Specifics of applying topic segmentation algorithms to scientific texts

K. Boyarsky, N. Gusarova, N. Dobrenko, E. Kanevsky, N. Avdeeva

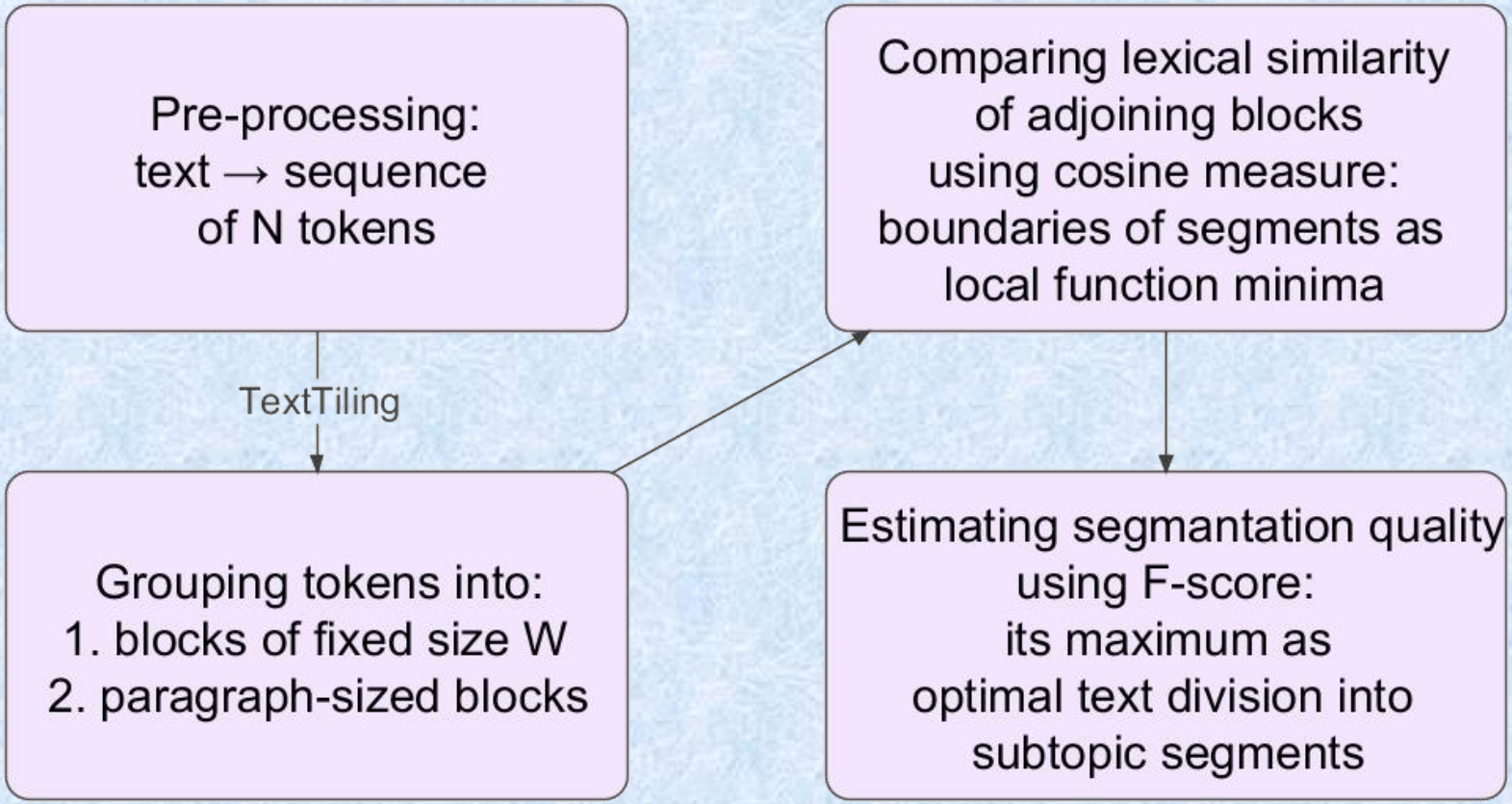
Исследование специфики применения алгоритмов тематической сегментации для научных текстов

К.К. Боярский, Н.Ф. Гусарова, Н.В. Добренко, Е.А. Каневский, Н.А. Авдеева

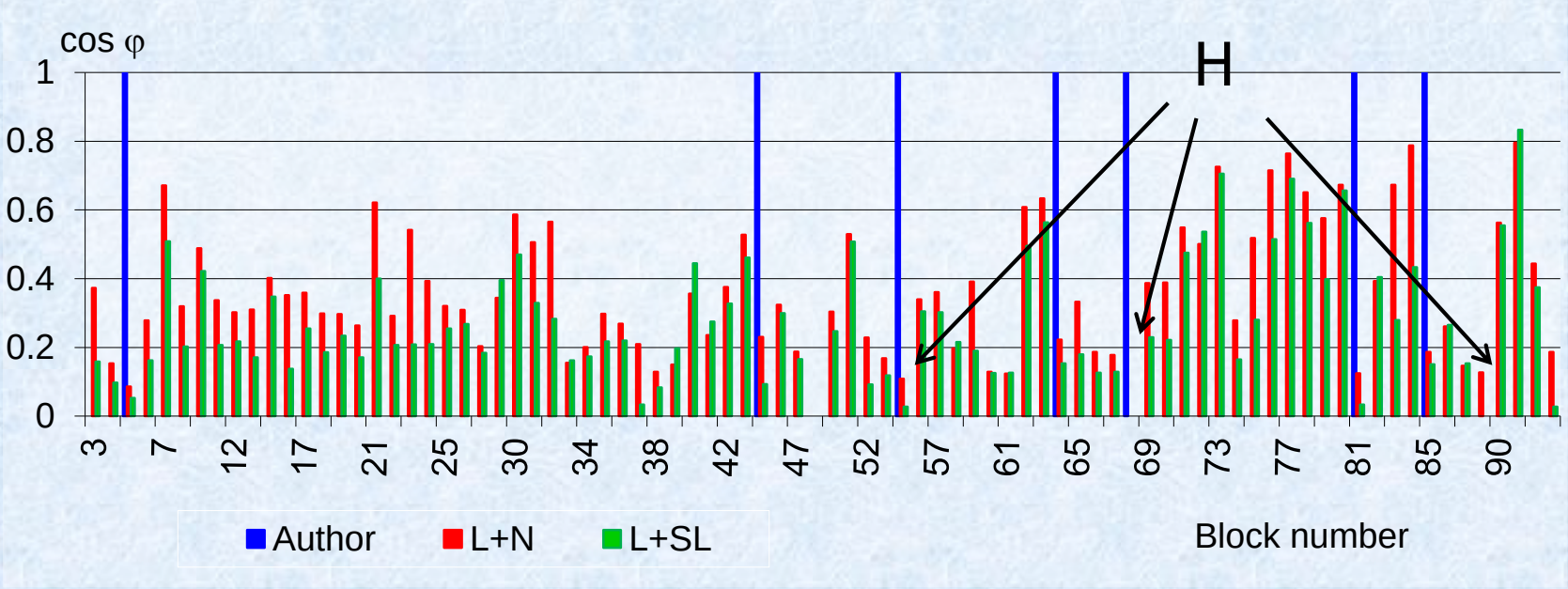
Peculiarities of scientific texts

- **Limited text size:** statistical methods are hardly suitable for them
- **Specific terminology and notation:** frequency distribution of terms does not match the one of general vocabulary
- **Vocabulary of low contrast:** repeated terms throughout the text
- **Thematic unity of the text with gradient junctions to another subtopic:** the boundaries of subtopic shift can conflict with author text segmentation

TextTiling: basic method of segmentation

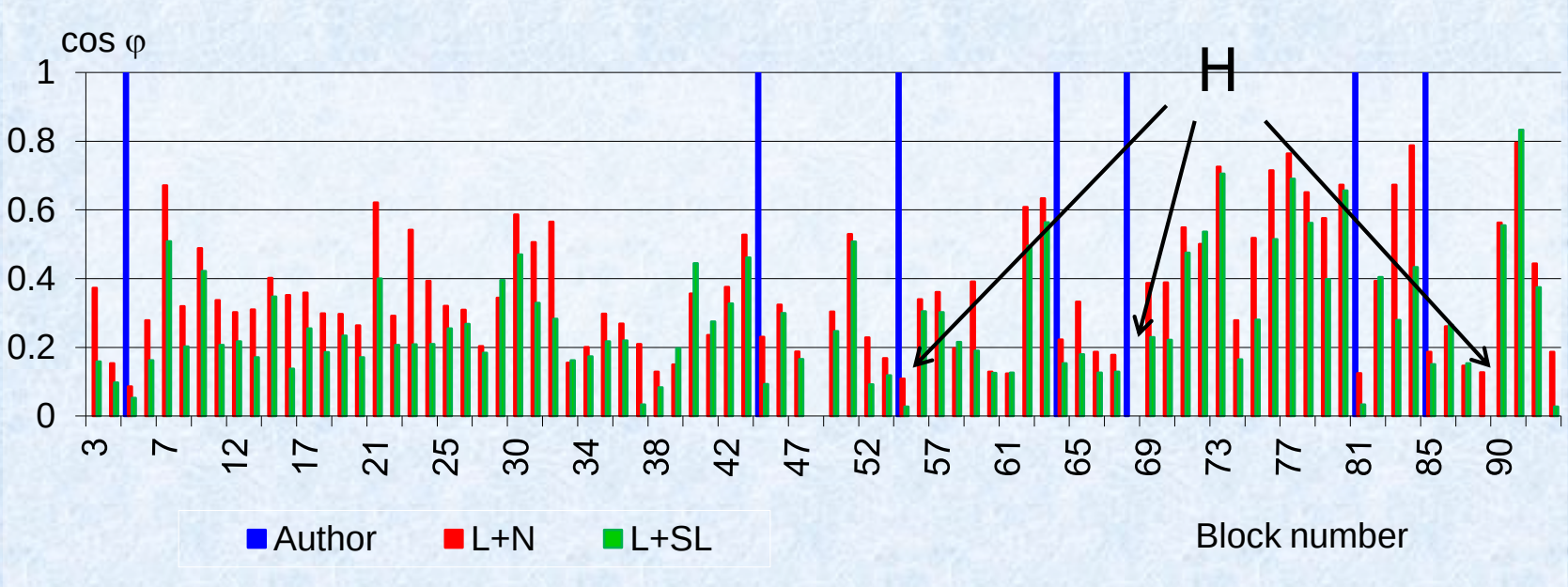


Author and automatic segmentation



$$\cos\varphi_i = \frac{\sum w_{i-1} w_i}{\sqrt{\sum w_{i-1}^2}\sqrt{\sum w_i^2}}$$

Author and automatic segmentation

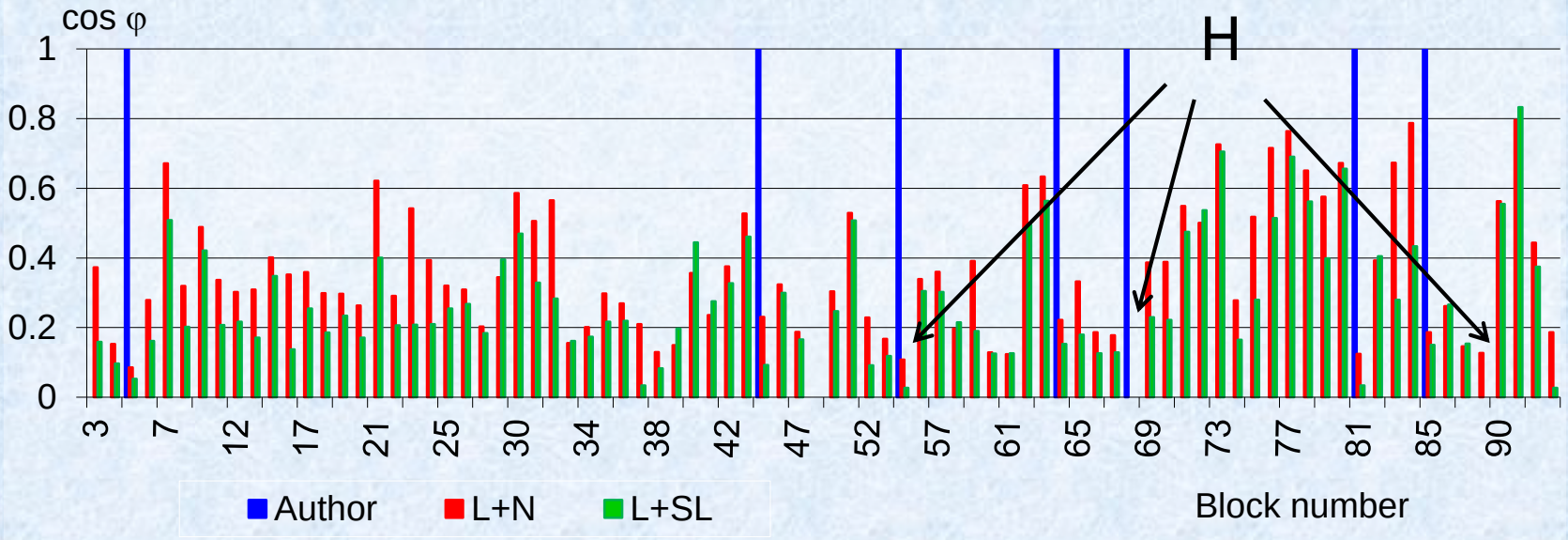


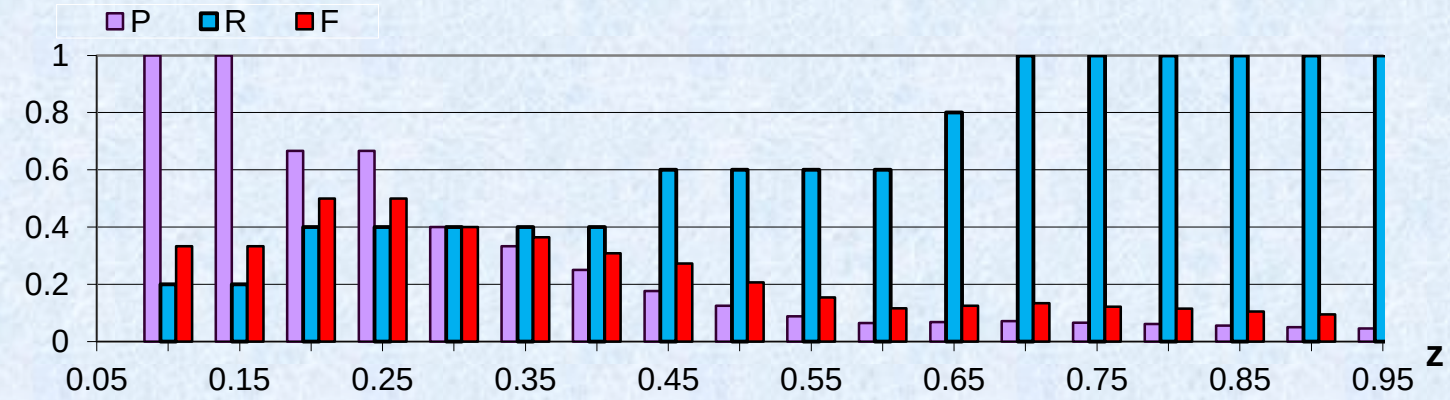
Precision $P = \frac{TP}{H}$

Recall $R = \frac{TP}{TP + FN}$

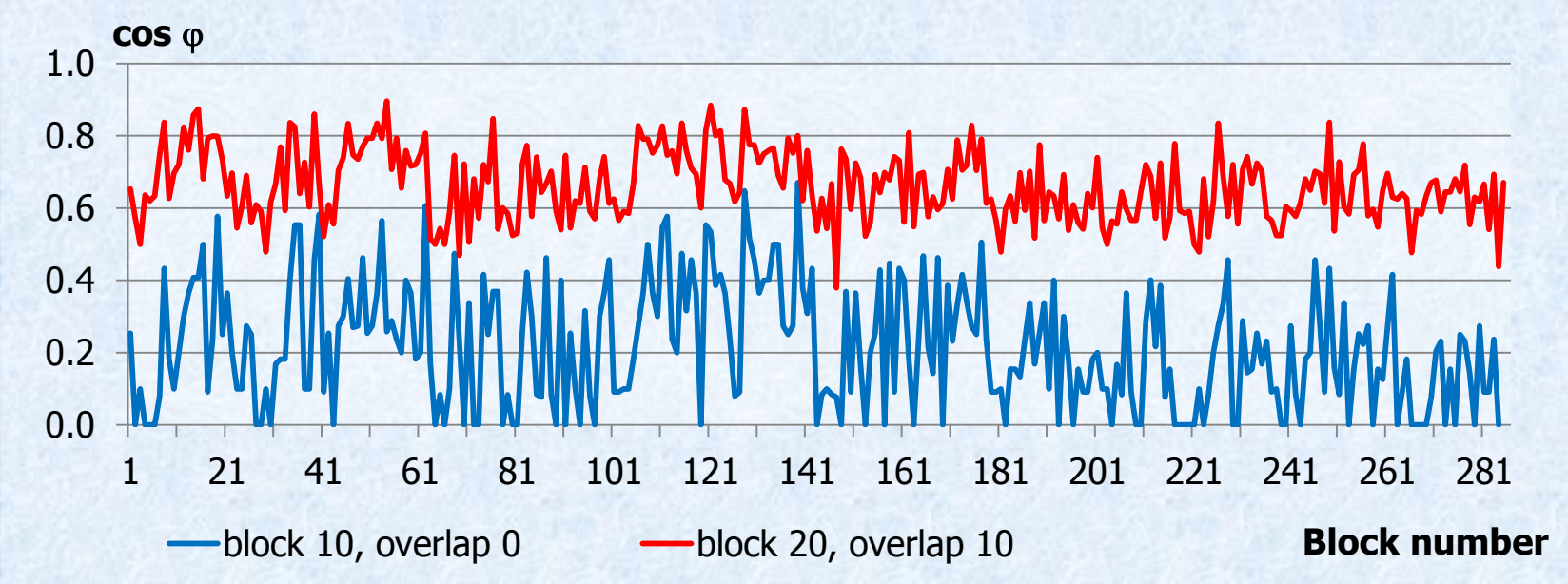
F-score $F = \frac{2 * P * R}{P + R}$

Author and automatic segmentation



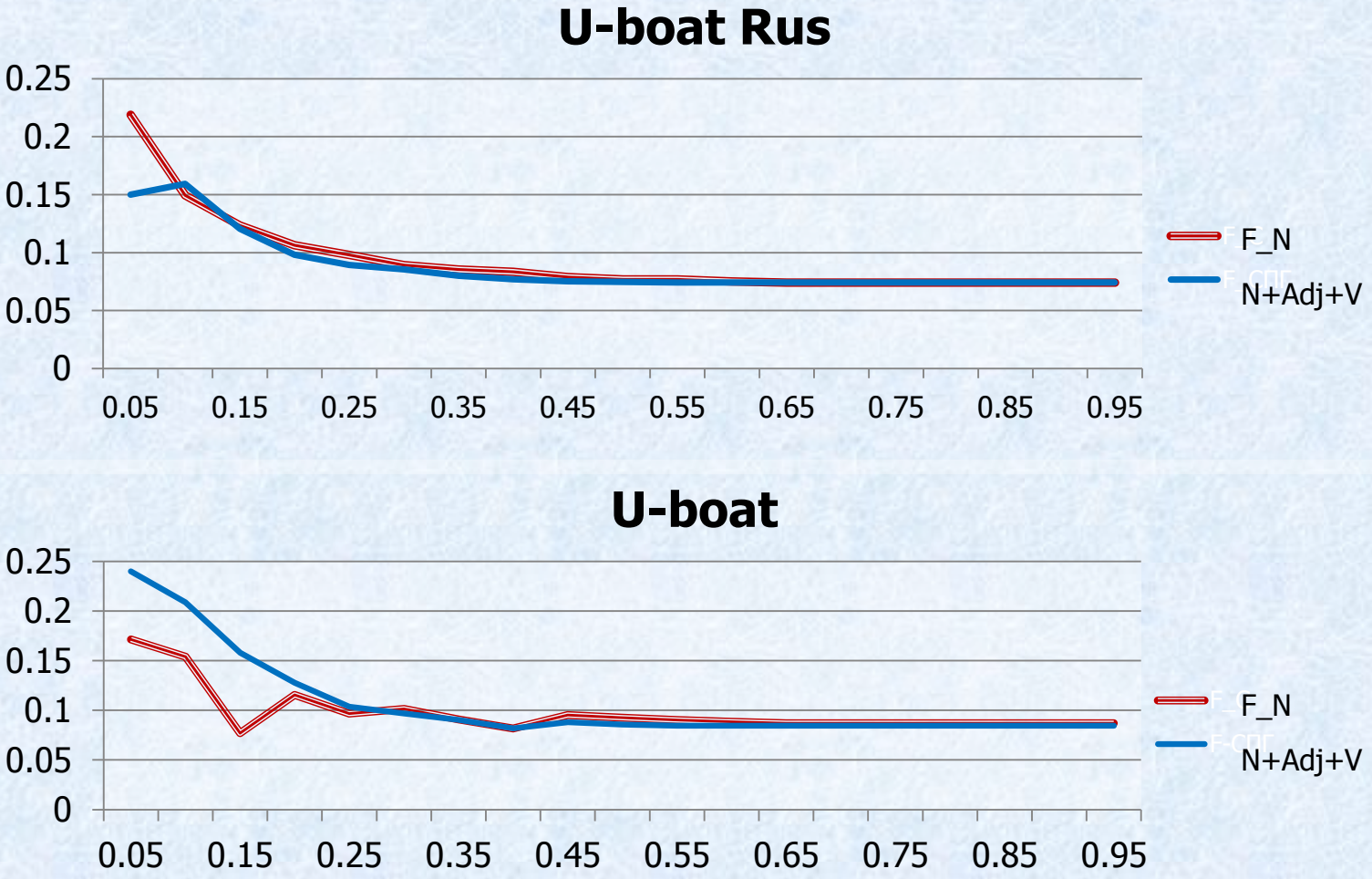


Ways to divide text into segments



Block size	10	25	40	paragraph
F-score	0.06	0.04	0.03	**0.17**

Lemmatization and vocabulary selection



Lemmatization and vocabulary selection

Title of the text	F-score	
	L+N	L+N+Adj+V
U-boat Rus (rus)	**0.21**	0.16
U-boat (eng)	0.17	**0.24**
Romme Rus (rus)	**0.22**	0.21
Romme (fr)	0.44	0.46
News (rus)	0.60	

Construction of vector-based semantic space

Ленточки бескозырок матросов кайзеровского флота имели надпись прописными печатными буквами, вышитыми золотой или серебряной канителью

Ribbons of peakless caps for the Kaiserliche Marine had gilt and silver thread block lettering

Пилотка кроилась из темно-синего плотного сукна, обычно с черной или темно-синей подкладкой из искусственного шелка

The field cap was cut from fine-quality navy blue cloth wool, usually with a black or dark blue cotton or artificial silk lining

$$\cos \varphi = 0$$

Construction of vector-based semantic space

Semantic classifier: 190 000 lexemes => 1700 classes

Ленточки бескозырок матросов кайзеровского флота имели надпись прописными печатными буквами, вышитыми золотой или серебряной канителью

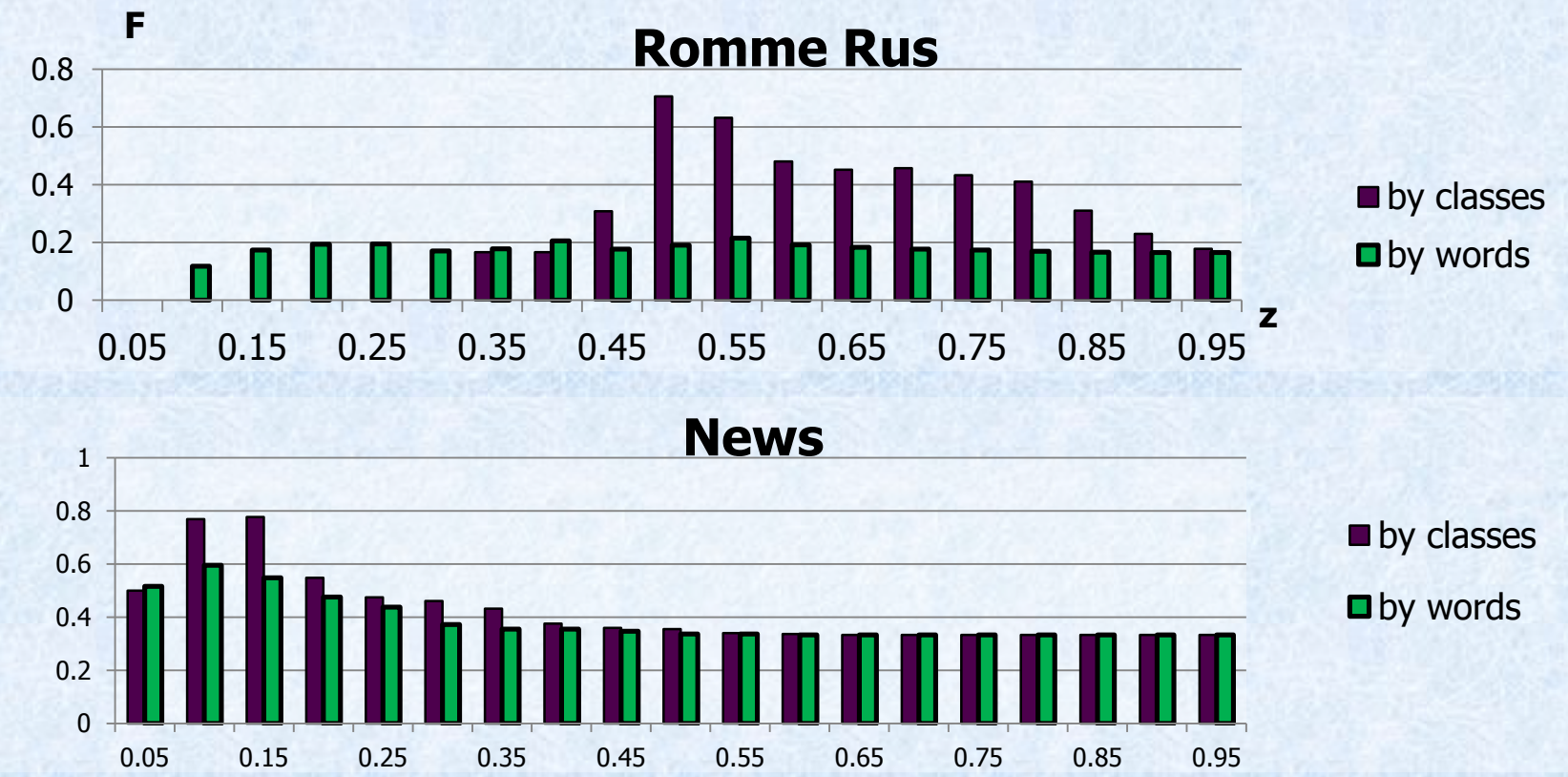
Ribbons of peakless caps for the Kaiserliche Marine had gilt and silver thread block lettering

Пилотка кроилась из темно-синего плотного сукна, обычно с черной или темно-синей подкладкой из искусственного шелка

The field cap was cut from fine-quality navy blue cloth wool, usually with a black or dark blue cotton or artificial silk lining

$\cos \varphi = 0.71$

F-score values of the analysis “by words” and “by classes”



Text	U-boat Rus	Romme Rus	News
By words	0.21	0.22	0.60
By classes	0.50	0.70	0.78

Thank you for your attention!

K. Boyarsky boyarin9@yandex.ru

N. Gusarova natfed@list.ru

N. Dobrenko graziokisa@yandex.ru

ITMO University, Saint-Petersburg

E. Kanevsky kanev@emi.nw.ru

N. Avdeeva assoul@yandex.ru

Saint-Petersburg Institute for Economics and Mathematics, RAS