ISSN 2411-1562

STUDIA LINGUISTICA, 2020. – Vol. 17 : 112-127

Стаття надійшла до редакції 26.08.2020
Перевірено на плагіат 03.09.2020 р.
унікальність – 93.71%
https://doi.org/10.17721/StudLing2020.17.112-127

УДК 81`33+811.133. 1:004.658

ФРАНЦУЗЬКІ КОРПУСИ УСНОГО МОВЛЕННЯ: ТЕОРЕТИЧНІ І ПРАКТИЧНІ АСПЕКТИ

Ірина Володимирівна Страшко, *i.v.strashko@npu.edu.ua*
канд. філософ. наук, докторантка
Національний педагогічний університет імені М. П. Драгоманова

Розглядаються питання появи, розвитку, поширення і використання французьких корпусів усного мовлення. Аналізується специфіка зібрання, компонування та опрацювання усних даних. Вибір французької дослідницької традиції обумовлений тим, вона є недостатньо відомою в українських корпусних розвідках. Зазначається, що у порівнянні з писемними корпусами, розроблення усних французькими дослідниками відбулося із затримкою, пов'язаною головним чином з технічними причинами. Укладачі ранніх збірок звукових текстів керувалися власними правилами запису, транскрибування та збереження, тому сьогодні отримати доступ до них майже неможливо. На підставі аналізу наявних мовленнєвих корпусів встановлено, що вони вирізняються гетерогенністю, різноманітністю цілей створення, невизначеністю хронологічних меж. Відмінності в епістемологічних орієнтаціях дослідників, різниця у знаннях та інструментах, неоднорідність самих корпусів означають й різноманітність методологічних підходів до їх конституювання і експлуатації. Наголошується, що недостатність великих і багаторівневих корпусів усного мовлення, як у кількісному вираженні, так і з точки зору їх якості та наукової надійності, безпосередньо пов'язана із забезпеченням їх реалізації і залежить від поєднання наукових, технологічних та інституційних чинників. Характерна для усних корпусів проблема транскрибування включає технологічні, теоретичні й інтерпретаційні питання. Як висновок, підкреслюється, що конституювання і експлуатація французьких мовленнєвих корпусів не зводиться лише до запису голосу і не обмежується суто технічними аспектами: вони набувають значення у відкритості і доступності своїх даних.

Ключові слова: корпус усного мовлення, французька мова, звуковий файл, транскрипція, анотація.

ISSN 2411-1562

STUDIA LINGUISTICA, 2020. – Вип. 17: 112-127

FRENCH ORAL SPEECH CORPORA: THEORETICAL AND PRACTICAL ASPECTS

Iryna V. Strashko, *i.v.strashko@npu.edu.ua*
PhD, Postdoctoral Student
National Pedagogical Dragomanov University

The paper describes the issues of the origin, development, distribution and use of French oral speech corpora. The specifics of collecting, constituting and treatment of oral data were also analyzed. The choice of the French investigation tradition was caused by the fact that it is not well-known in Ukrainian corpus research. It was noted that the development of oral speech corpora by French researchers occurred with a delay, mainly due to technical reasons. The compilers of early collections of sound texts followed their own rules of recording, transcribing and saving, so today it is almost impossible to use them. Based on the analysis of the available speech corpora, it was found that they are characterized by heterogeneity, a variety of purposes for their creating, including scientific ones, and blurred chronological boundaries. Dissimilarity in researchers' epistemological orientations, differences in knowledge and tools, corpora's heterogeneousness involve a diversity of methodological approaches to their constitution and usage. It is worth mentioning that the insufficiency of big and multi-level oral speech corpora in terms of their quantity, quality and scientific reliability, is directly related to the conditions of their implementation and depends on a combination of scientific, technological and institutional factors. Oral data treatment involves transcription, which includes technological, theoretical and interpretation issues. In conclusion, it should be emphasized that the constitution and the use of French speech corpora are not limited to voice recordings and purely technical aspects as they acquire importance in the openness and availability of their data.

Key words: oral speech corpus, French language, sound file, transcription, annotation.

ФРАНЦУЗСКИЕ КОРПУСА УСТНОЙ РЕЧИ: ТЕОРЕТИЧЕСКИЕ И ПРАКТИЧЕСКИЕ АСПЕКТЫ

Страшко Ирина Владимировна, *i.v.strashko@npu.edu.ua*
канд. философ. наук, докторантка
Национальный педагогический университет имени М. П. Драгоманова

Рассматриваются вопросы возникновения, развития, распространения и использования французских корпусов устной речи. Анализируется специфика сбора, компоновки и обработки устных данных. Выбор французской исследовательской

Французькі корпуси усного мовлення: теоретичні і практичні аспекти [Французские корпуса устной речи:теоретические и практические аспекты] (Українською / На укр. яз.)

традиции обусловлен тем, что она недостаточно известна в украинских корпусных изысканиях. Отмечается, что по сравнению с письменными, разработка устных корпусов французскими исследователями произошла с задержкой, связанной главным образом с техническими причинами. Составители ранних сборников звуковых текстов руководствовались собственными правилами записи, транскрибирования и сохранения, поэтому сегодня получить доступ к ним почти невозможно. На основании анализа имеющихся речевых корпусов установлено, что они отличаются гетерогенностью, разнообразием целей создания, размытостью хронологических границ. Различия в эпистемологических ориентациях исследователей, разница в знаниях и инструментах, неоднородность самих корпусов означают и разнообразие методологических подходов к их конституированию и эксплуатации. Отмечается, что недостаточность больших и многоуровневых корпусов устной речи, как в количественном выражении, так и с точки зрения их качества и научной надежности, непосредственно связана с обеспечением их реализации и зависит от сочетания научных, технологических и институциональных факторов. Характерная для речевых корпусов проблема транскрибирования включает технологические, теоретические и интерпретационные вопросы. В заключение подчеркивается, что конституирование и эксплуатация французских речевых корпусов не сводится только к записи голоса и не ограничивается сугубо техническими аспектами: они приобретают значение в открытости и доступности своих данных.

Ключевые слова: *речевой корпус, французский язык, звуковой файл, транскрипция, аннотация.*

Проблематика створення й експлуатації корпусів усного мовлення привертає увагу дослідників у сфері лінгвістики, психології, антропології, етнології, соціології, не тільки з позицій вивчення їх розробки і застосування, різноманітності і направленості, обробки усних даних, але й аналізу типів вимови, інтонаційної варіативності, поведінки мовця, його приналежності до певної соціальної групи.

Огляд стану розробленості проблеми в українських корпусних розвідках свідчить про наявність дослідницьких прогалин щодо аналізу усних корпусів в цілому і франкомовних зокрема, що й зумовлює **актуальність** їх розгляду.

Об'єктом дослідження є французькі корпуси усного мовлення.

Предметом – теоретичні та практичні аспекти конституювання і експлуатації французьких мовленнєвих корпусів.

*© **Strashko I. V. [Strashko I. V.],** i.v.strashko@npu.edu.u*
French Oral Speech Corpora: Theoretical and Practical Aspects [Francuz'ki korpusy usnogo movlennja: teoretychni i praktychni aspekty] (in Ukrainian)

Мета статті – стислий огляд історії розробок французьких усних корпусів, специфіки їх структурування і використання, проблем опрацювання звукових даних.

Роботи французьких науковців над створенням корпусів, як усних, так і писемних, мало відомі в українському науковому дискурсі. Між тим, багато з них заслуговують на увагу. Слід назвати праці O. Baude, C. Benzitoun, C. Blanche-Benveniste, P. Cappeau, F. Gadet, J. M. Debaisieux, B. Laks та ін., в яких піднімаються питання появи, розвитку, поширення і збереження французьких корпусів усного мовлення; аналізується потенціал, результати і проблеми їх використання; розглядаються етичні, правові та технічні аспекти збору та аналізу усних даних.

Усні корпуси, як правило, складаються із аудіо записів та їх анотацій, які включають акустичні, фонетичні, лінгвістичні дані. Олів'є Бод визначає корпуси усного мовлення як "упорядковані колекції записів усних та мультимодальних лінгвістичних творів" [Baude et al. 2006, p. 19]. Ксав'є Норс, говорячи про мовленнєві корпуси, бачить в їх основі не просто колекцію записів людського мовлення, а певним чином "побудований" об'єкт. На його думку, існуючі засоби обробки даних (оцифрування, транскрипція, індексація) не тільки дозволяють їх зберегти, а й переводять їх на новий статус стосовно досліджень і розробок [op.cit., p. 11]. На наш погляд, найбільш точним є визначення О. Ф. Крівнової, яка описує усний корпус як «структуровану сукупність мовних фрагментів, що забезпечена програмними засобами доступу до них» [Кривнова 2006]. У рамках нашої роботи ми будемо дотримуватися наступного трактування корпусів усного (або як їх ще називають – звукового) мовлення: це упорядкована сукупність аудіо записів, зібраних у відповідності до попередньо визначених критеріїв, яка поєднує первинні і вторинні дані і оснащена відповідними програмними засобами доступу.

П. Каппо і Ф. Гаде констатують, що усні корпуси зазвичай набагато менші за обсягом, ніж письмові, що пов'язане з трудомістким збиранням усних даних [Cappeau, Gadet, 2007]. Проте, варто зауважити, що трудомісткість не обмежується тільки етапом збору: побудова усного корпусу є вартісною процедурою, яка вимагає часу і залежить від поєднання наукових, технологічних та інституційних чинників.

До того ж, розмір корпусів та одиниць, що їх складають, змінюється в залежності від типу дослідження і вимірюється одиницями часу та кількістю

графічних слів у транскрипції. Так, наприклад, для проведення розвідок в галузі фонетики, фонології чи просодії достатньо оперувати звуковими одиницями обмеженої тривалості, тоді як вивчення лексики чи аналіз співвідношення між мовленням та іншими мовними явищами вимагатиме більш розвиненого і більш диверсифікованого корпусу [Baude et al. 2006].

На відміну від корпусів писемних текстів, робота над якими активно велась починаючи с середини двадцятого століття, усне мовлення протягом тривалого часу було маргіналізованим, як у французькому мовознавстві, так і у корпусних розвідках [Beneviste & JeanJean 1987; Cappeau, Gadet 2007]. Відтак, типологія французьких усних корпусів представлена дещо гірше, аніж текстових. Водночас дослідження корпусів розмовного мовлення "повністю оновили науки про мову, […] дозволили сформулювати нові гіпотези про нормальне і патологічне функціонування мови і стали важливим компонентом діалогу між лінгвістами та спеціалістами з комп'ютерних технологій" [Baude et al. 2006, p. 25]. (Тут і далі переклад мій – І. Страшко). Відставання у розробці мовленнєвих корпусів пояснюються переважно технічними причинами, оскільки на відміну від тривалої писемної традиції, режими збереження звуку налічують менше ніж півтора століття, а нові технології, що дозволяють комп'ютерну обробку звукових даних, знаходяться поки що на зорі свого розвитку. Поза тим, запис усного мовлення складніше адаптувати до оброблення, тоді як написання нормалізується шляхом його подання навіть у вигляді рядка. Це вже продукт транскрипції незалежно від джерела її продукування: ментального чи звукового сигналу [Abouda, Baude 2006, p. 2]. Однак, на думку О. Бода, було б неправильно звести те незначне визнання, яке надається звуковим корпусам, виключно до технічних проблем звукозапису, компонування, архівування і розповсюдження даних. Насправді, акцентує дослідник, йдеться про статус усного мовлення, оскільки французька, особливо нормативна писемна, довгий час була не тільки єдиним загальновизнаним об'єктом наукових досліджень, а й єдиним легітимним об'єктом соціального простору, визнаним культурним капіталом нації. Водночас, зазначає О. Бод, протягом останніх років ситуація змінилася як у науковій сфері, так і в соціальному просторі [Baude 2004].

Аудіоархів, створений Фердинандом Брюно і датований 1911 роком, є свого роду першим кроком у тривалій традиції досліджень розмовної французької мови, заснованих на корпусах. Наступними можна вважати заснування фонотеки Музею людини у 1932 році, фундацію Національної

© Strashko I. V. [Strashko I. V.], i.v.strashko@npu.edu.u
French Oral Speech Corpora: Theoretical and Practical Aspects [Francuz'ki korpusy usnogo movlennja: teoretychni i praktychni aspekty] (in Ukrainian)

фонотеки у 1938 році (згодом інтегровану до Аудіовізуального департаменту Національної бібліотеки Франції), розроблення Роже Девінєм так званих "фольклорних круїзів" та створення Національної енциклопедії діалектів, говірок та старовинних пісень Франції. Подальшим кроком у розвитку досліджень усного розмовного мовлення стали етнографічні колекції, зібрані Національним музеєм народного мистецтва та традицій та Центром етнології Франції у1939 році [Baude 2004; Baude et al. 2006]. Укладачі ранніх збірок звукових текстів керувалися власними правилами запису, транскрипції та збереження, тому сьогодні отримати доступ до них майже неможливо. Зазвичай ці колекції були невеликого розміру, налічували лише кілька годин запису, а пошук інформації міг здійснюватися лише вручну. Одним із перших французьких усних корпусів, започаткованих у 1970-х роках, був проєкт дослідницької групи GARS (фр. Groupe Aixois de Recherche en Syntaxe, згодом CorpAix). З огляду на сучасні стандарти, це зібрання було дуже недосконалим: звукозаписи різної якості, не систематизовані анонімізація та експлуатаційні дозволи, відсутність вирівнювання тексту / звуку, не сумісний з інструментами запиту формат документів тощо [Caddéo, Sabio 2017]. Пізніше науковці дослідницького центру DELIC (фр. DEscription Linguistique sur Corpus), заснованого у 1999 році, відновили зібрані дослідницькою групою GARS дані, які стали інтегральною частиною нового корпусу CorpAix. У даний час результатом роботи цього наукового колективу є кілька корпусів, які стосуються окремих проєктів і мають різні характеристики (якість записів, авторизація тощо): CorpAix (фр. Corpus d'Aix-en-Provence – раніше GARS) налічує 1 700 000 слів, є транскрипція; CRFP (фр. Corpus de Référence du Français Parlé) включає близько 440 000 слів і 134 звукозаписи; Corpus DELIC (у розробці з 2000 р.) – 560 000 слів [Baude et al. 2006].

Розвиток систем розпізнавання мовлення, комп'ютеризованих баз даних, обов'язковий формалізм їх структури, стандартизація, можливість обміну даними, безсумнівно, відкрили шлях до збільшення франкомовної корпусної бази і кількості корпусних досліджень. Заснування у 2011 році Консорціуму усних і мультимодальних корпусів (http://ircom.huma-num.fr/site/accueil.php) свідчить про зростаючий інтерес французьких науковців до цієї галузі. Створений для організації та розвитку усних і мультимодальних корпусів та допомоги дослідникам у застосуванні необхідних інструментів для розробки загальних еталонних стандартів, консорціум сприяє розвитку існуючих фондів, покращенню їх відкритості,

© Страшко І. В. [Страшко И. В.], i.v.strashko@npu.edu.u

об'єднанню та сумісності. Платформа ORTOLANG (www.ortolang.fr), наприклад, нараховує 172 звукових ресурси (тривалістю від кількох хвилин до кількох сотень годин), які ґрунтуються на різному матеріалі і створені для вирішення різноманітних дослідницьких завдань.

З огляду на те, що корпуси звукового мовлення висвітлюють фактичне використання мови, більшість з них – це записи історій, діалогів, інтерв'ю, пісень, церемоній, анкетувань, поздоровлень, радіопередач тощо. Це ті джерела, які стимулюють розвиток лінгвістичних досліджень в галузі лексики, просодії, морфосинтаксису, діалектної варіативності, комунікативної інтеракції, соціолінгвістики; розробку автоматизованих систем розпізнавання мовлення тощо. Вони все більше експлуатуються й в освітніх цілях, включаючи викладання другої мови. Серед найбільш відомих назвемо корпус CLAPI (фр. Corpus de LAngues Parlées en Interaction). Це мультимедійна база даних, яка містить аудіо матеріали, записані в автентичних ситуаціях комунікативної взаємодії. CLAPI (http://clapi.ish-lyon.cnrs.fr) пропонує вільний доступ до 40 корпусів, набір пошукових інструментів з мультимедійним відображенням результатів. TCOF (фр. Traitement de Corpus Oraux en Français) – база даних усних корпусів сучасної континентальної французької мови. Створення проєкту було ініційоване бажанням науковців зберегти усні корпуси, зібрані у 80-90-ті роки для особистих дослідницьких цілей. Сьогодні TCOF (https://www.cnrtl.fr/corpus/tcof/) складається з двох основних категорій звукозаписів, тривалістю від 5 до 45 хвилин і більше: діалоги дорослих та дітей віком до 7 років та записи діалогової інтеракції дорослих. CFPP2000 (фр. Corpus de Français Parlé Parisien) – корпус розмовного мовлення районів Парижа і передмістя (http://cfpp2000.univ-paris3.fr/), початок якого датується 2005-2006 рр., сьогодні містить більше 700 000 слів. Корпуси ESLO (фр. Les Enquêtes sociolinguistiques à Orléans), розроблені на основі соціолінгвістичних опитувань. Наразі база ESLO (http://eslo.huma-num.fr/index.php) надає доступ до звукозаписів та їх транскрипцій корпусів ESLO1 та ESLO2, а в перспективі й корпусів програми "Мови в контакті". Заслуговують на увагу також корпус сучасного усного мовлення PFC (фр. Phonologie du Français Contemporain) і вибірковий корпус ESLO-MD (фр. Enquêtes Socio-Linguistiques à Orléans: Corpus Micro-Diachronie).

Великі мовленнєві корпуси, як документувальні джерела національної мови, мають загальнокультурне значення і є цінним інструментом аналізу мовних практик для проведення мовної, освітньої та соціальної політики.

© Strashko I. V. [Strashko I. V.], i.v.strashko@npu.edu.u

Розвиток усних корпусів, їх збереження і поширення робить доступною і почутою лінгвістичну спадщину Франції в її правдивості, різноманітності і багатстві [Baude et al. 2006].

При конституюванні усного корпусу розрізняють первинні і вторинні дані, різниця між якими є суттєвою, зокрема для диференціації рівнів інтерпретації і можливості повернення до первинних даних та їхньої доступності. Однак, ця різниця не повинна нівелювати того факту, що будь-який запис є результатом як технічних, так і теоретичних рішень: вибір часу запису та розмежування записаного сегмента, вибір кадру та оптики для відео, розташування та орієнтація мікрофона для аудіо тощо [Baude et al. 2006]. Метадані, як інформаційно-бібліографічні, так і типологічні, документують корпус. Вони мають бути структурованими таким чином, щоб надати якнайповнішого уявлення про корпус та продемонструвати послідовність його зібрання, компонування і аналізу, а саме: методів і форматів запису, характеристик взаємодії, властивостей та доступності первинних і вторинних даних. Саме ця остання інформація є засадничою для забезпечення його доступності, а також для пояснення критеріїв відбору й організації даних, а отже, й обмежень, які часто відсутні в доступних корпусах. Фактично, підкреслюють Л. Абуда і О. Бод, саме уточнення меж корпусу (цілей, умов виробництва, контексту використання, соціологічної інформації, жанру тощо) дозволяє оцінити його репрезентативність [Abouda, Baude 2006].

Розглядаючи дигітальні технології обробки звуку, дослідники відмічають той факт, що оцифровані звукозаписи дають змогу швидко “перегортати” звук, як це можна зробити, гортаючи сторінки писемного твору [Baude et al. 2006, p.81]. Водночас цифрові техніки, які, завдяки досконалому характеру зроблених копій, не тільки революціонізували доступ до усних корпусів, а й знищили поняття оригіналу, яке тепер стосується не стільки носія, скільки самих даних, включаючи вибір формату для забезпечення ідентичності відтворення і гарантування їх стійкості. Тим самим вони ліквідували орієнтири, які до того часу визначали межі колекцій. Звідси неможливість розрізнення першого запису, який вважається “оригінальним”, та наступних копій усного корпусу [Там само].

Наведені міркування корелюють з позицією І. Віснер, яка обстоює ідею, що усні корпуси надають доступ не до спонтанного і автентичного усного мовлення, а до записів і транскрипцій висловлювань, викладених в усній формі.

© Страшко І. В. [Страшко И. В.], i.v.strashko@npu.edu.u
Французькі корпуси усного мовлення: теоретичні і практичні аспекти [Французские корпуса устной речи:теоретические и практические аспекты] (Українською / На укр. яз.)

У своїх міркуваннях дослідниця виходить з того, що діакодичні варіації мовлення залежать від каналу або засобу комунікації. Звідси її висновок про те, що, оскільки ці записи мають різні функціональні та комунікативні модальності, усність записаних та транскрибованих мовних творів залежить від середовища спілкування і є за своєю суттю діакодичною, а не діамезичною [Wissner 2012, p. 250]. У цьому зв'язку концептуально значимим є відомий вислів М. Маклюена – "The medium is the message", – про те, що сам канал комунікації певним чином впливає на сприйняття реальної дійсності людиною. Інакше кажучи, людина може зрозуміти зміст носія комунікації, що є повідомленням, і не звернути ніякої уваги на характер цього носія – як інше, або ще одне повідомлення. Отже, у логіці сказаного, сам запис отримує операційний характер у процесі його конкретного застосування і є не тільки засобом передавання інформації, а й характеризується власною змістотворною тенденцією. Це означає, що відповідність аудіо запису усному тексту є відносною і реалізується у межах функціональної та смислової еквівалентності. М. Креч теж виходить з констатації того, що надійність і достовірність звукового сигналу ніколи не є абсолютною і рідко незіпсованою. Більше того, вона не є ні можливою, ані навіть бажаною, стверджує вона. У своїй аргументації дослідниця звертає увагу на наступні положення. Насамперед, транскрипція – це робота фонетиста, який завжди переслідує певні цілі. Результатом транскрибування є сама транскрипція, що передає не зображення предмета, який, як вважається, вона передає, а продукт усної мовленнєвої діяльності, тобто розмовний текст. Останній є лише неповним вербальним слідом, оскільки він відрізаний від звукового сигналу разом з тим, що містить вербально: просодичним позначенням (фр. *le signifiant prosodique*). Відтак візуальне сприйняття, створюючи візуальні очікування, що активуються переходом до графіки, замість того, щоб конструювати значення, дезінтегрує його і повертає читача назад до образу незавершеного продукту. Однак, визнає М. Креч, спроби зіставити те, що потрібно "показати", з тим, що запис "пропонує почути", та дослідження, завдяки яким були розроблені інструменти, що дозволяють відокремити погрішності і невідповідності, тепер здатні врахувати цей "непоправний семіотичний розрив" [Krötsch 2007, p. 37-38].

Аналізуючи повтори в усному вживанні через візуальне сприйняття, яке дає транскрипція, М. Креч говорить про те, що повторення одночасно гарантує прогресію, зв'язність, зрозумілість і інтерпретабельність усного

тексту. Пов'язане з явищами розриву мовленнєвого ланцюга, повторення, що саме сприймається як розрив, у візуальному сприйнятті транскрипту, навпаки, забезпечує безперервність. Проте, переконана дослідниця, "почути" повтори не можна. На її думку, без транскрипту слухач не звернув би на них ніякої уваги. Їх вдається вловити лише завдяки активному, багаторазовому, не залученому прослуховуванню, яким володіє фонетист. Відтак М. Креч задається питанням стосовно того, що ж насправді порівнюється. Чи не порівнюється з письмовим текстом транскрипт, взятий за розмовний текст? Але "розмовний, плюрисеміологічний, контекстуально детермінований текст просто не може порівнюватись з письмовим текстом", –- стверджує вона. У контексті цих міркувань дослідниця робить висновок про те, що "без транскрипції немає "розмовної мови", що підкреслює її евристичну цінність" [Krötsch 2007, p. 39].

Дослідники [Кривнова 2001 и др.; Baude et al. 2006] виокремлюють чотири стрижневих аспекти конституювання та експлуатації корпусів усного мовлення: технічні, змістовні, структурні та виконавські. Технічний аспект створення мовленнєвого корпусу стосується питань, що пов'язані з акустичними і технічними умовами запису мовного матеріалу. Сюди входять: вибір режиму цифрового кодування, визначення і підбір необхідної кількості мікрофонів, формат звукових файлів, акустичне середовище запису, тип каналу зв'язку тощо. Змістовна сторона зібрання усного корпусу є засадничим етапом, який передбачає множинний вибір: мовців, текстового матеріалу, типів інформації. Вибір мовців (їх кількість, стать, вік, освіта, професія, соціальне становище, діалектні відмінності мовлення, дефекти мовлення, тип вимови тощо), ситуацій запису (лабораторія, кімната звукозапису, громадські місця з великою кількістю мовців) та текстового матеріалу (зв'язні, монотематичні, політематичні тексти) ґрунтується на цілях створення корпусу: репрезентативний, спеціалізований, навчальний, ілюстративний тощо.

Цільовим призначенням корпусу визначається й вибір мовних зразків (слів, речень, текстів, прикладів дискретного, спонтанного, безперервного мовлення з різним ступенем спонтанності) і способів представлення звукового матеріалу (орфографічний запис, фонемний запис, фонетична транскрипція фактичної вимови мовця, акустико-фонетична розмітка звукового сигналу). Структурний аспект стосується організації і представлення корпусної інформації у форматі, зручному для її розміщення, зберігання, пошуку і

використання. Виконавський аспект включає питання автоматизації і стандартизації етапів створення усних корпусів. На відміну від письмових, мовленнєві корпуси стикаються з проблемою розробки стандартів для транскрипції мовних сигналів, із встановленням набору транскрипційних символів, угод про розмітку сигналів, які задають рівні транскрипції (акустичний, фонетичний, фонематичний, просодичний тощо) [Кривнова 2001 и др; Кривнова 2006; Baude et al. 2006]. Корпуси, створені з метою синтаксичного аналізу, зазвичай представлені у базовій транскрипції, без позначень вимови, тривалості мовчання чи пауз, інтонаційних схем, екстралінгвістичних факторів. Це ті свідчення які, проте, є суттєвими для розмовної або інтерактивної роботи. І навпаки, транскрипції, встановлені для прагматичного аналізу, можуть містити інформацію, менш значиму для синтаксичного (напр. довжина пауз) [Cappeau, Gadet, 2007]. У деяких випадках при вільній зміні форми, коли мова йде про інформативний контент і зберігається тільки зміст записів, більш придатними будуть терміни *транспонування* або *адаптація*, які використовуються в журналістиці і соціології, коли сортуються дані, вирізаються уривки, видаляються особливості усного мовлення (повторення, вагання, кашель, сміх тощо). Великі корпуси розмовного мовлення транскрибовані, як правило, стандартним орфографічним написанням, що пояснюється необхідністю зробити їх читання легкодоступним. Цим обумовлюється й наявність декількох варіантів: з адаптацією або без, з розділовими знаками або без, із зазначенням пауз або без, зі збереженням індивідуальних особливостей розгортання мовленнєвого потоку [Baude 2006]. Більшість програмного забезпечення призначене для мультитранскрипції, тому пропонує гнучку деталізацію та неоднорідне подання (перший рядок транскрипції може містити орфографічний запис, другий – фонетичну транскрипцію, третій може бути пристосованим для конкретного пошуку тощо) [Baude 2004]. До того ж, обробку деяких мовних явищ не завжди можна автоматизувати. Відтак процес стає часозатратним і вимагає кропіткої кваліфікованої експертної роботи для досягнення відповідного ступеня надійності.

Питання використання усних корпусів не обмежуються лише технічним вибором: різноманітність цілей, у тому числі й наукових, пов'язаних зі створенням, експлуатацією, збереженням і розповсюдженням корпусів, відмінності у епістемологічних орієнтаціях дослідників, різниця у знаннях та інструментах, гетерогенність самих корпусів означає й різноманітність

© Strashko I. V. [Strashko I. V.], i.v.strashko@npu.edu.u
French Oral Speech Corpora: Theoretical and Practical Aspects [Francuz'ki korpusy usnogo movlennja: teoretychni i praktychni aspekty] (in Ukrainian)

методологічних підходів. Справа не тільки у недостатності великих репрезентативних корпусів усного мовлення – як у кількісному вираженні, так і з точки зору їх якості та наукової надійності – справа ще й у диспропорційності сфер застосування, оскільки переважають дослідження з фонології чи просодії. Іншим важливим чинником, який зумовлює і, в кінцевому підсумку, призводить до перерахованих проблем, є фрагментація розвідок і стандартів для збору, збереження та кодифікації: дослідження часто проводяться без екстралінгвальних цілей, які могли б забезпечити відкрите використання усних фондів. Дана обставина актуалізує проблему вимог до хронологічних меж корпусів усного мовлення. Л. Абуда і О. Бод пов'язують її з тим, що мовленнєві корпуси досить часто формуються відповідно до спеціальних цілей і визначених завдань, чим обумовлюється їх остаточність, завершеність: наприклад, розпізнавання голосу або інтелектуальний аналіз тексту. Проблематичність даної ситуації посилюється й недосконалістю каталогізації та опису ресурсів, а саме: стандартизації, індексації та консультування [Abouda, Baude 2006]. Відсутність або обмеженість відповідного інструментарію (середовище з належною ємністю зберігання, обладнання для оцифрування, відповідне програмне забезпечення), людських ресурсів (для переведення даних в цифровий формат, їх вирівнювання, розшифрування), можливості співробітництва з фахівцями з різних галузей (зокрема, зі спеціалістами в галузі комп'ютерних технологій, дослідниками і інженерами, що спеціалізуються на автоматичній обробці мови) для оцінювання потреб і просування технічної реалізації проєкту теж утруднюють конституювання корпусів.

Виникають труднощі і при транскрипційній роботі із зібраними звуковими джерелами. За твердженням К. Бензітуна, перша проблема − це відсутність пунктуаційного оформлення, а отже, графічних "речень". Використання системи пунктуації писемного мовлення, на його думку, є недостатнім для транскрибування усного слова, оскільки воно не відповідає чітко ідентифікованим знакам, а його позначення вже саме по собі становить неявний синтаксичний аналіз. Більш того, паузи, переконаний дослідник, теж не можуть розглядатися як знаки пунктуації [Benzitoun 2004, p. 14]. Справді, паузи сприяють передаванню синтаксичних і смислових відносин, як правило виступаючи як сигніфікативний засіб, сегментуючи мовленнєвий потік. Проте не кожен темпоральний інтервал можна трактувати як цілком делімітативний, оскільки не слід забувати про існування фізіологічних, хезитативних пауз, які

характерні, наприклад, для спонтанного мовлення. Іншим проблемним аспектом є акустична варіативність звукових одиниць, спричинена коартикуляцією, психофізіологічним станом мовця, технічними параметрами запису (характеристики мікрофона, його розміщення). Лінійність перебігу усного мовлення, яка породжує типові для нього способи організації і продукування висловлювання (самоперебивання, повторення слів або груп слів, переформулювання, незакінчені вислови, дискурсивні маркери та ін., які значно рідше фіксуються у писемних корпусах) теж спричиняє ускладнення у транскрибуванні усного тексту. Надійним критерієм для сегментації не може бути й інтонація мовця, особливо у спонтанному мовленні, оскільки вона не є абсолютно відповідною пунктуації. У такому разі транскрибування стає інтерпретаційним, а фонетист, ґрунтуючись на власній перцепції, власному розумінні смислу почутого при членуванні мовленнєвого потоку, змушений робити свій пунктуаційний вибір. При цьому сенс може змінитися або бути втраченим. У цьому контексті цікавими є приклади, наведені Ж.-М. Дебезьє щодо явища "подвійного прослуховування" у випадку розрізнення двох приголосних і двох голосних: "*c'est /**les/des/** sages-femmes qui vont accueillir ben des gens …", "…on va /**les, le**/ suivre tous les jours"*), двох омофонів: "*donc /**c'est**, **ces**/ deux filles qui font ça…*". У цьому разі неможливо розрізнити два трактування, оскільки при транскрибуванні реконструюється й значення. Слід враховувати й складність співвідношення правопис-мовлення у французькій мові, що вимагає відповідної підготовки спеціалістів [Debaisieux 2005, p. 17]. Прагнення наблизити транскрипцію до вимови призводить до нестандартних орфографічних рішень, акцентують П. Каппо і Ф. Гаде, і як приклад наводять написання *is viennent* та *aez*, зафіксовані Л. Форе у корпусі Sérignan замість відповідно *ils viennent* і *allez* [Cappeau, Gadet 2007, p. 107]. Такі випадки невнормованого написання, підкреслюють дослідники, породжують проблеми вибору способу транскрибування і застосування відповідних технологій. І якщо перший може мати наслідком дискредитування деяких інформантів, якого неможливо уникнути, зважаючи на навички читання в культурі грамотності, то другий, технологічний аспект, пов'язаний з автоматичними процедурами обробки мовного матеріалу, стосується питань відповідності, достовірності і надійності даних, а отже й можливості їх експлуатації, оскільки змінені форми правопису не можуть бути ідентифіковані за допомогою комп'ютерного пошуку і в кінцевому підсумку будуть втрачені. Ілюстрацією наведених міркувань слугує існуюча різноманітність

транскрибувань слова alors (*alors, aor, ar, or,* або *lors*), яка суттєво утруднює пошук необхідної форми [Там само].

Як **висновок**, можемо сказати, що конституювання корпусу усного мовлення не зводиться лише до запису голосу. Його експлуатація теж не обмежується суто теоретичними і технологічними аспектами: він набуває значення у відкритості і доступності своїх наукових, технічних, звукових даних. Усі ці елементи у сукупності складають нерозривну єдність, що надає корпусу цілісності та цінності. З плином часу французькі корпуси усного мовлення множаться, еволюціонують технології комп'ютерної обробки звукових даних, зростає значення, яке надається корпусам у наукових галузях, розвиваються транскрипція, анотація і розмітка, інструменти їх збереження і поширення. Осмислення досвіду французьких науковців щодо конституювання і експлуатації корпусів усного мовлення сприятиме виведенню українських корпусних розвідок на новий рівень розвитку. Перспективним напрямком подальших досліджень видається компаративний аналіз французьких корпусів усного мовлення.

Література

1. Кривнова О. Ф., Захаров Л. М., Строкин Г. С., "Речевые корпусы (опыт разработки и использование)", *Труды международного семинара Диалог,* (2001).

2. Кривнова, О. Ф., "Области применения речевых корпусов и опыт их разработки", *Тр. XVIII Сессии Российского акустического общества РАО*, Таганрог (2006).

3. Lofti Abouda, Oliver Baude, "Constituer et exploiter un grand corpus oral: choix et enjeux théoriques. Le cas des ESLO", (2006).

4. Lofti Abouda, Marie Skrovec, "Pour une micro-diachronie de l'oral: le corpus ESLO-MD", SHS Web of Conferences. - EDP Sciences, (46), (2018): 11004.

5. Oliver Baude, "Les corpus oraux entre science et patrimoine. L'expérience de l'Observatoire des pratiques linguistiques", (2004).

6. Olivier Baude, Claire Blanche-Benveniste, Marie-France Calas, Paul Cappeau, Pascal Cordereix, et al.. *Corpus oraux, guide des bonnes pratiques*. CNRS Editions, Presses Universitaires Orléans, 2006.

7. Benzitoun, Christophe, "L'annotation syntaxique de corpus oraux constitue-t-elle un problème spécifique", *In Actes de la conférence RECITAL,*(April 2004): 13-22.

8. Benzitoun, Christophe, Karen, Fort, Benoît, Sagot. "TCOF-POS: un corpus libre de français parlé annoté en morphosyntaxe." *JEP-TALN 2012 - Journées d'Études sur la Parole et conférence annuelle du Traitement Automatique des Langues Naturelles*, Jun 2012, Grenoble, France: 99-112.

9.Blanche-Benveniste, Claire, and Colette Jeanjean. *Le français parlé: transcription et édition.* Éditions Interco, 1987.

© Страшко І. В. [Страшко И. В.], *i.v.strashko@npu.edu.u*
Французькі корпуси усного мовлення: теоретичні і практичні аспекти [Французские корпуса устной речи:теоретические и практические аспекты] (Українською / На укр. яз.)

10. Claire Blanche-Benveniste. “Constitution et utilisation d’un grand corpus, Grands corpus: diversité des objectifs, variété des approches”. *Revue Française de Linguistique Appliquée* 4,1(1999): 65-74.

11. Sandrine Caddéo, Frédéric Sabio, “Le Groupe Aixois de Recherche en Syntaxe et les recherches actuelles sur le français parlé”, *Repères DoRiF* n.12 - *Les z'oraux - Les français parlés entre sons et discours* - Coordonné par Enrica Galazzi et Marie-Christine Jamet, DoRiF Università, Roma juillet 2017, http://www.dorif.it/ezine/ezine_articles.php?id=340

12. André, Virginie, and Emmanuelle Canut. “Mise à disposition de corpus oraux interactifs: le projet TCOF (traitement de corpus oraux en francais).” *Pratiques. Linguistique, littérature, didactique* 147-148 (2010): 35-51.

13. Cappeau, Paul, Gadet, Françoise, “L'exploitation sociolinguistique des grands corpus”, Revue française de linguistique appliquée, 12(1), (2007): 99-110.

14. Debaisieux, Jeanne-Marie, “Les corpus oraux: situation, exploitation linguistique, bilan et perspectives”, Scolia, Université des sciences humaines Strasbourg, (2005): 9-40.

15. Jacobson, Michel, “Corpus oraux en linguistique de terrain. Traitement automatique des langues”, ATALA 45(2), (2004): 63-88.

16. Krötsch, Monique, “Répétition et progression en français parlé”. Linx. *Revue des linguistes de l’université Paris X Nanterre*, (57), (2007): 37-46.

17. Mondada, Lorenza, “Pratiques de transcription et effets de catégorisation”, *Cahiers de praxématique*, (39) (2002): 45-75. DOI : https://doi.org/10.4000/praxematique.1835

18. Wissner, Inka, “Les grands corpus du français moderne: des outils pour étudier le lexique diatopiquement marqué”, *SKY Journal of Linguistics*, 25(2012): 233-272.

19. CLAPI: Corpus de LAngue Parlée en Interaction. http://clapi.univ-lyon2.fr/

20. ESLO: Enquêtes Sociolinguistiques à Orléans, Université d’Orléans. http://eslo.huma-num.fr/index.php

21. ORTOLANG:Open Resources and TOols for LANGuage. www.ortolang.fr

22. PFC: Corpus Phonologie du Français Contemporain. http://www.projet-pfc.net/

23. TCOF: Traitement des Corpus Oraux en Français. http://www.cnrtl.fr/corpus/tcof/

References

1. Krivnova, O. F., Zakharov, L. M., & Strokin, G. S. (2001). «Rechevyye korpusy (opyt razrabotki i ispol'zovaniye) [Speech corpora (development experience and use)]». *Trudy mezhdunarodnogo seminara Dialog*. (in Russ.).

2. Krivnova, O. F. (2006). «Oblasti primeneniya rechevykh korpusov i opyt ikh razrabotki [Scopes of speech corpora and experience in their development]». *Tr. XVIII Sessii Rossiyskogo akusticheskogo obshchestva RAO. Taganrog* (in Russ.).

3. Lofti Abouda, Oliver Baude, “Constituer et exploiter un grand corpus oral: choix et enjeux théoriques. Le cas des ESLO”, (2006).

4. Lofti Abouda, Marie Skrovec, “Pour une micro-diachronie de l’oral: le corpus ESLO-MD”, *SHS Web of Conferences. - EDP Sciences*, (46), (2018): 11004.

*© **Strashko I. V.** **[Strashko I. V.]**, i.v.strashko@npu.edu.u*
French Oral Speech Corpora: Theoretical and Practical Aspects [Francuz'ki korpusy usnogo movlennja: teoretychni i praktychni aspekty] (in Ukrainian)

5. Oliver Baude, “Les corpus oraux entre science et patrimoine. L'expérience de l'Observatoire des pratiques linguistiques”, (2004).

6. Olivier Baude, Claire Blanche-Benveniste, Marie-France Calas, Paul Cappeau, Pascal Cordereix, et al.. *Corpus oraux, guide des bonnes pratiques.* CNRS Editions, Presses Universitaires Orléans, 2006.

7. Benzitoun, Christophe, “L’annotation syntaxique de corpus oraux constitue-t-elle un problème spécifique”, *In Actes de la conférence RECITAL,*(April 2004): 13-22.

8. Benzitoun, Christophe, Karen, Fort, Benoît, Sagot. “TCOF-POS: un corpus libre de français parlé annoté en morphosyntaxe.” *JEP-TALN 2012 - Journées d’Études sur la Parole et conférence annuelle du Traitement Automatique des Langues Naturelles*, Jun 2012, Grenoble, France: 99-112.

9.Blanche-Benveniste, Claire, and Colette Jeanjean. *Le français parlé: transcription et édition.* Éditions Interco, 1987.

10. Claire Blanche-Benveniste. “Constitution et utilisation d’un grand corpus, Grands corpus: diversité des objectifs, variété des approches”. *Revue Française de Linguistique Appliquée* 4,1(1999): 65-74.

11. Sandrine Caddéo, Frédéric Sabio, “Le Groupe Aixois de Recherche en Syntaxe et les recherches actuelles sur le français parlé”, *Repères DoRiF* n.12 - *Les z'oraux - Les français parlés entre sons et discours* - Coordonné par Enrica Galazzi et Marie-Christine Jamet, DoRiF Università, Roma juillet 2017, http://www.dorif.it/ezine/ezine_articles.php?id=340

12. André, Virginie, and Emmanuelle Canut. “Mise à disposition de corpus oraux interactifs: le projet TCOF (traitement de corpus oraux en francais).” *Pratiques. Linguistique, littérature, didactique* 147-148 (2010): 35-51.

13. Cappeau, Paul, Gadet, Françoise, “L'exploitation sociolinguistique des grands corpus”, Revue française de linguistique appliquée, 12(1), (2007): 99-110.

14. Debaisieux, Jeanne-Marie, “Les corpus oraux: situation, exploitation linguistique, bilan et perspectives”, Scolia, Université des sciences humaines Strasbourg, (2005): 9-40.

15. Jacobson, Michel, “Corpus oraux en linguistique de terrain. Traitement automatique des langues”, ATALA 45(2), (2004): 63-88.

16. Krötsch, Monique, “Répétition et progression en français parlé”. Linx. *Revue des linguistes de l’université Paris X Nanterre*, (57), (2007): 37-46.

17. Mondada, Lorenza, “Pratiques de transcription et effets de catégorisation”, *Cahiers de praxématique*, (39) (2002): 45-75. DOI : https://doi.org/10.4000/praxematique.1835

18. Wissner, Inka, “Les grands corpus du français moderne: des outils pour étudier le lexique diatopiquement marqué”, *SKY Journal of Linguistics*, 25(2012): 233-272.

19. CLAPI: Corpus de LAngue Parlée en Interaction. http://clapi.univ-lyon2.fr/

20. ESLO: Enquêtes Sociolinguistiques à Orléans, Université d’Orléans. http://eslo.huma-num.fr/index.php

21. ORTOLANG:Open Resources and TOols for LANGuage. www.ortolang.fr

22. PFC: Corpus Phonologie du Français Contemporain. http://www.projet-pfc.net/

23. TCOF: Traitement des Corpus Oraux en Français. http://www.cnrtl.fr/corpus/tcof/

© Страшко І. В. [Страшко И. В.], *i.v.strashko@npu.edu.u*