The First One-Million Corpus for the Belarusian NooJ Module

Ivan Reentovich[1(✉)], Yuras Hetsevich[1], Valery Voronovich[2], Evgenia Kachan[2], Hanna Kozlovskaya[2], Angelina Tretyak[2], and Uladzimir Koshchanka[3]

[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.

Keywords: Corpora · Belarusian NooJ-module · Statistical analysis · Part-of-Speech tagging · Machine-learning algorithms · Levenshtein algorithm · The machine-learning model · Counter-check · Spelling errors · Concatenation-in-paradigm · Unknown words search · Known words search · Clustering · Belarusian N-Corpus · Text processing

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.

T. Okrut et al. (Eds.): NooJ 2015, CCIS 607, pp. 3–15, 2016.
DOI: 10.1007/978-3-319-42471-2_1

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 pre-processing 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 Levenshtein 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.

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

№	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:** 338**)**
2	The **"purity" check** of the developed corpus: *1. Unknown or unidentified symbols;* *2.* ***[In addition]:*** *words with apostrophe* ***(')****;* 3. ***[In addition]:*** *words with hyphen/* ***(–)*** *symbol.* *2, 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)* ***пад'*** *ехаць (incorrect);* *2)* ***цёмна-*** *сіні (incorrect).* *(In both cases the words are divided by the* ***space*** *and NooJ can't identify them as one unit).* **E.g.** *1)* ***пад'****ехаць (correct);* *2)* ***цёмна-****сіні (correct)*]	1. There are NO ANY **unknown** or **unidentified symbols** in the given corpus (all texts must be encoded to UTF-8). 2. **5818** 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: — **25 615 occurrences** of all wordforms with hyphen/ (–) symbol (The special NooJ-grammar **SearchWordFormsWithHyphen.fst** has been applied in this case); — **19 728 occurrences** of all unique (1 occurrence per match) wordforms with hyphen/ (–) symbol (The special NooJ-grammar **SearchWordFormsWithHyphen.fst** has been applied in this case); — **23 663 occurrences** of all wordforms with hyphen/ **(–)** symbol using special **<WORDFORMWITHHYPHEN> query**. (After the realization of Linguistic Analysis throughout the corpus, applying necessary NooJ resources **general_be.nod [4, 5]** and **SearchWordFormsWithHyphen.nog**); — **18 351 occurrences** of all unique (1 occurrence per match) wordforms with hyphen/ `–` symbol, using special **<WORDFORMWITHHYPHEN> query**. (Such results were got after the realization of Linguistic Analysis throughout the corpus, applying necessary NooJ resources **general_be.nod** and **SearchWordFormsWithHyphen.nog**).
3	The **wordforms** total is counted in the developed corpus	Wordforms (ALL) (**<WF>** = *1 884 971*) Wordforms (ALL, unique [1 occurrence per match]) (**<WF>** = *197 712*)
4	The **unknown words search** is realized from the developed corpus	Unknown words (ALL) (**<UNK>** = *140 235*) Unknown words (ALL, unique [1 occurrence per match]) (**<UNK>** = *50 186*)
5	Creation of the dictionary of unknown (**<UNK>**) unique word usage [*.dic]	File *UNKNOWNS.dic* (*49 749* unknown words (after the necessary correction of the unique **<UNK>** = *50 186* result))
6	Creation of the dictionary of naturalized lexical- and grammatical information for the whole list of unknown unique words (on the basis of the file *UNKNOWNS.dic*)	**[**The **more extended "purity" check** of the developed corpus. (The **"disposal" of** Roman alphabet letters and other problematic cases in texts of the corpus)**]**

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: *1 mil c_BE.dic (197 712* ***unique occurrences)***
9	The **punctuation marks search** is realized from the developed corpus	Punctuation marks (ALL) (**<P>** = *605 512*) Punctuation marks (ALL, unique [1 occurrence per match]) (**<P>** = *61*)
10	The **digrams search check** is realized from the developed corpus	Digrams *Concordance* (**Digrams** = *1 787 305* [occurrences]) *Dictionary* (*1 mil cor VB_(digrams).dic*)
11	The **Part of Speech tagging of words** is realized within the developed corpus	Parts of Speech (**<NOUN>** = *523 193*) –>ALL (**<NOUN>** = *58 376*) –> unique [1 occurrence per match] (**<VERB>** = *326 739*) –> ALL (**<VERB>** = *45 682*) –> unique [1 occurrence per match] (**<ADJECTIVE>** = *194 682*) –> ALL (**<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.

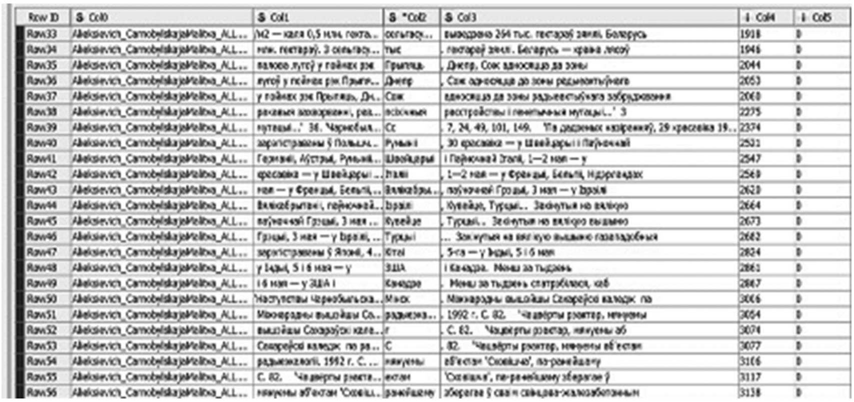
3.2 Statistical Analysis of the Text Corpus '*1 million corpus.noc*'

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)).
2. The search of wordforms (all wordforms, which are present in the corpus) using special queries (<WF>, <UNK>, <DIC>, <NOUN>, <VERB>, <ADJECTIVE>, <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 **main attribute of a wordform** 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	S Col0	S Col1	S *Col2	S Col3	I Col4	I Col5
Raw33	Aleksievich_CarnobylskajaMalitva_ALL...	[illegible]	[illegible]	[illegible]	1918	0
Raw34	Aleksievich_CarnobylskajaMalitva_ALL...	[illegible]	[illegible]	[illegible]	1946	0
Raw35	Aleksievich_CarnobylskajaMalitva_ALL...	[illegible]	[illegible]	[illegible]	2044	0
Raw36	Aleksievich_CarnobylskajaMalitva_ALL...	[illegible]	[illegible]	[illegible]	2053	0
Raw37	Aleksievich_CarnobylskajaMalitva_ALL...	[illegible]	[illegible]	[illegible]	2060	0
Raw38	Aleksievich_CarnobylskajaMalitva_ALL...	[illegible]	[illegible]	[illegible]	2275	0
Raw39	Aleksievich_CarnobylskajaMalitva_ALL...	[illegible]	[illegible]	[illegible]	2374	0
Raw40	Aleksievich_CarnobylskajaMalitva_ALL...	[illegible]	[illegible]	[illegible]	2521	0
Raw41	Aleksievich_CarnobylskajaMalitva_ALL...	[illegible]	[illegible]	[illegible]	2547	0
Raw42	Aleksievich_CarnobylskajaMalitva_ALL...	[illegible]	[illegible]	[illegible]	2569	0
Raw43	Aleksievich_CarnobylskajaMalitva_ALL...	[illegible]	[illegible]	[illegible]	2620	0
Raw44	Aleksievich_CarnobylskajaMalitva_ALL...	[illegible]	[illegible]	[illegible]	2664	0
Raw45	Aleksievich_CarnobylskajaMalitva_ALL...	[illegible]	[illegible]	[illegible]	2673	0
Raw46	Aleksievich_CarnobylskajaMalitva_ALL...	[illegible]	[illegible]	[illegible]	2682	0
Raw47	Aleksievich_CarnobylskajaMalitva_ALL...	[illegible]	[illegible]	[illegible]	2824	0
Raw48	Aleksievich_CarnobylskajaMalitva_ALL...	[illegible]	[illegible]	[illegible]	2861	0
Raw49	Aleksievich_CarnobylskajaMalitva_ALL...	[illegible]	[illegible]	[illegible]	2867	0
Raw50	Aleksievich_CarnobylskajaMalitva_ALL...	[illegible]	[illegible]	[illegible]	3006	0
Raw51	Aleksievich_CarnobylskajaMalitva_ALL...	[illegible]	[illegible]	[illegible]	3064	0
Raw52	Aleksievich_CarnobylskajaMalitva_ALL...	[illegible]	[illegible]	[illegible]	3074	0
Raw53	Aleksievich_CarnobylskajaMalitva_ALL...	[illegible]	[illegible]	[illegible]	3077	0
Raw54	Aleksievich_CarnobylskajaMalitva_ALL...	[illegible]	[illegible]	[illegible]	3106	0
Raw55	Aleksievich_CarnobylskajaMalitva_ALL...	[illegible]	[illegible]	[illegible]	3117	0
Raw56	Aleksievich_CarnobylskajaMalitva_ALL...	[illegible]	[illegible]	[illegible]	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 system[1]. This dictionary was meant to "train" all possible word paradigms by the above-named 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.)

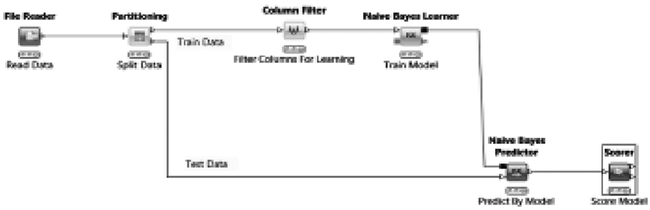


Fig. 2. Train model

   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	Col0	Col1
Row1750	адасечко	чко
Row1751	адасобліваўся	ўся
Row1752	адась	ась
Row1753	адасю	асю
Row1754	адася	ася
Row1755	адасё	асё
Row1756	адасік	сік
Row1757	адасічак	чак
Row1758	адацыстрацыі	ыі
Row1759	адаў	даў
Row1760	адаўнаму	аму
Row1761	адб	адб
Row1762	адбірае	рае
Row1763	адбаліць	ць
Row1764	адбараніў	ніў
Row1765	адбахаў	хаў

Fig. 3. Unknown data

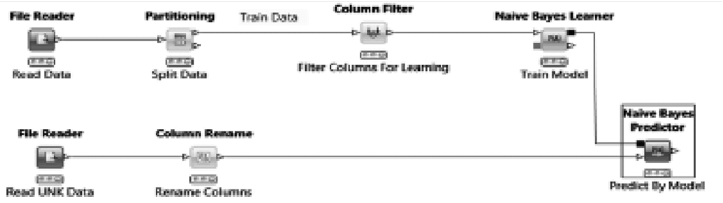


Fig. 4. POS prediction

Row ID	Seq	Ending	PartOfSpeech
Row1750	адасечко	чко	OTHER
Row1751	адасобліваўся	ўся	VERB
Row1752	адась	ась	OTHER
Row1753	адасю	асю	OTHER
Row1754	адася	ася	VERB
Row1755	адасё	асё	OTHER
Row1756	адасік	сік	OTHER
Row1757	адасічак	чак	OTHER
Row1758	адацыстрацыі	ыі	OTHER
Row1759	адаў	даў	OTHER
Row1760	адаўнаму	аму	OTHER
Row1761	адб	адб	OTHER
Row1762	адбірае	рае	VERB
Row1763	адбаліць	ць	VERB
Row1764	адбараніў	ніў	VERB
Row1765	адбахаў	хаў	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).

Table 2. Possible variants of data check results

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

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.)

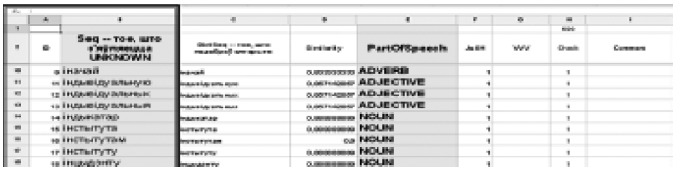


Fig. 6. Unknown words in the summary table of the Part-of-Speech Tagging Countercheck

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 (**магілёўскага**, **устаголдзем**);
– 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.)

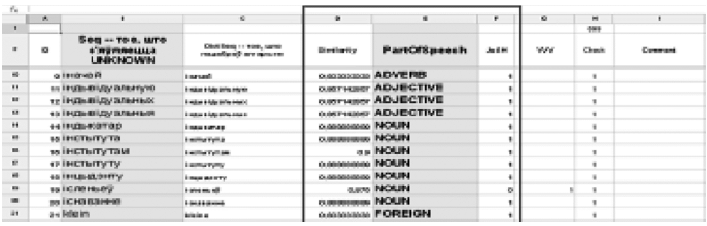


Fig. 7. Linguists' checkout of wordforms recognized by Levenshtein algorithm in the summary table of the Part-of-Speech Tagging countercheck

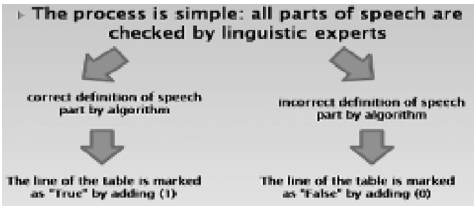


Fig. 8. The process of Parts-of-Speech tagging countercheck

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.)

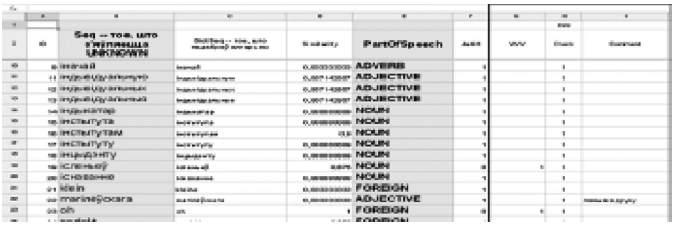


Fig. 9. Editing the results of linguist's checkout of wordforms recognized by Levenshtein algorithm in the summary table of the Part-of-Speech Tagging countercheck

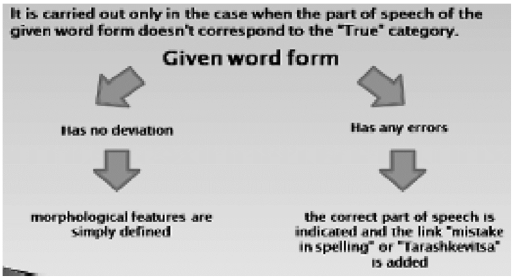


Fig. 10. The process of editing

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.)

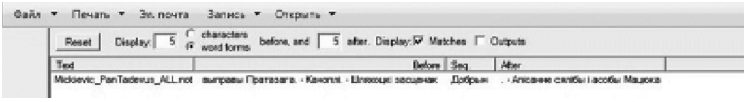


Fig. 11. Addressing to the context

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.)

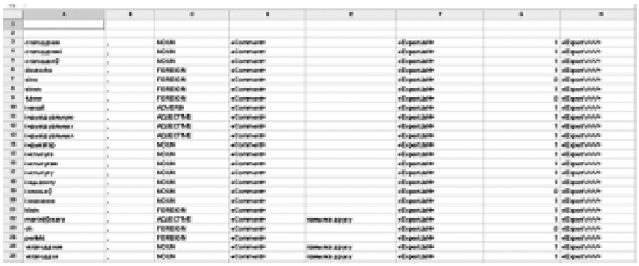


Fig. 12. The Concatenation-in-Paradigm results

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.)

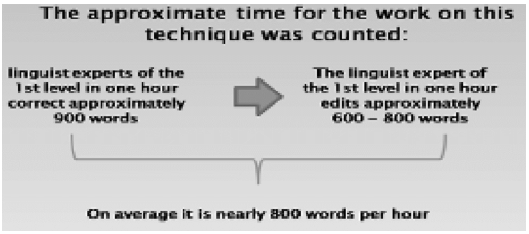


Fig. 13. The approximate time count for the work on the technique

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.

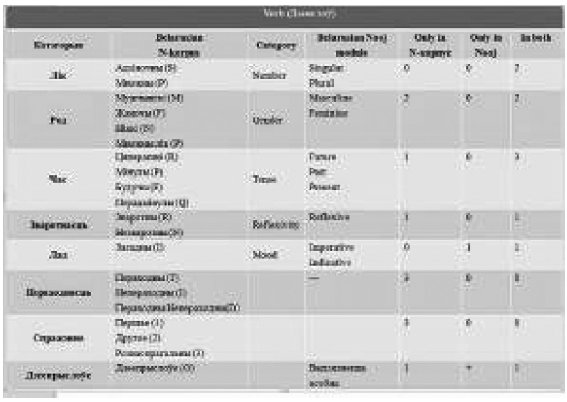


Fig. 14. Morphological characteristics of verb

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 morphological 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

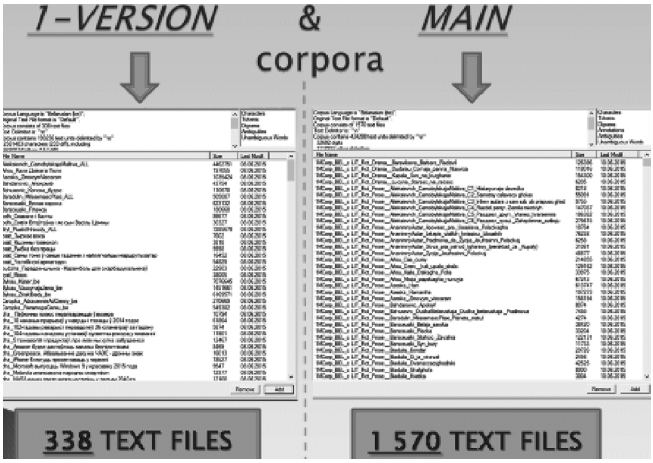


Fig. 15. 1-VERSION and MAIN corpora

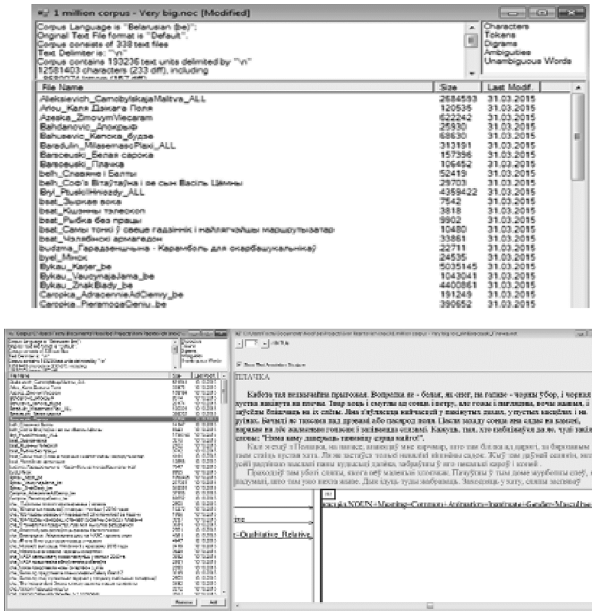


Fig. 16. Realized Linguistic analysis of the 1-VERSION corpus

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
3. Taylor, P.: Text-to-Speech synthesis. In: Taylor, P. (ed.) Text Decoding, pp. 89–92. Cambridge University Press, Cambridge (2009). Chapter 5
4. 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)
6. Аўтаматызаваная апрацоўка сімвальных выразаў у тэкстах для сістэмы сінтэзу беларускага маўлення. Беларускі N-корпус [Electronic resource] (2015). http://bnkorpus.info/. Accessed 17 May 2015
7. Corpus.by. Corpus.by [Electronic resource] (2015). http://www.corpus.by/. Accessed 08 May 2015