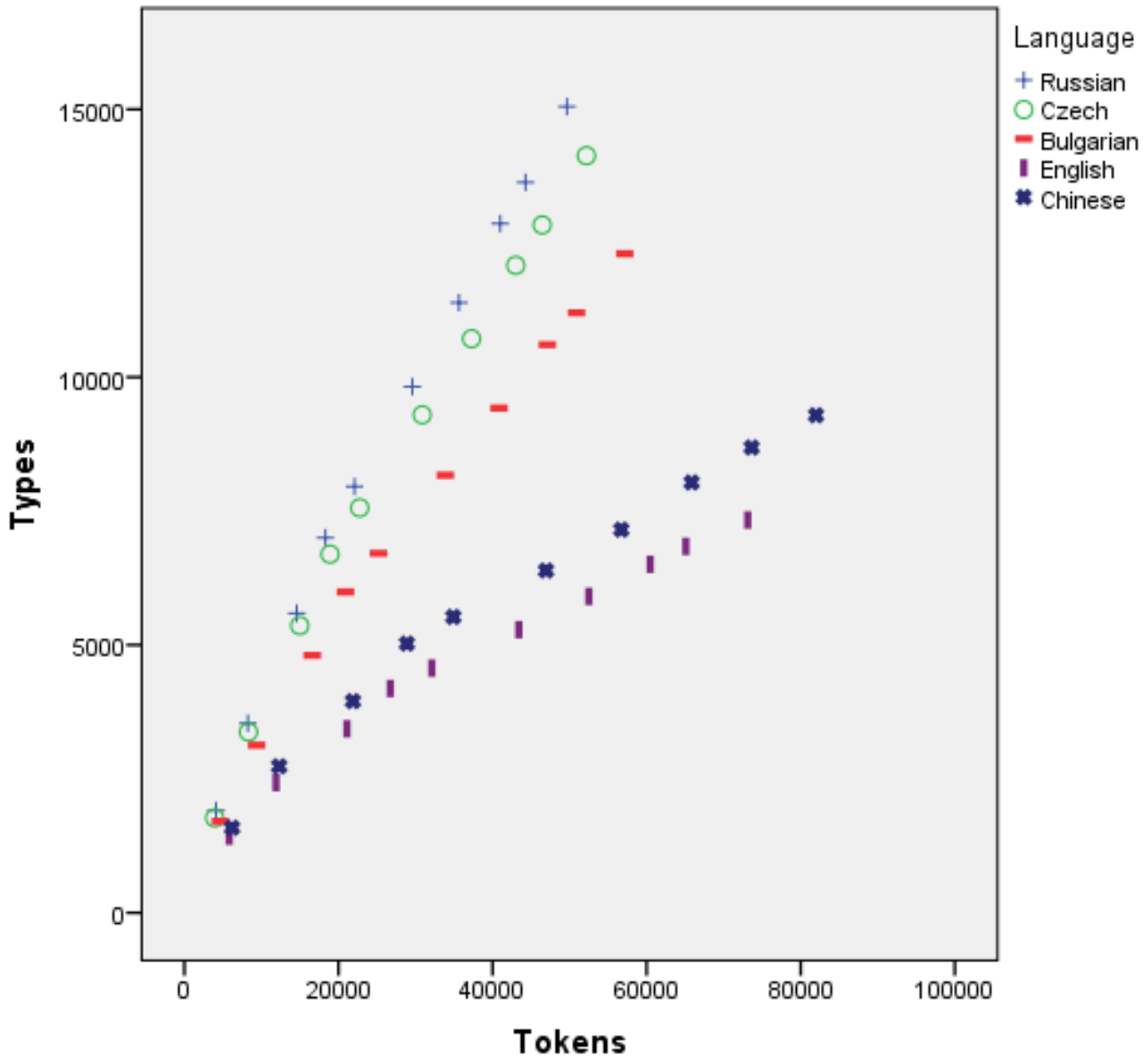
TYPES = a*TOKENS^b

$y = a*x^b$

a = 1



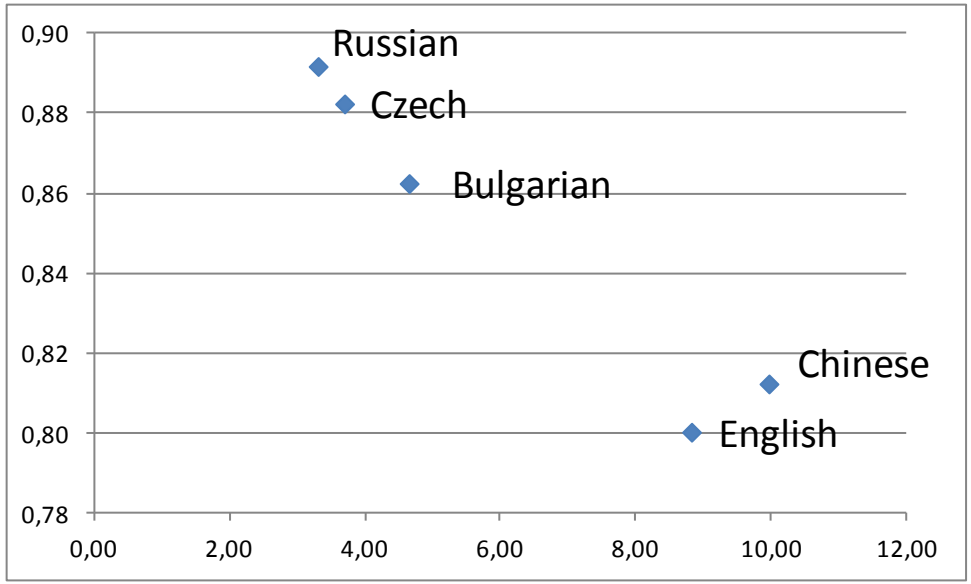
TTR in cumulated chapter of KZS: Russian, Czech, Bulgarian, English, Chinese



Parameter *b* of the TTR in cumulated chapter of KZS (How the steel was tempered)

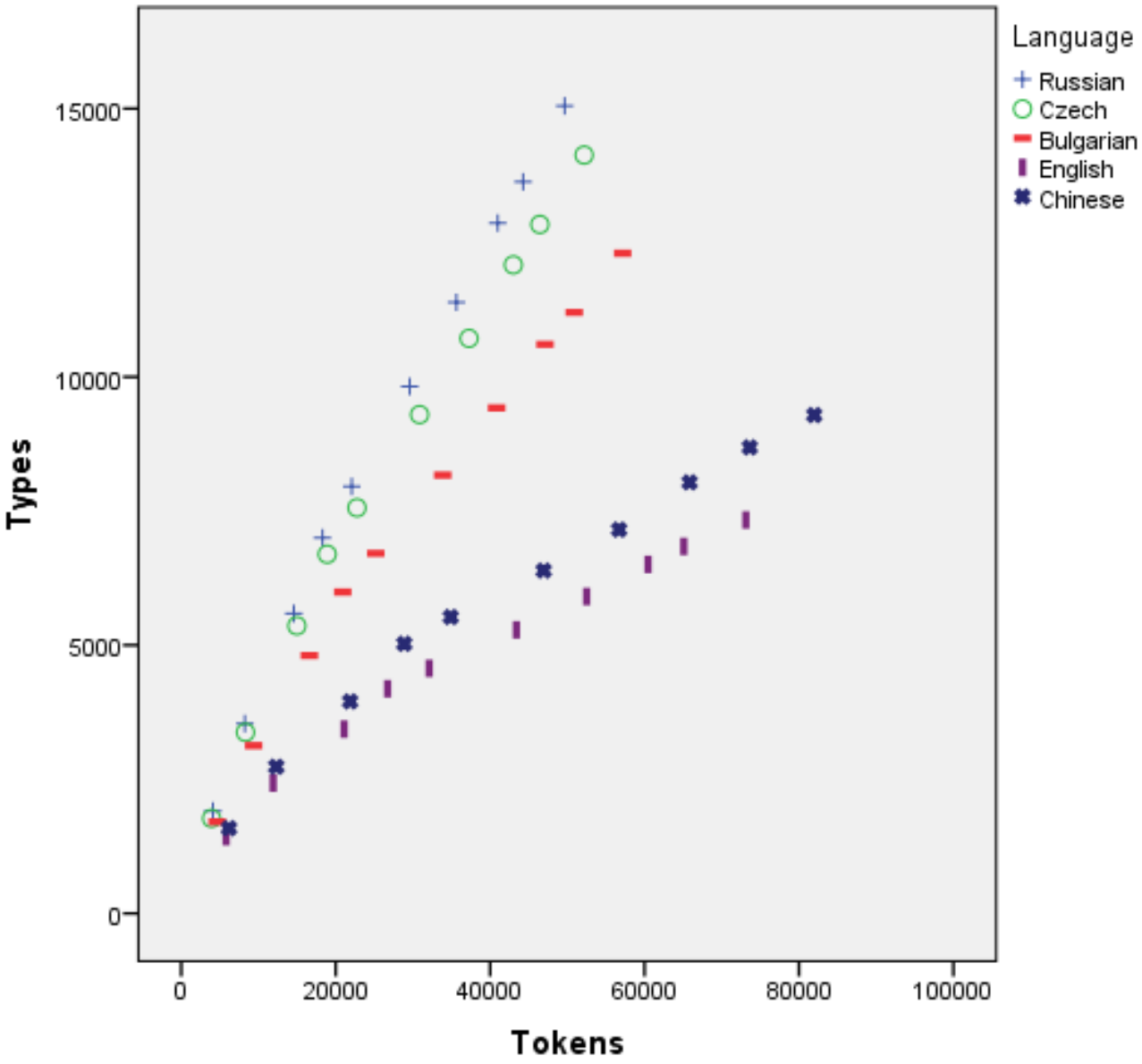
Language	Parameter a	Parameter b	R^2
Russian	1	0,89	0,99
Czech	1	0,88	0,98
Bulgarian	1	0,86	0,98
English	1	0,80	0,94
Chinese	1	0,81	0,96

Language	TTR
Russian	3,30
Czech	3,69
Bulgarian	4,65
Chinese	8,83
English	9,97

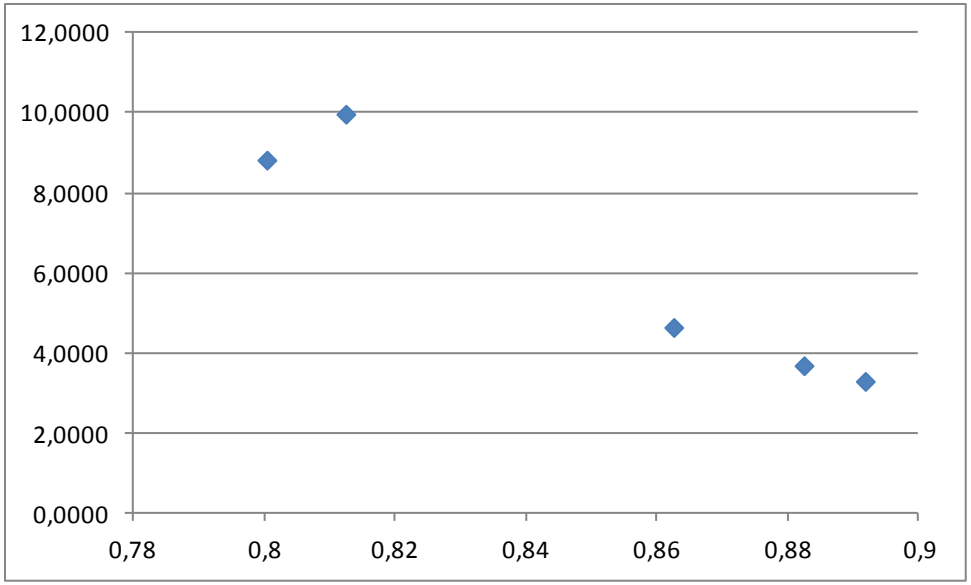


- no significant differences between Slavic languages, and between Chinese and English!
- significant differences between Chinese/English and Slavic Languages

Steepness of the TTR-curve



Interrelation of Parameter *b* and *TTR*



The higher parameter *b*, the lower the TTR.

Intergration and analysis of further characteristics

Word length
Number of morphemes

Parts of speech
Distribution

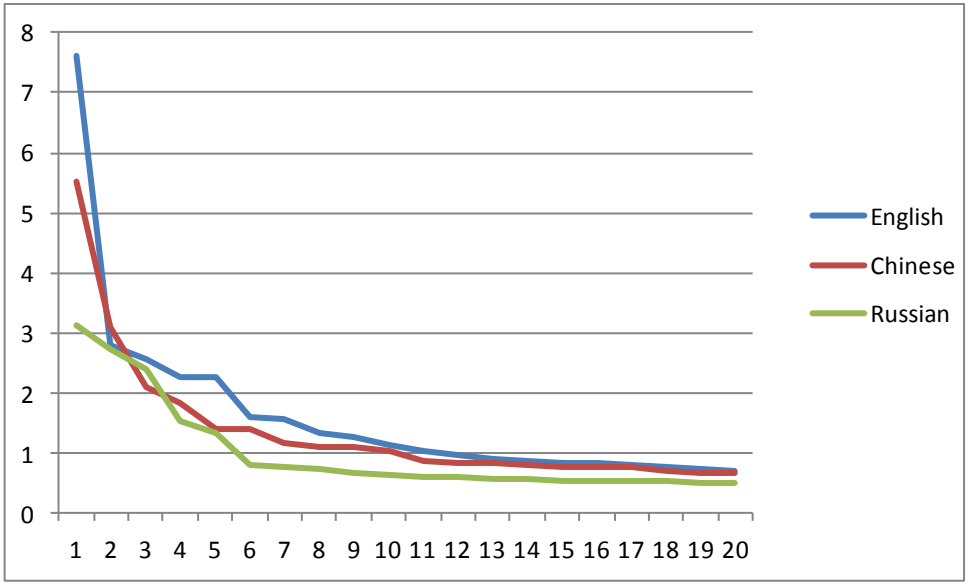
Parameters of
Zipf's law

frequency of
auto- and synsemantics

h-point of
rank frequency distribution

hapax legomena

Different behaviour of rank frequency distribution



most frequent word forms (%)

	English	Chinese	Russian
1	7,60	5,51	3,13
2	2,78	3,10	2,72
3	2,55	2,10	2,41
4	2,26	1,83	1,54
5	2,26	1,41	1,33

→ high percentage of auxilaries

→ different steepness of the curves

→ different location of the h-point

Summary:

- Quantitative methods can be applied successfully in language typology
- Importance of (parallel) text analysis
- analysis of word form types and word form tokens goes beyond the text level
- Type-Token Ration gives some information about the morphological structure of languages
- Interrelation to other text characteristics
- More systematic studies needed

Perspectives and boundaries of quantitative text and language analysis

- selection of linguistic properties and entities
- problems and challenges of measuring linguistic entitites and properties
- relevance of the deductive postulation of hypotheses
- „translation from the „language of linguistics“ into „language of statistics“
- statistical analysis of linguistic data is specific with many pitfals
- central problems are: homogeneity/heterogeneity of linguistic data, extreme skewness of linguistic data,
- interdisciplinary cooperation is required

BUT: every empirical/quantitative study = opens news problems = new title in a mosaic

Selected literature:

Altmann, Gabriel (1980): Prolegomena to Menzerath's law. In: Rüdiger Grotjahn (ed.): *Glottometrika 2*. Bochum: Brockmeyer (Quantitative Linguistics, 3), S. 1–10.

Bußmann, Hadumod (Hg.) (2008): Lexikon der Sprachwissenschaft. Vierte, durchgesehene und bibliographisch ergänzte Auflage. Stuttgart: Kröner.

Köhler, Reinhard (2005): Gegenstand und Arbeitsweise der Quantitativen Linguistik. In: R. Köhler, G. Altmann und R.G. Piotrowski (Hg.): *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook.* Berlin, New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 27), S. 1–16.

Zipf, George K. (1935): *The Psycho-Biology of Language. An Introduction to Dynamic Philology.* Cambridge: M.I.T. Press.

Zipf, George K. (1949): *Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology.* Cambridge, MA: Addison-Wesley.

Language	h-point
Russian	63
Czech	64
Bulgarian	72
English	96
Chinese	103

$$a = \frac{N}{h^2}$$

Language	h-point	a
Russian	63	12,51
Czech	64	13,96
Bulgarian	72	10,07
English	96	7,93
Chinese	103	7,73