УДК 81’33

А. А. Барковіч
(Мінскі дзяржаўны лінгвістычны ўніверсітэт, Мінск)

ІНДЭКС ДЭРЫВАЦЫЙНАСЦІ: АСАБЛІВАСЦІ СЛОВАЎТВАРАЛЬНАЙ АКТЫЎНАСЦІ ЎСХОДНЕСЛАВЯНСКІХ МОЎ

Артыкул прысвечаны актуалізацыі індэкса дэрывацыйнасці ў кантэксце разгляду словаўтваральнай актыўнасці ўсходнеславянскіх моў. Выкарыстанне корпусных тэхналогій, як паказана ў артыкуле,

дазваляе рабіць прадуктыўныя абагульненні адносна сутнасці і спецыфікі дэрывацыйнага развіцця мовы, у тым ліку ў супастаўляльным аспекце. У дадзеным даследаванні, у прыватнасці, ахарактарызавана дэрывацыйная актыўнасць ўсходнеславянскіх моў з выкарыстаннем металінгвістычнага інструментарыю – індэкса дэрывацыйнасці.

Пры спробах стварыць праграмныя сродкі апрацоўкі тэкстаў становіцца асабліва відавочнай выразная неадпаведнасць класічных граматычных канонаў практыцы словаўжывання ў працэсе камунікацыі [1, с. 160]. Так адбываецца не толькі у кантэксце камп'ютарна-апасродкаванай камунікацыі, але і пры суправаджэнні традыцыйнай камунікацыі, зарыентаванай у спрыяльных сацыяльных ці прагматычных абставінах на высокі ўзровень адпаведнасці норме. Характэрнымі сведчаннямі недастатковай аб'ектыўнасці і, калі называць рэчы сваімі імёнамі, штучнасці граматык з'яўляюцца пошукі актуальнага фармату *машыннага перакладу, заснаванага на правілах*, нестандартная для тэорыі мовы прэзентацыя граматычных мадэлей у *канкардансерах* і г. д.

У мовах пераважна сінтэтычнага складу (напрыклад, усходнеславянскіх рускай, украінскай і беларускай мовах) намінацыйныя асаблівасці маўлення непарыўна звязаны з дэрывацыйнымі. Менавіта тандэм намінацыйна-дэрывацыйных тэндэнцый вызначае напрамак мадыфікацыйнага развіцця мовы. Адметныя рысы стану моўнай сістэмы і фактары, якія фарміруюць паказчыкі яе напаўнення, у многім вызначаюць логіку распрацоўкі катэгорый. Камп'ютарна-апасродкаваны сегмент метамоўнай прасторы характарызуецца функцыянальнасцю надзвычайных статыстычна і камп'ютарна сумяшчальных магчымасцей дыскурсу.

Лічбавы характар даных апрацоўкі тэкстаў, якія складаюць аб'ектыўны бок камп'ютарна-апасродкаванага дыскурсу, дазваляе рабіць цэлы шэраг высноў высокай дакладнасці адносна, у першую чаргу, фармальна выражаных рыс тэкстаў. Прыкладам вырашэння прыватнай лінгвістычнай задачы ў такім ключы можа служыць кваліфікацыя зместу тэкстаў камп'ютарна-апасродкаванага дыскурсу ў частцы ўстанаўлення ступені лінгвістычнай, у тым ліку граматычнай, разнастайнасці фрагмента маўлення. Алгарытм такога даследавання патрабуе распрацаванасці граматычнага «кодэкса» мовы для адпаведнага праграмнага забеспячэння. У дадзеным кантэксце тыповай навуковай задачай будзе, напрыклад, даследаванне такога спецыфічнага аб'екта моўнай рэальнасці, як *ідыялект*.

Даследаванні прыкладнога характару мэтазгодна праводзіць на базе дастаткова прадстаўнічых масіваў тэкстаў. Сведчанні распрацоўкі адпаведных металінгвістычных інструментаў прысутнічаюць у працах аўтарытэтных замежных спецыялістаў. Так, разнастайнасць задзейнічанага ў корпусных рэсурсах слоўнікавага складу ўжо паспяхова вызначалася з дапамогай «суадносін “словаформа-токен”» (англ. – *Type-Token Ratio*, або *TTR*): «Мера слоўнікавай разнастайнасці ў корпусе ў лічбавым выражэнні роўная суме словаформ, падзеленай на суму токенаў. Чым бліжэй суадносіны да 1 (або 100 %), тым шырэй варыятыўнасць слоўніка. Гэта статыстыка не дазваляе непасрэдна параўноўваць паміж сабой корпусы розных памераў» [2]. Адпаведны эксперымент праводзіўся на матэрыяле англійскай мовы, але матэрыял іншамоўных тэкставых масіваў таксама дасягальны для падобнай апрацоўкі.

Пры гэтым для больш сінтэтычных, напрыклад, рускай ці беларускай моў даследаванні, разгорнутыя па методыцы вылічэння *індэкса*, а не каэфіцыента, характарызуюцца больш высокай рэлевантнасцю: аб’ектыўным паказчыкам для разгалінаванай сістэмы словазмянення славянскай мовы з’яўляецца арыентацыя менавіта на ўзровень *лемы*. Разам з тым, вызначэнне суадносных – без жорсткай абумоўленасці аб’ёмам тэксту ці гіпертэксту – характарыстык актуальна для даследаванняў абсалютнай большасці, напрыклад індаеўрапейскіх моў. Тэрмін *лема* часта выкарыстоўваецца паралельна з іншым тэрмінам – *лексема*. Іх карэктнае ўжыванне лёгка дыферэнцыраваць, калі разглядаць паняцці ў аспекце дыхатаміі «мова / маўленне». *Лема* як атрыбут маўленчай рэалізацыі і ‘сукупнасць форм слова’ відавочна істотна адрозніваецца ад *лексемы* – абстрактнай адзінкі моўнай сістэмы і ‘паняцця’.

Асобна трэба сказаць аб карэктным выкарыстанні тэрміна *словаўжыванне*, які набыў шматзначнасць і разнапланавасць. Традыцыйнае разуменне гэтага тэрміна як ‘ужыванне лексемы ў розных значэннях’ у прыкладной лінгвістыцы валодае больш шырокім аб’ёмам семантыкі – ‘любыя выпадкі ўжывання формы слова ў тэксце’. Тэрмін *словаўжыванне* менавіта ў прыкладным разуменні актыўна выкарыстоўваецца ў корпусных даследаваннях. Так, у праекце *InterCorp: projekt paralelních korpusů Filozofické fakulty Univerzity Karlovy v Praze* заяўлена наяўнасць каля *7 588 000* рускіх, *1 493 000* украінскіх і *1 308 000* беларускіх *словаўжыванняў*, або токенаў (англ. *token* – ‘словаўжыванне’). Зрэшты, у дадзеным праекце і іншыя славянскія мовы прадстаўлены дастаткова шырока: *26 879 000* балгарскіх, *12 625 000* харвацкіх, *2 664 000* македонскіх, *47 640 000*

польскіх, *40 108 000* славацкіх, *33 741 000* славенскіх, *6 972* сербскіх і, натуральна, *99 547 000* чэшскіх адзінак з агульнай колькасці ў *867 287 000* словаўжыванняў [3]. Інфармацыя вельмі красамоўная: гэта свайго роду рэйтынг, які ўскосна адлюстроўвае ўзровень развіцця прыкладной лінгвістыкі ў кожнай асобна ўзятай славянскай краіне.

Іншы шматмоўны корпусны праект *ParaSol: A Parallel Corpus of Slavic and other languages* быў створаны Рупрэхтам фон Вальдэнфельсам і Роландам Маерам у Інстытуце славянскіх моў і літаратур Бернскага ўніверсітэта і Інстытуце славістыкі Рэгенсбургскага ўніверсітэта ў Швейцарыі [4]. На фоне *18 085 532* словаўжыванняў у рубрыцы «Іншыя мовы» (англ. – *other languages*) паказчык у *7 646 832* токены для славянскіх моў дастаткова высокі. У гэтым рэсурсе прадстаўлены тэксты на рускай мове ў выглядзе *3 637 357* токенаў і *78 997* лем; на ўкраінскай – *1 017 057* токенаў і *33 563* лемы; на беларускай – *482 467* токенаў і *24 131* лема [4].

Для ацэнкі *парадыгматычнай* разнастайнасці лексічнага складу корпуса мэтазгодна параўнанне колькасці *словаўжыванняў* корпуса і *лем*, да якіх яны належаць. Такія суадносіны ў лічбавым выражэнні вызначаюцца як **індэкс дэрывацыйнасці** (англ. – *Derivability Index*, ці $\boldsymbol{I_D}$**)** – паказчык, які адлюстроўвае суадносіны колькасці словаўжыванняў тэксту (сукупнасці тэкстаў) да колькасці словаўжыванняў лемы [5]. У корпусе *ParaSol* для рускай мовы індэкс $I_D \approx 46{,}04$ (*3 637 357* токенаў / *78 997* лем), украінскай ≈ *30,3*, беларускай ≈ *19,99*. Несумненна, што дэрывацыйная дынаміка дыскурсіўнай практыкі будзе залежаць ад тэкставых асаблівасцей, але рэпрэзентатыўны корпус шэраг дазваляе дастаткова аб'ектыўна вызначыць сістэмныя ў агульнамоўным кантэксце характарыстыкі [6]. У нашым прыкладзе тэксты корпуса *ParaSol* у адпаведнасці з патрабаваннем рэпрэзентатыўнасці корпуснага дыскурсу валодаюць падобнай і параўнальнай метатэкставай якасцю.

Вышэйадзначанае дазваляе зрабіць высновы аб тым, што дэрывацыйная актыўнасць рускага маўлення вышэйшая за аналагічныя паказчыкі ўкраінскага і беларускага маўлення – пры ўсім падабенстве ўсходнеславянскіх моў і граматык. Другі вынік у такім параўнанні – у ўкраінскай мовы [7]. Беларуская мова вызначаецца невысокімі паказчыкамі дэрывацыйнай актыўнасці, саступаючы ў дадзеным аспекце іншым усходнеславянскім мовам – рускай і ўкраінскай. Зразумела, дэрывацыйная характарыстыка той ці іншай мовы не з'яўляецца абсалютным паказчыкам і павінна разглядацца ў кантэксце шырокага лінгвістычнага аналіза. Есць усе падставы меркаваць, што параўнальна невысокія дэрывацыйныя паказчыкі на практыцы

кампенсуюцца параўнальна большай прадуктыўнасцю моўных сродкаў іншага кшталту – і наадварот.

Спіс выкарыстанай літаратуры

1. Barkovich, A. Informational Linguistics: The new communicational reality / A. Barkovich. – Newcastle upon Tyne : Cambridge Scholars Publishing, 2020. – 271 p.
2. McEnery, T. Corpus Linguistics: Method, theory and practice / T. McEnery, A. Hardie. – Cambridge : Cambridge Univ. Press, 2012. – 294 p.
3. InterCorp: projekt paralelních korpusů Filozofické fakulty Univerzity Karlovy v Praze [Elektronický zdroj]. – Datum odběru : http://www.korpus.cz/intercorp/?req= page:info. – Datum ošetření : 30.09.2020.
4. ParaSol: A Parallel Corpus of Slavic and other languages [Electronic resourse]. – Mode of access : http://parasol.unibe.ch. – Date of access : 30.09.2020.
5. Барковіч, А. А. Металінгвістычная індэксацыя ў камп’ютарна-апасродкаваным дыскурсе / А. А. Барковіч // Беларуская лінгвістыка. – Мн. : Беларуская навука, 2015. – Вып. 74. – С. 79-87.
6. Barkovich, A. Meta-Description of Derivational Relations: Specifics of System Representation / A. Barkovich // Mundo Eslavo. – Universidad de Granada, 2018. – № 17. – P. 7-25.
7. Барковіч, А. А. Метамоўная характарыстыка камп’ютарна-апасродкаванага дыскурсу : дыс. … д-ра філал. навук : 10.02.19 / А. А. Барковіч. – Мінск, 2016.