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

Resources for Identification of Cues with Author's Text Insertions in Belarusian and Russian Electronic Texts

Yury Hetsevich, Tatsiana Okrut, Boris Lobanov and Yauheniya Yakubovich

Abstract

The aim of this paper is to give a general outline of the ongoing work in the processing of cues with text insertions by the author. We describe the main stages of the characters' gender identification in Belarusian and Russian electronic texts. The identification is based on punctuation marks and the detection of gender indicators.

Introduction

A literary text may include not only the author's words but also the words of other characters. As such, creators of audiobooks often use different speakers or synthesised voices so that a text reflects more closely the unique speech characteristics of the characters. It stands to reason, then, that the manual marking of a text for reading by different speakers or speech synthesisers takes a lot of time.

Text-to-speech systems (TTS systems) may serve as an alternative way of creating audiobooks. In only a short amount of time, such systems are able to create an electronic audio file from an electronic text.

Nowadays some ideas for this kind of text analysis already exist. For example, the online system Text Analysis Demo[1] makes it possible to

[1] http://www.alchemyapi.com/api/demo.html

identify characters and their statements[2] in a text, which is an important factor in the identification of the speakers' gender in TTS systems. A group of European scientists has also developed algorithms for character identification and automatic determination of the roles with the help of NooJ syntactic grammars [3]. As for the Slavic languages, the work of Croatian scientists on direct speech identification should be noted [4], although they do not consider the problem of gender identification. Programs for the creation of audiobooks such as MP3book2005[3] and AUDIOBOOK[4] are also quite developed in this area. These have special inbuilt units for logical analysis of dialogues, which can provide markings for distinguishing between the words of the author and various characters in a dialogical text. In AUDIOBOOK, steps were taken to read dialogues in character, but the program does not cover all of the cases. It does not take into account cue structures with more than one insertion of the author's words. In addition, it is unable to identify the gender of a character for indicators such as 'verb + masculine noun' combination in the author's words:

– Трэба напісаць 'яць', – адказвае вучань.

(– We should write 'яць', – the pupil (he) answers.)

It should also be noted that AUDIOBOOK works better with Russian and English speech engines, and that the units for logical analysis of dialogues do not include work with any other languages but Russian.

Thus, our main goal is to develop algorithms for direct speech processing and to provide identification of the characters' gender using the insertions of the author's words in direct speech. The algorithms can subsequently be used in a TTS system.

Data collection and processing procedures for identifying the gender of a character

The authors have developed 3 types of syntactic grammars for direct speech processing – one for all direct speech identification and two for gender-dependent direct speech identification (masculine and feminine gender detection).

We have selected some texts in Belarusian and Russian to identify the specific features of dialogical text (the so-called training set). As NooJ processes a text at the paragraph level, for the development of algorithms

[2] http://www.alchemyapi.com/api/entity/quotations.html

[3] http://mp3book2005.ru/1. htm

[4] http://kom-pas.narod.ru/audiobook_net.htm

each text paragraph was separately analysed by experts in the following order (see examples in Table 1):

— marking the paragraphs with direct speech (figure 1 was used as a label)

— marking the cues of male and female characters

— marking the words of a character and the author's words (italics – the characters' words, regular font – the author's words)

Marks of direct speech	**Gender of speaker**	**A paragraph of the Belarusian training text**
1	F	- *Бацька вады,* - шэптам сказала Майка.
1	M	- *Бацька вод,* - паправіў Алесь. - *Вось так і Дняпро пачынаецца недзе.*
1	F	- *Жывая вада,* - сказала Яня.
0	-	І яна апусцілася на калені і зламала пальчыкамі крыштальную паверхню.
1	F	- *Піце. Будзеце жыць сто год...*

Table 1 – An excerpt from the Belarusian training corpora with manual marking for direct speech

On the basis of manual text analysis, the following syntactic structures were revealed in direct speech:

Direct speech apart from the author's text:
– C (! | !! | !!! | ? | ?! | … | .).
Direct speech followed by the author's text:
– C (, | ! | !! | !!! | ? | ?! | … | .) – A (… | .).
Direct speech with one or more insertions of the author's text:
– C (, | ! | !! | !!! | ? | ?! | … | .) – A (, | … | . | : | .) – C (, | ! | !! | !!! | ? | ?! | … | .) (– A (, | … | . | : | .) – C (, | ! | !! | !!! | ? | ?! | … | .)).

The structures contain the following annotations: C – the words of a character (speaker), A - the author's text, brackets (,) – the beginning and the end of a choice set with punctuation marks, | - symbol *or* (separation of punctuation marks in a choice set).

Grammar for direct speech identification

The data obtained was used to develop a NooJ syntactic grammar for the automated identification of all paragraphs containing direct speech (DS_All) (Figure 1). The grammar has the same view for both the Belarusian and Russian languages. Its main parts – graphs Speaker and Author – serve to identify the character's words and the author's words, respectively.

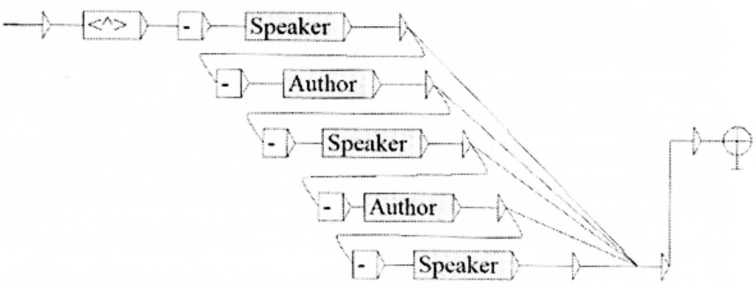
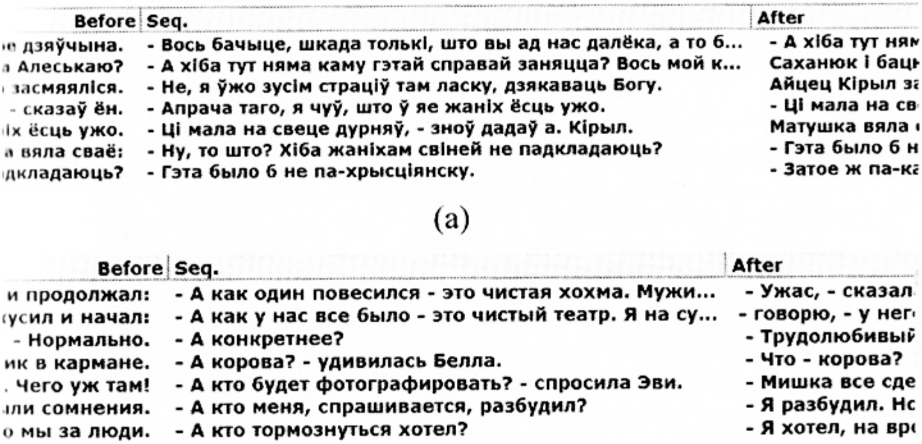


Figure 1 – A general view of the DS_All grammar (identical for Belarusian and Russian)

According to the grammar DS_All, any direct speech sentence starts with a dash followed by the characters' words; Speaker (numbers, different word forms and punctuation marks may be included), which in turn may end with a full stop, comma, exclamation mark, question mark, or a combination of these punctuation marks (combinations with quotation marks are also possible). If there is no author's text after the character's words, the grammar stops. This is how the structure of the first type is identified. If there is any combination of punctuation marks with a dash after the character's words, the grammar continues working and enables the graph Author (various word forms and such punctuation marks as a comma, a full stop or brackets may be included). This is how the structure of the second type is identified.

The grammar can identify from one to two text insertions by the author in direct speech through processing the algorithm in consecutive order: Speaker–Author–Speaker. This provides the identification of direct speech with the structure of the third type.

DS_All can be applied sequentially to any Belarusian or Russian electronic text through NooJ. The results of the application of the grammars are presented in the form of a concordance in Figure 2.



Before	Seq.	After
е дзяўчына.	- Вось бачыце, шкада толькі, што вы ад нас далёка, а то б...	- А хіба тут ням
а Алеськаю?	- А хіба тут няма каму гэтай справай заняцца? Вось мой к...	Саханюк і бац
засмяяліся.	- Не, я ўжо зусім страціў там ласку, дзякаваць Богу.	Айцец Кірыл з
- сказаў ён.	- Апрача таго, я чуў, што ў яе жаніх ёсць ужо.	- Ці мала на св
іх ёсць ужо.	- Ці мала на свеце дурняў, - зноў дадаў а. Кірыл.	Матушка вяла
а вяла сваё:	- Ну, то што? Хіба жаніхам свіней не падкладаюць?	- Гэта было б н
дкладаюць?	- Гэта было б не па-хрысціянску.	- Затое ж па-к

(a)

Before	Seq.	After
и продолжал:	- А как один повесился - это чистая хохма. Мужи...	- Ужас, - сказал
кусил и начал:	- А как у нас все было - это чистый театр. Я на су...	- говорю, - у нег
- Нормально.	- А конкретнее?	- Трудолюбивый
ик в кармане.	- А корова? - удивилась Белла.	- Что - корова?
. Чего уж там!	- А кто будет фотографировать? - спросила Эви.	- Мишка все сде
или сомнения.	- А кто меня, спрашивается, разбудил?	- Я разбудил. Н
о мы за люди.	- А кто тормознуться хотел?	- Я хотел, на вр

(b)

Figure 2 – The results of applying the DS_All grammar to Belarusian (a) and Russian (b) texts

Grammars for the characters' gender identification from the author's text

In the Belarusian and Russian languages, singular past tense verbs may have gender attributes, for example, *паправіў* 'he corrected' and *сказала* 'she said' for Belarusian, *высказался* 'he spoke' , *высказалась* 'she spoke' for Russian. As such verbs may often occur in the author's commentaries to direct speech, and some nouns have themselves gender attributes, they may serve as gender indicators and be considered suitable for the gender identification of the characters.

On the basis of the grammar DS_All, two separate grammars were developed – one for masculine gender identification (DS_M), and one for feminine gender identification (DS_F). For this purpose, we have modified the graph Author and added resources for gender identification (Fig. 3). In Figure 3, one can see the subgraph VERBSmasculine. It includes a list of masculine verbs, which were selected at the stage of manual marking of texts in the Belarusian and Russian languages. A similar list of verbs was created within the subgraph VERBSfeminine for feminine gender identification.

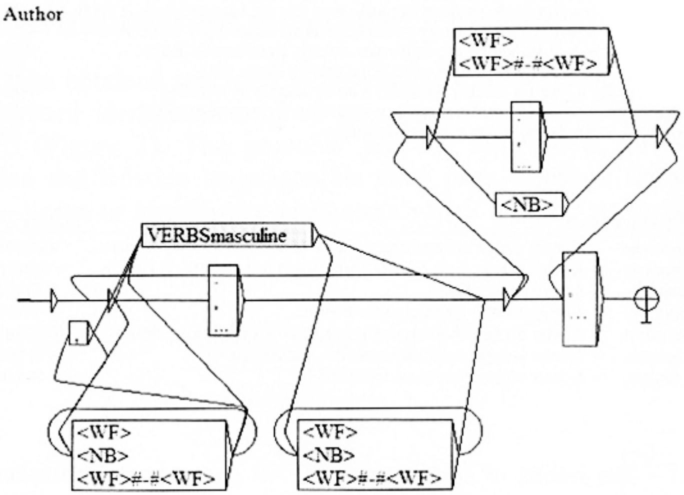
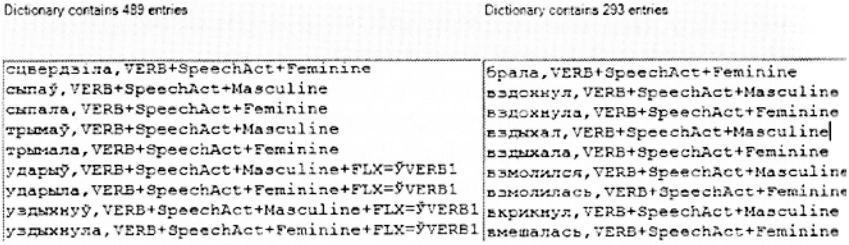


Figure 3 – The subgraph Author, DS_M

Subsequently, we created dictionaries for the Belarusian and Russian languages (Figure 4), where verbs were presented in pairs in the masculine and feminine forms.

It should be noted that in the Belarusian dictionary for verbs that have 'y' at the start of a word (like in *уздыхнула*), a special paradigm ЎVERB1 was added. It takes into account the transition of 'у' into 'ў' after vowels.



```
Dictionary contains 489 entries

сцвердзіла,VERB+SpeechAct+Feminine
сыпаў,VERB+SpeechAct+Masculine
сыпала,VERB+SpeechAct+Feminine
трымаў,VERB+SpeechAct+Masculine
трымала,VERB+SpeechAct+Feminine
ударыў,VERB+SpeechAct+Masculine+FLX=ЎVERB1
ударыла,VERB+SpeechAct+Feminine+FLX=ЎVERB1
уздыхнуў,VERB+SpeechAct+Masculine+FLX=ЎVERB1
уздыхнула,VERB+SpeechAct+Feminine+FLX=ЎVERB1
```

(a)

```
Dictionary contains 293 entries

брала,VERB+SpeechAct+Feminine
вздохнул,VERB+SpeechAct+Masculine
вздохнула,VERB+SpeechAct+Feminine
вздыхал,VERB+SpeechAct+Masculine
вздыхала,VERB+SpeechAct+Feminine
взмолился,VERB+SpeechAct+Masculine
взмолилась,VERB+SpeechAct+Feminine
вкрикнул,VERB+SpeechAct+Masculine
вмешалась,VERB+SpeechAct+Feminine
```

(b)

Figure 4 – Excerpts from the Belarusian (a) and Russian (b) dictionaries of verbs in masculine and feminine forms

Thus, instead of the subgraphs VERBSmasculine and VERBSfeminine, special tags (categories) were used: VERB, SpeechAct (a semantic mark, verbs as comments to direct speech) and Masculine/Feminine (Figure 5).

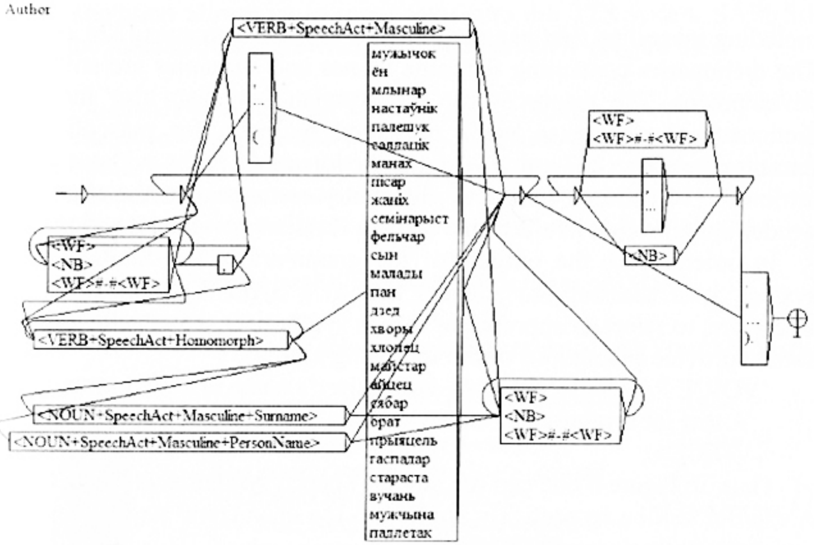


Figure 5 – The subgraph Author, DS_M for Belarusian

In the development process we also investigated the problem of grammatical homoforms. For instance, the verb form *кажа* 'say' (the present tense form of the Belarusian verb *казаць* 'to say') can refer both to the masculine and to the feminine genders, which makes it impossible to use such verbs separately in the identification of the characters' gender. To solve this problem, an additional 'verb-noun' combination of graphs was created, where the first graph with <VERB+SpeechAct+Masculine> annotation includes the grammatical homoforms of verbs which refer to the act of speaking, and the other graph represents a list of nouns having gender attributes (eg *дзед* 'grandfather') (the graphs are represented in Figure 5). Separate dictionaries were created for first names and surnames. These are linked to the grammars DS_M and DS_F through the following sequence of tags: <NOUN + SpeechAct + Masculine + PersonName>, <NOUN + SpeechAct + Feminine + PersonName>, <NOUN + SpeechAct + Masculine + Surname>, and <NOUN + SpeechAct + Feminine + Surname>.

In all, concerning dictionary resources for the Belarusian module, we have created a dictionary of past tense verbs presented by pairs for masculine and feminine genders (489 entries); a dictionary of verbs in the present and future tenses that serve as comments to direct speech but do not allow us to identify the character's gender (14 entries); dictionaries including masculine first names (676 entries) and surnames (324 entries). The dictionaries containing feminine names and surnames are still under development. The resources for the Russian language also include a dictionary of past tense verbs presented by pairs for masculine and feminine genders (297 entries) and a dictionary of verbs without gender attributes (14 entries). The other dictionary resources for the characters' gender identification in electronic texts in Russian are under development.

In order to use the outputs of the grammars in the SAPI 5.1 TTS system, it is necessary to adapt a text to a SAPI TTS XML format[5]. Therefore, to select an appropriate speech synthesiser, a syntactic grammar should provide annotations of the following kind:

<VOICE Required='name=[a synthesiser's name in a TTS system]'>
…A text for synthesis…
</VOICE>.

Thus, in Figure 6 one can see that the speech synthesisers BorisBel and AlesiaBel will be respectively applied to the characters' words (Speaker) and to the author's words (Author).

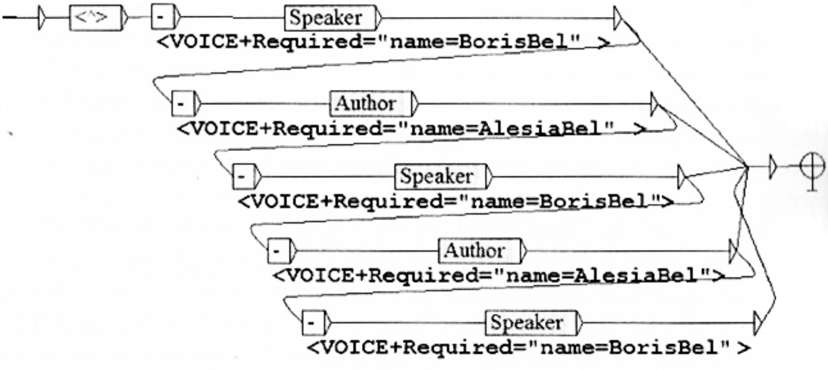
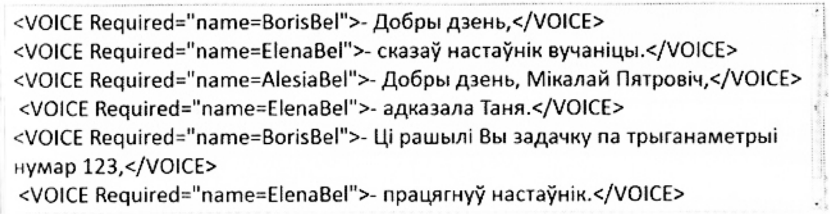

<^>
-
Speaker
<VOICE+Required="name=BorisBel" >
-
Author
<VOICE+Required="name=AlesiaBel" >
-
Speaker
<VOICE+Required="name=BorisBel">
-
Author
<VOICE+Required="name=AlesiaBel">
-
Speaker
<VOICE+Required="name=BorisBel" >

Figure 6 – The DS_All grammar following adaptation for SAPI 5.1

After being processed by the grammars DS_M and DS_F adapted for SAPI 5.1, the sentences in Table 1 will be annotated as in Figure 7. A female voice, AlesiaBel, is applied to the author's words, and voices ElenaBel and BorisBel are used for the female and male characters' words.

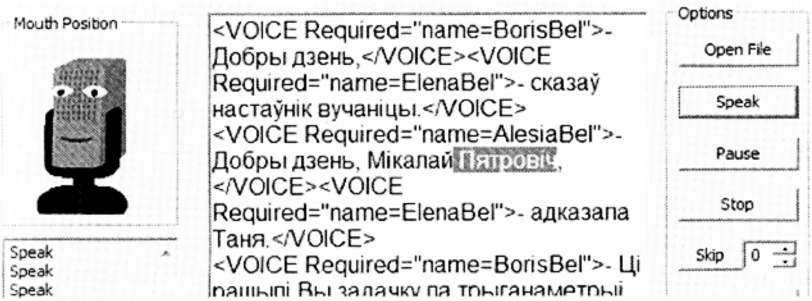
[5] XML TTS Tutorial (SAPI 5.3), http://msdn.microsoft.com/en-us/library/ms717077%28v= vs.85%29.aspx

Such annotation allows us to input texts into the TTS system SAPI 5.1, where the indicated voices switch over automatically (Figure 8).



```
<VOICE Required="name=BorisBel">- Добры дзень,</VOICE>
<VOICE Required="name=ElenaBel">- сказаў настаўнік вучаніцы.</VOICE>
<VOICE Required="name=AlesiaBel">- Добры дзень, Мікалай Пятровіч,</VOICE>
 <VOICE Required="name=ElenaBel">- адказала Таня.</VOICE>
<VOICE Required="name=BorisBel">- Ці рашылі Вы задачку па трыганаметрыі
нумар 123,</VOICE>
 <VOICE Required="name=ElenaBel">- працягнуў настаўнік.</VOICE>
```

Figure 7 – The sentences from Table 1 after being annotated with VoiceXML tags



```
<VOICE Required="name=BorisBel">-
Добры дзень,</VOICE><VOICE
Required="name=ElenaBel">- сказаў
настаўнік вучаніцы.</VOICE>
<VOICE Required="name=AlesiaBel">-
Добры дзень, Мікалай Пятровіч,
</VOICE><VOICE
Required="name=ElenaBel">- адказапа
Таня.</VOICE>
<VOICE Required="name=BorisBel">- Ці
рашылі Вы задачку па трыганаметрыі
```

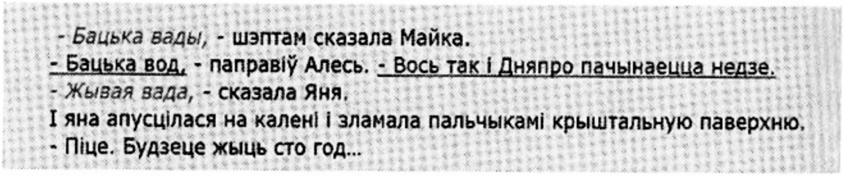
Figure 8 – Speech synthesis of the annotated sentences

Multi-coloured marking of a text

A further application for the above-mentioned annotation is the multi-coloured marking of a text for visual presentation of the author's words and of the female and male characters' words. Such marking may be used by an editor in order to quickly analyse direct speech in a text and to select an optimal number of speech synthesisers or speakers.

To provide multi-coloured marking, the authors have developed the VoiceXmlToColorReplacer software. The program processes VoiceXML-files and allows the conversion of VoiceXML-tags of speech synthesisers into HTML-tags with different styles of direct speech visual presentation.

After a text passes through the VoiceXmlToColorReplacer software, the characters' cues and the author's insertions are marked in different colours: namely, the author's words (AliesiaBel) are in black, the male characters' cues (BorisBel) – in blue, and the female characters' cues (ElenaBel) – in red (Figure 9).



- *Бацька вады,* - шэптам сказала Майка.
- Бацька вод, - паправіў Алесь. - Вось так і Дняпро пачынаецца недзе.
- *Жывая вада,* - сказала Яня.

I яна апусцілася на калені і зламала пальчыкамі крыштальную паверхню.

- Піце. Будзеце жыць сто год...

Figure 9 – An excerpt of a Belarusian text with multi-coloured marking of direct speech

Evaluation

Evaluation is conducted using precision, recall and F-measure. The results of the evaluation are given in Tables 2 and 3 for the Belarusian and Russian languages, respectively. In evaluation, we use the following categories: N – cues identified as appropriate by linguistic experts; M – cues correctly processed by the grammar; L – all cues found by the grammar.

The Belarusian test set includes 23,867 word forms; 955 paragraphs; 481 paragraphs with direct speech; 233 paragraphs with the author's text insertions, where 165 are cues of male characters and 68 are cues of female characters. The Russian test set contains 34,056 word forms; 2,669 paragraphs; 1,658 paragraphs with direct speech; 551 paragraphs with author's text insertions, where 456 are cues of male characters and 95 are cues of female characters.

Grammar Name	Precision (P)	Recall (R)	F-measure, %
	(M/L)	(M/N)	2*P*R*100 / (P+R)
DS_All	461/462 = 0,995	461/481 = 0,958	97,6
DS_M	143/145 = 0,986	143/165 = 0,866	92,2
DS_F	57/58 = 0,982	57/68 = 0,838	90,4

Table 2 – Evaluation of the Belarusian Syntactic Grammars' Performance for Direct Speech and Identification of Characters' Gender

Grammar Name	Precision (P)	Recall (R)	F-measure, %
DS_All	1628/1658 = 0,982	1628/1658 = 0,982	98,2
DS_M	339/339 = 1	339/456 = 0,743	85,3
DS_F	90/90 = 1	90/92 = 0,978	98,9

Table 3 – Evaluation of the Russian Syntactic Grammars' Performance for Direct Speech and Identification of Characters' Gender

Conclusion

In this paper we have presented our ongoing work on direct speech processing. We can conclude that rather good operating results have been obtained and that the algorithms developed have shown themselves suitable for use in combination with a TTS system. However, the resources developed still require some improvement: further work needs to be done on the extension of dictionary resources of verb-indicators identifying gender, the extension of the punctuation base (dash and quotation types etc), and the expansion of the test corpora.

Acknowledgements

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

References

Jurić, Tereza, Marija Stupar, Damir Boras. 2012. Direct Speech Recognition in Text. In *Selected Papers from the NooJ 2011 Intern. Conf. Automatic Processing of Various Levels of Linguistic Phenomena,* Vučković K., Bekavac B., Silberztein M. Eds. Cambridge Scholars Publishing, Newcastle, 2011, pp.122–127.

Lendvai, Piroska, Tamás Váradi, Sándor Darány, and Thierry Declerck. 2010. Assignment of character and action types in folk tales. In *Selected Papers from the NooJ 2010 Intern. Conf. Formalising Natural Languages with NooJ*, Gavriilidou Z., Chatzipapa E., Papadopoulou L., Silberzstein M. Eds. Democritus University of Thrace, Greece. pp.102–111.