DOI https://doi.org/10.36059/978-966-397-464-4-27

Дрига І. М.
кандидат філологічних наук, старший науковий співробітник,
доцент кафедра тюркології
Київського національного університету імені Тараса Шевченка
м. Київ, Україна

СУЧАСНИЙ СТАН КОРПУСНОЇ ЛІНГВІСТИКИ В ТУРЕЧЧИНІ

Продемонстрований насьогодні значний поступ корпусної лінгвістики у світі не оминув і Туреччину, де так само протягом останніх 20 років було створено та запропоновано до послуг мовознавців щонайменше три турецькомовних корпуси різного призначення. Метою даного дослідження є апробація корпусів турецької мови у порівнянні з традиційно вживаними (https://app.sketchengine.eu, Brown Corpus, BNC тощо), розуміння міри прогресування корпусної лінгвістики в Туреччині від анкарського періоду Сергєя Ніренбурга[1] та доробку Кемаля Офлязера[2] [Atalay, Oflazer…], а також порівняння операбельності новостворюваних корпусів тюркських мов, в тому числі тих, що їх розробляють в Україні.

Національний корпус турецької мови (TUD, Türkçe Ulusal Derlem, https://v3.tnc.org.tr) – надійне джерело для цитувань реальних вживань сучасної турецької мови, був розроблений за підтримки проєкту 108K242 Ради Туреччини з науково-технічних досліджень (TÜBİTAK) [Aksan… 2012; Aksan, Özel, Yılmazer, Demirhan 2016]. TUD – збалансований та репрезентативний масив текстів, що складається з 50 мільйонів слововживань сучасної турецької мови, створений по шаблону British National Corpus і включає тексти якомога більшої кількості різних жанрів та галузей приблизно 24-25-річного періоду з 1990 по 2014 рр. 98% наповнення корпусу становлять писемні тексти,

[1] Професор Політехнічного інституту Ренсселіра, Трой, Нью-Йорк. У 2000-х рр. Ніренбург виконував обов'язки директора Інституту передових досліджень мовної інженерії для маловивчених (природніх) мов під егідою НАТО (Анкара, Туреччина, пізніше – Батумі, Грузія, 2007).

[2] Кемаль Офлязер – колишній професор комп'ютерної лінгвістики університетів Сабанджи (Стамбул) та Бількент (Анкара), нині – Carnegie Mellon University в Катарі, займається мультидіалектним паралельним корпусом для арабської мови. Сфера інтересів – розробка та навчання ШІ для турецької мови, статистичний машинний переклад, використання ШІ для навчання мови.

2% – приклади, зібрані з розмовної турецької мови. Для користувачів, які хочуть отримати доступ до даних корпусу та робити запити, існує процедура авторизації та отримання доступу, після чого їм надається можливість безкоштовного онлайн – доступу через доволі зручний інтерфейс, який, утім, відрізняється від визнаних в ЄС та вимагає ознайомлення з інструкцією та взірцями запитів для тюркських мов. TUD наразі є найбільшим в Туреччині мовним корпусом та дає найбільші можливості пошуку, поетапно представлені кінцевому споживачеві, атож був обраний Управлінням підтримки дослідних програм TÜBİTAK (ARDEB) гідним увійти до каталогу кращих підтриманих проектів з соціально-гуманітарних наук «Історії успіху», який було видано до 50-річчя Ради. Таким чином, TUD використовується як розлоге реферативне джерело не лише колами лінгвістів, але й є довідковою базою для посадових осіб та урядовців, науково-освітніх та державних установ, органів преси та авторів публікацій з інформатики, методології освіти, науки і техніки; для промоції та популяризації використання турецької мови у світі, викладання турецької мови як іноземної та проведення наукових досліджень і зокрема написання курсових, дипломних та дисертаційних робіт в галузі тюркології.

Турецький текстовий корпус Середньосхідного ун-ту ODTÜ (http://medid.ii.metu.edu.tr) створений з метою виявлення слововжитку писемних текстів і так само фінансований проектом 107E156 TÜBİTAK у 2007-2011 рр. Після проходження реєстрації та надсилання письмової згоди використовувати корпус виключно в академічних цілях дослідники отримують заархівований файл програми METU Workbench, яку можна запустити за умови наявності на комп'ютері відповідної версії Java Runtime Environment (JRE). В корпусі з виконаною текстовою розміткою заявлено приблизно 2 млн. токенів зі збереженням таких тегів як тип та авторство, місце та рік видання тексту, і без збереження розбивки на параграфи тощо. Підкорпус у приблизно 450-500 тис слововживань має семантичну розмітку та становить собою дискурсний банк – браузер, де сполучники були анотовані разом з пов'язаними кластерами тексту, таким чином розширюючи операбельність даними з рівня речення до рівня дискурсу. Це включає анотації до приблизно 8400 дискурс – зв'язків, вибудованих на 500-тисячному субкорпусі з розподілом по жанрах двохмільйонного корпусу METU, що дає змогу досліджувати структурні та семантичні аспекти зв'язків дискурсу. Шар анотаційних даних генерується у форматі XML за допомогою символьних індексів. Інструмент анотації має користувацький інтерфейс, завдяки якому сполучники та аргументи підсвічуються різними кольорами.

Нарешті, 3-й корпус Мерсінського ун-ту (http://turkcederlem.mersin.edu.tr), здебільшого опрацьовуваний в проектах Бюлента Озкяна, відрізняється тим, що надає і лексичну, і морфологічну розмітку текстам, відповідно зробивши можливим складання словників колокацій, відслідковування активізації та зникнення варіантів абсолютних синонімів за періодами, різні конотації, сполучуваність різних частин мови. Зокрема, дослідження часто використовуваних віддієслівних частин турецької мови (дієприкметники– дієприслівники – віддієслівні імена / герундиви) за синтаксичними та функціональними якостями допомагає ідентифікувати морфологічну структуру слова та особливості їх використання, надає інформацію про їх розподіл за стратами (різні типи тексту), систематичні ознаки та використання в ідіомах тощо.

Всі три корпуси дають базу порівняльних даних для проведення компаративних досліджень з корпусами кримськотатарської мови (Ленара Кубедінова – Радован Гарабік, http://turk.translate.tatar; https://www.ctcorpus.org/index.php/uk/).

Література:

1. Aksan, Y., Aksan, M., Koltuksuz, A., Sezer, T. Mersinli, M., Demirhan, U. U., Yılmazer, H., Kurtoğlu, Ö., Atasoy, G. Öz, S., Yıldız, İ.2012. Construction of the Turkish National Corpus (TNC). Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012). Istanbul: Turkiye. URL: http://www.lrecconf.org/ proceedings/lrec2012/pdf/991_Paper.pdf

Aksan, Y., Özel,S. A., Yılmazer, H., & Demirhan U. 2016. The Turkish National Corpus (TNC): Comparing the Architectures of v1 and v2. 2016, 32-37.

2. Türkçe Ulusal Derlemi Web sayfası: URL: https://www.tnc.org.tr/tr/
3. Aksan, D. (1996). Türkçenin sözvarlığı. Ankara: Engin Yayınevi.
4. Aksan, D. (1980a). Sözcükbilim. Dilbilim ve Dilbilgisi Konuşmaları I, 51-74.
5. Derlem Dilbilim ve Sözcük Anlambilim. Doğan Aksan anısına. Yayına haz.: prof. Dr. Mustafa Aksan, dr. Gülsüm Atasoy. Toros Üniversitesi Yayınları, 2021. e-ISBN: 978-605-9613-05-7. 149 s.
6. Atalay Nart B., Oflazer Kemal, Say Bilge. The Annotation Process in the Turkish Treebank. URL: https://catalog.ldc.upenn.edu/ docs/LDC2015T11/turkish/turkishtreebank.pdf (12.12.2024)
7. Oflazer Kemal, Bilge Say, Dilek Zeynep Hakkani-Tur¨ , and Gokhan ¨ Tur¨ . 2003. Building a turkish Treebank //Anne Abeille,´ editor, Building and Exploiting Syntactically-annotated Corpora. Kluwer Academic Publishers.

8. Oflazer Kemal. Two-level description of Turkish morphology // Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics, April 1993.. A full version appears in Literary and Linguistic Computing, Vol.9 No.2, 1994.

DOI https://doi.org/10.36059/978-966-397-464-4-28

Іванов В. М.
аспірант кафедри українознавства, культурології та історії науки
Національного технічного університету
«Харківський політехнічний інститут»;
викладач
Харківської духовної семінарії та Вищих Свято-Володимирських
православних богословських курсів (м. Київ);
секретар
Харківського міського центру захисту історико-культурної спадщини
ГО «Українське товариство охорони пам'яток історії та культури»
м. Харків, Україна

КАРАЇМСЬКА ГРОМАДА М. ХАРКОВА: ІСТОРІЯ І СЬОГОДЕННЯ

Закон "Про корінні народи України" закріплює за караїмами статус народу, що не має іншої батьківщини, крім України.

За минуле століття їх порозкидало по країнах і континентах. Чисельність і без того маленької нації зменшилася з 13 600 у 1913 році до 1890 осіб у світі в 2002-му. В Україні, за останнім переписом 2001 року, мешкало 1196 караїмів, з яких половина − у Криму.

Скрізь, де розселялися караїми по Україні, вони створювали "маленький Крим": будували храми − кенаси, біля них мідраші − школи.

«Караїми були одним з найбагатших народів до революції 1917 року. Кількість мільйонерів на душу населення більше, ніж у інших. Навпроти магазину Габая потім перебував торговий дім братів Кальфа. Один з них, Марк, дав найбільше грошей на будівництво цієї будівлі кенаси і був її першим габбаєм − старостою, зараз би сказали − завгоспом. Дивимося пожертвування: 10 тисяч рублів. А що таке 10 тисяч у дорадянську добу? Хороший будинок у центрі міста можна було купити», − розповідає теперішній газзан Харківської кенаси