фінансової санації підприємства, другий – створення умов для розвитку високопродуктивного та ефективного індустріального парку; організація 100-150 нових підприємств на 100 тисячах квадратних метрів виробничих площ; досягнення ними через 3-4 роки щорічного рівня продажу у межах 150 мільйонів гривень. </s>.

Взагалі в процесі кодування корпусу на лексичному рівні рішення про пріоритети елементів TEI варто приймати окремо для кожного конкретного випадку відповідно до специфіки тексту та його стильового навантаження, оперуючи адаптованими до української мови прийомами тегування.

Література

1. Демська-Кульчицька О.М. Основи національного корпусу української мови. – К., 2005. – 219 с.
2. Корпусна лінгвістика / В. А. Широков, О. В. Бугаков, Т. О. Грязнухіна та ін.. – К.: Довіра, 2005. – 471 с.
3. Лозова Н.Є. Тегування власних назв організацій в Національному корпусі української мови //Лексикографічний бюлетень: Зб. наук. праць. – К, 2007. – Вип. 16. – С. 88–91.
4. Словник іншомовних слів. / За. ред. О.С. Мельничука. – К.: Гол. ред. УРЕ, 1977. – 776 с.
5. Сучасна українська літературна мова. / За ред. І. К. Білодіда. – Кн.: 5. Лексика і фразеологія. – К.: Наук. думка, 1973. – 439 с.
6. Сучасна українська літературна мова / За ред. М. Плющ. – К.: Вища шк., 2001.
7. Сучасна українська мова / За ред. О. Д. Пономарева, – К.: Либідь, 2001. – 400 с.
8. Тищенко О. М. Тегування дат в Національному корпусі української мови//Лексикографічний бюлетень: Зб. наук. праць. – К, 2007. – Вип. 16. – С. 82–87.
9. Українська мова. Енциклопедія. – К.: Укр. енциклопедія, 2000. – 750 с.
10. http://www.tei.org

КОРПУСИ ЖЕСТОВИХ МОВ ГЛУХИХ: СВІТОВИЙ ДОСВІД

© Оксана Тищенко, 2008
к. філол. н., Інститут української мови НАН України (Київ)
УДК 161.2.81'322.221.24+ 72'22

У статті здійснено огляд інтернет-ресурсів, присвячених корпусам жестових мов. Виявлено основні мотиви й цілі створення таких анотованих баз даних у США, Нідерландах, Німеччині, Японії та Греції, проаналізовано основні принципи добору й анотування відеоінформації.

Автоматичне розпізнавання й машинне статистичне опрацювання жестової мови глухих, створення відповідних систем оброблення даних –

надзвичайно важливі проблеми, які потрібно розв'язати передусім задля полегшення спілкування людей, що мають вади слуху. Для створення, тестування й удосконалення таких систем необхідна певна сукупність лінгвістичної інформації про жестову мову (ЖМ) глухих – корпус ЖМ.

Отже, актуальність створення корпусів ЖМ спричинена їхньою запитуваністю серед дослідників ЖМ в усьому світі, яких спонукає до цього проблема оптимізації навчального й комунікативного процесів серед глухих дітей і дорослих та їхніх неглухих родичів та вчителів. Так, Девід Горовітц (David Horowitz), співробітник Каліфорнійського університету (США), зазначив у 2002 р., що немає корпусу жодної жестової мови. Проблема, наголошував дослідник, полягає в нестандартності формату транскрибування, анотування, зберігання даних мови глухих [5]. На сьогодні ситуація в цій галузі значно змінилася – світовий досвід свідчить про розроблення теоретичних і технічних передумов для створення таких сукупностей жестових мов й функціонування їх в режимі онлайн.

У *Каліфорнійському університеті* створено систему розмічування жестової мови на морфологічному рівні й триває робота з даними американської та нідерландської жестових мов. Вірогідно, що створений корпус у майбутньому буде впроваджено до міжнародного архіву CHILDES – бази даних дитячої мови. Питання створення корпусу ЖМ, зокрема методів анотування жестів, вивчає також Урсула Беллуджі (Ursala Bellugi) (*Інститут Салка* в Каліфорнії, США) [5].

У *Бостонському університеті* група з вивчення лінгвістики американської жестової мови (American Sign Language (ASL)) створила частково доступний в онлайні **корпус американської жестової мови**. Усі відеофайли, отримані внаслідок тривимірного знімання, анотовані передусім з погляду лінгвістики ЖМ, а не її співвіднесеності з вербальною мовою.

Активно розробляють аспекти створення корпусу жестової мови в Нідерландах. Задекларовано роботу в онлайні **Корпусу нідерландської жестової мови (Corpus NGT)** – результату роботи відділу лінгвістики *Університету Рейдбод Неймеген* у 2006–2008 рр., зокрема Онно Красборна (Onno Crasborn), Елс ван дер Коій (Els van der Kooij), Йохана Роса (Johan Ros), Інге Цвітсерлод (Inge Zwitserlood). На час написання цієї статті сторінка *корпус* [The corpus] ще не функціонувала. Доступними для ознайомлення є сторінки *вступ* (загальні відомості про створення корпусу), *методологія* (теоретичні засади добору відеоматеріалу), *технологія* (відомості про інструментарій: технічні характеристики відеокамер, програмне забезпечення зберігання й анотування даних, напр., програм IMDI Editor, IMDI Browser, LAMUS, ELAN, Download & Demo), *контакти*. Як зазначено у вступі,

Корпус становить анотована база даних – відеоматеріали, отримані внаслідок синхронізованої роботи відеокамер [8].

Лінгвістично анотовані дані трьох жестових мов – нідерландської (Sign Language of the Netherlands (NGT)), британської (British Sign Language (BSL)) та шведської (Swedish Sign Language (SSL)) – становлять основу відкритого, динамічного **Корпусу Європейського культурного спадку онлайн (ECHO)**, який діє з 2004 р. в мережі Інтернет. У його створенні взяли участь *Стокгольмський університет, Міський лондонський університет* та *Університет Рейдбод Неймеген* [8]. Відеоматеріали фіксують жестові тексти близько п'яти різних оповідань кожною з мов, інтерв'ю носіїв жестової мови, також додано жестовомовну поезію британською та нідерландською мовами. Насамкінець до корпусу ECHO введено два анотовані фрагменти корпусу німецької жестової мови (German Sign Language (DGS)) (корпусу, створеного Йенсом Хессманом (Jens Heßmann) у 2001 р., про який ітиметься далі). Відеофайли супроводжено транскрипцією (ELAN-файли) та метаданими (IMDI-файли). Загалом цей тримовний корпус за необхідного програмного забезпечення можна використовувати для перекладу вербальних текстів на жестову мову.Керує цим проектом Онно Красборн (Onno Crasborn).

За даними нідерландського *Інституту психолінгвістики Макса Планка*, лінгвістична інформація про ЖМ є частиною доступного в Інтернеті **Корпусу жестових мов світу**, над яким у 2001 р. почала працювати група з дослідження типології жестової мови. Усього заплановано дослідити 37 жестових мов світу і впровадити цю інформацію до Sign Language Typology Database – **Бази даних типології жестової мови**. Пілотний проект корпусу (анотованої інформаційної бази) підготовлено до 2006 р.

Відомості на офіційній веб-сторінці інституту свідчать, що жестові мови в цьому корпусі представлені близько десятьма годинами відеозаписів, організованих у менші відеочастини – файли. Кожен із файлів супроводжують метадані: інформація про місце, час запису, про мовця (комуніканта), тему розмови, а також про технічні дані відеодокумента. Деякі відеодані споряджені транскрипцією (частково або повністю), анотування здійснено за допомогою розробленої в Інституті програми. Нелінгвістична інформація про кожну ЖМ організована як ієрархічна послідовність: країна, область та ін.

До аналізованого корпусу входять корпуси жестових мов таких країн, як Китай (корпус, опис письмовою китайською мовою, опис анотування зразків у цифровому оформленні), Індія (корпус, еквівалентні конструкції), Індонезія (жестова мова Балі), Росія (корпус, сегментальна, внутрішньосегментальне й надсегментальне морфологічне анотування), Південна Корея (корпус, система

гендерного маркування, лінгвістичні особливості поетичної мови жестів), Туреччина (корпус) [9].

Німецька ЖМ (German Sign Language (GSL) представлена в **Корпусі Йенса Хессмана** (Jens Heßmann (2001)), а також у German Sign Language Corpus of the Domain Weather Report «Phoenix» – **Німецькому корпусі повідомлень про погоду жестовою та вербальними мовами «Фенікс».**

Корпус Jens Heßmann оснований на даних інтерв'ю жестовою мовою, має кілька тисяч речень, як ми згадували, частково він доступний на веб-сайті ECHO. Однак мотивом створення ще одного жестовомовного корпусу «Фенікс» став суто прикладний аспект: удосконалення автоматичного розпізнавання ЖМ та перекладу вербальної мови жестовою і навпаки. За свідченням авторів «Фенікса» (Яна Бунгерота (Jan Bungeroth), Даніеля Штайна (Daniel Stein), Філіппа Дройва (Philippe Dreuw), Мортеца Цаеді (Morteza Zahedi), Германа Нея (Hermann Ney)) [6], співробітників відділу інформатики в *Ахенському університеті,* корпус Jens Heßmann для цього має надто широкий домен і заскладний для автоматичного вивчення, що спричинено ще й наявністю великої кількості імпліцитних змістів у текстових зразках ЖМ.

Натомість повідомлення про погоду жестовою мовою забезпечують: обмеженість предметної області корпусу, сталість синтаксичних структур, відносну сталість лексичного складу, майже відсутнє використання міміки. Усе це сприяє досягненню основної мети створення корпусу з обмеженим доменом «Фенікс» – відпрацювання систем автоматичного розпізнавання ЖМ.

Відповідно корпус «Фенікс» анотований передусім з погляду його співвіднесеності з вербальною мовою, на перший план висунуто gloss notation, словесний запис, коли жестовий вислів передано словесними еквівалентами жестів. Корпус складається з трьох частин: відеофайли повідомлень жестовою мовою, двомовна частина (жестовою та німецькою вербальною), де жестова інформація (у формі gloss notation – словесного запису) супроводжуються вербальною інформацією – переклад німецькою вербальною, та одномовна частина (усі тексти повідомлень німецькою вербальною) [6].

Корпус грецької жестової мови (Greek Sign Language Corpus for HCI) – проект *Інституту оброблення мови й мовлення* в Афінах – призначений для створення методології анотування даних і вироблення на цій основі адекватної лінгвістичної моделі для людино-машинного інтерфейсу. Корпус має три основні частини: список анотацій (лем) стосовно мануальних та немануальних жестів; сукупність спеціально розроблених жестових висловів,

які повною мірою репрезентують особливості структурно-семантичної системи мови; тексти, які відбивають специфіку безпосереднього вільного спілкування жестовою мовою.

Анотування здійснено на двох рівнях: фонологічному (параметри жестів) та граматичному, на рівні речення (межі речення, межі фрази тощо) і може розширюватися згодом на всіх рівнях [4].

Японська жестова мова репрезентована в **Анотованому корпусі японської жестової мови** (Annotated Japanese Sign Language Corpus), розробленому в *Центральній науково-дослідній лабороторії* в Хітачі такими вченими, як Атсуко Коїзамі (Atsuko Koizumi), Хірохіко Сагоа (Hirohiko Sagawa) та Мазеру Такеучі (Masaru Takeuchi). Мотивом створення цього корпусу, як і в попередніх випадках, є розроблення й тестування перекладних систем ЖМ. Мета авторів корпусу японської ЖМ – з'ясувати основні правила граматики ЖМ.

Корпус становлять 2800 анотованих висловів жестовою мовою. Основними вимогами до добору мовного матеріалу були такі: наявність протиставлення в реченнях, різні форми висловлення тієї самої інформації, повторення тієї самої інформації різними комунікантами – носіями ЖМ, повторення інформації тим самим комунікантом. В основі інтерв'ю 360 японських речень, які охоплюють основні граматичні явища.

Докладно розроблена методика процесу відеознімання, перевірки, повторного знімання тощо дає підстави стверджувати достовірність результатів. Дані зібрано у два етапи: ретрансляція готових граматичних конструкцій жестовою мовою (360 конструкцій і 2500 їх відповідників) та вільний опис запропонованої ситуації (300 висловів).

Перший етап лінгвістичного анотування – транскрибування мануальних жестів, що мають лексичне значення (методом наведення словесних еквівалентів, глос), другий етап – опис немануальних жестів (міміки, рухів голови, тулуба та ін.), що становлять граматичний компонент жестового вислову.

Досконалим є інструментарій корпусу. Так, користувач може отримати таку інформацію: синхронізоване тривимірне відеозображення (з можливістю спостереження в чотирьох зонах), анімоване зображення ЖМ (на основі введення даних від спеціального пристрою рукавички), переклад японською вербальною мовою, масштаб часу (перетягуванням курсору), сегментація відеоінформації, анотації мануальних та немануальних жестів. Під час роботи з корпусом користувач може отримувати потрібну інформацію за заданими параметрами, порівнювати необхідні дані тощо.

Самі автори, як і їхні колеги в інших країнах, у процесі роботи з корпусом досліджують лінгвістику ЖМ та роблять відповідні теоретичні висновки [7].

Отже, аналіз наявних у мережі Інтернет даних про корпуси жестових мов у різних країнах засвідчує, що мотивом побудови таких сукупностей анотованих даних передусім є розроблення електронних перекладних систем. Конкретні цілі: вивчення насамперед лінгвістичних особливостей ЖМ (внутрішньої та зовнішньої форми її лексичних та граматичних засобів), а також стилістичних, соціальних тощо характеристик; створення універсальної мови моделювання й розпізнавання жестових текстів; розроблення адекватного інструментарію для опису жестовомовної структури; удосконалення методик збору й аналізу мовної інформації. Зазвичай дослідження в цій галузі підтримують на державному рівні.

Наявні у світовій практиці корпуси ЖМ мають певні відмінності в плані структури корпусу, обсягу й характеру інформації, її добору та способу анотування (у теоретичних засадах й технічній реалізації). Корпуси можуть бути одномовні (на основі даних однієї національної жестової мови) та багатомовні, як, наприклад, ECHO.

Однак принцип побудови корпусів та їх репрезентації в мережі Інтернет засвідчує єдність підходів учених у цій галузі корпусних досліджень: корпус жестової мови – це анотована сукупність відеоданих (бажано синхронізованих тривимірних) жестової мови. Інформація може бути, з одного боку, результатом штучного моделювання висловів (як граматично репрезентативна), з іншого – вільним текстом живого спілкування комунікантів (природних носіїв жестової мови). Суть анотування жестових текстів полягає в а) записі глосами – вербальними еквівалентами; б) маркуванні на фонологічному та синтаксичному рівнях; в) перекладі вербальною мовою. Зазвичай корпуси відкриті й динамічні, тобто змінюватися може як обсяг даних, так і характер анотацій. Структура корпусу обов'язково передбачає опис методологічних та технологічних параметрів.

У перспективі плануємо зробити докладніший аналіз корпусної термінології в галузі жестової мови та принципів анотування відеоданих.

Література

1. Crasborn O., Els van der Kooij, Daan Broeder & Hennie Brugman. Sharing sign language corpora online: proposals for transcription and metadata categories. In: Proceedings of the LREC 2004 Satellite Workshop on Representation and processing of sign languages, Oliver Streiter & Chiara Vettori, eds. – 2004. – Pp. 20–23.
2. Crasborn O., Han Sloetjes, Eric Auer & Peter Wittenburg. Combining video and numeric data in the analysis of sign languages within the ELAN annotation software. In Chiara Vettori (ed.) Proceedings of the 2nd workshop on the

representation and processing of sign languages: lexicographic matters and didactic scenarios. Paris: ELRA. – 2006. – Pp. 82–87.
3. Crasborn O. & Maya de Wit. Ethical implications of language standardisation for sign language interpreters. In: J. Mole (ed.) International perspectives on interpreting. Selected proceedings from the Supporting Deaf People online conferences 2001–2005. – 2005. – Pp. 112–119.
4. http://dianoema.zenon.gr/documents/publications/Abstract_submitted_hcii2007
5. http://torvald.aksis.uib.no/corpora/2002-2/0215.htm
6. http://www-i6.informatik.rwth-aachen.de/~bungeroth/
7. http://www.irit.fr/ACTIVITES/EQ_TCI/EQUIPE/dalle/cognitique/Documentation/CorpusJapAnnote-318.pdf
8. http://www.let.ru.nl/corpusngt/scientific/index.html
9. http://www.mpi.nl/world/SignLang/WEB-FINAL/sl-world.htm