



Register in Russian: Multi-dimensional analysis vs. functional stylistics

Roland Meyer & Luka Szucsich

Institut für Slawistik, Humboldt-Universität zu Berlin

September 24, 2018

Slavicorp 2018 Panel on Corpus-based Language Variation, Praha

Overview

Functional styles vs. registers

Register-related linguistic features in Russian

Factor analysis on RNC.standart

Multidimensional registers vs. functional styles

Variationism?

Conclusion

Traditional functional styles (1)

- Register is known as *functional style* in the tradition of Russian grammar/linguistics (cf. Kožina 1992, Kožina et al. 2010, Lapteva 2003)
- Coarse classification:
 - conversational style (*разговорный стиль*)
 - scientific style (*научный стиль*)
 - official or business style (*официально-деловой стиль*)
 - journalistic style (*публицистический стиль*)
 - artistic style (*художественный стиль*)

Traditional functional styles (2)

- Functional styles are identified by linguistic features/markers, but also by “stylistic and speech-related characteristics” (encompassing extra-linguistic categories and text type, cf. Kožina et al. 2010: 287)
- Linguistic means (features, markers, lexical choices) basically listed, only weakly correlated to styles and not related to each other.
- Problems: partially vague criteria, markers which are virtually impossible to search or annotate:
-k- suffix: *Lenin-k-a* < *Biblioteka im. Lenina*; but also: *lopat-k-a* 'small shovel, shoulder blade', *perestroj-k-a* 'reconstruction', *laborant-k-a* 'female lab assistant', *nauk-a* 'science')

Register-related markup in the Russian National Corpus

- Russian National Corpus gold standard (= offline disambiguated version): ~ 1.6 Mio orth. toks. hand-annotated for *lemma*, *pos*, *morphology* and *sphere*, *type*, *topic*, *style* (per text)
- «Морфологический стандарт Национального корпуса русского языка», courtesy of the RNC
- *style*:
 - individual / neutral / official / regional / lower / special
 - strongly correlated with *sphere*
- *sphere* : functional styles, slightly more fine-grained:
 - official-business (*официально-деловая*)
 - industrial-technological (*производственно-техническая*)
 - journalistic (*публицистика*)
 - commercials (*реклама*)
 - private oral speech (*устная непубличная речь*)
 - public oral speech (*устная публичная речь*)
 - educational-scientific (*учебно-научная*)
 - artistic (*художественная*)

Text types in RNC.standart

- 42 in an open, opportunistic classification
- partly very coarse: *drama, movie, feuilleton*
- partly very fine-grained, e.g. following Zemska (1973) – *microdialogue in the supermarket / at home / in passing, conversation at home / leisure / at meeting / phone ...*
- extreme range of token counts, e.g. conversation|phone: 178, microdialogue|in passing: 180, conversation: 57925, memoires: 23998
- ranging from 1 text per type (*obituary, advice, resolution*) to >200 (*article*)
- reduction: 29 types, 510 texts, 980493 word forms (tokens excluding punctuation)

Register

- Registers are language varieties governed/determined by functional and situational characteristics (Biber 1988, 1995, 2009; Biber & Conrad 2009)
- Language varieties are characterized by relevant linguistic features constituting dimensions of variation
- Biber's (1988) dimensions:
 - (i) Involved vs. informational production
 - (ii) Narrative vs. non-narrative discourse
 - (iii) Situation-dependent vs. elaborated reference
 - (iv) Overt expression of argumentation
 - (v) Abstract vs. non-abstract style.

Selected variables and variants

- main clause polar questions: *razve, neuželi, li* (main cl.), \emptyset (intonation)
- causal adverbial clauses: *ibo, potomu čto, tak kak*
- conditional adverbial clauses: *esli, ezheli, koli*
- partitive: masc “second genitive” in *-u* vs. (standard, less archaic) *-a*
- instrumental: sing. fem. nouns in *-a*: *-oj/-ej* vs. *-oju/-eju*
- adjectival predicate: long form vs. short form adjective
- [ongoing annotation:]
 - dynamic situations: light verb+NP vs. lexical verb
 - subject coreference: \emptyset -pro vs. NP vs. personal pronoun

Examples

- (1) Nautro podnjalsja Krasnopërov, vypil **čajju**, xotel
morning-adv rose K.-nom drank tea-part.u wanted
uxodit'.
leave
'In the morning, K. drank up tea, wanted to leave.'
- (2) sidja na divane so stakanom **čaja** ili čego pokrepče
sitting on sofa with glass-ins tea-part.a or something stronger
'sitting on the sofa with a glass of tea or something stronger'

Non-variationally analyzed features

- marked diminutives: *-en'k-*, *-on'k-* – *xorošen'kij* 'good-dim.', *kuxon'ka* 'kitchen-dim.'
- marked augmentatives/derogatives: *-un* nouns, *-ovat-* adjectives/adverbs – *gryzun* 'biter, chewer', *mnogovato* 'too much'
- verbs and deverbal nouns with prefix *vos-* (Church Slavonic origin)
- various particles, adverbs, complementizers: *vot*, *poskol'ku*, *itak*, *slovno*, *xot'*, *sledovatel'no*, *nu*, *ved'*, *vvidu*, *kasatel'no*, *skvoz'*, *pust'*, *nasčët*, *odnako*
- N/V ratio
- modal predicatives: *možno* 'is-possible/allowed', *nel'zja* 'is-impossible/disallowed', *nado* 'is-necessary'
- gerunds ("adverbial participles")
- internationalisms: *-izm*, *-acija* nouns

Method

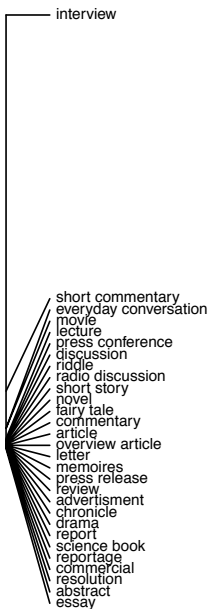
- import into R; semi-automatic annotation
- taking into account many exceptions and qualifications
 - -ovat- adjectives except *šiškovatyj* 'knobby', *kločkovatyj* 'ragged, flocky', *vinovatyj* 'guilty'
 - vos- archaisms including *voskresenie* 'resurrection', but not *voskresen'e* 'Sunday' etc.
- external manual annotation and recoding: Thanks to Natalia Graulich, Laura Perlitz, Luka Szucsich and Aleksej Tikhonov
- exploratory factor analysis (Biber 1988, 1995, 2009):
 - automatic identification of components (3 or more) which best capture the variance
 - identification of the relevant features (loading > .35) and their contributions to each factor
 - calculation of mean dimension scores per text type for each factor
 - functional interpretation of dimensions

Exploratory factor analysis

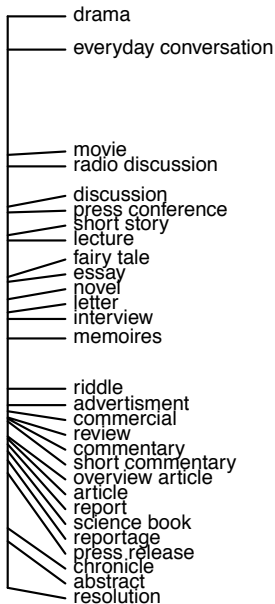
- 3 factors (sufficient, $p=3.44e-294$), 40 features, Loadings (orange $|\gt .35|$):

feature	factor1	factor2	factor3
Ccause.ibo	-.010	-.036	.017
Ccause.potomu.cho	-.007	.393	.008
Ccause.tak.kak	-.014	-.088	-.023
Lex.ezheli	-.001	.052	-.008
Lex.itak	-.006	.069	-.006
Lex.kasatelqno	-.001	.057	-.020
Lex.koli	-.003	.016	-.001
Lex.naschet	.000	.176	.075
Lex.nu	.010	.654	-.105
Lex.odnako	-.025	-.186	.105
Lex.poskolqku	-.019	-.085	-.009
Lex.pustq	-.007	.090	.006
Lex.skvozq	-.001	.077	.243
Lex.sledovatelqno	-.014	-.124	-.020
Lex.slovno	.011	-.093	1.014
Lex.vedq	-.015	.195	.015
Lex.vot	.000	.739	-.094
Lex.vvidu	-.008	-.051	.030
...
Sy.Gerund	-.034	.016	.221
Sy.N.V	-.006	-.083	-.003
Sy.Noun	.999	.004	.008
Sy.Pred	.513	-.004	-.002
Sy.Verb	.022	.581	.084
YNQ.bare	.999	.014	.008
YNQ.li.HS	-.011	.070	-.022
YNQ.neuzheli	.006	-.146	.724
YNQ.razve	-.001	.143	.120

Text types factor 1: interactive vs. informational



Text types factor 2: oral vs. literal



Example – factor 2

Lex.vot

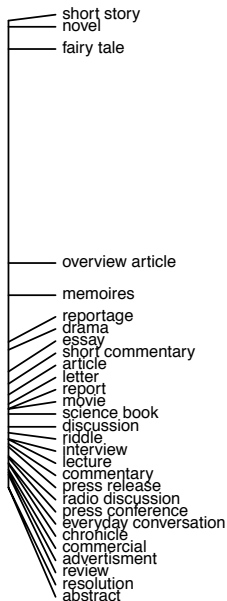
- (3) A **vot** èto ja.
and *ptcl* this I
'And so this is me.'

Lex.nu

- (4) **Nu** ja tože ne slepoj.
ptcl I also not blind
'Well, I am not blind either.'

- discourse / discourse-segmenting particles
- contextual coherence

Text types factor 3: narrative vs. non-narrative



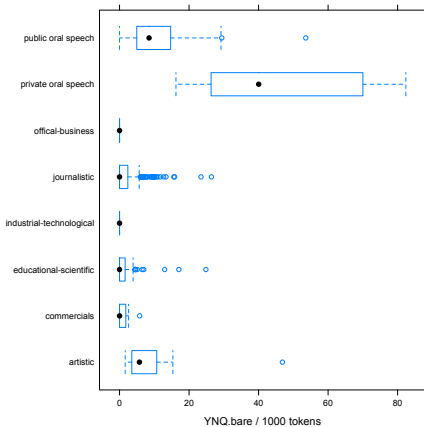
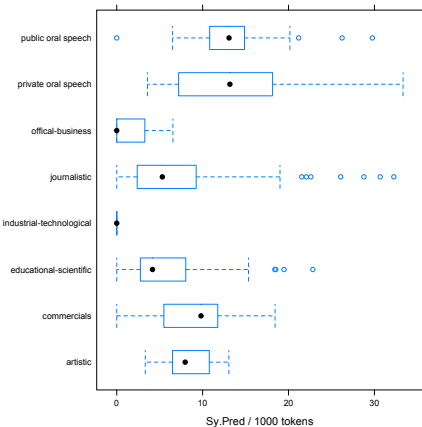
Example – factor 3

YNQ.neuzheli

- (5) **Neuželi** net zdes' ličnosti, čtoby xot' malost' na
ptcl not-exist-3sg here person that-sbj at-least small-amount to
lorda smaxivala? [memoires]
lord-acc throw-away
'Isn't there anyone here who would waste a little bit of money
for the lord?'

- indirect, emotional way of asking (often: yourself)
- requests rebuttal of a negative presupposition

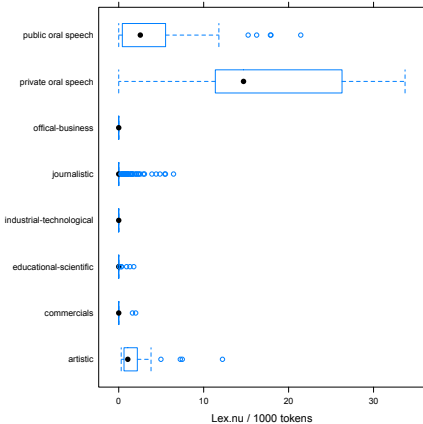
Relevant features across RNC “spheres” (Factor 1)



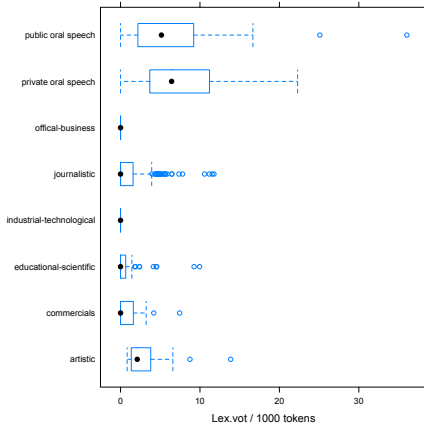
- not distinctive for functional style (except for lower frequency in official-business)

- (stylized) oral
- private vs. public oral

Relevant features across RNC “spheres” (Factor 2)

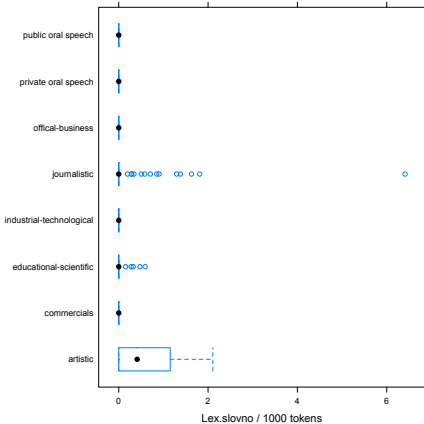


• (stylized) oral

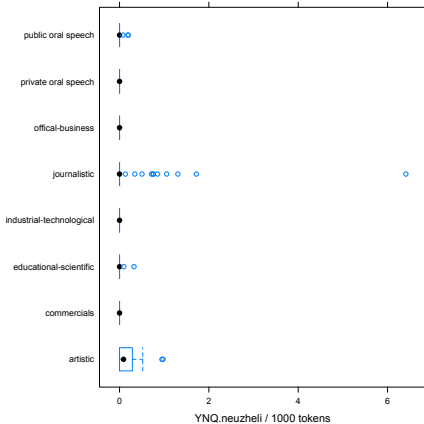


• conversational style

Relevant features across RNC “spheres” (Factor 3)



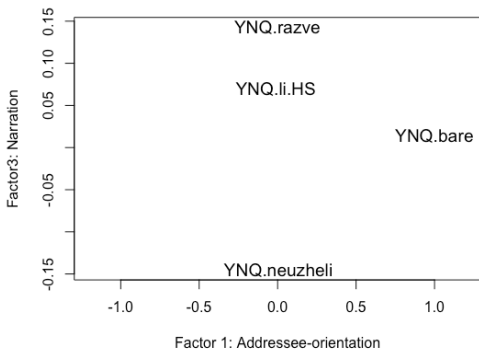
- only in artistic style



- unclear; low overall frequency

Variationist variables

- Build the grouping of variants into variables into the analysis?
- E.g., normalize not to a fixed number of tokens (1000), but to the sum of occurrences of the variable
- Or directly compare the loadings of the variants of a variable:



Conclusion

- exploration of explorative factor analysis (to be explored further)
- traditional functional styles offer too little information and are too vague
- inspection of textual data is essential for the interpretation of the factors
- variationism: Comparing the loadings of the variants of a variable in different dimensions can be instructive

Conclusion

- exploration of explorative factor analysis (to be explored further)
- traditional functional styles offer too little information and are too vague
- inspection of textual data is essential for the interpretation of the factors
- variationism: Comparing the loadings of the variants of a variable in different dimensions can be instructive

Thank you!

roland.meyer@hu-berlin.de

Come to the VallanCo Autumn School!

 *International Autumn School*

Variation <https://hu.berlin/varlanco>

25–27 Oct 2018
Humboldt-Universität zu Berlin

in Language Corpora

Douglas Biber (USA)
Václav Cvrček (Czech Republic)
Maciej Eder (Poland)
Stefan Evert (Germany)
Anke Lüdeling (Germany)

CENTRAL  **DAAD**  Deutscher Akademischer Austauschdienst
German Academic Exchange Service

 **DFG**  **DFG**  **DFG**  **DFG**  **DFG**  **DFG**  **DFG**  **DFG**  **DFG**  **DFG**  **DFG**  **DFG**  **DFG**  **DFG**  **DFG**  **DFG**  **DFG**  **DFG**  **DFG**  **DFG**  **DFG**  **DFG**  **DFG**  **DFG**  **DFG**  **DFG**  **DFG**  **DFG** **DFG**