# Processing of Quantitative Expressions with Measurement Units in the Nominative, Genitive, and Accusative Cases for Belarusian and Russian

Yury Hetsevich and Alena Skopinava

The United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, Belarus
{yury.hetsevich,skelena777}@gmail.com

**Abstract.** This paper outlines an approach to the stage-by-stage solution of the computer-linguistic problem of the processing of quantitative expressions with measurement units by means of the linguistic processor NooJ. The focus is put on the nominative, genitive, and accusative cases for Belarusian and Russian. The paper gives a general analysis of the problem providing examples not only for Belarusian and Russian, but also for English.

**Keywords:** NooJ, text-to-speech synthesis, text processing, Belarusian, Russian, quantitative expressions, measurement units

## 1 Introduction

In order to make text interfaces more “natural”, systems of human computer interaction should be able to voice electronic texts. High-quality text-to-speech synthesis cannot be achieved without excellent performance of text processing. There are plenty of objects in texts which require a specific way of treatment: tables, formulas, program codes, etc. This research concentrates on the processing of quantitative expressions with measurement units (QEMUs).

Let us take the following sentence in Belarusian as an example: *Маса Зямлі складае* $5.9736 \times 10^{24}$ *кг* ‘The mass of the Earth is equal to $5.9736{\times}10^{24}$ kg’. Obviously, there is one quantitative expression with a measurement unit, namely $5.9736{\times}10^{24}$ кг ‘$5.9736{\times}10^{24}$ kg’. The purpose is to develop algorithms and resources which will allow expanding this expression into an orthographic form: *пяць цэлых дзевяць тысяч семсот трыццаць шэсць дзесяцітысячных на дзесяць у дваццаць чацвёртай ступені кілаграмаў* ‘five point nine thousand seven hundred thirty-six multiplied by ten raised to the twenty-fourth degree kilograms’. The problem is not easy to solve due to the enormous variety of ways in which QEMUs are expressed in writing. Plenty of these ways differ within various language systems. The main difficulty lies in the need to correctly define the categories of case, number, and gender, and to coordinate all the elements of the expression. Both in Belarusian and Russian, six cases, two numbers, and three genders are grammatically possible, which is different from English in which there are only two cases, and the ending of the word after the numeral depends only on the number. By now the authors of this paper have focused on the nominative, genitive and accusative cases.

Previously in 2009 a team of Croatian linguists developed algorithms which identify dimensional expressions of length, square and numerical ranges for Croatian and English [1]. Much has been achieved by European developers of the Numeric Property Searching service [2], and Quantalyze semantic annotation and search service [3]. Later in 2013, Belarusian researchers demonstrated finite-state automata which identify, analyse, and classify QEMUs for the Belarusian and Russian languages [4]. The paper's authors have decided to go in another research direction, in particular turning QEMUs (namely combinations of digits, symbols, and letters) into orthographically correct sequences of words which agree in gender, number, and case according to the grammar rules of the Belarusian and Russian languages.

## 2 Characterizing the Problem of the Processing of QEMUs

At first it is important to decide which tokens should be viewed as measurement units generally. Every scientific field has its own terminological apparatus, objects, subject area, methods of research, etc. For instance a survey is one of research methods in sociology [5]. Sometimes results of surveys are represented with the help of the well-known unit *percent* but often sociologists use such words as *a person, a male, a child, a woman, a student, an employee, an American, etc.*, for example: *5 млн. чал* '5 mln ppl' (should be turned into *пяць мільёнаў чалавек* 'five million people'). At the same time these "units" are not relevant for astronomy or mathematics. This research aims to develop a system which processes QEMUs used as standards for measuring certain physical quantities according to the International System of Units (SI) [6]. At the same time some frequently used and often abbreviated words, such as Belarusian/Russian *штука* or *шт* 'piece' or 'pc' are not going to be ignored either.

Another difficulty is the variety within one and the same language system. For example, the Belarusian language possesses one more system of spelling, which is called Taraškievica (or Belarusian classical orthography). Nowadays the modern and classical systems co-exist, so it is important to take both of them into consideration. Thus, the full list of variants for the Belarusian word секунда 'second' (the SI base measurement unit of time) will be the following: *с, сек, сэк, секунда, секунды, секундзе, секунду, секундай, секундаю, секундзе, секунд, секундаў, секундам, секундамі, секундах, сэкунда, сэкунды, сэкундзе, сэкунду, сэкундай, сэкундаю, сэкундзе, сэкунд, сэкундаў, сэкундам, сэкундамі, сэкундах* – 27 variants. By analogy, there is American English, British English, etc. Thus, we have American *meter-meters*, and British *metre-metres* [5].

It is important to note that the problem of the QEMUs processing can be complicated not only by language peculiarities, but also by the human factor (orthographic mistakes, misprints), and the machine properties (the encoding of power signs, software limitations, etc.).

To sum up, the problem of QEMUs processing is not as easy as it may seem in the beginning. At the same time it requires to be solved because QEMUs can be found in electronic texts of almost any thematic domain in various spheres of everyday life: starting from culinary recipes, and ending with scientific data from space satellites and probes.

## 3 Creating the Algorithm for the Processing of QEMUs

To solve the problem, by the moment we have developed a syntactic grammar (Fig. 1) consisting of over 350 graphs which can be applied to any electronic texts in Belarusian and Russian. The grammar is represented by the finite-state automaton which has been created by means of the visual graphic editor built in the linguistic processor NooJ [7]. For the present the grammar covers three cases out of six: nominative (the graphs with *Nom*), genitive (the graphs with *Gen*), and accusative (the graphs with *Acc*).

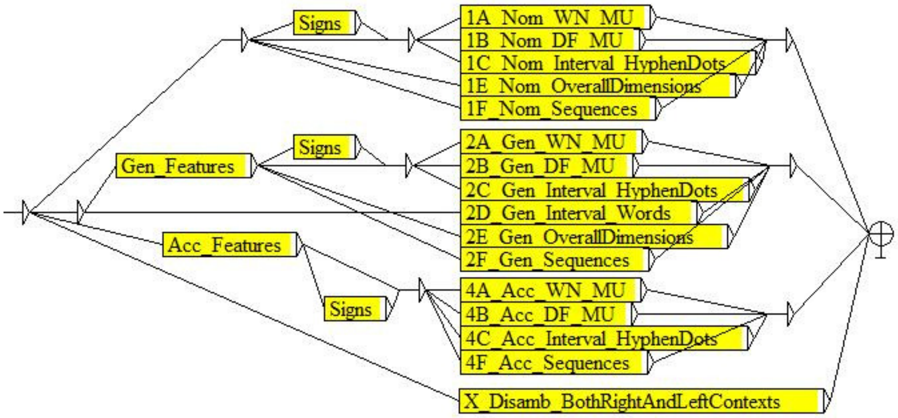


**Fig. 1.** The main graph of the algorithm for the processing of QEMUs in Belarusian/Russian

As one can see, the names of some graphs start with numbers followed by capital Latin letters. Since there are over 350 graphs we had to foresee a convenient way of the graphs ordering. There are six cases in Belarusian and Russian according to the modern grammar rules of these languages. Traditionally they are listed in the following order: $1^{st}$ nominative, $2^{nd}$ genitive, $3^{rd}$ dative, $4^{th}$ accusative, $5^{th}$ instrumental, $6^{th}$ prepositional. That is why we have put 1 before *Nom*, 2 before *Gen*, and 4 before *Acc*. Latin letters are used as codes for phenomena described by one or another graph. The phenomena are also specified at the end of the graphs' names. The *A*-graphs are for QEMUs in which numeral descriptors are expressed by whole numbers: *678 мм* '678 mm'. The *B*-graphs are used when numeral descriptors are decimal fractions: *18.0005 кілапаскаля* '18.0005 kilopascals'. The *C*-graphs work out for intervals connected either with a hyphen or dots: *125...1000 метров* '125...1000 meters'. The graph with *D* processes intervals formed with certain prepositions: *ад 5 да 6 мілірэнтгенаў* 'from 5 to 6 milliroentgens'. The algorithmic branches with *E* process expressions which describe overall dimensions: *240x707x1500 дм* '240x707x1500 dm'. The F-graphs are necessary for QEMUs in which there are homogeneous numeral descriptors: *около 2, 3, 5 ампер* 'nearly 2, 3, 5 amperes'. As one can notice in Fig. 1, there are also branches named *Signs*, *Gen_Features*, and *Acc_Features*. The *Signs*-branch works out when the algorithm finds one of the following sequences: -+± *плюс мінус* (for Belarusian); -+± *плюс минус* (for Russian). Under *Features* we mean word or symbol indicators of one or another case. For instance, after Belarusian *больш* за 'more than' or Russian *более чем* the expression is used in the accusative case, for example: *больш за 6 A* will transform into *больш за шэсць_ампераў*. At the same time Belarusian *больш* 'over' or Russian *более* will make the algorithm behave in a different way, and the phrase will take the genitive case: *больш 6 A* will be turned into *больш шасці ампераў*.

Within the subgraphs the algorithm fulfils plenty of checking operations. At first it focuses on the processing of numeral quantifiers, including not only whole numbers (*99837680, ± 2*) but also decimal fractions (*678,0000009*, ≈*567.7800*), numbers raised to one or another power (*1,43128*$\times10^{15}$), dimensions (Fig. 2) (*10*×*20*×*60*), and combinations of digits with letters (*3892 тыс*, *90 млрд*).

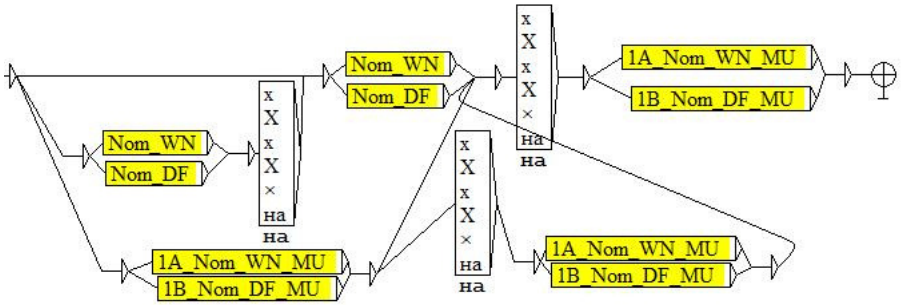
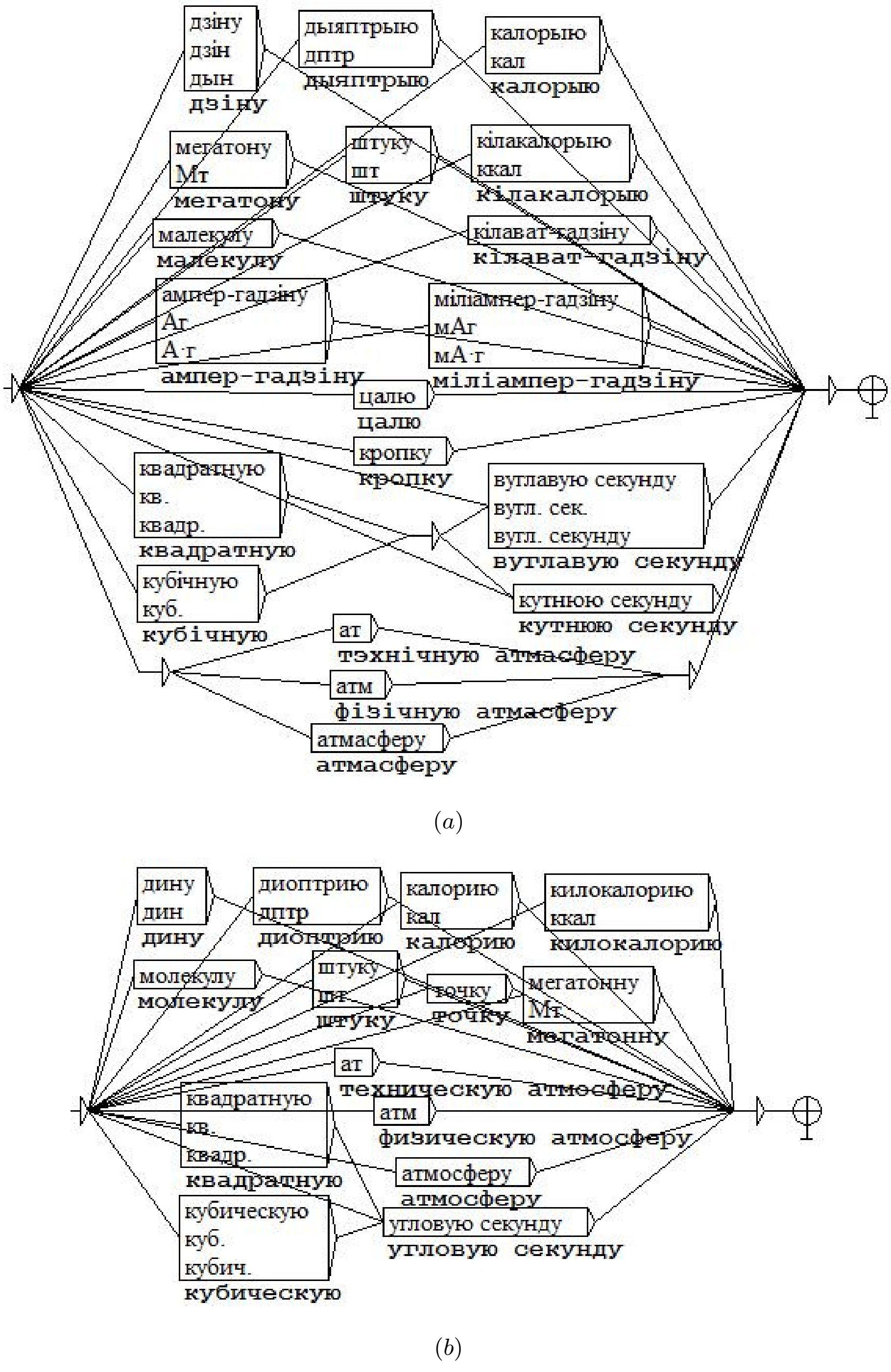


**Fig. 2.** The subgraph for the processing of QEMUs which denote overall dimensions for Belarusian and Russian

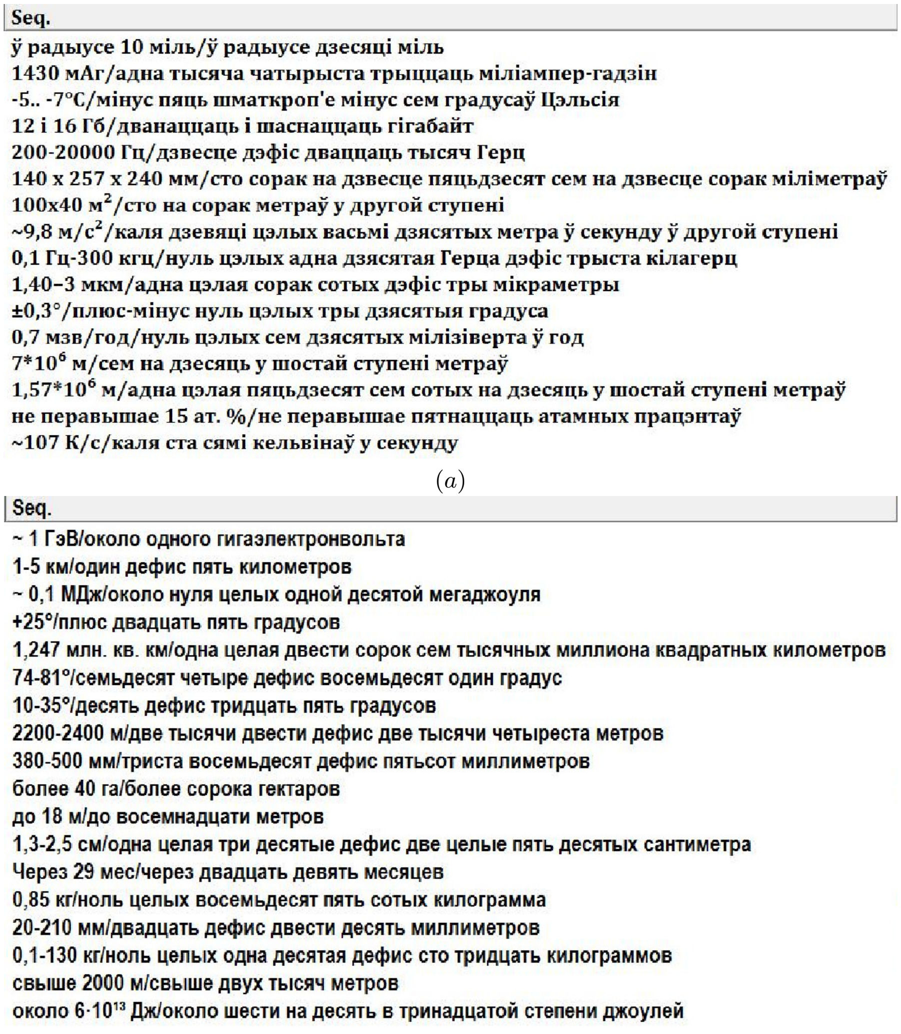
What concerns the graphs for measurement units, they process SI base units, SI derived units, some units which belong to other systems, and some tokens which can be used for measuring or even counting in general: *штука* 'piece', *диоптрия* 'dioptre', *цаля* 'inch', etc. They have been collected separately within subgraphs which contain *Extra* in their names (Fig. 3). So it is possible "to switch them off" if there ever arises the necessity to do this. We have also added



(*a*)

(*b*)

**Fig. 3.** The subgraph for the processing of QEMUs which denote overall dimensions for Belarusian and Russian



| Seq. |
|---|
| ў радыусе 10 міль/ў радыусе дзесяці міль |
| 1430 мАг/адна тысяча чатырыста трыццаць міліампер-гадзін |
| -5.. -7°С/мінус пяць шматкроп'е мінус сем градусаў Цэльсія |
| 12 і 16 Гб/дванаццаць і шаснаццаць гігабайт |
| 200-20000 Гц/дзвесце дэфіс дваццаць тысяч Герц |
| 140 х 257 х 240 мм/сто сорак на дзвесце пяцьдзесят сем на дзвесце сорак міліметраў |
| 100х40 $м^2$/сто на сорак метраў у другой ступені |
| ~9,8 $м/с^2$/каля дзевяці цэлых васьмі дзясятых метра ў секунду ў другой ступені |
| 0,1 Гц-300 кгц/нуль цэлых адна дзясятая Герца дэфіс трыста кілагерц |
| 1,40–3 мкм/адна цэлая сорак сотых дэфіс тры мікраметры |
| ±0,3°/плюс-мінус нуль цэлых тры дзясятыя градуса |
| 0,7 мзв/год/нуль цэлых сем дзясятых мілізіверта ў год |
| 7*$10^6$ м/сем на дзесяць у шостай ступені метраў |
| 1,57*$10^6$ м/адна цэлая пяцьдзесят сем сотых на дзесяць у шостай ступені метраў |
| не перавышае 15 ат. %/не перавышае пятнаццаць атамных працэнтаў |
| ~107 К/с/каля ста сямі кельвінаў у секунду |

(*a*)

| Seq. |
|---|
| ~ 1 ГэВ/около одного гигаэлектронвольта |
| 1-5 км/один дефис пять километров |
| ~ 0,1 МДж/около нуля целых одной десятой мегаджоуля |
| +25°/плюс двадцать пять градусов |
| 1,247 млн. кв. км/одна целая двести сорок сем тысячных миллиона квадратных километров |
| 74-81°/семьдесят четыре дефис восемьдесят один градус |
| 10-35°/десять дефис тридцать пять градусов |
| 2200-2400 м/две тысячи двести дефис две тысячи четыреста метров |
| 380-500 мм/триста восемьдесят дефис пятьсот миллиметров |
| более 40 га/более сорока гектаров |
| до 18 м/до восемнадцати метров |
| 1,3-2,5 см/одна целая три десятые дефис две целые пять десятых сантиметра |
| Через 29 мес/через двадцать девять месяцев |
| 0,85 кг/ноль целых восемьдесят пять сотых килограмма |
| 20-210 мм/двадцать дефис двести десять миллиметров |
| 0,1-130 кг/ноль целых одна десятая дефис сто тридцать килограммов |
| свыше 2000 м/свыше двух тысяч метров |
| около 6·$10^{13}$ Дж/около шести на десять в тринадцатой степени джоулей |

(*b*)

**Fig. 4.** Excerpts from the results of applying the algorithm to the Belarusian (*a*) and Russian (*b*) text corpora

indicators of powers which can be used either after a measurement unit (in this case a power is expressed by a superscript number or a sequence of letters: [1] *кубічны* ‘cubic’, *кв* ‘sq’, *квадратный* ‘square’), or before it (then a power can be indicated only by a sequence of letters, and the word order is reverse: *5 метраў кубічных*). The analysis of text corpora has shown that most often powers are used with measurement units of length or distance. So it has been decided not to overload the algorithm with unnecessary checks.

It is important to note that our finite-state automaton (Fig. 1) represents both the working model of the NooJ grammar [7], and the algorithm at once, so it is unnecessary to perform additional programming transformations in order to apply it in practice. Fig. 4 illustrates some results which have been received after the processing of Belarusian (*a*) and Russian (*b*) texts by our finite-state automata.

We have performed an evaluation test of the algorithm for Belarusian on the material of the text corpus with 100,000 word usages. They cover various scientific and non-scientific thematic domains: astronomy, biology, botany, chemistry, cosmology, culinary, defence technology, geography, history, law, mineralogy, news, physics, space travel science, etc. The corpus belongs to the specific NooJ-format *.noc*. Linguistic experts have counted the total number of QEMUs in the corpus: $N = 1765$. The quantity of all expressions processed by the algorithm is $L = 1430$; the number of those which have been correctly processed is $M = 1464$. The calculations have showed the following results: precision ≈ 83 %, recall ≈ 81 %, and F-score ≈ 82 %.

## 4 Conclusion

The main goal of the paper – to develop the algorithm for the processing of quantitative expressions with measurement units in the nominative, genitive, and accusative cases in Belarusian and Russian electronic texts – has been achieved. On the whole the algorithm processes whole numbers (±999 999 999 999), decimal fractions ($\pm999999999999, 999999999$) combined with 120 measurement units which belong to the classification of the International Bureau of Weights and Measures, units which belong to other systems, and some tokens which can be conditionally called units. In order to construct and test the algorithm, the computer-linguistic development environment NooJ has been used. The evaluation testing of the developed finite-state automaton has proved its rather high quality. At present we continue working in this direction, and further we plan to cover the dative, instrumental, and prepositional cases.

## References

1. Bekavac, B., Agić, Ž., Šojat, K., Tadić, M.: Units of Measurement Detection Module for NooJ. In: Mesfar S., Silberztein M. (eds.) NooJ 2009 International Conference and Workshop. Finite State Language Engineering, pp. 121–127. Centre de Publication Universitaire, Tunisia (2010)
2. Numeric Property Searching in Derwent World Patents, `http://www.stn-international.com/numeric_property_searching.html`
3. Quantalyze service, `https://www.quantalyze.com/en/`
4. Hetsevich, Yu., Skopinava A.: Identification of Expressions with Units of Measurement in Scientific, Technical & Legal Texts in Belarusian and Russian. In: Proceedings of the Workshop on Integrating IR technologies for Professional Search, `http://ceur-ws.org/Vol-968/irps_6.pdf` (2013)
5. American Heritage Science Dictionary, `http://dictionary.reference.com`
6. International System of Units (SI), `http://www.bipm.org/en/si/`
7. NooJ, `http://www.nooj4nlp.net/pages/nooj.html`