УДК 81'33

КОРПУС УКРАЇНСЬКОГО ТРАНСКРИБОВАНОГО УСНОГО МОВЛЕННЯ: ЗАСАДИ СТВОРЕННЯ

Плахотнікова О. Ю.
Київський національний університет імені Тараса Шевченка

У статті описано особливості створення Корпусу українського транскрибованого усного мовлення з використанням комп'ютерної програми ELAN. Особливу увагу звернуто на характеристику основних етапів і принципів формування корпусу усного мовлення.

Ключові слова: *корпус усного мовлення, українське мовлення, анотація, спрощена фонетична транскрипція, комп'ютерна програма ELAN.*

Плахотникова Е. Ю. Корпус украинской транскрибированной устной речи: принципы создания. *В статье рассматриваются особенности создания Корпуса украинской транскрибированной устной речи с помощью компьютерной программы ELAN. Особое внимание уделяется характеристике основных этапов и принципов формирования корпуса устной речи.*

Ключевые слова: *корпус устной речи, украинская речь, аннотация, упрощённая фонетическая транскрипция, компьютерная программа ELAN.*

Plakhotnikova O. Yu. The corpus of transcribed Ukrainian speech: principles of construction. *A corpus-based approach is one of the leading methods for speech analysis in modern linguistics nowadays. Ukrainian corpus resources mostly provide material for linguistic studies of written texts.*

Key words: *speech corpus, Ukrainian speech, annotation, broad phonetic transcription, ELAN computer program.*

Постановка проблеми та обґрунтування актуальності її розгляду. Сучасний стан розвитку українського мовознавства передбачає застосування комп'ютерних технологій для аналізу мовлення; сьогодні одним із провідних є корпусно базований підхід до опрацювання значного за обсягом мовленнєвого матеріалу. Корпусні ресурси, створені в Україні, забезпечують здебільшого матеріал для лінгвістичних досліджень писемних текстів. Проте ґрунтовні фонетичні дослідження українського мовлення потребують фонетично анотованих корпусів літературного усного мовлення, розміщених у відкритому доступі. Саме тому на базі навчальної лабораторії експериментальної фонетики Інституту філології Київського національного університету імені Тараса Шевченка (далі – ЛЕФ КНУ) створено Корпус українського транскрибованого усного мовлення (далі – Корпус, КУТУМ) із фонетичною анотацією, що містить транскрипцію на основі як кириличної, так і латинської графіки (символів Міжнародного фонетичного алфавіту, IPA) [3, 40].

Аналіз останніх досліджень і публікацій. Корпус як основне поняття корпусної лінгвістики в широкому розумінні означає зібрання текстів природної мови, придатне для загального лінгвістичного аналізу [8]. Проблемам корпусної лінгвістики присвячено конференції та збірники наукових праць за їх матеріалами (зокрема Міжнародна конференція з прикладної корпусної лінгвістики, International Conference on Applied Corpus Linguistics); також існують спеціальні фахові видання для висвітлення здобутків у галузі корпусної лінгвістики (наприклад, "International Journal of Corpus Linguistics", "Corpus Linguistics and Linguistic Theory", "Corpora"). Корпусний підхід не обмежується лише створенням ресурсів на матеріалі писемних текстів. Окремою сферою корпусної лінгвістики є укладання корпусів усного мовлення.

Метою статті є опис особливостей формування КУТУМ на базі комп'ютерної програми ELAN.

Завдання статті – 1) розгляд основних етапів створення Корпусу; 2) опис процедур пошуку в КУТУМ; 3) окреслення перспектив використання здобутків КУТУМ у наукових дослідженнях.

У статті подано модель створення корпусу усного мовлення на базі комп'ютерної програми ELAN; теоретичні засади, покладені в основу цієї моделі, можуть бути використані для укладання корпусу усного мовлення будь-якої мови. Це зумовлює актуальність роботи.

Виклад основного матеріалу дослідження. Створення КУТУМ передбачало три основні етапи: 1) загальне проектування КУТУМ; 2) збирання експериментального матеріалу дослідження; 3) опрацювання корпусної інформації (зокрема створення анотаційних файлів для аудіозаписів на базі комп'ютерної програми ELAN).

Загальне проектування КУТУМ. Одним із важливих завдань цього етапу було обґрунтування структурних характеристик КУТУМ. Набір індивідуальних базових характеристик для КУТУМ зумовлений класифікаціями сучасних електронних корпусів усного мовлення. Отже, КУТУМ – це корпус українського мовлення з такими базовими структурними характеристиками: 1) дослідницький: призначений для вивчення функціонування сучасної української мови у фонетичному аспекті з можливістю виходу на широкий спектр лінгвістичних досліджень укра-

© Плахотнікова О. Ю. Корпус українського транскрибованого усного мовлення: засади створення

їнського мовлення; 2) фрагментний: містить фрагменти аудіотекстів, дібрані за визначеними принципами відбору текстових даних до корпусу (читані тексти художнього стилю українського мовлення і фрагменти спонтанного українського мовлення) для досягнення репрезентативності корпусної інформації; 3) динамічний: передбачено постійне поповнення бази корпусних аудіотекстів і створення відповідних анотаційних файлів у форматі *.eaf за визначеною процедурою; у такий спосіб можна спостерігати варіативність фонетичної системи української мови; 4) синхронний: Корпус представлений звукозаписами, що відображають стан сучасного українського мовлення; 5) мономовний: тексти, що увійшли до Корпусу, є результатом мовленнєвої діяльності носіїв української мови; 6) фонетично анотований: відповідно до вимог Стандартів кодування корпусу, усі текстові дані забезпечені анотацією, де кожний звуковий фрагмент супроводжується відповідною інформацією про його фонетичну специфіку (у форматі спрощеної фонетичної транскрипції двох видів – засобами кириличного алфавіту й засобами IPA) та в орфографічному записі.

Збирання експериментального матеріалу дослідження. Джерелом мовленнєвого матеріалу для створення КУТУМ обрано аудіозаписи українського літературного мовлення. Сьогодні фрагменти корпусу усних текстів загальною тривалістю 123 хв., що слугували матеріалом КУТУМ, – це 35 аудіозаписів у *.wav-форматі мовлення викладачів, аспірантів і студентів Інституту філології Київського національного університету імені Тараса Шевченка, а також одного актора й одного депутата Верховної Ради України (2 аудіозаписи – спонтанне мовлення, решта – читане мовлення) і 59 анотаційних файлів у форматі *.eaf, створені для аудіозаписів на базі комп'ютерної програми ELAN.

Більшість аудіозаписів, відібраних для КУТУМ (33 аудіозаписи загальною тривалістю понад 116 хв.), належать до акустичного фонду ЛЕФ КНУ; ці аудіотексти були записані в студії звукозапису ЛЕФ КНУ з використанням динамічного мікрофона упродовж 2003–2011 рр. Спеціально для КУТУМ у 2014 р. у ЛЕФ КНУ здійснено додатковий запис тривалістю 4 хв. 52 сек. із залученням диктора – викладача Інституту філології (це прочитаний текст добірки «Неймовірні факти про Україну», що становить стиль засобів масової інформації). Ще один аудіозапис 2009 р. взято з бази даних нашого попереднього дослідження; це зразок спонтанного професійного мовлення [1]. Особливу увагу звернуто на дотримання дикторами норм літературного мовлення, адже досліджувався саме літературний вияв української мови.

Добір дикторів для КУТУМ здійснено за кількома критеріями з урахуванням аспекту збалансованості корпусу: а) за статтю: загалом для КУТУМ використано аудіозаписи 10 дикторів (6 – жінок, 4 – чоловіків); б) за віком диктори належать до різних вікових груп: 18–25 рр., 30–45 рр. і 50–65 рр.; в) за рідною мовою: українська є рідною мовою всіх дикторів; г) за мінімальним діалектним впливом: у жодного з дикторів явища діалектного мовлення не простежуються систематично; ґ) за вищою освітою: усі диктори мають повну або неповну вищу освіту; що ж до професійної належності дикторів, то вона пов'язана з публічним мовленням і філологією.

Базу електронних текстів Корпусу становлять прочитані дикторами фрагменти текстів таких функціональних стилів української літературної мови: художнього (фрагменти оповідань М. Коцюбинського «Ялинка» та новела "Intermezzo"), наукового (фрагмент лекції Ліни Костенко «Гуманітарна аура нації, або Дефект головного дзеркала», фрагмент тексту «Філософія Григорія Сковороди» з книги В. Горського «Історія української філософії: курс лекцій»), стилю засобів масової інформації (фрагменти інтернет-статті «Неймовірні факти про Україну»). Спонтанне мовлення охоплює розмовно-побутовий стиль (1 аудіозапис мовлення аспірантки Інституту філології) та усне професійне мовлення (1 аудіозапис мовлення депутата Верховної Ради України).

Опрацювання корпусної інформації. Для КУТУМ використано аудіозаписи у форматі *wav (моно); аудіотексти відредаговано для подальшого користування в Корпусі. На цьому етапі також здійснено метаопис аудіозаписів Корпусу, що об'єднував як змістові елементи даних (відомості про автора тексту, назву тексту, прізвище диктора), так і формальні (номер аудіозапису, прізвище автора анотації, розширення файлу); ці відомості введено вручну. Усі аудіозаписи розміщено в одній папці, що й становить базу даних КУТУМ (кожному аудіозапису відповідає анотаційний файл(и) з відповідною назвою); це зумовлено особливостями роботи комп'ютерної програми ELAN, оскільки без аудіозапису, розташованого в одній папці з анотаційним файлом *eaf, перегляд анотаційного файлу неможливий (буде відсутній звуковий сигнал).

Заради зручності опрацювання масиву даних і побудови анотацій з орфографічними записами створено базу електронних текстів, що співвідносяться з відповідними аудіозаписами. Електронні версії художніх текстів узято з мережі Інтернет (надалі тексти проходили філологічну перевірку й корегування, зокрема скорочення фрагментів). Орфографічний запис текстів аудіозаписів виконаний з урахуванням мовлення дикторів.

Для етапу створення анотаційних файлів аудіозаписів Корпусу використано комп'ютерну програму ELAN – відкритий безплатний програмний ресурс для розроблення складних анотацій, створений в Інституті психолінгвістики імені Макса Планка (The Language Archive, м. Неймеген, Нідерланди) [7, 847; 9]. Ця програма апробована у світовій практиці: на її основі вже сформовано чимало мультимедійних корпусів у всьому світі [3, 40]. Загальні принципи створення анотаційних файлів для аудіозаписів у програмі ELAN простежено в проекті «Рассказы о сновидениях и другие корпуса звучащей речи» [5] і на сайті Інституту психолінгвістики імені Макса Планка в Неймегені [9].

Створення анотаційного файлу в програмі ELAN для кожного аудіозапису (медіафайлу) відбувається в режимах розмітки, сегментації й транскрипції. Загалом можна окреслити чотири етапи створення анота-

ційного файлу аудіозапису: 1) визначення структури рівнів анотації в режимі розмітки; 2) сегментація звукового сигналу в режимі сегментації та введення анотацій у режимі розмітки; 3) транскрибування аудіотексту в режимі транскрипції; 4) перевірка й корекція транскрипційних записів (виправлення помилок і двозначностей) [2, 241]. Етапи створення транскрипції аудіотексту в програмі ELAN докладно описані в науковій статті [2, 240–242]; зміни й доповнення щодо принципів і особливостей транскрибування усного мовлення в програмі ELAN викладено в іншій науковій публікації [3].

Для нашого Корпусу також створено документаційне забезпечення, в якому, зокрема, описано особливості роботи над транскрибуванням аудіозаписів. Методику сегментації й транскрибування акустичного сигналу на базі комп'ютерної програми ELAN апробовано в навчальній практиці з експериментальної фонетики студентів 2 курсу спеціальності «Прикладна лінгвістика» Інституту філології Київського національного університету [4], а також у межах спецкурсів студентів-магістрів Інституту філології (спеціальність «Українська мова і література, іноземна мова») та Київського національного лінгвістичного університету (спеціальність «Прикладна лінгвістика»).

КУТУМ плануємо поширювати на електронних носіях інформації, зробити доступним для користувачів глобальної мережі Інтернет (на основі HTML-версій анотаційних файлів, а також із використанням медіафайлів у форматі *.wav та анотаційних файлів у форматі *.eaf).

Опції пошуку в КУТУМ на базі комп'ютерної програми ELAN. У комп'ютерній програмі ELAN передбачено такі функції пошуку в анотаційних файлах у форматі *.eaf: 1) пошук у межах одного анотаційного файлу та відповідне відображення отриманих результатів; 2) пошук у кількох анотаційних файлах, розташованих в одній папці, з можливістю структурувати пошуковий запит [6, 333]. Для корпусу, що містить транскрибоване усне мовлення, актуальним є пошук фонетичних явищ у межах транскрипційних записів анотаційних файлів із розширенням *.eaf. При цьому важливо правильно формулювати пошуковий запит, він має відповідати передусім транскрипційним позначенням, використовуваним у Корпусі (якщо йдеться про пошук у транскрипційних записах).

З метою здійснення простого пошуку в межах одного анотаційного файлу КУТУМ у режимі розмітки необхідно зайти в меню «Пошук» та обрати опцію «Знайти (і замінити)». У результаті відкриється діалогове вікно «Діалог пошуку», у якому можна вказати необхідні критерії пошуку (зокрема врахування регістру, інтервал анотацій, пошук на конкретних рівнях анотацій тощо) [6, 333–337]. Відображення результатів пошуку відбувається в діалоговому вікні «Діалог пошуку» (див. рис. 1). Зокрема, тут буде запропоновано таку інформацію: кількість анотацій, що містять зазначену одиницю пошуку; рівень, що містить анотацію; повний зміст кожної анотації, де міститься вказана одиниця пошуку; час початку, закінчення й загальної тривалості кожної анотації, що відповідає критеріям пошуку. Тобто результат пошуку відображений у вигляді конкордансу (списку слововживань у контекстуальному оточенні) [6, 338].

Після послідовного виконання кількох пошукових запитів за один сеанс користувач має можливість переглянути автоматично створену історію пошуку. Ця опція дає змогу здійснювати пошук у межах усіх уведених запитів. Результати пошукових запитів можна зберігати й експортувати у вигляді текстового файлу [6, 339–344].

З метою доступу до опції пошуку в кількох анотаційних файлах необхідно в режимі розмітки у відкритому анотаційному файлі зайти в меню «Пошук» та обрати опцію «Пошук у кількох файлах EAF». У діалоговому вікні потрібно вказати добірку файлів, у якій буде здійснено пошук (усі файли обов'язково мають бути розміщені в одній папці), а також об'єкт пошуку [6, 346–347]. Результати пошуку сформовані у вигляді конкордансу (див. рис. 2). Указане діалогове вікно містить такі поля для показу знайдених анотацій: номер анотації за порядком у переліку результатів пошуку; назва файлу, що містить знайдений результат; рівень, що містить результат; анотації, що розташовані перед або після анотації, яка відповідає параметрам пошуку; анотація, що є результатом пошуку; час початку, закінчення й тривалості анотаційного сегмента [6, 348].

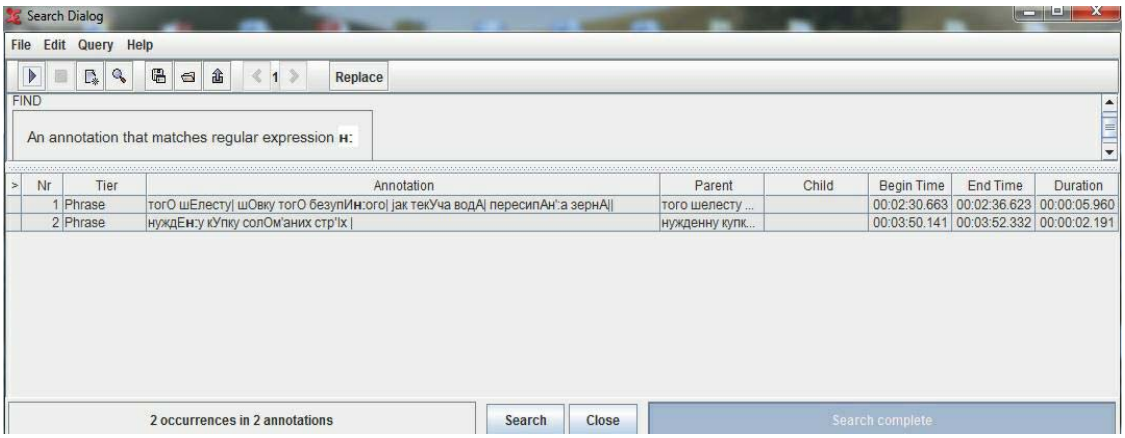
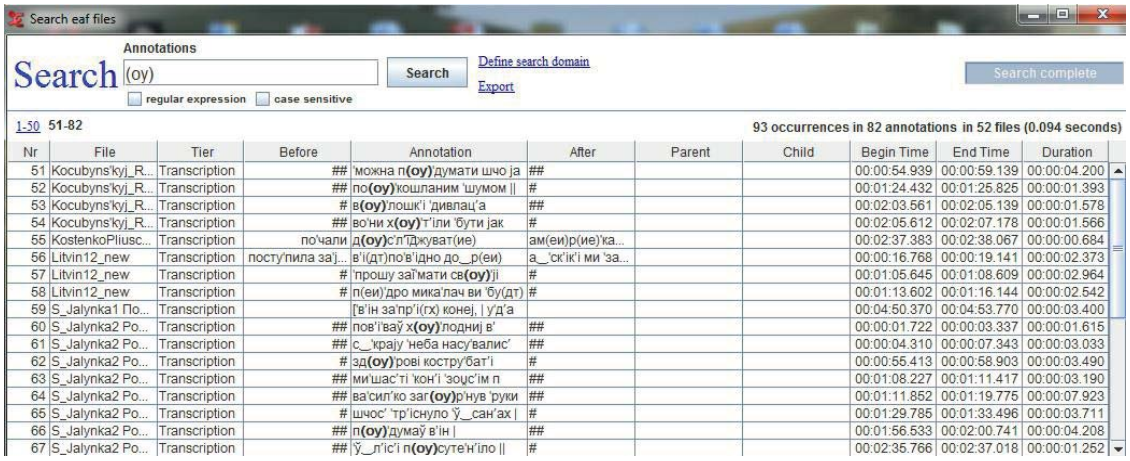


Рис. 1. Результати простого пошуку в межах одного анотаційного файлу


Search eaf files

Annotations

Search (oy) Search Define search domain Export Search complete

regular expression case sensitive

1-50 51-82 93 occurrences in 82 annotations in 52 files (0.094 seconds)

Nr	File	Tier	Before	Annotation	After	Parent	Child	Begin Time	End Time	Duration
51	Kocubyns'kyj_R...	Transcription	##	'можна п**(оу)**'думати шчо ја	##			00:00:54.939	00:00:59.139	00:00:04.200
52	Kocubyns'kyj_R...	Transcription	##	по**(оу)**'кошланим 'шумом ||	#			00:01:24.432	00:01:25.825	00:00:01.393
53	Kocubyns'kyj_R...	Transcription	#	в**(оу)**'лошк'і 'дивлац'а	##			00:02:03.561	00:02:05.139	00:00:01.578
54	Kocubyns'kyj_R...	Transcription	##	во'ни х**(оу)**'т'іли 'бути јак	#			00:02:05.612	00:02:07.178	00:00:01.566
55	KostenkoPliusc...	Transcription	по'чали	д**(оу)**с'л'ідкуват(ие)	ам(еи)р(ие)'ка...			00:02:37.383	00:02:38.067	00:00:00.684
56	Litvin12_new	Transcription	посту'пила заj...	в'і(дт)по'в'ідно до_р(еи)	а_'ск'ік'і ми 'за...			00:00:16.768	00:00:19.141	00:00:02.373
57	Litvin12_new	Transcription	#	'прошу заї'мати св**(оу)**'ji	#			00:01:05.645	00:01:08.609	00:00:02.964
58	Litvin12_new	Transcription	#	п(еи)'дро мика'лач ви 'бу(дт)	#			00:01:13.602	00:01:16.144	00:00:02.542
59	S_Jalynka1 По...	Transcription		['в'ін за'пр'і(гх) конеj, | у'д'а				00:04:50.370	00:04:53.770	00:00:03.400
60	S_Jalynka2 Po...	Transcription	##	пов'і'ваў х**(оу)**'лодниj в'	##			00:00:01.722	00:00:03.337	00:00:01.615
61	S_Jalynka2 Po...	Transcription	##	с_'крају 'неба насу'валис'	##			00:00:04.310	00:00:07.343	00:00:03.033
62	S_Jalynka2 Po...	Transcription	#	зд**(оу)**'рові кostру'бат'і	#			00:00:55.413	00:00:58.903	00:00:03.490
63	S_Jalynka2 Po...	Transcription	##	ми'шас'ті 'кон'і 'зоuс'ім п	##			00:01:08.227	00:01:11.417	00:00:03.190
64	S_Jalynka2 Po...	Transcription	##	ва'сил'ко заг**(оу)**р'нув 'руки	##			00:01:11.852	00:01:19.775	00:00:07.923
65	S_Jalynka2 Po...	Transcription	#	шчос' 'тр'існуло 'ў_сан'ах |	#			00:01:29.785	00:01:33.496	00:00:03.711
66	S_Jalynka2 Po...	Transcription	##	п**(оу)**'думаў в'ін |	##			00:01:56.533	00:02:00.741	00:00:04.208
67	S_Jalynka2 Po...	Transcription	##	'ў_л'іс'і п**(оу)**суте'н'іло ||	#			00:02:35.766	00:02:37.018	00:00:01.252

Рис. 2. Результати пошуку в кількох анотаційних файлах

Висновки та перспективи подальших досліджень у цьому напрямі. Корпус українського транскрибованого усного мовлення є першим українським мультимедійним корпусом усного українського мовлення, укладеним із використанням комп'ютерної програми ELAN. Застосування програми ELAN для роботи з медіафайлами Корпусу дає можливість розробити складні багаторівневі анотації аудіозаписів, зокрема створити спрощену фонетичну транскрипцію аудіотекстів на основі кириличної графіки і транскрипцію із застосуванням символів IPA. Комп'ютерна програма ELAN дає змогу здійснювати пошукові запити різних типів у межах анотаційних файлів КУТУМ (користувач отримує статистичну інформацію на базі КУТУМ, яку можна використати для ілюстрації фонетичних явищ українського мовлення). Матеріали мультимедійної бази даних КУТУМ можуть бути використані в таких сферах: 1) для наукових фонетичних досліджень українського літературного мовлення; 2) для навчальних матеріалів з української мови (як для носіїв мови, так й іноземців, які вивчають українську), зокрема під час створення мультимедійних орфоепічних словників. У майбутньому КУТУМ можна буде розширювати, додаючи нові аудіозаписи із залученням інших дикторів, оскільки в цьому проекті передбачена можливість поповнення Корпусу і створення відповідних анотаційних файлів у форматі *.eaf із використанням двох видів транскрипції – на основі кириличної й латинської графік (транскрипції засобами IPA).

ЛІТЕРАТУРА

1. Плахотнікова О. Ю. Асиміляційні процеси в українському мовленні (на матеріалі виступів депутатів Верховної Ради України) / О. Ю. Плахотнікова // Мовні і концептуальні картини світу. – 2011. – Вип. 37. – С. 191–195.

2. Плахотнікова О. Ю. Використання програми Elan в роботі зі звукозаписами корпусу українського усного мовлення / О. Ю. Плахотнікова // Українське мовознавство. – Харків, 2014. – Вип. 44. – Ч. 1. – С. 238–243.

3. Плахотникова Е. Ю. Особенности транскрибирования украинской устной речи в программе ELAN / Е. Ю. Плахотникова // Балтийский гуманитарный журнал. – 2015. – № 4 (13). – С. 40–43.

4. Програма навчальної практики для студентів ІІ курсу спеціальності «Прикладна лінгвістика та англійська мова» / [О. Зубань, З. Дудник, О. Бас-Кононенко, О. Плахотнікова]. – К. : ВЦ «Київський університет», 2016. – 31 с.

5. Рассказы о сновидениях и другие корпуса звучащей речи [Электронный ресурс]. – Режим доступа : http://www.spokencorpora.ru/.

6. Hellwig B., et al. ELAN – Linguistic Annotator. Version 5.0.0-beta [Електронний ресурс]/ B. Hellwig // The Language Archive, MPI for Psycholinguistics, Nijmegen, The Netherlands. – Режим доступу : http://www.mpi.nl/corpus/manuals/manual-elan.pdf.

7. Lausberg H. Coding gestural behavior with the NEUROGES-ELAN system / H. Lausberg, H. Sloetjes // Behavior Research Methods, Instruments, & Computers. – 2009. – № 41 (3). – P. 841–849.

8. Meyer Ch. F. English Corpus Linguistics: An Introduction / Ch. F. Charles. – Cambridge : Cambridge University Press, 2002. – 168 p.

9. The Language Archive. ELAN [Електронний ресурс]. – Режим доступу : http://tla.mpi.nl/tools/tla-tools/elan.