Today’s arguments are yesterday’s circumstantials: a corpus-study of Russian valency patterns

Maria Ovsjannikova
Institute for Linguistic Studies, St.Petersburg
masha.ovsjannikova@gmail.com

“Slavic Corpus Linguistics: The Historical Dimension”
Tromsø, April 21-22, 2015

Prepositional argument encoding: examples

- *смотреть на* ‘look at’

(1) *Я **смотрю** на Ксению.*

- *победа над* ‘victory over’

(2) ***Победу** над ней приближают усилия многих специалистов.*

- *виновный в* ‘culpable of’

(3) *Она считает ее **виновной** в этом плохом конце…*

Criteria for data selection

- non-spatial (i.e. abstract) meaning of the preposition
- the choice of the preposition is specified by the head

- In terms of traditional descriptions:
  - maximally strong government («максимум связи») in terms of [Peshkovskij 1928/2001]
  - “predictive obligatory government with syntactic relations” («предсказующая обязательная связь с синтаксическими отношениями») in terms of [Beloshapkova 1977]

Synchronic view

- Such uses are sometimes discussed in terms of “semantically empty”, or “purely functional”, cf. [Vinogradov 1947/2001].
- The meaning of the preposition in such cases can be induced from the meaning of semantically coherent group of lexemes governing it, cf.:

защищать	‘protect’
лечить	‘heal’
спрятать	‘hide’
избавить	‘spare’
etc.	

от

“verbs meaning ‘to prevent smth undesirable or threatening’, which is indicated by the phrase *от*+Gen” [Zolotova 2006]

Diachronic view

- Lexicalization:
  - “there exists a gradient from less lexicalized to more lexicalized prepositional verbs (= which class includes what I call prepositional encoding of arguments – *M. O.*), differentiated by their degrees of fusion and idiomaticization” [Brinton, Traugott 2005: 128].
  - Prepositions in combinations where the verbs are said to govern them “lose their independence from the verb and are somehow subsumed under its meaning” [Lehmann 1982/1995: 89].

Goals and data

- Are there any recurring diachronic mechanisms that lead to the creation of valency patterns? If yes, how can we empirically detect them?

- Data:
  - texts from the XVIII-XX cc. from the Russian National Corpus (www.ruscorpora.ru)

Method

- Problem:
  - The modern speakers’ judgment on the argumental vs. circumstantial status of a participant can be misleading even if the argument encoding strategy is stable over time.

- Method:
  - Inspect the lexical distribution, or profile, of the encoding strategy: arguments are known to be more liable to lexical restrictions, whereas circumstantials freely combine with open classes of lexemes, cf. [Apresjan 1974].

Technique

- Extract a large number (1000-2000) of random examples for preposition X.
- Extract from the sample all the uses of the preposition in one of its meanings as defined by the general type of situation and properties of participant and determine the head lexeme.
- Compare the profiles of head lexemes for this me for different periods covered by the corpus (usually three periods).
- Disadvantages of the technique:
  - time-consuming
  - largely based on semantic analysis, i.e. on intuition

Results

- I will try to generalize major trends in the development of lexical head + preposition combinations
by showing several typical cases.

Results: lexical distribution 1

- *от* ‘from’ encoding the undesirable or threatening participant:

(4) *…сии иноверцы* 〈…〉 *живут в* ***безопасности*** *<u>от обид и оскорблений</u>…* [неизвестный. О Вгаабисах // Вестник Европы, Часть 19, № 1, 1805]

‘…these heterodoxes live in safety from offences and insults…’

Results 1: lexical distribution 1

- Inspecting the lexical range:

Lexical heads in 1720-1770 sample	Token frequency	N of lexemes with a given token frequency
избавить	7	1
скрыть, *спасти*	6	2
освободить	5	1
воздержаться, *удержать*	4	2
защитить, *избавление*, *освободиться*, *свободный*, *скрыться*	3	5
отводить, *охранять*, *спасение*, *уволить*, *утаить*	2	5
безопасность, *воздержание*, *защититься*, *избавиться*, *избежать*, *излечиться*, ***исключить***, ***надежный***, *облегчение*, ***невинный***, *остерегать*, *охранительный*, *отрешиться*, *помощь*, *предохранять*, *простить*, *принять*, *сохранять*, *спастись*, ***тень***, *уберечься*, *увольнение*	1	22
Overall number of types and tokens	79	38

Results 1: lexical distribution 1

- Inspecting the lexical range:
  - lexemes unlikely to subcategorize for the “threatening participant”: *невинный* ‘innocent’, *тень* ‘shadow’ in the sample for the XVIII c.:

(5) *…под теми деревьями, ⟨…⟩ делающими мне, якобы в знак благодарности своей, приятную* ***тень*** *<u>от жара солнечного</u>…* [Я. П. Шаховской. Воспоминания (1766-1777)]

‘…under those trees that make for me, as though thanking me, pleasant shade <u>from the heat of the sun</u>…’

Lexical heads in 1790-1840 sample	Token frequency	N of lexemes with a given token frequency
скрыть	12	1
избавиться, освободить, сохранить, спасти,	7	4
удержаться	6	1
защита, свободный, укрыться	5	3
защитить	4	1
воздержаться, очиститься, предохранить, спастись	3	4
избавление, ***молитва****, освободиться, охранять, предостеречь, сберечь, скрытен, таить*	2	8
безопасность, воздержание, исправить, избежать, исцелиться, крыться, лечиться, ***льгота****, облегчить, оградить, отделаться, охраниться, очистить,* ***покой****, предостеречься, предосторожность, прятать, сбережение, скрыться, спасение, убежище,* ***уволить****, удержать, уклонить, уклониться, устоять, хранить, чист*	1	28
Overall number of types and tokens	121	50

Results 1: lexical distribution 1

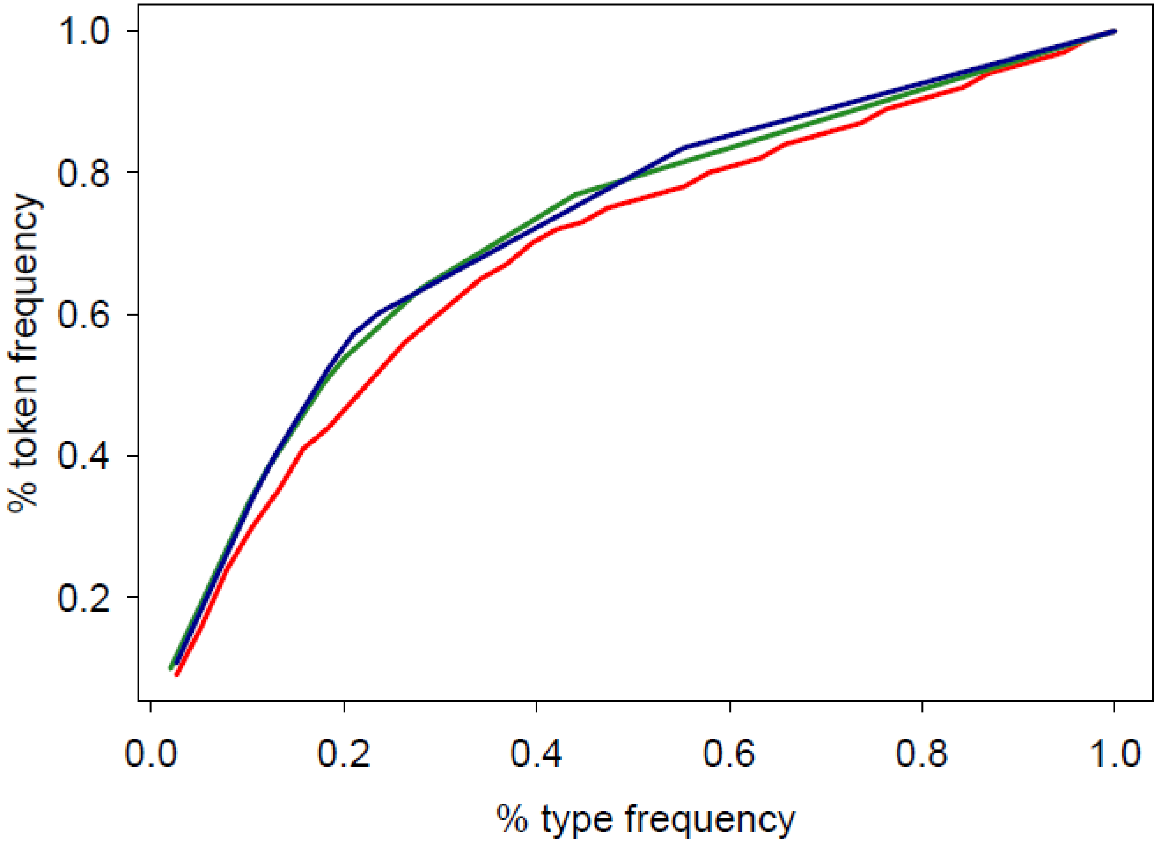
- Inspecting the lexical range:
  - lexemes unlikely to subcategorize for the “threatening participant”: *покой* ‘rest’, *молитва* ‘prayer’ in the sample for the second period.

(6) *…а <u>от мертвецов и выходцев из того света</u>*
есть у меня ***молитвы****...* [Н. В. Гоголь. Вий (1835-1841)]
‘…and from the deadmen and ghosts I have prayers…’

Lexical heads in 1980-2000 sample	Token frequency	N of lexemes with a given token frequency
освободить	11	1
защитить, избавить, освободиться	8	3
избавиться	7	1
свободный, спасти	6	2
скрыть	5	1
очистить	3	1
избавление, лекарство, лечить, ограждать, освобождение, отделаться, предостеречь, прятать, скрыться, спастись, удержаться, уклониться	2	12
воздержание, воздержаться, гарантировать, защита, защититься, лечиться, обезопасить, очиститься, применяться, расчищать, расчищаться, тайна, уберечься, уклонение, укрытие, укрыться, утаить	1	17
Overall number of types and tokens	103	38

Results 1: lexical distribution 1

Cumulative token frequency plotted against cumulative type frequency for the three samples of *om*:
18th c. is shown in red,
19th c. – in green,
20th c. – in dark blue.



Inspired by [Goto, Say 2009].

Results 1: lexical distribution 1

- Token frequencies of verbs found with *om* encoding threatening participants become less evenly distributed over time:
  the shape of the graph becomes more concave upward.

- In the distributions for the previous periods we observe lexemes that are unlikely to subcategorize for a threatening participant.
  With these lexemes threatening participants might have been used as circumstantial elements.

Results 1: lexical distribution 2

- *перед* ‘in front of’ encoding the Standard of comparison:
  - with lexical heads implying scale:

(7) *Российский язык* ***избыточествует*** *<u>перед прочими</u> для некоторых предлогов седьмым особливым падежом, который без них нигде не употребляется.* [М. В. Ломоносов. Российская грамматика (1755)]

‘The Russian language is redundant <u>in front of other languages</u> in that it has for some prepositions the special seventh case, which is not used anywhere without them’.

Results 1: lexical distribution 2

- *перед* ‘in front of’ encoding the standard of comparison:
  - with comparative forms of adjectives and adverbs:

(8) *…отчего в нем* ***изобильнее*** *бывают соляные частицы* *<u>перед масличными</u>.* [Иван Лепехин. Дневные записки (1768-1769)]
‘…for this reason in it (grapes – *M. O.*) saline particles tend to be more abundant <u>in front of butyric</u>’.

- Comparative forms constitute an open class of potential lexical heads, i.e. in such cases *перед* is not subcategorized for by the lexeme.

Lexical heads in 1720-1800 sample	Token frequency	N of lexemes with a given token frequency
преимущество	16	1
отменный	4	1
выгода, *ничто*	3	2
выгодный, ***лишний***, ***лучше***, *предпочтительно*, *тягость*	2	5
бог, ***вглубленный***, *все плюнь*, *возвысить*, *великий*, *избыточествовать*, ***изобильнее***, ***малосмыслен***, ***мал***, ***младенец***, ***меньше ужасен***, ***менее***, ***неудобность***, *ничтожен*, *особливый*, *нищ духом*, *отличаться*, *переменен*, *отягощен*, *преимуществовать*, *разница*, *старший*, ***счастлив***, *стоить*	1	24
Overall number of types and tokens	62	33

Results 1: lexical distribution 2

- The use of *перед* with comparative and superlative forms of adjectives and adverbs decreases over time:

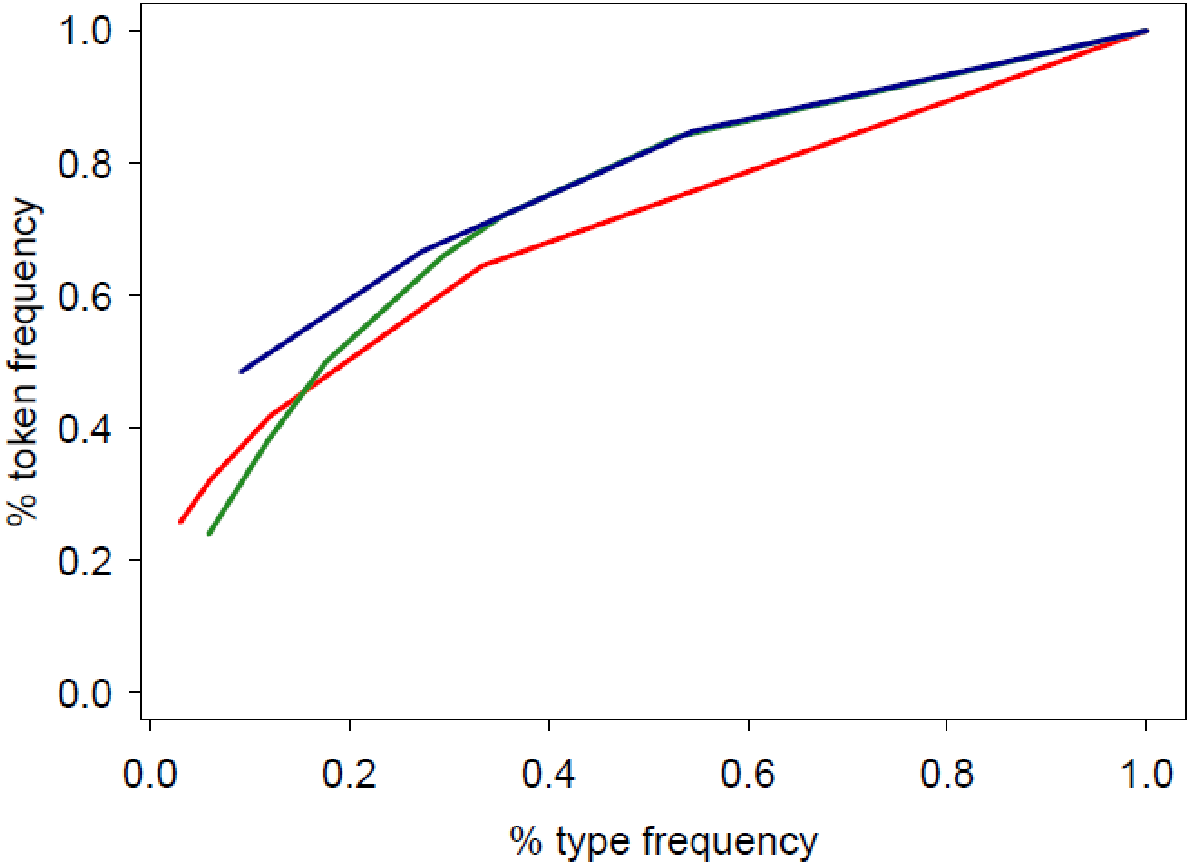
Period	**compar and superl forms with *перед*/*пред***	**N of compar and superl forms**	**N of uses with *перед* / *пред* per 10 000 of compar and superl**
< 1850	25	70 791	3,53
1850-1900	12	174 751	0,69
1970-2000	0	293 876	0

The differences between the first and the second and the second and the third periods are statistically significant: χ^2=25.2 , df = 1, p << 0.01, Exact Fisher test (two-tailed), p << 0.01.

Results 1: lexical distribution 2

Cumulative token frequency plotted against cumulative type frequency for the three samples of *перед*:
18th c. is shown in red,
19th c. – in green,
20th c. – in dark blue.

Note the difference in the relative frequency of the first frequent verb = the location of the left end of the line on the y-axis.



Results 1: lexical distribution 2

- The distribution of *перед* as shown by the graph becomes more concave over time.
- *Перед* encoding the Standard of comparison evidently lost part of its uses – those in which it was used with comparative forms, which importantly are an open class of lexical heads.
- Over time, *перед* encoding the Standard of comparison comes to be used in a much more narrow range of contexts.

- But maybe it’s just the way Russian syntax developed?...

Comparison 1: lexical distribution 3

- *на* encoding the Stimulus of perception:

(9) *Лиза пошла, но глаза её сто раз* ***обращались*** <u>*на Эраста*</u>...

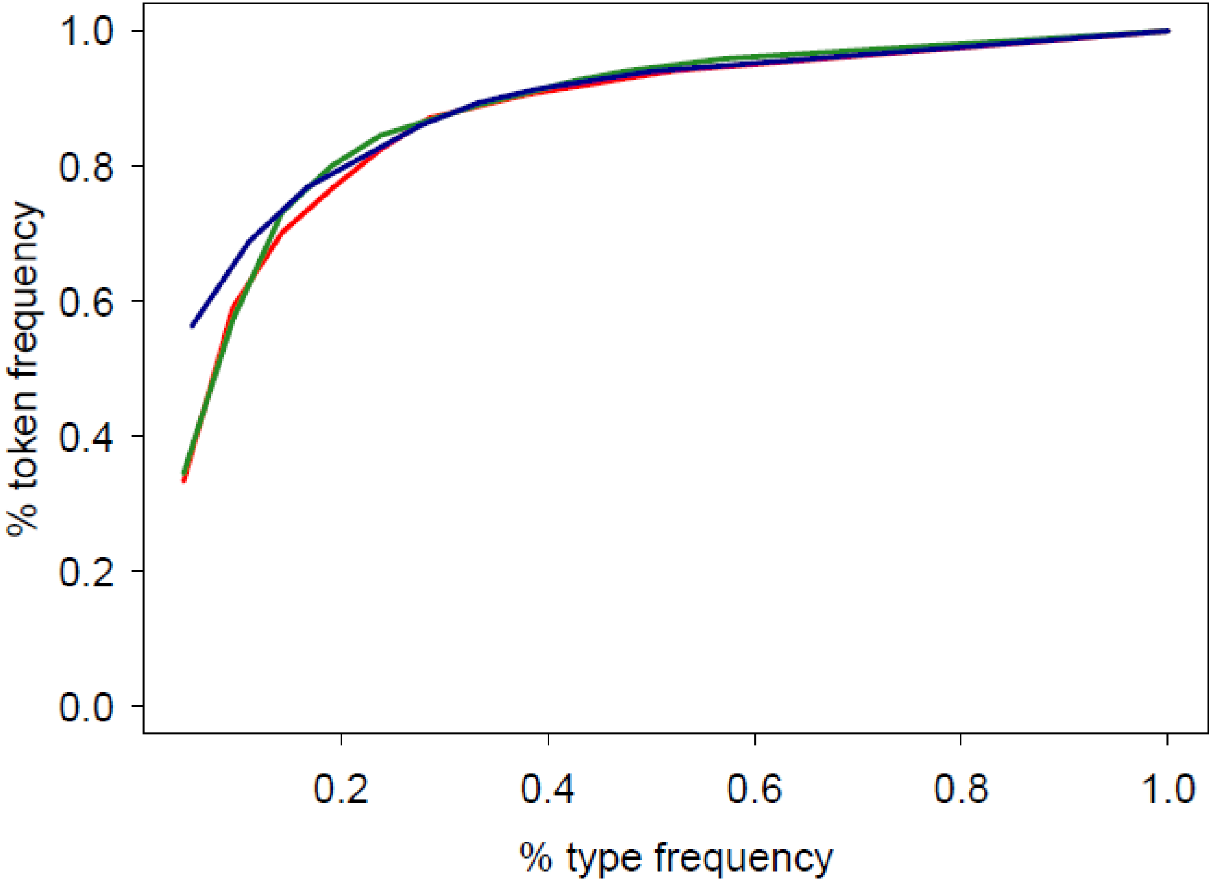
[Н. М. Карамзин. Бедная Лиза (1792)]

‘Liza went, but her eyes turned at Erast hundred times…’

Results 1: lexical distribution 3

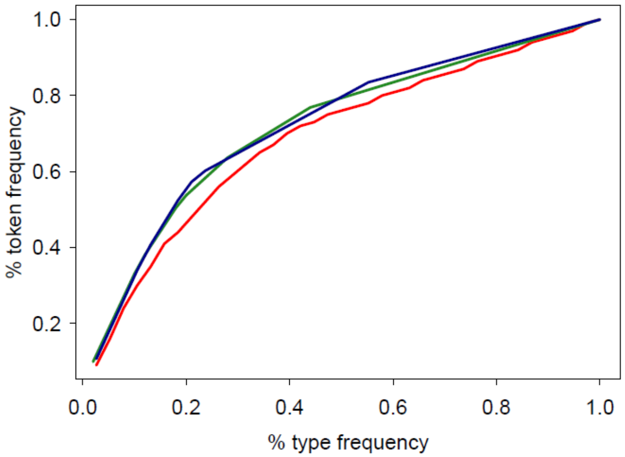
Cumulative token frequency plotted against cumulative type frequency for the three samples of *на*:
18th c. is shown in red,
19th c. – in green,
20th c. – in dark blue.

Almost no change in the structure of the frequency distribution!

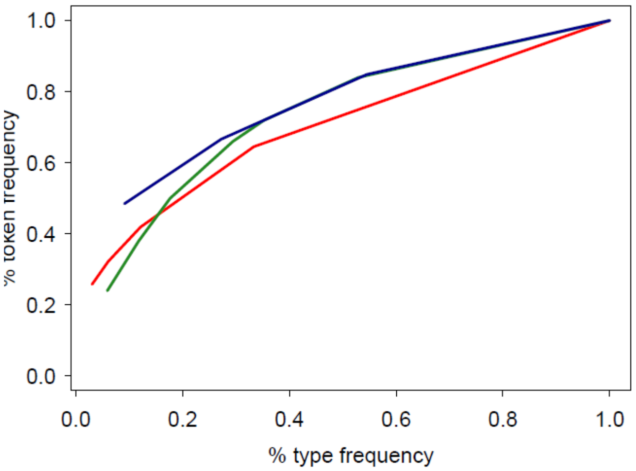


Comparison: lexical distributions

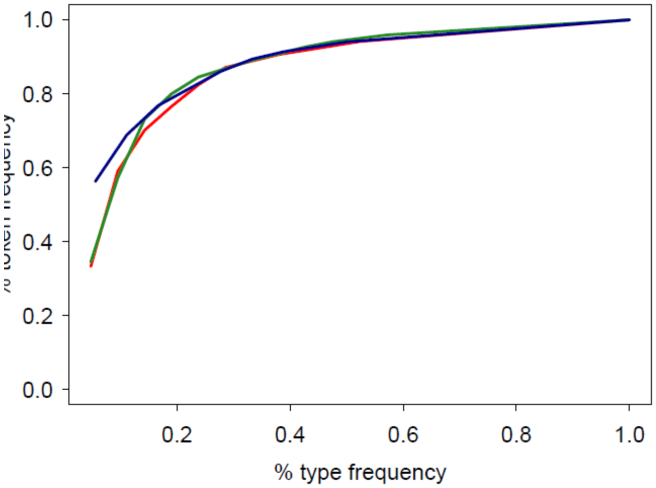
от with Threat



перед with Standard



на with Stimulus



- The structure of the lexical distribution for *на* encoding Stimulus did not undergo such a noticeable change as is observed in the two other cases.
- The major part of tokens for *на* in all periods is contributed by 3-4 very frequent lexemes, while in the other groups the distribution is more even.

Results and comparison: productivity

- Potential productivity measure from [Baayen 1993; 2009]:

> “The potential productivity of a rule is estimated by its hapax legomena in the corpus divided by the total number of its tokens N(C) in the corpus: P= V(1,C,N)/N(C). This ratio, known as the category-conditioned degree of productivity (Baayen 1993), estimates the growth rate of the vocabulary of the morphological category itself” [Baayen 2009: 8].

- What is the ratio of hapax legomena to the total number of uses of the preposition in a given meaning in the observed samples?

Summary 1: lexical distribution and productivity

- Potential productivity (N of hapaxes/N of tokens) for the three cases:

	от **+ ‘threat’**	***перед*** **+ Standard**	***на*** **+ Stimulus**
18th c.	0.28	0.35	0.06
19th c.	0.23	0.16	0.04
20th c.	0.17	0.15	0.06
Sign. level	p ≈ 0.065	p ≈ 0.01	p ≈ 0.99

  - Cochran-Armitage trend test for proportions, independence_test() in {coin}, [Hothorn 2014].

- Substantial decrease in potential productivity for *от* and *перед*, no significant change for *на*, which is unproductive already in the 18th c.

Results 2: syntactic bondedness

- Check (one of) the most frequent items of the lexical distribution for syntactic bondedness with the preposition.
- Intervening material (circumstantials, subject, etc.):
  - easy to capture
  - applicable to all types of heads

(10) *Дайте* ***избавиться*** *мне <u>от сего несносного бремени</u>: вы знаете теперь все мои несчастия.* [Д. И. Фонвизин. Сидней и Силли, или благодеяние и благодарность (1769)]

‘Let me get rid of this unhappy burden: now you know of all my miseries’.

Results 2: syntactic bondedness

- The differences in the distributions of the examples with intervening material in the samples for the three basic lexemes of the classes:
  - Distances from *избавиться* to *от* measured in the number of intervening groups:

	0		1		2		3	
< 1800	47	0.83	7	0.12	2	0.04	1	0.02
1970-2000	247	0.91	23	0.08	2	0.01	0	0

The difference between the distributions for the two periods is marginally significant, Mann-Whitney U test, $p \approx 0.053$.

Results 2: syntactic bondedness

- Distances from *преимущество* to *перед* measured in the number of intervening groups:

	0		1	
< 1850	49	0.79	13	0.21
2003	69	0.92	6	0.08

The difference between the distributions for the two periods is significant, Mann-Whitney U test, p ≈ 0.03.

- To compare, for *смотреть* with *на* the difference between the periods is not significant (Mann-Whitney U test, p ≈ 0.17).

Results 2: syntactic bondedness

- Dependency relations:

(11) *Графиня Головкина завидовала* ***преимуществам****, которыми* ***пользовалась*** *княгиня Долгорукая* <u>*пред прочими дамами*</u>*...*

[Е. Ф. Комаровский. Записки графа Е.Ф.Комаровского (1830-1835)]

‘Countess Golovkina was envious about the advantages that enjoyed princess Dolgorukaja as compared to other ladies…’

- The syntactic head of *перед* (*пользоваться*) and its “semantic” head (*преимущество*) are in different clauses.
- Dependency relations are particularly revealing for the arguments of nominal heads.

Summary 2: syntactic bondedness

- Changes in syntactic bondedness (the frequency of uses with some material intervening between the head and the dependent) corroborate the generalizations made on the basis of lexical distributions:
  - The prepositional uses that were shown to decrease in productivity show significant changes in syntactic behaviour over time.

Summary 2: syntactic bondedness

- Both in terms of changes in lexical distribution and syntactic behaviour some “prepositional meaning” move from more circumstantial properties to more argumental properties:
  - lexical range centers around a small group of lexemes
  - syntactic connection with the head becomes tighter
  - *перед* + Standard and *от*+Threat moves to become more argumental, *на*+Stimulus is more argumental already in the earliest texts in the RNC and it remains argumental

To make the picture more complex…

- Other “sources” of valency patterns:
  from more to less wide-spread and generalizable:
  - analogy, including the extention of valency patterns to new lexemes (derived or borrowed), cf. [Barðdal 2008]
  - metaphorical extention + lexicalization:

отражаться на+Loc	‘reflect on’	vs.	‘influence’
отражаться в+Loc	‘reflect in’	vs.	‘show itself’

  - calquing

Conclusions: Taking grammaticalization into perspective

- Taking a step back to the point when the preposition acquires a new meaning in the course of grammaticalization.
- This meaning spreads over a number of contexts, where the prepositional phrase in this meaning can be more of less widely used circumstantially.
- That is the stage we observe for *от* encoding Threat and *перед* encoding Standard of comparison.
- In many contexts it can be later replaced by newly grammaticalized means with similar meanings and be lost, cf. the competing means of Standard of comparison encoding: groups with *чем* ‘than’, *над* with *преимущество*, *по сравнению с* ‘in comparison to’…
- Still, in some contexts they are lexicalized and retained, thus resulting in idiomatic valency patterns.
- In what kind of contexts they are retaind and why, is the question for future study.

References

Apresjan Ju. D. Leksicheskaja semantika. Sinonimicheskie sredstva jazyka. Moscow, 1974.

Baayen R. H. Corpus linguistics in morphology: morphological productivity // Luedeling, A., Kyto, M. (eds.). Corpus Linguistics. An international handbook. Mouton De Gruyter, Berlin, 2009. P. 900-919.

Baayen R. H. On frequency, transparency, and productivity// Booij, G. E. , van Marle, J. (eds). Yearbook of Morphology 1992, Kluwer Academic Publishers, Dordrecht, 1993.P. 181-208.

Barðdal J. Productivity: Evidence from Case and Argument Structure in Icelandic. Amsterdam: John Benjamins, 2008.

Beloshapkova V. A. Sovremennyj russkij jazyk. Sintaksis. Moscow, 1977.

Brinton, L. J., Traugott E. C. 2005. Lexicalization and language change. Cambridge: CUP.

Goto K. V., Say S. S. Chastotnye xarakteristiki russkix refleksivnyx glagolov// Kisseleva K. L., Plungjan V. A. (eds.). Korpusnye issledovanija po russkoj grammatike. Moscow, 2009. P. 184-223.

Lehmann Ch. Thoughts on Grammaticalisation. München: Lincom Europa, 1982/1995.

Peshkovskij A. M. Russkij sintaksis v nauchnom osveshchenii. Moscow, 2001.

Vinogradov V. V. Russkij jazyk (Grammaticheskoe uchenie o slove). 8th ed. Moscow, 1947/2001.

Zolotova G. A. Sintaksicheskij slovar. Repertuar elementarnyx edinic russkogo jazyka. Moscow, 2006.

Thank you!