United Institute of Informatics Problems
of the National Academy of Sciences of Belarus

Processing of Quantitative Expressions with Units of Measurement in Scientific Texts as Applied to Belarusian and Russian Text-to-Speech Synthesis

Alena Skopinava, Yuras Hetsevich, Boris Lobanov

Scopes of Application

❖ **Corpora and Database Management Systems, Libraries, Information Retrieval Systems:**

- to formulate extended search queries,
- locate specific expressions on the Internet,
- support automatic text annotation and summarization;

❖ **Publishing Institutions:**

- to automatically locate specified lists of expressions with MUs,
- classify resulting expressions as SI units, their derivatives or units out of the SI,
- check quickly if the extended names of units are used correctly;

❖ **Text-to-Speech Synthesis Systems:**

- to generate orthographically correct texts and their tonal and prosodic peculiarities;

The main goal is

to develop algorithms and linguistic resources in order to **identify**, **classify** and **generate** measurement units (MU) and quantitative expressions (QE) with them on the material of hand-crafted text corpora for Belarusian and Russian;
+ to prove its importance for correct intonational marking.

SI derived unit of speed with a multiple prefix

Цягнік рухаўся з хуткасцю 200 км/г у Бекасава.
Поезд двигался со скоростью 200 км/ч в Бекасово.
‘The train was running at the speed of 200 km/h to Bekasovo.’

дзвесце кіламетраў у гадзіну
двести километров в час
‘two hundred kilometres per hour’

Three sets of algorithms for BE and RU:

1. ***Identification*** *&* ***Classification*** of QE with MU according to **the SI**

2. ***Identification*** *&* ***Classification*** of QE with MU according to **word formation**

3. ***Processing*** of QE with MU into **orthographical words**

Difficulties

- **The language-dependent origin;**

Bel. “ на” = Rus. “час” = Eng. “hour” = Ger. “Stunde”…

- **A great variety of numeral quantifiers and names of units, both in writing and formation;**

- intricate agreement:

25 метраў ‘25 meters’, 21 метр_ ‘21 meter’, 23 метры ‘23 meters’

- synonymy of written forms:

2000 метраў ‘2000 meters’ = 2000 м ‘2000 m’ = 2 10^3 м ‘2 10^3 m’ = 2 кіламетры ‘2 kilometers’ = 2 км ‘2 km’ = …

http://www.nooj4nlp.net/pages/nooj.html

Scientific-technical text corpora

BE

File Name	Size	Last Modif.
Kosmas_1_bel	53887	11.07.2012
Kosmas_2_bel	72228	11.07.2012
Kosmas_3_bel	77845	11.07.2012
Kosmas_4_bel	53016	11.07.2012
Kosmas_5_bel	51073	11.07.2012

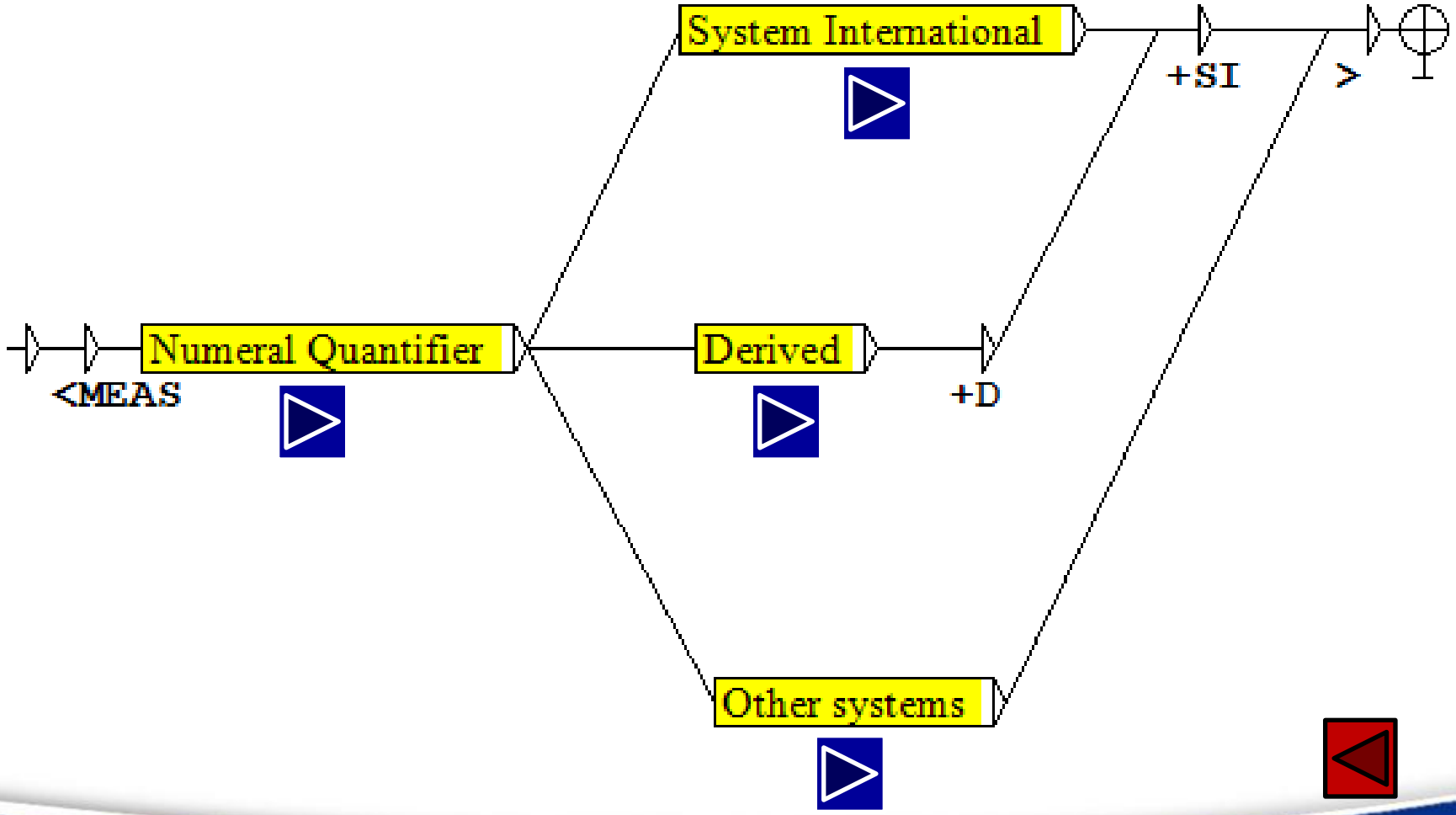
Кампаніяй DigitalGlobe эксплуатуецца КА высокага разрашэння QuickBird-2, які быў выведзены на арбіту вышынёй 450 км у 2001 г. Забяспечвае атрыманне панхраматычных малюнкаў з разрашэннем 0,64 м і мультыспектральных з разрашэннем 2,44 м у паласе захопу 16,6 км. Час актыўнага функцыянавання – 7 гадоў.
Францыя валодае двума КА SPOT (SPOT-4 і SPOT-5). Спадарожнік SPOT-4 функцыяну е з 1998 г. і забяспечвае атрыманне здымкаў з

RU

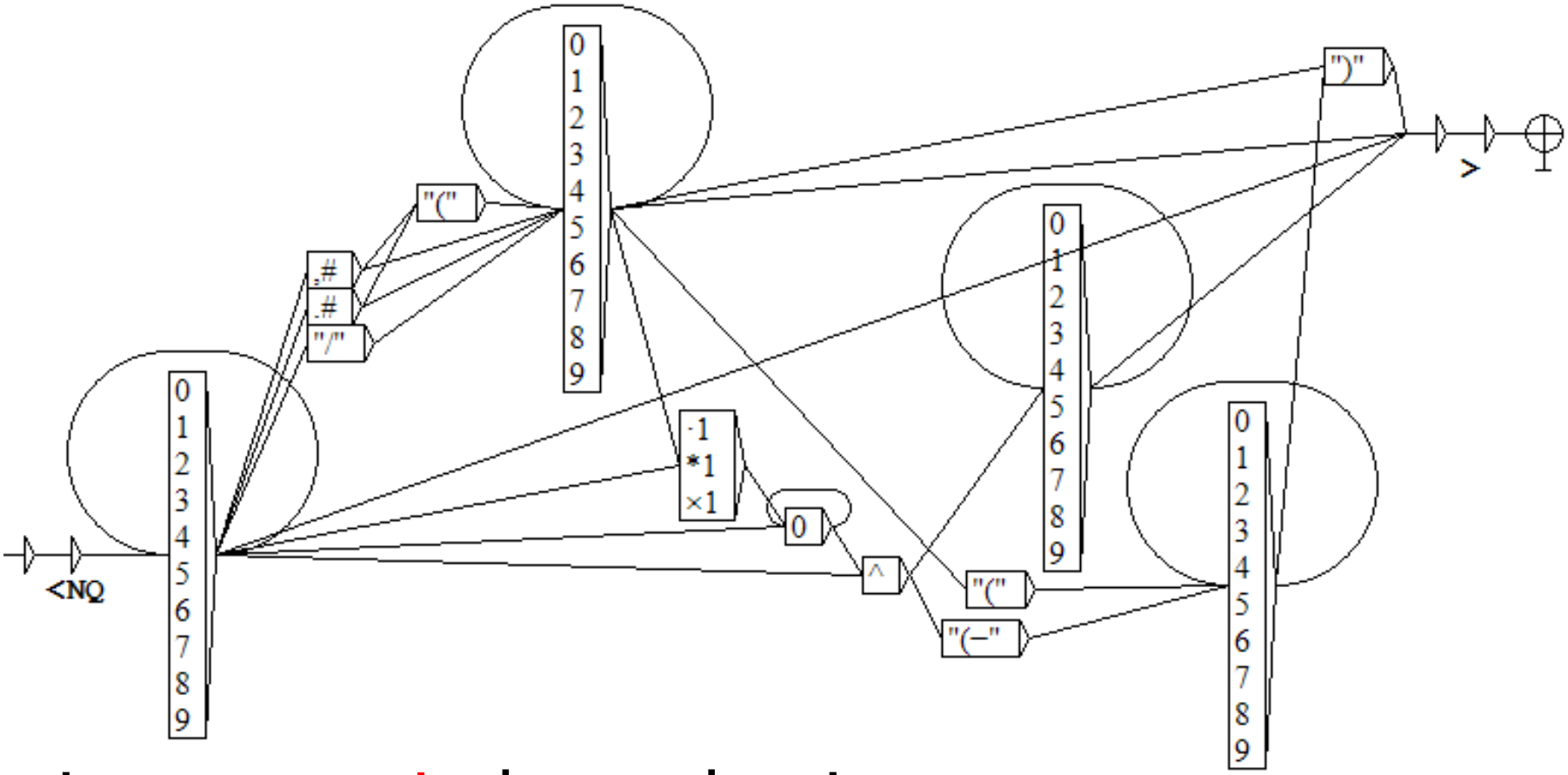
File Name
БИОЛОГИЯ
БИОТЕХНОЛОГИЯ
БИОФИЗИКА
БИОХИМИЯ
БОТАНИКА
ВУЛКАНОЛОГИЯ
ГЕНЕТИКА
ГЕОГРАФИЯ_пер...

эксперимента единственный выживший на глубине 2 см экземпляр имел 6 мм в длину и 3 - в ширину, тогда как самые крупные из 50 особей, выживших на глубине 20 см, достигали лишь 1.3 мм в длину и 0.77 мм в ширину. Следует отметить, что в каждой рамке находилось более тысячи семян. Первоначально проросли 13 % семян на глубине 1 см и 60% -

1. Identification & Classification of QE with MU according to the SI



Identification of Numeral Quantifiers



Language-Independent!

Results of NQ Identification

BE

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)	.

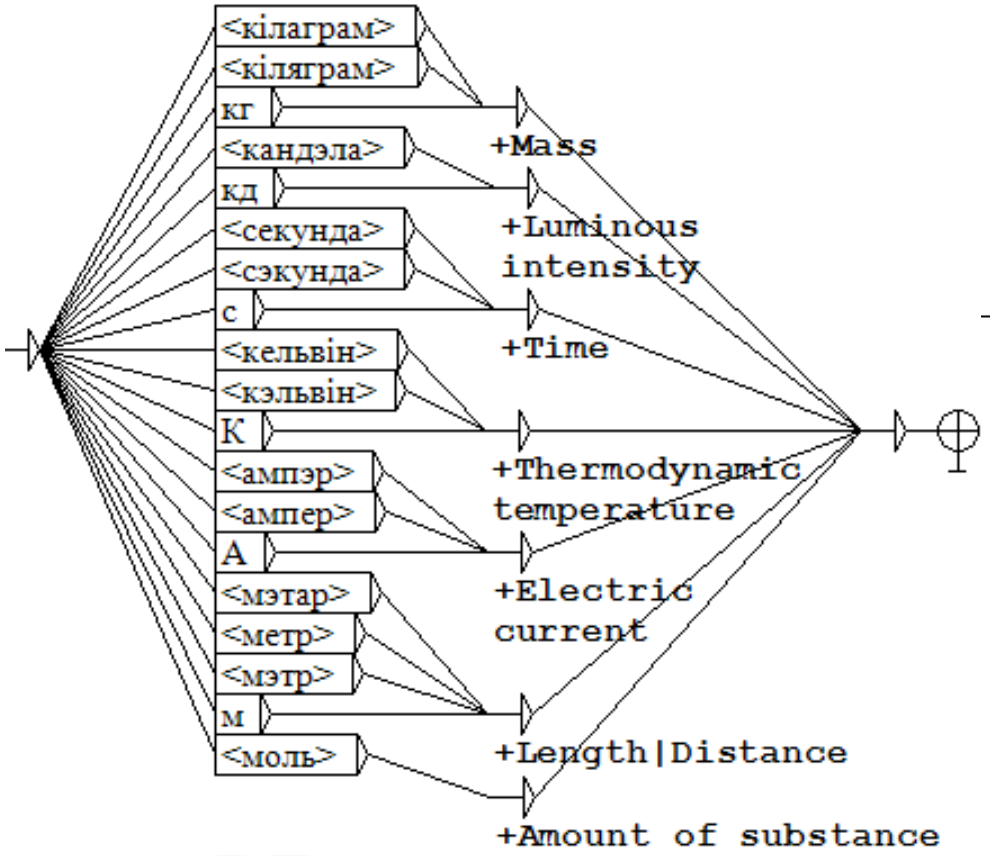
RU

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

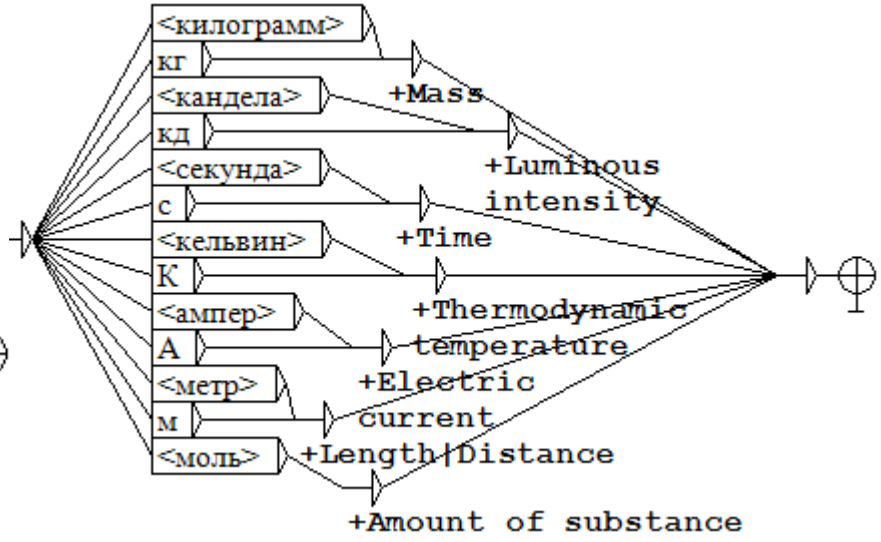
EN

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

Identification of the SI basic MU

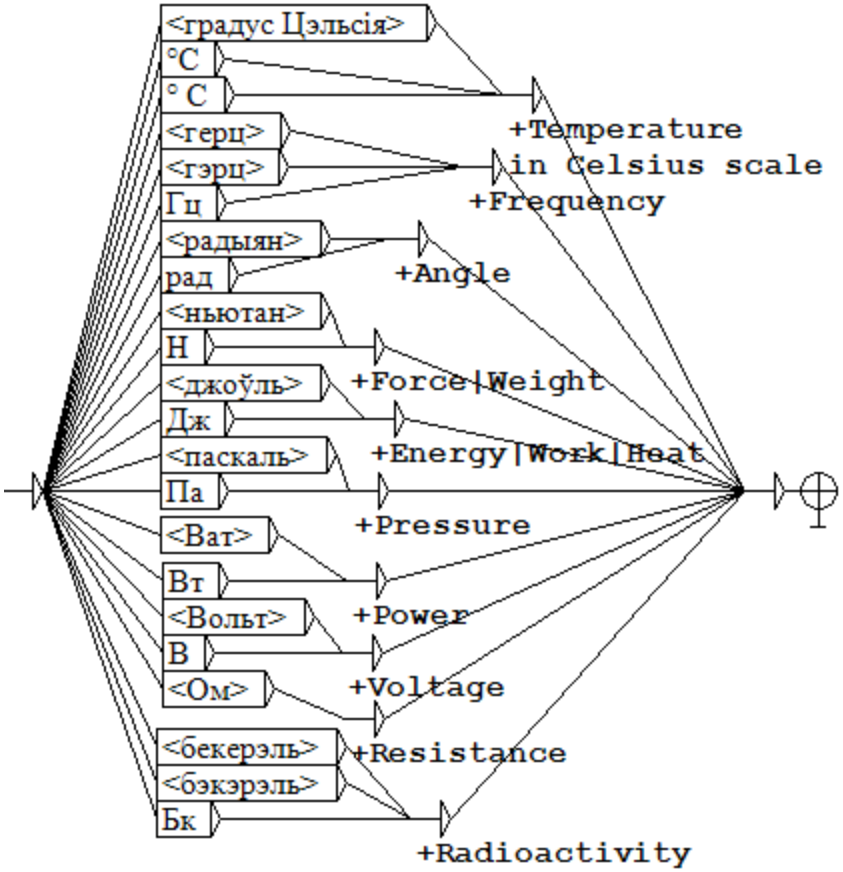


BE

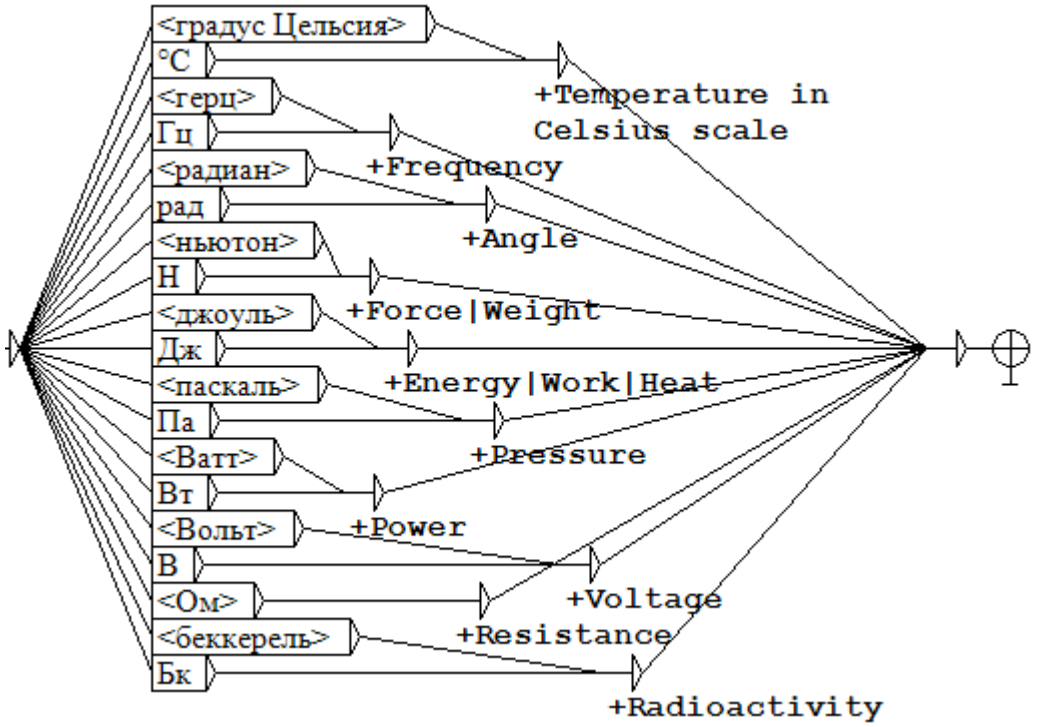


RU

Identification of the SI derived MU

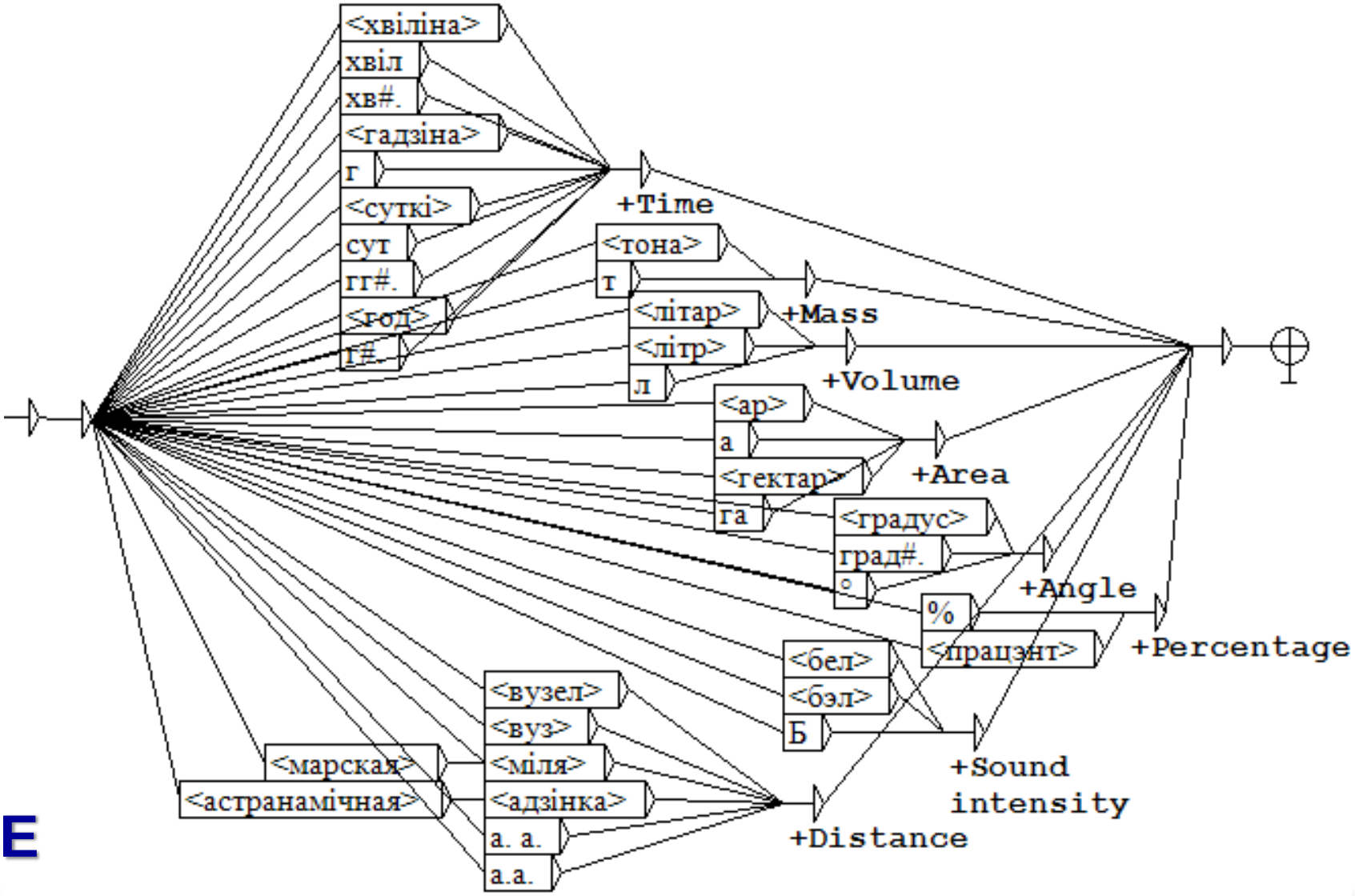


BE



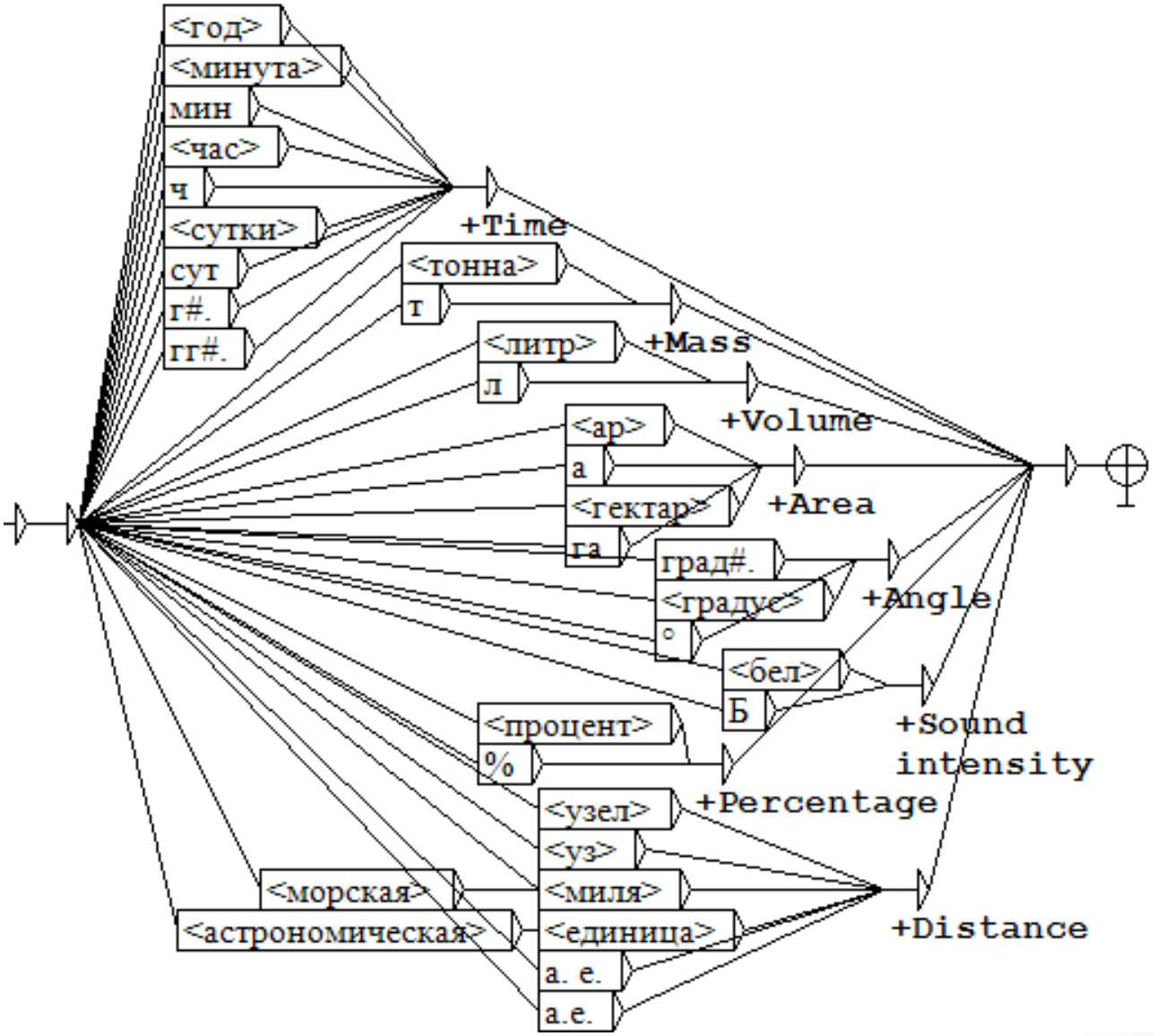
RU

Identification of MU out of the SI



BE

Identification of MU out of the SI



RU

Results of identification of QE with MU according to the SI for Bel & Rus

<MEAS>

BE

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

EN

1m
0,1 Hz
8 t
year 2005
74 degrees

RU

Before	Seq.	After
температуре	109 К/<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° С/<MEAS+Temperature in Celsius scal...	, а халькоге

109 K
200 000 l
33 years
5
600 C

Results of identification of QE with MU according to the SI for Bel & Rus

<MEAS+SI+D>

BE

Before	Seq.	After
МЭВ, 2-200 МЭВ, 2-30 кэв,	0,1 Гц	-300 кгц, 0-50 кгц; перыядычнасць
ткавай 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

EN

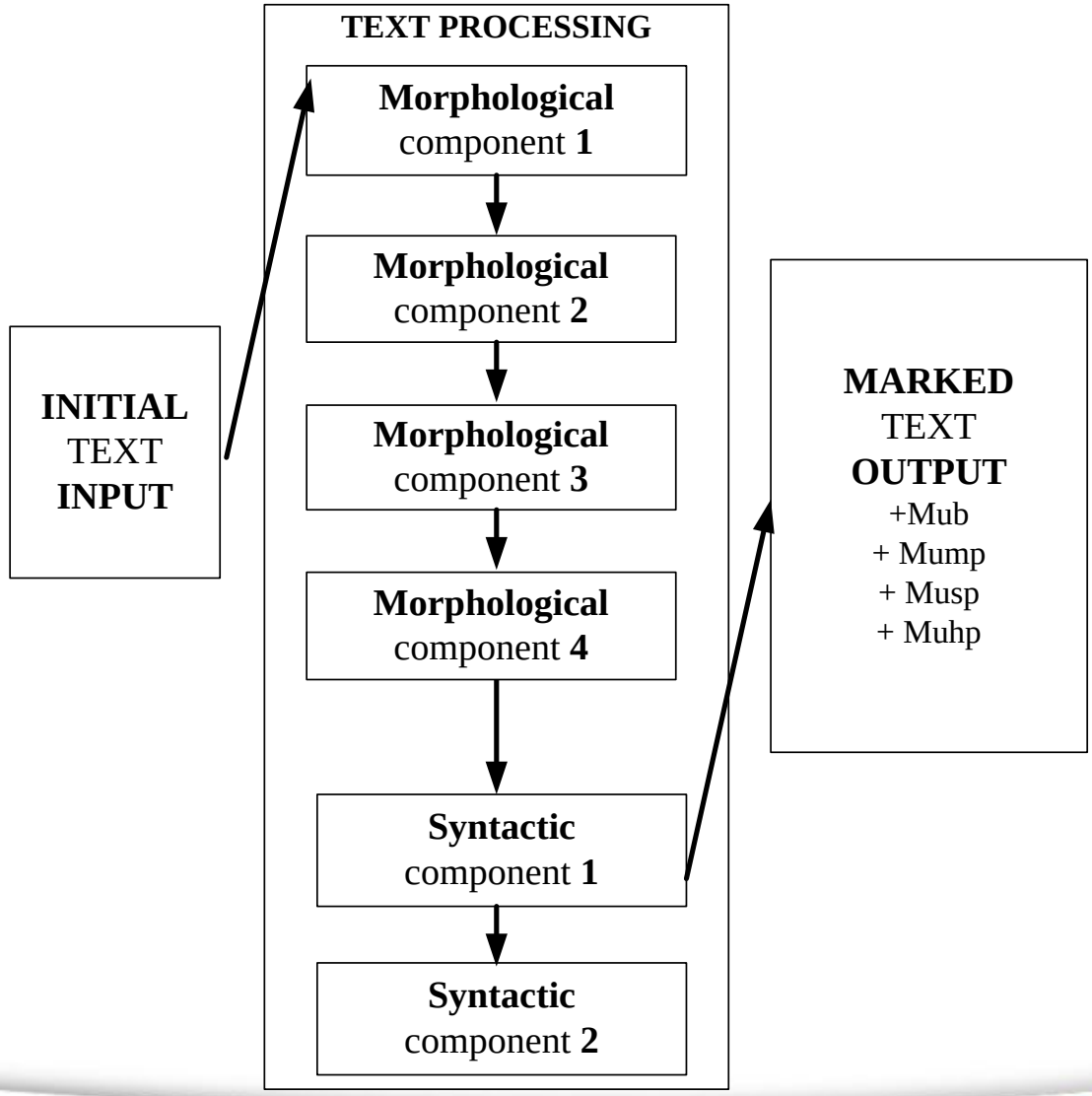
0,1 Hz
400 Pa
0 C
50 C
200 C
260 C

RU

Before	Seq.	After
МэВ, 2–200 МэВ, 2–30 кэВ,	0,1 Гц	–300 кГц, 0–50 кГц;
лазер накачки мощностью	25 Вт	возбуждает лазер
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

0,1 Hz
25 W
4,25 W
16,44 J
30,0 J
209,3 W

2. Identification & Classification of QE with MU according to Word Formation



Excerpts from the NooJ dictionaries of basic stems

BE

```
Б,ABBREVIATION+Mbase
байт,NOUN+FLX=БАЙТ+s2+UNAMB+Base
бекерэль,NOUN+FLX=АВАЛЬ+s6+UNAMB+Base
біт,NOUN+FLX=БАЙТ+s2+UNAMB+Base
В,ABBREVIATION+Mbase
Вт,ABBREVIATION+Mbase
ват,NOUN+FLX=БАЙТ+s2+UNAMB+Base
вольт,NOUN+FLX=БАЙТ+s2+UNAMB+Base
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
Гц,ABBREVIATION+Mbase
гц,ABBREVIATION+Mbase
```

RU

```
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
В,ABBREVIATION+Mbase
гектар,NOUN+FLX=АБАЖУР+s5+UNAMB+Base
га,ABBREVIATION+Mbase
герц,NOUN+FLX=ГЕРЦ+s2+UNAMB+Base
Гц,ABBREVIATION+Mbase
год,NOUN+FLX=ГОД+sN+UNAMB+Base
г,ABBREVIATION+Mbase
гг,ABBREVIATION+Mbase
град,ABBREVIATION+Mbase
грамм,NOUN+FLX=АНГСТРЕМ+s3+UNAMB+Base
```

Describing Inflectional Classes…

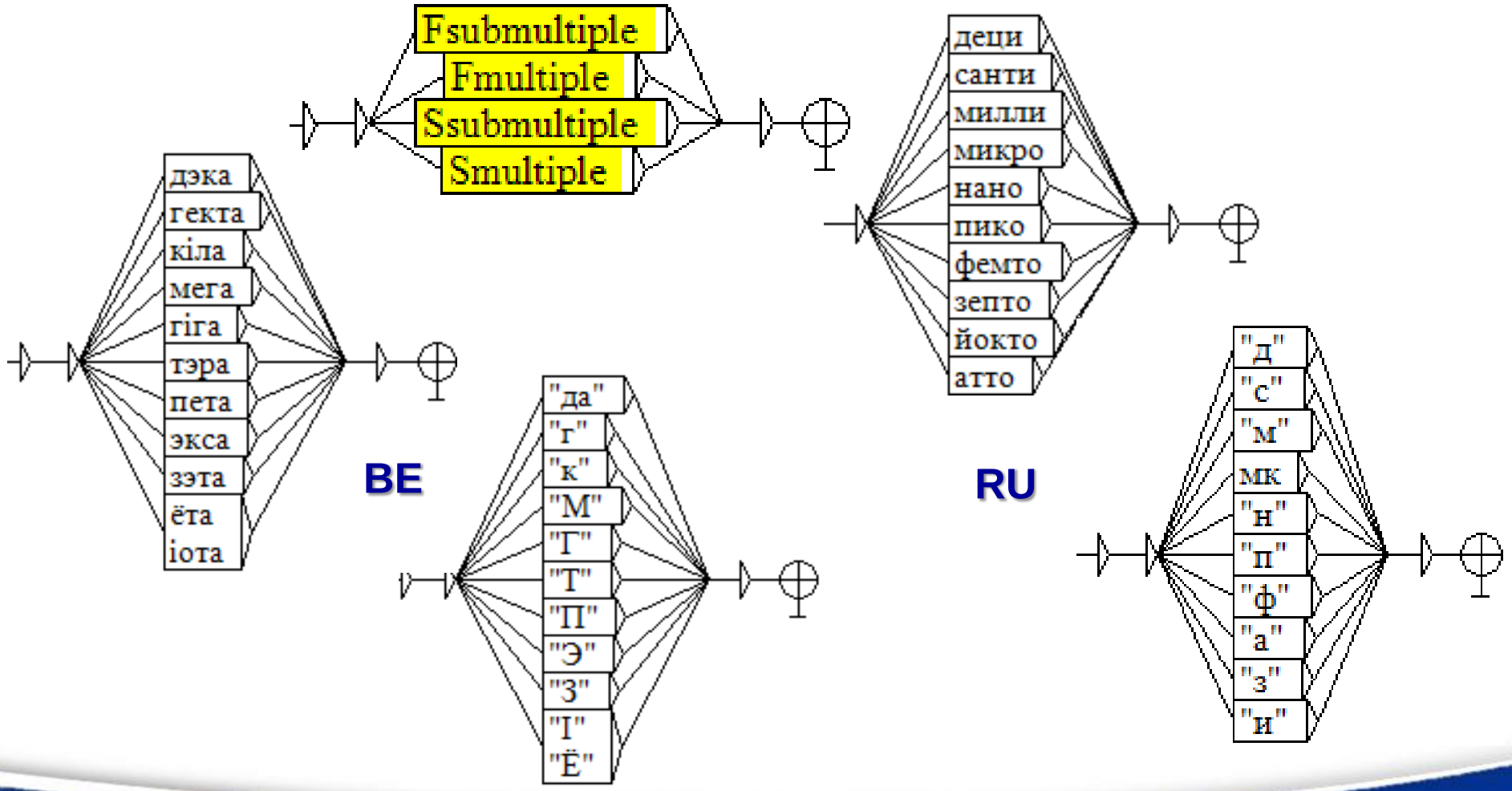
BE

```
ГЕКТАР =
<E>/Accusative+Common+Inanimate+Masculine
+ <E>/Common+Inanimate+Masculine+Nominative
+ <E>а/Common+Genitive+Inanimate+Masculine
+ <E>аў/Common+Genitive+Inanimate+Masculine+Plural
+ <E>ам/Common+Inanimate+Instrumental+Masculine
+ <E>ам/Common+Dative+Inanimate+Masculine+Plural
+ <E>амі/Common+Inanimate+Instrumental+Masculine+Plural
+ <E>ах/Common+Inanimate+Masculine+Plural+Prepositional
+ <E>ы/Common+Inanimate+Masculine+Prepositional
+ <E>у/Common+Dative+Inanimate+Masculine
+ <E>ы/Accusative+Common+Inanimate+Masculine+Plural
+ <E>ы/Common+Inanimate+Masculine+Nominative+Plural;
```

RU

```
АНГСТРЕМ =
<E>/Common+Genetive+Inanimate+Masculine+Plural
+ <E>/Accusative+Common+Inanimate+Masculine+Singular
+ <E>/Common+Inanimate+Masculine+Nominative+Singular
+ <E>а/Common+Genetive+Inanimate+Masculine+Singular
+ <E>ам/Common+Dative+Inanimate+Masculine+Plural
+ <E>ами/Common+Inanimate+Instrumental+Masculine+Plural
+ <E>ах/Common+Inanimate+Masculine+Plural+Prepositional
+ <E>е/Common+Inanimate+Masculine+Prepositional+Singular
+ <E>ов/Common+Genetive+Inanimate+Masculine+Plural
+ <E>ом/Common+Inanimate+Instrumental+Masculine+Singular
+ <E>у/Common+Dative+Inanimate+Masculine+Singular
+ <E>ы/Accusative+Common+Inanimate+Masculine+Plural
+ <E>ы/Common+Inanimate+Masculine+Nominative+Plural;
```

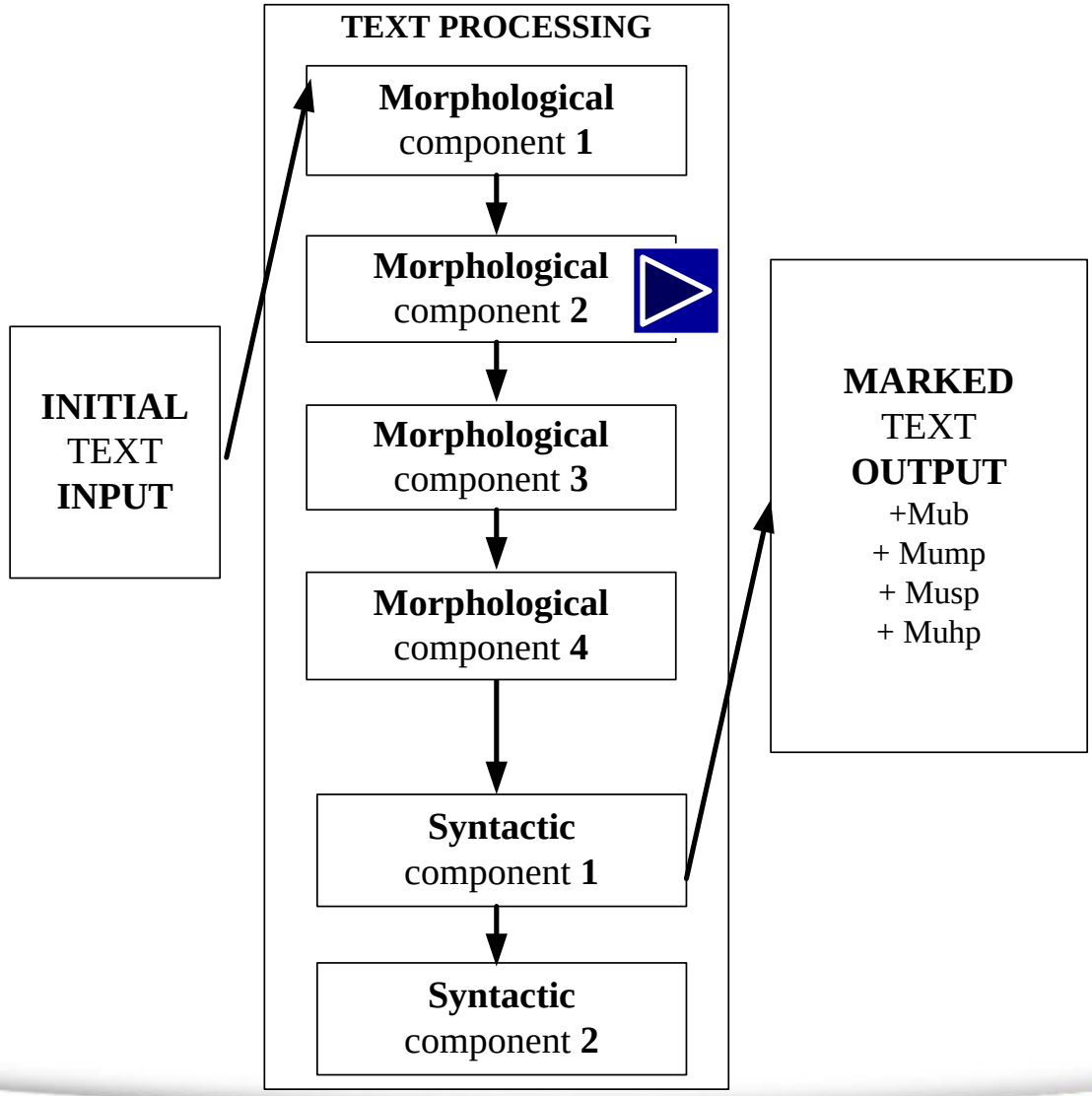
Databases of Prefixes according to the SI



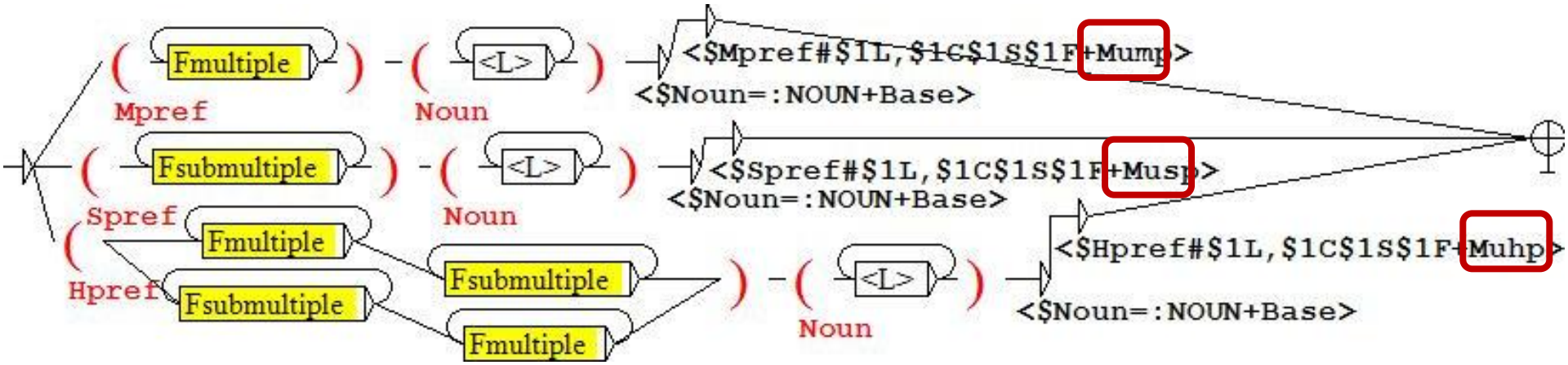
Word-formative classification of MU:

1. with full-form stems and without prefixes
   метр, Герц, Ом *(eng.* ***meter, Hertz, Ohm****)*
2. with full-form stems and full-form prefixes
   нанафарады, міліампер
   (eng. ***nanofarads, milliampere****)*
3. with full-form stems and shortened prefixes
   кБайт *(eng.* ***Kbyte****)*
4. with shortened stems and without prefixes
   Дж, га, Па *(eng.* ***J, ha, Pa****)*
5. with shortened stems and shortened prefixes
   км, дл, гПа *(eng.* ***km, dL, hPa****)*

2. Identification & Classification of QE with MU according to Word Formation



M2 identifies full-stem MU with full multiple and/or submultiple prefixes



M2-operation results

BE **<NOUN+Muhp>**

Before	Seq.	After
дзвюх ці больш прыставак (напр.,	мікрамегафарад	) не дазваляецца.

EN

micromegafarads

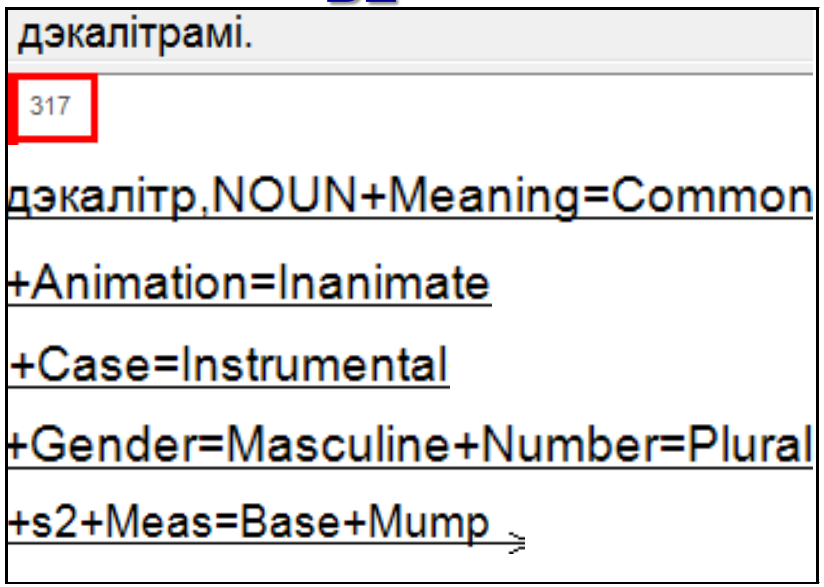
RU **<NOUN+Mump>**

Before	Seq.	After
на расстоянии несколько сотен	километров	. Первый вариант ударного
кусков может достигать нескольких	килограммов	. Куски брони поражают
излучения мощностью в сотни	мегаватт	. Проблема в том
составляет уже десятки тысяч	мегагерц	, что соответствует волнам
своих жестких дисках тысячи	гигабайт	информации, третьи подкл
энергии лазерного излучения порядка	мегаджоуля	(106 Дж) и кпд

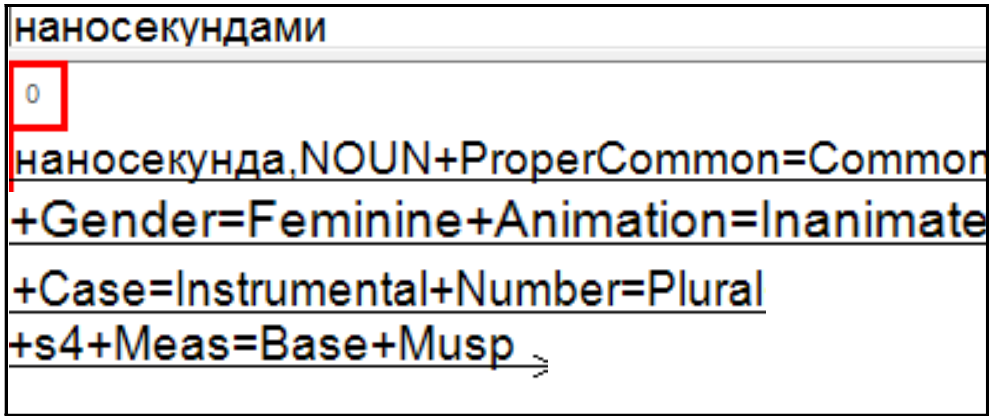
kilometres
kilograms
megawatt
megahertz
gigabyte
megajoule

Examples of annotated word forms

BE



RU

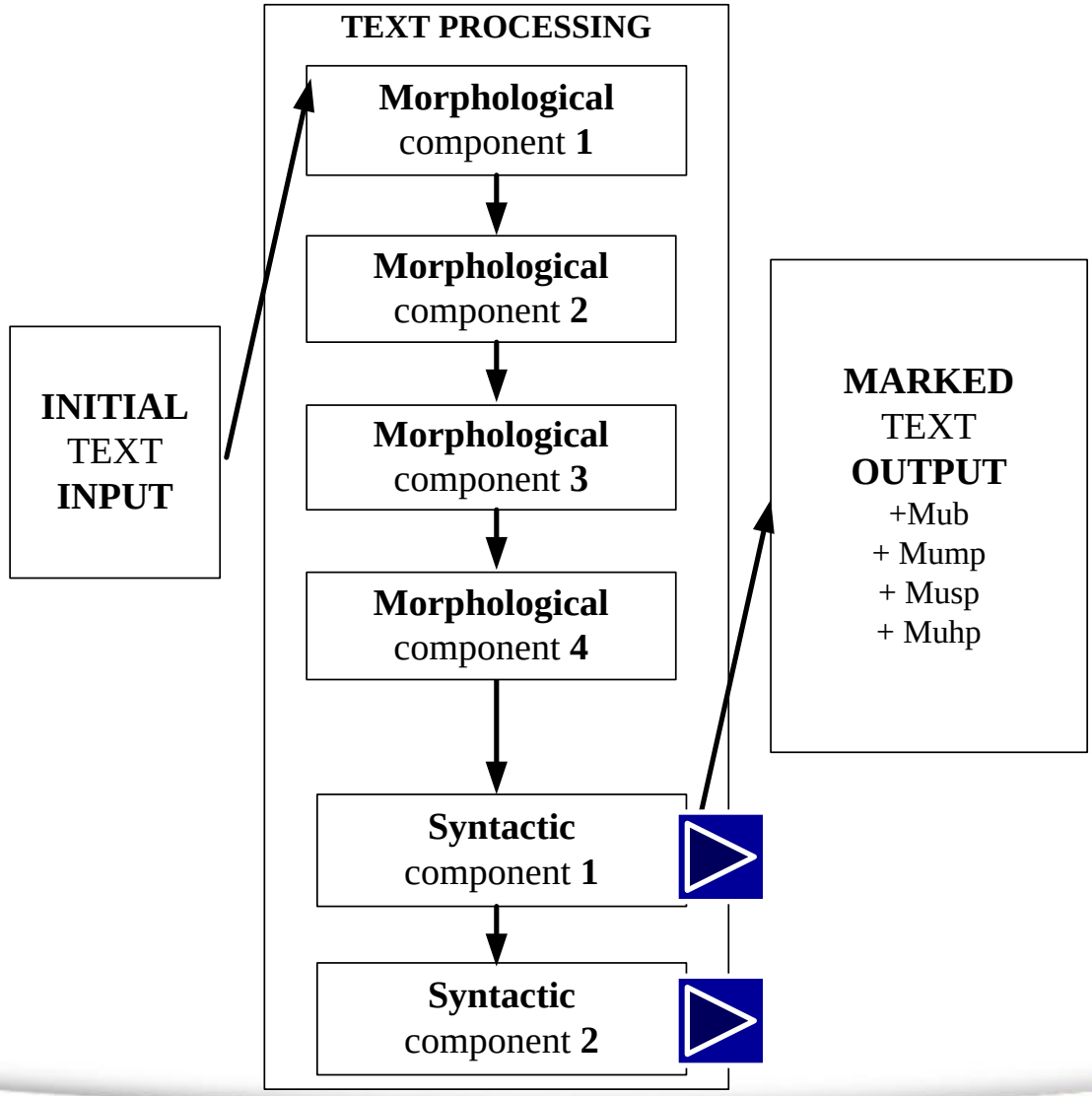


There are **no** word forms “дэкалітрамі” (Eng. decaliters) “наносекундами” (Eng. nanoseconds) **in the resource dictionary**.

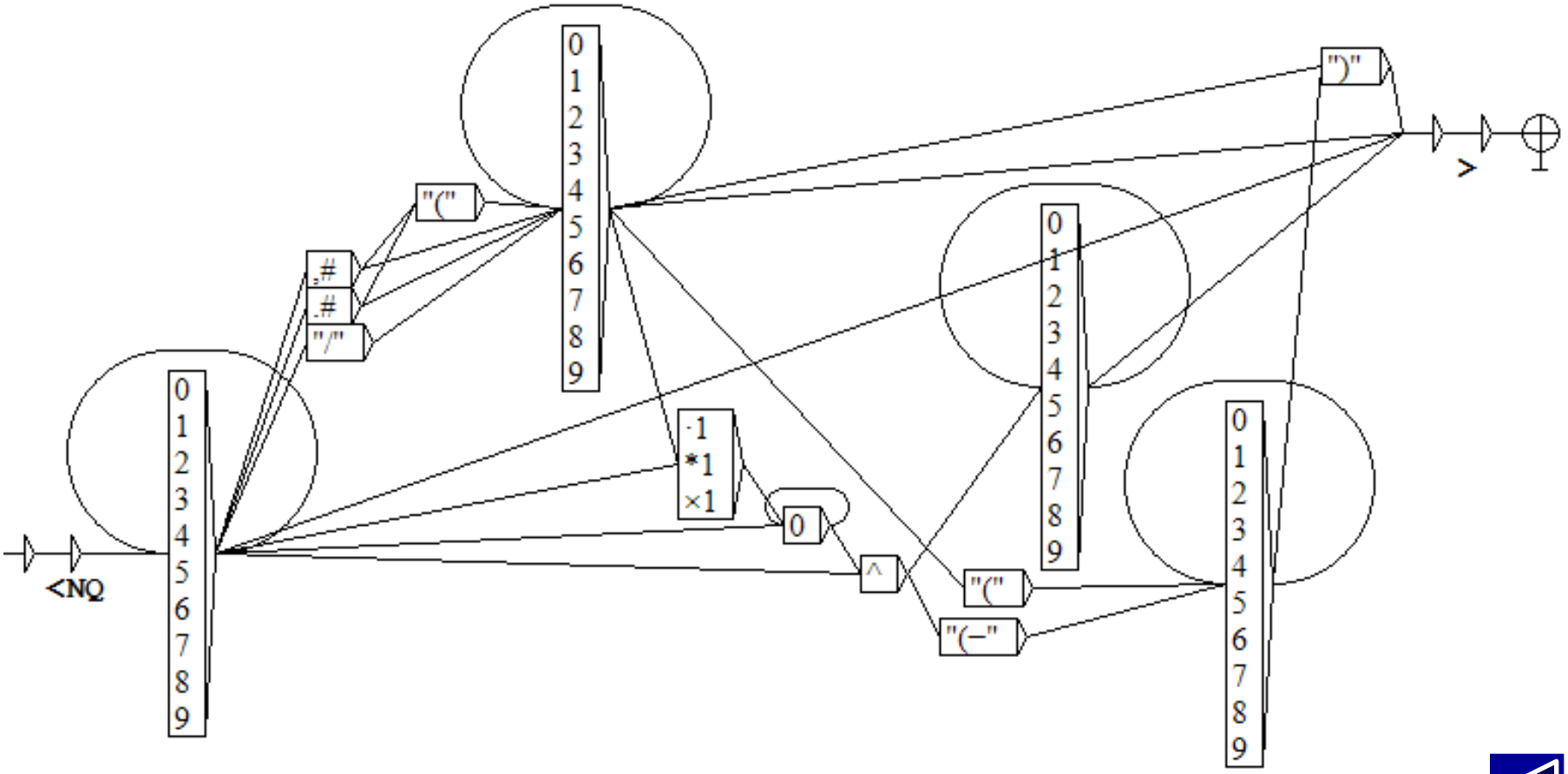
BUT!

All the morphological **data are preserved** by the algorithm (according to the grammatical characteristics of the basic stem).

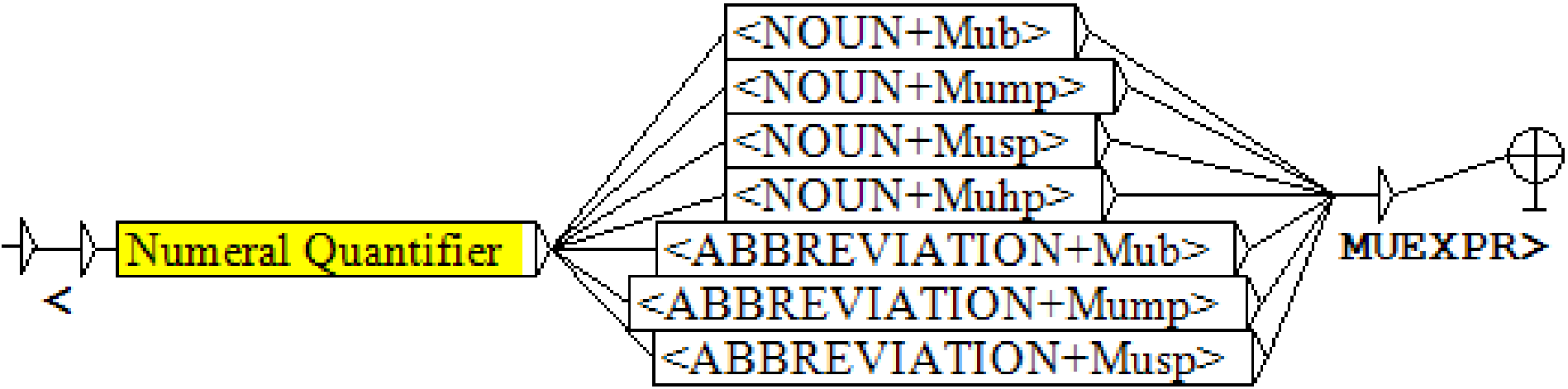
2. Identification & Classification of QE with MU according to Word Formation



S1 for Numeral Quantifiers is the same.



S2 collects M1-M4 data and identifies QE with MU



Excerpts of the results obtained with the S2

BE

<MUEXPR>

Before	Seq.	After
масы - грам (0,001 кг).	31 мкТл	(3,1×10^(-5) Тл) - напружанасць
йна вар'іруецца зблізку	2,4 мЗв	у год. 1 Н ёсць
на апору з сілай	9.81 Н	. Прыбліжэнне, што 1 кг адпавяд
нанафарадамі (пішуць	60 000 пф	, а не 60 нф; 2 000 мкф
асць шара з радыусам	1 сантыметр	, змешчанага ў вакуум. 1 сантым
Mbit(альбо проста Mb).	1 мегабіт	= 1000^2 біт = 10^6 біт = 10000
святло ў вакууме за (	1 / 299 792 458) секунды	. Метр быў упершыню ўведзены
азон - 40 кэв-3 МЭВ, 2-	200 МЭВ	, 2-30 кэв, 0,1 Гц-300 кгц, 0-50 кг
ды тэлевізара - парадку	20 кілаэлектронвольт	. Энэргіі касьмічных прамянёў - а
сьмічных прамянёў - ад	1 мегаэлектронвольта	да1000 тэралектронвольтаў.

EN

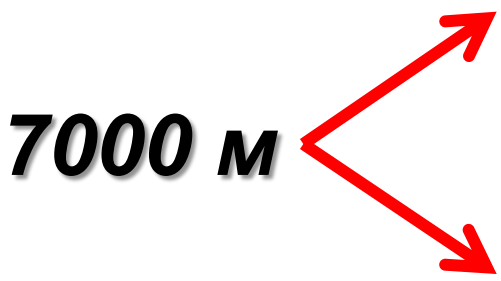
31 μT
2,4 mZv
9.81 N
60 000 pF
1 sm
1 megabit
(1/299 792 458) seconds
200 MeV
20 kiloelectronvolts
1 megaelectronvolt

RU

Before	Seq.	After
организм ток не превышал	1 мА	. На человека токи статиче
могут сказать «файл в	100 килобайт	»). При обозначении скоро
тивление величиной от 1 до	100 МОм	, чтобы протекающий чере
омегафарад пикотеравольт	13 йоттайоктограммов	Каждая строка содержит
до 64 Мбит/с) и	137,4 МГц	(метровый диапазон, фор
онширский изумруд» массой	1383,95 каратов	. Изумруды выращивают
ремя жизни мюонов - около	2.2 мкс	- осложняет задачу созда
а которой оказалась равной	22 фемтограммам	(1 фг = 1•10^(-15) г). . Мю
то они оказались равными	8.1·10^21 Дж	(уменьшение массы ледн
высота 670 км - наклонение	98,00 град	. Срок активного существо

1mA
100 kilobytes
100 Mohm
13
yottayoktograms
137,4 MHz
1383,95 carats
2.2 μs
22 femtograms
8.1·10^21 J
98,00 deg

3. Generating QE with MU into orthographical words

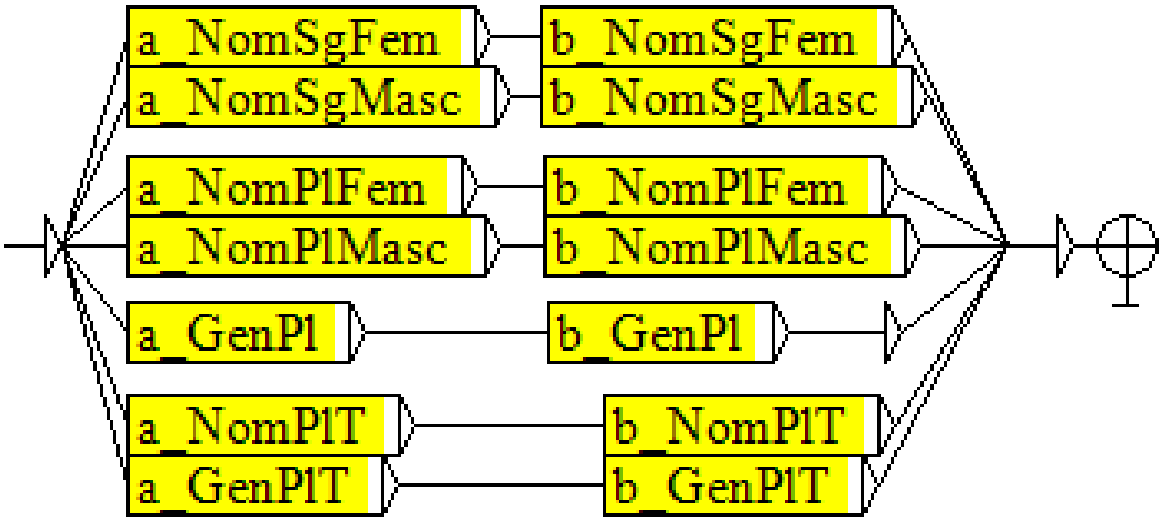


сем тысяч метраў BE

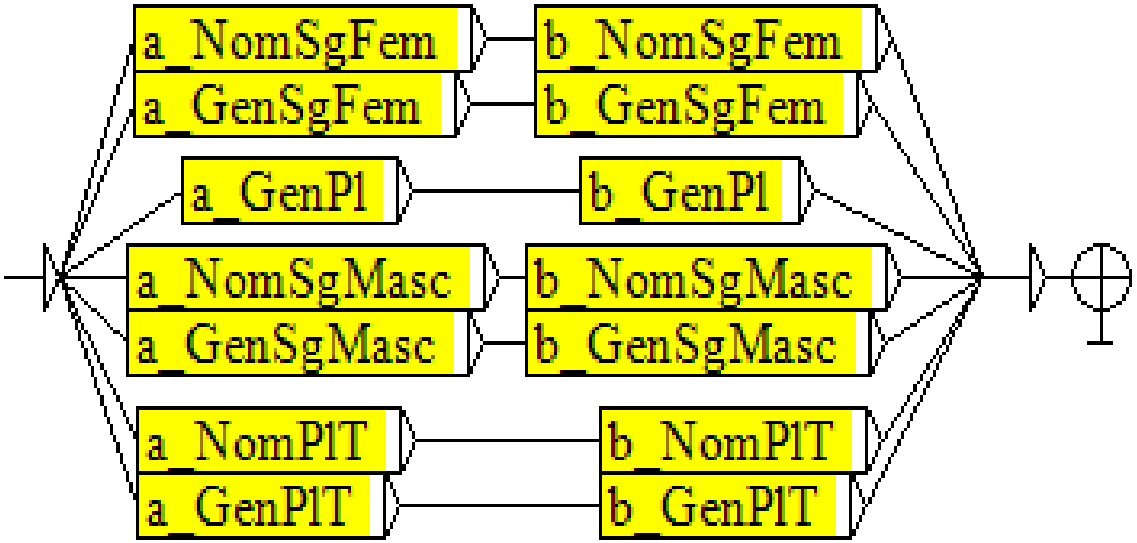
семь тысяч метров RU

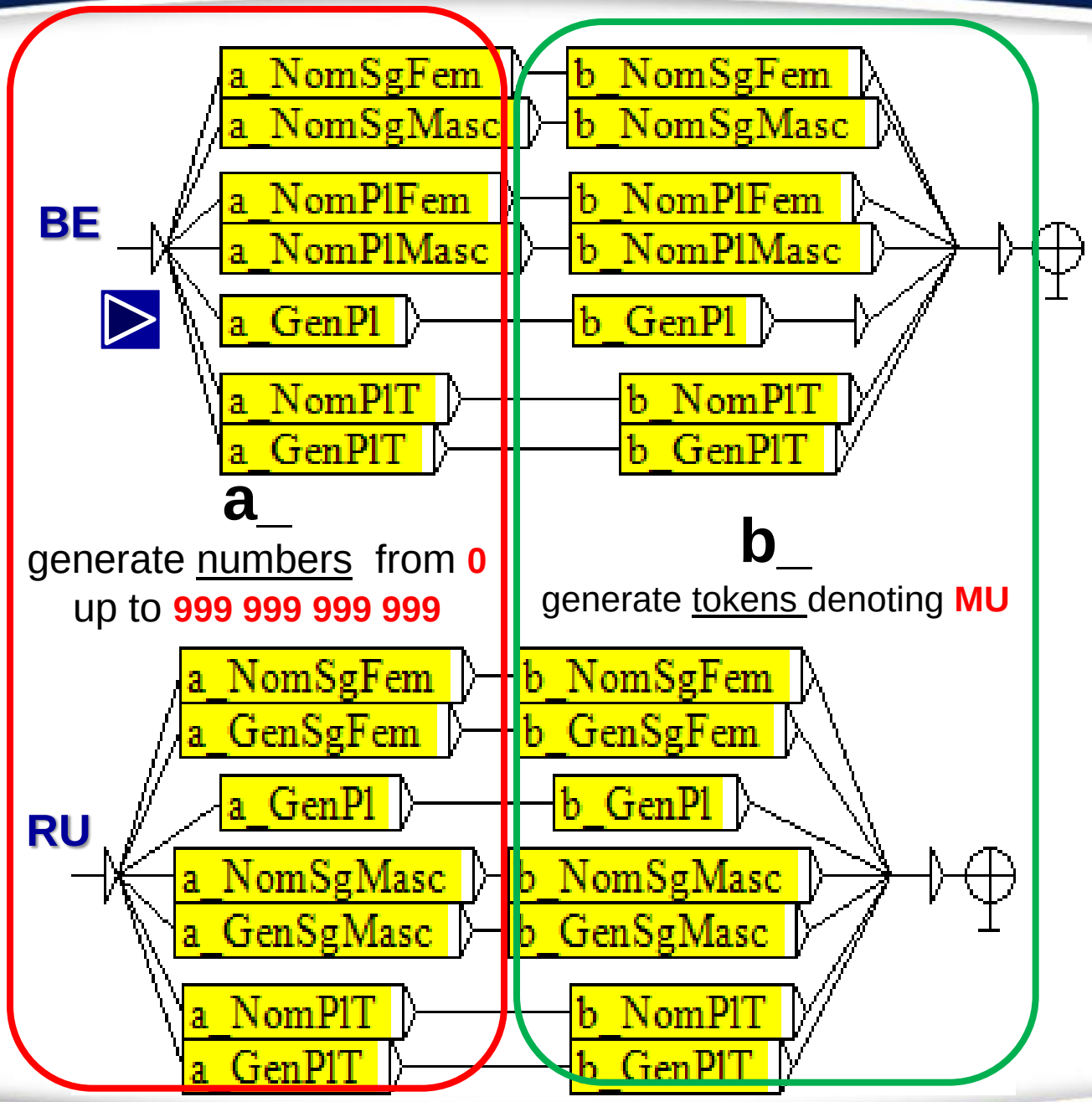
7000 m → seven thousand metres EN

BE


a_NomSgFem
a_NomSgMasc
a_NomPlFem
a_NomPlMasc
a_GenPl
a_NomPlT
a_GenPlT
b_NomSgFem
b_NomSgMasc
b_NomPlFem
b_NomPlMasc
b_GenPl
b_NomPlT
b_GenPlT

RU


a_NomSgFem
a_GenSgFem
a_GenPl
a_NomSgMasc
a_GenSgMasc
a_NomPlT
a_GenPlT
b_NomSgFem
b_GenSgFem
b_GenPl
b_NomSgMasc
b_GenSgMasc
b_NomPlT
b_GenPlT


BE
a_NomSgFem
a_NomSgMasc
a_NomPlFem
a_NomPlMasc
a_GenPl
a_NomPlT
a_GenPlT
b_NomSgFem
b_NomSgMasc
b_NomPlFem
b_NomPlMasc
b_GenPl
b_NomPlT
b_GenPlT
a_
generate numbers from 0
up to 999 999 999 999
b_
generate tokens denoting MU
RU
a_NomSgFem
a_GenSgFem
a_GenPl
a_NomSgMasc
a_GenSgMasc
a_NomPlT
a_GenPlT
b_NomSgFem
b_GenSgFem
b_GenPl
b_NomSgMasc
b_GenSgMasc
b_NomPlT
b_GenPlT

Peculiarities of the inflection of nouns after numerals

1. After number ***1*** (or with *1* as a final digit) nouns take endings of the Nominative Singular ***(NomSg)***. QE will proceed to branches **1** or **2**, depending on the gender of nouns, in particular *Masculine* ***(Masc)*** or *Feminine* ***(Fem)***.

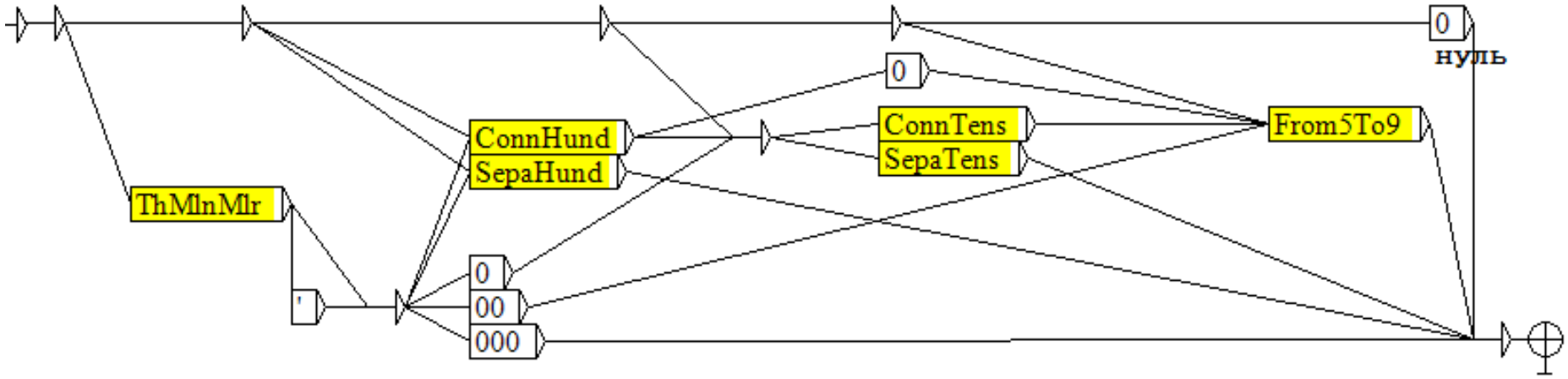
2. After numbers ***2, 3, 4*** (or with *2*, *3* or *4* as a final digit) nouns take the Nominative plural ***(NomPl)*** in Belarusian, whereas in the Russian these numbers require nouns in the Genetive singular ***(GenSg)***. Depending on the gender, QE will move to branches **3** or **4**.

3. Numbers **from 5 to 19** and **round** numbers (or with them as final digits) require nouns in the Genitive plural ***(GenPl)*** in both languages. QEs will follow the **5**th branch. Branches **6** and **7** are for pluralia tantum nouns.

4. Special inflection is demanded by **pluralia tantum nouns**: ***сутки*** and ***сутки*** (Eng. *twenty-four hours*). They take branches **6** or **7** .

a_GenPl

from 0 to 999 999 999 999
Genitive plural



BE

Scheme of the BE ramified algorithmic complexes

Triad 1
0-999

Triad 2
to 999999

Triad 3
to 999 999 999

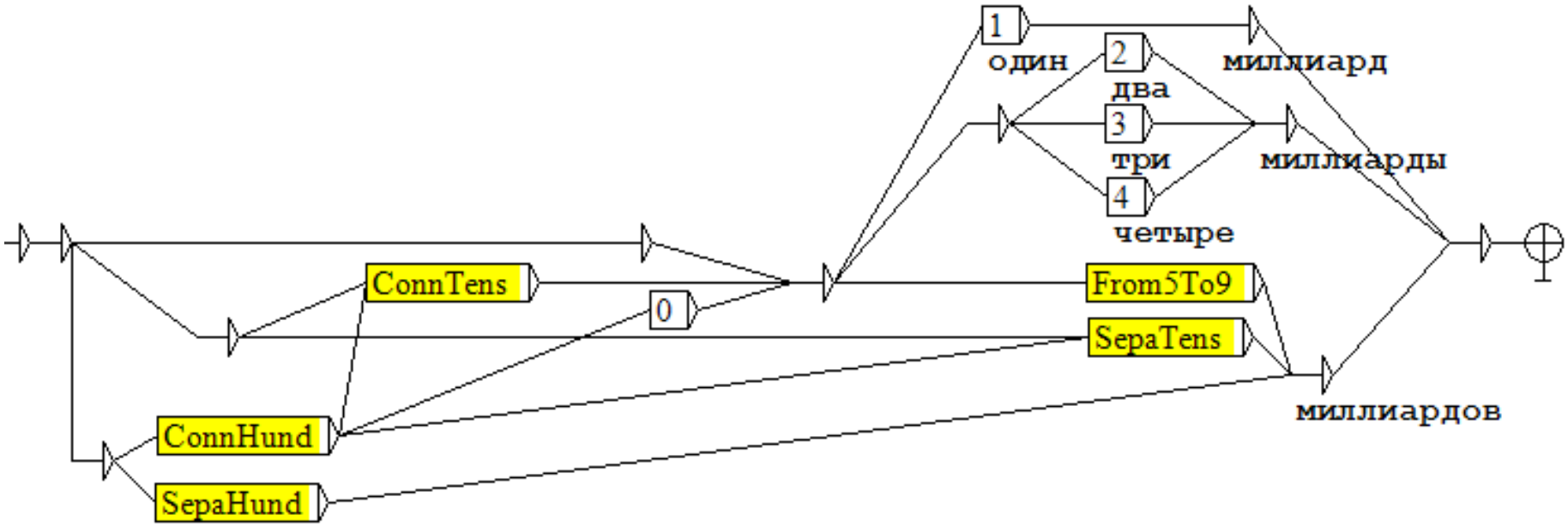
Triad 4
to 999 999 999 999

...

- Main
  - a_GenPl
    - ConnHund
    - ConnTens
    - From5To9
    - SepaHund
    - SepaTens
  - ThMlnMlr
    - ConnHund
    - ConnTens
    - From5To9
    - MlnMlr
      - ConnHund
      - ConnTens
      - From5To9
      - Mlr
        - ConnHund
        - ConnTens
        - From5To9
        - SepaHund
        - SepaTens
      - SepaHund
      - SepaTens
    - SepaHund
    - SepaTens

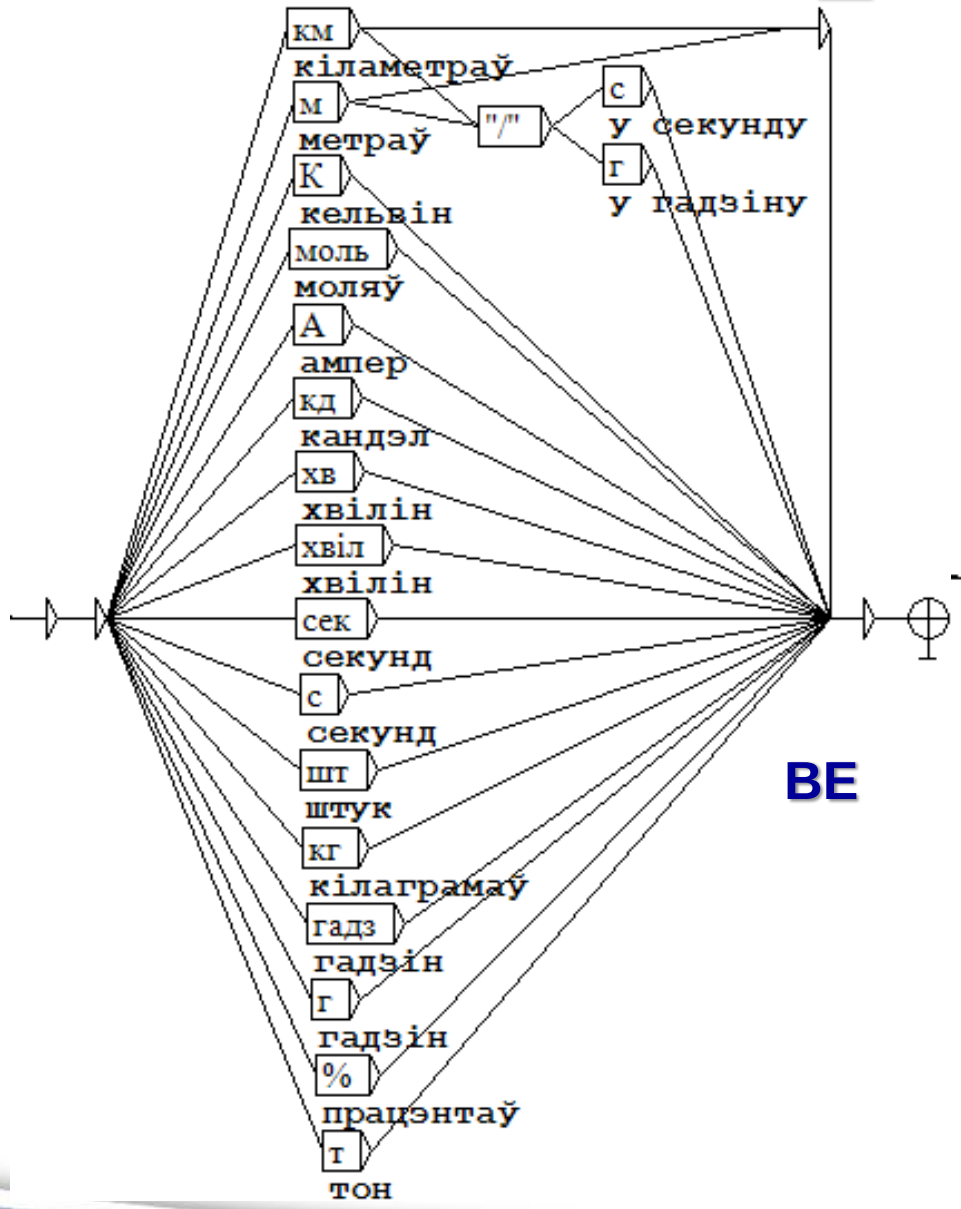
Mlr

It generates any whole number from the class of **billions**.

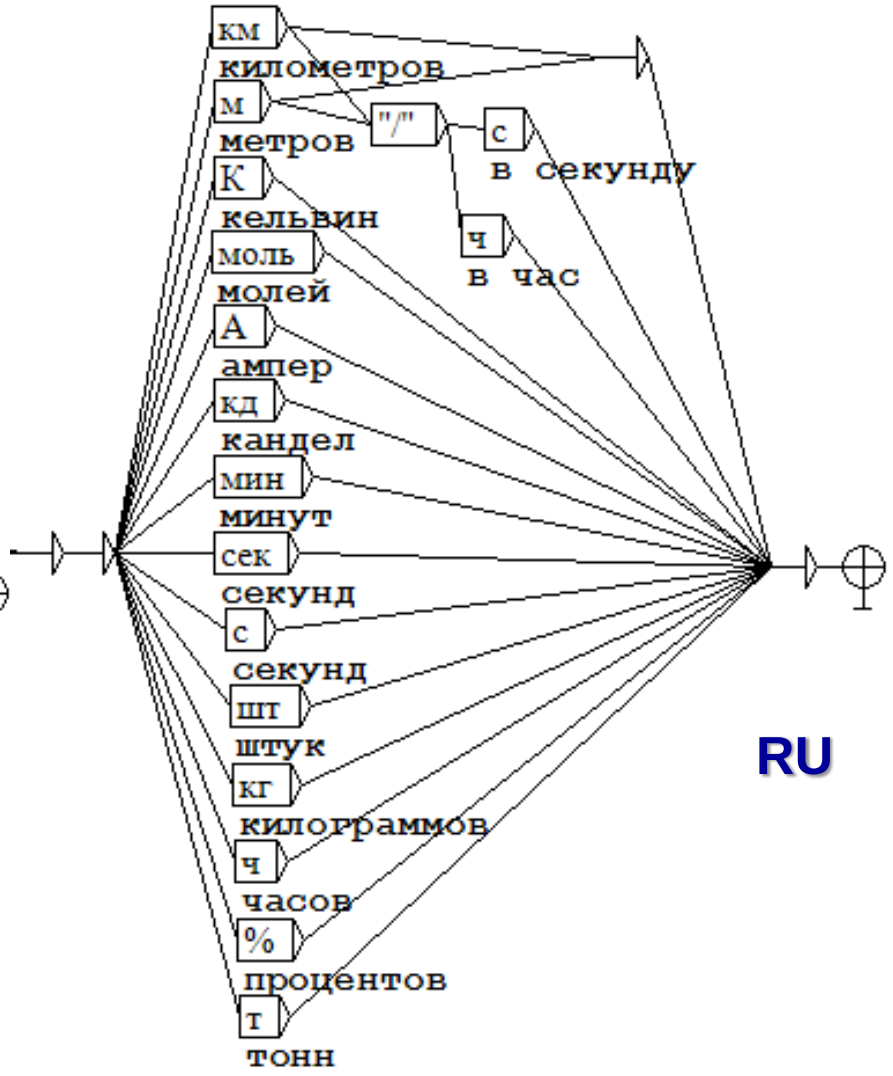


RU

b_GenPl


км
кіламетраў
м
метраў
"/"
с
у секунду
г
у гадзіну
К
кельвін
моль
моляў
А
ампер
кд
кандэл
хв
хвілін
хвіл
хвілін
сек
секунд
с
секунд
шт
штук
кг
кілаграмаў
гадз
гадзін
г
гадзін
%
працэнтаў
т
тон

BE


км
километров
м
метров
"/"
с
в секунду
ч
в час
К
кельвин
моль
молей
А
ампер
кд
кандел
мин
минут
сек
секунд
с
секунд
шт
штук
кг
килограммов
ч
часов
%
процентов
т
тонн

RU

Excerpts of results of generating QE with MU into orthographical words

BE

Seq.
15 т/пятнаццаць тонаў
123 кг/сто дваццаць тры кілаграмы
5678904 К/пяць мільёнаў шэсцьсот семдзесят восем тысяч дзевяцьсот чатыры кельвіны
123 А/сто дваццаць тры амперы
6785672 м/шэсць мільёнаў семсот восемдзесят пяць тысяч шэсцьсот семдзесят два метры
787879 сут/семсот восемдзесят сем тысяч восемсот семдзесят дзевяць сутак
6761 сут/шэсць тысяч семсот шэсцьдзесят адны суткі
999999994 моль/дзевяцьсот дзевяноста дзевяць мільёнаў дзевяцьсот дзевяноста дзевяць тысяч дзевяцьсот дзевяноста чатыры молі

RU

6661с/шесть тысяч шестьсот шестьдесят одна секунда
77700 т/семьдесят семь тысяч семьсот тонн
800009кд/восемьсот тысяч девять кандел
120202 мин/сто двадцать тысяч двести две минуты
8600км/ч/восемь тысяч шестьсот километров в час
903 м/с/девятьсот три метра в секунду

The regular expression for a search for QE

(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>x<NB>)|(<NB>x<NB>')|
(<NB>x<NB>x<NB>)|(<NB>.<NB>x<NB>.<NB>)|(<NB>,<NB>×<NB>)|(<NB>¸<NB><NB>)|

(<WF><NB>)

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
(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

Distribution of various ways of expressing QE in thematically distinct texts

Text	Physics	Space Travel Science	Geography	Military Equipment	Mineralogy	Botany	Transportation Communications	History
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 (Prosodic) marking

in sentences with QE/MU

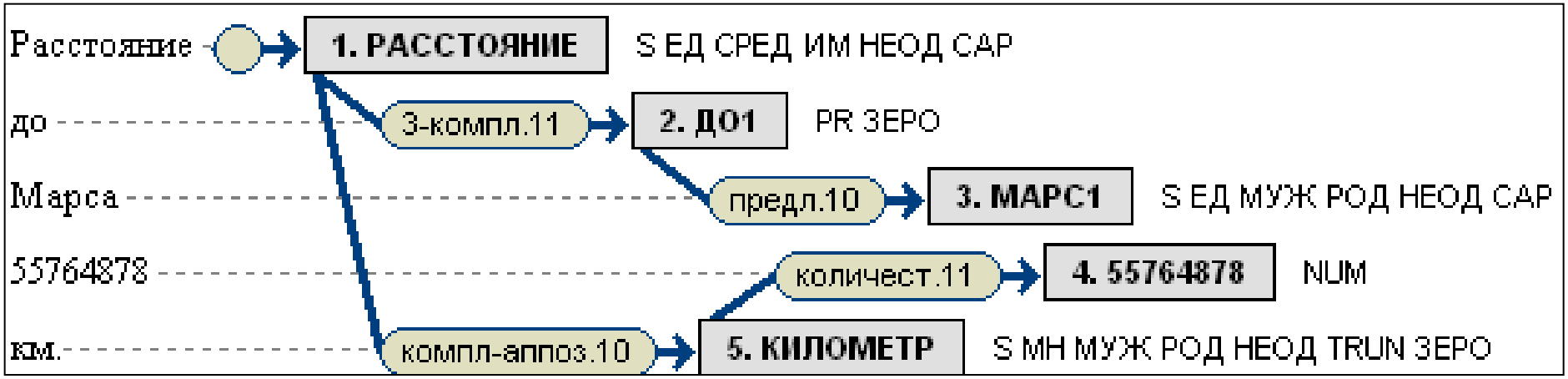
- dividing into syntagmas
- marking emphatically highlighted words
- indicating syntagmas with accent units
- creating a melodic contour of each syntagma

BUT!

First texts are reduced to
a normalized **orthographic** form.

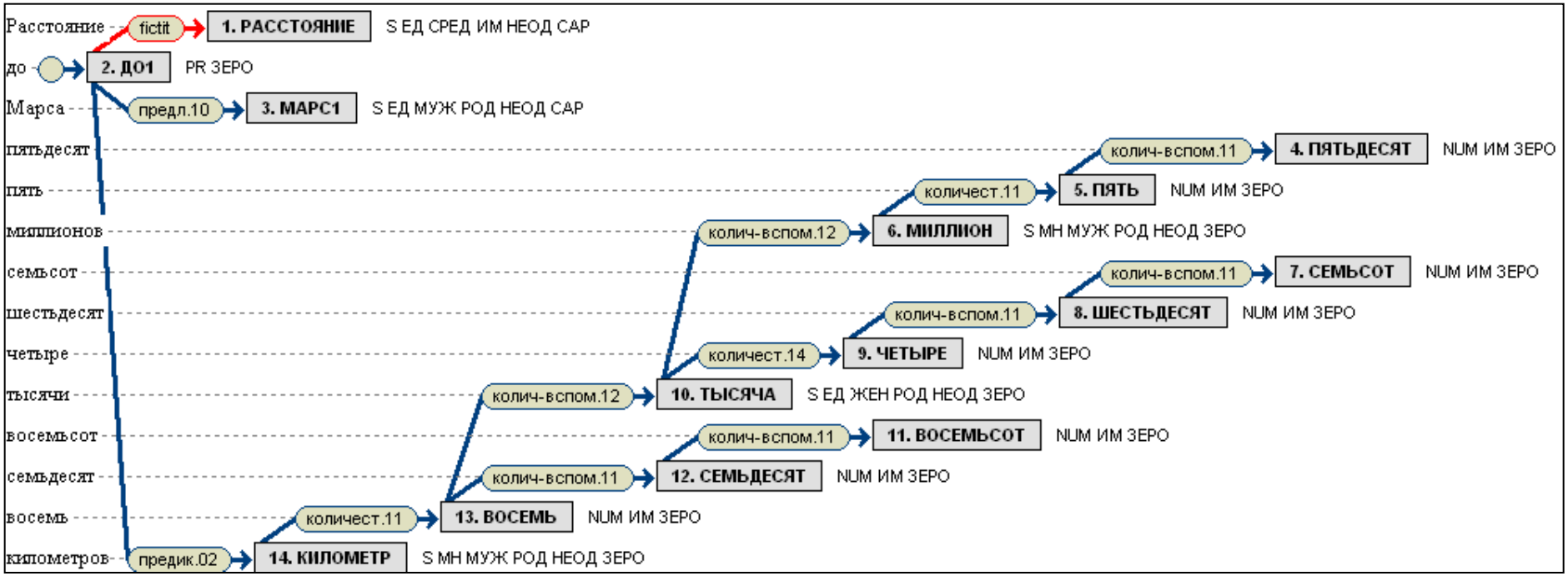
Расстояние до Марса 55764878 км

‘The distance to Mars is 55764878 km’



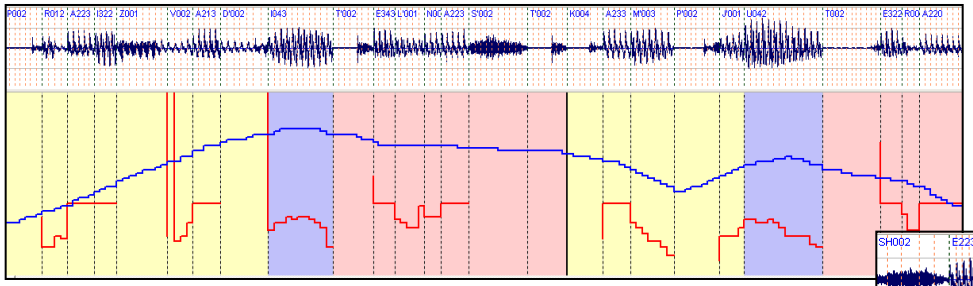
Расстояние до Марса пятьдесят пять миллионов семьсот шестьдесят четыре тысячи восемьсот семьдесят восемь километров

*'The distance to Mars is **fifty-five million seven hundred sixty-four thousand eight hundred seventy-eight kilometers**'*



Производительность компьютера **65.5 Мбит/с** была достигнута

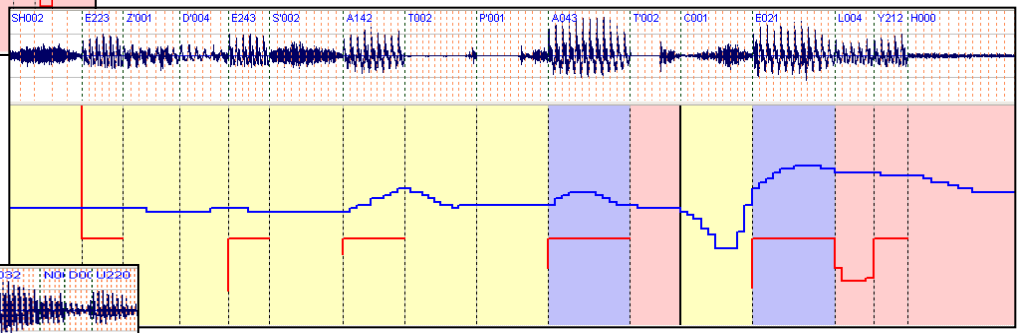
*‘The computer performance **65.5 Mbit/s** was achieved’*



1 C4

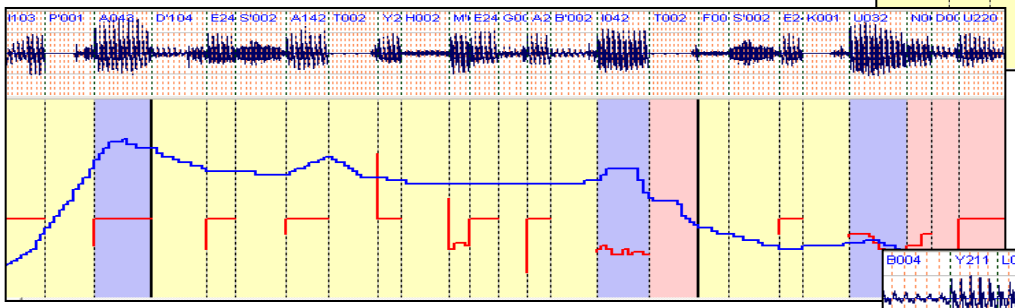
производи+тельность

компью+тера /



2 C1

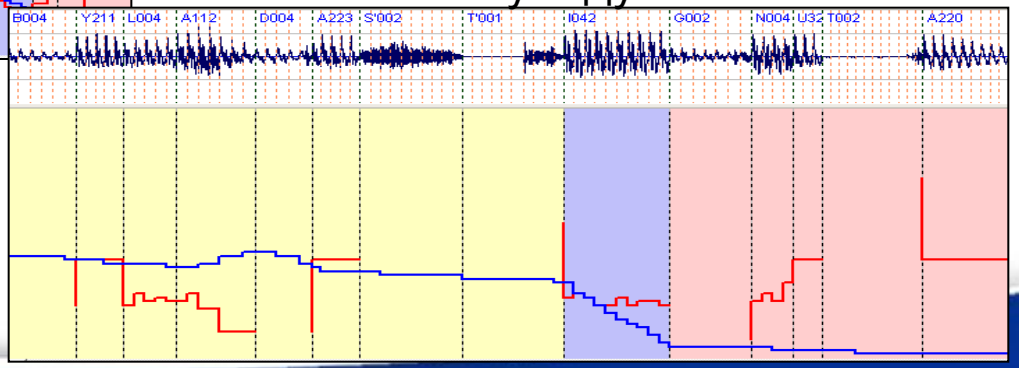
шестьдеся=т пя+ть/це+лых/



3 C3

и= пя+ть/деся=тых

мегаби+т/вЪсеку+нду/



4 P4

была= дости+гнута/

Conclusion

- Three sets of NooJ grammars for BE and RU have been obtained:

1. ***Identification*** *&* ***Classification*** of QE with MU according to **the SI**

2. ***Identification*** *&* ***Classification*** of QE with MU according to **word formation**

3. ***Generating*** QE with MU into **orthographical words**

- QE with MUs can get the correct intonation marking **only after they are properly generated**, i. e. expanded into orthographical words.

Future work includes:

- generating numeral quantifiers, expressed by **fractional** and **decimal** numbers

- **disambiguation**, e. g., in such cases when algorithms “confuse” some units (the same initial letter ***г*** for ***год*** **‘year’,** ***грам*** **‘gram’** ***на*** **‘hour’**

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

THANK YOU
FOR ATTENTION !

E-mail contacts:

Yury.Hetsevich@gmail.com
skelena777@gmail.com
lobanov@newman.bas-net.by