Emmerich Kelih

(Graz)

Are parallel text corpora a reliable source for the study of analytism/synthetism? – Some results from Slavic languages

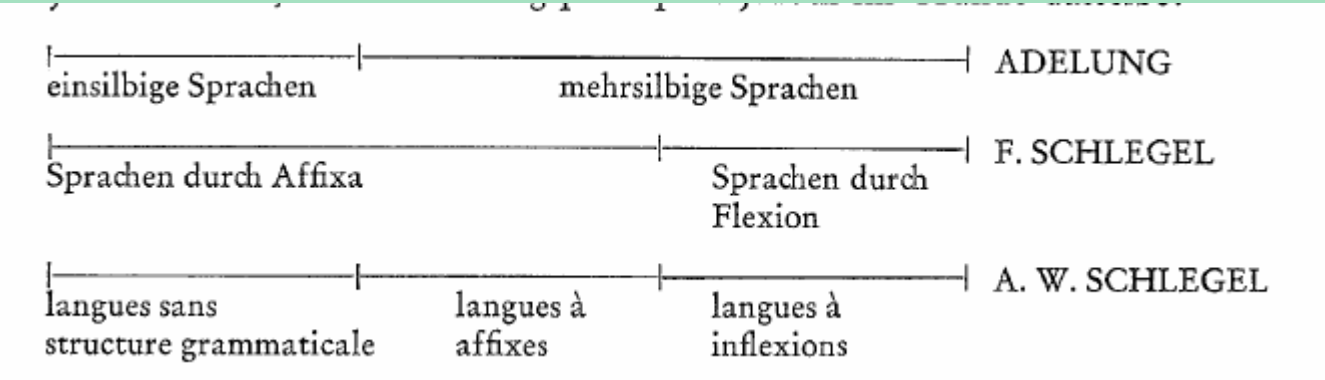
- Institut für Slawistik, Universität Graz
- http://www-gewi.uni-graz.at/quanta/ [Graz-project on Quantitative Text-Analysis]
- http://www.uni-graz.at/emmerich.kelih/ [emmerich.kelih@uni-graz.at]

Goals and purpose of the study

Morphological classification of Slavic languages

1. Analytism (A) and Synthetism (S): linguistic meaning of this terms
2. „State of the art“ in Slavic and quantitative linguistics
3. Indicators of (A) and (S)
   - 3.1. Word lenght (WL)
   - 3.2. Zipf‘s law: linguistic meaning of parameters *a* and *c*
4. Used Slavic parallel corpora for the study of (A) and (S)
5. First empirical results for WL and Zipf‘s law
6. Summary and conclusion

Morphological classification (19th century linguistics)



further development:

- isolating, agglutinative, polysynthetic, inflectional (fusional) languages
- morphologically „formed“ and „unformed“ languages
- analytic and synthetic types

Specifying the meaning of (A) and (S)

- expression of **"grammatical" information** via **separate, autonomous grammatical words**

- Transfer of grammatical information from the morphological to the syntactical level

- Preference for „multiple morpheme constructions" instead of „simple" inflectional constructions

Potential characteristics of (A) and (S)

(1) Developments and properties of the case system (diachronic perspective: loss of cases, increase/decrease of syncretism)

(2) Loss of agreement

(3) Processes of grammaticalisation

(4) Changes in preposition system

(5) Expression of comparison (synthetic vs. analytic forms)

(6) Morphological structure of the tense system (periphrastic forms of the future tense etc.)

(7) Properties of word formation (cf. Hinrichs 2000a: 92)

Some examples from different Slavic languages:

Comparision in Slovene:

I. Synthetic type:
Positive – comparative – superlative forms
lep – lep**ši** – **naj**lep**ši** (nice)
težak – tež**ji** – **naj**tež**ji** (heavy)
bogat – bogate**jši** – **naj**bogate**jši** (wealthy)

II. Analytic type:

(for polysyllables adjectives, ending with -en, -av, -ast, -at and colour terms)

zaželen – **bolj** zaželen – **najbolj** zaželen (requested, desired)
duševen – **bolj** duševen – **najbolj** duševen (mental)
muhast – **bolj** muhast – **najbolj** muhast (bothersome)

→ two different coding techniques are used !

Morphological form of the future tense in Serbian

1. Analytic type:

Enclitic form of verb „хтети“ will + Infinitive

ћу	**ћемо**	
ћеш	**ћете**	+ радити (to work), питати (to ask)
ће	**ће**	

2. Infinitive – ти + enclitic form of хтети

виде**ћемо**, пита**ћу**, ради**ћете**

[но: доћи ћу]

3. Analytic type: enclitic form of хтети + da-construction + present tense

то ћемо да видимо

и он ће да дође

студенти неће да се врате до јесени ...

→ use of different techniques !

Russian as an anti-analytical language?

• high frequency of preposition constructions

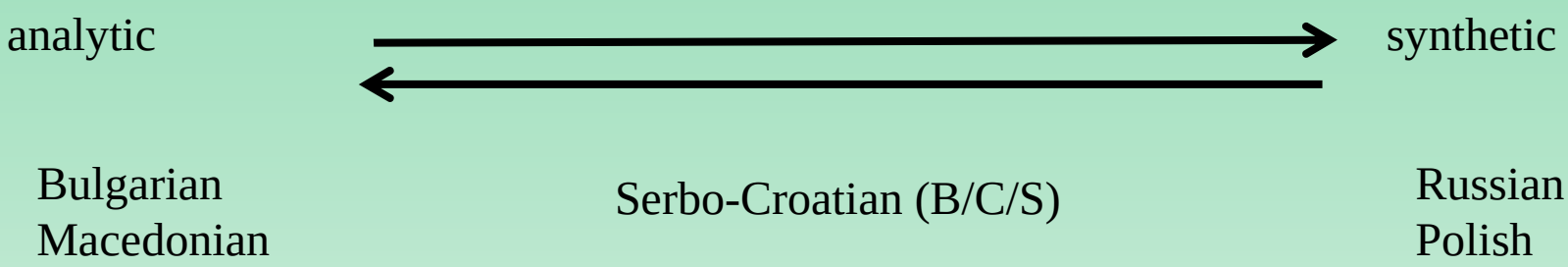
операция по пересадке сердца
план по выпуску vs. план выпуска
инженер по акустике vs. *akustični инженер
учебник по математике vs. учебник математики

„… eine schon fast parasitäre Polysemierung dieser Universalpräposition … bei der der potentielle Zuwachs an Analytismus einen gering zu veranschlagenden Nebeneffekt darstellt“ (Weiss 2004: 269)

• high frequency of abbreviations

• tendency to use compounds

What about (A) and (S) in Slavic languages?



It is a common place of Slavic linguistic that Bulgarian and Macedonian are strongly analytic languages; on the other hand Russian and Polish are strongly synthetic ones. This point of view is shared by many linguists.

Between these two groups of languages Serbo-Croatian as a typologically mixed language is located; Serbo-Croatian is synthetic in regard to the noun morphology and analytical in regard to the verb morphology.
(Hinrichs 2000: 91)

Open problem?

Which parameters and/or characteristics are the basis for this claim?

Constructive approach to (A) (S):

→(A) and (S) as gradual and relative concepts

→ (A) and (S) should be analysed by the means of quantitative methods

→ used language material should guarantee the comparability of the results (contrastive analysis)

→ Empirical material: Parallel text corpora of Slavic languages
1. «Как закалялась сталь» „How the steel was tempered" (KZS) 12 Slavic languages
2. «Мастер и Маргарита» „The Master and Margarita" (MiM) in 11 Slavic languages

→ **Word length as indicator of (A) (S) (**Greenberg 1960)

→ **Study of word form frequencies**

→ **rank frequency distributions**

→ **Parameters from Zipf's law**

→ **(Analysis of hapax legomena)**

Characteristics of the used corpora

- **How the steel was tempered** (KZS): written in 1932 by N.Ostrovskij, socialist realism novel 10 chapters in 12 languages (Belorussian, Russian, Ukrainian, Czech, Polish, Slovak, Upper-Sorbian, Bulgarian, Macedonian, Croatian (ijekavica), Serbian (ekavica), Slovenian)
- **Master i Margarita** (MiM), written by M. Bulgakov; full text in 11 Slavic languages (no translation for Sorbian)

→ „Input“: orthographic word forms

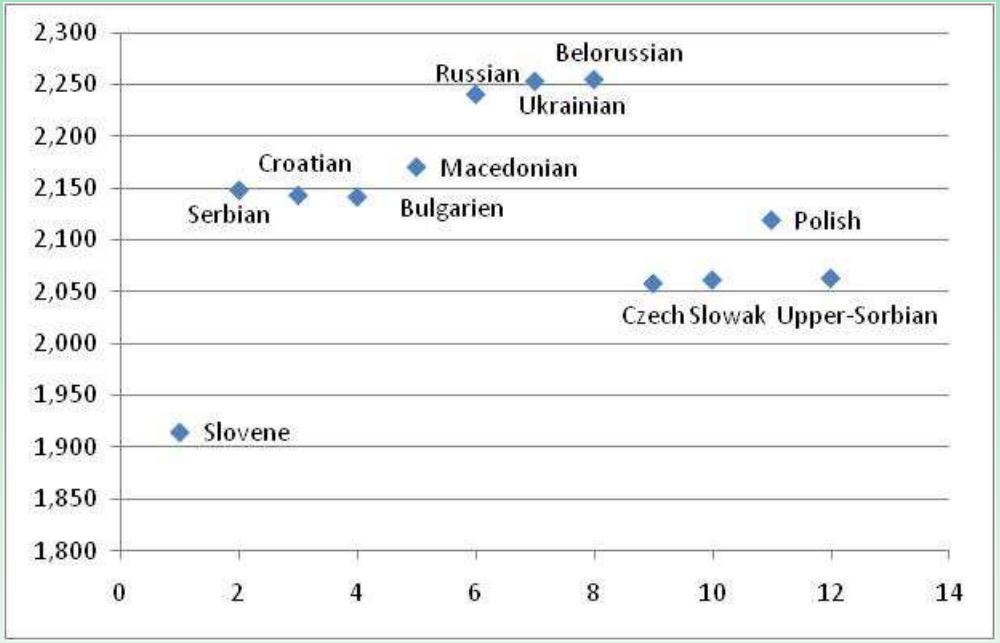
→ Word length (WL): measured by the number of syllables

→ Length of tokens = syntagmatic level

→ Automatic analysis is performed

Token-length “KZS”

Nr.	language	Tokens	WL/Syl.
1	Slovene	62655	1,914
2	Serbian	56230	2,148
3	Croatian	56424	2,143
4	Bulgarian	57174	2,142
5	Macedonian	58837	2,171
6	Russian	49675	2,241
7	Ukrainian	49612	2,254
8	Belorussian	50010	2,255
9	Czech	52180	2,058
10	Slovak	52099	2,062
11	Polish	52737	2,119
12	Sorbian	58484	2,063

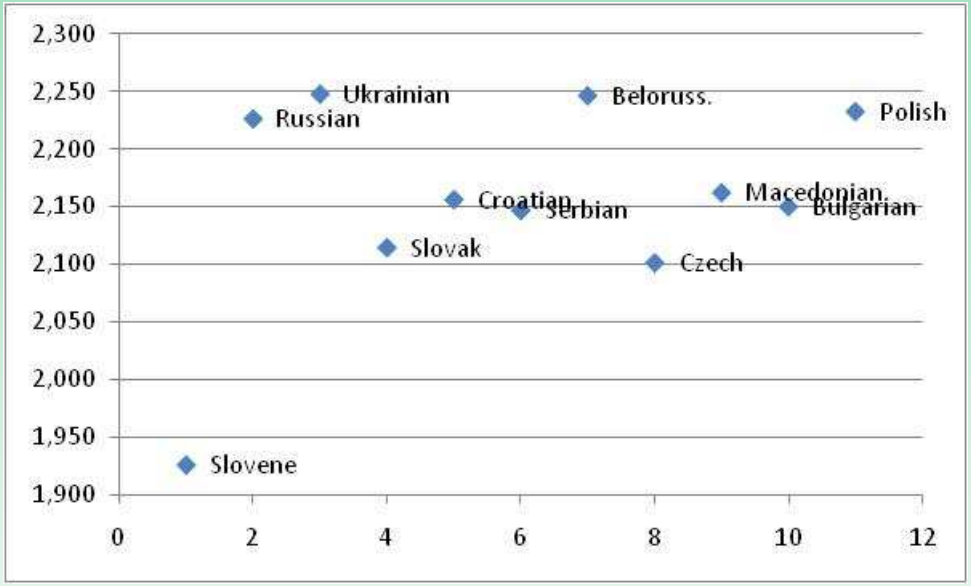


→The longer the tokens, the more synthetic the language !

→ typological row of (A) and (S) of Slavic languages !

Analysis of KZS + MiM : WOL-Tokens

Language	Tokens	WOL/Syl
Slovene	194728	1,926
Russian	163426	2,227
Ukrainian	162345	2,248
Slovak	163789	2,115
Croatian	179964	2,156
Serbian	183648	2,147
Beloruss.	158815	2,247
Czech	160812	2,102
Macedonian	190852	2,162
Bulgarian	181554	2,150
Polish	172161	2,233



analytic

synthetic

Slo.	Cz.	Sk.	Serb.	Bulg.	Cro.	Mac.	Rus.	Pol.	Belor.	Ukr.
1,93	2,10	2,11	2,15	2,15	2,16	2,16	2,23	2,23	2,25	2,25

(KZS + MiM)

Slo.	CZ.	SK	Sorb.	Pl.	Bulg.	Cro.	Serb.	Mac.	Rus.	Ukr.	Belor.
1,91	2,06	2,06	2,06	2,12	2,14	2,14	2,15	2,17	2,24	2,25	2,26

(KZS)

Bulgarian
Macedonian

Serbo-Croatian (B/C/S)

Russian
Polish

(A) and (S) in Slavic languages: Word length

Slo.	Cz.	Sk.	Serb.	Bulg.	Cro.	Mac.	Rus.	Pol.	Belor.	Ukr.
1,93	2,10	2,11	2,15	2,15	2,16	2,16	2,23	2,23	2,25	2,25

on syntagmatic level

Preliminary results:

Macedonian/Bulgarian hardly can be considered as strongly analytic languages

Slovene, Czech, and Slovak are analytic languages

Eastern Slavic languages are in fact synthetic languages!

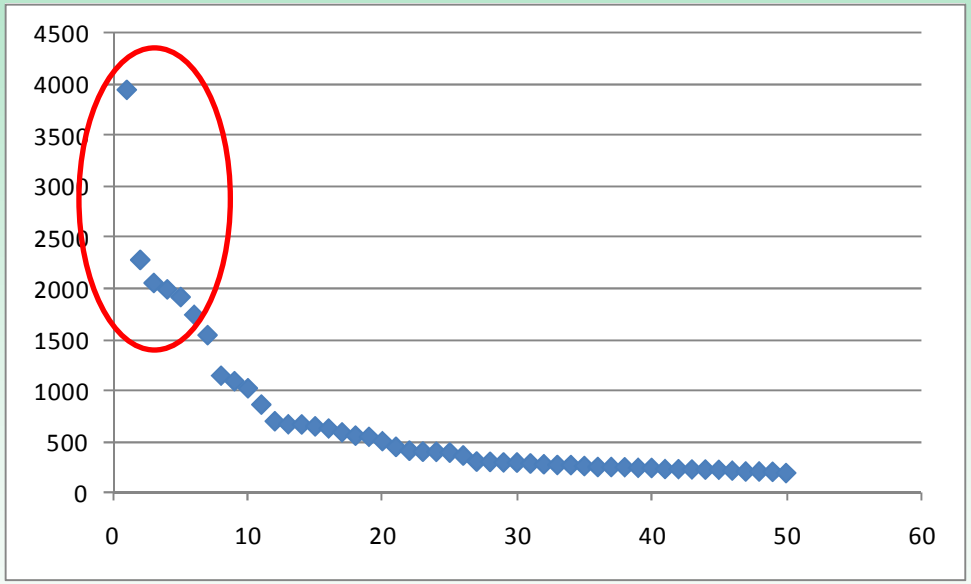
II. Study of word form frequencies

→ „Input“: orthographic word forms

→ frequency “dictionary” of types

→ analysis of the rank frequency distribution

N	Word	Freq.	%
1	І	1532	3,0717
2	НА	1136	2,2777
3	НЕ	934	1,8727
4	З	885	1,7745
5	Ў	775	1,5539
6	У	773	1,5499
7	ШТО	559	1,1208
8	ЁН	504	1,0105
9	ДА	467	0,9364
10	ЯГО	441	0,8842
11	А	392	0,786
12	ЗА	361	0,7238
13	Я	347	0,6958
14	ЯК	292	0,5855
15	АЛЕ	246	0,4932



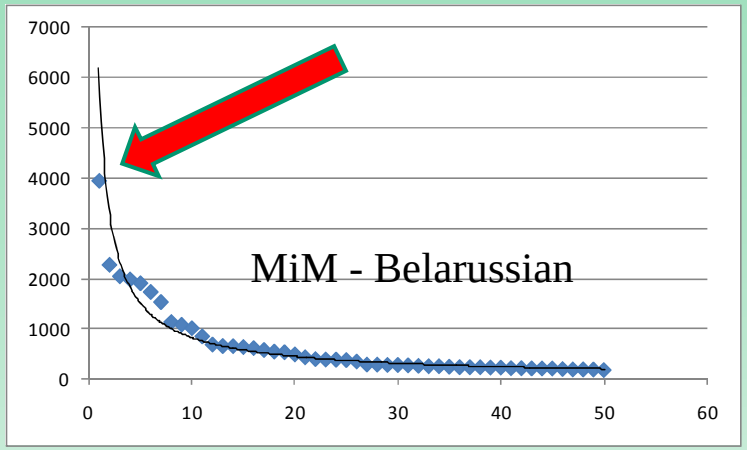
Zipf‘s law

$$f(r) = c / r^{a}$$

→ Linguistic meaning of most frequent word forms (p1) ?

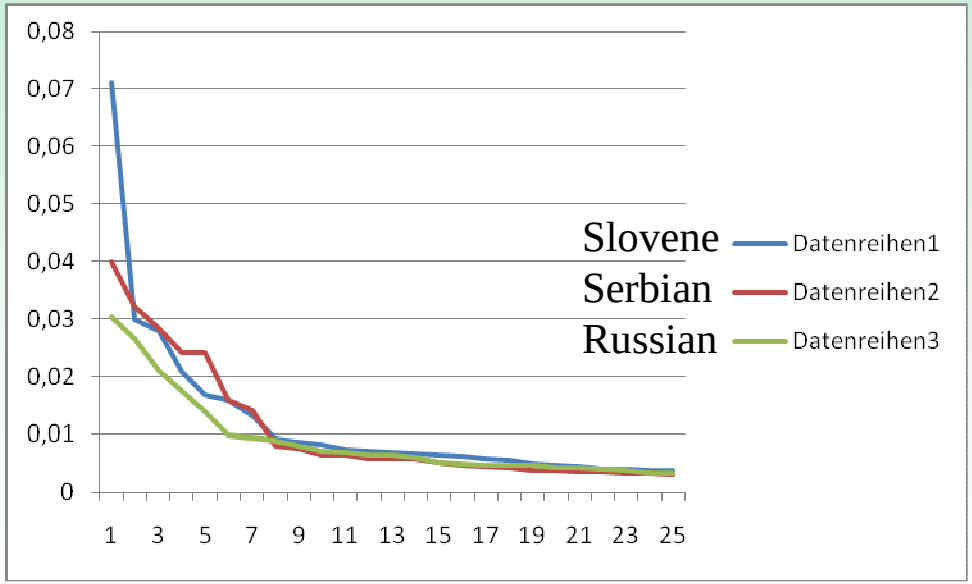
→ Interpretation of parameter *c* of Zipf’s law ?

Parameter ***c*** from Zipf‘s law



$$f(r) = c / r^{a}$$

Parameter c = „starting point“ of the curve
Parameter c = steepness

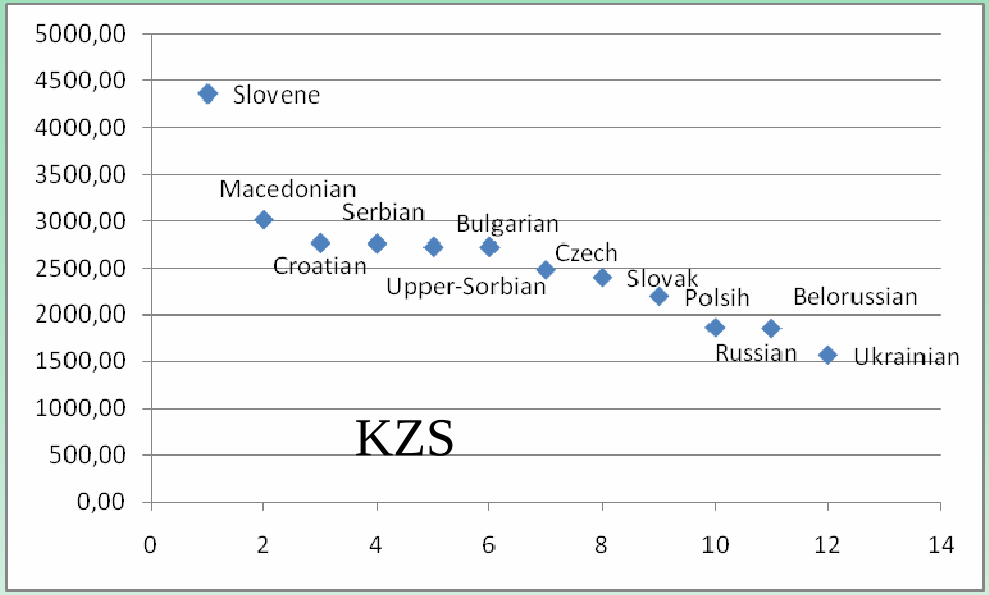


Parameter	**c**
Slovene	4358,26
Serbian	2756,89
Russian	1863,85

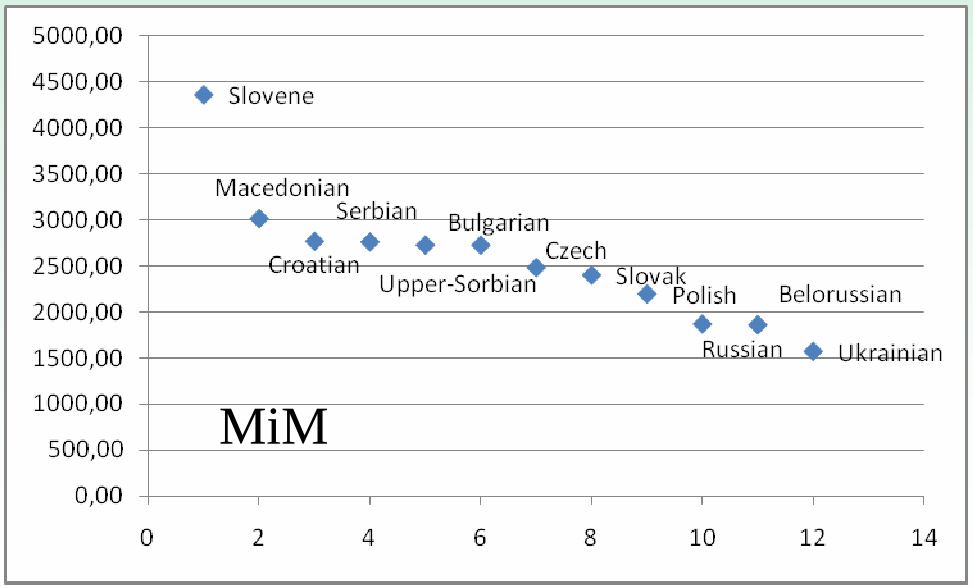
Different exploitation of high frequency words, which are in general synsemantics and “carriers” of grammatical information.

Parameter **c: Results**

	Language	c
1	Slovene	4358,26
2	Serbian	2756,89
3	Croatian	2764,28
4	Bulgarian	2722,46
5	Macedonian	3010,62
6	Russian	1863,85
7	Ukrainian	1565,28
8	Belorussian	1854,25
9	Czech	2477,64
10	Slovak	2395,46
11	Polsih	2191,49
12	Upper-Sorbian	2722,77



Language	c
Slo.	9888,96
Croatian	8185,78
Macedonian	7329,65
Serbian 2	6518,92
Bulgarian	6493,81
Serbian	6471,67
Czech	5576,01
Slovak	5532,36
Russian	5455,56
Polish	4985,22
Belorussian	4719,07
Ukrainian	4444,07



Indicator (A) and (S) in Slavic languages: Word length

analytic ⟵ ⟶ synthetic

Slo.	Cz.	Sk.	Serb.	Bulg.	Cro.	Mac.	Rus.	Pol.	Belor.	Ukr.
1,93	2,10	2,11	2,15	2,15	2,16	2,16	2,23	2,23	2,25	2,25

on syntagmatic level

(A) and (S) in Slavic languages: Parameter *c*

Slo.	Cro.	Mac.	Bulg.	Serb. 2	Serb.	Rus.	Sk.	Pol.	Ukr.	Cz.	Belor.
7123,61	5471,34	5046,96	4752,22	4620,69	4167,76	3966,60	3693,31	3690,34	3583,42	3570,65	3455,28

Similarities and differences
in dependency of used parameters slightly different typological “row”
Main result:
Slovene seems to be analytic in both respects
Macedonian and Bulgarian are analytic regarding parameter *c*
Belorussian and other Eastern Slavic languages are rather synthetic languages
Slovak and Czech show quite a flexible behaviour

Results

1. parallel texts are a reliable source for the analysis of (A) (S)
2. word length can be used as indicator of (A) (S)
3. Parameters of Zipf‘s law also gives some information about (A) (S)

Perspectives:

1. Analyses of further parallel texts are required (especially of non-fictional texts, non-Russian translations)
2. use of alternative indicators, like

$$A=\frac{c}{(V-HL/2)^a}$$ Popescu/Altmann (2009)

3. *c*, *a* … Parameters of Zipf‘s law

HL = «hapax legomena»

V = Vocabulary size

… enough work for the next years ☺ !

References (selected)

Bossong, Georg (2001): Die Anfänge typologischen Denkens im europäischen Rationalismus. In: Haspelmath, Martin; König, Ekkehard Oesterreicher Wulf; Raible, Wolfgang (Hg.): Language Typology and Language Universals/ Sprachtypologie und sprachliche Universalien. Berlin u.a.: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 20,1), S. 249–264.

Coseriu, Eugenio (1972): Über die Sprachtypologie Wilhelm von Humboldsts. Ein Beitrag zur Kritik der sprachwissenschaftlichen Überlieferung. In: Hösle, Johannes (Hg.): Beiträge zur vergleichenden Literaturgeschichte. Festschrift für Kurt Wais zum 65. Geburtstag. Tübingen: Niemeyer, S. 107–135.

Hinrichs, Uwe (2000a): Prolegomena zu einer Theorie des Analytismus I. Anhand der Sprachen in Ost- und Südosteuropa. In: Hinrichs, Uwe; Büttner, Uwe (Hg.), S. 83–105.

Hinrichs, Uwe (2000b): Prolegomena zu einer Theorie des Analytismus II anhand der Sprachen in Ost- und Südosteuropa. In: Hinrichs, Uwe; Büttner, Uwe (Hg.), S. 107–128.

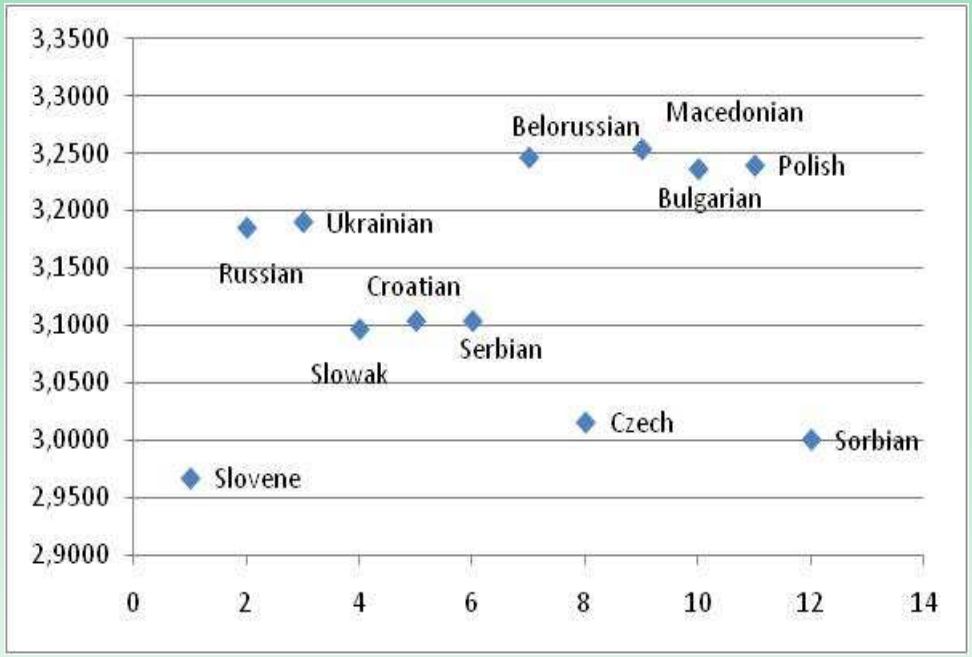
Hinrichs, Uwe; Büttner, Uwe (Hg.): Die Südosteuropa-Wissenschaften im neuen Jahrhundert: Akten der Tagung vom 16. - 19.10.1999 an der Universität Leipzig. Wiesbaden: Harrassowitz, S. 107–128.

Hinrichs, Uwe (2004): Zum Analytismus im Serbischen. In: Hinrichs, Uwe (Hg.): Die europäischen Sprachen auf dem Wege zum analytischen Sprachtyp. Wiesbaden: Harrassowitz (Eurolinguistische Arbeiten, 1), S. 293–302.

Greenberg, Joseph H. (1960): A quantitative approach to the morphological typology of language. In: International Journal of American Linguistics, H. 26, S. 178–194.

“KZS”’: Types-length

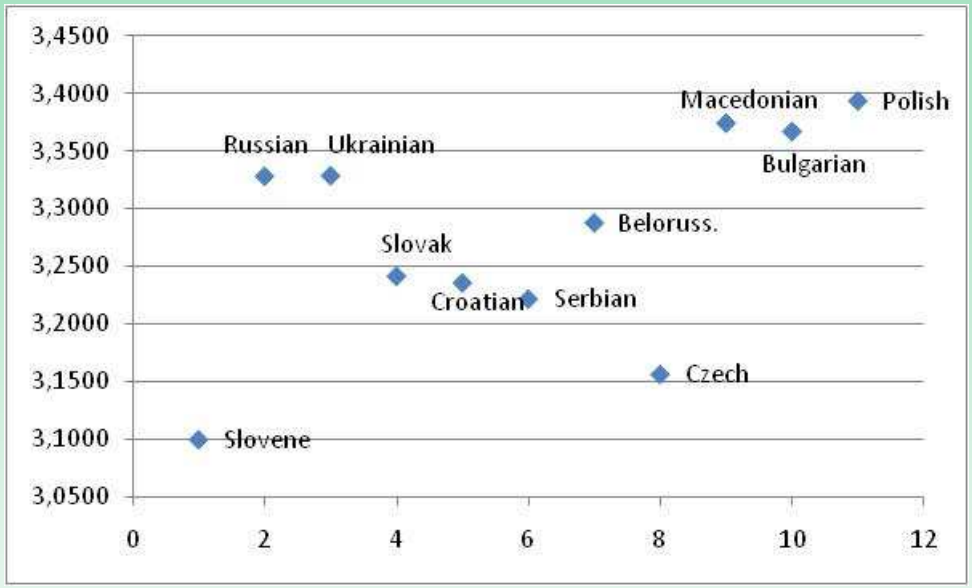
Nr.	language	Types	Wol/Syl
1	Slovene	13946	2,9667
2	Serbian	13642	3,1037
3	Croatian	13737	3,1038
4	Bulgarian	12308	3,2358
5	Macedonian	11465	3,2531
6	Russian	15053	3,1849
7	Ukrainian	14645	3,1899
8	Belorussian	14858	3,2459
9	Czech	14136	3,0155
10	Slovak	14027	3,0967
11	Polish	14978	3,2390
12	Sorbian	14574	3,0006



→ Bulgarian/Macedonian as synthetic languages !

Further material KZS + MiM: WOL-Types

Slovene	28419	3,0999
Russian	32874	3,3284
Ukrainian	34256	3,3289
Slovak	30885	3,2417
Croatian	29859	3,2359
Serbian	29851	3,2219
Beloruss.	30819	3,2880
Czech	31714	3,1565
Macedonian	24013	3,3745
Bulgarian	25761	3,3672
Polish	32711	3,3938



analytic synthetic

Slo.	Cz.	Serb.	Cro.	Sk.	Belor.	Rus.	Ukr.	Bulg.	Mac.	Pol.
3,10	3,16	3,22	3,24	3,24	3,29	3,33	3,33	3,37	3,37	3,39

Slo.	Sorb.	Cz.	Sk.	Serb.	Cro.	Russ.	Ukr.	Bulg.	Pol.	Belor.	Mac.
2,97	3,00	3,02	3,10	3,10	3,10	3,18	3,19	3,24	3,24	3,25	3,25

KZS

Word length = number of syllables/number of types

Word length = number of syllables / number of tokens

→ text length is playing an important role

General linguistic problems of analysing 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 process of alignment)

Орфографическая словоформа/словоупотребление

Автоматический анализ Type/Token

Teplé jarné slniečko práve zapadalo, keď sa pri Patriarchových rybníkoch zjavili dvaja občania. Prvý **z** nich – štyridsiatnik **v** sivom letnom obleku – bol nízky, zavalitý brunet **s** plešinkou, stľapčený klobúčik držal **v** ruke **a** hladko vyholenú tvár mu zdobili ozrutné okuliare **s** čiernym kosteným rámom. Druhý – plecnatý, ryšavkastý **a** strapatý mládenec **s** pepitovou čiapkou, zacapenou von **z** čela – mal na sebe károvanú rozhalenku, dokrkvané biele nohavice **a** čierne mokasíny.

Цепочка знаков между пробелами = 1 Token

→68 Tokens (словоупотреблений)

→ 62 Types (словоформ)

N	Word	Freq.	%	Lemmas
1	A	3	4,41	
2	S	3	4,41	
3	V	2	2,94	
4	Z	2	2,94	
5	BIELE	1	1,47	
6	BOL	1	1,47	
7	BRUNET	1	1,47	
8	ČELA	1	1,47	
9	ČIAPKOU	1	1,47	
10	ČIERNE	1	1,47	
11	ČIERNYM	1	1,47	
12	DOKRKVANÉ	1	1,47	
13	DRUHÝ	1	1,47	
14	DRŽAL	1	1,47	
15	DVAJA	1	1,47	

Резюме (предварительные результаты)

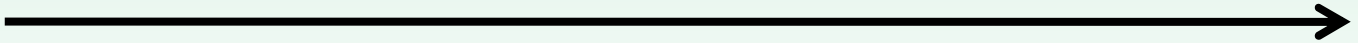
Индикатор A/S в славянских языках:

На синтагматическом уровне

аналитический → синтетический

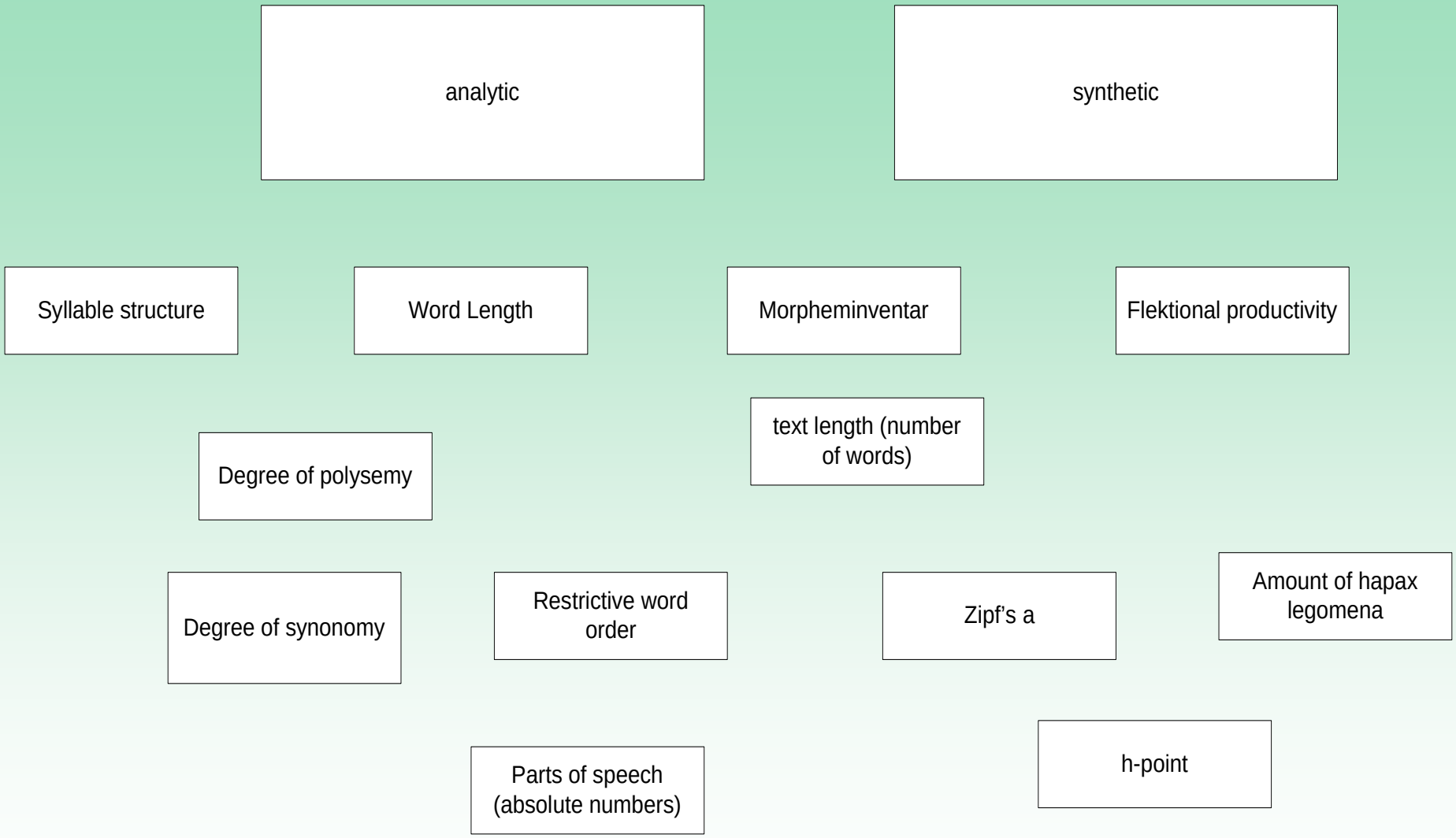
Slo.	**CZ.**	**SK**	**Sorb.**	**Pl.**	**Bulg.**	**Cro.**	**Serb.**	**Mac.**	**Rus.**	**Ukr.**	**Belor.**
1,91	2,06	2,06	2,06	2,12	2,14	2,14	2,15	2,17	2,24	2,25	2,26

На парадигматическом уровне



Slo.	**Sorb.**	**Cz.**	**Sk.**	**Serb.**	**Cro.**	**Russ.**	**Ukr.**	**Bulg.**	**Pol.**	**Belor.**	**Mac.**
2,97	3,00	3,02	3,10	3,10	3,10	3,18	3,19	3,24	3,24	3,25	3,25

Theoretical perspective: „Synergetic“ control cycle/ interrelations



Linguistic problems

1.Definition word, word form etc. & problems of word delimitation
2. meaning of “separate” and “autonomous”

- there is no universal valid definition of the concept word and wordform, therefore a morphological classification in terms of (A) and (S) is impossible (Schwegler 1990: 46)

- universal principles of word delimitation are only a theoretical construct of language typology (Haarman 2004: 73)

- „Much that has been written about the word is decidedly eurocentric“. (Dixon/Aikhenvald 2002: 2)