The First One-Million Corpus for the Belarusian NooJ Module

Ivan Reentovich^{1(K)}, Yuras Hetsevich¹, Valery Voronovich², Evgenia Kachan², Hanna Kozlovskaya², Angelina Tretyak², and Uladzimir Koshchanka³

¹ United Institute of Informatics Problems, Minsk, Belarus mwshrewd@gmail.com, Yury.Hetsevich@gmail.com ² Belarusian State University, Minsk, Belarus gamrat.vvv@gmail.com, evgeniakacan@gmail.com, malavita3000@gmail.com, angelina_tret@mail.ru ³ The Centre for the Belarusian Culture, Language and Literature Research, National Academy of Sciences of Belarus, Minsk, Belarus koshul@gmail.com

Abstract. In this article the first one-million corpus for the Belarusian NooJ module is represented. The given corpus has been built up of texts, patched up into sections by different subject categories. From the broad list of possible subject categories in the sections the corpus focuses on fiction, historic, medical, scientific, sociological literature, etc. Given a great number of similar subject categories, the first one-million corpus can be considered as a first subject collection of texts for the Belarusian NooJ module.

The text corpus is expected to be suitable for research in the following aspects: word polysemy processing of various texts, polysemic punctuation marks processing, and a new lexical items search.

The first one-million corpus for the Belarusian NooJ module can be fully applicable in many fields of linguistic research.

1 Introduction

The chosen subject "First One-million Corpus for the Belarusian NooJ module" is among the most important, integral parts of future research in the field of speech recognition and synthesis. It is really a great step on the way to new investigations of the Belarusian language in the world of NooJ and linguistics.

The purpose of this paper is to introduce the elaboration, creation stages as well as stages of "deep" analysis and practical application of the first one-million corpus for the Belarusian NooJ module in the context of different aspects and approaches.

Besides, the first one-million Belarusian corpus for the Belarusian NooJ module will be applicable for solving the following fundamental tasks: optimizing and expanding the development of high-quality linguistic algorithms for the electronic text preprocessing in the TTS (Text-to-Speech) system.

Two Belarusian corpora were developed for NooJ [1] – the 1-VERSION corpus (1 million corpus.noc) and the MAIN corpus (First 1MLN Corpus for the Belarusian NooJ Module.noc). To make the process of corpus creation more productive, a special (descriptive) algorithm has been worked out.

2 Descriptive Algorithm for the First One-Million Corpus for the Belarusian NooJ Module

The main work on corpus compilation and analysis with the help of this algorithm was fulfilled on the basis of 1-VERSION corpus (Table 1).

According to this algorithm, the 1-VERSION corpus has been built-up of 338 unarranged text units, the MAIN –of 1 570 text units, patched up into sections of different subject categories (As is seen from A Appendix (Fig. 15)). From the broad list of possible subject categories in the sections, the MAIN corpus focuses on fiction, historical, medical, scientific, sociological literature.

3 The Dictionary of Naturalized Lexical- and Grammatical Information for the Whole List of Unknown Unique Words (File *UNKNOWNS.dic*)

3.1 "Purity" Check of the Corpus '1 million corpus.noc'

To get better results on the task of the Dictionary of Naturalized Lexical- and Grammatical Information creation, it is necessary to realize the more extended "purity" check of the corpus '1 million corpus.noc'. From this corpus, with the help of Levenshtein algorithm [2], the search of wordforms with the high (0.8) level of similarity of one wordform to another was realized. As a result, comparing the created dictionaries of known (about 150 000) and unknown (about 50 000) wordforms, the authors have found almost 30 % of similar wordforms, which, as a matter of fact, must belong to known wordforms, despite the fact that the NooJ program has recognized them as unknown.

Below, the general problem points are given in terms of "purity" check of the whole text corpus, which may rather effectively be solved using the abovementioned Levensh-tein algorithm:

- 1. the occurence of Latin letters in many words in the texts of the Belarusian corpus;
- 2. dialectal words of the Belarusian language;
- 3. Russian words;
- 4. orthographic mistakes;
- 5. different letter case processing.

N₂	Action	Result
1	Text material collecting and subject specification for prospective corpora	The draft corpus for the Belarusian NooJ module is made. (The name is 1 million corpus – Very big[.noc]). Chosen subjects: without the subject specification. (TEXT TOTAL: <u>338</u>)
2	The "purity" check of the developed corpus:	 There are NO ANY <u>unknown</u> or <u>unidentified symbols</u> in the given corpus (all texts must be encoded to UTF-8).
	 Unknown or unidentified symbols; [In addition]: words with apostrophe ('); [In addition]: words with hyphen/ (-) symbol. 3 [All possible word occurrences with apostrophe (') and hyphen (-) must be searched in texts of the corpus because the NooJ program is unable (in the process of Linguistic Analysis) to parse them correctly (in a context such words are incorrectly lexically divided): e.g. 1) nad' exaups (incorrect); 2) uëmna- cini (incorrect). (In both cases the words are divided by the space and NooJ can't identify them as one unit). 	 2. <u>5818</u> wordforms with apostrophe (*) are found in the given corpus. (The special NooJ-grammar SearchWordFormsWithApostrophe(FIRST VERSION).nog has been applied in this case) 3. In the given corpus we detected: <u>25 615 occurrences of all wordforms with hyphen/(-) symbol</u> (The special NooJ-grammar SearchWordFormsWithHyphen.fst has been applied in this case); <u>19 728 occurrences of all unique (1 occurrence per match) wordforms with hyphen/(-) symbol</u> (The special NooJ-grammar SearchWordFormsWithHyphen.fst has been applied in this case); <u>23 663 occurrences of all wordforms with hyphen/(-) symbol</u> (The special NooJ-grammar SearchWordFormsWithHyphen.fst has been applied in this case); <u>23 663 occurrences of all wordforms with hyphen/(-) symbol</u> using special <a b="" check<="" href="https://www.www.www.www.www.www.www.www.www.w</th></tr><tr><th></th><th>E.g. 1) пад'ехаць (correct);
2) цёмна-сіні (correct)]</th><th>necessary NooJ resources general_be.nod and
SearchWordFormsWithHyphen.nog).</th></tr><tr><th>3</th><td>The wordforms total is counted
in the developed corpus</td><td>Wordforms (ALL)
(WF> = 1 884 971)
Wordforms (ALL, unique [1 occurrence per match])</td></tr><tr><th>4</th><td>The unknown words search is realized from the developed corpus</td><td>(<WF> = 197 712)
Unknown words (ALL)
(<UNK> = 140 235)
Unknown words (ALL, unique [1 occurrence per match])
(<UNK> = 50 186)</td></tr><tr><th>5</th><th>Creation of the <u>dictionary of</u>
<u>unknown</u> (<UNK>) unique
word usage [*.dic]</th><th>File UNKNOWNS.dic
(<u>49 749</u> unknown words (after the <u>necessary</u> correction of the
unique <math>\langle UNK \rangle = 50 \ 186</math> result))</th></tr><tr><th>6</th><td>Creation of the <u>dictionary of</u>
<u>naturalized lexical- and</u>
<u>grammatical information</u> for the
whole list of unknown unique
words (on the basis of the file
<u>UNKNOWNS.dic</u>)</td><td>[The more extended " purity"=""> of the developed corpus. (The "disposal" of Roman alphabet letters and other problematic cases <u>in texts of the corpus</u>)]

 Table 1. Descriptive algorithm of the first one-million corpus for the Belarusian NooJ module

7	The known words search is realized from the developed corpus	Known words (ALL) (< DIC > = 1 750 447) Known words (ALL, unique [1 occurrence per match]) (< DIC > = 148682)
8	Realization of the developed corpus time consuming operation of lexical information selection	The whole dictionary is created on the basis of the given corpus: <i>1 mil c_BE.dic (197 712 unique occurrences)</i>
9	The punctuation marks search is realized from the developed	Punctuation marks (ALL) $(\langle \mathbf{P} \rangle = 605 \ 512)$
	corpus	Punctuation marks (ALL, unique [1 occurrence per match]) ($\langle \mathbf{P} \rangle = 61$)
10	The digrams search check is	Digrams
10	realized from the developed	<u>Concordance</u> (Digrams = 1 787 305 [occurrences])
	corpus	<u>Dictionary</u> (1 mil cor VB_(digrams).dic)
		Parts of Speech
		(< NOUN > = 523 193) ->ALL (< NOUN > = 58 376) -> unique [1 occurrence per match]
11	The Part of Speech tagging of	(<verb></verb> = 326 739) -> ALL (<verb></verb> = 45 682) -> unique [1 occurrence per match]
11	words is realized within the	$(\langle ADJECTIVE \rangle = 194\ 682) \rightarrow ALL$
	developed corpus	(ADJECTIVE > = 38 218) -> unique [1 occurrence per match]
		(ADVERB> = 172935) -> ALL (ADVERB> = 3 869) -> unique [1 occurrence per match]

These issues are also solved in the NooJ program, though it takes far more time and effort, because the program has to process a large scope of information.

3.2 Statistical Analysis of the Text Corpus '<u>1 million corpus.noc</u>'

The following steps have been taken at this stage:

- 1. The corpus '*1 million corpus.noc*' Linguistic Analysis (As is seen from A Appendix (Fig. 16)).
- The search of wordforms (all wordforms, which are present in the corpus) using special queries (<WF>, <UNK>, <DIC>, <NOUN>, <VERB>, <ADJEC-TIVE>, <ADVERB>).
- 3. Export of the matches into text files (*.txt).
- 4. Text files with exported data are stored in a special database, where the unique wordform clustering was realized and the number of wordforms was counted.

3.3 Machine-Learning Algorithms Application for the Part-of-Speech Tagging [3] of Unknown Wordforms

1. The <u>main attribute of a wordform</u> for the Part-of-speech tagging process with the machine-learning algorithms, specified for the purpose, was *three ending letters* of each wordform in the dictionary of unknown wordforms. (As is seen from Fig. 1.)

Row ID	\$ Col0	\$ Col1	\$ "002	\$ Col3	-i Col4	1 008
Raw33	Aleksievich_CarnobylskajaMelitxa_ALL	142 - KERR 0,5 HOM FORTEL	007676700X	выводрана 254 тыс. гактараў зания. Баларусь	1918	
Row34	Aleksievich_CamobylskajoMolitva_ALL	ном. гостараў. 3 сольтася		. гастараў заклі. Беларусь — хража лісоў	1946	0
Raw35	Alleksievich_CarnobylskajaMalitva_ALL	BERGED JYTE? Y FORMER SHE	(puttle	Днятр. Сок адносяцар да зоны	2044	0
Raw35	Aleksievich_CamobylskajoMolitxa_ALL	лугој у пойнах рис Прыги	(Dwirp	Саж односяция до зоны рединовти/него	2053	0
Raw37	Aleksievich_CarnobylskajeMelitxa_ALL	у гойнах эн Прылязь, Дн.	Саж	адносяцая да заны радыевстыўнага забрудказоння	2060	
Raw38	Aleksievich_CarnobylskajaMaltxa_ALL	рачавые захвореени, рез	acticianase	ресстройствы і ленетыльная мутацыі' З	2275	
Raw39	Aleksievich_CarnobylskajaMaltxa_ALL	потацыя 36. Чарнобыл	Ct	7, 24, 49, 101, 149. 'Ta gagewark stolpesser), 29 xpeczelea 19	2374	8
Raw-90	Abeloievich_CarnobylskajaMaltxa_ALL	аврагістраваны ў Польциц	Румані	, 30 еросанія — у Шанії, прогинії Паўночнай	2531	0
Raw41	Alleksievich_CarnobylskajaMalitxa_ALL	Геривні, Аўстрыі, Руньні	LIDONALEPH	i Fisikovasa i Iranii, 12 kan y	2547	
Raw42	Aleksievich_CamobylskajeMalova_ALL	фесовіка — у Швей, роси	hani	, 1—2 ная — у Францыя, Бельпі, Нідэргандах	2569	B
Raw-13	Alleksievich_CarnobylskajeMelitxa_ALL	ная - у Срендый, Бельпіц	Важиры.	, паўчачнай Гродыі, 3 мат — у Ізроілі	262.0	9
Row44	Alieksievich_CamobylskajaMalitxa_ALL	Влякабритані, паўночнай.	boaini	, Кувейце, Туркиц., Закінутыя на вялікую	2664	0
Raw45	Aleksievich_CamobylskajeMeltoa_ALL	паўначнай Грэцыі, 3 мел	6,803,0	. Туркыі. Зекінутыя на яклікро вышынно	2673	
Raw46	Aleksievich_CamobylskajaMalitxa_ALL	Грэцыі, 3 мея — у браїні,	Турцыі	Закінутыя на вялікую вышыню газальдобныя	2662	
Row47	Alleksievich_CamobylskajaMelitxa_ALL	зерогістравены ў Японі, 4	KTRI I	, 5-га — у Індыі, 516 мая	2824	
Raw48	Aleksievich_CarnobylskajaMaltxa_ALL	Y 14344, 516 Har - Y	ALLE	і Канадра. Мона за тырхны	2861	
Raw49	Alleksievich_CamobylskajaMaltxa_ALL	16 ная — у ЗДА I	694509	. Менци за тъдзенъ статроблася, каб	2867	
Raw50	Alleksievich_CarnobylskajeMalitxa_ALL	Наступствы Чарнобыльска.	MHDK.	. Маккародны вышойшы Сакареўскі каледн: па	3066	
Raw51	Alleksievich_CamobylskajaMalibia_ALL	Макерадны выклайшы Со	PARK3	. 1992 г. С. 82. Чезебрты рожтер, нечусны	3054	
Raw52	Aleksievich_CamobylskajeMalitxa_ALL	вышайшы Сахараўскі кале	1	. С. 82. Чацекрпи ровстар, някуены аб	3074	B
Raw53	Aleksievich_CamobylskajoMolitxa_ALL	Секерейскі каледжі пе ра	c	. 82. "Чецэёрты рэсктер, некуюты еб'ектан	3077	
Raw54	Alleksievich_CarnobylskajeMolitiva_ALL	радыкихарті, 1992 г. С	нануены	аб'яктан 'Сковкича', па-раняйшану	3166	
Raw55	Aleksievich_CamobelskajeMeltoa_ALL	С. 82. Чазнірты рикта	er784	'Ozerane', ne-paveiuswy steparae (3117	0
Raw 56	Alleksievich CarnobelskajaMalitva ALL	нинусны аб'ектан "Сосель.	STRIERIN	зберегае ў сваін скінцава-жалозабеганным	3138	0

Fig. 1. The excerpt of the NLP system database with data from NooJ

- 2. The following algorithms are being applied:
 - Decision Tree;
 - Clustering;
 - Neural Network.
- 3. Firstly, the dictionary [5, 6] of known wordforms was downloaded into the system¹. This dictionary was meant to "train" all possible word paradigms by the abovenamed algorithms. In other words, the algorithms' *learning* is realized, and 30 % of known wordforms were taken for *its* realization. (As is seen from Fig. 2.)

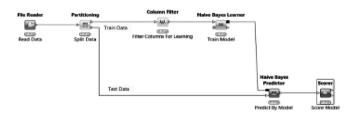


Fig. 2. Train model

Then, 70 % of checked remaining data were realized by the derived model of machine-learning algorithms to verify the proficiency of the model estimated as rather high.

4. After that, the existing dictionary of unknown wordforms (UNKNOWNS.dic) was "passed" through the machine-learning model. The results produced rather high degree of correct assignment of unknown wordforms to one or another part of speech.

¹ The Part-of-Speech Tagging process can be realized not only in one particular NLP system but also in many others (including integrated interactive systems), where these three algorithms, mentioned above, can be applied.

And that, even at the elementary level (here, the main wordform attribute for the realization of Part-of-Speech Tagging process, i.e. *three ending letters* of each wordform, is meant), confirms the effectiveness of the given machine-learning model. (As is seen from Figs. 3, 4 and 5.)

Row ID	S Col0	S Coll
Row1750	адасечко	400
Row1751	адасобліваўся	ýca
Row1752	адась	ась
Row1753	адасю	0036
Row1754	адася	aca
Row1755	адасё	acë
Row1756	адасік	Cik.
Row1757	адасічак	4684
Row1758	адациістрацыі	цыі
Row1759	адаў	даў
Row1760	адаўнаму	arry
Row1761	адб	адб
Row1762	адбірае	pae
Row1763	agfanius	iць
Row1764	адбараніў	niÿ
Row1765	адбахаў	xaý

Fig. 3. Unknown data



Fig. 4. POS prediction

Row ID	& Seq	-S Ending	S PartOfSpeech
Row1750	адасечко	MKD	OTHER
Row1751	адасобліваўся	ýce	VER8
Row1752	адась	acı.	OTHER
Row1753	адасю	800	OTHER
Row1754	адася	aca	VERB
Row1755	адасё	acë	OTHER
Row1756	agacix	cix	OTHER
Row1757	адасінак	MBK.	OTHER
Row1758	адациістрацыі	upati	OTHER
Row1759	адаў	489	OTHER
Row1760	адаўнаму	any	OTHER
Row1761	адб	8,05	OTHER
Row1762	ag6ipae	pae	VERB
Row1763	адбаліць	iųs	VERB
Row1764	адбараніў	miÿ	VER8
Row1765	адбахаў	xarÿ	VERB

Fig. 5. Predicted POS results

5. There are possible variants of data check results by using the aforementioned model (through the example of VERB; other parts of speech are considered as OTHER) (Table 2).

Variants	Actual	Model prediction
True positive	VERB	VERB
True negative	OTHER	OTHER
False positive	OTHER	VERB
False negative	VERB	OTHER

Table 2. Possible variants of data check results

4 Part-of-Speech Tagging Countercheck

The Part-of-Speech Tagging Countercheck on unknown words was realized with the help of Levenshtein algorithm (on basis of the file UNKNOWNS.dic).

One more task was to work out the dictionary of unknown words usage. The assignment for developers is the maximum reduction of the dictionary sizes and determination of unknown words values for their further correction and inclusion in the dictionary of the one-million Belarusian corpus. (As is seen from Fig. 6.)

	*		4		4				
1								899	
•	۰	Seg Toe, uite Organismuse UNKNOWN	Shirt has a real, serve requiring a real process	Similarity.	PartOfSpeech	A 64	wv	Ounit	Communit
		inavail	and and a second	0.000000000	ADVERB				
**	-	TROUBLESY INTERVICE	Industria and spin	0.067142867	ADJECTIVE	4			
	12	I KOMBION DUBHSHI	Industrial and mate		ADJECTIVE				
-	- 10	FOREIDUSY SCREEKER	manufacture and	0.057142857	ADJECTIVE	- 4			
	- 14	inguestap	100121-00	0.0000000000	NOUN				
•	- 18	I HCTIMTY TB	arrary18	0.0000000000	NOUN	4			
	+0	INCOMPTING AND A DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION OF A DESCRI	and the second s	0.0	NOUN	4			
÷.,	17	INCLUTIVITY	Bertwity by			1			
		ingstonry.	PROPERTY.	0.0000000000	NOUN				

Fig. 6. Unknown words in the summary table of the Part-of-Speech Tagging Countercheck

The main feature of the algorithm applied to the Belarusian one-million corpus is that the algorithm does not change the words in texts after editing, but makes it possible for users to see comments on various mistakes made in the texts incorporated in the Belarusian one-million corpus.

The words included in the dictionary were classified in groups of the unknown for various reasons:

- The words written in the Latin alphabet or having some Latin letters (a bavyazkov, atrgml_vayutsets, an akhoplepa);
- The words written by a tarashkevitsa, i.e. substandard spelling which, however, is used by a rather large number of people, especially the Internet users. The existence of alternative spelling is caused by the historical reasons (абараназдольнасьці, абвешчаньня, абвясьцілі);
- Words with spelling errors (абслўгоўванню, абяцаюдь);

- Words with recognition errors after scanning (**тагілёўскага**, **vстагоддзем**);
- Words of foreign languages (perfekt, deutsche, eine);
- Proper nouns (Дзятлава, Анатолія), etc.

The main objective at the stage of unknown words recognition is the definition of their morphological characteristics, i.e. assignment of parts of speech value to 49 749 wordforms. The algorithm of Levenshtein revealed parts of speech of unknown words, picked up a possible correct form of the usage, and also gave an index of probability of correct forms. (As is seen from Figs. 7 and 8.)

	*		0				0	10	
								-0009	
	٥	Seg Tee, MTe s'aganeuus UNKNOWN	Oblition in tee, same reconfigur@ are specified	Dethetty	PartOf8peech	368 M	w	Chesh	Common
6		ino-oni	t served.	0.003033333	ADVERB				
	8.9	індывідуальную	I make they are server	0.007140807	ADJECTIVE				
	12	індывідуальных	Frequent Fight Strike Heads	0.057140507	ADJECTIVE				
12	40	індывіду вланыя	Logical Lings and Logical Control	0.057142067	ADJECTIVE	•			
••	++	наннатар	Lines Landy	0,000000000	NOUN				
	90	інстытута	Learner years	0.00000000000					
	10	інстытутам	Leading Type and		NOUN	1			
£2	47	інстытуту	Lastratymy	0.000000000	NOUN				
-	+0	ingeatorry	I may act to by	0.000000000					
	10	ісленьеў	have as all	0.070	NOUN	0		1 1	
	10	ICH083HH0	100310000	0.000000000		1			
21	24	Hele im	brinin a	0.0030303030	FOREIGN				

Fig. 7. Linguists' checkout of wordforms recognized by Levenshtein algorithm in the summary table of the Part-of-Speech Tagging countercheck

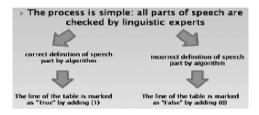


Fig. 8. The process of Parts-of-Speech tagging countercheck

The stage of manual editing is carried out after computer-assisted Part-of-Speech definition, i.e. the algorithm can correctly reveal all parts of speech on formal grounds. The algorithm is simple: all parts of speech are checked by linguistic experts. In case of the correct Part-of-Speech definition by the algorithm this line of the table is marked as "truly" (1). In an opposite case – "lie} (0). (As is seen from Figs. 9 and 10.)

									1
•								199	
		Seq roe, uno s'rrinneulla UNKNOWN	Shotbarg ton, sito manifest) nor spic no	Sudarity	PartOfSpeech		~~~	0 mail	terment
6		inna-sa il	inseat	0.4003030000	ADVERE				
	- 14	Instanti Cay Britsmyrth	ingenting and the	0,007140007	ADJECTIVE	1			
6	10	International Copy Statements	Index Page articles (0.007140807	ADJECTIVE				
	10	ingenities and	Index Page 2014 (1997)	0.007140907	ADJECTIVE				
-	54	ingeneration	transmap.	0.0000000000	NOUN				
÷	145	інстытута	more water	0.0000000000				1	
	140	інстытутам	between the second s	0.8	NOUN	1		1	
	107	increativity'	betrarying .	0,00000000		1		1	
	100	iestaugo ery	Inspecto wity	0.00000000	NOUN	1		1	
	100	IC/RHANY	Los ann all	0,075	NOUN			1 1	
-	100	IC HD BOHHRD	to protect	0.000000000	NOUN			1	
		ide in	a fa inc	0.000000000					
	- 200	maninti@cxana	ana nina filosa na	0,000000000					NUMBER OF STREET
-		oh	uk.		FOREIGN				
		and should be			ECONSIDER AND				

Fig. 9. Editing the results of linguist's checkout of wordforms recognized by Levenshtein algorithm in the summary table of the Part-of-Speech Tagging countercheck

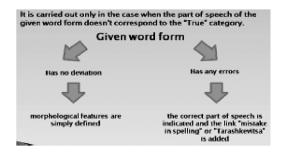


Fig. 10. The process of editing

If the part of speech of a specified wordform doesn't correspond to the validity, the editing stage comes, namely indications of the correct part of speech. If the corresponding wordform has no deviations (without spelling errors, unclear symbols, and also not foreign-language words), in this case morphological features are simply defined. If a word has a wrong spelling, the correct part of speech is indicated and the link "mistake in spelling" is specified. The same happens to the words written in a tarashkevitsa only with another link: "Tarashkevitsa".

The parts of speech noted by the "NULL" category are mainly proper names and therefore are defined by the algorithm described above: indication of the correct part of speech and assignment to this line of "true" value. (As is seen from Fig. 11.)



Fig. 11. Addressing to the context

In case the word meaning is not clear or causes doubts, it is necessary to address to a context, namely to the corpus.

At the end of this stage the quantity of unknown words was decreased that allowed to pass to the following stages of the first Belarusian one-million corpus improvements. (As is seen from Fig. 12.)

	1							
C		1						
8		-						
8			NORM	alorman		allopeniates		Allquerante
н.	-managemi		adjusts.	alorment		allepartation	4.	Annuelle
6	concert?		MO181	alignment:		ellipper-activ		diam'r.
	distant a		FOREIGN	all and an and the		of a particular		diam'r.
۲.	- alives		PORECA	al companie		off open tights	4	diam'r.
	- since		FORECe	+1		ellippential@	1.	diam'r.
н.	Mmr		reases	Contractor		ell'epochiette		diamont
۰	t-end		ACVER6	+Comments		-figerials-		diamonth
	Transmitte community		AGR/714	+Contraction		- Experiment	1.	digent and the
ŵ.	THE OWNER DRIVEN N		ADD/CTM8	+Contractor		+Experiate-	1)	digences.
101	Indiana and and a		ADDITION	-torestella		-tippet and	1.	digutation
۰.	INCOME INC.		NOUN	-0000-000		- Copertains	15	digent and
ŧ.	includy(s)		MORE	-tormetti		-dispersation	1)	deennoo
÷	instanty inst		adam.	alorment		disperiate	4.	diperiori
а.	isensey y		NOUN	alarman		all questions		diqueron the
6	ing any		ACCERC.	all and a set of the		of the party states		digariant.
×.	(manual)		BOOM .	al companie		of openants		dipertant.
8	(management		BOUR .	al company to		ellopertially		diam'r.
н.	winter .		FORECH	elone-		ellippential#*	1	diam'r.
	metellener		HORITME	(Company)	1010010-001017	ell'epochiette	1	diamone
	+		POREA	+100000-0000		-Breatlant		diamonth
а.	period .		PERSON	+Contraction		- Separate -	1.	digest state
	101103108		NORM	+tormeth	1000103-00011	-Experiment	1	digenters.
ж.	with the set		NORM	-domments	TORSE INT. SPIRTS	-dispertation	1.	digenone.

Fig. 12. The Concatenation-in-Paradigm results

According to the resulting data, the special Concatenation-in-Paradigm list was made after the countercheck of recognized by the Levenshtein algorithm unknown words (previously exported from the NooJ-dictionary file UNKNOWNS.dic)) in order to create the additional NooJ general_be.nod dictionary. (As is seen from Fig. 13.)

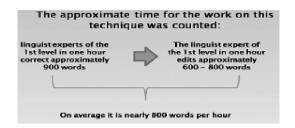


Fig. 13. The approximate time count for the work on the technique

This calculation allows to predict the work on this technique in the future and to estimate degree of overall performance in comparison with other techniques.

5 Comparison of Lexical and Grammatical Base of the Belarusian N-Corpus [6] with Dictionary Properties' Definition File of the Belarusian NooJ Module

In a similar manner a comparison of lexical and grammatical base of a Belarusian N-corpus with dictionary properties' definition file of the Belarusian NooJ module was made. The Belarusian N-korpus is the first widely available general Belarusian corpus. The Belarusian N-korpus currently contains ~50,000 texts (~30,000,000 tokens) taken from fiction, newspapers, journals and on-line editions. The texts of the corpus are grammatically annotated and contain metatextual information.

		Werb Cla	ina anjiji			
Erroropan	Debarandan 24 kargan	Category	Belarmian Nesj medule	Only in N-angaryz	Only in Nex)	in test
	Automatics Manager(P)	Newbor	Singular Placal	0	\$	<u>80</u>
Pa	Myresseni (M) Rosena (7) Altari (8) Marazar da (7)	Quarker	Moraka Tentahe	3	\$	Z . ().
No.	Chempsonie (R) Minyme (P) Systeme (R) Chempsonie (R)	Tras	Paran Par Pose	1		2
Зыретниках	Supermu(R) Meanpermu(N)	Reflection	Reflectes	1		l.
.7ba	Januara (3	Moot	Experience Industry	0	1	11) 1
llopeazamecan	Experiment(T) Herepension(T) Experiment(Exepterpension(T))		-	3		<u>E</u> :
Спрасние	Cleptone (1) Approxe (2) Promocepton mena (2)			1	P	<u>1</u>
Taxaber 1954	Линирыскоје (0)		Бальклантан асобла	1	•	

Fig. 14. Morphological characteristics of verb

The comparative analysis was performed on the morphological characteristics of different parts of speech listed in dictionaries of both programs. After the structure analysis of both Belarusian N-corpus and the Belarusian NooJ module, it can be concluded that the programs have quite developed system of characteristics of speech parts, but nevertheless some categories need to be improved, what was found out in the process of comparing lexical and grammatical bases. Comparison of the morphological characteristics of Verb is presented on Fig. 14.

6 Conclusion

In conclusion, the first one-million corpus for the Belarusian NooJ module is suitable for research in the following aspects:

- 1. Words polysemy processing in texts of different subjects;
- 2. Polysemic punctuation marks processing;
- 3. New lexical items search.

Besides, the one-million corpus is valuable for solving other important tasks:

- Conduction of several experiments in order to specify the **syntactic and morpho**logical grammar use efficiency of texts of each subject in the corpus, at minimum as well as maximum level.
- Taking thorough measures in order to create the *subject domain generator*. (This will be then very useful for the formation of special subject-oriented NooJ dictionaries.)
- The usage of the given corpus (in the most extent) in the process of Text-to-Speech synthesis with the help of available programs [7], required for such a process, and also when testing newly created applications.
- Carrying-out of a comparative analysis of this corpus with the same corpora in other languages (taking into account all necessary rules, language features in texts of each current corpus, various possible emerging issues, while building syntactic and morphological grammars, etc.).

Thus, it is essential that the first one-million corpus for The Belarusian NooJ Module has practical application in any line of linguistic research. In the near future the corpus is planned to be expanded up to approximately 5–10 million words.

Acknowledgements. Many thanks to T. Okrut, J. Baradzina, A. Fiodarau for their help in revising the language of this paper.

A Appendix

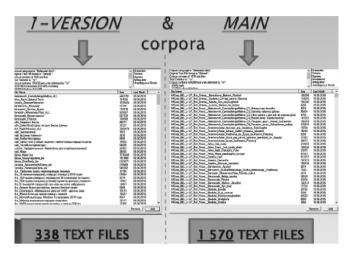


Fig. 15. 1-VERSION and MAIN corpora

*= 1 million corpus - Very		
Corpus Language is "Belarus		 Characters
Orginal Text File format is "D		Tokens
Corpus consists of 338 text fil	**	
Text Delimiter is: "Vs"	when the first states and the states	Ambiguities
Corpus contains 193236 text		Unambiguous Words
12581403 characters (233 dl	nj, makaleng	- 1
Fie Name		Size Last Modf. A
Alekslevich Carnobylskajah	faitura ALL	2584593 31.03.2015
Arlou Kana Dakara Dona	10110 June	120535 31.03.2015
Azeska_ZmovymViecaram		622242 31.03.2015
Bahdanovic Anoxosite		25930 31.03.2015
Bahusevic Kencka 6yase		68630 31.03.2015 a
Baradulin MissemascPlani	ALL.	313191 31.03.2015
Barsceuski, Benas capoka		157396 31.03.2015
Barapeuski Плачка		106452 31.03.2015
beh. Crappine i Santui		52419 31.03.2015
beh Coo's Bita/ta/va i se	CHIH BACITA LIMMAN	29703 31.03.2015
Bryl PtuskillHiniozdy ALL		4359422 31.03.2015
bast_3upras soks		7542 31.03.2015
bsat_Kwawaa Taneoxon		3818 31.03.2015
best_Pulicka Geo npauar		9902 31.03.2015
	адаіннік і найлягчайшы маршоутызатар	10480 31.03.2015
beat_Vons6woki apware.por		33861 31.03.2015
budzma_Fapagaamunema	Карамболь для окарбашукальнікаў	22711 31.03.2015
byel_Minox		24535 31.03.2015
Bykau_Karjer_be		5035145 31.03.2015
Bykau_VaucynajaJama_be		1043041 31.03.2015
Bykau_ZnakBiady_be		4400861 31.03.2015
Caropka_AdracennieAdOen		
	γy_pe	191249 31.03.2015
Carooka PeranogaCariu.t	ing and a second se	390652 31.03.2015
Carcoka PeranogaCariu b		390652 31.03.2015
Carooka Pieramoos Ceniu J		390652 31.03.2015
Carooka Peramoos Ceniu. 3 Carooka Peramoos Ceniu. 3 Marcina Marcina Marcina Santa Marcina Carooka Marcina Carooka Marci		390652 31.03.2015
Carooka PeranosaCenu.b	K K	390652 31.03.2015
Carooka Peramoos Ceniu. B Carooka Peramoos Ceniu. B Carooka Internet Control (Section Sec Peramo Carooka Peramo Carooka P	Comparing the second seco	390552 31.03.2015
Carcoka Peramosa Cenu. J Carcoka Peramosa Cenu. J Carcoka Peramona Internet Carco Antonio Internet Carco Antonio I	Comparing the second seco	390552 31.03.2015
Carcoka Peramosa Cenu. J Carcoka Peramosa Cenu. J Carcoka Peramona Internet Carco Antonio Internet Carco Antonio I	Constraints and a second	390552 31.03.2015
	Net and the second sec	390652 31.03.2015
Consider _ Persona Consultation for Construction of the construction of the person of the second	N TRANSPORTER AND A CONTRACT OF A DECEMBER AND A D	390552 31.03.2015
	N Linear have been been been been been been been be	200552 31.03.2015
	Al and a second s	390652 31.03.2015
	N insurance with a second sec	390652 31.03.2015
	Net and the second sec	200552 31.03.2015
	Net and the second sec	390652 31.03.2015 инститистика и постатование с нарада у бор, у портада в сокрытия на - борых, на - опоста постатова, нача на нала и 1. опостава с начала на наука, на и токата и начата на на 1. опостава с начала на наука, на начата на начата на на на опостатова на нача на нача на нача на нача на нача на на опостато на нача на нача на нача на нача на нача на на опостато на нача на нача на нача на нача на нача на на нача на нача на нача на нача на нача на нача на на нача на нача на нача на нача на нача на нача на нача на на нача на нача на нача на нача на нача на нача на нача на на нача на на нача на нача на на нача на на на нача на нача на нача на нача на нача на нача на нача на на нача на нача на нача на нача на нача на нача на нача на на нача на нача на нача на нача на нача на нача на на на нача на нача на нача на нача на нача на нача на нача на на нача на нача на нача на нача на нача на нача на нача на на нача на нача на на нача на нача на на нача на нача на на нача на на на на на на нача на нача на нача на нача на на на нача на нача на нача на нача на нача на нача на на нача на нача на
	Ne ne service and	200552 31.03.2015
	Ne ne se	390652 31.03.2015
	No manufacture and a second an	390652 31.03.2015
	Ne ne service and rest and res and res and rest and rest and rest and rest and rest and rest	390652 31.03.2015
	Ne	390652 31.02.2015
	Ne	390652 31.02.2015
	Ne ne service de la construir	390652 31.02.2015
	Ne ne version de la construir	390652 31.02.2015
	No ne se	390652 31.02.2015
	be The second	390652 31.02.2015
	Ne	390652 31.03.2015 инститистите проготорые собщение собщ
	Ne ne crate prese version en	390652 31.02.2015

Fig. 16. Realized Linguistic analysis of the 1-VERSION corpus

References

- 1. NooJ: a linguistic development environment [Electronic resource] (2015). http:// www.NooJ4nlp.net/. Accessed 08 May 2015
- 2. The Levenshtein-Algorithm [Electronic resource] (2015). http://www.levenshtein.net/. Accessed 24 Sept 2015
- Taylor, P.: Text-to-Speech synthesis. In: Taylor, P. (ed.) Text Decoding, pp. 89–92. Cambridge University Press, Cambridge (2009). Chapter 5
- Hetsevich, Y.: Overview of Belarusian and Russian dictionaries and their adaptation for NooJ. Hetsevich, Y., Hetsevich, S. In: Vučković, K., Božo, B., Max, S. (eds.) Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the NooJ 2011 International Conference, pp. 29–40. Cambridge Scholars Publishing, Newcastle (2012)
- 5. Hetsevich, Y.: Accentual expansion of the Belarusian and Russian NooJ dictionaries. Hetsevich, Y., Hetsevich, S., Lobanov, B., Skopinava, A., Yakubovich, Y. In: Donabédian, A., Khurshudian, V., Max, S. (eds.) Formalising Natural Languages with NooJ : Selected Papers from the NooJ 2012 International Conference, pp. 24–36. Cambridge Scholars Publishing, Newcastle (2013)
- Аўтаматызаваная апрацоўка сімвальных выразаў у тэкстах для сістэмы сінтэзу беларускага маўлення. Беларускі N-корпус [Electronic resource] (2015). http://bnkorpus.info/. Accessed 17 May 2015
- Corpus.by. Corpus.by [Electronic resource] (2015). http://www.corpus.by/. Accessed 08 May 2015