POLSKA AKADEMIA NAUK
INSTYTUT JĘZYKA POLSKIEGO

POLONICA

TOM XXXIII

KRAKÓW 2013

Spis treści

POLONICA XXXIII
PL ISSN 0137-9712

PAWEŁ RUTKOWSKI, SYLWIA ŁOZIŃSKA, JOANNA FILIPCZAK,
JOANNA ŁACHETA, PIOTR MOSTOWSKI

Jak powstaje korpus polskiego języka migowego (PJM)?

1. Wstęp

Niniejszy tekst poświęcony jest pierwszemu w historii korpusowi polskiego języka migowego (PJM), który od roku 2010 powstaje w Pracowni Lingwistyki Migowej Uniwersytetu Warszawskiego (PLM UW — zob. http://www.plm.uw.edu.pl)[1]. PJM to niezależny od polszczyzny pod względem leksykalnym i gramatycznym język naturalny, którym posługuje się od blisko 200 lat kilkudziesięciotysięczna społeczność polskich Głuchych[2]. Artykuł prezentuje trudności związane z tworzeniem korpusów języków wizualno-przestrzennych, szczegółowe rozwiązania techniczne i metodologiczne stosowane w projekcie realizowanym przez PLM UW oraz możliwe zastosowania zgromadzonych danych.

[1] Różne aspekty prac nad koncepcją i tworzeniem korpusu były w tym czasie finansowane ze środków Fundacji na rzecz Nauki Polskiej (projekty o numerach 1/2009 i F1/09/P/2013 w ramach programu FOCUS), Fundacji Współpracy Polsko-Niemieckiej (*Stiftung für deutsch-polnische Zusammenarbeit* — nr grantu: 13232/10/IF) oraz Narodowego Centrum Nauki (projekt o numerze 2011/01/M/HS2/03661 w ramach programu HARMONIA). Pragniemy tu również odnotować ważne wsparcie merytoryczne, które PLM UW otrzymała od Instytutu Niemieckiego Języka Migowego i Komunikacji Niesłyszących (IDGS, *Institut für Deutsche Gebärdensprache und Kommunikation Gehörloser*) z Uniwersytetu w Hamburgu. Dzięki życzliwości kierującego nim prof. Christiana Rathmanna wiele rozwiązań technicznych wypracowanych przez zespół IDGS zostało wykorzystanych także w projekcie korpusu PJM.

[2] W przypadku nie medycznego, a językowo-kulturowego podejścia do głuchoty (opierającego się na dostrzeżeniu w niej przede wszystkim fundamentu tożsamości społecznej osób od najwcześniejszego dzieciństwa posługujących się językiem migowym jako pierwszym i podstawowym narzędziem komunikacji) określenie *Głuchy* zapisujemy wielką literą, analogicznie do pisowni *Kaszub* czy *Ślązak*. Ta konwencja ortograficzna rozpowszechniła się w ostatnich dekadach w wielu krajach świata, por. np. angielskie rozróżnienie *Deaf* (o członku społeczności Głuchych) — *deaf* (o osobie lub zwierzęciu z uszkodzeniem słuchu).

2. Metodologia korpusowa w lingwistyce migowej

Lingwistyka migowa to jedna z młodszych dziedzin językoznawstwa — rozwija się dopiero od lat 50. ubiegłego wieku. Jest to także dyscyplina niezwykle dynamicznie zyskująca na znaczeniu — obecnie badania w zakresie komunikacji Głuchych są prowadzone w ponad 100 ośrodkach na całym świecie (zob. http://lrwiki.ldc.upenn.edu, dostęp 14.10.2013). Niewątpliwie rozwój badań nad językami wizualno-przestrzennymi jest powiązany z rewolucją informatyczną ostatnich dekad i rozwojem lingwistyki korpusowej. Wykazanie owej zależności musi być poprzedzone krótkim wprowadzeniem dotyczącym natury języków migowych. Należy przede wszystkim podkreślić, że są to w pełni rozwinięte języki naturalne, które od języków fonicznych różnią się przede wszystkim kanałem przekazu i odbioru (wizualno-przestrzennym w miejsce wokalno-audytywnego).

Modalność ta sprawia z kolei, że w językach migowych znajdziemy cechy nieobecne w językach fonicznych. Wielość dostępnych artykulatorów (w produkcji znaku migowego biorą udział nie tylko dłonie, ale także twarz, głowa, tułów czy inne części ciała, np. przedramiona lub — w wyjątkowych sytuacjach — nogi) przekłada się na możliwość przekazu symultanicznego, czyli artykulacji dwóch lub nawet kilku sygnałów w tym samym czasie. Gramatyczną funkcję zyskuje ekspresja mimiczna oraz przestrzenność i dynamiczność znaku. Istotą komunikacji migowej jest realizowany w trójwymiarowej przestrzeni ruch i z tego względu niemożliwy jest zapis dyskursu migowego w postaci ściśle linearnej (charakterystycznej dla zapisów języków fonicznych). Języki wizualno-przestrzenne nie posiadają zatem pisma w tradycyjnym rozumieniu, choć istnieją ich notacje tworzone zarówno do celów czysto akademickich (m.in. HamNoSys, tj. Hamburg Sign Language Notation System — wypracowany w Hamburgu system transkrypcji fonetycznej znaków, zob. Hanke 2004), jak i wykorzystywane w edukacji Głuchych (metoda SignWriting, zob. www.signwriting.org). Pełne oddanie właściwości komunikatu migowego, a tym samym sporządzenie jego dokładnego opisu, jest jednak możliwe wyłącznie poprzez rejestrację znaków w postaci nagrań wideo. Wspieranie analizy zebranych w ten sposób danych przez wykorzystanie komputerów prowadzi zaś w prosty sposób do narzędzi lingwistyki korpusowej. Wśród najważniejszych atrybutów tej metodologii wymienia się m.in.:

- podejście empiryczne, ukierunkowanie na badanie autentycznych i naturalnych tekstów języka,
- wykorzystywanie do analizy danych korpusowych,
- wykorzystanie analizy ilościowej i jakościowej (por. Biber, Conrad, Reppen 1998).

Zdefiniowane w ten sposób podejście naukowe wydaje się najtrafniejszym wyborem w wypadku badań nad komunikacją Głuchych. Analizowanie naturalnych i autentycznych tekstów jest szczególnie ważne dla zachowania rzetelności i obiektywizmu prac badawczych. Języki foniczne są najczęściej opisywane przez swoich rodzimych użytkowników.

Języki migowe są zaś nadal badane przede wszystkim przez osoby słyszące (będące nienatywnymi użytkownikami), które nie mogą opierać się wyłącznie na własnej intuicji językowej. Języki te są także bardzo młode (najstarsze z nich, jak PJM, powstały nie więcej niż kilkaset lat temu, zob. Hollak, Jagodziński 1879) i nie w pełni zgramatykalizowane, niejednokrotnie także silnie zróżnicowane regionalnie czy socjologicznie (zob. Lucas 2003). Z tego względu ich analiza oparta wyłącznie na wiedzy samego badacza będzie niepełna i subiektywna. Jedynie dane korpusowe pozwolą na uzyskanie wiarygodnych wyników. Wiarygodność ta jest także gwarantowana przez ścisłą zależność badań od analiz jakościowych i ilościowych. Dzięki nim możliwe jest np. określenie, które zjawiska są dla danego języka najczęstsze czy najważniejsze.

Powyższe uwarunkowania sprawiają, że metodologia korpusowa jest obecnie stosowana w analizach wielu języków migowych. W najbardziej znaczących i największych ośrodkach badań nad komunikacją Głuchych powstają korpusy wykorzystywane później w celu szczegółowego opisu gramatyki danego języka czy opracowania słownika migowo-fonicznego. Obecnie na całym świecie istnieje 35 korpusów osiemnastu języków migowych (niemieckiego, szwajcarskiego, austriackiego, holenderskiego, francuskiego, hiszpańskiego, brytyjskiego, duńskiego, szwedzkiego, irlandzkiego, islandzkiego, amerykańskiego, australijskiego, nowozelandzkiego, koreańskiego, malijskiego, kata kolok — jednego z języków indonezyjskich, oraz polskiego, zob. Konrad 2008, Barberà 2012: 7–9). Jednocześnie należy podkreślić, że wciąż ograniczona liczba korpusów migowych wynika z faktu, iż zbieranie danych tego typu jest procedurą niezwykle czasochłonną i kosztowną. Trudności te można porównać z problemami, jakie wiążą się ze zbieraniem korpusów języka mówionego. Warto zaznaczyć, że dane mówione to zazwyczaj jedynie niewielka część korpusów języków fonicznych. W Narodowym Korpusie Języka Polskiego dane mówione stanowią około 10 procent całych zasobów (30 mln słów), jednak jedynie 1 900 000 (nieco ponad 6 procent) z nich to dane mówione konwersacyjne, czyli zebrane na potrzeby NKJP transkrypcje rozmów. Resztę stanowią dane medialne (transkrypcje audycji radiowych i programów telewizyjnych, około 3 procent) oraz tzw. mówione inne, tj. głównie stenogramy posiedzeń Sejmu RP i sejmowych komisji śledczych (blisko 91 procent, Pęzik 2012: 39). Twórcy NKJP uzasadniają, że proporcje te są efektem kosztowności i pracochłonności transkrypcji oraz anotacji danych mówionych, a podobny stosunek danych z języka mówionego i pisanego można znaleźć w innych korpusach narodowych. Korpusy języka migowego w całości składają się z danych „mówionych" (niepisanych), zatem z założenia są projektami wieloletnimi i kosztownymi. Jako przykład może posłużyć korpus niemieckiego języka migowego (DGS), którego czas realizacji jest planowany na lata 2009–2023 i który do tej pory otrzymał fnansowanie w kwocie 8,5 mln euro (http://www.sign-lang.uni-hamburg.de). Względy te sprawiają, że jedynie niewielka część korpusów migowych to zbiory dokumentujące zapis języka całej społeczności (tzw. korpusy ogólne, obecnie 9 z 35 istniejących korpusów migowych). Znacznie częściej pojawiają się korpusy skupiające się na wybranych aspektach języka (np. klasyfikatorach czy uzgodnieniach), rejestracji wariantów socjolingwistycznych i dialektów bądź konkretnej dziedzinie

słownictwa (np. słownictwie technicznym, medycznym, prawnym, religijnym czy psychologicznym), których kolekcjonowanie jest mniej czasochłonne i tańsze.

3. Korpus PJM

3.1. Informacje wstępne

Prace nad gromadzeniem tekstów polskiego języka migowego są prowadzone od około 3 lat. W tym relatywnie krótkim czasie korpus PLM stał się największym zaanotowanym korpusem migowym świata (6794 leksemy, 202 208 jednostek tekstowych, stan na 14.10.2013). Jest to korpus jednojęzyczny, ogólny, całościowy (prezentuje zebrane teksty w całości, a nie ich próbki o określonej długości), synchroniczny (obrazuje aktualny stan języka, nie zaś teksty historyczne czy zróżnicowane czasowo), anotowany i uzupełniony o metadane.

3.2. Sesja nagraniowa i jej uczestnicy

Wszystkie dane wchodzące w skład korpusu PJM powstają specjalnie na potrzeby tego projektu. Zespół PLM UW nie korzysta z żadnych wcześniej zgromadzonych tekstów języka migowego. Dane pozyskiwane są podczas sesji nagraniowych organizowanych najczęściej w siedzibie PLM. Do udziału w nagraniu zapraszani są rodzimi użytkownicy PJM powyżej 18. roku życia, urodzeni i mieszkający na terenie Polski. Kandydaci na informatorów wypełniają ankietę zawierającą pytania dotyczące m.in. momentu nabycia języka migowego, głuchoty w rodzinie, edukacji itp. Dzięki tym informacjom możliwe jest zaangażowanie do badań wyłącznie osób, dla których PJM jest pierwszym i podstawowym narzędziem komunikacji od wczesnego dzieciństwa.

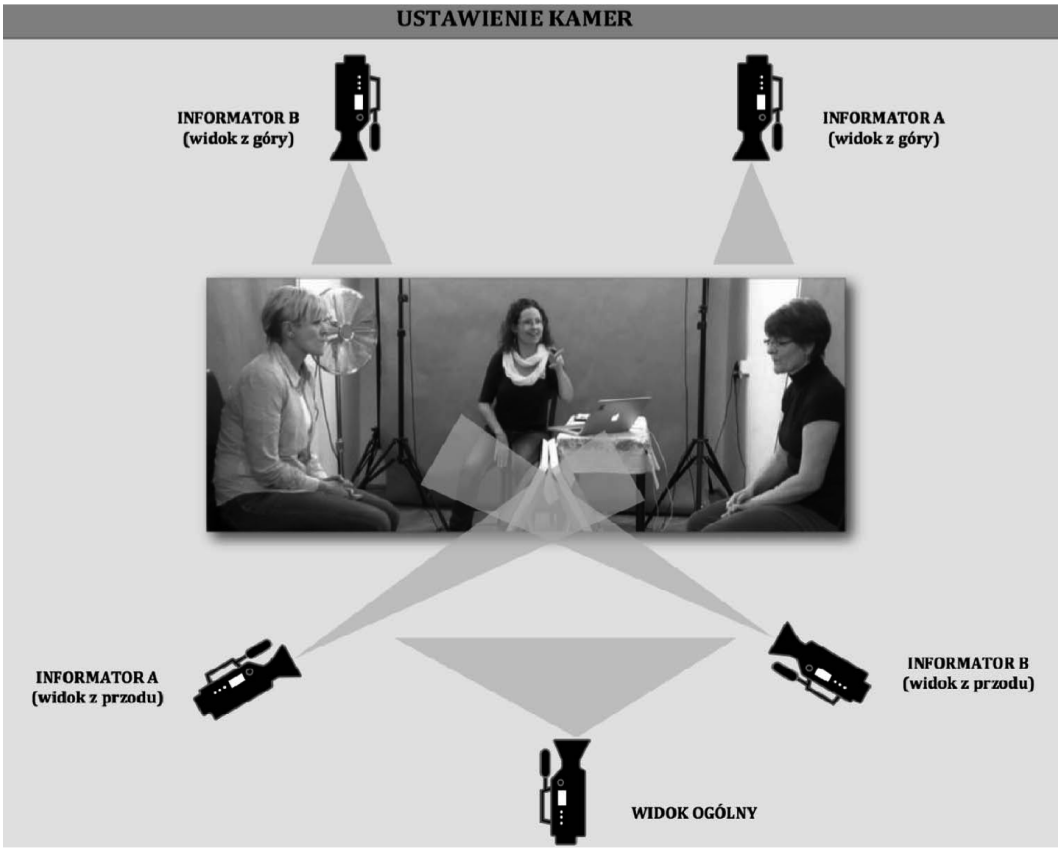
Ankieta pozwala także kontrolować, czy w korpusie zostają zachowane proporcje wieku, płci, miejsca zamieszkania czy wykształcenia informatorów, tak by w żadnej z tych kategorii nie istniało znaczące niedoszacowanie czy nadmiar. Zgodnie z założeniami w nagraniach do korpusu PJM ma bowiem wziąć udział równa liczba kobiet i mężczyzn, reprezentujących różne grupy wiekowe, o zróżnicowanym wykształceniu i pochodzeniu. Zalecenie to wynika z dążenia do reprezentatywności korpusu, tj. tego, by zebrany materiał językowy odzwierciedlał rzeczywisty stan języka dla całej społeczności, która się nim posługuje (McEnery, Wilson 1996: 30). Jak dotąd w nagraniach korpusowych wzięły udział osoby w wieku od 18 do ponad 90 lat, przy czym najmniej licznie reprezentowane są osoby w kategoriach wiekowych 18–30 lat, 60–70 lat oraz powyżej 70. roku życia (łącznie około 20 procent badanych). Wśród informatorów znajdują się osoby z wykształceniem od podstawowego (3 procent) poprzez zawodowe (61 procent, w tym w osoby w trakcie edukacji na tym poziomie), średnie (30 procent, w tym w osoby w trakcie edukacji na tym poziomie), wyższe (4,5 procent) aż po wyższe ze stopniem doktora (1,5 procent). Badani pracują intelektualnie, fizycznie, pozostają na rencie lub emeryturze bądź są bezrobotni. Pochodzą z rodzin, w których oprócz nich nie ma osób głuchych (24 procent), jest jedna

osoba niesłysząca (24 procent), dwie (17 procent), trzy (13 procent), cztery (6 procent) lub więcej. Jedynie 1,5 procent badanych ma w rodzinie 6 i więcej osób głuchych. 72 procent informatorów opanowało język migowy do 7. roku życia (czyli nie później niż w momencie rozpoczęcia edukacji), pozostali — w następnych kilku latach. Badani pochodzą z ponad 20 miejscowości w Polsce — małych i dużych miast oraz wsi.

Obecnie dokonano nagrań 80 osób, planowane są jednak dalsze sesje w kolejnych latach. W nagraniach do korpusów ogólnych bierze zawsze udział co najmniej kilkudziesięciu informatorów (od 40 w korpusie irlandzkim do 330 w korpusie niemieckim, Konrad 2008). W korpusach porównawczych, specjalistycznych itp. liczba informatorów waha się najczęściej od kilku do kilkudziesięciu osób.

Istotną kwestią podczas doboru osób do nagrań jest oddzielenie faktycznych użytkowników PJM od głuchych posługujących się językiem miganym (tzw. systemem językowo-migowym, dalej SJM), czyli wizualno-przestrzennym subkodem polszczyzny opracowanym w latach 60. ubiegłego wieku (por. np. Fabisiak 2010). SJM, system sztuczny, będący połączeniem znaków języka migowego i polskiej gramatyki, nie jest przedmiotem zainteresowania lingwistyki migowej. Jako narzędzie komunikacji używane m.in. w szkołach dla niesłyszących i w większości tłumaczeń telewizyjnych jest jednak wśród głuchych bardzo rozpowszechniony. Niejednokrotnie zdarza się także, że sami niesłyszący nie mają świadomości, czy posługują się językiem miganym czy migowym. Zagadnienie to wymaga zatem zwiększonej uwagi u osób zajmujących się doborem informatorów.

Sesja nagraniowa obejmuje zawsze parę informatorów. Dzięki temu możliwe jest uzyskanie naturalnych dialogów, a tym samym także różnorodnych konstrukcji składniowych — nie tylko samych zdań twierdzących, ale także pytań, rozkaźników, elementów o funkcji fatycznej. Dialogowy charakter sesji pomaga również jej uczestnikom szybciej przyzwyczaić się do obecności kamer i innych części wyposażenia studia nagraniowego — osoby nagrywane pojedynczo czują się bowiem zazwyczaj skrępowane koniecznością migania do obiektywu. Nagrania do korpusu PJM wykonywane są z 5 kamer jednocześnie, z trzech różnych perspektyw (zob. ryc. 1), zatem oswojenie się z ich obecnością jest szczególnie trudne. Każda z perspektyw ma jednak swoje uzasadnienie badawcze — rzut frontalny to oczywiście podstawa do rozpoznawania i analizowania znaków użytych przez uczestników nagrania, ale rzut z góry („z lotu ptaka") jest niezwykle użyteczny przy określaniu dystansu między dłońmi a ciałem osoby migającej (dystans ten może odgrywać istotną rolę gramatyczną). Poza pozytywnym wpływem na eliminowanie stresu i barier komunikacyjnych związanych z wykorzystaniem tak wielu kamer obecność drugiego informatora pomaga również przezwyciężyć znużenie sesją, której czas trwania wynosi zazwyczaj około 5–6 godzin.



Ryc. 1. Organizacja studia nagraniowego

W studiu nagraniowym oprócz informatorów obecny jest także głuchy moderator sesji, który czuwa nad jej przebiegiem i wyjaśnia niezrozumiałe kwestie, nie ingeruje jednak w wypowiedzi samych nagrywanych (choć może np. zachęcać ich do rozszerzenia czy uzupełnienia poszczególnych opinii). Oprócz wymienionej trójki w studiu znajduje się wyłącznie technik obsługujący kamery (także niesłyszący). W sesji nie biorą zatem udziału żadne osoby słyszące, które mogłyby swoją obecnością wpłynąć na sposób migania informatorów. Językiem kontaktowym jest jedynie PJM.

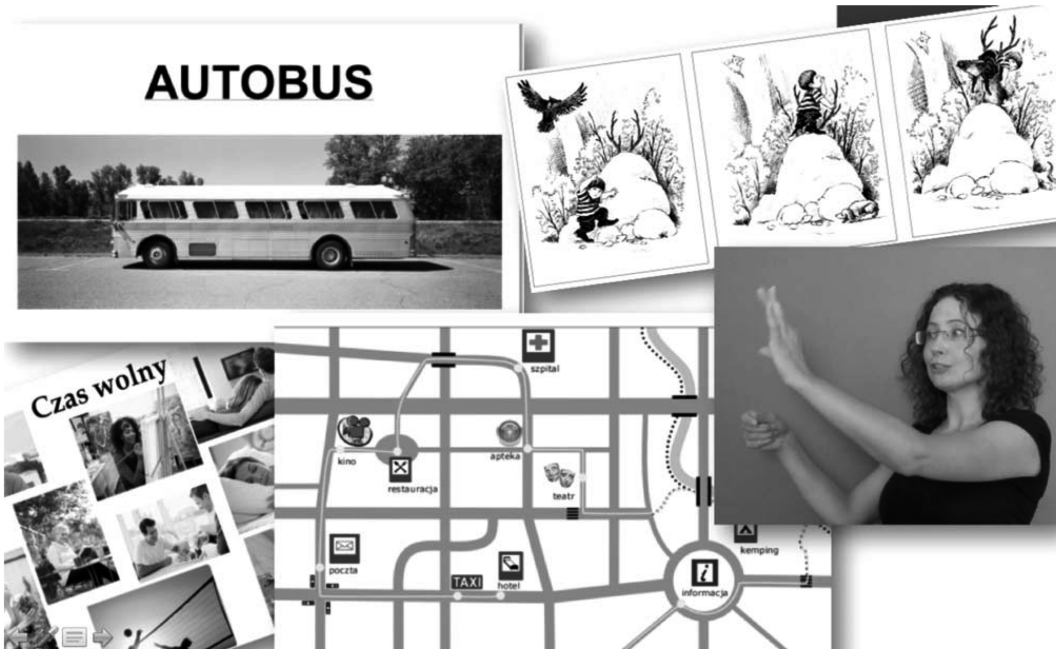
Przed rozpoczęciem sesji nagraniowej moderator prezentuje informatorom prawne aspekty udziału w nagraniu, m.in. związane z koniecznością wyrażenia przez nich zgody na utrwalanie i rozpowszechnianie wizerunku. W przypadku nagrań do korpusów migowych kwestie etyczne nabierają szczególnego znaczenia, języki wizualno-przestrzenne nie pozwalają bowiem na prezentację anonimowych danych — konkretny komunikat językowy będzie zawsze ściśle związany z wizerunkiem jego autora. Sytuacja ta wymaga zatem, aby uczestnicy nagrania mieli pełną świadomość wagi podejmowanej decyzji (por. m.in. Crasborn 2002).

3.3. Elicytacja danych

Gromadzenie danych do korpusów migowych musi zawsze uwzględniać dwa podstawowe wymogi — naturalności i różnorodności (leksykalnej i gramatycznej) zbieranego materiału. Z tego względu niewskazane jest z jednej strony pozostawienie informatorom pełnej swobody wypowiedzi, z drugiej — zbytnie ingerowanie w dialog między badanymi czy przedstawianie im gotowych polskich tekstów do przetłumaczenia na język migowy (tu dodatkowo pojawia się ryzyko interferencji językowej i wprowadzania do komunikacji języka miganego). Optymalnym rozwiązaniem tego problemu jest zastosowanie techniki elicytacji danych, tj. szeregu różnorodnych zadań zakładających pewną wolność wypowiedzi badanego i jednocześnie pomagających pozyskać pożądane konstrukcje i słownictwo. Elicytacja jest obecnie najczęściej stosowaną metodą zbierania danych migowych.

Na procedurę elicytacyjną w korpusie PJM składa się ponad dwadzieścia różnorodnych zadań wymagających od informatorów m.in.:

- relacjonowania obejrzanych historyjek obrazkowych i filmów (m.in. kreskówek „Tweety i Sylwester" oraz „Baranek Shaun", filmów z udziałem Charliego Chaplina, klipów filmowych z ruchem, historyjek rysunkowych dołączanych do gumy do żucia „Kaczor Donald" i innych);
- przedyskutowania wybranych tematów ważnych dla społeczności Głuchych, tematów związanych z historią najnowszą, rozwojem języka migowego itd.;
- swobodnej wypowiedzi z udziałem moderatora (np. zaprezentowania dowcipu, opowiedzenia o swoim doświadczeniu głuchoty, edukacji, rodzinie, miejscu pochodzenia, zaangażowaniu w życie lokalnej społeczności niesłyszących) bądź bez niego (podczas jednego z zadań moderator opuszcza studio nagraniowe, informując uprzednio badanych, że kamery pozostają włączone);
- udzielenia instrukcji dotyczących np. sposobu wykonania jakiejś czynności (kupna biletu lotniczego przez Internet, farbowania włosów, naprawy pralki) czy sposobu dojścia z punktu A do punktu B w przestrzeni miejskiej zilustrowanej mapą;
- wspólnych ustaleń, np. znalezienia terminu na spotkanie, odczytania znaczenia nietypowego znaku ostrzegawczego;
- zaprezentowania sposobu migania danego znaku (zadanie pozwalające uchwycić zróżnicowanie regionalne PJM).



Ryc. 2. Przykłady materiałów elicytacyjnych wykorzystywanych w korpusie PJM

Tak opracowane zadania elicytacyjne pozwalają na uzyskanie różnorodnego słownictwa z najważniejszych dziedzin życia, bogatych konstrukcji składniowych (twierdzeń, negacji, pytań, rozkaźników itp.) oraz informacji o gramatyce PJM (m.in. poprzez wprowadzenie zadań, w których konieczne jest gramatyczne wykorzystanie przestrzeni i ruchu).

Znaczna część procedury elicytacyjnej bazuje na metodzie wykorzystywanej podczas tworzenia korpusu niemieckiego języka migowego (Hanke, König, Wagner, Matthes 2010) oraz innych korpusów migowych, co w przyszłości umożliwi np. prowadzenie badań porównawczych nad różnymi językami migowymi.

Każde z zadań informatorzy wykonują w parze, relacjonując sobie obejrzane materiały, dyskutując na wybrane tematy, dokonując wspólnych ustaleń itp. Zadania zakładają aktywność obu informatorów jednocześnie (np. w czasie tematów dyskusyjnych) bądź naprzemiennie większą aktywność i bierność, w sytuacji gdy jeden z informatorów zapoznaje się z materiałem wywoławczym (np. filmem) i relacjonuje go drugiemu, który może prosić o uściślenia, dopytywać itp. Materiałami wywoławczymi są rysunki, klipy filmowe, plansze z polskimi słowami i odpowiadającymi im obrazkami, mapki. Z zasady w zadaniach unika się pośrednictwa języka polskiego. Gdy to konieczne, pojawia się on w materiałach wywoławczych najczęściej w postaci pojedynczych słów, nigdy zaś jako złożony tekst do tłumaczenia na PJM. Informatorzy nie znają wcześniej swoich zadań. Każde z nich poprzedzone jest instrukcją odtwarzaną przez moderatora — są to nagrane uprzednio filmy w PJM, takie same dla wszystkich uczestników nagrań. Gwarantuje to, że każdy informator otrzyma identyczną instrukcję, zaś sam moderator nie zapomni o istotnych szczegółach, które mogłyby znacząco wpłynąć na wykonanie zadania.

Sesja kończy się ewaluacją — uczestnicy nagrania proszeni są o komentarze dotyczące jego organizacji, przedstawianych materiałów itp., co pozwala na ciągłe doskonalenie procedury elicytacyjnej.

3.4. Anotacja pozyskanego materiału

Obecnie korpus PJM to ponad 500 godzin materiału filmowego w jakości HD, czyli około 15 TB danych (odpowiednik ponad 20 tys. płyt CD) pochodzących z 40 sesji nagraniowych. Statystyki te pozwalają umiejscowić projekt realizowany przez PLM UW na czele listy największych korpusów migowych świata. Na większość korpusów ogólnych składa się około 50–70 godzin nagrań; zaledwie w kilku przypadkach ich liczba przekracza 100 godzin (korpus brytyjskiego języka migowego — 180 godzin, korpus australijskiego języka migowego — 300 godzin, korpus francuskiego języka migowego — 454 godziny, korpus DGS — ponad 500 godzin; Konrad 2008).

Dane wideo zgromadzone podczas sesji nagraniowej korpusu PJM poddawane są wstępnej obróbce (obejmującej kompresję, usunięcie zbędnych fragmentów, np. przerw), a następnie umieszczane na serwerach UW (por. Mostowski 2013). Surowe dane z kamer zostają zarchiwizowane. W kolejnej fazie prac filmy wprowadzone na serwer zostają przeniesione do programu iLex — aplikacji stworzonej w Instytucie Niemieckiego Języka Migowego i Komunikacji Niesłyszących w Hamburgu do anotacji tekstów języka migowego i używanej obecnie w kilku projektach korpusowych (Hanke, Storz 2008). Dalsze prace nad zgromadzonym materiałem dokonywane są już w programie iLex. Kolejne etapy analizy danych to segmentacja materiału, czyli jego podział na pojedyncze znaki migowe, lematyzacja — przypisanie formie wyrazowej właściwej interpretacji słownikowej (lematu, leksemu), tagowanie (dodanie informacji np. o funkcji, jaką dana jednostka pełni w zdaniu, o jej właściwościach gramatycznych, o elementach niemanualnych znaku), transkrypcja artykulacyjna w notacji HamNoSys oraz tłumaczenie poszczególnych zdań w PJM na język polski. Anotacja tekstu zapewnia m.in. przeszukiwalność korpusu, w korpusach fonicznych dostępną dzięki ich formatowi elektronicznemu (McEnery 1996: 31). Pojedyncza glosa (polski odpowiednik znaku migowego) zbudowana jest z formy hasłowej (w postaci jednego lub kilku polskich wyrazów) oraz szeregu symboli opisujących układ obu dłoni (oznaczanych jako prawa — P i lewa — L) podczas wykonywania znaku. Taki zapis pozwala na szybkie identyfikowanie glos o podobnym znaczeniu, ale różnym kształcie. Wieloznaczność formy jest zaznaczana poprzez rozdzielenie poszczególnych znaczeń ukośnikiem (np. JEŚĆ/JEDZENIE/POSIŁEK/JADALNIA/ŚNIADANIE/KOLACJA P:E;L:Ø). Osobne symbole są stosowane dla oznaczenia np. klasyfikatorów (zob. Rutkowski, Łozińska 2011), gestów, wskazań, znaków kulturowych czy literowania (Filipczak 2013). Warto podkreślić, że anotacja jest procesem niezwykle czasochłonnym.

Aktualnie korpus PJM anotowany jest niemal wyłącznie w zakresie podstawowych glos lematyzacyjnych. Dalsze etapy anotacji zaplanowane są na kolejne lata realizacji projektu.

4. Zastosowania korpusu

Dane korpusowe stanowią wartość nie do przecenienia. Przede wszystkim umożliwiają prowadzenie rzetelnych i opartych na dużym materiale językowym badań lingwistycznych. Jak dotąd, dzięki materiałom korpusowym, w ramach prac badawczych PLM UW przeprowadzono m.in. analizy szyku zdaniowego PJM (Rutkowski, Łozińska, Łacheta, Czajkowska 2013), struktury konstrukcji nominalnych (Rutkowski, Czajkowska, Łacheta, Kuder 2013), repetycji (Filipczak, Mostowski 2013), relacji czasowych (Rutkowski, Łacheta, Łozińska, Mostowski 2012) czy zdań względnych (Rutkowski 2011). Trwają badania nad kolejnymi zagadnieniami oraz prace nad pierwszym w Polsce słownikiem PJM, którego brak jest szczególnie dotkliwie odczuwalny.

Korpus PJM ma także niezwykłą wartość jako skarbiec kultury Głuchych i archiwum języka migowego, zwłaszcza starszego pokolenia niesłyszących. W związku z brakiem szeroko używanej notacji PJM zbiór danych wideo jest jedyną drogą do utrwalenia obecnego kształtu języka migowego i zachowania go dla kolejnych pokoleń — zarówno dla członków społeczności Głuchych, jak i dla badaczy.

Korpus to również istotny czynnik zapobiegający deprecjacji języka migowego oraz wspierający emancypację samych Głuchych, którzy dzięki tego typu projektom powinni być utwierdzani w poczuciu wartości własnego języka.

5. Podsumowanie

Niniejszy artykuł przedstawia zarys projektu korpusowego realizowanego od trzech lat przez PLM UW. W tym czasie dokonano wyboru metody gromadzenia danych, opracowano procedurę elicytacyjną i wyposażono studio nagraniowe, a także skompletowano duże ilości danych, które czynią korpus PJM największym zaanotowanym korpusem języka migowego na świecie.

Przykłady innych krajów pokazują, że przedsięwzięcia tego typu skutkują zazwyczaj wieloletnimi i rozbudowanymi projektami badawczymi. Osiągnięte dotąd wyniki pozwalają zatem mieć nadzieję, że prace nad korpusem PJM przyczynią się do dalszego rozwoju lingwistyki migowej w Polsce.

Bibliografia

Barberà G., 2012, The meaning of space in Catalan Sign Language (LSC). Reference, specificity and structure in signed discourse, Barcelona.

Biber D., Conrad S., Reppen R., 1998, Corpus Linguistics: Investigating Language Structure and Use, Cambridge.

Crasborn O., 2002, Technological, linguistic and ethical challenges in sparing sign language data online within the ECHO Project, Nijmegen.

Fabisiak S., 2010, Języki migowe a lingwistyka korpusowa, Język Polski XC, z. 4–5, s. 338–345.

Filipczak J., Mostowski P., 2013, Repetition in Polish Sign Language (PJM): Discourse — grammar — information structure?, plakat na konferencji Theoretical Issues in Sign Language Research 11 — TISLR 11 (11 lipca 2013 r., University College London, Londyn, Wielka Brytania).

Filipczak J., 2013, Anotacja korpusu PJM, [w:] Lingwistyka przestrzeni i ruchu. Komunikacja migowa a metody korpusowe, red. P. Rutkowski, S. Łozińska, Warszawa.

Hanke T., 2004, HamNoSys — representing sign language data in language resources and language processing contexts, [w:] LREC 2004, Workshop proceedings: Representation and processing of sign languages, red. O. Streiter, C. Vettori, Paryż, s. 1–6.

Hanke T., König L., Wagner S., Matthes S., 2010, DGS Corpus & Dicta-Sign: The Hamburg Studio Setup, [w:] 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, Malta.

Hanke T., Storz J., 2008, iLex — A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography, [w:] Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages, red. O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood i E. Thoutenhoofd, Paryż.

Hollak J., Jagodziński T., 1879, Słownik mimiczny dla głuchoniemych i osób z nimi styczność mających, Warszawa.

Konrad R., 2008, The lexical structure of German Sign Language (DGS) as mirrored by empirical sign language lexicography of technical terms. A corpus-based lexicon model of DGS taking into account the iconicity of signs, Hamburg.

Lucas C., 2003, The Sociolinguistics of Sign Languages, Cambridge.

McEnery T., Wilson A., 1996, Corpus Linguistics, Edinburgh.

Mostowski P., 2013, Korpus polskiego języka migowego — zarys projektu, [w:] Lingwistyka przestrzeni i ruchu. Komunikacja migowa a metody korpusowe, red. P. Rutkowski, S. Łozińska, Warszawa.

Pęzik P., 2012, Język mówiony w NKJP, [w:] Narodowy Korpus Języka Polskiego, red. A. Przepiórkowski, M. Bańko, R.L. Górski, B. Lewandowska-Tomaszczyk, Warszawa, s. 37–47.

Rutkowski P., Łozińska S., 2011, O niedookreśloności semantycznej migowych predykatów klasyfikatorowych, [w:] Różne formy, różne treści, red. M. Bańko, D. Kopcińska, Warszawa, s. 211–223.

Rutkowski P., 2011, The syntax of relative clauses in Polish Sign Language (PJM): Some empirical, theoretical and methodological considerations, referat na konferencji Complex Sentences and Beyond in Sign and Spoken Languages (13 października 2011 r., Lichtenberg-Kolleg, Georg-August-Universität Göttingen, Getynga, Niemcy).

Rutkowski P., Czajkowska-Kisil M., Łacheta J., Kuder A., 2013, Analiza szyku przymiotnika jako przykład wykorzystania danych korpusowych w badaniach nad gramatyką PJM, [w:] Lingwistyka przestrzeni i ruchu. Komunikacja migowa a metody korpusowe, red. P. Rutkowski, S. Łozińska, Warszawa.

Rutkowski P., Łacheta J., Łozińska S., Mostowski P., 2012, The Iconicity of Temporal Reference in Sign Language: At the Intersection of Language, Discourse and Cognition, referat na konferencji CLDC 2012 — The 6th Conference on Language, Discourse, and Cognition (4–6 maja 2012 r., National Taiwan University, Tajpej, Tajwan).

Rutkowski P., Łozińska S., Łacheta J., Czajkowska-Kisil M., 2013, Constituent order in Polish Sign Language (PJM), referat na konferencji Theoretical Issues in Sign Language Research 11 — TISLR 11 (11 lipca 2013 r., University College London, Londyn, Wielka Brytania).

http://www.plm.uw.edu.pl

http://lrwiki.ldc.upenn.edu

http://www.sign-lang.uni-hamburg.de

www.signwriting.org

SUMMARY

The making of Polish Sign Language Corpus

The aim of this paper is to present an overview of the most important aspects of creating sign language corpora. It discusses selected technical, theoretical and methodological problems of data collection, includingv, among others, the selection of signing informants, the organization of recording sessions, the use of appropriate software, and the adoption of consistent elicitation and annotation conventions. The article also presents the significance of the Polish Sign Language (PJM) Corpus project for research on visual-spatial communication in Poland.