PROCESSING OF QUANTITATIVE EXPRESSIONS WITH UNITS OF MEASUREMENT IN SCIENTIFIC TEXTS AS APPLIED TO BELARUSIAN AND RUSSIAN TEXT-TO-SPEECH SYNTHESIS

Skopinava A. M. (skelena777@gmail.com),
Hetsevich Yu. S. (Yury.Hetsevich@gmail.com),
Lobanov B. M. (Lobanov@newman.bas-net.by)

United Institute of Informatics Problems of the NAS of Belarus, Minsk, Belarus

The article discusses problems of identification, analysis, classification (according to the International System of Units and separately according to word formation peculiarities), and processing of quantitative expressions (QE) with measurement units (MUs) as applied to text-to-speech synthesis by means of the linguistic processor NooJ[1] and specially collected legal, scientific and technical text corpora for the Belarusian and Russian languages. In addition to a general description of algorithms and resources for finding QE in Belarusian and Russian texts, the paper gives an overview of QE with MUs with regard to how their components could be written, i.e. digital descriptors, and MUs proper (five different types). It is shown that QE with MUs can get the correct intonation marking only after they are properly generated, i. e. expanded into orthographical words.

Key words: text-to-speech synthesis, NooJ, units of measurement, quantitative expressions, finite-state automata, generation of an orthographical text, identification, processing, intonation marking, Belarusian, Russian

Introduction

After the Belarusian and Russian NooJ modules [3] were obtained, it became possible to check and update experimental solutions to different linguistic tasks in application to text-to-speech synthesis [1, 2]. Synthesizers which use orthographic texts cope well with voicing orthographic words [7], but abbreviations, acronyms, numbers, symbols, etc. demand preprocessing into real words before they can be voiced.

The main purpose of this article is to describe approaches to identification and transformation of quantitative expressions (QE) with measurement units (MUs) into correct orthographic words in hand-crafted scientific, technical and legal text corpora for Belarusian and Russian; and to prove its importance for correct intonational marking of texts.

[1] http://www.nooj4nlp.net/pages/nooj.html

To give an example, Belarusian sequences like *123 мА* '123 mA' and *120 мА* '120 mA' have to be transformed by the synthesizers into sequences of words with intricate agreement: resp. *сто дваццаць тры міліампер**ы*** and *сто дваццаць міліампер*, because Belarusian (and Russian) numerals are declinable and can influence subsequent words (in our case measurement units), unlike English, where, e. g., a preposition before a numeral does not change anything in the voicing of MUs. For the present we deal with generating QE in the Nominative.

When dealing with QE with MUs, many difficulties arise. First, they are conditioned by a great variety of numeral quantifiers and names of units, both in writing and formation. Creating rules of complex expressions localization for all cases is practically impossible (that is exactly the reason why regular expressions are not the best way to obtain localization rules). In order to simplify this process, it is extremely important to use tools that allow users to easily modify previously-developed rules and add new ones. The international program NooJ is one such tool. It allows implementing sophisticated algorithms of searching for compound text fragments in Belarusian and Russian in the form of visual executable graphs.

Second, an expression with a MU is difficult to recognize and analyze (note a considerable number of digits, words with quantitative meaning with all their possible paradigmatic forms, names of metrological system units) without thoroughly prepared linguistic resources, i.e., dictionaries with all possible word forms, abbreviations, and rules for building derivative forms of measurement units. This is necessary, e. g., for proper treatment of expressions with units of length, written in various ways: *1 м (1 m), 31 метр (31 meters), 25 метраў (25 meters), 44 метры (44 meters)* [4].

Third, QE with MUs are language-dependent: in English *meter* and *mile* are abbreviated as *m*, while in Belarusian and Russian as м; even within largely similar Russian and Belarusian, names of measurement units differ in spelling — *гадзіна*, *час* 'hour'. Therefore, it is essential to make accurate provisions for each language.

Significant results have been achieved by European researchers and developers of the Quantalyze semantic annotation and search service[2], and Numeric Property Searching service in Derwent World Patents Index on STN[3]. However, language orientation is the reason why theoretical or practical results cannot be fully reusable for Belarusian or Russian. We view QE with MUs as combinations where each component requires a specific approach for successful identification.

Searching for and classifying QE with MUs according to the SI

In order to construct and test algorithms, four text corpora were formed for two domains: scientific, technical and legal (two for each language) (Fig. 1) [4]. According to the main graph (Fig. 2) of the obtained algorithms (for Belarusian and Russian they differ in some language-dependent subgraphs), any text fragment is initially checked in the 1st subgraph (Numeral Quantifier) if it has a compound numerical descriptor (Fig. 3).

---

[2] https://www.quantalyze.com/en/

[3] http://www.stn-international.com/numeric_property_searching.html

File Name	
Раздзел 21. Рух гужавых транспартных сродкаў, коннікаў і прагон жывёлы	186.2.1. 12 метраў для аўтамабіля, тралейбуса, прычэпа;
Раздзел 22. Карыстанне знешнімі светлавымі прыборамі і гукавымі сігналамі	186.2.2. 13,5 метра для аўтобуса з двума восямі, 15 метраў
Раздзел 23. Перавозка пасажыраў	для аўтобуса з больш чым двума восямі;
Раздзел 24. Перавозка грузаў	186.2.3. 18,75 метра для сучлененага аўтобуса, сучлененага
Раздзел 25. Буксіроўка механічных транспартных сродкаў	тралейбуса;
Раздзел 26. Асноўныя палажэнні аб допуску транспартных сродкаў да ўдзелу	
Раздзел 27. Абавязкі службовых і іншых асоб па забеспячэнні бяспекі дарожн	

a)

File Name	
Глава 10. Расположение транспортных средств на проезжей части дороги	89.2. автобусам и мотоциклам — не более 90 км/ч;
Глава 11. Скорость движения транспортных средств	89.3. автобусам, легковым и грузовым автомобилям при их
Глава 12. Обгон, встречный разъезд	движении с прицепом, грузовым автомобилям с технически
Глава 13. Проезд перекрестков	допустимой общей массой более 3,5 тонны на автомагистралях —
Глава 14. Пешеходные переходы и остановочные пункты маршрутных тран	не более 90 км/ч, на остальных дорогах — не более 70 км/ч;
Глава 15. Преимущество маршрутных транспортных средств	
Глава 16. Железнодорожные переезды	

b)

File Name	
Chapter 21. Traffic of animal-drawn vehicles, horseback riders and guiding	186.2.1. 12 meters for a motor vehicle, trolleybus, trailer;
Chapter 22. Use of external luminous and audible devices of vehicles	186.2.2. 13.5 meters for a bus with two axles, 15 meters
Chapter 23. Carriage of passengers	for a bus with more than two axles;
Chapter 24. Carriage of goods	186.2.3. 18.75 meters for an articulated bus, articulated
Chapter 25. Towing of power-driven vehicles	trolleybus;
Chapter 26. General provisions about admission of vehicles to participate i	

c)

Fig. 1. Fragments of legal text corpora for (a) Belarusian, (b) Russian, and (c) translated into English

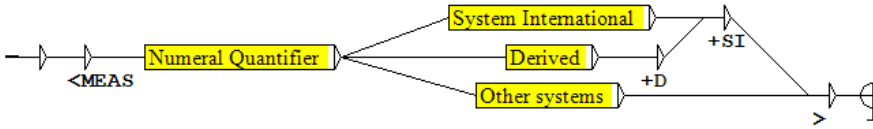


Fig. 2. The main graph of the algorithm for identification of QE with MUs

It should be noted that this subgraph works out not only for prime, decimal and fractional numbers in various forms of writing, but also for compound numerical combinations with exponential parts and periods. Some results of its work can be observed in the form of a concordance (Fig. 4). It should be emphasized that this subgraph is language-independent (Fig. 4c).

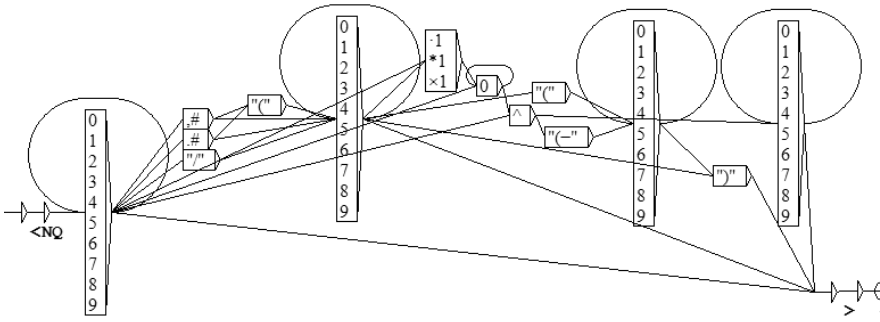


Fig. 3. The subgraph for identification of numbers and compound numerical combinations

Before	Seq.	After
электратэхнічнай камісіяй) IEC	60027	ужываецца пазначэнне Mbit
проста Mb). 1 мегабіт =	1000^2	біт = 10^6 біт = 1000000 біт
Mb). 1 мегабіт = 1000^2 біт =	10^6	біт = 1000000 біт. Дзесятков
Напрыклад: 1/6 = 0,166666... =	0,1(6)	; 1/7 = 0,1428571428... = 0,(14
= 0,1(6); 1/7 = 0,1428571428... =	0,(142857)	.

a)

Before	Seq.	After
автомагистралях - не более	110	км/ч, на
двумя осями; · -	18,75	метра для сочлененного
в среднем составляет	5·10^(−5)	Тл, а на
на экваторе (широта 0°) —	3,1·10^(−5)	Тл. 5. Ом — единица
бомбардировке Хиросимы: около	6·10^13	Дж. Энергия фотона
красного видимого света:	2,61·10^(−19)	Дж.

b)

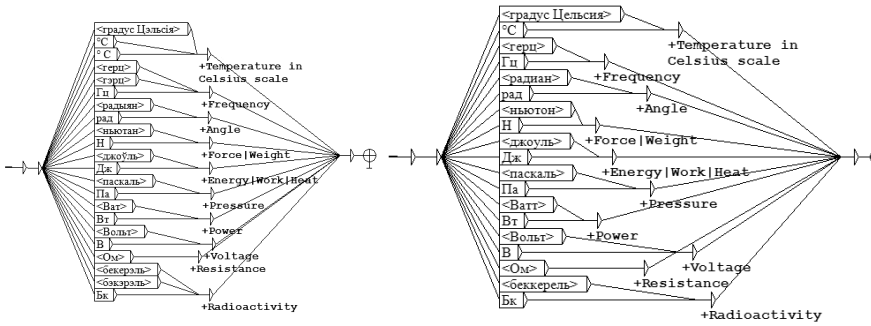
Before	Seq.	After
is equal to	6.24150974×10^18	eV (electronvolts). 1 joule
is equal to	2.3901×10^(−4)	kcal (thermochemical kiloc
defined as exactly	0.0254	m, and the
defined as exactly	453.59237	g. Also a
are equivalent to	1/100	. An integer such

c)

Fig. 4. Results of identifying complex numerical expressions in (a) Belarusian, (b) Russian, and (c) English texts

After the first subgraph has been processed, the algorithm proceeds to other subgraphs, which are connected to its output by means of respective transition lines. The subgraph *System International* identifies units according to the SI, e. g., *кілаграм* 'kilogram'; the subgraph *Derived* — SI derivatives (Fig. 5), such as *герц* 'hertz'; the subgraph *Other systems* — frequently used, but non-systemic units, such as *час* 'hour'. If any of the three subgraphs works out, the sequence of respective transition lines on the way to the main graph's output is indicated by markers. Let us draw up a list of some possible markers: *MEAS, MEAS+SI+…, MEAS+D+SI+…* . They correspond to the above-mentioned subgraphs' respective predestinations. Three dots in the last two markers can be replaced by special markers within a respective subgraph that works out. At the same time names of MUs (or their word forms) correspond to names of respective physical values (or their word forms). Take the word combination дадаць *3,3 молі* 'add 3,3 moles' as an example. The algorithm will recognize the following expression: *3,3 молі* '3,3 moles'. It will receive the following marker: *MEAS+SI+Amount of substance*. The marker enables one to identify exactly which subgraph works out and which unit of measurement is used. The code *MEAS* means that the expression *3,3 молі* '3,3 moles' contains a unit of measurement *молі* 'moles'. The code + *SI* informs

that the MU *молі* 'moles' belongs to the SI units. The code + *Amount of substance* means that *молі* 'moles' are used for measuring amounts of substances. The component *D* of the marker *MEAS+D+SI+…* requires the existence of the second distinct subgraph in order to separate expressions with MUs derived from the SI basic units, i. e., *degree Celsius, hertz, radian, newton, joule, pascal, watt, volt, ohm, becquerel*.



a) b)

Fig. 5. The subgraphs which identify expressions with SI-derived units for (a) Belarusian and (b) Russian

Such a flexible system of markers allows building search queries of different types: to find all expressions with MUs (Fig. 6); to find expressions without derived units *(<MEAS+SI-D>)* (Fig. 7); etc. Table 1 contains the search results in Fig. 6 and Fig. 7 translated into English and listed from top to bottom.

Before	Seq.	After
ашэнне –	1м/<MEAS+Length|Distance+SI>	(бач.),
2-30 кэв,	0,1 Гц/<MEAS+Frequency+D+SI>	-300 кгц
вую масу	8 т/<MEAS+Mass>	, вывед
Зямлі. У	2005 г./<MEAS+Time>	Іран зд
н і ўхілам	74 градусы/<MEAS+Angle>	. Затым

a)

Before	Seq.	After
ература	109 К/<MEAS+Thermodynamic temperature+...	В эт
стью ок.	200 000 л/<MEAS+Volume>	. Желе
а спустя	33 года/<MEAS+Time>	– и ег
вышало	5°/<MEAS+Angle>	), а пот
ре выше	600° С/<MEAS+Temperature in Celsius scal...	, а хал

b)

Fig. 6. Results of identification of QE with MUs in (a) Belarusian and (b) Russian

цыі на ўзроўні 1–	10 м	. Такая дэта
иаса перавышае	3600 кг	. Разліковы
масай меней за	10 кг	, а праз 10–2
– масай парадку	1 кг	, якія змогуц
ала парадку 150–	500 метраў	. У 70–80-х г

a)

– ОД % (вид.),	0,1 К	(ИК), 1д
разрешение –	1м	(вид.), 5
тупая 19 апреля	1904 с	больши
через каждые	30 секунд	трех бро
рез 30, а через	3 секунды	, то прям

b)

Fig. 7. Results of identification of QE with only SI-units of measurement on the request <MEAS+SI-D> in (a) Belarusian and (b) Russian

Table 1. Search results in Fig. 6, Fig. 7 translated into English

	Figure 6	Figure 7
a)	1m <MEAS+Length|Distance+SI> 0,1Hz <MEAS+Frequency+D+SI> 8 t <MEAS+Mass> year 2005 <MEAS+Time> 74 degrees <MEAS+Angle>	10 m 3600 kg 10 kg 1 kg 500 metres
b)	109 K <MEAS+Thermodynamic temperature+SI> 200 000 l <MEAS+Volume> 33 years <MEAS+Time> 5° <MEAS+Angle> 600°C <MEAS+Temperature in Celsius scale+D+SI>	0,1 K 1m 1904 30 seconds 3 seconds

Identification of MUs with metrological prefixes

First of all, the authors created necessary linguistic dictionaries *S* for Belarusian and Russian (Fig. 8). They contain some basic stems of MUs — complete nouns and their abbreviations. Each stem is marked by a respective attribute: either *Base* or *Mbase*. In addition, descriptions of full stems include indicators of respective inflectional classes. The dictionary *S* is obviously a language-dependent linguistic resource, unlike algorithms for identification of MUs with metrological prefixes, which are implemented as language-independent components. The next step was to develop language-dependent linguistic resources (*Fsubmultiple*, *Fmultiple*, *Ssubmultiple*, *Smultiple*) (Fig. 9). For MUs-formation either *multiple* or *submultiple* prefixes can be used. Besides, they can take a shortened (*S*-) or full (*F*-) form (Fig. 10) [5].

```
г,ABBREVIATION+Mbase
га,ABBREVIATION+Mbase
гг,ABBREVIATION+Mbase
гектар,NOUN+FLX=ГЕКТАР+s5+UNAMB+Base
герц,NOUN+FLX=АМПЕР+s2+UNAMB+Base
год,NOUN+FLX=ГОД+sN+UNAMB+Base
град,ABBREVIATION+Mbase
грам,NOUN+FLX=ГРАМ+s3+UNAMB+Base
```

a)

```
ампер,NOUN+FLX=АЛТЫН+s4+UNAMB+Base
А,ABBREVIATION+Mbase
байт,NOUN+FLX=АБАЖУР+s2+UNAMB+Base
бит,NOUN+FLX=АБАЖУР+s2+UNAMB+Base
Б,ABBREVIATION+Mbase
ватт,NOUN+FLX=АЛТЫН+s2+UNAMB+Base
Вт,ABBREVIATION+Mbase
вольт,NOUN+FLX=АЛТЫН+s2+UNAMB+Base
```

b)

Fig. 8. Dictionary resources of basic MUs' stems for (a) Belarusian and (b) *Russian*

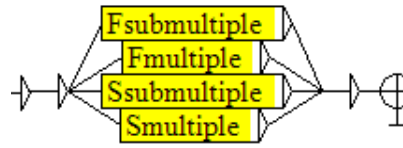


Fig. 9. Classifying metrological prefixes using a NooJ finite-state automaton

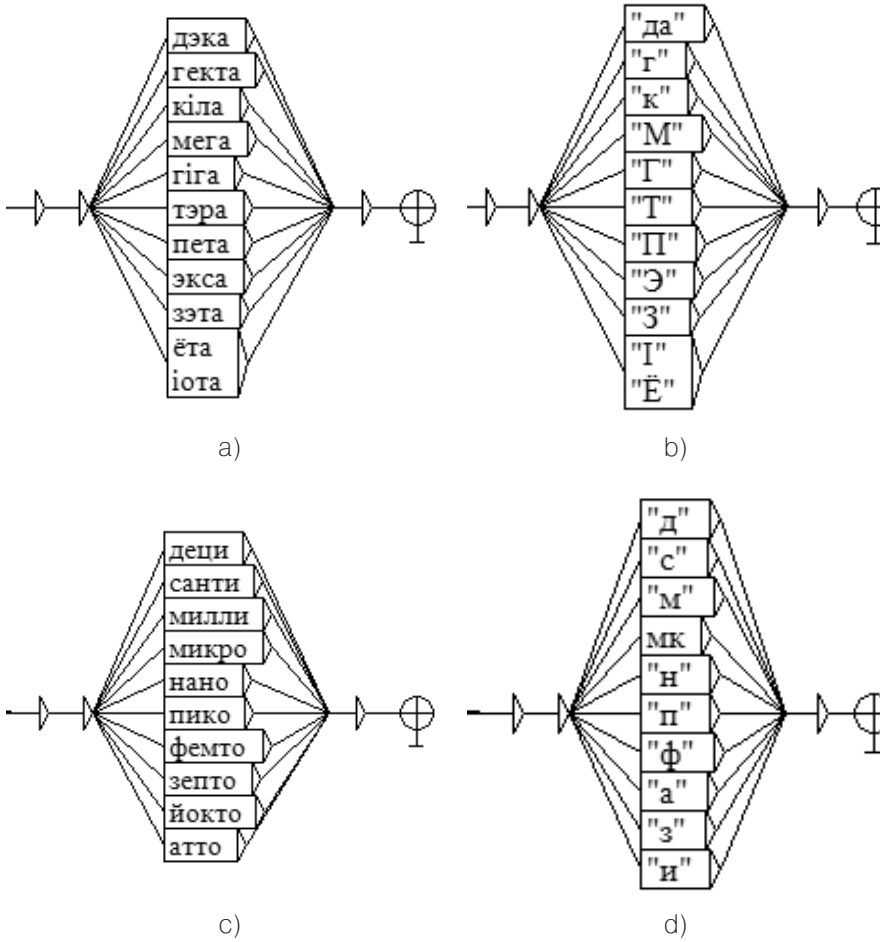


Fig. 10. Graphs for identification of (a) full-stem and (b) shortened-stem multiple prefixes for Belarusian, and (c) full-stem and (d) shortened-stem submultiple prefixes for Russian

The basic principle for the components became the following word-formative classification of MUs:

- MUs with full-form stems and without prefixes (*метр* 'meter', *Герц* 'hertz', *Ом* 'ohm');
- MUs with shortened stems and without prefixes (*Дж* 'J', *га* 'ha');
- MUs with full-form stems and full-form prefixes (*нанафарады* 'nanofarads', *міліампер* 'milliampere');
- MUs with full-form stems and shortened prefixes (*кБайт* 'Kbyte');
- MUs with shortened stems and shortened prefixes (*км* 'km', *дл* 'dL', *гПа* 'hPa').

Depending on word formation peculiarities, 4 morphological language-independent grammars M1-M4 (algorithms) were obtained. They use the dictionary *S* and

linguistic resources *Fsubmultiple, Fmultiple, Ssubmultiple, Smultiple*. For example, the morphological grammar M2 identifies MUs which are formed with the help of multiple and/or submultiple full-form prefixes (Fig. 11).

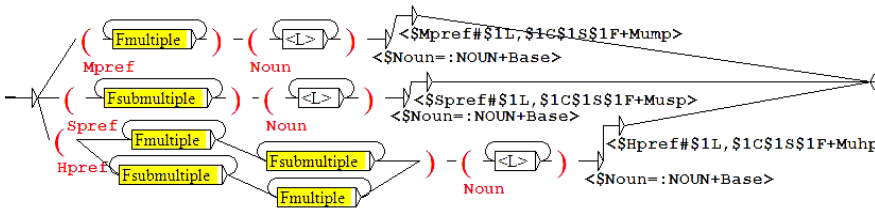


Fig. 11. The morphological grammar M2 which identifies MUs with full-form stems and full-form multiple and/or submultiple prefixes

As a result of its work, MUs may be given one of the following markers:

- *Mump* means that identified MUs have multiple prefixes;
- *Musp* implies that identified MUs have submultiple prefixes;
- *Muhp* denotes MUs which have several prefixes, e. g.: *мікрамегафарад* 'micro-megafarad'. According to the SI, such a way of formation is not common among MUs, so such words require a specified marker, so later they can be extracted from text within a list of mistakes.

Fig. 12 represents operation examples of the above-described morphological component. Note that the obtained morphological components enable the identified MU to inherit all grammatical and inflectional characteristics of initial words. E. g., the word *дэкалітрамі* 'deciliters' (in the Instrumental case) will remain the noun with all its inflectional endings and grammatical features, though the resource dictionary *S* does not contain it (Fig. 13).

Before	Seq.	After
несколько сотен	километров	. Первый вариант
ıгать нескольких	килограммов	. Куски брони пора
цностью в сотни	мегаватт	. Проблема в том
е десятки тысяч	мегагерц	, что соответствуе
іх дисках тысячи	гигабайт	информации, тре
лучения порядка	мегаджоуля	(106 Дж) и кпд

a)

Before	Seq.	After
пашырыць да 2	тэрабайт	! Паскаральнік
;еткі «усяго» 50	кілават	. Астатнія канст
кунд. Таму 425	кілаграмаў	рабочага цела
тую ж мэта 300	кілаграмаў	аргону штогод,
(магутнасцю да	мегавата	) (ілюстрацыя А
ўстаноўкі ў 200	мегават	. Шмат. Але зат

b)

Fig. 12. The resulting concordance of full-stem MUs on the request <NOUN+Mump> as applied to (a) Belarusian and (b) Russian

дэкалітр,NOUN+Meaning=Common
+Animation=Inanimate
+Case=Instrumental
+Gender=Masculine+Number=Plural
+s2+Meas=Base+Mump

a)

наносекунда,NOUN+ProperCommon=Common
+Gender=Feminine+Animation=Inanimate
+Case=Instrumental+Number=Plural
+s4+Meas=Base+Musp

b)

Fig. 13. Examples of annotated word forms for (a) Belarusian and (b) Russian

Finally, the algorithm proceeds to the syntactic grammar *S1* (Fig. 14). It accumulates all the markers from the text *T*, placed by means of the dictionary *S* and morphological grammars *M1-M4*. It works out only for QE with MUs (numerical descriptors in front of them). Numerical descriptors are identified by the inbuilt syntactic component *S1*. Each QE with MUs receives the marker <*MUEXPR*>. It enables users to create concordances of QE with MUs (Fig. 15).

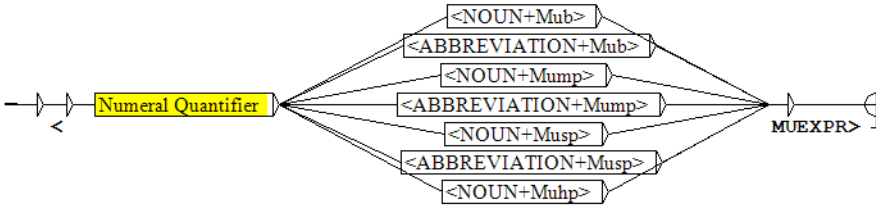


Fig. 14. The main syntactic component S2, which identifies QE with MUs

Before	Seq.	After
адзінку масы - грам (0,001 кг).	31 мкТл	(3,1×10^(-5) Тл) - напружанасць магні
звычайна вар'іруецца зблізку	2,4 мЗв	у год. 1 Н ёсць
на апору з сілай	9.81 Н	. Прыбліжэнне, што 1 кг адпавядае
дамі або нанафарадамі (пішуць	60 000 пф	, а не 60 нф; 2 000 мкф
ёмістасць шара з радыусам	1 сантыметр	, змешчанага ў вакуум. 1 сантыметр
Mbit(альбо проста Mb).	1 мегабіт	= 1000^2 біт = 10^6 біт = 1000000 біт.
святло ў вакууме за (	1 / 299 792 458) секунды	. Метр быў упершыню ўведзены
ны дыяпазон - 40 кэв-3 МЭВ, 2-	200 МЭВ	, 2-30 кэв, 0,1 Гц-300 кгц, 0-50 кгц
ай трубцы тэлевізара - парадку	20 кілаэлектронвольт	. Энэргіі касьмічных прамянёў - ад
энергіі касьмічных прамянёў - ад	1 мегаэлектронвольта	да1000 тэралектронвольтаў.

a)

Before	Seq.	After
организм ток не превышал	1 мА	. На человека токи статического
могут сказать «файл в	100 килобайт	»). При обозначении скоростей тел
ротивление величиной от 1 до	100 МОм	, чтобы протекающий через челове
икромегафарад пикотеравольт	13 йоттайоктограммов	Каждая строка содержит информа
до 64 Мбит/с) и	137,4 МГц	(метровый диапазон, формат АРТ
евонширский изумруд» массой	1383,95 каратов	. Изумруды выращивают искусств
время жизни мюонов - около	2.2 мкс	- осложняет задачу создания мюон
сса которой оказалась равной	22 фемтограммам	(1 фг = 1•10^(-15) г). . Мюоны, как
то они оказались равными:	8.1·10^21 Дж	(уменьшение массы ледников на
: - высота 670 км - наклонение	98,00 град	. Срок активного существования 1 г

b)

Fig. 15. Some results of identification of QE with MUs after processing (a) Belarusian and (b) Russian texts by means of the obtained morphological M1–M4 and syntactic S1 grammars

Generation of orthographical words from QE with MUs

In text-to-speech synthesis tasks it is important to develop algorithms not only for identification of definite expressions but also for their processing and transformation into orthographical word sequences. With this aim, grammars in the form of visual finite-state automata for Belarusian and Russian were worked out. As a result, for each language a ramified algorithmic complex of 21 graphs and subgraphs was obtained. Fig. 16 represents the structure of the main graph. Since QE with MUs consist of 2 components (numbers and nouns), it is required to work out separate graphs for their generation. This algorithm contains of graphs of 2 types. Those, which have names starting with *a_*, generate numbers from 0 to 999,999,999,999. All the rest, in particular the ones with *b_*, are intended for generation of nouns which denote MU.

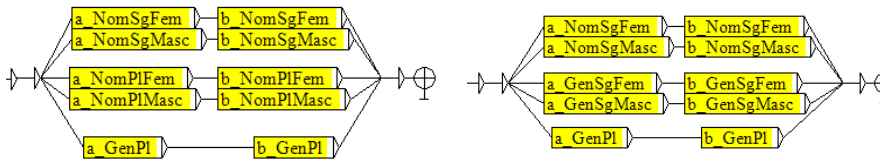


Fig. 16. The main graph of the algorithm which generates orthographical words from QE with MUs for (a) Belarusian and (b) Russian

QE with MUs pass from input to output by means of one of 5 ways in accordance with peculiarities of the inflection of nouns after numerals, in particular for the first 3 ways:

1. After number *1* (including numbers with *1* as a final digit) nouns take endings of the Nominative Singular *(NomSg)*. QE will proceed to one of the top branches, depending on the gender of nouns, in particular *Masculine (Masc)* or *Feminine (Fem)*.
2. After numbers *2, 3, 4* (including numbers with *2, 3* or *4* as a final digit) nouns take the Nominative plural *(NomPl)* in Belarusian, whereas in the Russian these numbers require nouns in the Genetive singular *(GenSg)*. Depending on the gender, QE will move to branches 3 or 4.
3. Numbers from 5 to 19 and round numbers (including numbers with them as final digits) require nouns in the Genitive plural *(GenPl)* in both languages. QEs will follow the 5th branch.

As an example, let us stop on the first branching of the algorithm, in particular the graph *a_GenPl* (Fig. 17). It generates any whole number from 0 to 999,999,999,999, which demands the Genitive plural form.

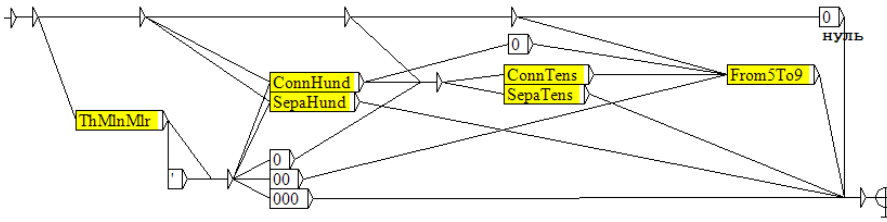


Fig. 17. The subgraph a_GenPl for Belarusian

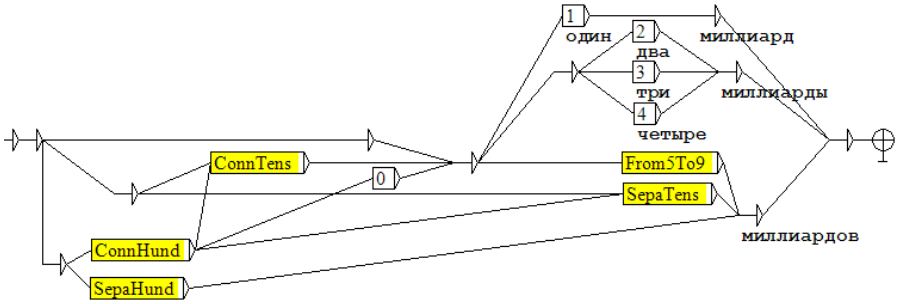
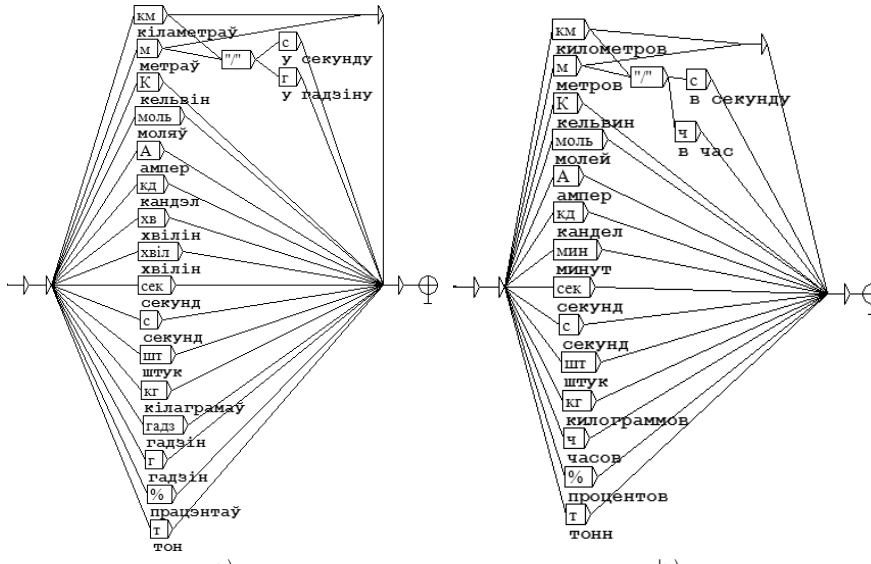


Fig. 18. The subgraph Mlr for Russian

The structure of the algorithm for generation of numbers resembles Russian dolls. At first the graph for numbers of the first triad (from 0 to 999) was obtained. It includes the inbuilt subgraph *ThMlnMlr* for the class of thousands or numbers with 2 triads (from 1,000 to 999,999). Inside of this subgraph the other one (*MlnMlr)* was placed for the class of millions or three-triads numbers; at last, the subgraph *Mlr* (Fig.18) for the class of billions or numbers with 4 triads (from 1,000,000,000 to 999,999,999,999) was worked out. Depending on research goals, the algorithm can be expanded by further triads. After generating numbers the algorithm proceeds to processing nouns which denote MUs. Concerning the last branch of the algorithm, it happens with the help of the graph *b_GenPl* (Fig. 19).

For the present, this subgraph can generate basic SI units and some frequently used ones. Thanks to the visuality of finite-state automata, the algorithm can be easily and rapidly improved by adding more MUs. In order to add a new unit, three case endings (mind the gender and number) should be added to the respective graphs. Variations of written forms should also be taken into account. For example, for the noun градус (shortened гр, ° — three variants; in English *degree*, shortened *deg*, °), one should add 3 respective word forms (градус, градусы, градусаў for Belarusian; and градус, градусы, градусов for Russian) for each variant into the following graphs: *b_NomSgMasc, b_NomPlMasc, b_GenPl* for Belarusian; *b_NomSgMasc, b_GenSgMasc, b_GenPl* for Russian.



a) b)

Fig. 19. The subgraph b_GenPl for (a) Belarusian and (b) Russian

Thus, language-dependent complexes of grammars for generation of orthographic words from QE with MUs have been obtained. Fig. 20 demonstrates some results of their operation.

700001г/семсот тысяч адна гадзіна
0 с/нуль секунд
777'700т/семсот семдзесят сем тысяч семсот тон
888'808хв/восемсот восемдзесят восем тысяч восемсот восем хвілін
2220020 хвіл/два мільёны дзвесце дваццаць тысяч дваццаць хвілін
444'014моль/чатырыста сорак чатыры тысячы чатырнаццаць моляў

a)

10120202 мин/десять миллионов сто двадцать тысяч двести две минуты
70000000071 шт/семьдесят миллиардов семьдесят одна штука
81234999 А/восемьдесят один миллион двести тридцать четыре тысячи девятьсот девяноста девять ампер
8600км/ч/восемь тысяч шестьсот километров в час
90673 м/с/девяноста тысяч шестьсот семьдесят три метра в секунду

b)

Fig. 20. Generation of orthographic words from QE with MU for (a) Belarusian and (b) Russian with the help of the developed algorithms

Variety of ways to express QE with MU in Belarusian and Russian texts

Since the practical goal is to identify MUs and generate expressions with them, a question inevitably arises: which ways of written forms should be taken into consideration? Thus, it is required to make a certain sample of QE in order to cover all the variety of ways of their expression in writing.

In order to attain this aim, a hand-crafted corpus (about 1 mln words) of Russian scientific texts (in the format NooJ) was used. With the help of statistical computations 8 texts with highest amounts of MUs were selected, in particular texts on: *military equipment (ME), geography (Geo), history (His), space travel science (STS), physics (Phy), mineralogy (Min), transportation communications (TC), botany (Bot)*. After that a special formula in the form of a regular expression was created (Fig. 21), starting with the simple (<WF><NB>).

Further the right and left contexts were analyzed, so the formula was constantly updated. Table 2 demonstrates what happens when the formula is applied to the selected texts, whereas Table 3 contains generalized data for them.

```
(от <NB> до <NB>)|(от <NB> до <NB><NB>)|(от <NB><NB> до <NB>)|(от <NB>,<NB> до <NB>,<NB>)|
(от <NB>,<NB> до <NB>)|(от <NB>.<NB> до <NB>.<NB>)|(от <NB> до <NB>,<NB>)|
(от <NB>–<NB> до <NB>–<NB>)|(от <NB>×<NB>-<NB> до <NB>×<NB>-<NB>)|
(от <NB>×<NB>–<NB> до <NB>×<NB>–<NB>)|(от <NB> до почти <NB>)|(<WF>|"~"|"=")(<NB>|(<NB><NB>)|
(<NB><NB><NB><NB><NB>)|(<NB><NB><NB><NB><NB>)|(<NB><NB><NB>)|(<NB>×<NB>)|(<NB>,<NB>)|
(<NB>–<NB>)|(<NB>—<NB>)|(<NB>-<NB>)|(<NB>.<NB>-<NB>.<NB>)|(<NB>.<NB>)|(<NB>–<NB>,<NB>)|
(<NB>.<NB>.<NB>)|(<NB>,<NB>–<NB>,<NB>)|(<NB>,<NB>–<NB>)|(<NB>,<NB>-<NB>)|(<NB>-<NB>,<NB>)|
(<NB>"/"<NB>)|(<NB>":"<NB>)|(<NB>":"<NB>-<NB>)|(<NB>":"<NB>":"<NB>)|(<NB>,<NB>×<NB>–<NB>)|
(<NB>×<NB>–<NB>)|(<NB>´<NB>)|(<NB>,<NB>´<NB>,<NB>)|(<NB>´<NB>,<NB>)|(<NB>·<NB>)|
(<NB>,<NB>·<NB>–<NB>)| (<NB>.<NB>•<NB>)|(<NB>•<NB>–<NB>)|(<NB>.<NB>•<NB>–<NB>)|
(<NB>.<NB>·<NB>)|(<NB>,<NB>·<NB>)|(<NB>·<NB>–<NB>)|(">"<NB>)|(<NB>">>"<NB>)|
(<NB>|(<NB><NB>)|(<NB>,<NB>)|(<NB>–<NB>)|(<NB>—<NB>)|(<NB>-<NB>)|(<NB>.<NB>)|
(<NB>,<NB>–<NB>,<NB>)|(<NB>,<NB>–<NB>)|(<NB>-<NB>,<NB>)|(<NB>":"<NB>)|
(<NB>,<NB>×<NB>–<NB>)|(<NB>,<NB>·<NB>)|(<NB>·<NB>–<NB>)|(<NB>x<NB>)|(<NB>x<NB>')|
(<NB>x<NB>x<NB>)|(<NB>.<NB>x<NB>.<NB>)|(<NB>,<NB>×<NB>)|(<NB>¸<NB><NB>)|
```

Fig. 21. The regular expression for a search for QE with MUs

Table 2. Excerpts: variety of QE, revealed with the help of the obtained formula

Formula's constituents	Examples of QE, found with the help of a certain constituent
(от <NB> до <NB>)	находится в пределах от 40 до 347 г/л. Несмотря лежит в диапазоне от 2000 до 3000 А, и следовательно удаленной на расстояние от 160 до 640 км от восточного мере удаления "Пионера-10" от 40 до 60 а.е. величина
(от <NB> до <NB><NB>)	находятся в диапазоне от 4000 до 14 000 МГц и даже вариации) с периодами от 10 до 10 000 лет, сосредоточенными в
(от <NB><NB> до <NB>)	см (частотный интервал от 600 000 до 1000 МГц) относится к
(от <NB>,<NB> до <NB>,<NB>)	семь камней массой от 94,45 до 4,39 кар каждый. Кроме между рельсами пути от 1,52 до 1,68 м) типичны для кристалла со скоростями от 0,01 до 0,001 скорости света; их
(от <NB>,<NB> до <NB>)	видов токсичность составляет от 0,01 до 1 %. Практически это означает
(от <NB>.<NB> до <NB>.<NB>)	диапазоне длин волн от 1.3 до 1.7 мкм. На подложке

Formula's constituents	Examples of QE, found with the help of a certain constituent
(от <NB> до <NB>, <NB>)	отношениях ионных радиусов от 1 до 0,732 (рис. 4,а). При С-диапазоне и от 12 до 12,7 ГГц в Q
(от <NB> – <NB> до <NB> – <NB>)	выемчатые. Их длина от 1–2 до 30–40 см. Самые длинные длиной волны l от 10–3 до 10–8 м. Этот диапазон
(от <NB> ×<NB> – <NB> до <NB> × <NB> – <NB>)	с удельным сопротивлением от 5×10-8 до 8×10-5 Ом×м. Композиционные
(от <NB> × <NB> – <NB> до <NB> × <NB> – <NB>)	в разных материалах: от 3×10–6 до 2×10–5 см. Магнитный поток
(от <NB> до почти <NB>)	током (при этом от 50 до почти 100 % его энергии превращается
(<WF>|»~»|»=»)<NB>	Мировом Океане составляет около 550 млрд. тонн в излучения с l ~ 10 Å не существует до цели L = 1000 км, получим ограничение затем разгоняются до энергии 5 МэВ на линейном крупного «суперматерика» Го... Около 160 млн. лет назад
(<WF>|»~»|»=») (<NB><NB>)	отравлений растениями страдают примерно 15 000 человек. Для домашних ядерных взрывов суммарной силой 10 000 Мт в центральных и в среднем на 386 063 км от центра В 1990 она насчитывала приблизительно 900 000 верующих, в основном установки. В 9.50 на высоте 15 800 м Волков - первым в

Table 3. Distribution of various ways of expression of QE in different texts

Text	*Phy*	*STS*	*Geo*	*ME*	*Min*	*Bot*	*TC*	*His*
Number of variations, *a*	51	23	22	19	18	16	14	9
Number of numeral expressions, *b*	2841	2245	2765	9961	3668	1407	2066	4198

Intonation marking in sentences which contain QE with MUs

Text-to-speech synthesis requires an automatic procedure of building current contours of melody, sound intensity, phoneme and pause duration, which is based on the analysis of certain properties of sentences according to rule-based prosodic marking. Prosodic marking of sentences implies their division into syntagmas, marking emphatically highlighted words, indicating syntagmas with accent units, and creating a melodic contour of each syntagma in accordance with certain rules. Solutions to these problems by means of in-depth syntactic analysis are thoroughly discussed in [6, 8]. Texts, when being synthesized, are first reduced to a normalized orthographic form. Next they undergo a complete syntactic analysis, performed by the parser ЭТАП-3 'ETAP-3'. The parser (1) divides texts into separate sentences; (2) for each sentence it builds treelike syntactic structures; (3) using special rules, which can be applied to ready syntactic structures, it sets boundaries among speech syntagmas

and emphatically highlighted components. The system Мультыфон 'Multiphone' [7] processes this information and, depending on syntactic types, determines a melodic contour and duration of pauses between syntagmas. Prosodic and intonation marking of sentences which contain QE with MUs can be carried out by the method proposed in the [6, 8]. However, before syntactic analysis of such sentences it is required to generate QE into orthographic words.

Indeed, Fig. 22 gives an example of syntactic analysis of the following sentence: *Расстояние до Марса 55764878 км* 'The distance to Mars is 55764878 km'. As a result of processing this sentence by the system ETAP-3 in accordance with the rules discussed in [6], at the output of the system the following information for synthesis is received: *Расстоя́ние до <EMP t="*4">Ма́рса</EMP> 55764878 км*. Thus, the system suggests no additional partitioning of this sentence into syntagmas. After *55764878 км* '55764878 km' is generated into orthographical words, we have the following result: *Расстояние до Марса пятьдесят пять миллионов семьсот шестьдесят четыре тысячи восемьсот семьдесят восемь километров* 'The distance to Mars is fifty-five million seven hundred sixty-four thousand eight hundred seventy-eight kilometers'. The syntactic analysis of this sentence can be observed in Fig. 23.

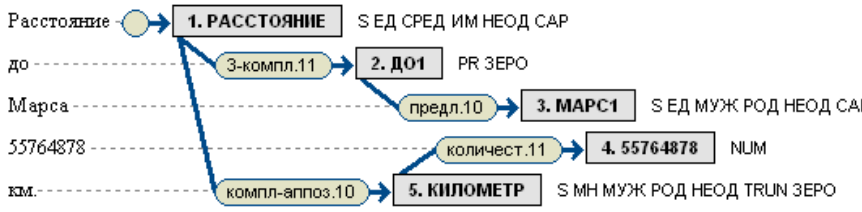


Fig. 22. Syntactic analysis of the sentence *Расстояние до Марса 55 764 878 км*

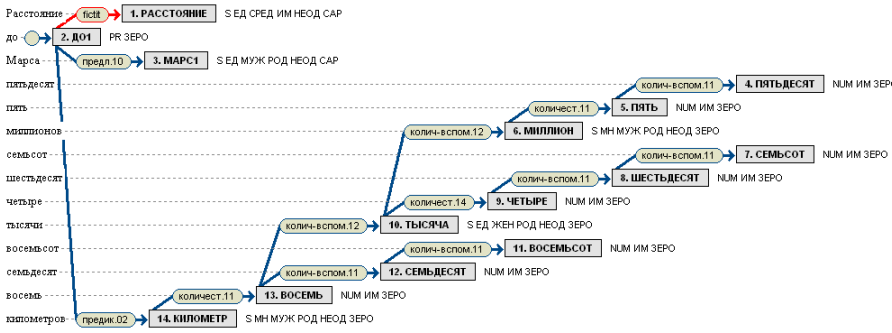


Fig. 23. The syntactic analysis of the sentence *Расстояние до Марса пятьдесят пять миллионов семьсот шестьдесят четыре тысячи восемьсот семьдесят восемь километров*

At the output of the system ETAP-3 the following information for the text-to-speech synthesizer MULTIPHONE is obtained: *Расстоя`ние <EMPt=”*16”> до </EMP> <EMPt=»*4»> Ма`рса</EMP> пятьдеся`т пя`ть <EMPt=»*4»> миллио`нов </EMP> семьсо`т шестьдеся`т четы`ре <EMPt=”*4”> ты`сячи </EMP> восемьсо`т се`мьдесят во`семь <EMPt=»*4»> киломе`тров </EMP>*. According to this information the Multiphone forms 4 syntagmas with melodic contours C01, C3, C3_1, P4 (emphatically highlighted words are indicated with the «+» sign).

1 *C01* *расстоя+ние/доЪма+рса/*
2 *C3* *пядеся=т пя+ть/миллио+нов/*
3 *C3_1* *семьсо=т шездеся=т четы=ре ты+сячи/*
4 *P4* *восемьсо=т се=мьдесят во=семь киломе+тров/*

Fig. 24 demonstrates another example of syntactic analysis, in particular for the following sentence: *Производительность компьютера 65,5 Мбит/с была достигнута* “The computer performance 65,5 Mbit/s was achieved”. The algorithm identifies *65,5 Мбит/с* ‘65,5 Mbit/s’. After processing into orthographical words it gets the following form: *Производительность компьютера шестьдесят пять целых и пять десятых мегабит в секунду была достигнута* ‘The computer performance sixty-five point five megabits per second was achieved’.

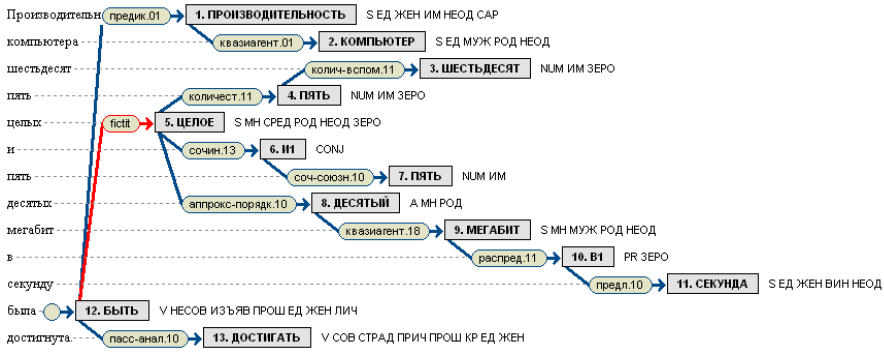
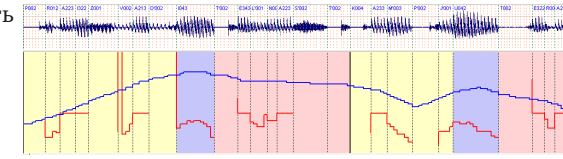


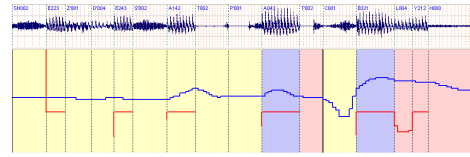
Fig. 24. The syntactic analysis of the sentence *Производительность компьютера шестьдесят пять целых и пять десятых мегабит в секунду была достигнута*

According to the data obtained by the ETAP-3, the MULTIPHONE forms 4 syntagmas (Fig. 25):

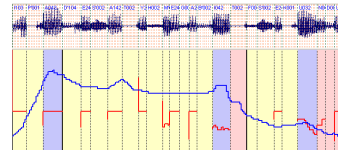
1 C4 производи+тельность компью+тера /



2 C1 шестьдеся=т пя+ть/ це+лых/



3 C3 и= пя+ть/деся=тых мегаби+т/вЪсеку+нду/



4 P4 была= дости+гнута/

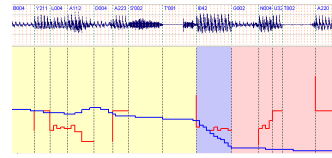


Fig. 25. Syntagmas and melodic contours for *«Производительность компьютера шестьдесят пять целых и пять десятых мегабит в секунду была достигнута»*

Conclusion

It can be concluded that the main goal of this research — to develop appropriate algorithms which *identify* quantitative expressions with various MUs and *generate orthographic texts* for the Belarusian and Russian languages for scientific, technical and legal text corpora — has been achieved. The results can be applied in any branches of science connected with information retrieval systems and text-to-speech synthesis. The resulting algorithms are created in the form of finite-state automata through a set of syntactic grammars within the powerful linguistic processor NooJ, which helps to build up formal grammars without requirements for special knowledge of programming. The automata demonstrate how the algorithms work and indicate how they can be further updated in order to improve their accuracy. Future work includes:

- disambiguation, e. g., in such cases when algorithms "confuse" some units (the same initial letter г for год 'year', грам 'gram', гадзі́на 'hour';
- developing algorithms that will identify numeral quantifiers expressed not only by numbers (mathematical objects), but also by numerals (parts of speech).

References

1. *Hetsevich Yu. S., Hetsevich S. A.* (2012), Overview of Belarusian and Russian dictionaries and their adaptation for NooJ, Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the NooJ 2011, Dubrovnik, pp. 29–40.

2. *Hetsevich Yu. S., Hetsevich S. A., Lobanov B. M.* (2012), Belarusian and Russian linguistic modules processing for the system NooJ as applied to text-to-speech synthesis, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012" [Komp'juternaja Lingvistika i Intellektual'nye Tehnologii: po Materialam Mezhdunarodnoj Konferentsii "Dialog 2012"], Bekasovo, pp. 198–212.
3. *Hetsevich Yu. S., Hetsevich S. A., Lobanov B. M., Yakubovich Ya.* (2012), Belarusian module for NooJ, available at: http://www.nooj4nlp.net/pages/belarusian.html
4. *Hetsevich Yu. S., Skopinava A. M.* (2012), Identification of Expressions with Units of Measurement in Scientific, Technical and Legal Texts in Belarusian and Russian [Ідэнтыфікацыя выразаў з адзінкамі вымярэння ў навукова-тэхнічных і прававых тэкстах на беларускай і рускай мовах], Development of information and the state system of scientific and technical information (DISTI-2012): Reports of the XI International Conference [Razvitie Informatizatsii i Gosudarstvennoj Sistemy Nauchno-Tehnicheskoj Informatsii (RINTI-2012): Doklady XI Mezhdunarodnoj Konferentsii], Minsk, pp. 260–265.
5. *Hetsevich Yu. S., Skopinava A. M.* (2013), Components for Identification of Quantitive Expressions with Measurement Units in Belarusian and Russian Texts [Кампаненты ідэнтыфікацыі колькасных выразаў з адзінкамі вымярэння ў тэкстах на беларускай і рускай мовах], Open Semantic Technologies for Intelligent Systems (OSTIS–2013): Proceedings of the III International scientific and technical conference [Otkrytye Semanticheskie Tehnologii Proektirovanija Intellektual'nyh Sistem (OSTIS–2013): Materialy III Mezhdunarodnoj Nauchno-Tehnicheskoj Konferentsii], Minsk, pp. 319–328.
6. *Iomdin L. L., Lobanov B. M., Hetsevich Yu. S.* (2011), The talking ETAP. Using the ETAP parser in Russian speech synthesis [Govorjashhij "ÈTAP". Opyt Ispol›zovanija Sintaksicheskogo Analizatora Sistemy ÈTAP v Russkom Rechevom Sinteze], Proceedings of the International Conference "Computational Linguistics and Intellectual Technologies" (Dialog'2011) [Trudy Mezhdunarodnoj Konferentsii "Komp›juternaja Lingvistika i Intellektual›nye Tehnologii" (Dialog'2011)], Bekasovo, pp. 269–279.
7. *Lobanov B. M., Tsirulnik L. I.* (2008), Computer speech synthesis and cloning [Komp›juternyj sintez i klonirovanie rechi], Belarusian Science [Belorusskaja Nauka] Publ., Minsk.
8. *Lobanov B. M., Iomdin L. L.* (2009), Syntactic Correlates of Prosodically Marked Elements of the Sentence and their Role in the Tasks of Text-To-Speech Synthesis [Sintaksicheskie Korreljaty Prosodicheski markirovannyh èlementov predlozhenija i ih rol› v zadachah sinteza rechi po tekstu], Proceedings of the International Conference "Computational Linguistics and Intellectual Technologies" (Dialog'2009) [Trudy Mezhdunarodnoj Konferentsii "Komp›juternaja Lingvistika i Intellektual›nye Tehnologii" (Dialog'2009)], Bekasovo, pp. 339–348.