COMPARISON OF HIGH-FREQUENCY NOUNS FROM THE PERSPECTIVE OF LARGE CORPORA

Maria Khokhlova

St.Petersburg State University

m.khokhlova@spbu.ru

Large Corpora

- Automatic collection of data from the Internet [Kehoe, Renouf 2002; Kilgarriff, Grefenstette 2003; Belikov, Selegey, Sharoff 2012];
- 1 mln tokens → 100 mln tokens → 500 mln tokens;
- New questions:
  - What can we see with big data and how does it affect the results?
  - Do linguists actually need large corpora or their appetites can be satisfied with less data?
  - Can small corpora be viewed as little big ones?

Large Russian Corpora

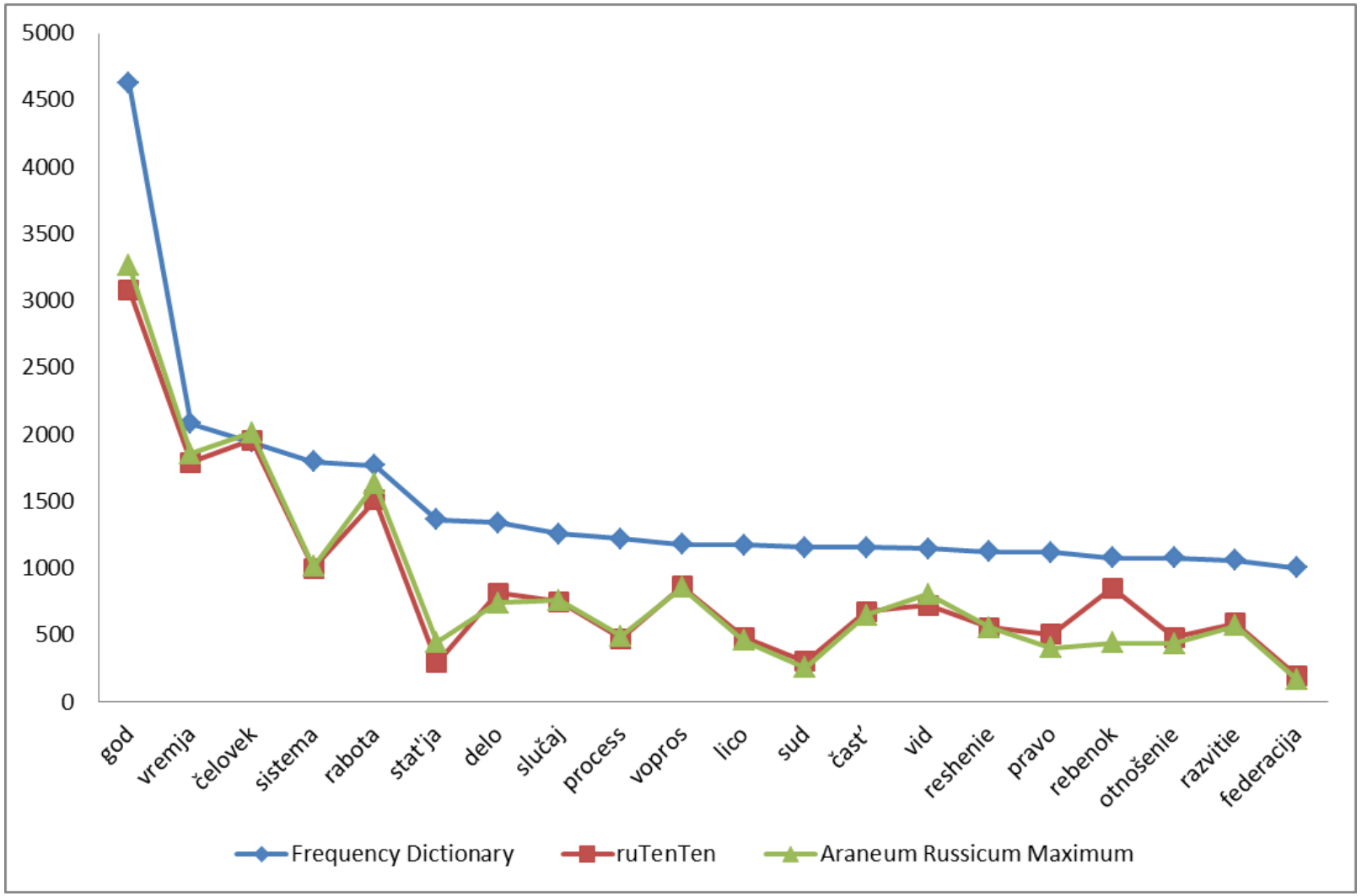
- "Manual" or traditional corpora:
  - RNC;
  - HANCO;
  - …
- "Automatic" corpora:
  - General Internet-Corpus of Russian;
  - Aranea project;
  - Russian Web corpus;
  - ruTenTen.

Data

- ruTenTen (18.28 bln tokens);
- Araneum Russicum Maximum (13.7 bln tokens);
- The majority of Russian texts in web corpora come from news websites, blogs, commercial websites, social media groups etc.
- Frequency Dictionary [Lyashevskaya. Sharoff 2009].
- High-frequency nouns selected from non-fiction texts and social and political journalism.

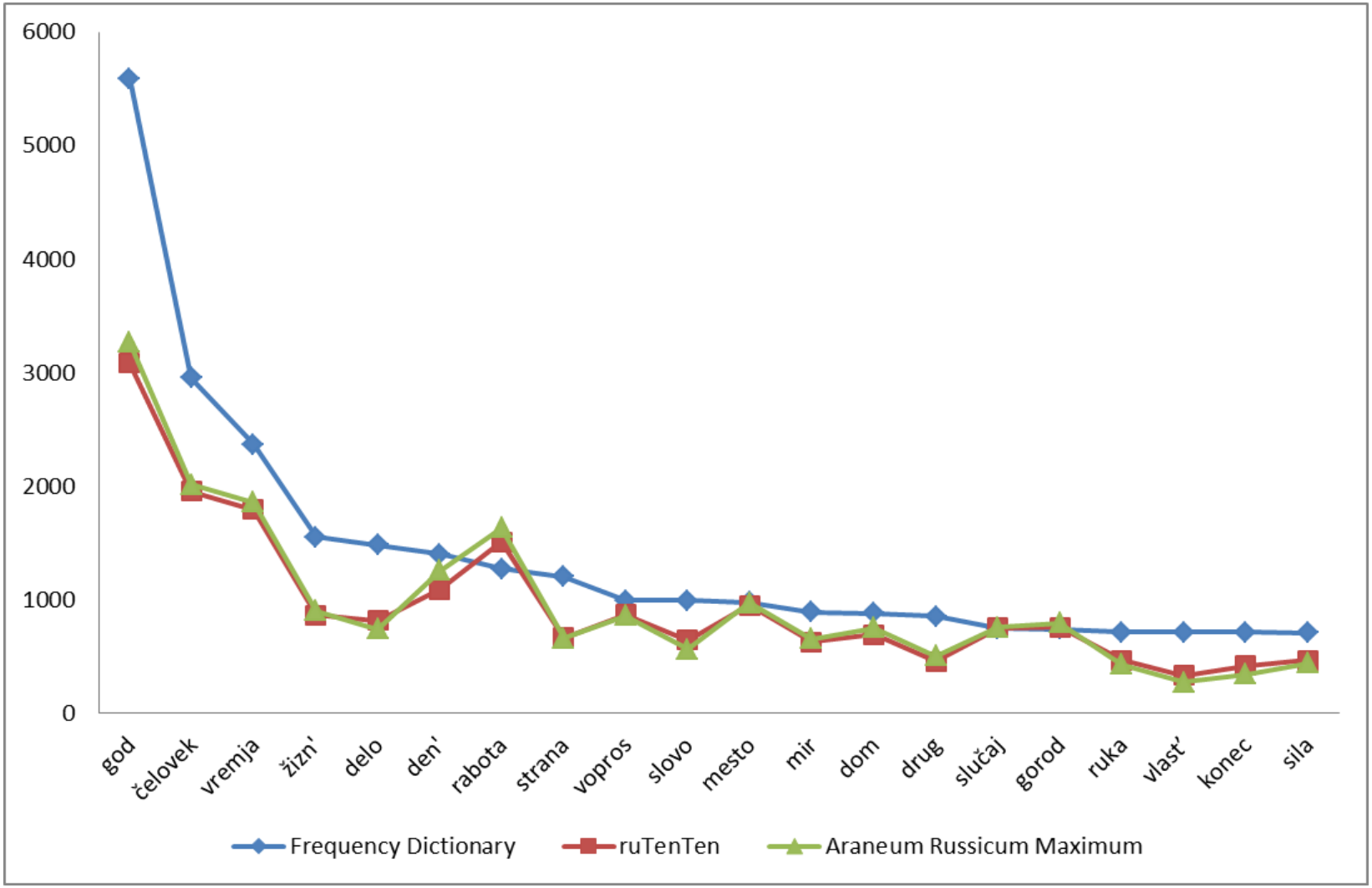
Lemma	Translation	Frequency (ipm)		
		Frequency word list for non-fiction in the Frequency Dictionary	Corpora	
			ruTenTen	Araneum Russicum Maximum
god	year	4624.2	3080.0	3263.0
vremja	time	2080.5	1791.0	1857.0
čelovek	man, person	1945.3	1956.0	2012.0
sistema	system	1798.0	999.0	1011.0
rabota	job, work	1766.4	1510.0	1632.0
stat'ja	article, clause	1363.0	294.1	446.8
delo	affair, business	1339.5	814.0	741.0
slučaj	case	1259.0	752.0	758.0
process	process	1221.8	474.0	491.9
vopros	question	1180.9	866.0	855.0
lico	face, person	1175.9	483.7	458.1
sud	court	1153.9	303.2	255.7
čast'	part	1153.8	677.0	650.3
vid	kind, aspect	1147.9	723.0	806.0
reshenie	decision	1122.3	558.0	556.3
pravo	right	1117.6	507.2	405.1
rebënok	baby, child	1078.4	850.0	443.1
otnošenie	relation	1077.5	481.2	438.4
razvitie	development	1059.6	587.0	570.6
federacija	federation	1003.1	198.4	168.0

Frequency distribution of nouns on the non-fiction word list



Lemma	Translation	Frequency (ipm)		
		Social & political journalism word list in the Frequency Dictionary	Corpora	
			ruTenTen	Araneum Russicum Maximum
god	year	5589.50	3080.0	3263.0
čelovek	man, person	2950.10	1956.0	2012.0
vremja	time	2364.60	1791.0	1857.0
žizn'	life	1548.40	865.0	899.0
delo	affair, business	1482.00	814.0	741.0
den'	day	1397.80	1089.0	1253.0
rabota	job, work	1272.40	1510.0	1632.0
strana	country	1203.90	662.0	657.6
vopros	question	992.00	866.0	855.0
slovo	word	989.70	645.0	563.3
mesto	place	976.10	950.0	970.0
mir	world, peace	887.80	626.0	655.5
dom	house, home	879.70	689.0	751.0
drug	friend	850.90	452.3	500.7
slučaj	case	744.30	752.0	758.0
gorod	city, town	738.50	757.0	792.0
ruka	arm, hand	713.00	466.7	430.5
vlast’	power	711.80	330.0	273.9
konec	end	710.80	417.8	344.4
sila	strength	709.80	467.5	438.2

Frequency Distribution of Nouns on the Social and Political Journalism Word List



Conclusion and Next Work

- Large corpora reflect the language of the web.
- The **Araneum Russicum Maximum** appears to be slightly more consistent with the Frequency Dictionary than the **ruTenTen** corpus in describing high-frequency nouns.
- For the given high-frequency nouns there is a very strong association between the data obtained on two corpora. Hence it can be supposed that there is no difference between the automatically crawled corpora in case of high-frequency lexemes.
- Both corpora show quite a high correspondence with the Frequency dictionary. The data selected from the Frequency dictionary were based on the Russian National Corpus and therefore the obtained results reveal a close correlation between traditional and web-corpora.
- Our next work will be targeted at other parts of speech as nouns can be thematically biased, and their frequencies can depend on types of texts and thus differ dramatically even among corpora compiled within the same methodology.

Thank you!