# FORMALISING NATURAL LANGUAGES WITH NOOJ 2014

Edited by

Johanna Monti
Max Silberztein
Mario Monteleone
Maria Pia di Buono

# Formalising Natural Languages with Nooj 2014

Edited by

Johanna Monti,
Max Silberztein,
Mario Monteleone
and Maria Pia di Buono

Selected papers from the NooJ 2014 International Conference
University of Sassari, 3-5 June 2014

Cambridge Scholars Publishing

Formalising Natural Languages with Nooj 2014

Edited by Johanna Monti, Max Silberztein, Mario Monteleone
and Maria Pia di Buono

This book first published 2015

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library



ISBN (10): 1-4438-7558-9
ISBN (13): 978-1-4438-7558-5

**Part III: Applications**

# CONVERTING QUANTITATIVE EXPRESSIONS WITH MEASUREMENT UNITS INTO AN ORTHOGRAPHIC FORM, AND CONVENIENT MONITORING METHODS FOR BELARUSIAN

ALENA SKOPINAVA, YURY HETSEVICH, AND JULIA BORODINA

## Abstract

*This paper describes a NooJ syntactic grammar developed for recognising quantitative expressions with measurement units (QEMU) and converting them into the grammatically correct orthographic form in Belarusian. In addition to a general description of the grammar, the paper suggests methods for easy monitoring of the results received by means of the developed grammar. These methods involve the replacement of QEMUs in an initial document with their resulting orthographically-correct equivalents in an exported XML-document which can be used further in different applications.*

## 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 solving various computer-linguistic problems. By 'computer-linguistic problem', we mean a task which refers to electronic texts, and concerns the identification, classification, and processing of sequences of letters, digits, and symbols. Solving the problem means developing a program for preliminary text processing.

At the international NooJ (Saarbrücken, 2013) and Dialogue (Bekasovo, 2013) conferences, we demonstrated solutions to several computer-linguistic problems which concern QEMUs. In particular, we gave a detailed overview of syntactic grammars and linguistic resources which identify, analyse, and classify QEMUs. All of these were built in the form of finite-state automata with the help of the linguistic processor NooJ and its built-in visual graphic editor. So far three complementary algorithmic blocks have been built for the Belarusian language. They allow:

- identification and classification of QEMUs according to the system of the International Bureau of Weights and Measures (expressions with SI-basic, SI-derived, and non-systemic measurement units)
- classification of QEMUs according to word formation peculiarities (full or shortened, with multiple or submultiple prefixes)
- expansion of QEMUs into orthographic words.

Although much has already been done, there is still room for further improvements. Problems concerning QEMUs are not so easy to solve due to the enormous variety of ways in which they are expressed in writing.

Moreover, many of the ways they are expressed differ within various language systems.

## Difficulties in Belarusian

There are some difficulties which must be taken into consideration in order to develop an accurate grammar. Let us begin with the most difficult cases. The first difficulty arises in the linguistic category of case: there are six cases in Belarusian (nominative, genitive, dative, accusative, instrumental, and prepositional), while, in English, for example, there are only two cases, common and possessive. As a result, a context can influence how words agree within one expression. In addition, numerals also influence the case of the nouns which follow them. Thus, the quantitative expression *1 хв.* '1 min.' has 6 forms in Belarusian: *1 хв. – адн**а** хвілін**а**; каля 1 хв. – каля адн**ой** хвілін**ы**; на 1 хв. – на адн**ой** хвілін**е**; больш за 1 хв. – больш за адн**у** хвілін**у**; жыць 1 хв. – жыць адн**ой** хвілін**ай**; аб 1 хв. – аб адн**ой** хвіліне*. In English, the equivalent expression '1 min. – 1 minute' will always remain unchanged, regardless of the context.

Secondly, word endings in Belarusian depend not only on the category of case but also on the categories of number and gender. Table 1 illustrates the endings taken by the numeral *один* 'one' in the nominative case. In Belarusian this numeral takes different endings depending not only on the case but also on the gender of a noun which follows it (Table 1).

| Belarusian | English |
|---|---|
| 1 ст. = адно стагоддзе **(neuter)**<br>1 хв. = адна хвіліна **(feminine)**<br>1 м. = адзін_ метр **(masculine)**<br>1 сут. = адны суткі **(pluralia tantum)** | 1 c. = one century<br>1 min. = one minute_<br>1 m. = one meter_<br>1 d. = one day_ |

**Table 1 – Declension of QEMUs containing the numeral '1' in Belarusian and English**

Thirdly, in addition to word declension, another difficulty is the variety within one language system. For example, the Belarusian language possesses a second system of spelling, which is called the *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' (ie the SI-basic 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. This phenomenon can be compared with lexical variants within American English and British English. Thus, according to the World English Dictionary, there are American forms (meter-meters), and British forms (metre-metres).

Finally, the problem of processing QEMUs is complicated by homonymy. For instance, the abbreviation *г* (in Belarusian) can stand for four different measurement units: *гадзіна, год, грам* 'hour, year, gram', and sometimes even *градус* 'degree'.

## Construction of the Grammar

Last year an algorithm for the nominative case was created. As well as the improvement of the algorithm by the addition of more measurement units and more models which can be processed, the algorithm is now able to handle QEMU sequences and intervals. However, the most important achievement is the processing of two more cases: genitive and accusative. The analysis of the NooJ-corpus of scientific and technical texts has shown that these cases are the ones used most frequently.

The grammar is fully self-containable and works without any dictionaries applied. It contains 351 graphs; therefore, we had to come up with a convenient way of ordering the graphs (Figure 1). Traditionally in Belarusian, the six cases are listed in a certain order, in particular: 1$^{st}$ is nominative, 2$^{nd}$ genitive, 3$^{rd}$ dative, 4$^{th}$ accusative, 5$^{th}$ instrumental, and 6$^{th}$ prepositional. This is why we have put *1* before *Nom*, *2* before *Gen*, and *4* before *Acc*. The capital Latin letters in the names of graphs signify a model of QEMU described by the graph. The model is specified by abbreviations at the end of the names of the graphs. For instance, the letter *A* in the name '1A_Nom_WN_MU' stands for a QEMU with a numeral descriptor expressed by a whole number *WN_MU*. With such ordering, we receive an algorithmic tree which is clear and easy to work with.

Other graphs in the grammar describe mathematical signs which can be found in front of numeral quantifiers, as well as the most probable prepositions and other pieces of the remaining context. The latter are stored in the graphs 'Gen_Features' and 'Acc_Features'.

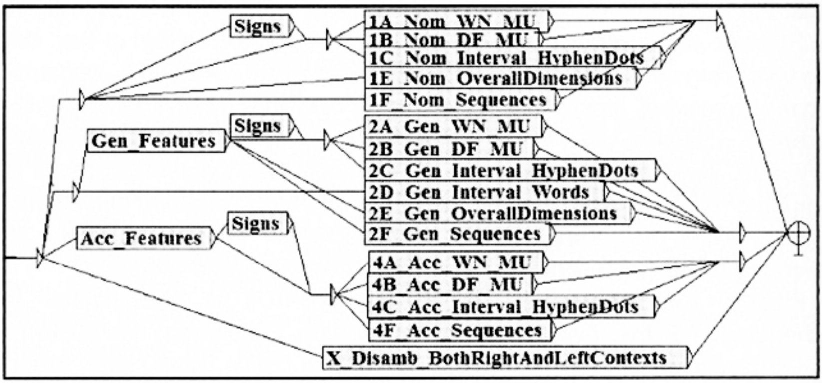


Figure 1 – General view of the grammar for QEMU processing

For a more detailed view of the work of the grammar, let us look at what is inside the graph '1A_Acc_WN_MU' (Figure 2). The graph has been created for QEMU which are used in the nominative case, and which contain a numeral quantifier expressed by a whole number. Graphs of the A-group (whose names start with *a_...*) process numeral quantifiers according to the required grammatical form: singular or plural; feminine or masculine; nominative, genitive, or accusative cases. Graphs of the B-group, in turn, describe measurement units of each grammatical form and class (basic SI, derived from SI, and out of SI).

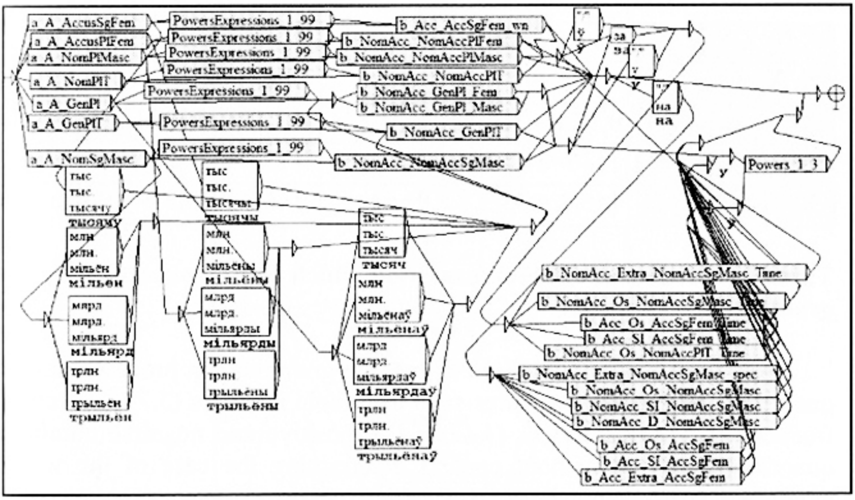


Figure 2 – Graph for QEMUs with a whole number in the nominative case

For example, the expression *74 градусы* '74 degrees' is processed within this graph. Since *74* is a whole number and the expression takes the nominative case, the grammar will use this exact graph. As a result of the processing, *74 градусы* '74 degrees' is converted into the orthographic form *семдзесят чатыры градусы* 'seventy-four degrees'.

This example can be represented as a common model: *WXY*, where *X* is any numeral quantifier, *Y* is a measurement unit and *W* is a context determining the grammatical case. This model of the formation suits the majority of QEMUs, but there are some other models which the grammar can process (Table 2).

| Model | Example | English Translation |
|---|---|---|
| X Y | 12 % \| 40-50 тыс. м | 12 % \| 40-50 thousand m. |
| X Y/Y | 0,5144444 м/с | 0,5144444 m/s |
| X-[… ..]X Y | 1-1,5 года \| +13... +19 °С | 1-1,5 years \| +13... +19 °C |
| X-[… ..]X Y/Y | 0,1-5,7·$10^{-2}$ м/с | 0,1-5,7·$10^{-2}$ m/s |
| ~[+-±><]X Y | ±0,3° \| > 6 Зв | ±0,3° \| > 6 Sv |
| ~[+-±<>]X Y/Y | ~107 К/с \| ~9,8 м/с$^2$ | ~107 C/s \| ~9,8 m/s$^2$ |

| | | |
|---|---|---|
| X,[i] X Y | 2 і 4 метры | 2 and 4 meters |
| X, X, X Y | 5, 6, 7 шт | 5, 6, 7 pc |
| X Y - X Y | 0,1 Гц - 300 кГц | 0,1 Hz - 300 kHz |
| X × X Y | 1136×640 пікселяў | 1136×640 pixels |
| X × X × XY | 146,8 × 75,3 × 8,9 мм | 146,8 × 75,3 × 8,9 mm |

**Table 2 – Models of QEMU formation which can be processed by the grammar**

Apart from the models listed above, the grammar can process numeral quantifiers of various structures such as: whole numbers (*5,791*), decimal fractions (*0.5144*), intervals (*+13...+19*), positive and negative numbers, quantifiers with or without context determining the case of the whole expression (*~20*), exponents ($0.1\text{-}5.7\cdot 10^{-2}$), etc.

The graphs starting with *b_...*, describe measurement units. The grammar can process about 120 measurement units belonging to various classes: SI-basic units (*кілаграм* 'kilogram', *ампер* 'ampere'), SI-derived units (джоуль 'joule', *радыян* 'radian'), units belonging to other systems (*гектар* 'hectare', *градус* 'degree'), and additional words which are not officially considered units but can still be used for measuring (*штука* 'piece'). These classes are illustrated by the names of the graphs: *SI*, *D*, *OS*, and *EXTRA* respectively. See Figure 3 for more examples of measurement units from Belarusian.

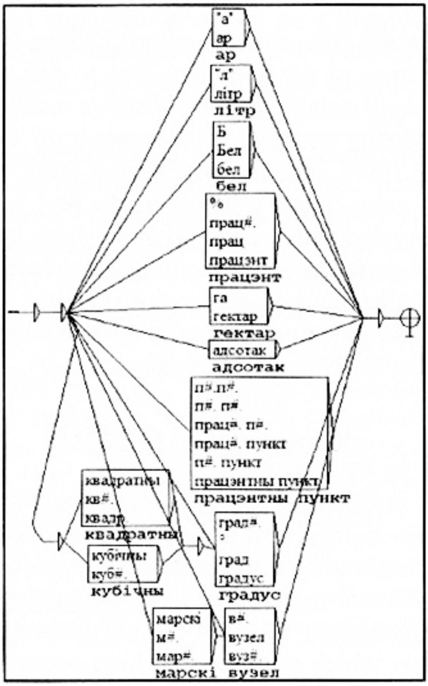


Figure 3 – Examples of measurement units belonging to other systems in Belarusian

## Quality Evaluation

As we have already mentioned, the algorithm consists of 351 graphs. It can process a variety of numbers: up to 12-digit, whole and fractional, positive and negative, exponents, intervals, etc. Also, the algorithm is capable of recognising more than 120 measurement units, including an extra category containing units which can be used for measuring but are not officially considered units eg: *штука* 'a piece'; or units which are relatively seldom used eg *дыяптрыя* 'a diopter' etc.

The quality evaluation test has been performed on the material of a text corpus of over 100 thousand tokens containing 1,765 QEMU, and it has shown quite good results: precision 86%, recall 82%, and F-measure 84%. For now the grammar covers three linguistic cases out of six, but the performance of the grammar is quite accurate: QEMU tend to be most frequently used in the following cases: nominative, genitive and accusative – the ones which our grammar covers.

However, one of the difficulties mentioned above is homonymy, which affects the performance of the grammar. Due to full homonymy of some phrases, there is a small number of false-positive results (approx. 1%). The context of the following example *у пастанове ад 29.03.1996 г* 'in the resolution of 29.03.1996' clarifies that the last letter stands for the word *год* 'year' (but not *грам* 'gram'); however, the part *3.1996 г* is processed separately and converted into an orthographic form as *тры цэлыя адна тысяча дзевяцьсот дзевяноста шэсць дзесяцітысячных грама* 'three point one thousand nine hundred ninety nine grams'.

Thus, the grammar performance can be improved by adding the rest of the cases and solving the homonymy issues.

## Replacer for QEMU

In the previous sections we have observed results with the help of the 'Outputs' checkbox in the NooJ concordance mode. This is extremely useful for the intermediate monitoring of performance at the development stage. However, different researchers have different goals and objectives concerning text extraction, recognition or processing. Therefore, we suggest a technology which turns the resulting outputs into a form which can be further handled by another application and can be understood by any user. A description of the process is as follows:

First, we need to add special markers to the main graph of the algorithm (Figure 4). With their addition, the correct grammatical forms which have been previously generated as comments become part of the XML tag. Later these markers will let us replace the initial quantitative expressions with their full equivalents.

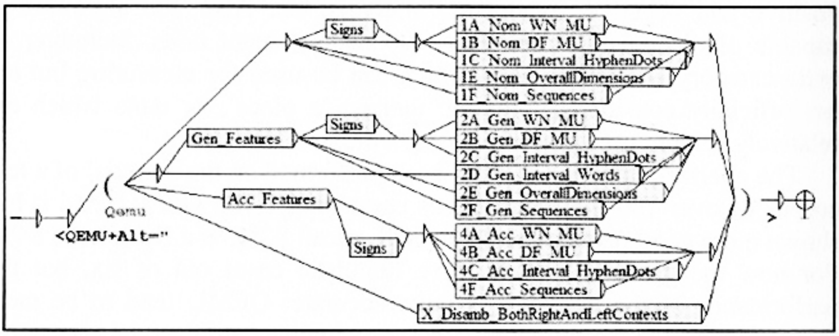


Figure 4 – General view of the grammar with XML markers added

Second, we must apply the grammar as a linguistic resource. At the stage of lexical analysis conducted on the corpus, QEMUs are identified and annotated with XML tags. Figure 5 shows the annotated text.

```
Amazon будзе дастаўляць заказы беспілотнікамі
Амерыканская інтэрнэт-крама Amazon плануе запусціць
дастаўку пасылак з дапамогай Дронаў.
Пра гэта гаворыцца ў паведамленні кампаніі.
Дроны змогуць дастаўляць пасылкі вагой <QEMU
Alt="да дзвюх цэлых трох дзясятых кілаграма ">да
2,3 кг</QEMU>, якія цяпер складаюць <QEMU Alt="каля
васьмідзесяці шасці працэнтаў ">каля 86%</QEMU>
усіх заказаў на Amazon, сказаў генеральны дырэктар
кампаніі Джэф Безас у інтэрв'ю тэлеканалу CBS.
Дастаўка будзе ажыццяўляцца <QEMU Alt="ў радыусе
дзесяці міль ">ў радыусе 10 міль</QEMU> (прыблізна
<QEMU Alt="шаснаццаць кіламетраў ">16 км</QEMU>) ад
складоў інтэрнэт-крамы, максімальны час ад
афармлення заказу да атрымання тавару <QEMU
Alt="складзе трыццаць хвілін ">складзе 30
хвілін</QEMU>. Авіадастаўка будзе ўключаная ў
```

Figure 5 – A piece of text with QEMU annotations in Belarusian

The third step is to export the annotated text in the XML format. We see in Figure 5 that QEMUs have received the markers and have been converted into their full forms, but the abbreviated form of the expression is still preserved. Annotated text can then be exported as an XML document.

In order to complete the replacement of the initial expression, we launch and apply a specially written script program in the PEARL language, which is called 'Replacer for QEMU'.

An input file must be an annotated text in the txt–format. It is placed in the folder 'in' (Figure 6). After launching the 'Start.bat', the program starts to work. The output htm-file appears in the folder 'out'. The final result can be seen in Figure 7. All the initial sequences of signs, digits, and letters are now transformed into linguistically correct word expressions. In addition, their font colour is automatically changed to red and underlined.

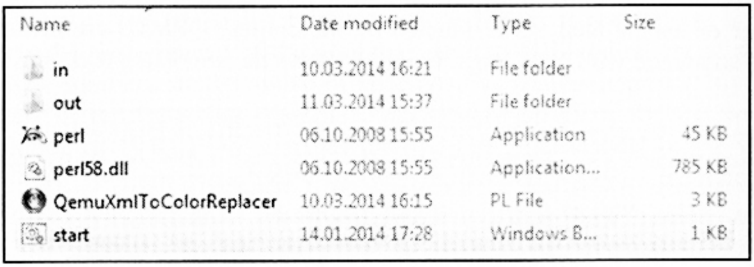


Figure 6 – Program 'Replacer for QEMU'

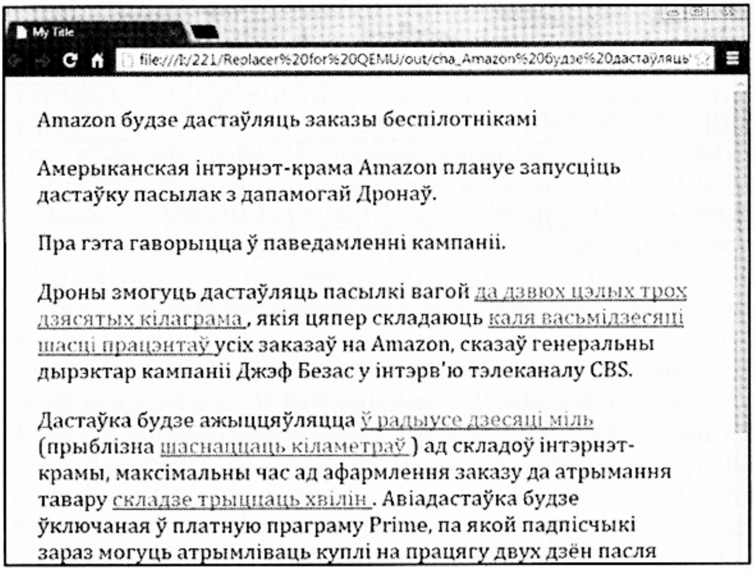


Figure 7 – Resulting text with underlined QEMUs

Anyone who knows the basics of programming or HTML can easily change the style of a replaced text to any other by editing the <FONT…> line in the PEARL code.

## Conclusion and Future Work

Quantitative expressions with measurement units can be found in electronic texts of almost any thematic domain in various areas of everyday personal and professional life: culinary recipes; information about goods in online-shops or on labels; weather forecasts; transport schedules; commentaries on sport events; ticket prices; tourist guides; voiced transmission of scientific data from space satellites and probes; architectural descriptions; measurements of human body indicators, such as temperature, pressure, pulse, sugar, weight, cholesterol etc.

As a result, we have created an algorithm which transforms QEMUs into orthographical words in Belarusian. The grammar describes several particular models of QEMU formation, as well as an enormous variety of numeral quantifiers and measurement units. The grammar has been tested on a corpus containing 100,000 tokens, and it has shown relatively good performance, with precision and recall rates over 80 %. Nevertheless, the algorithm can be further improved. In order to reach the highest level of precision and recall, we plan to add the rest of the cases, more types of numeral quantifiers, measurement units, and structural models.

## Acknowledgements

We would like to thank the linguist Adam Morrison for his help in revising the language of this paper.

## References

Collins English Dictionary – Complete & Unabridged 10th Edition [Electronic resource]. – HarperCollins Publishers, 2011. – http://dictionary.reference.com/browse/sociology.

Hetsevich Yuras and Alena Skopinava. 2013. 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)

—. 2013. Transforming quantitative expressions with measurement units into orthographical words for text-to-speech synthesis to Belarusian and Russian. In *Вестник МГЛУ. Сер. 1, Филология*, pp. 133–144.

Skopinava Alena, Yuras Hetsevich, and Boris Lobanov. 2013. Processing of quantitative expressions with units of measurement in scientific texts as applied to Belarusian and Russian text-to-speech synthesis. In *Компьютерная лингвистика и интеллектуальные технологии: По материалам Международной конференции «Диалог»*, 29 мая – 2 июня 2013: Bekasovo, Russia, pp. 634–651.

The American Heritage Science Dictionary [Electronic resource]. – 2014. – Mode of access: http://dictionary.reference.com/browse/sociology. – Date of access: 03.03.2014.