АЎТАМАТЫЧНАЯ ДЭЛІМІТАЦЫЯ ТЭКСТАВАЙ ІНФАРМАЦЫІ ДЛЯ БЕЛАРУСКАЙ МОВЫ

AUTOMATIC DELIMITATION OF TEXT INFORMATION FOR THE BELARUSIAN LANGUAGE

Я.С. Зяноўка[1)], М.В. Супрунчук[2)], А.А. Бакуновіч[3)], М.А. Казлова[4)], Д.І. Латышэвіч[5)], Ю.С. Гецэвіч[6)]

Ya. Zianouka[1)], M. Suprunchuk[2)], A.A. Bakunovich[3)], M.A. Kazlova[4)], D. Latyshevich[5)], Yu. Hetsevich[6)]

[1)]АІПІ НАН Беларусі,
Мінск, Беларусь, *evgeniakacan@gmail.com*
[2)]АІПІ НАН Беларусі,
Мінск, Беларусь, *suprunchuk@gmail.com*
[3)]АІПІ НАН Беларусі,
Мінск, Беларусь, *bakunovich.andrei@gmail.com*
[4)]АІПІ НАН Беларусі,
Мінск, Беларусь, *margaryta.kazlova@gmail.com*
[5)]АІПІ НАН Беларусі,
Мінск, Беларусь *david.latyshevich@gmail.com*
[6)]АІПІ НАН Беларусі,
Мінск, Беларусь, *yuras.hetsevich@gmail.com*
[1)]UIIP of NASB,
Minsk, Belarus, *evgeniakacan@gmail.com*
[2)]UIIP of NASB,
Minsk, Belarus, *suprunchuk@gmail.com*
[3)]UIIP of NASB,
Minsk, Belarus, *bakunovich.andrei@gmail.com*
[4)]UIIP of NASB,
Minsk, Belarus, *margaryta.kazlova@gmail.com*
[5)]UIIP of NASB,
Minsk, Belarus, *david.latyshevich@gmail.com*
[6)]UIIP of NASB,
Minsk, Belarus, *yuras.hetsevich@gmail.com*

Дадзены артыкул прысвечаны аналізу аўтаматычнай сегментацыі тэксту на сінтагмы беларускай літаратурнай мовы. Апісаны распрацаваны прататып інтанацыйнага працэсара, асноўнай задачай якога з'яўляецца вызначэнне мінімальных сэнсавых адзінак у сказе на аснове знакаў прыпынку і злучнікаў.

Ключавыя словы: аўтаматычная дэлімітацыя; сінтагма; інтанацыйныя межы; інтанацыйны працэсар; сінтаксіс; пунктуацыя.

This article is devoted to the analysis of automatic textual segmentation into syntagmas of the Belarusian literary language. The developed prototype of an intonation processor is described, the major task of which is to determine the minimum semantic units in a sentence based on punctuation marks and conjunctions.

Keywords: automatic delimitation; syntagma; intonation boundaries; intonation processor; syntax; punctuation

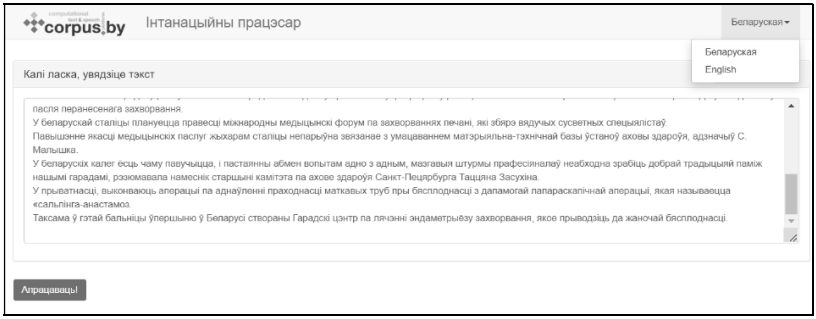
Уводзіны. У сучаснай навуцы пытанне аб зыходнай адзінцы будовы і ўспрымання маўлення не мае адназначнага рашэння менавіта таму, што разглядаецца на аснове розных падыходаў і прынцыпаў. Аднак вядома, што маўленне мае сінтагматычную прыроду і ўяўляе сабой ланцужок не асобных слоў, а адзіных лексічных комплексаў, што фарміруюць сінтагмы. Сінтагма – канкрэтная індывідуальна-аўтарская моўная структура, якая прадстаўляе інтанацыйнае, граматычнае і сэнсавае адзінства некалькіх слоў. Адэкватнае ўспрыманне маўлення забяспечваецца аўтарскай дэлімітацыяй, а правільнае разуменне тэксту дасягаецца дакладным аднаўленнем яго сінтагматычнай структуры і інтанацыі.

Як паказаў агляд наяўных у літаратуры даных, узаемадзеянне фактараў, якія ўплываюць на межы сінтагм, вывучана недастаткова [1]. Выяўлены хутчэй тэндэнцыі, і нават для заўважных канцэпцый няма дастаткова фармалізаванага апісання тэкставых варыяцый, характарыстыкі якіх маглі б служыць ключамі для аўтаматычнай расстаноўкі інтанацыйных межаў і паўз. Няма таксама дастатковых статыстычных даных, якія дазволілі б аддзяляць нарматыўнае (нейтральнае) прачытанне тэксту ад дапушчальнага ці памылковага. Асноўным доказам з'яўляецца сцвярджэнне, што кампаненты інтанацыі звязаны з тымі ці іншымі асаблівасцямі сінтаксічнай структуры сказа [2]. У той жа час узаемаадносіны інтанацыі і сінтаксісу нельга звесці да паралелізму сінтаксічных і інтанацыйных сродкаў. Інтанацыйных мадэляў, дынамічных і меладычных структур, якія маюць фаналагічнае значэнне, заўсёды менш, чым сінтаксічных мадэляў. Адны і тыя ж інтанацыйныя сродкі мовы выкарыстоўваюцца для выражэння розных сінтаксічных значэнняў у розных маўленчых сітуацыях.

З прычыны дамінуючай ролі сінтагмы агульным законам сэнсавага чляненне маўлення з'яўляецца асаблівая, уласцівая кожнай канкрэтнай мове, рытміка-меладычная будова сінтагмы і маркіраванне сінтагматычных межаў. Адным з шляхоў вырашэння пастаўленай праблемы з'яўляецца распрацоўка канкрэтных метадаў і алгарытмаў аналізу і апрацоўкі інтанацыйных асаблівасцей натуральнага маўлення, яго аўтаматычнага сінтагматычнага падзелу і рэалізацыі ўсіх інтанацыйных канструкцый зададзенай мовы, што прывядзе да аўтаматызаванага прайгравання адвольнага тэксту з манерай чытання чалавека, а не штучнай сістэмы [3]. Такім чынам, на падставе дадзенага падыходу супрацоўнікі лабараторыі распазнавання і сінтэзу маўлення АІПІ НАН Беларусі [4] распрацавалі праграмны прадукт *"Інтанацыйны*

працэсар"[5], які аналізуе сінтаксічную структуру тэкставай інфармацыі і вызначае інтанацыйныя межы сінтагм. Дадзены сэрвіс знаходзіцца ў адкрытым доступе на платформе для апрацоўкі тэкставай і гукавой інфармацыі для розных тэматычных даменаў Corpus.by [6].

Асноўная частка. Прынцып функцыянавання сістэмы прадстаўлены на малюнку 1, згодна з якім уваходны тэкст паслядоўна праходзіць лінгвістычную апрацоўку на аснове ўсіх знакаў прыпынку і злучнікаў, якія сустракаюцца ў тэксце. Сэрвіс просты ў выкарыстанні. Пасля кліка кнопкі *"Апрацаваць!"* уваходны тэкст разбіваецца на сінтагмы (малюнак 2) у выглядзе спіса слоў з пазнакай іх граматычнага класа і ўказаннем знакаў прыпынку. У хуткім часе плануецца цалкам дапрацаваць сэрвіс для вызначэння ўсіх тыпаў сінтагм, заснаваных на сінтактыка-семантычным падыходзе (вылучэнне фразеалагізмаў і сінтагм, аб'яднаных сэнсавымі адносінамі без абмежавання знакамі прыпынку).



Рыс. 1. Інтэрфейс прататыпа «Інтанацыйны працэсар»

Сэрвіс таксама прадстаўляе наступныя выніковыя даныя:

У акне *"Вынік тэкставай апрацоўкі"* пералічаны паслядоўны спіс слоў, у якім выводзяцца даныя аб часцінамоўнай прыналежнасці кожнай адзінкі, знойдзенай сістэмай у тэксце: для слова вызначаюцца часціны мовы і іх тэгі, якія з'яўляюцца скарочаным абазначэннем марфалагічных прыкмет слова (малюнак 2). Кожны тэг надаецца слову згодна з асобным слоўнікам, убудаваным у платформу Corpus.by, а менавіта:

– sbm1987 – «Слоўнік беларускай мовы. Арфаграфія. Арфаэпія. Акцэнтуацыя. Словазмяненне / пад рэд. М.В. Бірылы. – Мінск, 1987».

– noun2013 – назоўнікі, згодна з выданнем «Граматычны слоўнік назоўніка / навук. рэд. В.П. Русак. – Мінск : Бел. навука, 2013».

– adjective2013 – прыметнікі, згодна з выданнем «Граматычны слоўнік прыметніка, займенніка, лічэбніка, прыслоўя / навук. рэд. В.П. Русак. – Мінск : Бел. навука, 2013».

– numeral2013 – лічэбнікі, згодна з выданнем «Граматычны слоўнік прыметніка, займенніка, лічэбніка, прыслоўя / навук. рэд. В.П. Русак. – Мінск : Бел. навука, 2013».

– pronoun2013 – займеннікі, згодна з выданнем «Граматычны слоўнік прыметніка, займенніка, лічэбніка, прыслоўя / навук. рэд. В.П. Русак. – Мінск : Бел. навука, 2013».

– verb2013 – дзеясловы, згодна з выданнем «Граматычны слоўнік дзеяслова / навук. рэд. В.П. Русак. – Мінск : Бел. навука, 2013».

– adverb2013 – прыслоўі, згодна з выданнем «Граматычны слоўнік прыметніка, займенніка, лічэбніка, прыслоўя / навук. рэд. В.П. Русак. – Мінск : Бел. навука, 2013».

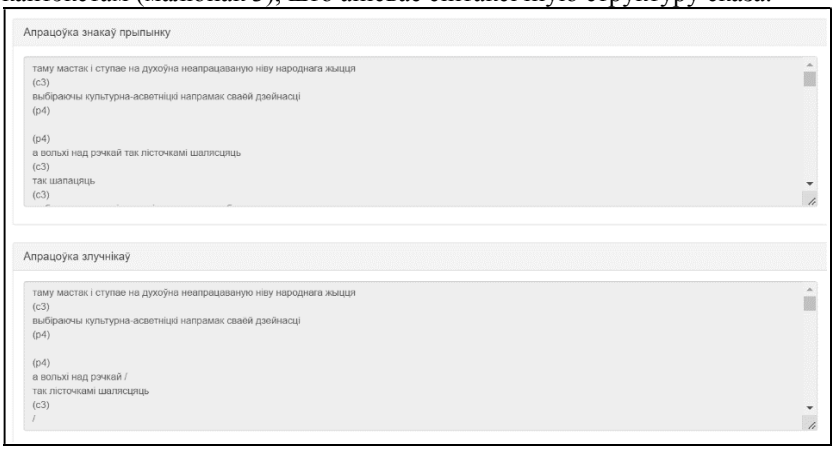
– zalizniak – «Грамматический словарь русского языка: Словоизменение / А.А. Зализняк. – Москва : Русский язык, 1980. – 880 с.».

Рыс. 2. Вынікі тэкставай апрацоўкі «Інтанацыйным працэсарам»

Таксама праграма выдае дадатковыя даныя лінгвістычнай апрацоўкі. Так, напрыклад, карыстальнік можа азнаёміцца з ўсімі невядомымі словамі ў акне *"Невядомыя токены"* і са спісам амографаў у акне *"Амографы"* (малюнак 2). У акне *"Апрацоўка знакаў прыпынку"* адлюстраваны спіс пунктуацыйных сінтагм з указаннем камп'ютарнай інтанацыйнай мадэлі (у дужках). Паколькі пунктуацыйныя сінтагмы могуць быць рознай даўжыні, пад пунктуацыйнай фразай маецца на ўвазе любая паслядоўнасць словаформаў або фанетычных слоў, раздзеленых знакамі прыпынку. Пры абазначэнні тыпа інтанацыі сінтагмы

ўлічваюцца не толькі знакі прыпынку, але і найбліжэйшы кантэкст у тэксце.

Акно *“Апрацоўка злучнікаў”* дэманструе разбіццё сказаў на лексічныя сінтагмы згодна са злучнікам як фармальным паказчыкам інтанацыйных межаў. Выкарыстоўваючы прыведзеныя даныя, карыстальнік можа азнаёміцца са спісам знакаў прыпынку, злучнікамі і іх фармальным маркерам, якія сустрэліся ў тэксце, з левым і правым кантэкстам (малюнак 3), што апісвае сінтаксічную структуру сказа.



Рыс. 3. Прамежкавыя даныя апрацоўкі знакаў прыпынку і злучнікаў ва ўваходным тэксце

Заключэнне. Такім чынам, першапачатковым этапам сінтагматычнай дэлімітацыі і інтанацыйнай апрацоўкі ўваходнага тэксту з’яўляецца вызначэнне межаў сінтагм, што магчыма праз выкарыстанне сэрвіса *“Інтанацыйны працэсар”*. Практыка стварэння аўтаматычных сістэм сінтэзу і распазнавання маўлення паказвае, што найважнейшымі ключамі для вызначэння інтанацыйных межаў пры агучванні тэксту з'яўляюцца сінтактыка-семантычныя адносіны ўнутры сказа. Аднак для аўтаматычнай дэлімітацыі тэксту неабходны аўтаматычны сінтаксічны аналіз, які на бягучы момант цалкам не рэалізаваны. У той жа час распрацоўка канкрэтных прыкладанняў для розных моў сведчыць аб тым, што для расстаноўкі інтанацыйных межаў аднаўленне поўнай сінтаксічнай структуры сказа неабавязковае. Прызнана, што арганізацыя прасадычных складнікаў прасцейшая, чым звязаныя з імі сінтаксічныя залежнасці. Стварэнне прататыпнай сістэмы “Інтанацыйны працэсар” для падзелу тэкставай інфармацыі на мінімальныя сэнсавыя адрэзкі даказвае дадзены факт. Сэрвіс ажыццяўляе пошук сінтагм праз знакі

прыпынку і злучнікі. На наступным этапе адбудзецца аналіз алгарытмаў на аснове тэсціравання існуючых лінгваакустычных рэсурсаў, якія пакрываюць усю разнастайнасць сінтаксічных канструкцый беларускай літаратурнай мовы. Гэта дазволіць выявіць недакладнасці ў прадстаўленым прататыпе, якія будуць выпраўлены ў найбліжэйшы час.

БІБЛІЯГРАФІЧНЫЯ СПАСЫЛКІ

1. Зяноўка, Я.С. Перспектывы развіцця аўтаматычнай апрацоўкі вуснага маўлення / Я.С. Зяноўка, М.В. Супрунчук, Ю.С. Гецэвіч // Лингвистика, лингводидактика, лингвокультурология: актуальные вопросы и перспективы развития / Белорусский гос. ун-т ; редкол.: Н. А. Куркович (гл. ред.) и др.. – Минск : БГУ, 2022. – С. 252-257.
2. Зяноўка, Я.С. Прынцыпы аўтаматычнага вызначэння інтанацыйных партрэтаў беларускага маўлення / Я. С. Зяноўка // Лингвистика, лингводидактика, лингвокультурология: актуальные вопросы и перспективы развития : материалы IV Междунар. науч.-практ. конф., Минск, 19–20 марта 2020 г. / редкол.: О.Г. Прохоренко (отв. ред.) [и др.]. – Минск : БГУ, 2020. – С. 472-477.
3. Zianouka, Y. Automatic Generation of Intonation Marks and Prosodic Segmentation for Belarusian NooJ Module / Y. Zianouka, Y. Hetsevich, D. Latyshevich, Z. Dzenisiuk // 15th International Conference, NooJ 2021 / France ; ed. Magali Bigey, Annabel Richeton, Max Silberztein, Izabella Thomas. – Besançon : Springer, Cham, 2022. – P. 231-242.
4. Лабараторыя распазнавання і сінтэзу маўлення [Электронны рэсурс]. – 2022. Рэжым доступу: http://ssrlab.by/. – Дата доступу: 11.07.2017.
5. Інтанацыйны працэсар [Электронны рэсурс]. – 2023. Рэжым доступу : https://corpus.by/IntonationalProcessor/?lang=be. – Дата доступу : 29.01.2023.
6. Платформа для апрацоўкі тэкставай і гукавой інфармацыі для розных тэматычных даменаў corpus.by // [Электронны рэсурс]. – 2022. Рэжым доступу : http://corpus.by/ . – Дата доступу : 12.01.2023.

АВТОМАТИЧЕСКОЕ ПОСТРОЕНИЕ ЧАСТОТНОГО СЛОВАРЯ

AUTOMATIC CONSTRUCTION OF FREQUENCY DICTIONARY

Ю.Ю. Красовская

J.J. Krasowskaja

Белорусский государственный университет,
Минск, Беларусь, *hisradaar@gmail.com*
Belarusian State University,
Minsk, Belarus, *hisradaar@gmail.com*

Частотный словарь включает в себя слова или словоформы какого-либо текста или корпуса текстов и частоту их встречаемости в исследуемом корпусе текстов. Составление частотного словаря посредством программного обеспечения позволяет автоматизировать процесс. Автоматизация с применением лемматизации позволит упростить долгий про-