Emmerich Kelih

(Graz)

How the Steel was tempered

– A New Slavic Parallel Corpus

- 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]

1. Introduction
2. Purposes and goals of parallel corpora research
3. Slavic Parallel Corpora: Overview
4. Description: Kak zakaljalas‘ stal‘ – How the steel was tempered (KZS)
5. Quantitative features and characteristics of KZS

“Application” fields of Parallel Corpora

1. Typological and cross-linguistic studies
2. Linguistic structure of translations
3. Hypotheses from quantitative and synergetic linguistics
4. Inter-lingual readability

“Available” Slavic Parallel Corpora (selected)

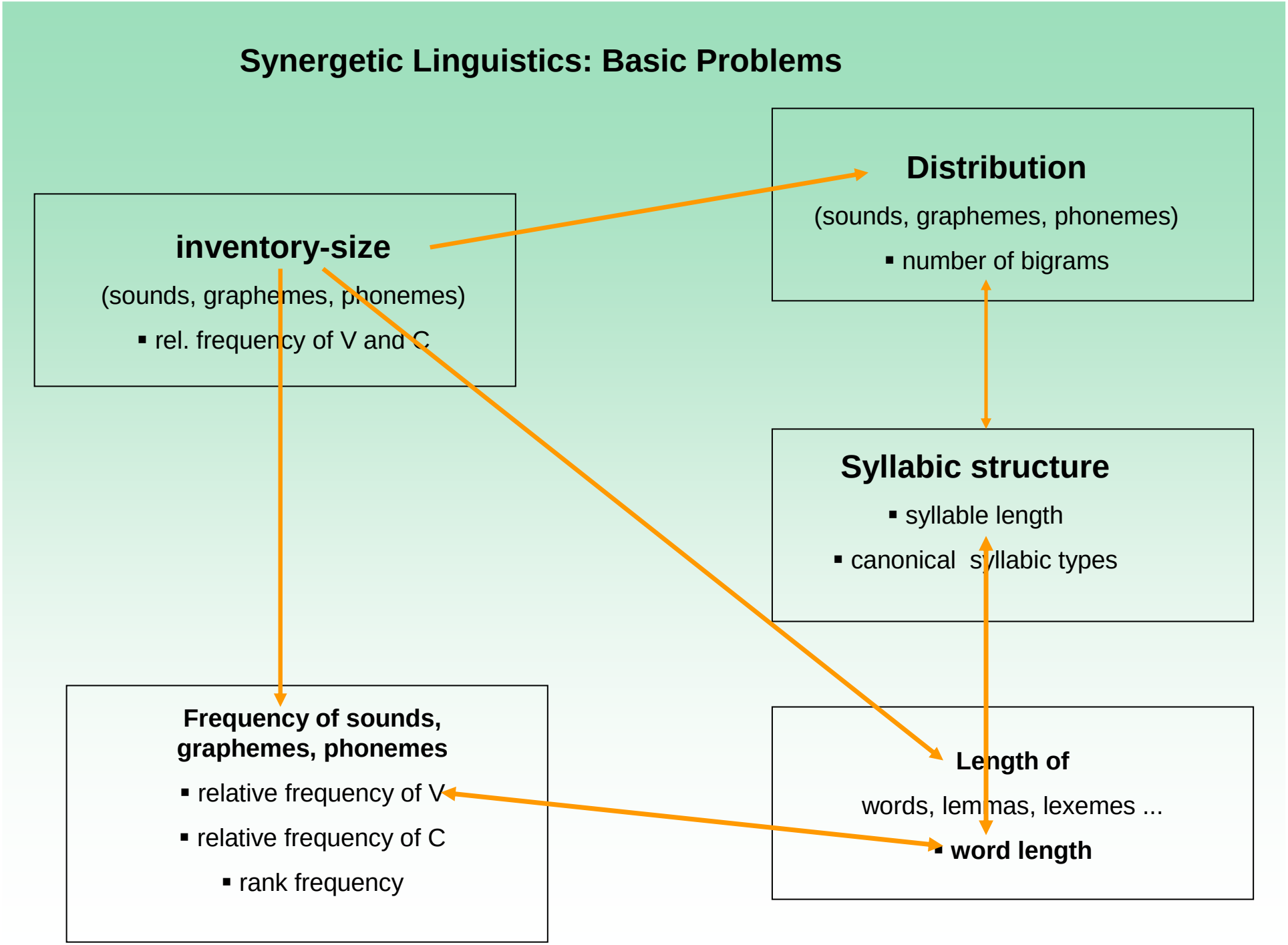
- **Multext East**: George Orwells „1984“ with translations into Bulgarian, Croatian, Czech, Resian, Russian, Serbian and Slovene

- ***ParaSol*: A Parallel Corpus of Slavic and other languages**
  = formerly “Regensburg Parallel Corpus of Slavic Languages”
  includes different text form Slavic languages
  „Master i Margarita“ (M. Bulgakov) in 7 Slavic languages
  „Solaris“ (St. Lem) in 6 Slavic languages

- **Smaller projects** (restrictions to 2, 3 ... languages, special word lists etc.

→ Parallel corpora for all (almost all) Standard Slavic Languages?

Synergetic Linguistics: Basic Problems



Empirical check of synergetic hypotheses

- Need for data from many Slavic Languages !

- Parallel corpora as reliable material for cross-linguistic analyses: a-priori low inter-lingual heterogeneity !

- Additionally analysis of (native)non-translated texts is required !

→Building up the parallel corpus **„Kak zakaljalas‘ stal’** – How the steel was tempered (KZS) in 12 Slavic languages !

→ Filling up the „Master i Margarita“ translations: Slovene, Serbian, Croatian, Slovak and Ukrainian !

Basic characteristics of KZS

- Text: “Kak zakaljalas’ stal’/How the steel was tempered”
- Socialist realism novel from 1932-1934
- Author: N.A. Ostrovskij
- 10 chapters (scanned, OCR, plain text) for 12 Languages
- Belarusian, Ukrainian, Russian
- Czech, Polish, Slovak, Upper-Sorbian,
- Bulgarian, Macedonian, Croatian (ijekavian), Serbian (ekavian) and Slovenian translations

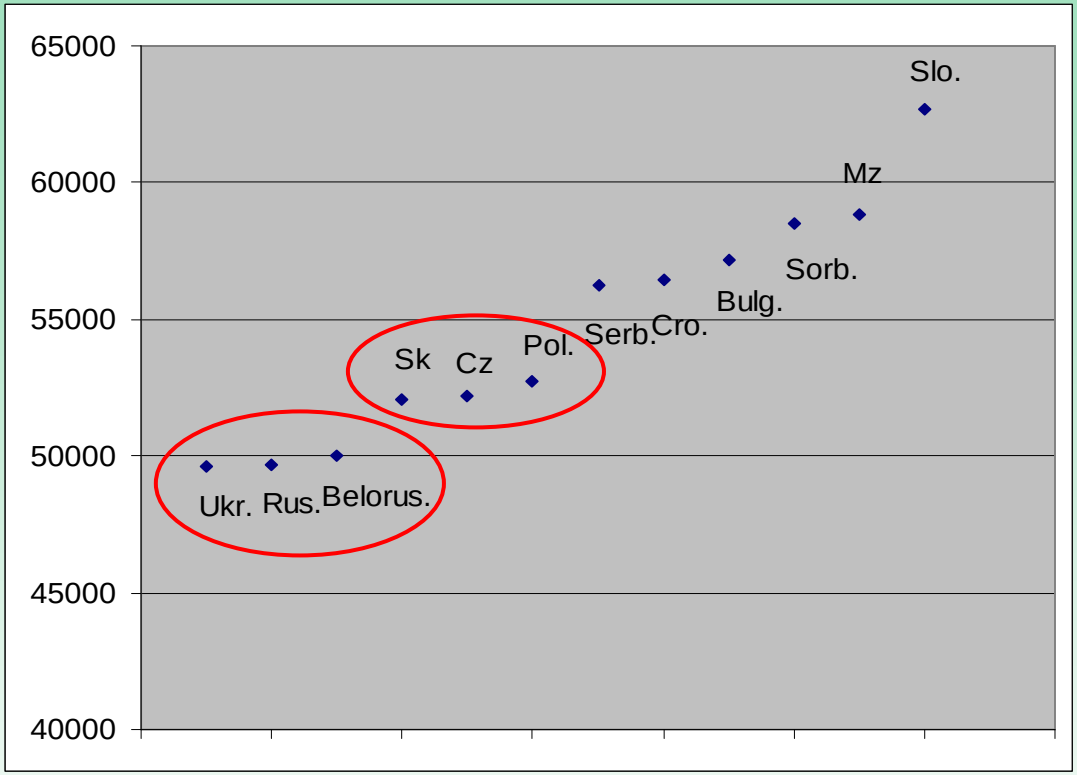
→simple text pre-processing has been done !

→ no tagging and no annotation up to now !

→ first quantitative Analyses !

Analysing the text length: Number of Tokens

Nr.	language	Tokens
1	Slovene	62655
2	Serbian	56230
3	Croatian	56424
4	Bulgarian	57174
5	Macedonian	58837
6	Russian	49675
7	Ukrainian	49612
8	Belorussian	50010
9	Czech	52180
10	Slovak	52099
11	Polish	52737
12	Sorbian	58484



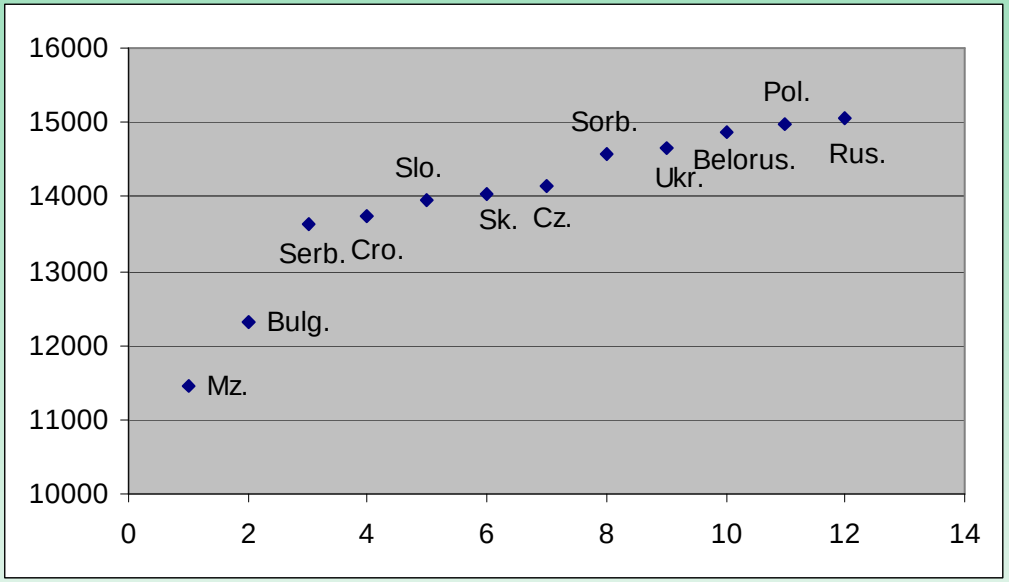
→ Different text length: max. 62655 (Slo.) vs. min. 49612 (Ukr.)

→ Language typological features responsible for this differences !

→ The higher the number of tokens, the more synthetic a language is !

Analysing the text length: Number of Types

Nr.	Sprache	Tokens
1	Slovene	13946
2	Serbian	13642
3	Croatian	13737
4	Bulgarian	12308
5	Macedonian	11465
6	Russian	15053
7	Ukrainian	14645
8	Belorussian	14858
9	Czech	14136
10	Slovak	14027
11	Polish	14978
12	Sorbian	14574

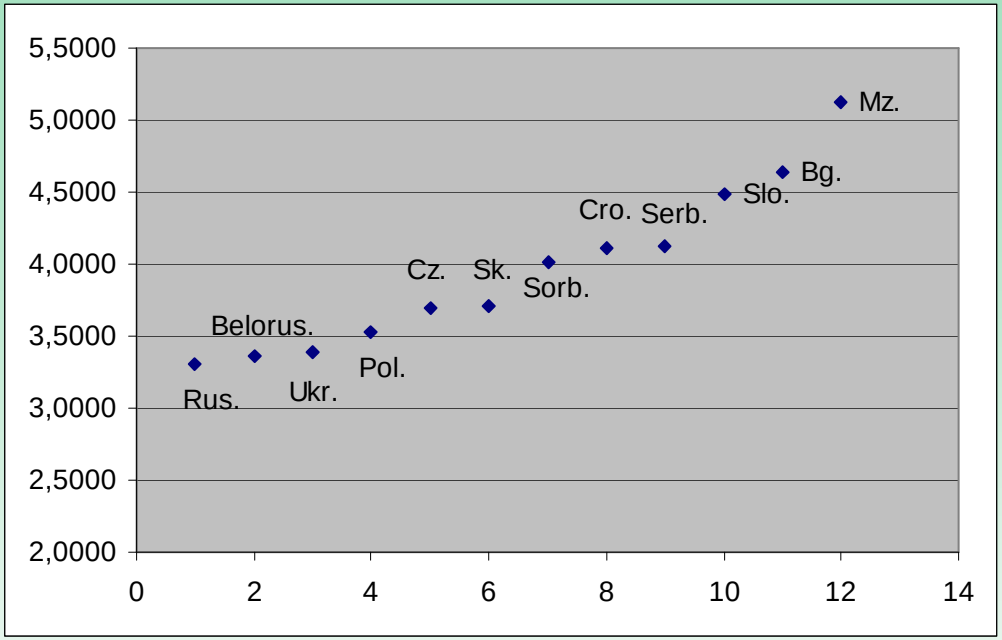


→ Different text length: max. 15053 (Rus.) vs. min. 11465 (Mz.)

→ The higher the number of types, the higher the morphological richness !

Final Step: Analysing the TTR = Types/Tokens

Language	TTR
Rus.	3,3000
Belorus.	3,3659
Ukr.	3,3876
Pol.	3,5210
Cz.	3,6913
Sk.	3,7142
Sorb.	4,0129
Cro.	4,1074
Serb.	4,1218
Slo.	4,4927
Bg.	4,6453
Mz.	5,1319



→ TTR as index of morphological richness/activity and degree of synthetism/analytism

→ typological row of Slavic languages

Summary

- parallel corpora as adequate basis for language typological studies
- differences of text length (types and tokens) not due the kind of translation, but in dependency of the language type
- TTR as indicator of morphological richness and analytism/synthetism

Perspectives

- further empirical evidence is required (parallel corpora and other)
- more systematic quantitative studies !
- need for cooperation, especially with regard to annotation and tagging of the texts !

Hvala za vašo pozornost !

Thank you for your attention !