КАРПОВСКИЕ НАУЧНЫЕ ЧТЕНИЯ

Выпуск 8

Часть I

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФИЛОЛОГИЧЕСКИЙ ФАКУЛЬТЕТ

КАРПОВСКИЕ НАУЧНЫЕ ЧТЕНИЯ

Сборник научных статей

Основан в 2007 году

Выпуск 8

В двух частях

Часть 1

Минск
«ИВЦ Минфина»
2014

УДК 80/81 (082)

В сборнике представлены материалы Восьмых Карповских научных чтений, посвященных памяти профессора В.А. Карпова — ученого, философа, поэта, оставившего заметный след в отечественной культуре.

Адресован филологам, философам, системологам, специалистам по прикладной и компьютерной лингвистике, а также студентам, магистрантам, аспирантам, интересующимся рассматриваемыми проблемами.

Рекомендовано
Ученым советом филологического факультета
Белорусского государственного университета
(протокол № 9 от 19 июня 2014 г.)

ISBN 978-985-7060-81-8 (ч. 1)
ISBN 978-985-7060-80-1

Секция 4: ПРИКЛАДНАЯ ЛИНГВИСТИКА В БЕЛАРУСИ: СОСТОЯНИЕ И ПЕРСПЕКТИВЫ РАЗВИТИЯ

Таким образом, морфологические признаки имени существительных китайского и русского языков сильно отличаются. В китайском языке не все существительные имеют категории рода, кроме собственных имен людей, но по их именам не определяется пол человека. Категория числа в русском языке проявляется в окончаниях слов. Изменение окончаний русских слов по падежам называется склонением. В китайском языке число существительного выражается суффиксом 们, а падежи отсутствуют, вместо их существуют строгие правила по порядку слов.

ЛИТЕРАТУРА

1. Ахманова, О.С. Словарь лингвистических терминов. – М.: Сов. энциклопедия, 1966.

2. Белошапкова, В.А. Современный русский язык: учеб. для филол. спец. ун-тов // В.А. Белошапкова, Е.А. Брызгунова, Е.А. Земская и др.; Под ред. В.А. Белошапковой [Электронный ресурс]. – Режим доступа: http://sci-book.com/yazyik-russkiy/sovremennyiy-russkiy-yazyik-ucheb-dlya-filol.html. – Дата доступа: 26.02.2014.

3. Корпус современного китайского языка [Электронный ресурс]. – Режим доступа: http://ccl.pku.edu.cn:8080/ccl_corpus/index.jsp?dir=xiandai – Дата доступа: 26.02.2014.

4. 刘守军 词类 (Категория слов) / 刘守军 [Электронный ресурс]. – Режим доступа: http://study.hhit.edu.cn/subject/CourseWare_Detail.aspx?TeachCourse WareID=1684. – Дата доступа: 04.03.2014.

5. 孙浩成 现代汉语语法大全 (Грамматика современного китайского языка) / 孙浩成 [Электронный ресурс]. – Режим доступа: http://wenku.baidu.com /view/0ea2c833-43323968011c9253.html. – Дата доступа: 04.03.2014.

6. 朱雪峰, 王惠 现代汉语量词与名词的子类划分 (Подкатегории счётных слов и существительных современного китайского языка) / 朱雪峰, 王惠 [Электронный ресурс]. – Режим доступа: http://www.tigernt.com/yyy23.php. – Дата доступа: 04.03.2014.

Ю.С.Гецэвіч, Т.І.Окрут, Я.А.Міхайлава (Мінск, НАН РБ, БДУ)

РАСПРАЦОЎКА ЛІНГВІСТЫЧНЫХ РЭСУРСАЎ ДЛЯ АЛГАРЫТМАЎ ІДЭНТЫФІКАЦЫІ РЭПЛІК ДЫЯЛОГАЎ У ЭЛЕКТРОННЫХ ТЭКСТАХ МАСТАЦКАЙ ТЭМАТЫКІ НА БЕЛАРУСКАЙ І РУСКАЙ МОВАХ

На сённяшні дзень стварэнне аўдыёкніг можа праходзіць двума шляхамі: праз запіс надыктоўкі тэксту чалавекам і праз аўтаматычнае стварэнне аўдыёфайлаў на аснове сінтэзу маўлення па тэксце. Абодва варыянты забяспечваюць у большасці сваёй аднагалосае агучванне, пры гэтым першы выпадак патрабуе шмат выдаткаў часу. Але, калі звярнуць увагу на звычайны мастацкі тэкст, то ён налічвае вялікую колькасць дыялогаў паміж рознымі персанажамі, у сувязі з чым з'яўляецца матываванае жаданне стваральніка аўдыёкніг выкарыстаць розных дыктараў ці сінтэзаваныя галасы, каб аўдыёкніга была больш набліжанай да адлюстравання ўнікальных характарыстык маўлення персанажаў.

У напрамку стварэння шматгалосага сінтэзу маўлення па тэксце (ССМТ) аўтары пачалі працу яшчэ летам 2013 года. У якасці асяроддзя

распрацоўкі алгарытмаў выкарыстоўвалася праграма NooJ [1]. NooJ — гэта міжнародная лінгвістычная праграма, якая дазваляе распрацоўваць сінтаксічныя і марфалагічныя граматыкі і тэставаць іх на вялікай колькасці тэкстаў. З іх дапамогай можна потым ствараць сінтаксічныя анатацыі і экспартаваць размечаны тэкст як файл XML для далейшай апрацоўкі. На дадзены момант здзяйсняецца аўтаматычнае шматгалосае агучванне дыялогаў з рэплікамі, дзе прысутнічаюць аўтарскія каментары, пры гэтым распрацаваныя алгарытмы былі адаптаваныя і для рускай мовы. Непасрэдна мае месца ідэнтыфікацыя роду персанажа па такіх паказчыках ва ўстаўках слоў аўтара, як дзеясловы мінулага часу адзіночнага ліку з прыкметамі роду (*сказаў*, *сказала*), уласныя імёны (*Алесь*, *Майка*), і назоўнікі, якія пазначаюць размоўцу (*бацька*, *дзяўчынка*).

Напачатку вялася ручная праца: быў сабраны корпус тэкстаў, у якім асобна былі пазначаныя ўсе рэплікі са ўстаўкамі слоў аўтара. Для іх адначасова пазначаўся род персанажа — мужчынскі, або жаночы, вызначаліся паказчыкі роду.

На гэтым этапе таксама былі выведзеныя ніжэй прыведзеныя структуры афармлення простай мовы дыялогаў з наступнымі абазначэннямі: П — словы персанажа; А — словы аўтара; дужкі (,) — пачатак і завяршэнне набору варыяцый знакаў прыпынку; | — сімвал або (раздзяленне знакаў прыпынку ў наборы варыяцый знакаў прыпынку).

1. Словы моўцы без слоў аўтара:

– П (! | !! | !!! | ? | ?! | … | .).

2. Словы моўцы са словамі аўтара ў канцы:

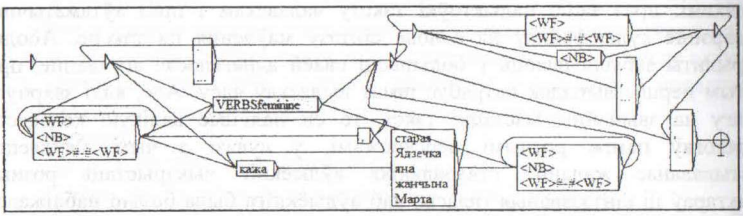
– П (, | ! | !! | !!! | ? | ?! | … | .) — А (… | .).

3. Словы моўцы з некалькімі аўтарскімі ўстаўкамі:

– П (, | ! | !! | !!! | ? | ?! | … | .) — А (, | … | . |: | .) — П (, | ! | !! | !!! | ? | ?! | … | .)

(– А (, | … | . |: | .) — П (, | ! | !! | !!! | ? | ?! | … | .)).

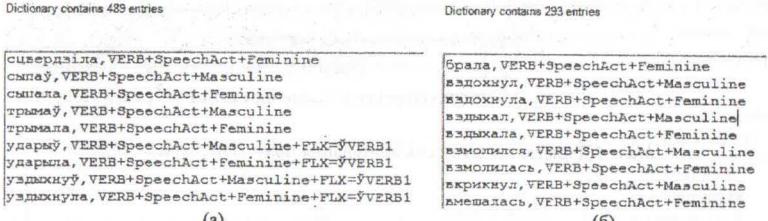
На аснове атрыманых дадзеных былі распрацаваныя алгарытмы DirectSpeechMasculine і DirectSpeechFeminine для пошуку рэплік са ўстаўкамі слоў аўтара, якія належаць персанажам мужчынскага ці жаночага роду. На малюнку 1 прадстаўлены падкампанент алгарытму ідэнтыфікацыі рэплік персанажаў жаночага роду.



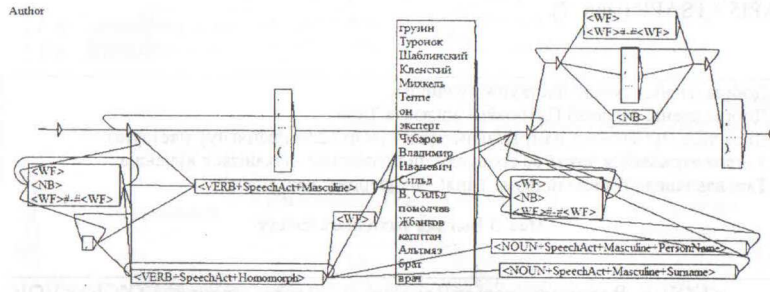
Мал. 1 Сінтаксічны падкампанент для вызначэння рэплік персанажаў жаночага роду

Ідэнтыфікацыя роду перасанажа ў дадзеных алгарытмах здзяйсняецца за кошт слоў-паказчыкаў роду, а непасрэдна праз падграфы VERBSmasculine і VERBSfeminine, якія ўключаюць спіс дзеясловаў мінулага часу адзіночнага ліку з наяўнымі атрыбутамі роду.

Па меры папаўнення спісу дзеясловаў-паказчыкаў роду былі створаныя асобныя лінгвістычныя рэсурсы ў выглядзе слоўнікаў для беларускай і рускай моў (мал. 2). У іх парамі прадстаўленыя дзеясловы мінулага часу ў формах для жаночага і мужчынскага роду. Такім чынам, замест падграфа VERBSmasculine цяпер прымяняюцца спецыяльныя тэгі (катэгорыі) SpeechAct (семантычная пазнака для дзеясловаў-каментароў простай мовы) і Masculine (мал. 3), адпаведна для падграфа VERBSfemenine — тэгі SpeechAct і Feminine. Акрамя таго, у алгарытме прадугледжана ідэнтыфікацыя роду персанажаў па спісу назоўнікаў, на аснове якіх на далейшым этапе таксама будуць створаныя слоўнікі.



Мал. 2. Слоўнікі дзеясловаў-паказчыкаў роду для беларускай (а) і рускай (б) моў

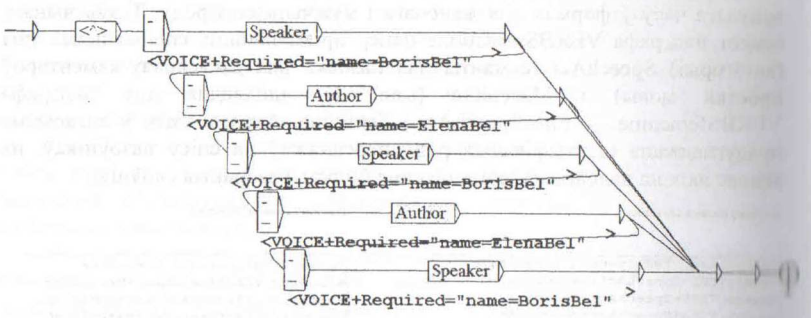


Мал. 3 Падграф Author у алгарытме для ідэнтыфікацыі рэплік персанажаў мужчынскага роду

Для таго каб працу гэтых граматык можна было выкарыстоўваць у ССМТ пад стандарт SAPI 5.1, то неабходна прывесці тэксты да выгляду SAPI TTS XML [2]:

<VOICE Required="name=[Назва голасу ў сістэме]">
…Тэкст для агучвання…
</VOICE>.

Для гэтага ў распрацаваныя алгарытмы былі дададзены маркеры абазначэння шляхоў, якія спрацавалі ў алгарытмах (Мал. 4). Маркеры настроены так, што непазначаны тэкст твору і словы аўтара чытаюцца голасам ElenaBel (Elena для рускай мовы), мужчынскія рэплікі голасам — BorisBel (Boris для рускай мовы), а жаночыя — AlesiaBel (Alesia для рускай мовы).



Мал. 4 Дададзеная функцыя анатавання праз VOICE-тэгі

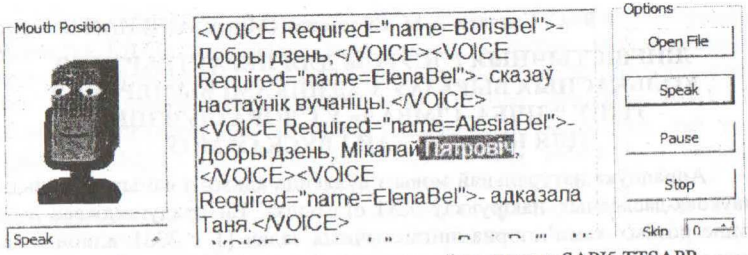
Так, напрыклад, тэкст на мал. 5 пасля апрацоўкі будзе выглядаць як на мал. 6. Пасля такой апрацоўкі тэкст можна выкарыстоўваць у праграме SAPI5 TTSAPP (мал. 7).

- Добры дзень,- сказаў настаўнік вучаніцы.
- Добры дзень, Мікалай Пятровіч,- адказала Таня.
- Ці рашылі Вы задачку па трыганаметрыі нумар 123,- працягнуў настаўнік.
- У мяне атрымаўся адказ 60 градусаў, ці правільна? - цікавілася вучаніца.
- Так, дакладна! Ты малайчына, Таня! - пацвердзіў настаўнік.

Мал. 5 Выгляд зыходнага тэксту

<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>

Мал. 6 Анатаваны праз VOICE-тэгі тэкст



Мал. 7 Аўтаматычнага пераключэнне галасоў у праграме SAPI5 TTSAPP

Па выніках тэсціроўкі створаных алгарытмаў на тэставым корпусе шматгалосае агучванне рэплік персанажаў са ўстаўкамі слоў аўтара ў электронных тэкстах здзяйсняецца з дакладнасцю больш за 75 %. Дэталёвая ацэнка распрацаваных алгарытмаў адлюстроўваецца ў табліцах 1 і 2, дзе N — вызначаныя экспертам правільныя рэплікі ў тэксце, M — правільна знойдзеныя алгарытмам рэплікі, L — усе знойдзеныя алгарытмам рэплікі.

Табліца 1

Ацэнка працы сінтаксічных алгарытмаў NooJ для беларускага корпусу тэкстаў

Назвы граматык	Дакладнасць (P)	Паўната (R)	Сярэдняя гарманічная велічыня (F1-measure), %
	(M/L)	(M/N)	2*P*R*100 / (P+R)
DirectSpeechMasculine	143/145 = 0,986	143/165 = 0,866	92,2
DirectSpeechFeminine	57/58 = 0,982	57/68 = 0,838	90,4

Табліца 2

Ацэнка працы сінтаксічных алгарытмаў NooJ для рускага корпусу тэкстаў

Назвы граматык	Дакладнасць (P)	Паўната (R)	Сярэдняя гарманічная велічыня (F1-measure), %
	(M/L)	(M/N)	2*P*R*100 / (P+R)
DirectSpeechMasculine	300/339 = 0,885	300/456 = 0,657	76,4
DirectSpeechFeminine	70/90 = 0,777	70/92 = 0,761	76,9

У далейшым плануецца павысіць дакладнасць працы распрацаваных алгарытмаў, а таксама распрацаваць кампанент ідэнтыфікацыі роду персанажаў па словах персанажаў.

ЛІТАРАТУРА

1. Лінгвістычны працэсар NooJ [Электронны рэсурс]. – 2002. – Рэжым доступу: http://www.nooj4nlp.net/pages/nooj.html. – Дата доступу: 01.07.2013.

2. XML TTS Tutorial (SAPI 5.3) // Microsoft Developer Network [Electronic resource]. – 2013. – Mode of access: http://msdn.microsoft.com/en-us/library/ms717077 %28v=vs.85%29.aspx. – Date of access: 29.07.2013.