Identification of Expressions with Units of Measurement in Scientific, Technical & Legal Texts in Belarusian and Russian

Alena Skopinava
United Institute of Informatics Problems,
National Academy of Sciences,
Minsk, Belarus
skelena777@gmail.com

Yury Hetsevich
United Institute of Informatics Problems,
National Academy of Sciences,
Minsk, Belarus
yury.hetsevich@gmail.com

Abstract

A study of an identifying process of expressions with metrological units according to the International System of Units for thematically distinct text corpora for Belarusian and Russian is reported here. The urgency of the problem is dictated by the ubiquity of units of measurement and their enormous variety. The resulting algorithms are created in the form of finite automata through a set of visual syntactic grammars. Such a method allows algorithms and resources to be updated much easier and far more quickly than, for instance, regular expressions. The algorithms carry out a search for expressions with measurement units, identify and classify them according to the SI. These practical results may find application in information search engines, libraries, publishing houses and speech synthesis systems.

1 Introduction

Units of measurement have truly been man's helpmates since ancient times. They are used in every branch of science as well as in daily life. It is metrological units that present a quantitative perspective of the world, whether we consider the building of the Egyptian pyramids or flights into space.

Units of measurement are a current object of numerous research projects, first of all, within the framework of a special discipline – metrology, as well as in the fields of mathematics, computer science, physics, coding theory, and, of course, corpora linguistics. Possible examples can be such frequently used expressions as: *120 км/г (120 km/h)* (speed), *345 м (345 m)* (distance), *12 мА (12 mA)* (amperage), etc.

Texts containing expressions with measured units require special algorithms of identification and processing in the following areas:

– corpora and database management systems, libraries, information retrieval systems (to formulate extended search queries, locate specific expressions on the Internet, support automatic textannotation and summarization);

– speech synthesis systems according to the text (to generate orthographically correct texts and their tonal and prosodic peculiarities);

– publishing institutions (to locate automatically specified lists of expressions with measured units, classify resulting

In: M. Lupu, M. Salampasis, N. Fuhr, A. Hanbury, B. Larsen, H. Strindberg (eds.): Proceedings of the Integrating IR technologies for Professional Search Workshop, Moscow, Russia, 24-March-2013, published at http://ceur-ws.org

expressions as SI units, their derivatives or units out of the SI, and finally check quickly if the extended names of units are used correctly).

However, when dealing with units of measurement, many difficulties arise. Firstly, 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. 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. Secondly, an expression with measured units is difficult to recognize and analyze (divide by the numeral quantifier (digits, parts of speech with quantitative meaning with all their possible paradigmatic forms) and the name of the metrological system) without thoroughly-prepared linguistic resources, that is, dictionaries with all possible word forms, abbreviations and rules for building derivative forms of measured units' names. This is necessary for proper localization, for instance, of the following expressions with units of length, recorded in various ways: *1 м (1 m), 31 метр (31 meters), 25 метраў (25 meters), 44 метры (44 meters)*. Thirdly, expressions with units of measurement are language-dependent: in English abbreviated *meter* and *mile* refer to *m*, while in Belarusian and Russian – *м*; even within similar Russian and Belarusian names of units differ in spelling. Therefore, it is essential to make accurate specifications for respective recognition of algorithms.

Some important achievements have already been realized in the Quantalyze semantic annotation and search service [QS13], and Numeric Property Searching service in Derwent World Patents Index on STN [NPS98]. However, both services cannot be applied to Belarusian or Russian text corpora. Also some steps in order to solve the above-mentioned problems were carried out in 2009 by a team of Croatian linguists, who managed to obtain algorithms to identify dimensional expressions of length, square and numerical ranges for the English and Croatian languages [Bek09]. Superficial coverage and interpretation of the subject area can be found in research conducted by several other European linguists. Still their work is more theoretically than practically based, is very descriptive and is concentrated on measured units not as separate objects but "certain occurrences of words and expressions as belonging to particular category of named entity" [Cun99]. Research workers of the Bulgarian Academy of Sciences and its Department of Sheffield University mention that their "observations on the linguistic nature of Slavonic NE [named entities] are based only on their general characteristics and on the general conclusions on their behavior in the text" [Pas02]. In practice, the named entity is semantically huge and is composed of other various, complex categories: "locations, persons, organizations, dates, times, monetary amounts and percentages" [Pas02] or, in other words, "persons, locations, organizations, time and numerical expressions" [Myk07]. So, all of them have to be treated separately if the chief aim is to obtain successful identification algorithms. A Bulgarian-Serbian research team particularized the term 'measure' as "a structure of a sequence of numbers written by words or digits followed by a measure indicator (kilometer, grade, mile, foot, etc.)" [Duš07] and represented it formally as a graph. Still, their practical results are limited only to the definite language systems (Bulgarian, Serbian), and, therefore, cannot be applied to Belarusian or Russian.

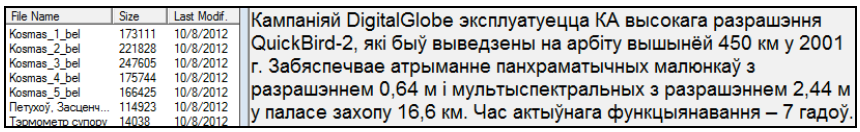
So, our research work views expressions with measured units as numerical word combinations where each component requires a certain approach for successful identification. Our goal is to develop algorithms and linguistic resources in order to identify and classify units of measurement and expressions with them on the material of hand-crafted scientific-technical and legal text corpora for the Belarusian and Russian languages.

The specificity of our work is not simply to describe the expressions with measurement units. Their enormous variety is the reason why regular expressions are not the best way to obtain localization rules. The use of visual methods of NooJ allows users to easily modify previously-developed rules and add new ones. The opportunities are endless for any language. We decided to demonstrate them using Belarusian and Russian, two Slavic languages which have much in common, but at the same time they differ. So do the units of measurement. Most of the resources necessary for their localization in complex text fragments are language-dependent, but the algorithm itself remains the same. It should be noted that not only the units of measurement are algorithmically described but also the numeral quantifiers that stand before them, which is extremely important for automatic comprehension of documents and information retrieval systems. The corpora for testing are also constructed by means of NooJ, as it can perform any syntactic or semantic analysis on partially or totally ambiguous texts. This fact confirms that all the results (concordances with units of measurement) are obtained only with the help of algorithms, rather than special tags or indices.

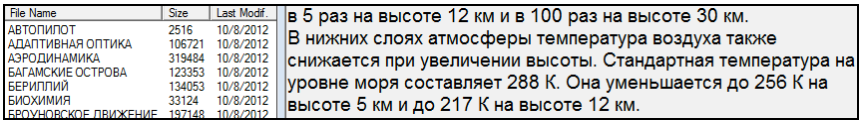
2 Finite-State Automata of NooJ for Identification of Measurement Units

In order to find a solution to the above-mentioned scientific problem, some practical results, already obtained while constructing the Belarusian and Russian modules of the international computer-linguistic program NooJ, were used [Het12a, Het12b]. This program allows to implement sophisticated algorithms for searching across compound text fragments in Belarusian and Russian in the form of visual executable graphs within finite-state automata [NooJ02].

For the construction and testing of algorithms, four text corpora were formed: two in Belarusian and two in Russian. They contain a wide array of expressions with units of measurement for two thematically distinct domains: scientific-technical (from the fields of astronomy, physics, geography, chemistry, aviation, space, history, energy, transport and communication) (figure 1) and legal (the traffic code of Belarus) (figure 2) [RRB07].

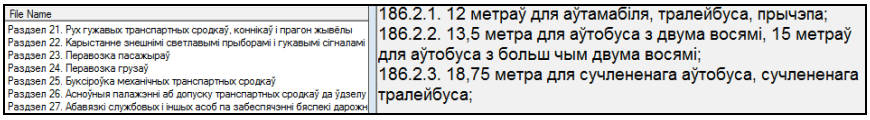


a)

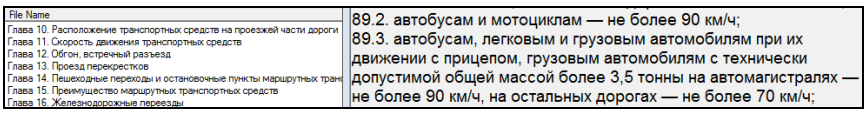


b)

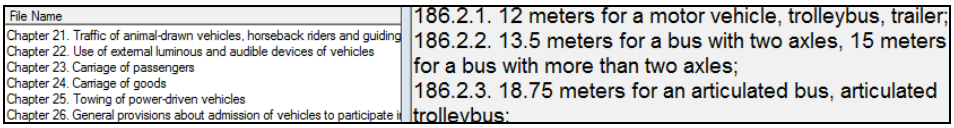
Figure 1: Fragments of scientific and technical text corpora for the *a)* Belarusian and *b)* Russian languages



a)



b)



c)

Figure 2: Fragments of legal text corpora for the *a)* Belarusian, *b)* Russian languages, and *c)* translated into English

Each of the resulting algorithms (figure 3) is presented as the main graph. From left to right, it contains the input (marked by an arrow) and output (marked by a circle with a cross inside), 4 subgraphs (each with a title), transition lines and markers. A graph or subgraph works out when a complete path from its input to output can be fulfilled on condition that all checks are executed. Algorithms for the Belarusian and Russian languages differ in some language-dependent subgraphs.

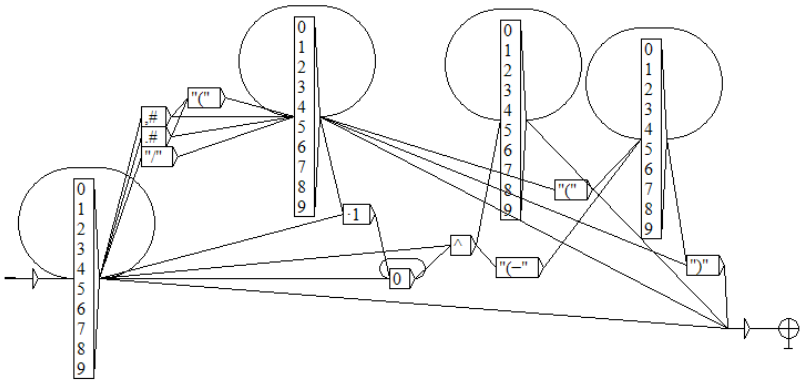


Figure 3: The subgraph which recognizes numbers and complex numeral expressions

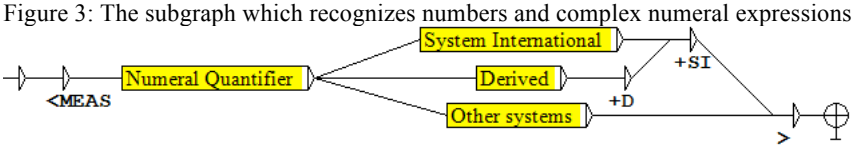


Figure 4: The main graph of the algorithm for identification of expressions with units of measurement in text corpora

According to the main graph, any text fragment is initially checked in the 1st subgraph (Numeral Quantifier) if it has a compound numerical descriptor (figure 4). It should be noted that this subgraph works out not only for prime, decimal and fractional numbers in various forms of recording but also for compound numeral expressions with exponential parts and periods. Some results of the work of the subgraph can be observed in the form of a concordance in figure 5. The extracted numerals are listed in the column *Seq*. The columns *Before* and *After* contain pieces of left and right contexts in which the extracted numerals are used. It should be emphasized that this subgraph is language-independent (see the example for English in figure 5c).

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

b)

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

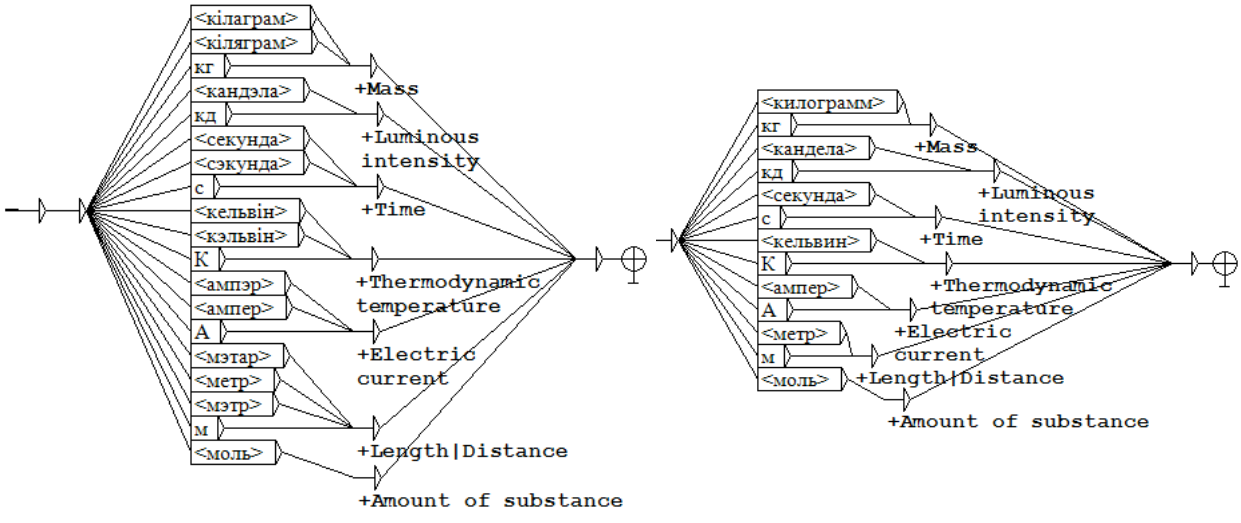
c)

Figure 5: Some results of the work of the subgraph, which identifies complex numeral expressions in texts in the *a)* Belarusian, *b)* Russian, *c)* English languages

After the first subgraph works out, the algorithm proceeds to checks of other subgraphs, which are connected to its output by means of respective transition lines. For comparison, the 6th figure represents subgraphs for Belarusian and Rus-

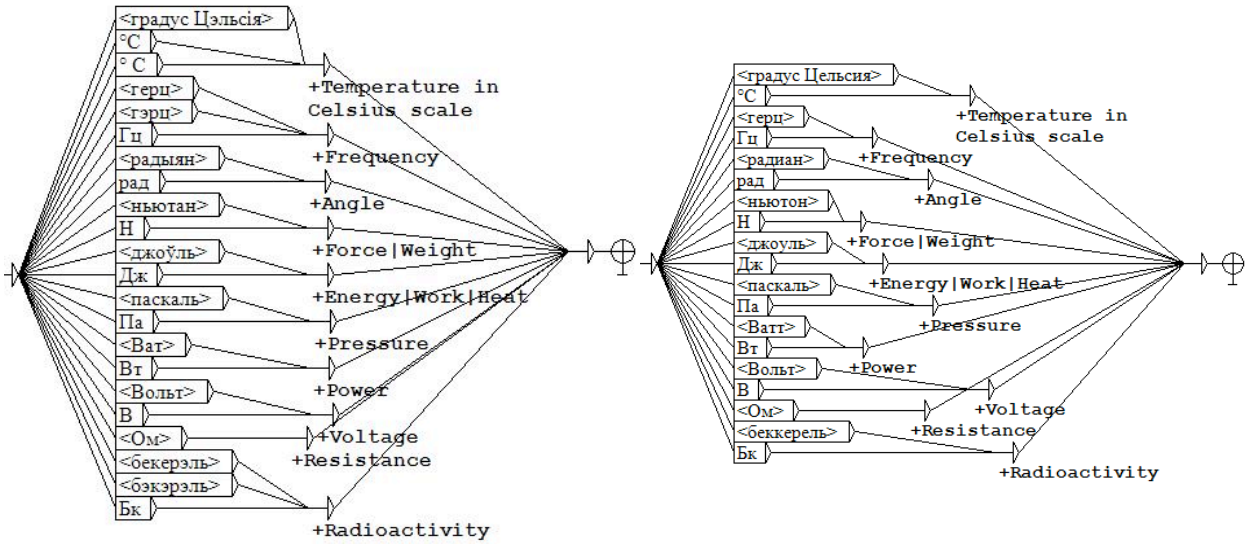
sian texts for recognition of the units of measurement within the International System of Units. These subgraphs are language-dependable. Though the subgraphs in figure 6 serve the same purpose and recognize the same units (*kilograms*, *candelas*, *seconds*, *kelvins*, *amperes*, *meters*, *moles*), they differ not only by fonts, but also by ways of writing the same units of measurement. For example, in Russian the electric current can be measured by *"А"* or *"Ампер"*. So in Russian there are 2 ways to express 1 unit, that is *ampere*. In Belarusian one can use 3 ways: *"А"*, *"Ампер"*, *"Ампэр"*. By the way, in English there are 3 ways as well: *A*, *ampere*, *amp*. It should be mentioned that the order of checks is not important, because all the checks within subgraphs are mutually exclusive. They help to search for and identify expressions with measured units that belong to the general classification of measurements of the International Bureau of Weights and Measures [BIPM06]. The subgraph *System International* identifies the units according to the SI, for example, *кілаграм* (*kilogram)*; the subgraph *Derived* – SI-derivative units, such as *герц (hertz)*; the subgraph *Other systems* – the most common, frequently used, but non-systemic units, such as *час* (*hour)*. If any of the above-mentioned three subgraphs works out, the sequence of respective transition lines on the way to the main graph's output is indicated by markers. Let's draw up a list of some possible outcomes of the markers in case the main graph works out: *MEAS* (stands for 'measurement'), *MEAS+SI+…*, *MEAS+D+SI+…*. These markers correspond to the above-mentioned subgraphs' respective predestinations. Three dots in the last two markers can be substituted for special markers within a respective subgraph that works out. In each subgraph, a name of a measured unit (or its word form) corresponds to the name of a respective physical value (or its word form). Let's take the word combination *"узяць 3,3 молі" (take 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 to identify exactly which subgraph works out and which units of measurement are used in the expression. 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 unit of measurement *"молі" (moles)* belongs to the SI units. The code + *Amount of substance* means that *"молі" (moles)* are used for measuring amount 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 units derived from the SI base units, such as *degree Celsius*, *hertz*, *radian*, *newton*, *joule*, *pascal*, *watt*, *volt*, *ohm*, *becquerel*. Its structure is shown in figure 7.



a) *b)*

Figure 6: The subgraphs which identify expressions with units within the SI for *a)* Belarusian and *b)* Russian texts



a) *b)*

Figure 7: The subgraphs which identify expressions with units derived from the SI base units for *a)* Belarusian and *b)* Russian texts

As a result, a flexible system of markers allows the user to build search queries of different types: to find all the expressions with units of measurement (figure 8); to draw a concordance of expressions with units of mass (on request *<MEAS+Mass>*) or length *(<MEAS+Length>)*, to determine expressions either with units, derived from the SI-units *(<MEAS+SI+D>)* (figure 10) or without them *(<MEAS+SI-D>)* (figure 9); to recognize expressions that do not belong to the SI *(<MEAS-SI>)* (figure 11); etc. Table 1 contains the search results in figures 8-11 translated into English and listed from top to bottom.

Before	Seq.	After
разрашэнне –	1м/<MEAS+Length|Distance+SI>	(бач.), 5
МЭВ, 2-30 кэв,	0,1 Гц/<MEAS+Frequency+D+SI>	-300 кгц,
гартавую масу	8 т/<MEAS+Mass>	, выведзе
вання Зямлі. У	2005 г./<MEAS+Time>	Іран зда
хвілін і ўхілам	74 градусы/<MEAS+Angle>	. Затым к

a)

Before	Seq.	After
температуре	109 K/<MEAS+Thermodynamic temperature+...	В этом сл
гимостью ок.	200 000 л/<MEAS+Volume>	. Железные
рии (а спустя	33 года/<MEAS+Time>	– и его сын
е превышало	5°/<MEAS+Angle>	), а потом на
ратуре выше	600° C/<MEAS+Temperature in Celsius scal...	. а халькоге

b)

Figure 8: The fragment with results of identification of expressions with measured units within scientific and technical corpora in the *a)* Belarusian and *b)* Russian languages

Before	Seq.	After
з разрашэннем да 1–	5 м	. З улікам камерцыйных
інфармацыі на ўзроўні 1–	10 м	. Такая дэталёвасць неабход
стартавая маса перавышае	3600 кг	. Разліковы тэрмін актыўнага
масай меней за	10 кг	, а праз 10–20 гадоў
гадоў – масай парадку	1 кг	, якія змогуць устанаўліваць
гэтым складала парадку 150–	500 метраў	. У 70–80-х гг

a)

Before	Seq.	After
в диапазоне от 10–	50 м	до нескольких километров
разрешение – ОД % (вид.),	0,1 К	(ИК), 1дБ (СВЧ
часа; пространственное разрешение –	1м	(вид.), 5 м (ИК
упоминается. Выступая 19 апреля	1904 с	большим докладом в
положения через каждые	30 секунд	трех броуновских частиц
через 30, а через	3 секунды	то прямые между

b)

Figure 9: The fragment with results of identification of expressions with ***only*** SI-units of measurement within scientific and technical corpora on request <*MEAS+SI-D*> in the *a)* Belarusian and *b)* Russian languages

Before	Seq.	After
МЭВ, 2-200 МЭВ, 2-30 кэв,	0,1 Гц	-300 кгц, 0-50 кгц; перыядычнасць
рэшткавай атмасферы складаў	400 Па	, працягласць плазмавага імпульсу
гэтага тэрмомэтра пры	0 °C	, і літара, якая
у дыяпазоне ад −	50 °C	да +200 °C. Залежнасьць
ад −50 °C да +	200 °C	. Залежнасьць супору ад
у дыяпазоне ад −	260 °C	да +1100 °C. Залежнасьць

a)

Before	Seq.	After
МэВ, 2–200 МэВ, 2–30 кэВ,	0,1 Гц	–300 кГц, 0–50 кГц; периодичность
лазер накачки мощностью	25 Вт	возбуждает лазер на
красителях выходной мощность	4,25 Вт	, который и дает
всех металлов теплоемкостью:	16,44 Дж	/(моль К) для
К) для ▯-Be,	30,0 Дж	/(моль К) для
С она составляет	209,3 Вт	/ (м К), что

b)

Figure 10: The fragment with results of identification of expressions with SI-derived units within scientific and technical corpora on request <*MEAS+SI+D*> in the *a)* Belarusian and *b)* Russian languages

Before	Seq.	After
на час да	5 хвілін	, а таксама больш
масай больш за	3,5 тоны	, колавыя трактары і
час больш за	5 хвілін	і стаянка транспартнага
масай больш за	3,5 тоны	, аўтобусаў, колавых трактароў

a)

Before	Seq.	After
на время до	5 минут	, а также более
более чем на	5 минут	по причинам, не
общей массой более	3,5 тонны	, колесные тракторы и
общей массой более	3,5 тонны	, автобусов, колесных тракторов

b)

Figure 11: The fragment with results of identification of expressions with units out of the SI units within legal corpora on request <*MEAS-SI*> in the *a)* Belarusian and *b)* Russian languages

Table 1: The search results in figures 8-11 translated into English

	Figure 8	Figure 9	Figure 10	Figure 11

a)	1m <MEAS+Length|Distance+SI> 0,1Hz <MEAS+Frequency+D+SI> 8 t <MEAS+Mass> year 2005 <MEAS+Time> 74 degrees <MEAS+Angle>	5 m 10 m 3600 kg 10 kg 1 kg 500 metres	0,1 Hz 400 Pa 0 °C 50 °C 200 °C 260 °C	5 minutes *(5 hvilin)* 3,5 tons 5 minutes 3,5 tons
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>	50 m 0,1 K 1m 1904 30 seconds (*30 sekund*) 3 seconds *(3 sekundi)*	0,1 Hz 25 W 4,25 W 16,44 J 30,0 J 209,3 W	5 minutes *(5 minut)* 5 minutes 3,5 tons 3,5 tons

3 Evaluation of the resulting algorithms

Evaluation of the resulting algorithms is based on the values of precision *(P)*, recall *(R)* and their average harmonic mean *(F-measure)* of the results of identification of expressions with units of measurement on the material of text corpora. Table 2 contains all the necessary data and calculations performed.

Table 2: Evaluation of the obtained results

Corpora	The number of expressions with measured units …			Criteria for evaluation, %		
	in total (N)	identified ***in total*** by the algorithms (L)	identified ***correctly*** by the algorithms (M)	*P*	*R*	*F-measure*
				$(M/L)\times100$	$(M/N)\times100$	$2\times P\times F/(P+R)$
Legal (Bel)	104	65	63	97	61	75
Legal (Rus)	107	66	64	97	59	73
Science & Technology (Bel)	692	404	393	97	57	72
Science & Technology (Rus)	811	478	476	99	59	75

For evaluation of algorithms, texts with 104 and 107 usages of units of measurement were selected respectively for the Belarusian and Russian languages within legal corpora, and 692 and 811 usages within scientific and technological corpora. This parameter is represented by the letter *N*. Then an expert checked the concordances built by the algorithms for selected tests. The total number of expressions with measured units and then the quantity of only correctly-identified combinations were counted separately and presented by the letters *L* and *M* respectively. The evaluation process showed that the algorithms developed possess on average 72% accuracy for each test corpus.

4 Conclusion

It can be concluded that the main goal of this research – to take the first steps toward developing appropriate algorithms that identify expressions with various measured units for the Belarusian and Russian languages for materials in 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. Though rather high results have already been achieved (more than 70%), there is still much room for further improvements. For example:

– taking into account a metrological system of prefixes as parts of the units' names (mille-, deci- , kilo-, giga-, etc.);

– disambiguation of multiple-valued expressions, for example, in such cases when algorithms "confuse" some units with each other (the same initial letter *'г'* for *'год' (year), 'грам' (gram)*, *'гадзі́на' (hour*) or some units with brands of vehicles (*МАЗ-4А*, not *4 amperes*);

– developing algorithms that will identify numeral quantifiers expressed not only by numbers (mathematical objects), but also numerals (parts of speech);

– identifying the plus and minus signs positioned in front of numeral quantifiers, disambiguating the minus, hyphen and dash signs;

– updating the base of measured units with seldom-used ones.

References

[NPS98] *Numeric Property Searching in Derwent World Patents Index on STN* [Electronic resource]. – 1998. – Mode of access: http://www.stn-international.com/numeric_property_searching.html. – Date of access: 05.02.2013.

[QS13] *Quantalyze semantic annotation and search service* [Electronic resource]. – 2013. – Mode of access: https://www.quantalyze.com/en/. – Date of access: 05.02.2013.

[Bek09] B. Bekavac. *Units of Measurement Detection Module for NooJ*. Conference on NooJ 2009, pp. 121-127, Tunisia (2009)

[Cun99] H. Cunningham. *Information Extraction: a User Guide (revised version), Research Memorandum CS-99-07.* Department of Computer Science, University of Sheffield (May, 1999)

[Pas02] E. Paskaleva, G. Angelova, M. Jankova, K. Bontcheva., H. Cunningham, Y. Wilks. *Slavonic Named Entities in Gate, Research Memorandum CS-02-01.* Department of Computer Science, University of Sheffield, Great Britain (2002)

[Myk07] A. Mykowiecka, A. Kupść, M. Marciniak, J. Piskorski. *Resources for Information Extraction from Polish texts.* Proceedings of 3rd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Poznan (2007)

[Duš07] V. Duško, C. Krstev, S. Koeva. *Towards a Complex Model for Morpho-Syntactic Annotation.* Proceedings of the Workshop Workshop on a Common Natural Language Processing Paradigm for Balkan Languages, 26 September 2007, Borovets, Bulgaria. In: Paskaleva, E., Slavcheva, M. (eds.), pp. 65-71 (2007)

[Het12a] Y. Hetsevich, S. Hetsevich. *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 Intern. Conf. In: Vučković, K., Bekavac, B., Silberztein, M. (eds.), pp. 29-40. Cambridge Scholars Publishing, Newcastle (2012)

[Het12b] Y. Hetsevich, S. Hetsevich, B. Lobanov. *Belarusian and Russian linguistic modules processing for the system NooJ as applied to text-to-speech synthesis.* Computational Linguistics and Information Technologies : materials of the Int. Conf. "Dialogue", Bekasovo, May 30 – June 3, 2012. Issue 11 (18), vol. 1, pp. 198-212. Moscow (2012)

[NooJ02] *Linguistic Processor NooJ* [Electronic resource]. – 2002. – Mode of access: http://www.nooj4nlp.net/pages/nooj.html. – Date of access: 10.11.2012.

[RRB07] *Rules of the Road of Belarus* [Electronic resource]. – 2007. – Mode of access: http://pdd.by. – Date of access: 10.11.2012.

[BIPM06] *International Bureau of Weights and Measures BIPM* [Electronic resource]. – 2006. – Mode of access: http://www.bipm.org/en/si/si_brochure/general.html. – Date of access: 10.11.2012.