

Piotr Mostowski

## Badania korpusowe a lingwistyka migowa

### 1. Wstęp

W poniższym rozdziale przedstawiono dwie, wydawać by się mogło — dość odległe, dziedziny: badania nad językami migowymi i lingwistykę korpusową. Analiza korpusów tekstów stała się w ostatnich latach bardzo ważnym elementem wielu studiów dotyczących komunikacji wizualno-przestrzennej, głównie za sprawą dostępności niezbędnych podstaw empirycznych. Na przykładzie realizacji konkretnych projektów naukowych zostały w niniejszym rozdziale opisane procedury wykorzystywania narzędzi korpusowych w badaniach nad językami Głuchych.

### 2. Metody badań nad językami migowymi

Lingwistyka migowa w dużej mierze korzysta z dorobku badań nad językami fonicznymi, przejmując zarówno koncepcje dotyczące istoty języka, jak i dotychczas stosowane narzędzia analityczne. Na języki migowe można zatem patrzeć z perspektywy generatywizmu, kognitywizmu, dystrybucjonizmu, socjolingwistyki itd. Niezależnie od obranej metody badawczej podstawowym problemem, z którym musi zmierzyć się językoznawca, jest wybór najwłaściwszej formy reprezentacji danych. Migowy aparat artykulacyjny obejmuje wiele części ciała — od dłoni, rąk i tułowia po głowę i twarz z jej wyrafinowaną mimiką. Artykulacja wizualno-przestrzenna polega na prezentowaniu trójwymiarowych układów, na które składają się poszczególne części ciała. Układy te mają zazwyczaj charakter dynamiczny. Na kształty znaków nakładają się też często kształty elementów niejęzykowych (Świdziński, Mikulska, 2003, s. 32). Ze względu na powyższe

właściwości optymalnym sposobem rejestracji danych migowych wydaje się zapis wideo.

Specyficzna sytuacja socjolingwistyczna (zdeterminowana przez brak normy literackiej i systematycznej edukacji prowadzonej w języku migowym) skutkuje dużym zróżnicowaniem sposobów migania poszczególnych Głuchych. Użytkownikom języków migowych zdarza się posługiwać w codziennej komunikacji hybrydami i pidżynami migowo-fonicznymi. Dodatkowe utrudnienie w badaniach stanowi fakt, że lingwistyka migowa jest relatywnie młodą dziedziną, niewykładaną zupełnie — choćby w szczątkowej formie — w szkołach dla głuchych, w związku z czym bardzo niewiele osób migających zna na tyle lingwistyczny aparat pojęciowy, by dokonywać naukowego opisu swojego języka. Dotychczas gros prac z zakresu lingwistyki migowej było tworzonych przez słyszących badaczy, dla których miganie jest językiem drugim/obcym. Przy formułowaniu tez nie mogli więc oni w pełni polegać na swojej intuicji. Ze względu na powyższe ograniczenia tylko stworzenie — według określonych zasad — korpusu tekstów migowych dostarcza reprezentatywnej ilości materiału badawczego i pozwala na dostrzeżenie tendencji, które można uznać za reguły gramatyczne danego języka wizualno-przestrzennego. Możliwość taka pojawiła się niedawno, kiedy to dzięki dynamicznemu rozwojowi techniki zbieranie i przetwarzanie olbrzymiej ilości danych stało się łatwe. Rewolucja komputerowa (Świdziński, 2006) właściwie wyeliminowała problem trwałego zapisywania tekstów języka migowego.

Metoda korpusowa jest jednak często krytykowana (por. Rudolf, 2004). Najwięcej uwag pojawia się od zwolenników teorii generatywnej, w tym samego jej twórcy Noama Chomsky'ego, który argumentował, że żaden korpus nie jest dobrym modelem kompetencji językowej. Jako że opisanie kompetencji językowej jest głównym celem gramatyki generatywnej, charakterystyczną cechą tego nurtu jest marginalizowanie korpusów, uważanych za zbiory niereprezentatywne (rzeczywiste teksty nie są bowiem reprezentacją całego języka). Przeciwnie podejście cechuje brytyjską tradycję lingwistyczną wyznaczoną pracami J. R. Firtha, M. Hallidaya i J. Sinclaira. Uznaje ona lingwistykę za naukę społeczną, która powinna mieć zastosowanie w praktyce; kładzie więc nacisk na autentyczne teksty, rzeczywiste przykłady użycia języka.

Chomsky podkreśla zbędność odwoływania się do danych empirycznych, skoro — jako rodzimi użytkownicy języka — możemy sięgnąć do własnej intuicji. W przypadku języków migowych, jak wspomniano, większość badaczy stanowią słyszący, w odniesieniu do których ten argument nie ma zastosowania (nie mają oni intuicji rodzimego użytkownika). Inny kontrargument przedstawił G. Leech (1992, ss. 112–113):

model oparty na lingwistyce korpusowej jest falsyfikowalny w tym znaczeniu, że może być przetestowany na nowej próbie materiału innej od tej użytej przy opracowywaniu modelu. W razie potrzeby może być zastąpiony modelem lepiej opisującym zgromadzone dane. [...] Obiektywność — model oparty na lingwistyce korpusowej może być wielokrotnie przetestowany przez niezależnych badaczy, którzy nie są w żaden sposób emocjonalnie związani z negatywnym bądź pozytywnym wynikiem badania.

Nierzadko intuicja badacza może zawodzić przy próbie oceny częstotliwości występowania pewnych konstrukcji, więc jedynie duże korpusy pozwalają uzyskać reprezentatywne dane statystyczne.

O popularności korpusów w lingwistyce migowej świadczy gwałtowny wzrost ich liczby na przestrzeni ostatnich lat. W dalszej części pracy zostanie opisanych jedynie kilka korpusów migowych. Będą to korpusy: ECHO, australijskiego języka migowego (auslan), brytyjskiego (BSL — *British Sign Language*), niemieckiego (DGS — *Deutsche Gebärdensprache*) i holenderskiego (NGT — *Nederlandse Gebarentaal*) oraz projekt Dicta-Sign. Wybór taki, choć arbitralny, podyktowany jest kilkoma względami. Na liście znalazły się przedsięwzięcia, które są najpełniej opisane, korzystają z nowatorskich rozwiązań przy zbieraniu danych, odznaczają się rozbudowaną anotacją lub skalą projektu. W osobnym rozdziale niniejszego tomu (*Korpus polskiego języka migowego — zarys projektu*) opisano założenia realizowanego obecnie projektu budowy korpusu polskiego języka migowego (PJM). W tym kontekście warto wspomnieć również o polskim projekcie badawczym, w którym po raz pierwszy wykorzystano narzędzia korpusowe do analizy PJM.

### 3. „Bajka o Czerwonym Kapturku” jako pierwszy polski korpus języka migowego

Na Wydziale Polonistyki Uniwersytetu Warszawskiego zrealizowano w latach 1999–2002 kierowany przez prof. Marka Świdzińskiego projekt naukowo-badawczy *Studia nad kompetencją językową i komunikacją niesłyszących* (KBN 1 H01D 012 16). Jest to ważne wydarzenie w historii lingwistyki migowej w Polsce między innymi z tego względu, że po raz pierwszy na potrzeby badań językoznawczych powstał swoisty „korpus” (zebrane materiały nie spełniają wszystkich wymogów stawianych korpusom) tekstów PJM. Stanowi go zbiór nagrań, które — z wyjątkiem jednego tekstu — nie zostały zaanotowane i nie są reprezentatywne. Z drugiej strony jest to pierwsza próba analizy PJM na podstawie zebranych tekstów, a nie intuicji

użytkowników. Należy oczywiście pamiętać, że zadaniem projektu było nie budowanie korpusu, ale opis języka migowego z perspektywy różnych nauk: lingwistyki, psychologii i historii. Sukcesem jest, że przy tej okazji udało się zebrać około trzech godzin nagrań, które do tej pory są wykorzystywane w innych przedsięwzięciach.

Projekt ten należy uznać za nowatorski jeszcze z jednego powodu. Po raz pierwszy została zaproponowana notacja tekstu migowego, która działała na poziomie składni bez oddawania szczegółów fonologicznych (Świdziński, Mikulska, 2003, ss. 57–58). Zanalizowany w ten sposób tekst to popularna bajka o Czerwonym Kapturku — ponadośmiominutowe nagranie, którego jedynym aktorem jest Piotr Tomaszewski. Zapis linearny składa się ze 179 wypowiedzi o zróżnicowanej długości. W oznaczeniach użyto znaków ASCII, tekst nie został zsynchronizowany z obrazem wideo. Bez wątplenia jest to jednak pierwszy przykład zaanotowanego zestawu danych PJM.

## 4. Zagraniczne korpusy migowe

Na świecie powstaje coraz więcej korpusów migowych. Oprócz dużych projektów realizowanych w kilku krajach europejskich gromadzone są również nagrania w Chinach, Japonii czy w krajach arabskich. Różnią się one od siebie znacznie — począwszy od procedur zbierania danych, a skończywszy na zakresie i sposobie anotowania. Należy pamiętać, że część z tych korpusów powstaje w konkretnym celu i podlega specjalnej obróbce, dlatego podobieństwa między nimi mogą być znikome. Dotychczas nie zostały wypracowane uniwersalne standardy dotyczące sposobu zapisywania nagranych materiału i jego opisywania. Każdy ośrodek naukowy używa odrębnych narzędzi i procedur. Dopiero od niedawna nagrania tworzone są z myślą o porównywaniu ich między różnymi krajami.

Przed przystąpieniem do prezentacji poszczególnych projektów niezbędne jest przedstawienie problemów, które wynikają z charakteru gromadzonych danych, a nie występują (lub są łatwo rozwiązywalne) przy tworzeniu korpusów języków fonicznych (pisanych).

### 4.1. Naturalność

Założeniem większości korpusów migowych jest rejestrowanie naturalnych i spontanicznych wypowiedzi migowych. Nie jest to łatwe w realizacji. W obecności kamer uczestnicy nagrań zmieniają swój sposób migania, starają się wyrażać poprawniej, wybierają oficjalny styl wypowiedzi.

Dodatkowo mogą nie używać pewnych znaków, które uważają za regionalizmy albo za znaki nieodpowiednie. Kolejnym czynnikiem mającym wpływ na ograniczenie płynności i spontaniczności wypowiedzi jest świadomość, że zebrane dane mogą zostać opublikowane. Przed podobną trudnością stają twórcy korpusów języka mówionego (w odróżnieniu od korpusów tekstów pisanych), którym też zależy na naturalności zebranych wypowiedzi. Badacze są jednak zgodni co do tego, że ograniczenia metodologiczne, prawne i etyczne nie pozwalają w pełni spełnić tego postulatu (Grucza, 2007, s. 112). Wymienia się trzy sposoby pozyskiwania materiału jak najbliższego naturalnej komunikacji:

- ukryta obserwacja i nagrywanie. Badacz z ukrycia rejestruje rozmowę. Metoda, która pozwala uzyskać rzeczywiście naturalny dialog, ze względów prawnych i etycznych jest jednak niedopuszczalna.
- udział i jednoczesna obserwacja. Badacz bierze udział w akcie komunikacyjnym, jednocześnie go rejestrując. Największy zarzut stawiany tej metodzie dotyczy możliwości wypaczenia danych przez obecność badacza i jego aktywne uczestnictwo w rozmowie.
- elicytacja. Metoda zbliżona do powyższej. Zakłada obecność badacza w rozmowie i jednoczesne próby sterowania procesem komunikacji za pomocą odpowiednich materiałów stymulacyjnych. Jest to najczęściej stosowany sposób przy tworzeniu korpusów migowych. W sesji nagraniowej uczestniczy głuchy moderator (niekoniecznie sam badacz), który prezentuje odpowiednio przygotowane zadania elicytacyjne i na tym właściwie kończy się jego rola.

Czasami stosuje się metodę podobną do ukrytej obserwacji — moderator wychodzi z pomieszczenia, pozostawiając włączone kamery, a uczestnikom nie przekazuje konkretnego zadania. Są oni jednak poinformowani, że ich rozmowa nadal będzie rejestrowana.

## 4.2. Lematyzacja

Spośród korpusów języków migowych wyróżniają się te, których zawartość została poddana lematyzacji, czyli procesowi przypisania każdej formie wyrazowej jej reprezentacji słownikowej (leksemu, lematu). Rozbudowuje to możliwości wyszukiwania i znacznie ułatwia na przykład tworzenie słowników. Lematyzacja nastrocza w wypadku języków wizualno-przestrzennych wielu trudności (por. rozdział *Anotacja korpusu PJM* autorstwa Joanny Filipczak w niniejszym tomie). Z technicznego punktu widzenia połączenie

poszczególnych wystąpień znaków z glosami leksemowymi można zrealizować na wiele sposobów. Najefektywniejsze rozwiązania stosowane obecnie na świecie to programy ELAN i iLex, z których drugi — jako wykorzystywany podczas anotacji korpusu PJM — został opisany we wspomnianym artykule Joanny Filipczak. ELAN (EUDICO Linguistic Annotator) powstał w Instytucie Maxa Plancka w Nijmegen w 2005 roku (Crasborn, Sloetjes, 2008). Program ten stanowi regularnie aktualizowaną, bezpłatną aplikację, która powstała na potrzeby anotowania tekstów. Nie była tworzona z myślą o wykorzystywaniu w pracach nad korpusami migowymi, jednak jej możliwości czynią z niej doskonale narzędzie do opisywania nagrań. Użytkownik ręcznie może przeprowadzać wielopoziomą anotację plików wideo i/lub audio. Praca odbywa się na warstwach zawierających tekst anotacji — z reguły jedna warstwa odzwierciedla jeden typ informacji, np. dotyczący mimiki. Warstwy mogą być grupowane bądź łączone ze sobą hierarchicznie. Możliwe jest tworzenie różnych restrykcji dotyczących samych warstw (przypisanie im określonego typu) bądź ich zawartości (stworzenie zamkniętej listy jednostek do wyboru). Dostępna jest zaawansowana wyszukiwarka pozwalająca na tworzenie zapytań z użyciem operatorów logicznych i przeszukiwanie wielu plików.

Działanie programu ELAN opiera się na systemie dokumentów — anotaator pracuje lokalnie (na pojedynczym komputerze). W osobnych plikach przechowywane są film i anotacja. Przy większej ilości danych narzuca to konieczność rygorystycznego przestrzegania zasad nazewnictwa i katalogowania.

### 4.3. ECHO

Korpus języków migowych powstały w ramach europejskiego programu ECHO (*European Cultural Heritage Online*) to pierwszy korpus elektroniczny w całości dostępny przez Internet (Crasborn i in., 2007). Ma on charakter wielojęzyczny — gromadzi nagrania tekstów niderlandzkiego języka migowego (NGT), brytyjskiego języka migowego (BSL) oraz szwedzkiego języka migowego (STS – *Svenskt teckenspråk*). Jednym z celów projektu było stworzenie bazy do badań porównawczych, dlatego w każdym z uczestniczących w nim krajów używano tych samych zadań elicytacyjnych (pięć bajek Ezopa, podstawowe słownictwo jako lista wyrazów, krótki wywiad oraz poezja migowa). Na każdy język przypadało od 2 do 4 użytkowników (kobiet i mężczyzn).

Przy tworzeniu korpusu po raz pierwszy pojawiły się komplikacje, które do tej pory są podstawowymi problemami w tego typu przedsięwzięciach. Ze względu na ich charakter można podzielić je na trzy kategorie:

- kwestie techniczne dotyczące zarówno przeprowadzania nagrań uczestników, jak i późniejszej anotacji,
- kwestie lingwistyczne dotyczące anotacji danych,
- kwestie prawne dotyczące wykorzystywania wizerunku osoby migającej.

Na potrzeby projektu ECHO do anotacji użyto programu ELAN. Przed rozpoczęciem anotacji spisano szczegółowe zasady dotyczące jej warstw i segmentacji materiału. Zadbano o podpisanie z każdym głuchym stosownej umowy, w której wyrażał on zgodę na opublikowanie swojego nagrania w Internecie.

Projekt zakończył się w 2003 roku, a cały zebrany materiał jest dostępny pod adresem [http://corpus1.mpi.nl/ds/imdi\\_browser/?openpath=MPI84302%23](http://corpus1.mpi.nl/ds/imdi_browser/?openpath=MPI84302%23).

#### 4.5. Auslan

Korpus australijskiego języka migowego (auslan) jest zdigitalizowanym archiwum nagrań 256 użytkowników języka migowego (Johnston, Schembri, 2006; Johnston, 2009). Składa się z dwóch części, które powstawały w różnych okresach:

- projekt badający socjolingwistyczne zróżnicowanie języka auslan (2003–2005),
- projekt związany z dokumentowaniem ginących języków (1100 klipów wideo, 5 miast x 20 uczestników x 3h, 2004–2006).

Nagrane zostało 300 godzin migania, co zajmuje około 3 TB danych. Informacje o zajmowanej przestrzeni dyskowej są istotne, ponieważ w przeciwieństwie do tradycyjnych korpusów, które mieszczą się na płycie DVD, zbiory nagrań wideo mogą zajmować gigantyczne ilości pamięci, co powoduje liczne problemy przy tworzeniu korpusów.

W sesjach nagraniowych do korpusu auslan uczestniczyli nie tylko rodzimi użytkownicy, lecz także osoby, które przyswoiły język migowy później, ale jeszcze w dzieciństwie (użytkownicy prawie natywni). Uczestnicy pochodzili z różnych miast Australii, podczas trzygodzinnej sesji mieli do wykonania kilka zadań, całość prowadził głuchy, który był znany osobom nagrywanym.

W każdym z miast (Sydney, Melbourne, Brisbane, Adelaide i Perth) osobami odpowiedzialnymi za wyszukanie informatorów byli głusi

współpracownicy, którzy doskonale znali lokalną społeczność niesłyszących. Do nagrań używano dwóch kamer cyfrowych — chociaż badacze informują o potrzebie większej ich liczby, przede wszystkim dostępności kamery nagrywającej z góry, która pokazuje wykorzystanie przestrzeni dookoła migającego.

Nagrany film był ostatecznie dzielony na klipy odpowiadające poszczególnym punktom elicytacyjnym, którymi były:

1. Zapoznanie się. Prezentacja własnych znaków imiennych i krótka rozmowa z prowadzącym sesję\*.
2. Narracja (tekst). Opowiadanie wcześniej poznanego tekstu. Uczestnicy na tydzień przed nagraniem dostawali bajki Ezopa w formie tekstu pisanego (*Żółw i zając*, *Pastuszek i wilk* — znana też pod tytułem *Chłopiec, który krzyczał „wilki!”*, „wilki!”), których opowiadanie ćwiczyli w domu. Podczas sesji migali ich treść partnerowi.
3. Narracja (wydarzenie). Prowadzący prosił uczestnika o opowiedzenie partnerowi ważnego wydarzenia ze swojego życia.
4. Ankieta (opinie). Prowadzący pytał uczestników o ich zdanie dotyczące najważniejszych problemów głuchej społeczności w Australii — stosowania implantów ślimakowych, dostępu do tłumaczy, nauczania języka migowego w szkole, przyszłości języka auslan.
5. Konwersacja. Pozostawieni sami sobie uczestnicy mieli ok. 15 minut na swobodną konwersację.
6. Narracja (film rysunkowy). Uczestnicy na przemian oglądali po dwa klipy wycięte z odcinka przygód Tweety'ego i Sylwestra (odcinek *Canary Row*, Warner Bros). Następnie opowiadali partnerowi obejrzaną historię.
7. Narracja (historyjka obrazkowa). Jeden z uczestników po obejrzeniu historyjki obrazkowej dla dzieci (*Frog, where are you?*, Mayer, 1969) opowiadał jej akcję partnerowi.
8. Narracja (tekst migowy). Jeden z uczestników opowiadał partnerowi obejrzaną historię, która zamigana była w języku auslan.
9. Opis (materiał wideo). Jeden z uczestników po obejrzeniu nagrania z kilkoma miniaturami opisywał partnerowi, co zobaczył. Materiał został dobrany pod kątem wywołania konkretnych znaków.

---

\* W językach migowych wszystkie nazwy własne są znakami (nieróżniącymi się formą artykulacyjną i sposobem funkcjonowania od znaków pospolitych). Dotyczy to także migowych odpowiedników imion i nazwisk. Każdy członek danej społeczności Głuchych ma tzw. przydomek (znak imienny), który został mu nadany przez innych. Nie musi on być w żaden sposób związany z imieniem lub nazwiskiem danej osoby w języku fonicznym.



10. Opis (obrazek). Jeden z uczestników opisywał po kolei 18 obrazków, na których przedstawione były różne scenki („odwracalne” — każda z postaci może być agensem, „nieodwracalne” — tylko jedna postać może być agensem, lokatywne — przedstawiające jakieś ułożenie obiektów w przestrzeni, nie akcję). Partner dostawał zestaw kart ze scenkami. Zadanie polegało na wybraniu właściwej scenki przedstawionej przez drugą osobę. Z pary kart tylko jedna przedstawiała odpowiedni obrazek.
11. Pytania („Znajdź różnice”). Każdy uczestnik dostawał jeden z dwóch obrazków, które nieznacznie różniły się między sobą. Zadaniem było ustalenie wszystkich różnic. Materiał dobrany pod kątem wywołania jak największej ilości pytań i zdań przeczących.

Anotacja klipów wykonywana jest w programie ELAN. W celu jak najlepszej standaryzacji oznaczeń powstały szczegółowe zasady używania skrótów i konwencji zapisu oraz szablon używanych warstw, który zawiera aż 49 pozycji, z czego minimalnie muszą być wypełnione trzy — glosa identyfikująca znak dla obu rąk i luźne tłumaczenie. Badacze sami przyznają, że nie jest oczywiste, które informacje należy uzupełniać w pierwszej kolejności w trakcie procesu anotowania. Model zakłada uzupełnianie pozostałych informacji w dowolnym czasie w zależności od ustaleń innych zespołów badawczych oraz od doraźnych potrzeb dotyczących korpusu. Opisywanie nagrań migowych jest procesem niezwykle żmudnym, co doskonale widać na przykładzie korpusu auslan. Od 2006 do końca marca 2011 roku zaanotowano na poziomie podstawowym 347 klipów z 1100 (31%), z czego 50 szczegółowo na potrzeby badań. Przetworzenie całości zajmie więc jeszcze wiele lat. Dane anotacyjne są dostępne tylko na prośbę innych ośrodków badawczych.

Każdy plik wideo oraz plik z anotacją posiada szczegółowe metadane:

- uczestnik (m.in. miejsce pochodzenia, płeć, wiek, wykształcenie),
- zawartość (użyte procedury elicytacyjne, zadania językowe),
- media (format i typ),
- projekt (nazwa, język, metodologia),
- sesja (m.in. nazwa zadania, uczestnicy),
- powiązane zasoby (obecność pliku z anotacją, typ anotacji).

Narzędzia wyszukiwania, które oferuje program ELAN, wraz z możliwością przeszukiwania IMDI pozwolą na tworzenie zaawansowanych kwerend łączących w sobie informacje gramatyczne z metadanymi. Jest to ogromnym ułatwieniem w przeprowadzaniu badań nie tylko językoznawczych, lecz

także socjolingwistycznych. Niedogodnością związaną z wykorzystaniem oprogramowania ELAN przy projekcie tej skali jest brak możliwości równoległego korzystania przez poszczególnych anotatorów ze wspólnej bazy znaczników. Każda z anotacji w projekcie powstaje niezależnie od innych, co wymaga ścisłego przestrzegania przez anotatorów przyjętych zasad (objętość podręcznika z zasadami anotacji korpusu auslan to prawie 70 stron).

#### 4.6. BSL

Projekt tworzenia korpusu brytyjskiego języka migowego trwał od stycznia 2008 roku do grudnia 2010 (Schembri, 2008). Jego zadaniem było opracowanie korpusu dostępnego online, który posłuży do badań wariacji fonologicznych, leksykalnych i socjolingwistycznych występujących w BSL. Wykorzystano dwa rodzaje tekstów: wytworzone spontanicznie oraz wywołane sztucznie (elicytacja).

W fazie pierwszej z ośmiu regionów (Londyn, Bristol, Birmingham, Manchester, Newcastle w Anglii oraz Walia, Szkocja, północna Irlandia) wybrano minimum 30 użytkowników języka migowego, co przełożyło się na 249 nagranych osób. W celu uzyskania zrównoważonego korpusu zadbano o proporcje między płcią informatorów, wiekiem, pochodzeniem, wykształceniem itp. Warunkiem koniecznym było przyswojenie BSL przed 7 rokiem życia. Nagrania przeprowadzano podczas 2–4 dwugodzinnych sesji. Faza druga obejmowała nagrania 100 natywnych użytkowników języka migowego, którzy przez dwie godziny wykonywali różne zadania elicytacyjne. Osoby do nagrań wyszukiwane były przez głuchych pracowników terenowych (osoby kontaktowe) — pary informatorów dobierano z zachowaniem ich podobnego wieku i pochodzenia (rozumianego jako co najmniej dziesięcioletni pobyt) z tego samego regionu.

Nagranie prowadziła osoba głucha, podczas całej sesji nie pojawiały się żadne osoby słyszące. Po pięciominutowym przedstawieniu się następowała około półgodzinna swobodna rozmowa w diadzie — unikano włączania się osoby prowadzącej. Za takim rozwiązaniem przemawia chęć eliminacji interferencji językowych, głusi bardzo często zmieniają swój sposób migania w celu dostosowania się do drugiej osoby (np. unikanie znaków regionalnych). Po rozmowie prowadzący przeprowadzał wywiad. Pytał się m.in. o biegłość w BSL i języku angielskim, o stosunek do głuchoty, pochodzenie, pozycję socjalną itp. Kolejne zadanie było całkowicie elicytowane. Informatorom prezentowano pięćdziesiąt kart z różnymi pojęciami. Starano się pokazywać jak najwięcej graficznych reprezentacji, jednakże w celu większej precyzyjności stosowano również angielskie podpisy. Pojęcia dobrane

były na podstawie wyników wcześniejszych badań dotyczących wariantowości w BSL. Podczas osobnej sesji nagrywano wypowiedzi wywołane tylko przy użyciu materiału elicytującego. Dobór zadań był związany z aspektami, które — jak zakładano — pojawią się w tekstach. Wykorzystano niektóre punkty z procedur elicytacyjnych korpusów auslan i NGT.

Do nagrań użyto czterech kamer cyfrowych umieszczonych na statywach. Korzystano również z oświetlenia studyjnego, gdy światło dzienne było niewystarczające. Anotacja odbywała się za pomocą programu ELAN.

Korpus BSL jest pierwszym po projekcie ECHO zamierzeniem opublikowania wszystkich zebranych danych w Internecie. Najczęściej dostęp do nagranych materiałów mają tylko pracownicy ośrodka badawczego. W związku z nowatorską koncepcją pojawiły się istotne kwestie prawne — jako że nie można przedstawić tekstu języka migowego oddzielnie od osoby, która go wykonuje, potrzebne jest zapewnienie sobie praw do wizerunku informatora, co pozwala na publikacje wideo.

#### 4.7. NGT

Korpus niderlandzkiego języka migowego (NGT) jest już zamkniętym projektem, który trwał w latach 2006–2008 (Crasborn, Zwitserlood, 2008). Zebrany materiał wraz z glosowaniem i translacją dostępny jest w całości w Internecie na licencji Creative Commons BY-NC-SA 3.0, która umożliwia wykorzystanie zawartości na użytek niekomercyjny z każdorazowym podaniem autora i źródła.

Głównym celem projektu było stworzenie korpusu do dalszych badań naukowych. Sfilmowano 92 informatorów z całego kraju w wieku od 17 do 84 lat, którzy posługują się NGT. Starano się unikać interferencji z holenderskim fonicznym, dlatego nagrania prowadziła osoba głucha. Początkowo planowano nagrania tylko 24 osób z dwóch regionów, ale z czasem zmieniono liczbę uczestników. Chętnych wyszukiwano za pomocą reklam na stronach dotyczących problematyki głuchych, przez spotkania w klubach dla głuchych oraz wskutek rozprzestrzenienia się informacji w społeczności osób niesłyszących. Gdy projekt stał się znany, wielu starszych głuchych zgłaszało się samych, tłumacząc, że chcą zarejestrować, a tym samym ocalić od zapomnienia, swój wariant NGT. Każdy ochotnik informowany był jeszcze przed nagraniem o późniejszej publikacji w Internecie oraz proszony o podpisanie zgody.

Podczas sesji nagraniowych próbowano uzyskać od osób komunikaty naturalne — nie przeszkadzano w spontanicznych rozmowach, ale używano też zestawu zadań elicytacyjnych zaczerpniętych z prac nad korpusem

auslan, dostosowując część z nich do holenderskich warunków. Wykorzystano m.in. kreskówkę *Tweety i Sylwester* oraz historyjkę *Frog, where are you?*. Przestrzegano też zasady jak najmniejszej ilości tekstu pisanego.

Kolejność procedury elicytacyjnej:

1. Zapoznanie się. Uczestnicy krótko przedstawiali się sobie nawzajem.
2. Narracja. Każdy z uczestników był proszony o opowiedzenie jednego (bądź kilku) znaczących wydarzeń ze swojej przeszłości.
3. Ankieta. Przedstawiano filmiki prezentujące kilka ważnych spraw dotyczących społeczności głuchych i niderlandzkiego języka migowego. Uczestnikom zadawane były pytania na ten temat oraz zawiązywana była dyskusja.
4. Narracja (bajka). Uczestnicy oglądali krótką bajkę w NGT, po obejrzeniu opowiadali jej historię własnymi słowami.
5. Narracja (komiks). Uczestnicy na zmianę opowiadali wydarzenia z obejrzonej wcześniej historyjki obrazkowej.
6. Narracja (skecz telewizyjny). Zadanie analogiczne do 5. Tym razem prezentowany był fragment programu telewizyjnego.
7. Narracja (książka z obrazkami). Każdy uczestnik oglądał kilka historyjek obrazkowych, potem opowiadał ich treść partnerowi.
8. Pytania (znajdź różnice). Każdy uczestnik dostawał jeden z obrazków, które nieznacznie różniły się między sobą. Zadaniem było wspólne ustalenie wszystkich różnic.

Informatorzy nagrywani byli w parach w celu uzyskania „naturalnego” migania. Na początku głuchy moderator wyjaśniał ideę samego korpusu oraz opisywał zadania elicytacyjne. Sesja trwała średnio około 4 godzin (trzy minisesje jednogodzinne rozdzielone przerwami), uzyskiwano jednorazowo od 2,5 do 5,5 godzin materiału nadającego się do dalszej obróbki. Nagrywano w kilku miejscach: na Uniwersytecie w Nijmegen, w szkołach dla głuchych, klubach dla głuchych. Nigdy nie korzystano z dodatkowego oświetlenia studyjnego, dlatego jakość poszczególnych klipów czasami bardzo się różni. Nagrania dokonywane były za pomocą czterech kamer cyfrowych uruchamianych ręcznie — dla potrzeb synchronizacji na początku trzykrotnie klaskano. Dwie kamery HDV rejestrowały osobno migających — wysoka jakość eliminuje potrzebę osobnych kamer skierowanych tylko na twarz. Dwie kamery miniHDV umieszczone podsufitowo zapisywały obraz wykorzystywania przestrzeni dookoła uczestników. W ten sposób przy minimalnych środkach uzyskano nagrania, które zbliżone są do efektu trójwymiarowości.

Następnym krokiem była obróbka techniczna. Powycinano zbędne sceny (np. moment zapoznawania się z zadaniem), użyto różnych metod kompresji, które mają ułatwić przechowywanie i dostęp. Nie wnikając w szczegóły techniczne, warto podkreślić, że wybór kodeków i użytych formatów przeprowadzany był pod kątem ich dostępności (tylko bezpłatne rozwiązania) oraz przyszłej użyteczności — wybrano najbardziej perspektywiczne metody dostępne w tamtym czasie.

Korpus NGT został opublikowany przez Max Planck Institute for Psycholinguistics z Nijmegen jako część zbioru korpusów językowych dostępnego w Internecie. W celu ułatwienia poruszania się po zawartości korpusu do nagrań dodano szereg metadanych, które mogą być przeglądane za pomocą licznych darmowych aplikacji (wszystkie autorstwa zespołu z Max Planck Institute for Psycholinguistics). Dodatkowe informacje dotyczą: regionu, płci i wieku osoby nagrywanej, daty nagrania oraz informacji o sesji z wyszczególnieniem kolejnych zadań elicytacyjnych. Starano się jednak w jak największym stopniu chronić informacje osobiste — część danych jest dostępna tylko po wyrażeniu zgody przez osobę zarządzającą korpusem. Dodatkowo nigdzie nie pojawiają się imiona, nazwiska i inicjały informatorów.

W przewidzianej grupie odbiorców korpusu oprócz lingwistów wyróżniono także psychologów, nauczycieli, słownikarzy, nauczycieli NGT i tłumaczy oraz głuchych, którzy mogą być zainteresowani zawartymi treściami kulturowymi. Mimo łatwego dostępu do danych przez przeglądarkę IMDI autorzy postanowili sukcesywnie tworzyć strony tematyczne, które ułatwią przeszukiwanie korpusu. Dodatkowym ułatwieniem dla osób korzystających są anotacja i translacja zebranego materiału.

Anotację w programie ELAN wykonali biegli użytkownicy NGT — nie były nią objęte jednak wszystkie pliki. Transkrypcja wideo — nawet najprostsza — jest niezwykle czasochłonna, dlatego zastosowano warstwy z tylko podstawowymi informacjami: glosami dla każdej z rąk oraz tłumaczeniem. Było to spowodowane również kwestiami organizacyjnymi — ograniczonym budżetem oraz przewidzianym czasem realizacji projektu (dwa lata). Do wykonania całej pracy powstał zespół czterech głuchych asystentów, którzy zostali odpowiednio przeszkoleni w zakresie standardów anotacji i obsługi ELAN. W pracy pomagali sobie dwujęzycznym słownikiem holendersko-migowym (NGT) z 5000 hasłami oraz słownikami holenderskimi (w tym słownikiem obrazkowym) do sprawdzania ortografii. Niestety, okazało się, że asystenci popełnili mnóstwo literówek oraz nie przestrzegali (z różnych powodów) narzuconych procedur opisywania znaków. Mimo częściowej korekty nadal nie można całkowicie polegać na napotkanych glosach. Jest to przykład bardzo istotnego zagadnienia — tego, kto i jak powinien dokonywać anotacji.

Mimo symplifikacji osiągnięte rezultaty nie są imponujące. Na 2375 plików anotacja dodana została jedynie do 160 (12 godzin migania, około 64 000 glos). Zastosowano jednak rozwiązanie, które pozwala w dowolnym momencie uzupełnić brakujące dane. Używając szablonu z jednakowym zestawem warstw oraz skryptów w języku Perl, do każdego pliku dodano „szkielet” anotacji — puste miejsca do wypełnienia i półautomatycznie przypisane metadane.

W celu ułatwienia nielingwistom korzystania z korpusu niektóre nagrania oprócz podstawowych glos mają dołączone tłumaczenie na język holenderski (foniczny). Zespół certyfikowanych tłumaczy dokonywał tłumaczenia symultanicznego — oglądając klipy, jednocześnie nagrywał wersję foniczną miganego tekstu. Udział w tym brali również studenci translatoryki, dla których był to rodzaj ćwiczeń. Tak powstałe pliki zostały dołączone do korpusu. Jest to novum dla korpusów migowych.

Tłumacz w przypadku natrafienia na trudniejszy fragment mógł konsultować się z kolegami, choć z reguły oglądał dany klip maksymalnie dwukrotnie. Możliwość przewijania klipu i wprowadzania poprawek sprawiła jednak, że proces tłumaczenia trwał średnio dziesięciokrotnie dłużej niż czas odtwarzania danego klipu. W wyniku tak dużej rozbieżności w stosunku do planowanych założeń nie można było dodać tłumaczenia do każdego nagrania. Nie udało się również — z kilku powodów, głównie braku właściwego oprogramowania — wykorzystać programów rozpoznających mowę i automatycznie wygenerować tekstu tłumaczenia.

Twórcy korpusu zastanawiają się, jak będzie wyglądała jego przyszłość. Słusznie podkreślają, że w dobie rozwoju korpusów migowych bardzo ważną kwestią jest standaryzacja glosowania, aby umożliwić wymianę materiałów między różnymi ośrodkami badawczymi. Dostrzegają również rozwój technologii, który pozwala już na pewną automatyzację, nadal jednak nie zapewnia zaawansowanych narzędzi do przeszukiwania tak dużej ilości nagrań. Kluczem do sukcesu pozostaje dobrze wykonana anotacja, która jest bardzo czasochłonna i kosztowna.

Podsumowując, trzeba przyznać, że korpus NGT jest interesującym przedsięwzięciem. Z jednej strony godne podziwu jest jego wykonanie oraz umieszczenie całej zawartości w Internecie (wraz z dokładną licencją użytkowania). Na użytek zbierania i obróbki materiału powstało wiele narzędzi, które nie są przypisane tylko do tego projektu, a więc mogą być zastosowane przez inne zespoły badawcze. Wykorzystano też rozwiązania wypracowane przy tworzeniu korpusów fonicznych, niektóre wzbogacono o funkcjonalności stricte „migowe”. Z drugiej strony zaskakuje ubogi repertuar anotacji. Glosy stosowane są niekonsekwentnie, pojawiają się opisy zaczerpnięte wprost z języka holenderskiego, dla których brak odpowiedników w języku

migowym. Tylko minimalna część materiału została sprawdzona przez drugą osobę, dlatego nie uniknięto literówek i innych błędów. Największym brakiem w opisie jest całkowite pominięcie części niemanualnej znaku (nawet gdy jest to komponent decydujący o znaczeniu). W przypadku znaku, w którym migający jednocześnie pokazuje znak i porusza głową, opisany został tylko mig, nie informacja przekazywana przez ruch głową (np. negacja). Nie jest to jednak wynikiem zaniedbania. Przy tworzeniu korpusu położono nacisk przede wszystkim na jak najszybsze przygotowanie nagrań do publikacji, wychodząc z założenia, że resztę informacji będzie można uzupełniać na bieżąco w zależności od potrzeb badawczych. Jest to więc korpus właściwie nieanotowany, który można transkrybować wielorako w zależności od tego, jakie dane są potrzebne.

#### 4.8. DGS

Powstający korpus niemieckiego języka migowego (DGS) jest pod wieloma względami największym przedsięwzięciem tego typu (Prillwitz i in., 2008; Blanck i in., 2010; Hanke i in., 2010; Nishio i in., 2010). Imponuje przewidywana liczba zebranego materiału, czas trwania projektu i szczegółowość anotacji. Cały projekt stworzenia korpusu i docelowej publikacji słownika DGS opartego na tym materiale przewidziany jest na 15 lat (2009–2023), a jego koszt to 8,5 mln euro. Ośrodkiem naukowym, który podjął się tego wyzwania, jest Instytut Niemieckiego Języka Migowego i Komunikacji Niesłyszących (niem. *Institut für Deutsche Gebärdensprache und Kommunikation Gehörloser*, IDGS) z Uniwersytetu w Hamburgu. Ze względu na czas trwania projektu i możliwości techniczne jest to największy projekt stworzenia korpusu języka migowego, jaki do tej pory realizowano, dlatego poddany będzie szczegółowemu opisowi.

Dwujęzyczny słownik migowo-niemiecki będzie pierwszą tak obszerną publikacją powstałą na bazie korpusu. Jest to bardzo ważny krok w dokumentowaniu, badaniach i nauczaniu niemieckiego języka migowego. Stanie się punktem odniesienia dla wszystkich, którzy są zainteresowani DGS — nie tylko lingwistów, lecz także tłumaczy, słyszających pracujących z głuchymi oraz samych głuchych, którzy chcieliby dowiedzieć się więcej o swoim języku ojczystym. Zebrane nagrania z terenu całego kraju będą opisane w dwóch językach i przedstawiane w różnorodnych kontekstach, dlatego rola tej publikacji jest trudna do przecenienia.

Przez około piętnaście lat nagranych ma być 328 informatorów (164 pary) z terenu całych Niemiec podzielonych na dwanaście regionów. Ustalono zestaw kryteriów, które muszą spełniać nagrywani. Oczywistymi

wymaganiami są pochodzenie z Niemiec i używanie DGS. Przy wyborze starano się o zachowanie równych proporcji między liczbą kobiet i mężczyzn. Ograniczono również liczbę głuchych artystów, lingwistów i nauczycieli DGS (łącznie nie więcej niż 20%). Wprowadzono też cztery kategorie wiekowe: 18–30, 30–45, 45–60 i 60+. Szczególny nacisk położono na grupę głuchych z głuchych rodzin — są to tak zwani głusi dynastyczni, których język migowy odznacza się wyjątkowym bogactwem i niewielkim wpływem języka fonicznego. Taka liczba osób ma zapewnić około 400 godzin materiału wideo, a w tym 2,25 miliona znaków, co pozwala porównywać ten korpus z korpusami języków fonicznych.

W naborze informatorów zastosowano stanowisko pośrednika, tzw. osobę kontaktową. W pierwszym etapie rekrutacji wybrano w sumie dwadzieścia głuchych osób ze wszystkich regionów nagrań. Były to postaci doskonale znające lokalną społeczność głuchych i obdarzone jej szacunkiem. Po dokładnym przeszkoleniu w Hamburgu ich zadaniem było znalezienie chętnych do nagrań oraz dalsze pośrednictwo w rozmowach z ochotnikami. Jest to rozwiązanie zbliżone do zastosowanego przy tworzeniu korpusu BSL, jednak wiążące się z większymi obowiązkami osób kontaktowych. Oprócz wyszukania informatorów były one odpowiedzialne za moderację sesji nagraniowej i rozwiązanie wszystkich ewentualnych problemów podczas jej trwania.

Na specjalną uwagę zasługuje rozwiązanie kwestii prawnych dotyczących ochrony wizerunku i danych osobowych. Aspekt prawny jest bardzo dużym problemem przy tworzeniu korpusu języka migowego — pod uwagę należy wziąć kwestie etyczne, ochronę danych osobowych i prawo autorskie. Cały materiał wideo będzie udostępniony na potrzeby badaczy i użytkowników słownika. Jak wspomniano, w przypadku tekstów języka migowego niewykonalny jest proces bezstratnego oddzielenia ich od nadawcy komunikatu, tym samym zawsze obecna jest osoba migająca. Każdemu informatorowi dostarczany był zestaw zawierający trzystronicowy dokument w języku niemieckim opisujący, na czym polega sesja nagraniowa, jaki typ danych będzie zbierany, kto będzie miał do niego dostęp w dalszej części projektu, do jakich celów wykorzystane będą nagrania. Oprócz tego głuchy otrzymywał półtorastronicową zgodę na wykorzystywanie danych, którą podpisywał podczas pierwszego dnia sesji w obecności kamer — jest to dodatkowe zabezpieczenie w przypadku późniejszych roszczeń. Zadbano o przekład na DGS — na dołączonej płycie DVD cały tekst dokumentów był migany przez biegłego użytkownika języka migowego. Informacje podlegające ochronie dzielą się na trzy grupy: dane kontaktowe (imię, nazwisko, adres itp.), metadane (wiek, wykształcenie, region pochodzenia) oraz nagrane wideo. Informator mógł wyrazić zgodę na całkowity dostęp do wszystkich danych bądź też na każdorazowy kontakt w przypadku ich użycia (np. udostępnienie nagrań osobom



trzecim do celów naukowych). Po zakończeniu sesji nagraniowej informatorzy mieli do niej pełny dostęp i mogli jeszcze podjąć decyzję o ewentualnych wykluczeniach pewnych momentów. Przyjmowano, że brak zastrzeżeń przez 4 tygodnie od momentu zakończenia filmowania oznacza pełną akceptację całości materiału. Do końca roku 2010, kiedy nagranych zostało już 51 informatorów, zostało wykluczonych 48 minut ze 125 godzin (0,64%). Są to głównie momenty pomiędzy zadaniami albo fragmenty, w których głusi podają prywatne informacje (np. o swoim znaku migowych, wyrażają opinie o innych głuchych ze środowiska), do wykluczenia zostały też zgłoszone wycinki, w których osoba poprawia sobie ubranie, wydmuchuje nos itp.

Sesje nagraniowe odbywały się w Hamburgu i dziesięciu innych miastach. Podział na regiony, z których zbierano dane, oraz wybór miejsc na studio mobilne związany był z obecnością szkół dla głuchych, wielkością głuchej społeczności i prozaiczną łatwością dotarcia na miejsce. W pomieszczeniu nagraniowym znajdowali się tylko głusi — informatorzy, moderator (osoba kontaktowa) i technik, który czuwał nad sprzętem. Para informatorów umieszczona była twarzami do siebie, w odległości około trzech metrów. Kamery kierowane były na poszczególne osoby, a dodatkowa nagrywała całą scenę. Materiały elicytacyjne wyświetlano na monitorach ustawionych przed migającymi na odpowiedniej wysokości, nieprzeszkadzającej w miganiu. Moderator siedział obok informatorów, mógł swobodnie prezentować kolejne zadanie i obserwować rozmowę, jednak poza tym nie udzielał się w żaden sposób, jedynie w wyjątkowych sytuacjach służył podpowiedziami.

Do rejestracji obrazu użyto w sumie siedmiu kamer — po trzy na każdego informatora oraz jedną, która obejmowała całą sytuację komunikacyjną. Dwie kamery HD zapewniały widok z przodu, a dwie kolejne nagrywały z perspektywy ptasiej (umieszczone centralnie nad głowami migających), która ułatwia rozpoznanie znaków i ukazuje sposób zagospodarowania przestrzeni wokół osoby. Dodatkowo do kamer nagrywających z przodu dołączono kamery stereoskopowe, z których obraz pozwala na stworzenie przekazu 3D i dalszą automatyczną analizę wideo. Siódma kamera była również kamerą HD, na jej obrazie widoczny jest także moderator, co pozwala anotatorom na szybkie zorientowanie się, co dzieje się w studiu. Podobnie jak w korpusie NGT wybór sprzętu HD eliminuje konieczność ustawienia dodatkowych kamer nagrywających w zbliżeniu. Takie ustawienie było wynikiem licznych testów. W przeciwieństwie do przywołanego korpusu NGT zarzucono pomysł delikatnego obrócenia osób migających i umieszczenia kamer z boku. Rozstaw ten powoduje, że głuchy kieruje swój komunikat raz do kamery, a raz do partnera — zaburza to rozmowę i wpływa na aspekty przestrzenne języka migowego. Rozważano również zwiększenie dystansu między osobami i umieszczenie kamer przed

nimi na podłodze. Jest to jednak rozwiązanie bardzo nienaturalne do prowadzenia konwersacji. Ostatecznie wybrany sposób jest optymalny — nie zakłóca rozmowy, a nagrania nie są zdeformowane.

Bardzo ważną postacią w procesie zbierania danych jest osoba kontaktowa, która jednocześnie zajmuje się moderacją sesji. Wybór osoby głuchej eliminuje możliwość interferencji z językiem fonicznym — podczas całego nagrania w studiu nie pojawia się żadna osoba słyszająca. Zadaniem moderatora jest sprawne przeprowadzenie nagrania — upewnienie się, że głuchy zrozumiał polecenie kolejnego zadania, oraz sprawdzenie, że odpowiedni materiał został wyświetlony na monitorach uczestników. W trakcie realizowania polecenia jego rola ogranicza się jedynie do pilnowania czasu przewidzianego na dany fragment i ingerowaniu tylko w koniecznych sytuacjach. Każda osoba kontaktowa przechodziła dokładne szkolenie, jak zachowywać się w trakcie nagrań. Sesja nagraniowa ma przewidziany czas trwania. Moderator posiada zestaw dodatkowych zadań, które może dołączyć, jeśli uczestnicy za szybko realizują polecenia — na tej samej zasadzie może niektóre punkty omijać, jeśli czasu jest już za mało. Przewidziany czas nagrania jednej pary informatorów to maksymalnie 6 godzin.

Sesja nagraniowa składała się z trzech etapów. W pierwszym (trwającym około 40 minut) przeprowadzany był wywiad z każdym z uczestników, w ten sposób uzyskiwano zestaw standardowych metadanych (zgodnych z wytycznymi IMDI). Etap drugi to swobodna rozmowa na kilka z zaproponowanych tematów, trwająca do 90 minut; pozyskiwano w ten sposób zestaw jednostek leksykalnych. W ostatnim etapie informatorzy musieli „rozwiązać” 20 zadań elicytacyjnych — przewidziano na to około 120 minut, próbowano zebrać jak najwięcej zjawisk fonologicznych, morfologicznych i składniowych.

Lista poszczególnych zadań elicytacyjnych:

1. Znak migowy. Prezentacja swojego znaku migowego i wytłumaczenie, skąd się wziął.
2. Dowcip. Opowiedzenie wcześniej przygotowanego dowcipu.
3. Opowiadanie o swoich kontaktach z kulturą Głuchych (np. wspomnienia ze szkoły). Brak filmiku wprowadzającego.
- 4.1a. Opowiadanie obejrzanej historyjki obrazkowej *Frog, where are you?*
- 4.1b. Opowiadanie obejrzanego odcinka kreskówki *Tweety i Sylwester*.
- 4.2a. Opowiadanie obejrzanego filmiku *Pear story*.
- 4.2b. Opowiadanie obejrzanego skeczu *Haushaltshilfe* (niemiecki serial komediowy dla głuchych).
5. Zadanie kalendarzowe. Każdy z uczestników widzi wyświetlony plan tygodnia (różny dla każdej osoby). Zadaniem jest zaplanowanie spotkania w dogodnym terminie.

6. Dyskusja. Uczestnicy proszeni są o przedyskutowanie jednego z kilku kontrowersyjnych tematów dotyczących społeczności głuchych i języka migowego.
7. Swobodna rozmowa. Moderator opuszcza studio na ok. 15 minut. Uczestnicy mogą w tym czasie rozmawiać na dowolne tematy — kamery cały czas ich rejestrują.
8. Izolowane znaki. Wyświetlone są plansze z obrazkami i/lub niemieckim słowem. Informatorzy mają za zadanie zamigać dane pojęcie. Wywołany materiał ma służyć do badania wariantów w DGS.
9. Opowiadanie historyjki *Vater und Sohn*. Zadanie dodatkowe w przypadku zbyt dużej ilości pozostałego czasu.
10. Znaki. Uczestnikom prezentowane są różne znaki informacyjne. Między sobą muszą zdecydować, jakie jest ich znaczenie.
11. Ważne wydarzenia z przeszłości. Wyświetlane są ważne wydarzenia z przeszłości. Informatorzy mają za zadanie opowiedzieć partnerowi, co robili, gdy dane wydarzenie miało miejsce.
12. Dziedziny. Wyświetlane są nazwy różnych dziedzin życia (np. polityka) wraz z powiązanymi ilustracjami. Uczestnicy prezentują (dowolne) skojarzenia z danym tematem.
- 13.1. Instrukcja. Informatorzy proszeni są o zamiganie partnerowi szczegółowej instrukcji, jak należy wykonać daną czynność (np. ugotowanie potrawy).
- 13.2. Opowiedzenie historii *Travel Story*. Uczestnik ma za zadanie opowiedzenie fabuły historyjki o pewnej wycieczce.
14. Regionalne atrakcje. Uczestnicy są proszeni o opowiedzenie partnerowi o specjalnościach swojego regionu (potrawach, zabytkach, zwyczajach itp.).
15. Opowiadanie filmu *Signs*. Uczestnik ogląda krótki film specjalnie zmontowany na potrzeby projektu (oryginalnie jest to reklama napoju Schweppes), a następnie opowiada partnerowi jego akcję.
16. Nowe i stare znaki. Zadanie opcjonalne. Rozmowa na temat starych znaków, które wyszły już z użycia.
17. Wydarzenia. Rozmowa na temat „głuchych” wydarzeń (np. igrzyska Głuchych).
- 18.1. Opowiadanie historyjki *Fire alarm*. Uczestnik opowiada partnerowi obejrzaną historyjkę (zamiganą przez natywnego użytkownika DGS).
- 18.2. Mapa. Zadaniem jest opisanie partnerowi usytuowania różnych obiektów w mieście (na podstawie wyświetlonej mapy) i sposobu dotarcia do nich.

Niektóre zadania mają na celu wydobycie konkretnego materiału badawczego, np. wariantów leksykalnych. Inne są próbą udokumentowania kultury Głuchych. Większość punktów poprzedzona jest projekcją klipu w DGS, w którym opowiedziane jest, czego dotyczy następne zadanie. W przypadku wątpliwości pomocą służy moderator. Taka procedura eliminuje jakikolwiek wpływ języka fonicznego. Część materiału (np. historia o chłopcu i żabie czy *Pear story*) była wykorzystywana w innych projektach badawczych, co pozwala na analizy porównawcze. Wszystkie zadania przetestowano w fazie pilotażowej. Sprawdzano, jak głusi radzą sobie z kolejnymi punktami, ile trwa ich realizacja, czy udaje się wywołać zamierzony materiał, czy zadania są odpowiednio ciekawe. Sesja (łącznie z przerwami) trwa około 7 godzin. Materiał elicytacyjny musi być na tyle interesujący i różnorodny, by pomimo zmęczenia uczestników nadal skłaniał do dyskusji. Wyniki dotychczasowych nagrań potwierdzają, że dobrano bardzo dobry zestaw.

Anotacja zebranych nagrań przewidziana jest na około 10 lat. Do jej wykonania posłuży program iLex. Następować będzie w dwóch etapach. Na początku dokonana zostanie segmentacja na znaki oraz tłumaczenie na język niemiecki. Każdemu wystąpieniu znaku przypisany zostanie typ (*token-type matching*) — jest to opisany wcześniej proces głosowania. Podstawowy podział przewiduje trzy rodzaje jednostek: jednostki leksykalne, znaki produktywne (klasyfikatory) i inne. Za pomocą notacji HamNoSys przedstawiony będzie kształt (forma) migu — dopuszczalne jest przeszukiwanie po fragmencie zapisu HamNoSys. Anotować będą studenci znający język migowy, ich zapis sprawdzany będzie przez natywnych użytkowników. Jest to tylko pierwszy krok w transkrypcji. Zakłada użycie bardzo prostego podziału wszystkich znaków. Lematyzacja dokonywana jest na potrzeby słownikowe, dlatego nie do końca odpowiada lematyzacji w rozumieniu lingwistycznym. W drugim etapie dodany ma być szczegółowy opis glos, ruchu warg (ang. *mouthing*) oraz mimiki. W większym stopniu uwzględniony ma być podział na frazy i ich kontekst. Również cała kategoria „znaków produktywnych” (m.in. klasyfikatorów) doczekać ma się rozbudowanej analizy.

W przypadku korpusu DGS od początku akcentowana jest jego użyteczność dla społeczności Głuchych. Jest to ważne zagadnienie, ponieważ język migowy oprócz przedmiotu opisu językoznawczego stanowi także główny nośnik kultury Głuchych. Korpus języka migowego (inaczej niż dowolnego języka fonicznego) zawsze będzie w takim samym stopniu archiwum tekstów kultury, co próbek empirycznych. Sami niemieccy głusi uczestniczący w projekcie wyszczególnili cztery sposoby wykorzystywania korpusu w swojej społeczności:

- rekreacja. Korpus jako źródło inspirującego materiału — zadania dotyczące dowcipów, konwersacji na różne tematy, historii z życia głuchej społeczności.
- edukacja. Korpus jako zbiór regionalizmów, różnych rejestrów migania oraz stylów wypowiedzi. Badania gramatyczne i słownikowe. Doskonały materiał do nauczania i uczenia się DGS.
- dokumentacja. Korpus jako materiał dla historyków, antropologów i socjolingwistów. Wycinek kultury i historii Głuchych i ich języka.
- identyfikacja. Korpus jako potwierdzenie tożsamości Głuchych i roli języka migowego. Pomoc dla głuchych w określeniu swojej tożsamości.

Zgromadzone informacje potwierdzają, że korpus DGS jest największym tego typu przedsięwzięciem w historii lingwistyki migowej. Wykorzystano najnowsze osiągnięcia techniki (nagrania 3D) i zbudowano odpowiednią infrastrukturę techniczną (studio, oprogramowanie, przestrzeń dyskowa). Zaprojektowano i przetestowano bogaty zestaw elicytacyjny, którego efektem będzie ponad 400 godzin nagrań wideo. Godna podziwu jest dbałość o szczegóły zarówno na poziomie zbierania, jak i przetwarzania materiału. Mocno podkreślany jest też aspekt kulturowy. Dokonania Instytutu w Hamburgu kształtują standardy tworzenia korpusów migowych, przy okazji pokazując, jak wielkim wyzwaniem jest próba przeniesienia standardów lingwistyki korpusowej na grunt języka migowego.

#### 4.9. Dicta-Sign

Dicta-Sign nie jest zamierzeniem stricte korpusowym (Hanke i in., 2011). Znalazł się w zestawieniu z powodu swojej interdyscyplinarności i obrazowania nowych tendencji w lingwistyce migowej. Jest to trzyletni projekt badawczy dotowany z funduszy Unii Europejskiej, który rozpoczął się w styczniu 2009 roku. Jego celem jest poprawa komunikacji online głuchych użytkowników. Doprowadzić do tego ma rozwój aplikacji webowych korzystających z języka migowego. Realizatorzy projektu zwracają uwagę, że cała wymiana informacji w Internecie odbywa się za pomocą pisanej odmiany języka fonicznego, co dla głuchego internauty nie jest sposobem najwygodniejszym, dlatego zamierzeniem Dicta-Sign jest automatyczne rozpoznawanie naturalnego migania oraz generowanie właściwych odpowiedzi. Końcowy produkt umożliwiłby następujące działanie: głuchy miga pewien komunikat do kamery, komputer rozpoznaje zamigane frazy i zapisuje je w specjalnym formacie (odmiana języka znaczników XML), który jest

zrozumiały dla animowanego awatara. Nasuwa się słuszne zastrzeżenie, że taki sam efekt można uzyskać, umieszczając nagranie wideo danej wiadomości. Po raz kolejny pojawia się kwestia niemożności rozdzielania wizerunku nadawcy od jego komunikatu. Przy publikacji na forum tekstów w języku polskim z reguły nie widać, jak wygląda rozmówca. Drugim argumentem przemawiającym za wyższością awatara jest możliwość wykorzystywania wewnętrznej reprezentacji tekstu migowego w automatycznych translatorach międzymigowych (tj. tłumaczących między dwoma różnymi językami migowymi). Oprócz tego renderowana postać miga w sposób „przezroczysty”, nie posiada indywidualnego stylu migania, który czasami utrudnia zrozumienie wiadomości. Dicta-Sign rozbija się więc na kilka pomniejszych problemów badawczych, dwa podstawowe to automatyczne rozpoznawanie języka migowego i awatar.

Projekt jest prowadzony w kilku ośrodkach w Europie, a równoległe badania dotyczą aż czterech języków migowych: brytyjskiego, niemieckiego, greckiego i francuskiego. Co najbardziej interesujące, jednym z pierwszych kroków było stworzenie kilkujęzycznego korpusu tekstów dotyczących podróżowania w Europie. Dobór materiałów elicytacyjnych zakładał w miarę zamknięte odpowiedzi obfitujące w leksykę dotyczącą podróży i zwiedzania innych krajów oraz bardzo duże użycie relacji przestrzennych. Zbierano nagrania z czterech docelowych języków migowych (starano się unikać języka migowego). Bardzo często realizowano ten punkt niejako przy okazji tworzenia korpusu danego języka migowego (por. zadanie elicytacyjne 18.2 w korpusie DGS).

Wyposażenie studia nagraniowego jest podobne do zestawu hamburskiego — każdy informator ma swój własny monitor, na którym prezentowany jest materiał stymulujący. Oprócz rozmówców w nagraniu bierze udział moderator, którego rola sprowadza się do czuwania nad przebiegiem sesji i ewentualnej pomocy. Poza tym powinien być „niewidoczny”. Czas trwania sesji przewidziany był na maksymalnie dwie godziny (co przekłada się na ok. godzinę i 15 minut „czystego” migania), a kolejność zadań wyglądała następująco:

1. Rozgrzewka (5 min). Swobodna konwersacja, zapoznanie się.
2. Komunikacja miejska (10 min). Każdy z informatorów dostaje mapkę pokazującą schemat komunikacji miejskiej w pewnym mieście. Ich zadaniem jest zaproponowanie różnych dróg pomiędzy dwoma zadanymi punktami.
3. Biuro podróży (11 min). Informator A jest pracownikiem biura podróży, który musi przedstawić swojemu klientowi (informator B) dwa różne miejsca na spędzenie wakacji, następnie pyta o doko-

nany wybór. Różne pary miejsc były prezentowane w poszczególnych sesjach.

4. Planowanie wakacji (7 min). Zadanie negocjacyjne, w którym uczestnicy muszą ustalić, gdzie i na jak długo pojadą na wakacje oraz jakim środkiem transportu. Dla utrudnienia każdy informator dostaje kalendarz z zaplanowanymi wydarzeniami — rozmówcy muszą więc dodatkowo liczyć się z wolnymi terminami na urlop.
5. Na lotnisku (5 min). Informator B (patrz zadanie 3) opisuje rozmówcy krok po kroku, jak przebiegają procedury na lotnisku.
6. Mapa miasta (9 min). Wyświetlone zostają mapy miasta, informator A na swojej ma zaznaczoną pewną trasę oraz mijane obiekty (np. poczta, kościół, rynek). Musi opowiedzieć partnerowi, którą przebiega trasa i co jest po drodze. Zadaniem informatora B jest zaznaczenie tych obiektów na swojej mapie.

Przerwa (5 min).

7. „Zawiedzione oczekiwania” (12 min). Każdy z informatorów widzi slajd, na którym pokazane są oczekiwania pewnej osoby co do nadchodzącego wyjazdu oraz to, co zastała na miejscu. Pominięte są jakiegokolwiek teksty — użyto zdjęć i rysunków. Migający mają opowiedzieć podobną historię, która ich spotkała.
8. Podróżowanie kiedyś i dziś (11 min). Dyskusja na temat zmian w sposobach podróżowania. Materiałem stymulującym są kolaże zdjęć pokazujących dawne i współczesne środki transportu.
9. Opowiadanie historii (10 min). Każdy z uczestników opowiada obejrzaną historię — jedna to historyjka obrazkowa, druga to zamigana opowieść o przygodzie podczas ostatnich wakacji.
10. Słownictwo (12 min). Na zmianę prezentowane są pojedyncze pojęcia związane z podróżami (zdjęcie, rzadko tekst). Uczestnicy muszą zamigać odpowiedni znak. Jeśli druga osoba zna inny wariant, może podać swoją wersję. Kategorie pojęć to: daty (12 slajdów), pojazdy (8 slajdów), kraje (31 slajdów), pogoda (7 slajdów).
11. Własne doświadczenia (10 min). Uczestnicy mogą opowiedzieć o swoich doświadczeniach związanych z podróżowaniem. Jest to zadanie opcjonalne.
12. Ewaluacja (6 min). Informatorzy proszeni są o komentarz dotyczący zakończonej sesji.

Warto zwrócić uwagę, że cały materiał jest łatwo adaptowalny na potrzeby poszczególnych krajów biorących udział w projekcie. Zmian wymagają drobne napisy pojawiające się w zadaniach, zamigana historia z punktu 9. oraz ewentualnie niektóre zdjęcia (ze względów kulturowych).

Zebrany materiał ma pomóc w zrealizowaniu zamierzonych celów, jednocześnie powstają „produkty pośrednie”, np. narzędzia do półautomatycznego tagowania i anotacji.

Trzema końcowymi rozwiązaniami mają być:

- tłumacz pojęć technicznych pomiędzy językami migowymi (SL-SL translator);
- wyszukiwanie przez zamigany znak (użytkownik miga do kamery pewien znak, następuje automatyczne rozpoznanie przez komputer i wyświetlenie ze słownika/korpusu wszystkich wystąpień poszukiwanego migu, ewentualnie podobnych znaków);
- Wiki w języku migowym (stworzenie encyklopedii na licencji Wiki, której artykuły publikowane będą w języku migowym — wyżej wspomniana technologia animacji awatarowej).

W czasie pisania niniejszego tekstu projekt Dicta-Sign jeszcze trwał, dlatego trudno przewidzieć, czy uda się sprostać wszystkim założeniom. Upublicznione fragmenty rezultatów pozwalają wierzyć w powodzenie przedsięwzięcia. Chociaż zapewne nie uda się uzyskać zadowalającej automatyzacji na żadnym z poziomów (rzecz do tej pory niewykonalna nawet dla języków fonicznych), ale bardzo przydatne są liczne narzędzia i wypracowane standardy, które powstają przy okazji, a wykorzystywane mogą być w innych projektach.

## 5. Perspektywy

Migowa lingwistyka korpusowa dopiero wypracowuje standardy zbierania nagrań i tworzenia korpusu. O ile procedury przygotowywania korpusów tekstowych zostały już wielokrotnie opisane i przedstawiono szczegółowo ich standardy, o tyle dopiero od niedawna podejmuje się próby wprowadzania wytycznych dla języków migowych. Jest to związane z historią — porównanie dotyczy ponad pięćdziesięciu lat korpusologii fonicznej (zakładając, że korpus Browna z 1961 roku jest pierwszy) z mniej niż dziesięcioma latami migowej (za pierwszy korpus przyjmijmy rezultaty projektu ECHO z roku 2003). Wychodząc naprzeciw temu zagadnieniu, powstał trzyletni projekt Sign Linguistics Corpora Network (SLCN) (Crasborn, 2010), którego zadaniem było zorganizowanie czterech warsztatów dotyczących kolejno:

- zbierania materiału (Londyn 2009),
- metadanych (Nijmegen 2009),



- anotacji (Sztokholm 2010),
- eksploatacji (Berlin 2010).

W prezentacji otwierającej pierwszą konferencję jeden z koordynatorów całego projektu przedstawił *explicite* cele działania w postaci pytań: Czego oczekujemy jako lingwiści? Jak współpracujemy z innymi badaczami (również pomiędzy dziedzinami)? Jakie są potrzeby technologiczne i jak możemy korzystać z dokonań innych dyscyplin? Podczas tego warsztatu omówiono procedury zbierania materiału do korpusów takich języków jak: auslan, NGT, BSL, DGS, LSF (franc. *langue des signes française* — francuski język migowy), jamajski migowy. Zapowiedziano także m.in. projekt stworzenia korpusu ASL (ang. *American Sign Language*, tj. amerykańskiego języka migowego).

Dodatkowe prezentacje dotyczyły zagadnień lingwistyki stosowanej, licencji Creative Commons oraz przetwarzania komputerowego.

Podczas drugiego spotkania zaprezentowano różne sposoby podejścia do tematu metadanych i metody radzenia sobie z nimi w różnych projektach. Opisane zostały głównie trzy rozwiązania: edytor IMDI (zastosowany w projekcie ECHO) i jego propozycja rozszerzenia pod kątem języków migowych, sposób organizacji metadanych w programie iLex oraz edytor ARBIL. Wszyscy uczestnicy podkreślali, że mimo nowych możliwości nie może istnieć korpus bez metadanych i odpowiedniej ich organizacji.

Temat anotacji był motywem przewodnim spotkania trzeciego. Używanie spójnego systemu opisywania nagrań jest niezmiernie ważne. Wypracowane metody z różnych projektów pozwalają zaoszczędzić wiele czasu i trudu oraz przekładają się na jakość produktu, jakim jest korpus. W perspektywie współpracy z innymi ośrodkami zachowanie pewnych dobrych wzorców anotacji umożliwi podjęcie wspólnych badań komparatystycznych. Omawiano również trudne sytuacje, w których możliwe są różne interpretacje tekstu migowego — najlepszym przykładem są znaki o wysokim stopniu ikoniczności.

Podczas warsztatu w Berlinie dokonano przeglądu zagadnień dotyczących metod archiwizowania nagrań, ich upubliczniania, a przede wszystkim możliwości wykorzystywania. Nie można marginalizować znaczenia aspektu kulturowego języka migowego. Żaden korpus języka migowego nigdy nie będzie tylko przedsięwzięciem lingwistycznym, lecz zawsze będzie też świadectwem kultury Głuchych, do którego będą oni chcieli mieć dostęp. Nie powinno się również zapominać o wykorzystywaniu korpusu w nauczaniu języka migowego jako języka obcego (L2). Architektura korpusu powinna pozwalać na wykorzystanie go w jak największej liczbie dziedzin.

Oprócz cyklu spotkań wynikiem projektu Sign Linguistics Corpora Network jest stworzenie strony na zasadach Wiki, na której każdy z partne-

rów może umieszczać i edytować artykuły poruszające tematykę korpusów języka migowego i pozostałych zagadnień związanych z głuchymi, ich językiem i kulturą.

Próby wypracowywania standardów i powstawanie korpusów wielojęzycznych (np. projekt Dicta-Sign) są jednym z bardzo silnych trendów w migowej korpusologii. Ze względu na duże podobieństwo języków migowych wynikające ze wspólnych cech gramatyki wizualno-przestrzennej dąży się do budowania korpusów, łatwo porównywalnych między sobą. Dzięki temu możliwe będzie uchwycenie uniwersaliów migowych i pełniejsza analiza gramatyczna. Równolegle, dynamicznie rozwijają się projekty automatyzacji w przetwarzaniu języka migowego. Od kilku lat trwają próby rozpoznawania języka migowego przez komputer — to problem z pogranicza kilku dziedzin (technologia 3D, *motion capture*, rozpoznawanie wideo, sztuczna inteligencja). Obecnie przy tworzeniu korpusu trzeba mieć na uwadze możliwość wykorzystania materiału do takich właśnie potrzeb, dlatego wskazane jest dokonywanie nagrań w odpowiednio wysokiej jakości. Nieustanny rozwój techniki powoduje, że być może i w tej dziedzinie niedługo dokona się przełom. Rewolucją, której rezultaty już widać, jest ekspansja filmów 3D. W ośrodku hamburskim do zestawu kamer dołączono dodatkową parę kamer stereoskopowych. Obniżanie się cen kamer cyfrowych i dostępność coraz bardziej zaawansowanych zestawów pozwala nawet w małych projektach na bardzo wysoką jakość materiału. Wniosek nie może być inny niż w przypadku korpusów tekstowych — wszystko wskazuje na nadchodzący rozkwit lingwistyki migowej w jej korpusowym wydaniu.

## Bibliografia

- Blanck, Dolly, Hofmann, Ilona, Jeziorski, Olga, König, Susanne, Langer, Gabriele, Rathmann, Christian (2010). *Uses of the DGS Corpus from a Deaf community perspective* (poster), 4th SLCN-Workshop (Exploitation). Berlin.
- Crasborn, Onno (2010). The Sign Linguistics Corpora Network: Towards standards for signed language resources. W N. Calzolari i in. (red.), *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, ss. 457–460. Paryż: ELRA.
- Crasborn, Onno, Mesch, Johanna, Waters, Dafydd, Nonhebel, Annika, van der Kooij, Els, Woll, Bencie, Bergman, Brita (2007). Sharing sign language data online. Experiences from the ECHO project. *International Journal of Corpus Linguistics* 12(4), ss. 537–564.
- Crasborn, Onno, Sloetjes, Han (2008). *Enhanced ELAN functionality for sign language corpora*. W O. Crasborn i in. (red.), *Proceedings of the 3rd Workshop on the Representation and Processing of Signed Languages: Construction and exploitation of sign language corpora*.

- International Conference on Language Resources and Evaluation*, ss. 39–43. Paryż: ELRA.
- Crasborn, Onno, Zwitserlood, Inge (2008). The Corpus NGT: an online corpus for professionals and laymen. W O. Crasborn i in. (red.), *Proceedings of the 3rd Workshop on the Representation and Processing of Signed Languages: Construction and exploitation of sign language corpora. International Conference on Language Resources and Evaluation*, ss. 44–49. Paryż: ELRA.
- Grucza, Sambor (2007). O konieczności tworzenia korpusów tekstów specjalistycznych. W S. Grucza (red.), *W kręgu teorii i praktyki lingwistycznej*, ss. 103–122. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego.
- Hanke, Thomas, König, Lutz, Wagner, Sven, Matthes, Silke (2010). DGS Corpus & Dicta-Sign: The Hamburg studio setup. W P. Dreuw i in. (red.), *LREC 2010. 7th International Conference on Language Resources and Evaluation. Workshop proceedings. W13. 4th Workshop on Representation and Processing of Sign Languages: Corpora and Sign Language Technologies. May 22/23 2010. Valletta — Malta*, ss. 106–109. Paryż: ELRA.
- Hanke, Thomas, Bowden, Richard, Braffort, Annelies, Coller, Christophe, Efthimiou, Eleni (2011). *Dicta-Sign — Use of a multilingual corpus to improve sign language technology* (poster), 4th International Workshop on Sign Language Translation and Avatar Technology (SLTAT). Berlin.
- Johnston, Trevor, Schembri, Adam (2006). *Issues In the creation of a digital archive of a signed language*. W L. Barwick, N. Thieberger (red.), *Sustainable data from digital fieldwork*, ss. 7–16. Sydney: University of Sydney Press.
- Johnston Trevor (2009). Creating a corpus of Auslan within an Australian National Corpus. W M. Haugh i in. (red.), *Selected proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering languages*, ss. 87–95. Sommerville, MA: Cascadilla Proceedings Project.
- Leech, Geoffrey (1992). Corpora and theories of linguistic performance. W J. Svartvik (red.), *Directions in corpus linguistics*, ss. 112–113. Berlin: Walter de Gruyter.
- Nishio, Rie, Hong, Sung-Eun, König, Susanne, Konrad, Reiner, Langer, Gabriele, Hanke, Thomas, Rathmann, Christian (2010). *Elicitation methods in the DGS (German Sign Language) Corpus Project*. W P. Dreuw i in. (red.), *LREC 2010. 7th International Conference on Language Resources and Evaluation. Workshop proceedings. W13. 4th Workshop on Representation and Processing of Sign Languages: Corpora and Sign Language Technologies. May 22/23 2010. Valletta — Malta*, ss. 178–185. Paryż: ELRA.
- Prillwitz, Siegmund, Hanke, Thomas, König, Susanne, Konrad, Reiner, Langer, Gabriele, Schwarz, Arvid (2008). DGS Corpus Project — Development of a Corpus Based Electronic Dictionary German Sign Language. W O. Crasborn i in. (red.), *Proceedings of the 3rd Workshop on the Representation and Processing of Signed Languages: Construction and exploitation of sign language corpora. International Conference on Language Resources and Evaluation*, ss. 159–164. Paryż: ELRA.
- Rudolf, Michał (2004). *Metody automatycznej analizy korpusów tekstów polskich*. Warszawa: Wydział Polonistyki.
- Schembri, Adam (2008). The British Sign Language Corpus project: Open access archives and the observer's paradox. W O. Crasborn i in. (red.), *Proceedings of the 3rd Workshop on the Representation and Processing of Signed Languages: Construction and exploitation of sign language corpora. International Conference on Language Resources and Evaluation*, ss. 165–169. Paryż: ELRA.

- Świdziński, Marek, Mikulska, Danuta (2003). Reprezentacja linearna tekstu Polskiego Języka Migowego. W M. Świdziński, T. Gałkowski (red.), *Studia nad kompetencją językową i komunikacyjną niesłyszących*, ss. 31–58. Warszawa: Uniwersytet Warszawski: Wydział Polonistyki, Wydział Psychologii, Polski Komitet Audiofonologii, Instytut Głuchoniemych im. ks. Jakuba Falkowskiego.
- Świdziński, Marek (2006). Lingwistyka korpusowa w Polsce — źródła, stan, perspektywy, *LingVaria 1*, ss. 23–32.