Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2017”

Moscow, May 31—June 3, 2017

TEXT NORMALIZATION IN RUSSIAN TEXT-TO-SPEECH SYNTHESIS: TAXONOMY AND PROCESSING OF NON-STANDARD WORDS

Cherepanova O. D. (cherepanova.od@gmail.com)

Moscow State University, Moscow, Russia

Alongside with ordinary words, natural-language text also contains non-standard words (NSWs), such as abbreviations, acronyms, dates, phone numbers, currency amounts etc. Before phonetizing these text elements in Text-to-Speech synthesis, it is necessary to normalize them by replacing them with an appropriate ordinary word or word sequence. NSWs are increasingly diverse and most of them require specific normalization rules. In this paper, we present a taxonomy of NSWs for the Russian language developed on the basis of news texts, software and car reviews and instruction manuals. We grouped NSWs that have similar normalization rules or patterns taking into account their graphic form and their context dependence. We propose five main groups of NSWs: abbreviations (including acronyms and initialisms), text elements containing numbers, special characters, foreign words written in the Latin alphabet and mixed-type non-standard words. In this work, we describe these NSW types and address the issue of their normalization in Russian Text-to-Speech synthesis.

Key words: Text-to-Speech-synthesis, text normalization, Russian

НОРМАЛИЗАЦИЯ ТЕКСТА В СИСТЕМЕ РУССКОЯЗЫЧНОГО СИНТЕЗА «ТЕКСТ-РЕЧЬ»: КЛАССИФИКАЦИЯ И ОБРАБОТКА НЕСТАНДАРТНЫХ ТЕКСТОВЫХ ОБЪЕКТОВ

Черепанова О. Д. (cherepanova.od@gmail.com)

МГУ им. М. В. Ломоносова, Москва, Россия

Ключевые слова: синтез речи по тексту, нормализация текста, русский язык

1. Introduction

To illustrate the subject of our research let us consider the following headline: *Nokia планирует купить разработчика ПО Comptel за €347 млн* In order to correctly pronounce this sentence, a Russian Text-to-Speech synthesizer has to normalize most of these words by completing the following operations: expand the abbreviation *млн* to *миллионов*, transform the number *347* into its graphical representation *триста сорок семь*, transliterate the names *Nokia* and *Comptel* written in the Latin alphabet, classify *ПО* as an initial abbreviation and insert the word *евро* instead of the symbol *€*. These text units requiring additional processing rules are called ***non-standard words*** or **NSWs** ([Black etal. 1999] and others). In TTS-synthesis, NSWs should be processed at the stage of text normalization in order to get at the output "a sequence of white-space separated accentuated orthographic words" [Krivnova 1998: 5]. There is a wide range of NSWs and their number increases with every new text: they include all types of abbreviations, dates, phone numbers, addresses, special characters etc. However, many of these NSW groups have similar normalization rules.

Fig. 1 presents a basic NSW normalization algorithm based on [Sproat etal. 2001: 304] and adapted to Russian as an inflected language (Figure 1). The data used at each step (hand-written rules, dictionaries, language models etc.) depend largely on the TTS-system and can vary for different domains.

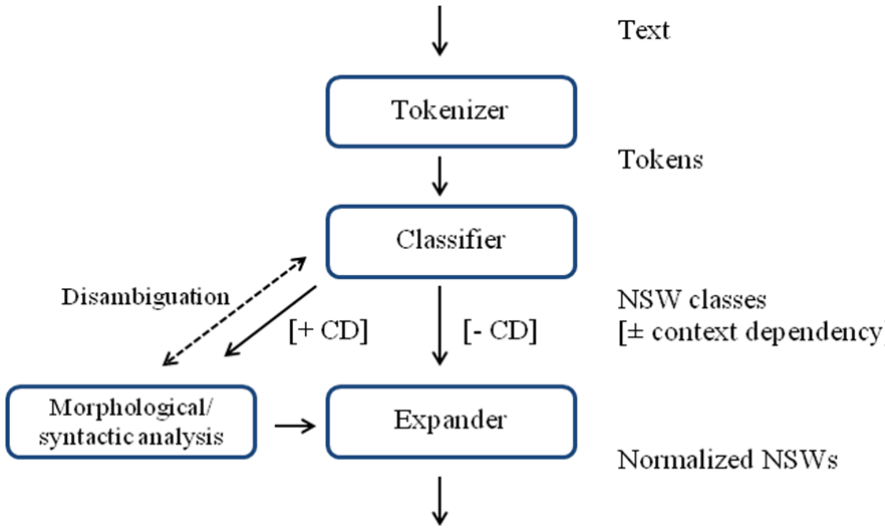

Fig. 1. Basic NSW normalization algorithm

In the following sections, we describe a NSW taxonomy designed to simplify text normalization as a stage of Russian TTS-synthesis.

2. Previous approaches

While NSW processing is described in detail for English TTS-systems (see [Black et al. 1999], [Olinsky, Black 2000], [Sproat et al. 2001]), as far as we know, there are practically no published works on this subject for the Russian language. The issue of NSW disambiguation in Russian TTS-synthesis is discussed in [Khomitsevich etal. 2013], and a detailed taxonomy of Russian abbreviations is presented in [Krivnova 1999]. A research on NSW normalization in inflected languages (on the example of Greek) was carried out by [Xydas et al. 2004] with an emphasis on NSW expansion rules.

One of the most detailed researches on NSW normalization is [Sproat etal. 2001]. The authors proposed a systematized NSW taxonomy and investigated several techniques of NSW normalization. Sproat's taxonomy provided the basis for our classification but had to be extended due to peculiarities of the Russian language. The following changes were introduced:

- a feature '± context dependency' was added to all NSW classes (since in Russian many of them require grammatical agreement within the sentence);
- a new NSW class was added for special characters (see Section 3.3);
- a new NSW class was added for words written in the Latin alphabet (Section 3.4).

3. A Taxonomy of NSWs

In our taxonomy, we tried to define NSW categories with similar normalization rules or patterns. Our taxonomy was developed on the basis of texts from news papers, car websites, software descriptions and instruction manuals.

Taking into account our data analysis and previous taxonomies mentioned above, we propose the following taxonomy of Russian NSWs: 1) abbreviations; 2) NSWs containing numbers; 3) special characters; 4) Latin alphabet words; 5) mixed-type NSWs. The classification is based on following NSW features: graphic form of NSWs, their potential normalization rules and their context dependency. The NSW taxonomy is summarized in Table 1.

Table 1. Taxonomy of Russian non-standard words

Class name	Context dependency	Examples
1. Abbreviations		
1.1. Shortened abbreviations	–	исполком, завлаб
1.2. Graphic abbreviations	+	филол., т.е., оз., 60 км/ч, 20 кг, пр-т
1.3. Initial letter abbreviations		
1.3.1. Initialisms	–	МГУ, СНГ
1.3.2. Acronyms	–	МГИМО, ГУМ
1.3.3. Mixed-type initial abbreviations	–	ЦСКА, ГИБДД
1.4. Mixed-type abbreviations	–	БелАЗ

Class name	Context dependency	Examples
2. NSWs containing numbers	+	12-ый *том*, *Иван* IV, 13:45 *часов*, 12 *тыс.* **€**
	–	*ауд.* 956, *тел.:* 8 (495) 123 45 67, *Android* 2.3
3. Special characters	+	$, **€**, **Y**, °
	–	+, –, ±, ≤, <, >, *, &, #, ~
4. Latin alphabet words	–	Windows, microSD;
5. Mixed-type NSWs	+–	№38-ФЗ, MP3-плеер

Note: 'Context dependency' means here the need for grammatical agreement of expanded NSWs (*у оз.* /озера/ *Селигер* 'near the Seliger lake' but *на оз.* /озере/ *Селигер* 'at the Seliger lake').

In the following sections, each NSW category is described in more detail.

3.1. Abbreviations

We use the term 'abbreviation' here in the broadest sense: it includes all kinds of abbreviations, acronyms and initialisms. There are several categories of abbreviations in Russian: ***1.1. shortened word combinations*** (*исполком*, *колхоз*); ***1.2. graphic abbreviations*** marked by a full stop, hyphen, slash or other graphical means (*т.е.*, *н.э.*, б/у); ***1.3. initial letter abbreviations*** formed by initial components of word combinations and pronounced in their shortened form: *МГУ* /эм-гэ-у/[1], *ГУМ* /гум/); ***1.4. mixed-type abbreviations*** (*БелАЗ*).

Most ***shortened word combinations*** (*роддом, детдом, телесеть, драмкружок, теракт, запчасть* etc.) are pronounced as ordinary words, and thus usually pose no problem in TTS-synthesis. However, abbreviations like *завлаб, местком* or *продмаг* pronounced in their shortened form might be difficult to understand; for better intelligibility they probably should be normalized to their expanded form (*заведующий лабораторией, местный комитет, продуктовый магазин*).

Graphic abbreviations have rather simple normalization rules: most of them are widely used, are generally included into dictionaries and, thus, can be verbalized in their full form. Cases of ambiguity can generally be resolved using the context: e.g., ***г.*** is expanded as /город/ 'city' if the previous word starts with a capital letter (*г. Москва*) or as /год/ 'year' after a number sequence (*2017 г.*). Graphic abbreviations can be formed in several different ways: a) by omitting the end of the word (*филол.* –/филологический/, *архит.*—/архитектурный/); b) by using initial letters of each word or syllable (*н. э., и т. д., л. с., пп., гг., вв.*); c) as abbreviations without a full stop (*км, м, кг, л, мл, т*), d) by omitting the middle of the word marked by a hyphen (*г-н, г-жа, пр-т*), e) marked by a slash (*н/Д*—/на-Дону/; *б/у*—/бывший в употреблении/). Here, a special group form abbreviations of physical units (*км/ч; об/мин; Мбит/с* etc.).

[1] In the present paper we indicate proposed graphic normalization forms of NSWs by slash signs (/).

As we can see, graphic abbreviations are very diverse and, according to our data analysis, they are more frequent than other abbreviation types. As M. Rovinskaya pointed out in her undergraduate's thesis, more than the half of abbreviations marked with a full stop have only one extension variant [Rovinskaya 1998: 6]. This makes it possible to cover most graphic abbreviations by means of a dictionary and local syntactic analysis.

There are two groups of ***initial letter abbreviations***: ***1.3.1 initialisms*** (pronounced one letter at a time: *СНГ* /эс-эн-гэ/, *ГДР* /гэ-дэ-эр/); ***1.3.2 acronyms*** (pronounced as one word: *ГУМ* /гум/, *ЛЭП* /лэп/); ***1.3.3 mixed-type initial abbreviations*** (one part is pronounced as a single word, and the other—as separate letters: *ЦСКА* /цэ-эс-ка/, *ГИБДД* /ги-бэ-дэ-дэ/). In our analyzed text data, 58% of initial letter abbreviations were acronyms.

Even though we can write quite simple normalization rules for this NSW class, there is still need for a dictionary. To begin with, there can be more than one way of pronouncing the same letter: the letter 'Ф' can be pronounced both as /эф/ (*РФ*, *ФСБ*) and /фэ/ (*ФБР*, *ФРГ*). Secondly, there is a range of exceptions: e.g., *США* /сэ-шэ-а/, *ТВ* /теле/, *МЮ* /манчестер юнайтед/, and the abbreviation *МСК* which formally might be classified as an initial letter abbreviation and is pronounced as /москва/ 'Moscow' or even as a whole sentence—/по московскому времени/ 'Moscow time'.

The stress in acronyms falls usually on the last syllable, but there are some exceptions (*НАТО*, *ЮНЕСКО* etc. [Krivnova 1998: 4]) that should be included in a dictionary.

Another challenge in Russian TTS-synthesis are abbreviations written in the Latin alphabet. According to the analyzed data, English acronyms and initialisms are widely used in Russian texts and compose 29% of all initial abbreviations. A normalization method for English words and abbreviations is proposed in Section 3.4.

3.2. NSWs containing numbers

Number sequences are pronounced in different ways depending on their function—and as we can see from Table 2, there is a wide range of their functions. We defined three distinctive features for NSWs containing numbers (hereinafter 'number sequences' or 'NS'): their ***context dependency***, ***verbalization format*** and ***number class***.

We distinguish three NS categories by their verbalization format: (A) NSs pronounced as one number; (B) NSs pronounced one number at a time; and (C) NSs with special verbalization formats.

There are three number classes NSs can be expanded with: 1) cardinal numbers; 2) ordinal numbers; and 3) collective numbers.

In compliance with these features we defined 17 categories of number sequences:

Table 2. Taxonomy of NSWs containing numbers

NS category	Class[2]	Form[3]	CD[4]	Examples
2.1. cardinal numbers	C	A	+	12 домов
2.2. numbers (excluding phone numbers)	C	B	–	ауд. 956
2.3. phone numbers	C	C	–	8 (495) 123 45 67; 123-45-67
2.4. addresses	C	C	–	д.1, к.2, кв.123; д.2/3
2.5. index numbers	C	B	–	123456
2.6. time indication	C	C	+	в 13:45; к 13:45
2.7. money amounts	C	C	+	$ 1,5; 1.20 руб; 12 тыс. €
2.8. percentage	C	A	+	29,99%; 50%
2.9. series numbers	C	B	–	Android 2.3; § 4.2.3.
2.10. multiplicative constructions 'number (GEN) + adjective'	C*	A	–	11-метровый; 4-кратный; 2,0-литровый
2.11. ordinal numbers	O	A	+	12-ый том; Иван IV; 1. …; 2. …
2.12. dates	O	C	+	2.05.06; 02/05; 2 мая 2006г.
2.13. years	O	A	+	2001г.; 2010/11 гг.; 60-е
2.14. fractions	O*	A	+	1/4 финала; 2/3 опрошенных
2.15. collective numbers	Col	A	–	5-ро друзей; 2-е суток
2.16. denumerate constructions	Col*	A	–	16-ричный режим; 2-ичная нумерация
2.17. multiplicative numbers	Col*	A	+	3-ной подогрев; 4-ной сальхов

In this article, we are not going to provide a detailed review for each of the NS classes since it is a rather broad area described for the Russian language, in particular, in [Azerkovich 2013]. However, let us take a closer look at the normalization schemes of some of these NSW categories.

For better intelligibility of **money amounts**, currency units should be verbalized: *$12,25* /двенадцать долларов и двадцать пять центов/ (but: *$12,25 млн* / двенадцать целых и двадцать пять сотых миллионов долларов/). Normalization rules of currency symbols are discussed in Section 2.4.

Even though there are only a few conventional **time formats** in literary Russian, there is still room for ambiguity. For example, such NSs as *19-30*, *19.30* or *19:30* can be used not only referring to time, but also to sport scores or prices. The form *1:30* can denote day time (/час тридцать/), the length of a phone call or race time (/минута

[2] Number class: C = cardinal numbers; O = ordinal numbers; Col = collective numbers

[3] Normalization format

[4] Context dependency

тридцать секунд/). In some cases, we can disambiguate NSs using context key words (currency names, time abbreviations etc.).

Most ordinal number NSs are pronounced as one number (group A). According to our text analysis, NSWs are expanded by ordinal numbers in date statements (*12 мая 2001 г., 2010/11 гг.*) or if there is a marked word ending (*12-ый том*, *60-е*). Ordinal numbers are also used in Roman numerals (*XIX век, Карл V, глава I, XX съезд КПСС*).

There are several **date formats**, but the standard format in Russia is 'day-month-year'. The year can be denoted both by two and four digits: *04.05.2006* and *04.05.06*. Numbers in date statements are separated by full stops (most frequent), hyphens or slashes: *04.05.2006; 04-05-2006; 04/05/2006*. In order to expand a date statement, the day number should be replaced by an ordinal number (/четвертое/), the month—by the corresponding month name in the genitive case (/мая/), and the year—by an ordinal number in the genitive case (/две тысячи шестого года/). According to our research, full date forms as listed above are used not very often. Usually, month numbers are already replaced by month names in texts (*4 мая* or *4 мая 2006 г.*). However, only the day number requires grammatical agreement with the context. Another date format that might cause difficulties is *4/5* (/четвертое мая/) since it could also denote a fraction. The normalization of slashes and other special characters is discussed in Section 3.3.

3.3. Special characters

There is a small group of context-dependent special characters requiring agreement in case and number. One of the most frequent is the percent sign **%** /процент*[5]/. It is sometimes used as an abbreviation in constructions like *10%-й раствор; 20%-я сметана*. The characters **§** /доллар*/, **№** /номер*/, ° /градус*/, ″ /дюйм*/ and currency symbols (€, **$**, £ etc.) are also context-dependent. It should be pointed out that currency symbols usually precede number sequences (€12), but should be pronounced after them: /двенадцать евро/. When used after words like *тыс., млн, млрд* etc., currency names should be normalized to their genitive form (*прибавить к 12 тыс. $ /долларов/*).

For context-independent special characters, the context can still be of paramount importance as it could be used for disambiguation. Thus, a **slash** can both denote a fraction (*2/3* — /две трети/), a division sign (*2/3 = 0.67* /два поделить на три/), a separator in physical units or date statements (*120 об/мин* — /оборотов в минуту/) and the meaning 'or' (*7 шт / 14 шт* — /семь штук **или** четырнадцать штук/). In an address, the token *2/3* can also be pronounced as /дом два дробь три/.

Semicolons are mostly used as punctuation marks, but can also denote **proportions** (*1:1000 /один **к** тысячи/; 50:50 /пятьдесят **на** пятьдесят*). **Superscript numbers** can be used as power exponents in physical units ($м^2$, $см^3$) or as footnotes. Even if there is no need in verbalizing the footnote number, a TTS-synthesizer should still recognize them in order to put in the footnote text in the right place.

A detailed taxonomy of special characters is described in Table 3.

---

[5] The symbol * marks context-dependent word endings.

Table 3. Taxonomy of special characters

Special character class	Examples	Pronunciation
1. Context-dependent		
§, №, %	§ 12, № 3, 20 %	/параграф* n/, /номер* n/, /n процент*/
Physical quantities (°,″)	+12 °C; −12° по Цельсию; экран 6″.	/(плюс/минус) n градус* (по Цельсию)/, /экран в n дюймов/)
Currencies	12 тыс. $, 12€, ¥ 150	/n тысяч долларов/, /n евро/, /n иен/
2. Context-independent		
Mathematical characters:		
+, −, ±	(+7); Google+; −0,5; ±12	/плюс/; /минус/; /плюс-минус/
*, /, :	2,5*20*5,3; 3/4; 7шт./14шт.; 1:100; 50:50; 60 км/ч	/на/; /три четвертых (три четверти)/; /или/; /к/; /в/
^, ²	2^6; *м²*	/x в степени n/; /метр квадратный/
=, >, <, ≤	2+2 = 4; 4>2	/равно/; /больше, чем/; /меньше, чем/; /больше или равно/
Other characters:		
&, #, @, ~, x, ©	Маркетинг&Реклама; #100; abc@web.de; 4x3; ~25%	/и/; /решетка/; /собака/; /на/; /приблизительно/
Footnotes	[1], (1), ², *, **	verbalization of the footnote text

The range of specialized characters used in texts is certainly much wider and largely depends on the topic. In the present paper, we listed only the most frequent special symbols occurring in Russian texts.

3.4. English words and word combinations in Russian texts

Nearly every Russian text contains words written in the Latin alphabet. These are mostly English names of companies, mass media, brands or software. In order to verbalize English words in a Russian TTS-system, we need to transform them into the graphic (or phonetic) system used by the synthesizer for ordinary Cyrillic words. One possible approach here is to use **orthographic transcription**. Provided that we have an IPA transcription for these English words[6], with the help of English-Russian orthographic transcription rules we can convert them into the Cyrillic alphabet with due

[6] There is a large number of publicly available IPA transcription programs for the English language. However, most of them are based on dictionaries and, therefore, not all organization names can be automatically transcribed.

regard to their pronunciation in English. This method is discussed in [Cherepanova 2016]; here we present only a few examples:

(1) Microsoft ['maɪkrəʊsɒft]—/ма+йкрософт/[7]

(2) British Airways ['brɪtɪʃ 'eəweɪz]—/бри+тиш э+рвэйс/.

3.4.1. Abbreviations in the Latin alphabet

There are different types of English abbreviations in Russian texts: acronyms and initialisms (*SMS, SIM, GPS, OS*), graphic abbreviations (*Ltd., Co., Inc.*), mixed-type abbreviations (*Mp3, 4G, microSD, DivX, MPEG-2, 3D, e-mail, iPhone*). Quite common are composed constructions where English abbreviations precede Russian words: *USB-накопитель, DVD-плеер, FM-передатчик, IP-адрес*.

As it seems, graphic abbreviations should be expanded to their full form and vocalized using the same rules as for ordinary English words. Numbers used in English word constructions are always pronounced in Russian.

The verbalization of English initialisms depends on several factors. Let us compare the following abbreviations used in Russian: *3D* /три-**дэ**/ and *DVD* /**ди**-ви-**ди**/; *диск С* /диск **цэ**/ and *CD* /си-**ди**/; *MP3* /эм-**пэ**-три/ and *IP* /ай-**пи**/. Presumably, the pronunciation of such initialisms depends on their length (single letters vs. letter sequences) and usage frequency. But even the same English initialism can be pronounced in different ways: the most common pronunciation of *HTML* is /эйч-ти-эм-эль/, but there is also a rather frequent informal variant /аш-тэ-эм-эль/. The issue of intelligible and natural verbalization of English abbreviations in Russian TTS-synthesis needs further investigation (for example, an analysis of the speech of Russian news readers).

4. Conclusion

In this paper, we presented a taxonomy of non-standard words requiring special normalization rules in Russian TTS-synthesis. This taxonomy might be used in text normalization tasks and help to systematize hand-written rules of NSW processing. The presented NSW list is not intended to be exhaustive: specialized texts might contain a wide range of topic-specific abbreviations, number sequences or special characters not mentioned here. Our goal was to systematize the main NSW categories in order to provide a basis for further investigations. Moreover, our research didn't include the issue of out-of-vocabulary word processing and of spelling mistakes. Such text elements also require additional processing rules at the stage of text analysis and, in particular, were included by Sproat in one of his NSW categories. This could be an issue for further research.

---

[7] Orthographic transcription is indicated by slashes (/), the sign '+' indicates the position of the stress.

References

1. *Azerkovich I. L.* (2013) Automatic identification of numbers and number groups in text normalization in Text-to-Speech synthesis [Avtomaticheskaya identifikatsiya tsifr i chislovykh grupp v protsesse normalizatsii teksta pri sinteze rechi] // Proc. XX International youth conference for students, postgraduate students and young scientists "Lomonosov-2013" [Materialy XX Mezhdunarodnoy molodezhnoy nauchnoy konferentsii studentov, aspirantov i molodykh uchenykh "Lomonosov-2013"], Moscow, MAKS Press.
2. *Black A., Chen S., Kumar Sh., Ostendorf M., Richard Ch., Sproat R., Yarowsky D.* (1999) Normalization of non-standard words, JHU99.
3. *Cherepanova O. D.* (2016) Verbalizing of English word combinations in Russian Text-to-Speech synthesis by means of orthographic transcription [Ozvuchivanie angloyazychnykh slovoupotrebleniy v sisteme russkoyazychnogo sinteza "tekst-rech" s pomoshchyu prakticheskoy transkriptsii], Online-materials of the conference "Dialogue-2016", available at: http://www.dialog-21.ru/media/3449/cherepanovaod.pdf.
4. *Khomitsevich O. G., Rybin S. V., Anichkin I. M.* (2013) Linguistic analysis in text normalization and disambiguation in Russian text-to-speech synthesis [Ispolzovanie lingvisticheskogo analiza dlya normalizatsii teksta i snyatia omonimii v sisteme russkoy rechi]. IzvestiyaD vysshikh uchebnykh zavedeniy. Priborostroyenie, No. 2, pp. 42–46.
5. *Krivnova O. F.* (1998) Automatic Text-to-Speech synthesis (second version with a women voice) [Avtomaticheskiy sintez russkoy rechi po proizvol'nomu tekstu (vtoraya versiya s zhenskim golosom)], Proc. Int. seminar in computational linguistics and its applications "Dialogue 1998" [Trudy mezhdunarodnogo seminara po komp'yuternoy lingvistike i yeyo prilozheniyam "Dialog 1998"], Moscow.
6. *Krivnova O. F.* (1999) Processing of acronyms in automatic Text-to-Speech synthesis [Obrabotka inizialnykh abbreviatur pri avtomaticheskom sinteze rechi], Proc. Int. seminar in computational linguistics and its applications "Dialogue 1999" [Trudy mezhdunarodnogo seminara po komp'yuternoy lingvistike i yeyo prilozheniyam "Dialog 1999"], Moscow.
7. *Olinsky C., Black A. W.* (2000) Non-standard word and homograph resolution for Asian language text analysis, INTERSPEECH, ISCA, pp. 733–736.
8. *Sproat R., Black A., Chen S., Kumar Sh., Ostendorf M., Richards Ch.* (2001) Normalization of non-standard words, Computer Speech and Language, Vol. 15, pp. 287–333.
9. *Rovinskaya M. M.* (1998) Recognition of full stop functions in automatic speech synthesis [Raspoznavanie funktsii upotrebleniya tochki pri avtomaticheskom sinteze rechi], undergraduate's thesis, MSU, Moscow.
10. *Xydas G., Karberis G., Kouroupertroglou G.* (2004) Text Normalization for the Pronunciation of Non-Standard Words in an Inflected Language, Proceedings of the 3rd Hellenic Conference on Artificial Intelligence (SETN04), Samos, Greece, May 5–8.