Morphosyntactic encoding strategies - A quantitative perspective

22.08. 2013
Quantitative Linguistic as integral part of General Linguistics

since 1994

since 2001

since 2008

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:


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 planned]
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**

![Rank frequency distribution graph](image)

![Type-Token-Ratio graph](image)

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 and fusional languages …

\[\downarrow\]

analytic vs. synthetic 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**

\[
\begin{array}{ll}
\text{po-smotr-im} & \text{vidě-ti} \\
\text{PRE.ASP.FUT.-see-1PL} & \\
\end{array}
\]

**Serbian: Analytic**

\[
\begin{array}{ll}
\text{ћe-mo} & \text{vide-ti} \\
\text{će-mo} & \\
\text{AUX.FUT-1PL.PRES} & \text{see-Inf.} \\
\end{array}
\]
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. Northern area: **Synthetic**
- West Slavic languages
- East Slavic languages
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?

What about English? What about Chinese?

- which language level should be analysed? (system vs. Text)
- which characteristics are taken for the classification?
- what about the degree of (A) and (S) within one language?
Example 1: Comparison in Slovene:

I. Synthetic (for frequently used adjectives)

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

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

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

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

- **muhast-∅** bothersome-M.Nom 'bothersome'
- **bolj muhast-∅** COMP bothersome-M.NOM. 'more bothersome'
- **najbolj muhast-∅** SUPERLAT-bothersome-M.NOM. 'most bothersome'
Example 2: Future tense in Serbian

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

<table>
<thead>
<tr>
<th>ми</th>
<th>हे-мо</th>
<th>виђе-ми</th>
</tr>
</thead>
<tbody>
<tr>
<td>ti</td>
<td>Ćе-мо</td>
<td>vide-ти</td>
</tr>
<tr>
<td>we</td>
<td>AUX.FUT-1PL.PRES.</td>
<td>see-Inf.</td>
</tr>
</tbody>
</table>

'we will see'

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

<table>
<thead>
<tr>
<th>виђе-ће-мо</th>
<th>ради-ће-те</th>
</tr>
</thead>
<tbody>
<tr>
<td>vide-ćе-мо</td>
<td>radi-ćе-те</td>
</tr>
<tr>
<td>see.INF-AUX-1PL</td>
<td>work.INF.-AUX-2PL</td>
</tr>
</tbody>
</table>

'we will see'

but: Infinitives with –ći \(\rightarrow\) доћи hy/doći ću (orthographical convention!)

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

<table>
<thead>
<tr>
<th>мо</th>
<th>हे-мо</th>
<th>да</th>
<th>виђимо</th>
</tr>
</thead>
<tbody>
<tr>
<td>to</td>
<td>Ćе-мо</td>
<td>da</td>
<td>vid-имо</td>
</tr>
<tr>
<td>this/that</td>
<td>will-1Pl</td>
<td>da-conjunction</td>
<td>see-1Pl</td>
</tr>
</tbody>
</table>

'that we will see'
Example 2a: Future tensen in Croatian

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

```
\underline{videt} \, \underline{\text{če-mo}}
see.INF \, AUX-1PL
'we will see'

\underline{radit} \, \underline{\text{če-te}}
work.INF \, AUX-2Pl
'you will work'
```

but: -ti at the infinitive, if immediately 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

\[
\begin{align*}
\text{учебник-О} & \quad \text{po} & \quad \text{математика-е} \\
\text{schoolbook-NOM.SG.} & \quad \text{about.PREP.} & \quad \text{mathematics-DAT.SG.} \\
\text{'maths school book'} & & \\
\text{план-О} & \quad \text{po} & \quad \text{выпуск-у} \\
\text{plan-Nom.Sg.} & \quad \text{about.PREP.} & \quad \text{production-DAT.SG.} \\
\text{'production plan'} & & \\
\end{align*}
\]

II. Synthetic

\[
\begin{align*}
\text{учебник-О} & \quad \text{математика-и} \\
\text{schoolbook-NOM.SG.} & \quad \text{mathematics-GEN.SG.} \\
\text{'maths school book'} & & \\
\text{план-О} & \quad \text{выпуск-а} \\
\text{plan-Nom.Sg.} & \quad \text{production-Gen.Sg.} \\
\text{'production plan'} & & \\
\end{align*}
\]
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 required, but the frequency of this types in a text is lower.
First results: Number of Tokens and Types

<table>
<thead>
<tr>
<th>Language</th>
<th>Tokens</th>
<th>Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Russian</td>
<td>49672</td>
<td>15053</td>
</tr>
<tr>
<td>Czech</td>
<td>52180</td>
<td>14136</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>57165</td>
<td>12303</td>
</tr>
<tr>
<td>English</td>
<td>73123</td>
<td>7333</td>
</tr>
<tr>
<td>Chinese</td>
<td>81982</td>
<td>9289</td>
</tr>
</tbody>
</table>
Type-Token-Ratio = TOKENS/TYPES

<table>
<thead>
<tr>
<th>Language</th>
<th>TTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Russian</td>
<td>3.30</td>
</tr>
<tr>
<td>Czech</td>
<td>3.69</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>4.65</td>
</tr>
<tr>
<td>Chinese</td>
<td>8.83</td>
</tr>
<tr>
<td>English</td>
<td>9.97</td>
</tr>
</tbody>
</table>

→ 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:

\[ \text{TYPES} = a \cdot \text{TOKENS}^b \]

\[ y = a \cdot x^b \]

\[ a = 1 \]
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)

<table>
<thead>
<tr>
<th>Language</th>
<th>Parameter a</th>
<th>Parameter b</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Russian</td>
<td>1</td>
<td>0.89</td>
<td>0.99</td>
</tr>
<tr>
<td>Czech</td>
<td>1</td>
<td>0.88</td>
<td>0.98</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>1</td>
<td>0.86</td>
<td>0.98</td>
</tr>
<tr>
<td>English</td>
<td>1</td>
<td>0.80</td>
<td>0.94</td>
</tr>
<tr>
<td>Chinese</td>
<td>1</td>
<td>0.81</td>
<td>0.96</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language</th>
<th>TTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Russian</td>
<td>3.30</td>
</tr>
<tr>
<td>Czech</td>
<td>3.69</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>4.65</td>
</tr>
<tr>
<td>Chinese</td>
<td>8.83</td>
</tr>
<tr>
<td>English</td>
<td>9.97</td>
</tr>
</tbody>
</table>

- no significant differences between Slavic languages, and between Chinese and English!
- significant differences between Chinese/English and Slavic Languages
Steepness of the TTR-curve
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 (%)

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Chinese</th>
<th>Russian</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7,60</td>
<td>5,51</td>
<td>3,13</td>
</tr>
<tr>
<td>2</td>
<td>2,78</td>
<td>3,10</td>
<td>2,72</td>
</tr>
<tr>
<td>3</td>
<td>2,55</td>
<td>2,10</td>
<td>2,41</td>
</tr>
<tr>
<td>4</td>
<td>2,26</td>
<td>1,83</td>
<td>1,54</td>
</tr>
<tr>
<td>5</td>
<td>2,26</td>
<td>1,41</td>
<td>1,33</td>
</tr>
</tbody>
</table>

→ 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 entities 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 pitfalls
- central problems are: homogeneity/heterogeneity of linguistic data, extreme skewness of linguistic data,
- interdisciplinary cooperation is required

BUT: every empirical/quantitative study = opens new problems = new title in a mosaic
**Selected literature:**


<table>
<thead>
<tr>
<th>Language</th>
<th>h-point</th>
<th>$a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Russian</td>
<td>63</td>
<td>12,51</td>
</tr>
<tr>
<td>Czech</td>
<td>64</td>
<td>13,96</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>72</td>
<td>10,07</td>
</tr>
<tr>
<td>English</td>
<td>96</td>
<td>7,93</td>
</tr>
<tr>
<td>Chinese</td>
<td>103</td>
<td>7,73</td>
</tr>
</tbody>
</table>