Slavic Corpus Linguistics: The Historical Dimension
Tromsø, Norway, April 21–22, 2015
UiT The Arctic University of Norway

Corpus as a tool in real-time socio-linguistics:
the spread of an innovation in the texts of Russian 19th-century writers

Sergey Say
(St.Petersburg State University / Institute for linguistic studies, RAS)
serjozhka@yahoo.com

Outline of the talk

- Problem
- Proposal
- Innovation: Instr. Sg. Fem. *oju* > *oj*
- Data collection and analysis
- Results & discussion

Outline of the talk

- **Problem**
- Proposal
- Innovation: Instr. Sg. Fem. *oju* > *oj*
- Data collection and analysis
- Results & discussion

Problem

- Diachronic changes are gradual: linguistic change is inseparable from variation
- Variation and change have several dimensions:
  - lexical;
  - areal;
  - stylistic;
  - **social => my focus today**

Problem

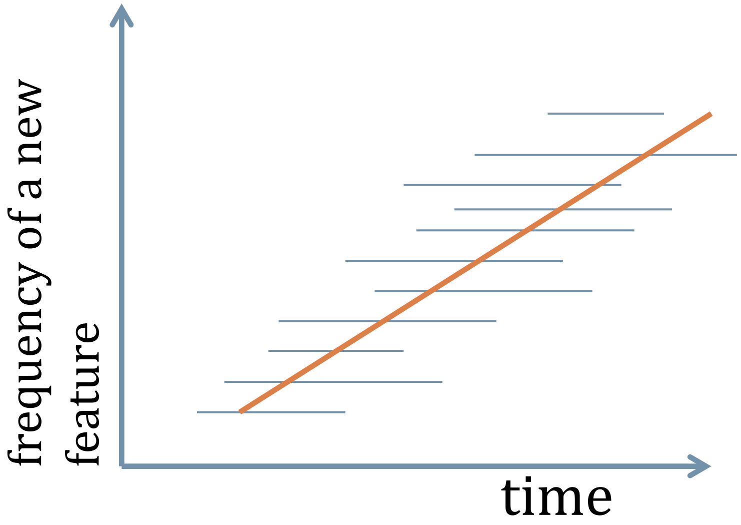
- How does (the frequency of) an innovation spread (increase) in the community?
- Two extreme possibilities:
  - Scenario 1: the speech of every individual speaker is relatively stable during his / her lifespan; the spread of innovations is largely due to **generational** shifts. It has been reported to be typical of sound change and morphological change.
  - Scenario 2: **communal** change. "All members of the community alter their frequency together" [Labov 1999: 84]. Individual speakers' age is irrelevant for language change.

Labov, William. 1999. Principles of linguistic change. Vol. 1. Internal factors. Oxford: Blackwell.

Problem

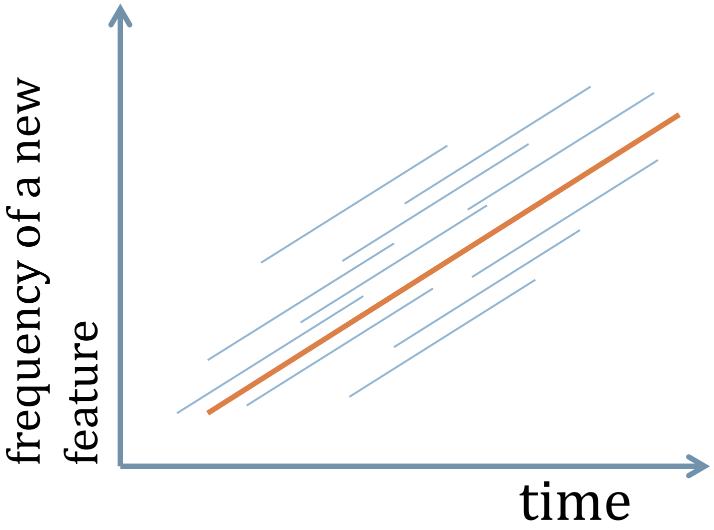
— Community

— Individual speakers



Putative radical scenario 1

Cf. critical age hypothesis
Basic phonological features?



Putative radical scenario 2

E.g. non-linguistically motivated changes (lexical frequency of *computer*, etc.)

Problem

Typical data in sociolinguistics

Typical data in historical linguistics

Synchronic Pattern	Interpretation	Individual	Community
flat	1. Stability	stable	stable
monotonic slope with age	2. Age-grading	unstable	stable
monotonic slope with age	3. Generational change	stable	unstable
flat	4. Communal change	unstable	unstable

Sankoff, Gillian. 2005. Cross-Sectional and Longitudinal Studies in Sociolinguistics. In Ammon et al. (eds.). Sociolinguistics: An International Handbook of the Science of Language and Society. Vol. 2. New York, Berlin: Walter de Gruyter, 1003–1013. Orginally in [Labov, 1994]

Problem

- A sociolinguist's approach to testing the apparent-time hypothesis: **real-time sociolinguistics**, that is, observation on the same community at multiple points of time
  - **Cross-sectional** (trend): same community (but different speakers)
  - **Longitudinal** (panel): same speakers observed repeatedly
- Longitudinal studies are generally superior, but
  - costly,
  - time-consuming,
  - "… somewhat impractical. Because sociolinguistics is such a young subfield within the linguistics discipline, there has arguably not been enough time to study individual speech communities over periodic increments and map notable changes" [Dannenberg 2000: 254].

Dannenberg, Clare J. 2000. Sociolinguistics in real time. American speech, 75.3. 254-257.

Outline of the talk

- **Problem**
- **Proposal**
- Innovation: Instr. Sg. Fem. *oju* > *oj*
- Data collection and analysis
- Results & discussion

Proposal

- Use (Russian) National Corpus (www.ruscorpora.ru) as a source of data in “post hoc” real-time socio-linguistics
- After all, (even great) writers are language users!
- The idea is not entirely new, but most previous studies are essentially cross-sectional, that is, no attempt is made to trace changes in individual speakers (= writers) during their lifespans (see [Sankoff 2005] for an overview)

Sankoff, Gillian. 2005. Cross-Sectional and Longitudinal Studies in Sociolinguistics. In Ammon et al. (eds.). Sociolinguistics: An International Handbook of the Science of Language and Society. Vol. 2. New York, Berlin: Walter de Gruyter, 1003–1013.

Proposal

- In just a few studies literary texts have been analyzed in a truly longitudinal perspective [Raumolin-Brunberg 1996; Arnaud 1998]
  - the corpora used are typically smallish;
  - such studies have never (?) been carried out on Russian (or Slavic) material

Arnaud, René. 1998. The development of the progressive in 19th century English: a quantitative survey. Language Variation and Change 10: 123-152.

Raumolin-Brunberg, Helena. 1996. Apparent time. In Nevalainen & Raumolin-Brunberg (eds). Sociolinguistics and Language History: Studies based on the Corpus of Early English Correspondence. Amsterdam: Rodopi. 93-109.

Proposal

- Advantages of the Russian National Corpus:
  - wealth of easily searchable data;
  - continued observation of same subjects
  - many authors are represented by texts that are chronologically separated by several decades
    - e.g. Bunin: 1881-1953, that is, **72 years**: much longer period than any real longitudinal survey can cover
  - texts are annotated for
    - author
    - date of creation
    - date of birth of the author (=> age at the time of writing)

Proposal

Disadvantages / disclaimers:

- socially very biased sample of speakers (definitely not representative of the entire community)
- lack of younger speakers / writers
- written texts only, conscious self-control is very likely
- relatively shallow in terms of language history
- experiments are not possible
- genre / style factors are hard to control for
- etc.

Outline of the talk

- **Problem**
- **Proposal**
- **Innovation: Instr. Sg. Fem.** ***oju*** **>** ***oj***
- Data collection and analysis
- Results & discussion

Innovation: Instr. Sg. Fem. *oju > oj*

- The loss of disyllabic Instr. Sg. endings in feminine nouns (and Fem.Sg. adjectives, and some pronouns):
  - *любовию*/l’ubov’iju/ > *любовью* /l’ubov’ju/
  - *деревнею* /d’ir’evn’iju/ > *деревней* /d’ir’evn’ij/

    ***рукою* |rukoju| > *рукой* /rukoj/**
  - *пустою* /pustoju/ > *пустой* /pustoj/
  - *мною* /mnoju/ > *мной* /мной/
- Instr. Sg. forms of (feminine) nouns from the *a*-class are in the focus of this study.

Соболевский, А.И. 1901. Лекции по истории русского языка, М. 1907. 4-ое изд. (non vidi).

Innovation: Instr. Sg. Fem. *oju > oj*

- Occurrences of syncopated forms with –*oj* are first registered in the 13th century [Sobolevskij 1907].
- However, the -*oju* > -*oj* process in still underway: disyllabic endings are sometimes used in contemporary texts.

(1) *Давно с этой* ***программою*** *пытаюсь подружиться, но все никак не сработаемся* [vk.com; 03.09.**2014**].

(2) [Этот персонаж] *оказался единственным, кто смог дожить до конца, ибо он возглавил этот конец, но* ***ценою*** *своей жизни* [НКРЯ; коллективный. Рецензии на фильм «V значит вендетта» (**2006-2010**)]

Innovation: Instr. Sg. Fem. *oju* > *oj*

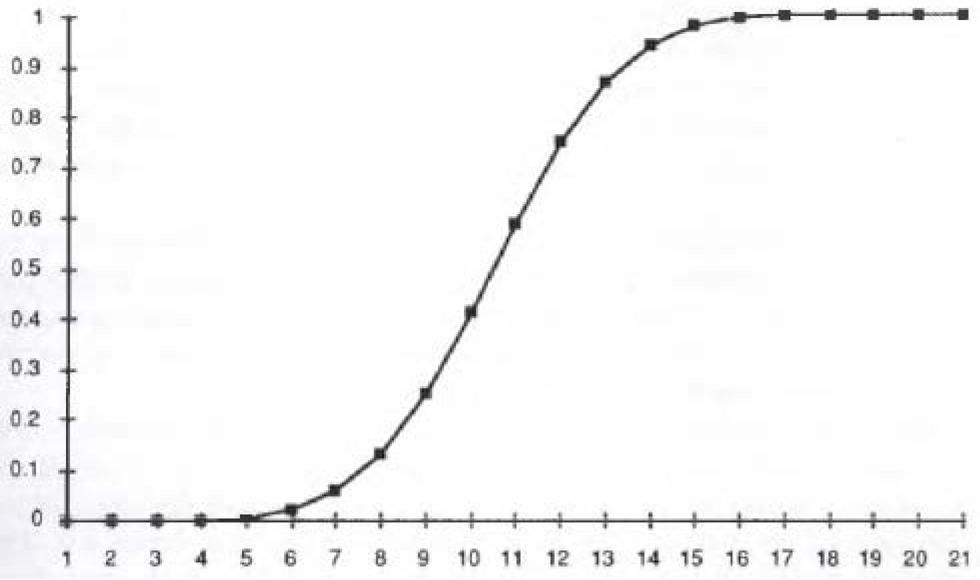
- The most rapid stage in this development was clearly the 19th century

Russian National Corpus (1701-2012):

	N (-*oj*)	N (-*oju*)	% (-*oj*)
≥ 1800	3321	9528	**26%**
1801-1820	2614	6365	**29%**
1821-1840	10521	16216	**39%**
1841-1860	31558	16852	**65%**
1861-1880	61853	18901	**77%**
1881-1900	69125	13266	**84%**
> 1900	868 161	40 312	**96%**

Innovation: Instr. Sg. Fem. *oju > oj*

- Typical **S-curve** development [Weinreich et al. 1968], cf. “sound changes begin in a slow rate, progress rapidly in midcourse, and slow down in their last stages” [Labov 1994: 65].



Weinreich, U., W. Labov, M. Herzog. 1968. Empirical foundations for a theory of language change. Austin: UTexas.

Labov, W. 1999. Principles of linguistic change. Vol. 1. Internal factors. Oxford: Blackwell.

Innovation: Instr. Sg. Fem. *oju > oj*

Advantages of the –*oju* > -*oj* change for my purposes:

- a very frequently observed variable;
- the most rapid phase was in the 19th century, which is well documented in the RNC;
- not directly related to semantics & syntax , hence not as sensitive to genre and style factors as many other variables;
- more or less independent of the lexical content;
- rather unconstrained variation, e.g. variants can easily co-occur in the same sentence:

(3) *Купец … вместо того, чтобы съесть, как полагается, пирожка с* ***тешечкой****, пирожка с* ***визигою****, с осетровой* ***щекой****, с налимьей* ***печенкою****, … — околоточного вкусил*. [В. М. Дорошевич. Дело о людоедстве (1900)]

Innovation: Instr. Sg. Fem. *oju > oj*

- Basic question: is this change **generational** (as expected for phonetic and morphological changes [Labov 1999: 84]), **communal** or somewhere between the two extremes?

Labov, W. 1999. Principles of linguistic change. Vol. 1. Internal factors. Oxford: Blackwell.

Innovation: Instr. Sg. Fem. *oju > oj*

- Other factors, although they do play an important role in this change, are disregarded in this study. See [Katlinskaja 1969] for the list of factors that correlate positively with the choice of –*oj*:
  - fiction > non-fiction
  - adjective > noun > pronoun
  - adnominal modifier > object > adverbial > predicative > agent (passive)
  - presence of other instrumental forms > absence of such forms
  - 5 syllables > 4 > 3 > 2 > 1
  - /o/ stressed > /o/ not stressed

Катлинская, Л.П. 1969. Условия варьирования флексий творительного имен женского рода в литературном языке конца XVIII – начала XIX века. Филологические науки, № 3. 110-119.

Outline of the talk

- **Problem**
- **Proposal**
- **Innovation: Instr. Sg. Fem. *oju* > *oj***
- **Data collection and analysis**
- Results & discussion

Data collection and analysis

- Focus on the 19th century
- Major writers:
  - all those who have at least 100.000 word tokens in the RNC + 5 other important writers (Griboyedov, V. F. Odoyevsky, Belinsky, A. K. Tolstoy, Sukhovo-Kobylin)
  - under the stipulation that at least **some** of the texts were created in the 19th century
- **50** writers overall
- Fairly conventional 20-year-long periods*: 1801-1820, 1821-1840, etc.

*There are some undesirable border effects: some texts are taken into account for two periods. This problem has to be solved later.

Data collection and analysis

- Raw data:
  - N(*oju*): the number of hits for the search query:
    - Instrumental Singular form,
    - ends in –***ою***,
    - feminine gender noun that
    - ends in –a in its basic form
  - N(*oj*): the number of hits for the search query:
    - Instrumental Singular form,
    - ends in –***ой***,
    - feminine gender noun that
    - ends in –a in its basic form

Data collection and analysis

- Some query results have been checked manually, noise ratio is negligibly low*.

$$\text{p}(\textit{oj}, \text{Writer, Period}) = \frac{N(oj,W,P)}{N(oj,W,P)+N(oju,W,P)}$$

e.g.

$$\text{p(oj, Herzen, 1841-1860)}=\frac{1247}{1247+137} \approx 0.90$$

The higher the p(oj) value, the more innovative (less conservative) is the author in the period.

*Besides, false hits are mostly feminine adjectives, which show a similar pattern of variation anyway.

Data collection and analysis

Fragment of the table with p(*oj*) values:

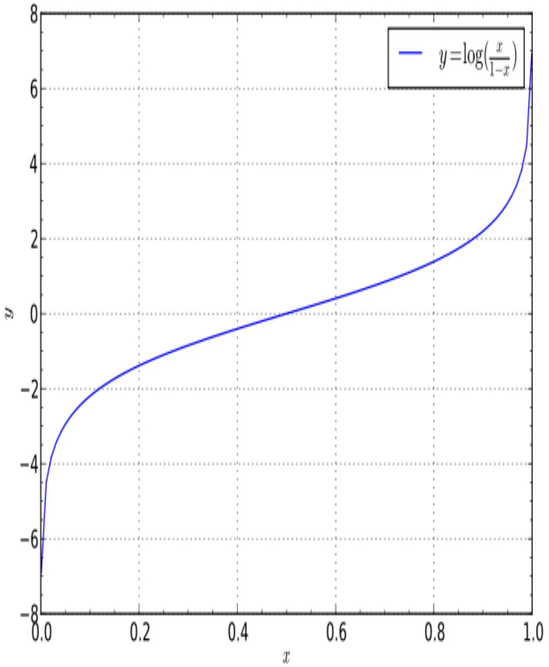
Writer	**date of birth**	**< 1801**	**1801-1820**	**1821-1840**	**1841-1860**	**1861-1880**	**1881-1900**	**>1900**
…								
Pushkin	1799		0,15	0,33				
Vladimir Dal	1801			0,60	0,81	0,82		
Odoyevsky	1803			0,24	0,39			
Gogol	1809			0,32	0,55			
Belinsky	1811			0,20	0,20			
Goncharov	1812			0,60	0,92	0,98	0,76	
Herzen	1812			0,70	**0,90**	0,96		
…								

Data collection and analysis

- Raw probabilities were transformed using logit.

$$L=\text{logit}(p(oj))=\log\left(\frac{p(oj)}{1-p(oj)}\right)$$

$$L(\text{Herzen, 1841-1860})=\log\left(\frac{0.90}{0.10}\right)=2.21$$



The rationale: logit captures the non-linear nature of probability curves. The difference between p=0.2 and p=0.1 is more significant than e.g. the difference between p=0.6 and p=0.5.

Data collection and analysis

- Measuring changes in individual writers during their lifespans.

$$\Delta(\text{Writer}, Period_{i+1}, Period_i) = \text{L}(\text{Writer}, Period_{i+1}) - \text{L}(\text{Writer}, Period_i)$$

e.g.

Δ(Herzen, 1841-1860, 1821-1840) = 2.21 – 0.82 = 1.39

A positive value of Δ is observed if the relative frequency of the innovative ending –*oj* increases.

Data collection and analysis

$$\bar{L}(Period) = \frac{\sum_{i=1}^{n} L(Writer_i, Period)}{n}$$

Period	$\bar{L}$
<1801 (19th century writers)	-1.51
1801-1820	-0.88
1821-1840	-0.45
1841-1860	0.85
1861-1880	1.68
1881-1900	2.01
> 1900 (19th century writers)	2.14

Data collection and analysis

- How conservative / innovative is a writer relative to other writers who are active in the same period?

Z-score(Writer,Period)=$\frac{L(Writer,Period)-\bar{L}(Period)}{\sigma(L(Period))}$,

where $\sigma(L(Period))$ is the standard deviation of L for the writers of the period.

Data collection and analysis

- For example:

L (Herzen, 1841-1860)≈ 2.2

$\bar{L}$ (1841-1860) ≈ 0.85

$\sigma\left(L(1841 - 1860)\right) \approx 1.29$

z-score (Herzen, 1841-1860) ≈ $\frac{2.21-0.85}{1.29} \approx \mathbf{1.06}$

This means that in the years 1841-1860 Herzen tends to use innovative –*oj* forms **more** frequently than other writers of this period, and his score is more than one standard deviation above the mean.

Data collection and analysis

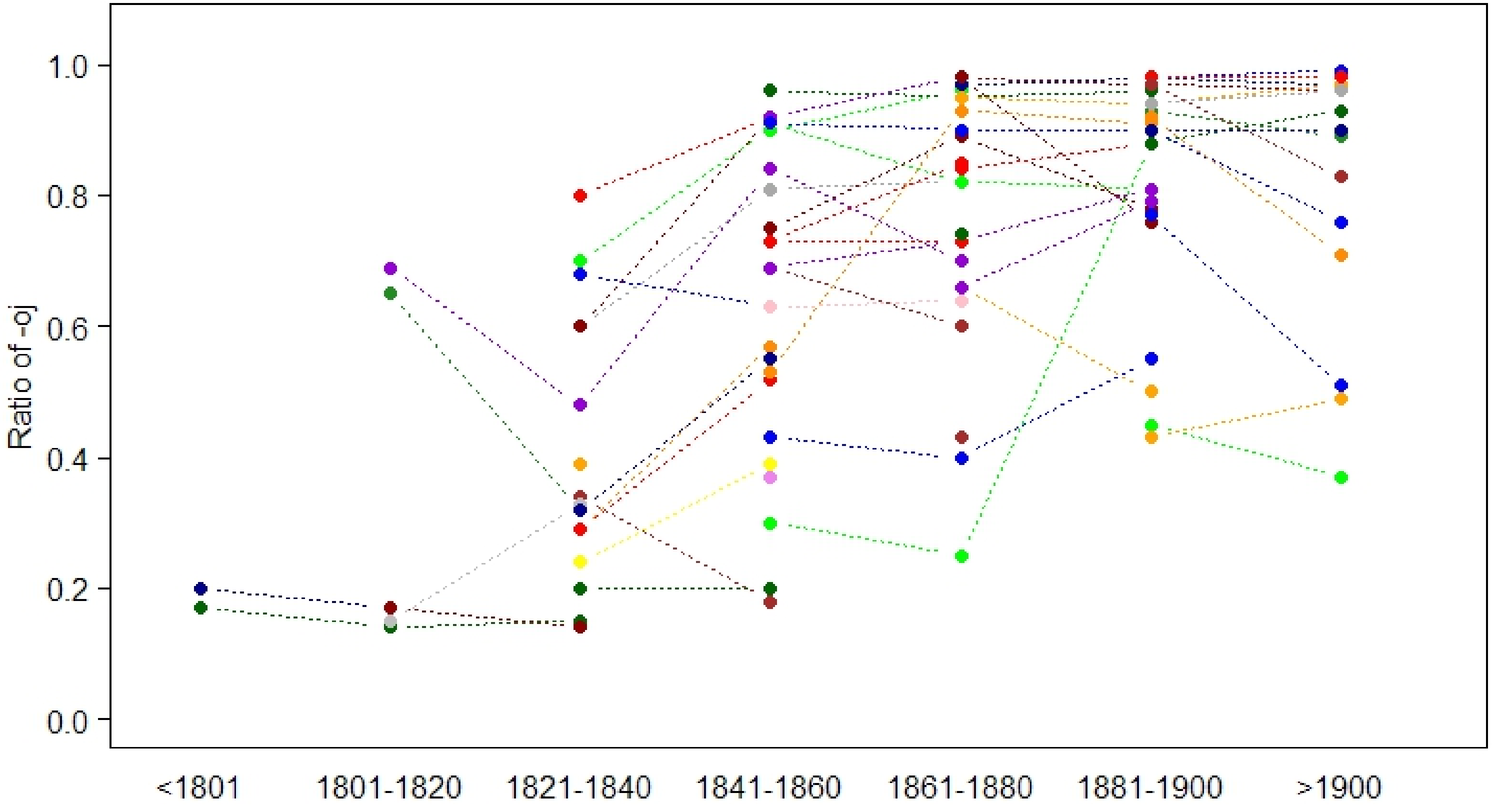
- Author's **age**: the 10th year of each 20-year-old period was used as an estimate of the age of the author for that period.
- E.g. Herzen's age for the period 1841-1860 was set at 38, his actual age in 1850/1851.

Outline of the talk

- **Problem**
- **Proposal**
- **Innovation: Instr. Sg. Fem. *oju* > *oj***
- **Data collection and analysis**
- **Results & discussion**

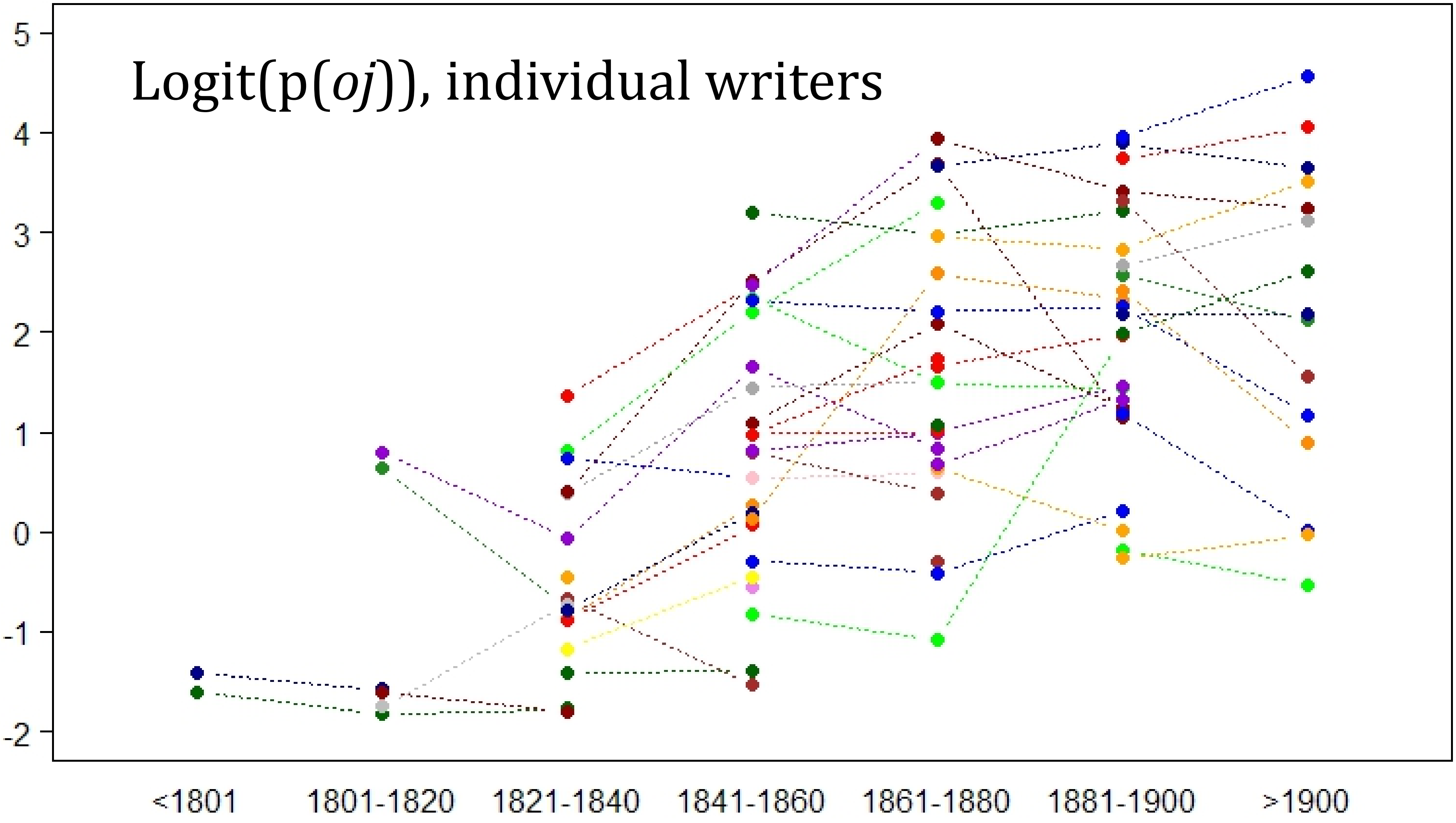
Results & discussion

Raw data: p(*oj*) for individual writers



Thanks to Masha Ovsjannikova for the graphs!

Results & discussion



Results & discussion

- There is a huge dispersion between individual writers, even if they belong to the same generation and write at more or less the same time.
- Conservativeness of individual writers with respect to -oju/-oj variation strongly correlates with conservativeness with respect to other types of changes [Tixomirov, in prepartion]:
  - pre- vs. postnominal possessors: *дом мой > мой дом*
  - *сей > этот*

Results & discussion

	year of birth	p(oj, 1861-1880)	z-score
A.K.Tolstoy	1817	0,64	-0.77
Sukhovo-Kobylin	1817	0,60	-0.91
Kostomarov	1817	0,43	-1.40
Buslaev	1818	0,40	-1.48
Turgenev	1818	0,82	-0.14
Melnikov-Pechersky	1818	0,95	0.90
Dostoyevsky	1821	0,89	0.28
Pisemsky	1821	0,85	0.03
Grigorovich	1822	0,93	0.63
Ostrovsky	1823	0,98	1.58
Saltykov-Schedrin	1826	0,73	-0.48
Leo Tolstoy	1828	0,90	0.37
Chernyshevsky	1828	0,25	-1.95
Leskov	1831	0,66	-0.74
RNC		**0,77**	

Results & discussion

- Expectedly, both the date of creation (r = 0.61) and the date of the author’s birth (r = 0.58) correlate strongly with the Logit(*oj*) values.

Results & discussion

- Most writers tend to show significant changes during their lifespans, that is far beyond the “critical age”.

$$\bar{\Delta}(\mathrm{L}(Writer_i, Period_{j+1})\text{-}(\mathrm{L}(Writer_i, Period_j))=\mathbf{0.16}$$

- This is an estimate* of the rate of change in individual writers (e.g. from 50% to 54% of –*oj* in 20 years).

$$\bar{\Delta}(\bar{L}(Period_{j+1})\text{-}(\bar{L}(Period_j))=\mathbf{0.42}$$

- This is an estimate of the rate of change in the whole community of writers (e.g. from 50% to 60.5% of –*oj* in 20 years).

*This might be a somewhat understated estimate because individual authors are not necessarily active throughout whole 20-year-long periods.

Results & discussion

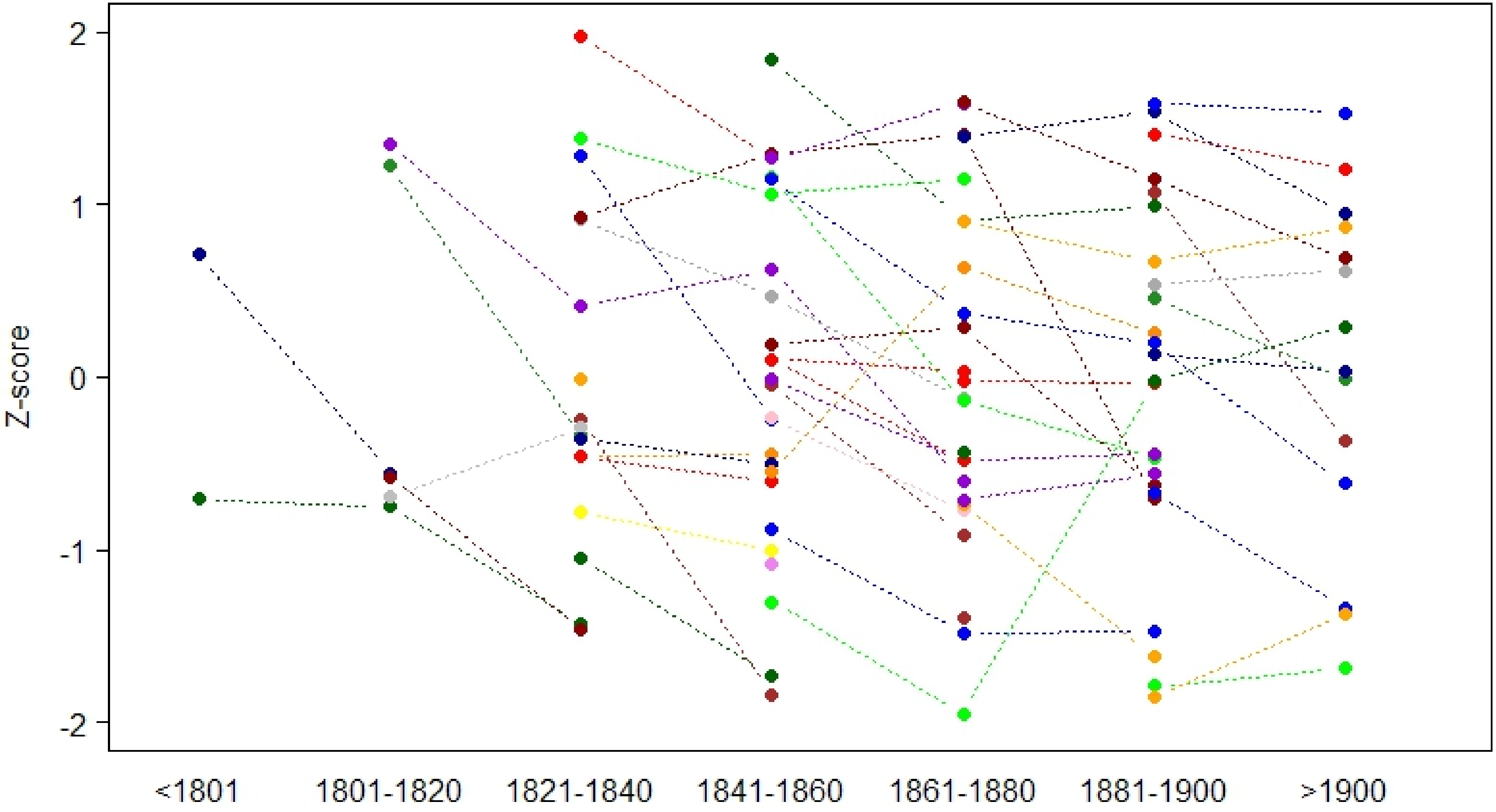
- Thus, in this case study, “individual speakers change over their lifespans in the direction of a change in progress in the rest of the community” [Sankoff 2005:1011].
- However, the **average** pace of change in individual writer is approximately two times lower than in the community in general.
- The data at hand allow us to reject the apparent-time hypothesis in its utter form (generational differences faithfully reflect stages of language change).
- The observed scenario is somewhere between generational change and communal change.

Results & discussion

- As a consequence, an almost exceptionless pattern: innovativeness of an individual writer **relative to the current period decreases** over time.
- That is, regardless of whether a writer starts as a conservator or innovator, later writings are almost always closer to the conservative end of the distribution (due to appearance of writers from younger generations).

Results & discussion

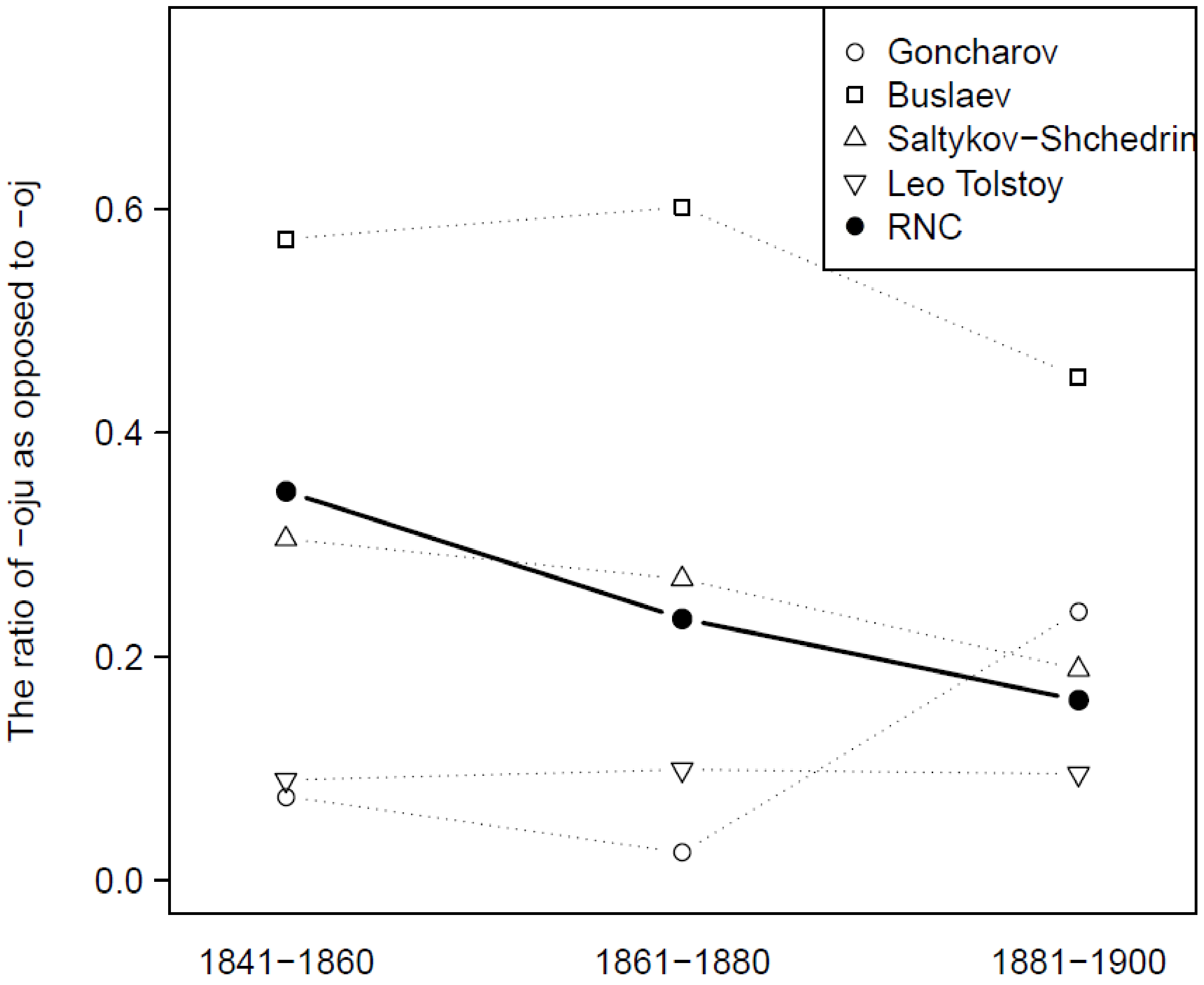
Changes in z-scores (“innovativeness”) with age



Results & discussion

- Writers are very heterogeneous not only in terms of absolute frequencies of the use of old / new forms, but also in terms of their inclination to follow the communal path of change.
- Leo Tolstoy, for example, is exceptional in that he shows a fairly stable p(-*oj*) from his earliest writings and up to the end of the 19th century (next slide).

Results & discussion

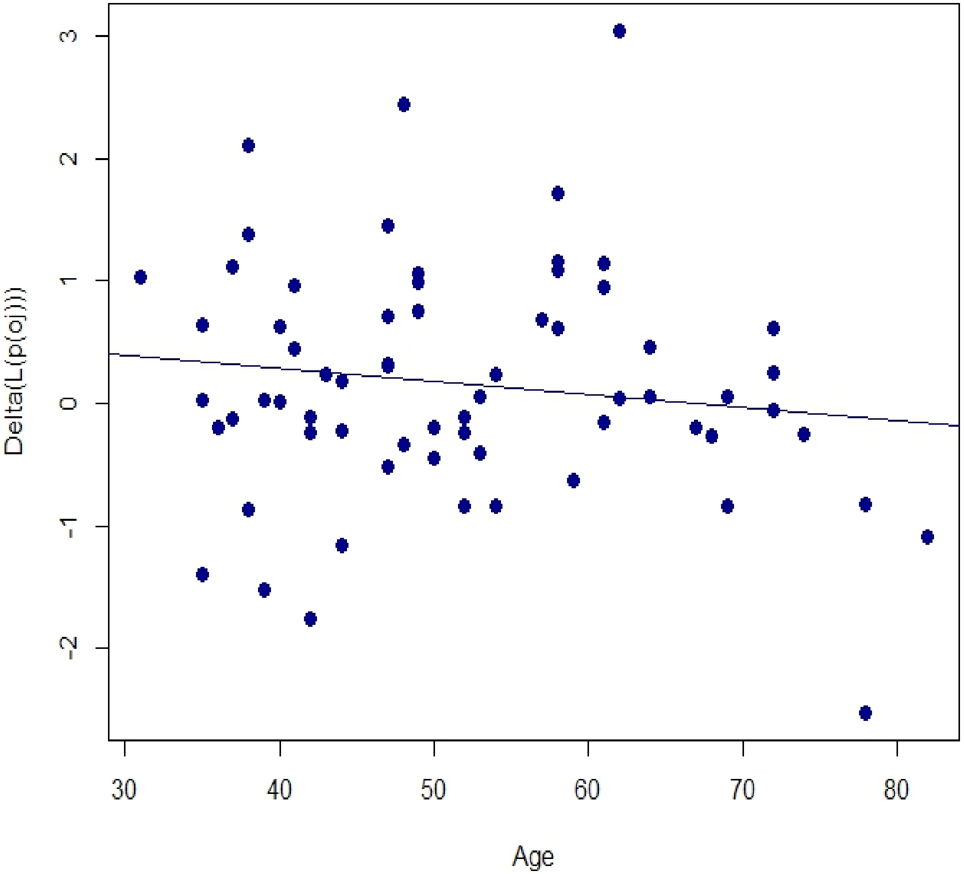


Results & discussion

Conformity of individual writers to the communal trend correlates negatively (R = -0.14) with age: older writers move in the general direction slower that younger writers, and sometimes even turn to more archaic patterns of use (see the case of Goncharov on the previous slide).

Results & discussion

Age and direction of change in individual writers (conformity to the communal trend)



Thank you!