Marek Świdziński
(Uniwersytet Warszawski,
e-mail: m.r.swidzinski@uw.edu.pl)
ORCID: 0000-0001-6518-2376

DOI: 10.33896/PorJ.2022.3.1

Paweł Rutkowski
(Uniwersytet Warszawski,
e-mail: p.rutkowski@uw.edu.pl)
ORCID: 0000-0001-7967-6959

KORPUS OGÓLNY JAKO MODEL DANEGO JĘZYKA NATURALNEGO: KORPUSY JĘZYKÓW FONICZNYCH A KORPUS POLSKIEGO JĘZYKA MIGOWEGO

WSTĘP

Celem niniejszego artykułu jest spojrzenie na Korpus Polskiego Języka Migowego (KPJM) z perspektywy teorii języka i współczesnej lingwistyki korpusowej. KPJM jest rezultatem przedsięwzięcia naukowego realizowanego od dekady w Pracowni Lingwistyki Migowej na Wydziale Polonistyki Uniwersytetu Warszawskiego [Rutkowski i in. 2013, 2014, 2017].[1] Pokażemy, czym są korpusy we współczesnej lingwistyce – jak są konstruowane, czemu i komu służą, jak się z nich korzysta. Przedstawimy tu pewien szczególny typ korpusu: korpus ogólny. Przyjrzymy się także najważniejszym podobieństwom i różnicom między korpusem języka wizualno-przestrzennego a korpusami języków fonicznych.

KORPUS JAKO ŹRÓDŁO EMPIRYCZNE

Język naturalny to niezwykła umiejętność ludzka: maszyna wytwarzania i odbioru komunikatów, narzędzie wszechstronnego porozumiewania się wewnątrz populacji. Strukturalista postrzega język jako parę <Słownik, Gramatyka>, dwa składniki kompetencji językowej. Odpowiada temu zdroworozsądkowa intuicja „słówek" i „regułek" w nauce języka obcego. Wytwory tego urządzenia komunikacyjnego nazywamy

[1] W latach 2014–2020 prace nad KPJM finansowane były ze środków Narodowego Programu Rozwoju Humanistyki Ministerstwa Nauki i Szkolnictwa Wyższego (tytuł projektu: „Wielopoziomowa anotacja lingwistyczna korpusu polskiego języka migowego (PJM)", partner zagraniczny: Trevor Johnston z Uniwersytetu Macquariego – Sydney, Australia).

tekstami. Lingwista opisujący język, nie mogąc zajrzeć do umysłu użytkownika, zwykle opisuje teksty, czyli wytwory nadawcy komunikatu – pisane, mówione czy migane.

Przez korpus rozumie się zbiór tekstów – zbiór zdań dowolnego języka zbudowany z rozmysłem, dla rozwiązania jakiegoś zadania badawczego.[2] Nie uznamy za korpus, powiedzmy, zawartości Biblioteki Kongresu USA czy jakiegoś serwera pocztowego. Korpus to źródło danych dla językoznawcy, który jest empirystą jak biolog, fizyk czy chemik. Rozważając problem – na przykład wyrażeń przyimkowych, nazw własnych, struktur z bezokolicznikiem itp., gromadzi materiał, sporządza sobie listę jednostek tekstowych, które chce objąć oglądem, i zapisuje je z kontekstem. Takie lokalne korpusy nierzadko służą wykonaniu jednego zadania, a po osiągnięciu tego celu nie są już wykorzystywane – tak jak próbka krwi pacjenta utylizowana po jej zbadaniu.

Ale istnieją też korpusy nielokalne, mające ambicję reprezentowania czegoś; w szczególności – *całego* języka. Są to korpusy ogólne. Korpusy takie budowali od dawna słownikarze, którym miały one dostarczać danych leksykalnych, a także ilustracji do artykułów hasłowych. Czasem ujawniali oni swój korpus, podając na przykład listę książek wykorzystanych jako podstawa tworzonego słownika. Taki korpus leksykograficzny mógł nie istnieć fizycznie.

Schyłek XX wieku przyniósł światu całą serię rewolucji technicznych. Najważniejsza z nich, i to w aspekcie cywilizacyjnym, to ta informatyczna – rozpowszechnienie komputerów osobistych, gigantyczne pamięci masowe oraz Internet. Teksty funkcjonują dziś w postaci cyfrowej, tak są przechowywane i tak są dostępne powszechnie, jeśli nie liczyć pewnych ograniczeń technicznych lub prawnych. Nastała era lingwistyki informatycznej, tworzenia i obsługi narzędzi dostępu do tekstu – narzędzi dla każdego, nie akurat dla naukowca. Nic dziwnego, że obserwujemy istną eksplozję przedsięwzięć korpusowych, które wobec braku limitów wielkości mogą być duże i bardzo duże. Stanowią one podstawę opisu lingwistycznego całego języka.

MODELOWANIE W NAUKACH EMPIRYCZNYCH

Uczony empirysta zdaje sprawę z różnych wycinków rzeczywistości, pokazuje, jak wyglądają i funkcjonują. Mówimy, że buduje model. Model to coś, co przypomina obiekt modelowany; coś, co jest od oryginału prostsze, wyrazistsze, łatwiejsze do zrozumienia. Model nie musi być czymś materialnym, jak kolejka elektryczna w pokoju dziecięcym, która odwzorowuje „prawdziwe" pociągi: bywa abstrakcyjną teorią albo

[2] Dobrym wprowadzeniem w świat lingwistyki korpusowej jest rozprawa M. Rudolfa [2004].

rachunkiem. Dla danego fragmentu świata można stworzyć nieograniczoną liczbę modeli.

Lingwista sporządza opis języka naturalnego na dwa sposoby. Jednym jest rekonstrukcja kompetencji językowej rodzimego użytkownika – nadawcy i odbiorcy: zbudowanie słownika i podanie instrukcji wytwarzania i rozbioru wyrażeń, czyli gramatyki. Jest to model użytkownika; dokładniej – reprezentacja sumy indywidualnych kompetencji. Język polski więc na przykład to ten, którego użytkownik posługuje się *takim* słownikiem i *taką* gramatyką, jakie zostały tu właśnie wykoncypowane przez lingwistę. Ten typ modelu zaproponowali europejscy strukturaliści w pierwszej połowie XX wieku.

Drugi typ to prezentacja wytworów działań użytkownika, czyli zbioru tekstów – produkowanych przezeń lub odbieranych. Na pytanie „Co to jest język X?" – na przykład polski – da się odpowiedzieć: „To zbiór wszystkich możliwych zdań języka X (polskiego) i tylko ten zbiór" i podeprzeć się pokazem polskiego korpusu. Powiemy o nim, że *reprezentuje* polszczyznę, tak jak uznamy na przykład grono respondentów ankiety preferencji wyborczych za reprezentatywne dla całej populacji. Ów drugi rodzaj modelu ujawnia nie tylko cechy jakościowe danego języka – jego leksykon i gramatykę, ale i własności kwantytatywne. Ten typ modelu wdrożyli przed stuleciem strukturaliści amerykańscy (dystrybucjoniści), podejmując trud opisywania umierających języków Indian Ameryki Północnej. Dla nich zbiór zapisanych tekstów takiego języka był jakby całym tym językiem. Podobnie traktuje się zasoby wszystkich zachowanych tekstów wymarłego języka, takiego jak sanskryt, staro-cerkiewno-słowiański czy staroislandzki. Interesujące, że korpusowe spojrzenia na język naturalny okazało się bardzo płodne, tak jakby dystrybucjoniści przewidzieli powstanie lingwistyki informatycznej.

Konstruując model pierwszego rodzaju, korzysta się zwykle z danych drugiego modelu – czyli z korpusu. Są językoznawcy, którzy pracują jawnie bez korpusu, ograniczając podstawę materiałową swych badań do intuicji rodzimego użytkownika języka – na przykład Noam Chomsky. Jednak w ostatnich latach nawet wśród badaczy odwołujących się do tradycji generatywizmu obserwujemy coraz większe zainteresowanie uzupełnianiem i weryfikowaniem intuicji pojedynczego użytkownika języka intersubiektywnym materiałem korpusowym [por. np. Oliviéri 2010]. Zauważmy, że w procesie naturalnego nabywania języka (*language acquisition*) obserwacja tekstów stanowi podstawę, aktywizuje, jak powiada N. Chomsky, wrodzoną gramatykę uniwersalną. Tak samo dzieje się przy opanowywaniu języka obcego.

Korpus ogólny jest więc modelem danego języka naturalnego: to suma produkcji rodzimych użytkowników. Mówi się o korpusie tego typu, że jest narodowy i ma to sens, choć bynajmniej nie w aspekcie nacjonalistycznym. Inna etykietka to *korpus referencyjny*. Ten drugi termin jest zdecydowanie lepszym wyborem w odniesieniu do KPJM, czyli korpusu

tekstów polskich Głuchych[3] – społeczności, która stanowi mniejszość językową, nie będąc jednakże mniejszością etniczną.

POLSKI JĘZYK MIGOWY

Polski język migowy (PJM) jest językiem pierwszym dla szacowanej na kilkadziesiąt tysięcy populacji niesłyszących Polaków. To język rodzimy czy macierzyński kilkuprocentowego podzbioru tej społeczności – potomków głuchych rodziców, Głuchych dynastycznych; język wyniesiony przez nich z domu. W mniejszym lub większym stopniu opanowują go również głuche dzieci rodziców słyszących – jeśli zetkną się z miganiem jeszcze w dzieciństwie, osiągają biegłość językową w zasadzie nieodróżnialną od rodzimej.

PJM nie ma nic wspólnego z polszczyzną. Reprezentuje specjalną klasę języków naturalnych. Jest to język wizualno-przestrzenny, nie wokalno-audytywny jak polszczyzna i tysiące innych języków fonicznych [por. Świdziński 2005; Czajkowska-Kisil 2014]. Języki migowe i foniczne mają takie same cechy funkcjonalne: stanowią narzędzie komunikacji uniwersalnej, nadawca i odbiorca wytwarzają i odczytują teksty. Tekst wizualno-przestrzenny przypomina tekst mówiony – trwa w czasie rzeczywistym, jest komponentem jednostkowego aktu mowy. Nie istnieje inna niż nagranie wideo postać utrwalona tekstu migowego. Tekst migowy jest trójwymiarowy – choć sekwencyjny (znaki wytwarzane są jeden po drugim), to często wykorzystujący też przekaz symultaniczny i łączenie kilku niezależnych artykulacji: bywa, że odrębne informacje przekazują ręka dominująca (dla osób praworęcznych – prawa), niedominująca, mimika, kierunek wzroku, wychylenie ciała, ruchy głową [Łozińska 2014b]. Tekst migowy jest wreszcie dynamiczny – z ruchem jako fundamentalnym składnikiem kształtu. Gramatyka PJM, przede wszystkim – składnia, jest zdecydowanie ikoniczna [Świdziński, Rutkowski 2014]. Tekst migowy wykorzystuje elementy świata jako składniki wyrażeń, co w językach fonicznych jest skrajnie rzadkie.

Ze względu na to, że PJM jest językiem relatywnie młodym, do niedawna nie był właściwie wykorzystywany w sytuacjach oficjalnych i w zorganizowanym nauczaniu. Populacja jego użytkowników jest dia-

[3] Rzeczownik *Głuchy* – analogicznie do słów takich jak *Kaszub* czy *Francuz* – zapisujemy wielką literą, mając na myśli członka wspólnoty, której fundamentem i spoiwem jest posługiwanie się tym samym językiem (w tym wypadku – językiem migowym). Jest to konwencja stosowana często w literaturze przedmiotu i promowana przez same środowiska migające. Kiedy nie jest istotne, czy osoby z ubytkiem słuchu, o których mowa, czują się członkami mniejszości językowo-kulturowej (a nawet – czy w ogóle znają język migowy), rzeczowniki *głuchy* lub *niesłyszący* zapisujemy w niniejszym artykule małą literą. Co oczywiste, we wszystkich użyciach przymiotnikowych stosujemy jedynie małą literę.

sporą rozproszoną [Świdziński 2015], nie ma on jednolitej normy. Mówiąc z pewną przesadą: PJM stanowi sumę idiolektów. Stąd jego opis lingwistyczny musi zdawać sprawę z wieloaspektowej różnorodności. Aby stworzyć reprezentatywny model struktury leksykalno-gramatycznej stojącej za komunikacją osób głuchych, coraz więcej badaczy języków migowych sięga obecnie po dane korpusowe [por. Fabisiak 2010; Rutkowski, Łozińska 2014].

REPREZENTATYWNOŚĆ I ZRÓWNOWAŻENIE

Wiele języków fonicznych doczekało się w ostatnich dekadach powstania korpusów referencyjnych. Mają takie korpusy język angielski (brytyjski i amerykański), francuski, niemiecki, czeski czy słoweński. W Polsce funkcjonuje Narodowy Korpus Języka Polskiego (NKJP), stanowiący wynik wspólnej inicjatywy Instytutu Podstaw Informatyki PAN, Instytutu Języka Polskiego PAN, Wydawnictwa Naukowego PWN oraz Zakładu Językoznawstwa Komputerowego i Korpusowego Uniwersytetu Łódzkiego, swobodnie dostępny sieciowo pod adresem nkjp.pl [por. Przepiórkowski i in. 2012]. NKJP jest dobrym modelem współczesnej polszczyzny. Przemyślana selekcja tekstów, podstawy gramatyczne i narzędzia informatyczne czynią zeń cenne narzędzie w najrozmaitszych przedsięwzięciach lingwistycznych. Jest on bardzo duży: w roku 2011 długość jego mierzona w segmentach liczyła 1,2 miliarda (odpowiadało to 800 mln słów); dziś jest to 1,8 miliarda segmentów. W tym świetle KPJM wydawać się może korpusem niewielkim. Obecnie składa się nań ok. 700 000 segmentów (które są wystąpieniami ok. 16 000 leksemów). Warto jednak zwrócić uwagę, że mimo iż KPJM – tak jak NKJP – pomyślany był jako korpus referencyjny, ze względu na proces powstawania porównywany powinien być nie tyle do korpusów języka pisanego, ile do korpusów języka mówionego.

Aby pewien zbiór tekstów okazał się dobrym modelem danego języka, nie może być przypadkowy. Korpus wymaga odpowiedniego zaprojektowania. Jest reprezentatywny, kiedy ujmuje język w całej różnorodności, włączając wszystkie odmiany danego języka i wszystkie typy tekstów; zrównoważony – kiedy każda odmiana i każdy typ tekstu ma w nim stosowny udział. Sformułowania powyższe brzmią, przyznać trzeba, mocno ogólnikowo. Zakładają (nieprawdziwie) istnienie jakichś klasyfikacji odmian i typów tekstów, z których każda może budzić obiekcje. Podobnie decyzje na temat ilościowego udziału w korpusie danej odmiany czy typu tekstu zależą od tylu różnych czynników, zarówno lingwistycznych, jak i pozajęzykowych, że muszą być arbitralne. W każdym razie lepsze są jakieś ograniczenia niż rozbudowa ekstensywna. Należy przy tym pamiętać, że różne korpusy bywają różnie tworzone.

Dwie zasadnicze odmiany sporej części języków naturalnych to mowa i pismo[4] (pomińmy tu problem innych zróżnicowań – regionalnych, chronologicznych, stylistycznych itp.). Niejeden lingwista teoretyk powie, że język mówiony i język pisany są odrębnymi językami, funkcjonującymi w różnych modalnościach, a więc mającymi drastycznie odmienną postać fizyczną tekstu. Tekst mówiony jest ulotny, jednostkowy, niepowtarzalny; tekst pisany – utrwalony. Odmiana mówiona jest prymarna zarówno przyrodniczo, jak i kulturowo: więcej mówimy, niż piszemy, więcej słuchamy, niż czytamy. Z całą jednak pewnością korpus referencyjny powinien obejmować obydwie odmiany, jeśli istnieją.

Korpusy ogólne języków fonicznych (w rodzaju NKJP) są jednak przede wszystkim zbiorami tekstów pisanych. Sporządzenie zbioru tekstów mówionych, tak samo jak ich obsługa, to zadania bardzo trudne ze względu na koszt i czas realizacji [por. Pęzik 2012a]. Typowa praktyka jest taka, że korpusy ogólne wykorzystują głównie prasę, z tekstami informacyjnymi i publicystyką (ok. 50%), i książki, czyli prozę, beletrystyczną oraz niefabularną (ok. 25%). Teksty mówione są tylko markowane; włącza się je już to w postaci pozyskiwanych skądś stenogramów, już to jako zapisy rzeczywistych nagrań audio; stanowi to zwykle ok. 10% zawartości korpusu. Można powiedzieć, że reprezentatywność i zróżnicowanie korpusu pisanego dla całego języka budzi wątpliwości. Dlatego też, jak zauważa P. Pęzik [2012a], twórcy NKJP uznali uwzględnienie jak największej liczby próbek języka mówionego (a w szczególności języka potocznej konwersacji) za jeden z kroków do zwiększenia reprezentatywności tego zasobu (komponent konwersacyjny NKJP jest obecnie największym korpusem polszczyzny mówionej).

W wypadku KPJM cały korpus – nie zaś jego niewielką część – należy uznać za zbiór danych konwersacyjnych. W przeciwieństwie do NKJP zawiera on materiał językowy, który był rejestrowany specjalnie na potrzeby projektu korpusowego. Jedyna możliwa postać fizyczna zapisu komunikacji migowej to klip wideo rejestrujący cały akt mowy, z jednym nadawcą i jednym odbiorcą. Na etapie analizy tekst migowy musi być jakby wypreparowany z całego aktu mowy. Nie istnieją wystarczająco obszerne zasoby odpowiednio nagranych i wykadrowanych klipów, z których można by stworzyć korpus PJM. Nie ma też mowy o nagrywaniu przypadkowych głuchych migających na ulicy. Korpus trzeba tworzyć od zera, a zapis aktu mowy musi być dokonywany z kilku perspektyw, co oznacza niezbędność zaawansowanego studia nagrań. Budowanie KPJM stanowiło ogromne przedsięwzięcie eksperymentalne, teoretyczne i inżynierskie.

[4] Warto w tym miejscu odnotować, że według najnowszej edycji bazy *Ethnologue* spośród 7139 opisanych tam języków naturalnych używanych obecnie na świecie tylko 4065 rozwinęło system zapisu [Eberhard, Simons, Fennig 2021].

W wielogodzinnych sesjach nagraniowych wzięło łącznie udział 150 Głuchych z całej Polski, reprezentujących wszystkie województwa i następujące grupy wiekowe: 18–30 lat – 15 kobiet i 13 mężczyzn, 31–40 lat – 17 kobiet i 14 mężczyzn, 41–50 lat – 12 kobiet i 22 mężczyzn, 51–60 lat – 17 kobiet i 10 mężczyzn, 61–92 lata – 14 kobiet i 16 mężczyzn. Budujący korpusy pisane wkładają wiele trudu, by zapewnić bogactwo odmian i typów włączanych tekstów. Aby zachować dane nielingwistyczne, wszywa się w teksty korpusu rozmaite metadane – informacje o źródle, autorze, wydawcy, dacie publikacji, typie publikacji itp. Analogicznie postąpili twórcy KPJM. W każdej sesji nagraniowej uczestniczyły dwie osoby używające od dzieciństwa PJM. Osobom tym przedstawiano na ekranach zadania do wykonania i tematy do przedyskutowania. Ze względu na bardzo zróżnicowany poziom polszczyzny wśród głuchych informatorów, a także z obawy przed interferencjami międzyjęzykowymi, podstawowym założeniem było, że uczestnicy nie posługują się żadnym innym językiem poza migowym; stąd wśród owych zadań nie znalazły się na przykład polecenia wymagające tłumaczenia polskiego tekstu pisanego. Takie „poddanie tematów" do rozmowy nazywane jest w lingwistyce migowej elicytacją. Materiały elicytacyjne opracowane zostały w taki sposób, by zebrane dane umożliwiały analizę wielu różnych aspektów słownika i gramatyki PJM, a także by mogły być punktem wyjścia badań porównawczych (część zadań została zaczerpnięta z projektów korpusowych z innych krajów) [zob. Rutkowski i in. 2013, 2014, 2017]. Każda sesja nagraniowa była *de facto* kilkugodzinną rozmową. Filmy nagrane na potrzeby KPJM można obejrzeć w *Otwartym Repozytorium Korpusu Polskiego Języka Migowego* [Wójcicka i in. 2020].

Konwersacyjny charakter danych zebranych w KPJM wpływa na ich właściwości tekstowe, gramatyczne i leksykalne. Korpusowe teksty migowe można scharakteryzować dokładnie tak samo, jak P. Pęzik [2012a] opisuje polszczyznę mówioną:

> Nie ulega wątpliwości, że autentyczny język konwersacyjny istotnie różni się od innych odmian polszczyzny. Z punktu widzenia badacza, dyskurs konwersacyjny cechuje pewna nieprzewidywalność, która częściowo wynika z przebiegu samej rozmowy, a częściowo także z okoliczności, w których jest ona prowadzona, i zdarzeń, które jej towarzyszą. Zamiast kompletnych składniowo wypowiedzi, które znamy z języka pisanego, w języku konwersacyjnym często mamy do czynienia z aproksymacją komunikowanych znaczeń (Lewandowska-Tomaszczyk 2012). Trudno jest w autentycznej polszczyźnie konwersacyjnej wskazać granice zdań składniowych, a w warstwie leksykalnej szczególnie często występują wyrazy, frazy i zbitki leksykalne, które pełnią funkcję dyskursywną i ułatwiają budowanie wypowiedzi ustnych w czasie rzeczywistym [2012a, 47].

Konwersacyjność tekstu przekłada się też m.in. na cechy frekwencyjne poszczególnych leksemów. Jak pokazuje P. Pęzik [2012a, 38-39], to, które wyrazy pojawiają się w polskich tekstach najczęściej, zależy w dużej mierze od odmiany języka. W tekstach pisanych najczęstszy jest

przyimek *w*, a w podkorpusie mówionym szczyt listy frekwencyjnej wygląda tak:

Wyraz	**Częstość**
to	86 960
nie	82 714
no	61 964
i	54 915
w	47 056

Wyrazy pełniące wielorakie funkcje dyskursywne są w mowie używane znacznie częściej niż w piśmie. Współbrzmi to ciekawie z obserwacjami dotyczącymi list frekwencyjnych KPJM. Jak pokazują P. Rutkowski i in. [2021], aż 14,68% wszystkich segmentów tekstu migowego wyodrębnionych w KPJM stanowią znaki wskazujące, które pełnią wiele funkcji dyskursywnych i wiążą się ściśle z kontekstem danego aktu mowy [por. Rutkowski, Czajkowska-Kisil 2010]. Jedynie 70,17% tekstu migowego to znaki leksykalne, czyli elementy, które można porównać do polskich rzeczowników, przymiotników, czasowników czy przysłówków, a na pozostałe 15,15% składają się znaki niemające oczywistych odpowiedników w językach fonicznych, m.in. konstrukcje klasyfikatorowe [por. Rutkowski, Łozińska 2011; Dziewanowska 2022], literowanie słów polskich za pomocą znaków daktylograficznych, nieznakowe gesty fatyczne. Nawet najczęstsze znaki leksykalne są o wiele rzadsze od wskazań – jak pokazujemy poniżej, żaden nie zbliża się nawet do 1% wszystkich segmentów w KPJM [za: Rutkowski i in. 2021].

Znak	**Odsetek wszystkich segmentów w KPJM**
MIGAĆ/MIGOWY	0,8%
TAK-SAMO/TEŻ	0,7%
JUŻ	0,7%
GŁUCHY	0,5%
WIDZIEĆ/PATRZEĆ	0,5%
DOBRZE	0,5%
MĘŻCZYZNA/CHŁOPAK	0,5%
WIEDZIEĆ/UMIEĆ	0,4%

DOSTĘP: KORPUS ANOTOWANY

Korpus jest użyteczny, jeśli zapewni się możliwość dotarcia do tego, co interesuje użytkownika. W czasach dawniejszych, przedkomputerowych, przeszukiwało się materiał ręcznie. Słownikarz sporządzał fiszki, porządkował je i przechowywał jako kartotekę. Współcześni leksykografowie konstruują dla swoich potrzeb konkordancje, czyli listy słów

z kontekstem (KWIC – *KeyWord In Context*). Są to techniki wizualizacji wyników przeszukiwania po kształtach. Zna to dobrze współczesny użytkownik komercyjnych programów edycyjnych czy internetowych wyszukiwarek. Interesujące, że osiem wieków temu, w roku 1230, powstała pierwsza konkordancja *Biblii* łacińskiej – spis imion proroków z odwołaniem do stosownych miejsc w tekście *Wulgaty*.

Ale kwerenda korpusu po kształtach to za mało. Napisy wchodzące w jego skład powinny być wzbogacone o jakieś informacje dodatkowe, tak jak robił to leksykograf, wprowadzając do własnego lokalnego korpusu glosy, odesłania, synonimy itp. Jednostką tekstową będącą obiektem opisu jest zazwyczaj słowo, a tekst tak wzbogacony nazywamy korpusem anotowanym. Kwerenda korpusu może więc dotyczyć rozmaitych cech tekstu – tych, które zostały ujawnione przez jego twórców.

W wypadku polszczyzny, języka wysoce fleksyjnego z silnie rozwiniętą homonimią, są to parametry morfologiczne i składniowe, takie jak część mowy, nazwa hasła, przypadek, rodzaj, liczba, osoba, czas, tryb itp. Można powiedzieć, że tym lepszy korpus ogólny (model drugiego rodzaju), im lepszy opis gramatyczno-słownikowy (model pierwszego rodzaju) zaktywizowany przy projektowaniu anotacji. Ponieważ wielki korpus nie może zostać oznakowany manualnie, pierwszym etapem pracy jest ręczna anotacja reprezentatywnego i zrównoważonego podzbioru korpusu. Proces ręcznego znakowania materiału korpusowego jest niezwykle pracochłonny, czasochłonny i kosztowny, więc w dużych korpusach referencyjnych poddawana jest mu tylko relatywnie niewielka część danych. Zwykle ręcznie anotowane podkorpusy nie przekraczają miliona segmentów [Degórski, Przepiórkowski 2012, 51–52]. Podkorpus o takiej wielkości wydzielili także twórcy NKJP. Oto przykład wypowiedzenia z tego korpusu:

Przez wiele lat trwała także współpraca ze znaną szeroko w świecie Olsztyńską Pantomimą Głuchych.

Dostało ono następujące anotacje:

Przez [przez:prep:acc:nwok]
wiele [wiele:num:pl:acc:m3:rec]
lat [rok:subst:pl:gen:m3]
trwała [trwać:praet:sg:f:imperf]
także [także:qub]
współpraca [współpraca:subst:sg:nom:f]
ze [z:prep:inst:wok]
znaną [znany:adj:sg:inst:f:pos]
szeroko [szeroko:adv:pos]
w [w:prep:loc:nwok]
świecie [świat:subst:sg:loc:m3]

Olsztyńską [olsztyński:adj:sg:inst:f:pos]
Pantomimą [pantomima:subst:sg:inst:f]
Głuchych [głuchy:adj:pl:gen:m1:pos]
. [.;interp]

Odczytajmy interpretacje trzech przykładowych słów:

- *trwała* [trwać:praet:sg:f:imperf]: to forma wyrazowa leksemu TRWAĆ o wartościach przeszlik (= składnik tradycyjnej formy przeszłej czasownika; to dla twórców NKJP osobna część mowy), liczba pojedyncza, rodzaj żeński, aspekt niedokonany;
- *lat* [rok:subst:pl:gen:m3]: to forma wyrazowa leksemu ROK o wartościach rzeczownik, liczba mnoga, dopełniacz, rodzaj męski nieżywotny;
- *w* [w:prep:loc:nwok]: to forma wyrazowa leksemu W o wartościach przyimek, rząd miejscownikowy, forma niewokaliczna (bez *e*).

Anotatorzy starają się przypisać słowu jedyną interpretację właściwą, a więc rozwiązywać homonimię w sposób mocny. Zauważmy, że słowo *trwała* to również kształt formy wyrazowej leksemu TRWAŁY o wartościach przymiotnik, liczba pojedyncza, rodzaj żeński, stopień równy, słowo *lat* – kształt formy rzeczownika LATO, a słowo *w* – kształt formy przyimka rządzącego biernikiem.

Podkorpus anotowany manualnie, treningowy, ale dostępny osobno, stanowi podstawę działań automatycznego tagera PANTERA, który przeprowadza anotację całego korpusu. Segmentuje on tekst na wypowiedzenia, przypisuje słowom opisy morfosyntaktyczne i w miarę skutecznie rozwiązuje homonimię.

Twórcy NKJP dostarczają narzędzi wyszukiwania. Są to przede wszystkim wyszukiwarki Poliqarp [Janus, Przepiórkowski 2007] i PELCRA [Pęzik 2012b]. Użytkownik może wprowadzać zapytania, proste lub złożone, dotyczące kształtów słów, lematów, lewego lub prawego kontekstu, różnych wartości parametrów, różnych ich kombinacji itp.

Oto wynik poszukiwania formy leksemu GŁUCHY (10 pierwszych przykładów):

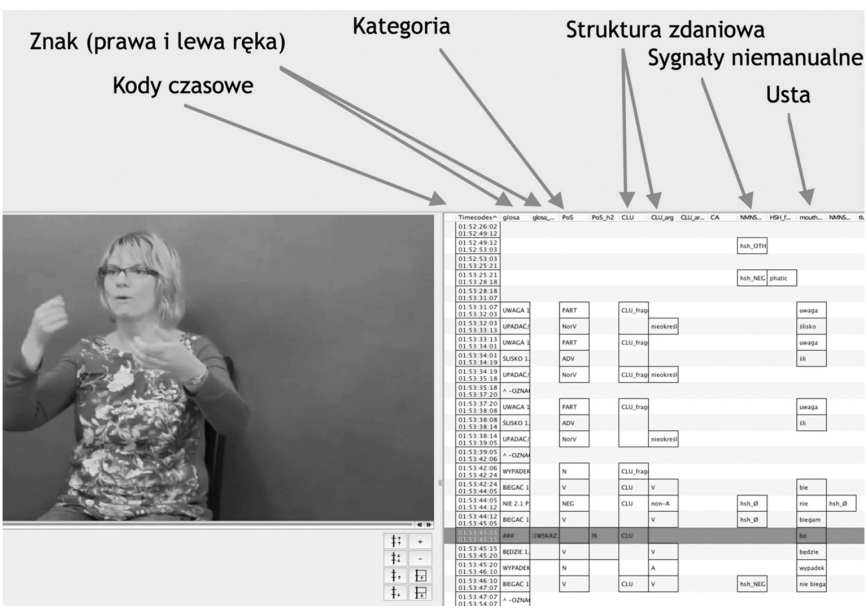
1.	szeroko w świecie Olsztyńską Pantomimą	**Głuchych** [głuchy:adj:pl:gen:m1:pos]	, zapoczątkowana w 1978 roku
2.	na igłę z zakłopotaniem.	**Głuche** [głuchy:adj:pl:nom:f:pos]	!... Sowy
3.	podłodze i rozpoczyna zabawę w	**głuchy** [głuchy:adj:sg:acc:m3:pos]	telefon. Na pożegnanie śpiewa
4.	sennym przywidzeniu, rozległ się	**głuchy** [głuchy:adj:sg:nom:m3:pos]	pomruk burzy. Przeskakiwaliśmy
5.	opuszczały go tłumy katolików,	**głuchych** [głuchy:adj:pl:gen:m1:pos]	na jego desperackie zapewnienia,
6.	, Martwy dla burzy i	**głuchy** [głuchy:adj:sg:nom:m1:pos]	dla gromu, Konaniem zwiędły
7.	klamką. W mieszkaniu panowała	**głucha** [głuchy:adj:sg:nom:f:pos]	cisza. - Wyszła gdzie
8.	razy, nieszczęsną, całkiem	**głuchą** [głuchy:adj:sg:acc:f:pos]	, okropnie postarzałą. Chciała
9.	reszta tip top – alkoholik	**głuchy** [głuchy:adj:sg:nom:m1:pos]	Gruzin – o Żydzie zapomniał
10.	Kurtz usłyszał zza bocznych drzwi	**głuche** [głuchy:adj:pl:acc:n:pos]	uderzenia i okrzyki bólu.

A to wynik poszukiwania sekwencji „forma pojedyncza leksemu GŁUCHY i jakiś przyimek”:

11.	prześladowca. Zadowolenie pozostało jednak	**głuche** [głuchy:adj:sg:nom:n:pos] **na** [na:prep:acc]	apele rozsądku, miało gdzieś
12.	jednym był niezachwiany i,	**głuchy** [głuchy:adj:sg:nom:m3:pos] **na** [na:prep:acc]	cierpkie uwagi podczas konferencji,
13.	w mrok czegoś pustego i	**głuchego** [głuchy:adj:sg:gen:m3:pos] **jak** [jak:prep:nom]	otchłań. Jedyna wolna przestrzeń
14.	z mety dostanie? Dwa	**głuche** [głuchy:adj:sg:nom:n:pos] **na** [na:prep:acc]	korpus i z glana w
15.	poważnie. Był ślepy i	**głuchy** [głuchy:adj:sg:nom:m3:pos] **na** [na:prep:acc]	to, co się z
16.	. Zrozumiała, że kiedy	**głuchy** [głuchy:adj:sg:nom:m3:pos] **we** [w:prep:loc:wok]	wsi Kolok otwartą dłoń kładzie
17.	zupełnie bezcelowy. Reynevan pozostał	**głuchy** [głuchy:adj:sg:nom:m3:pos] **na** [na:prep:acc]	logikę Szarleja, Szarleja nie
18.	stał pochylony, obojętny i	**głuchy** [głuchy:adj:sg:nom:m3:pos] **na** [na:prep:acc]	wszystko, zwrócił uwagę roztrzęsionych
19.	. Jedynie prezes Cichacz pozostał	**głuchy** [głuchy:adj:sg:nom:m3:pos] **na** [na:prep:acc]	tupot nóg. Pająkowska ze
20.	ciszy bezwzględnej, w lesie	**głuchym** [głuchy:adj:sg:loc:m3:pos] **między** [między:prep:inst]	Jastarnią i Helem. A

Znakowanie automatyczne nie wchodzi w grę w wypadku danych wideo (technologie rozpoznawania obrazu nie są wciąż wystarczająco zaawansowane). Z kolei bez anotacji materiał filmowy jest praktycznie bezwartościowy, bo niedostępny. Żadne przeszukiwanie po kształtach nie wchodzi w rachubę. Zarejestrowane klipy wideo zostały więc w procesie tworzenia KPJM poddane anotacji [Mostowski i in. 2018]. Anotatorami byli w większości głusi użytkownicy PJM. Ich zadanie polegało na sporządzeniu zapisu linearnego, a więc posegmentowaniu tekstu na drobniejsze jednostki tekstowe, zakodowaniu ich kształtów, przypisaniu polskich glos, zapisaniu wartości wielu różnych parametrów. Przykład wyniku tego procesu pokazujemy poniżej (wraz z oznaczeniem różnych poziomów anotacji).

Rys. 1. Oznakowany fragment filmu z KPJM



Tak zlinearyzowany i oznakowany tekst migowy – korpus anotowany – może być domeną poszukiwań, podobnie jak korpus języka pisanego; wyszukiwarki pracują na tekście jednowymiarowym, a nie przestrzennym; na tekście, nie na grafice. Chodzi o to, żeby KPJM był modelem tego języka, źródłem danych leksykalnych i gramatycznych.

PODSUMOWANIE: WYKORZYSTANIE I ZNACZENIE KPJM

Zaprojektowanie reprezentatywnego i zróżnicowanego korpusu PJM wymagało, po pierwsze, starannego dobrania głuchych użytkowników tego języka migowego, których konwersacje dostarczyły materiału do korpusu. Po drugie, konieczne było wykreowanie dla nich listy zadań elicytacyjnych. Tworzenie korpusu było poniekąd wymuszaniem tekstu wizualno-przestrzennego, ale nie ma innego sposobu na skompletowanie wystarczająco różnorodnego materiału empirycznego. Pozyskany w ten sposób zbiór nie mógł być tak obszerny jak korpusy pisanej odmiany języków fonicznych. Jednak warto podkreślić, że KPJM to obecnie jeden z dwóch największych zbiorów oznakowanych danych migowych na świecie (zbliżony rozmiar ma jedynie korpus niemieckiego języka migowego, który jednak jest wciąż rozbudowywany i docelowo ma obejmować 3 000 000 segmentów; pozostałe korpusy migowe są znacząco mniejsze, np. korpus niderlandzkiego języka migowego zawiera 145 000 wystąpień znaków). Na odnotowanie zasługuje też fakt, że pierwszy polski słownik przygotowany na początku lat siedemdziesiątych XX w. techniką komputerową i opublikowany w latach 1974–1977 w serii 5 tomów (dostępny w całości jako *Słownik frekwencyjny polszczyzny współczesnej* [Kurcz i in. 1990]) został opracowany na podstawie korpusu obejmującego 500 000 słów (w sensie graficznym), a zatem jeszcze niedawno zbiory o takiej wielkości uznawano za wystarczające także w wypadku modelowania języków fonicznych.

Cechy wspomniane powyżej odróżniają KPJM od korpusów takich jak NKJP. Jednak między migową a foniczną lingwistyką korpusową dostrzegamy też wiele podobieństw. W obu wypadkach chodzi o stworzenie zasobów, które pozwolą na opracowanie modelu danego języka, a w konsekwencji lepsze zrozumienie jego specyfiki leksykalnej i gramatycznej. Choć celem twórców KPJM nie było przeprowadzenie konkretnych analiz danych migowych, to już w trakcie realizacji zadania jego budowy opublikowane zostały dziesiątki prac naukowych wykorzystujących dane zeń zaczerpnięte, w tym dwa doktoraty [Łozińska 2014a; Kuder 2019] i pierwszy słownik PJM rejestrujący rzeczywisty uzus, czyli zasób leksykalny potwierdzony w danych korpusowych [Łacheta i in. 2016]. Wiele z tych publikacji dostępnych jest na stronie internetowej Pracowni Lingwistyki Migowej UW [www.plm.uw.edu.pl].

Należy podkreślić, że prace nad KPJM – poza wymiarem badawczym – znalazły także istotne zastosowania praktyczne. Od 2014 r. PLM UW tworzy na zlecenie Ministerstwa Edukacji Narodowej (obecnie Ministerstwa Edukacji i Nauki) multimedialne adaptacje podręczników szkolnych przeznaczone dla uczniów ze specjalnymi potrzebami edukacyjnymi (w tym niesłyszących i słabosłyszących). Każda z nich ma formę programu komputerowego dającego uczniowi dostęp do tysięcy plików wideo z tłumaczeniami migowymi wszystkich tekstów oryginalnego podręcznika. Przedsięwzięcie to nie byłoby możliwe bez rzetelnych podstaw lingwistycznych w postaci danych korpusowych na temat uzusu PJM.

Dane z KPJM zostały również wykorzystane w kształceniu akademickim. W październiku 2019 r. Uniwersytet Warszawski włączył do swej oferty dydaktycznej nowy kierunek studiów magisterskich – filologię polskiego języka migowego (FPJM). Większość przyjętych osób nie znała PJM na etapie rekrutacji. Zarówno w trakcie samej dydaktyki PJM, jak i podczas badań prowadzących do napisania prac magisterskich udostępniane studentom dane z KPJM okazują się bardzo ważnym i potrzebnym narzędziem (jako materiał szkoleniowy i obiekt analiz). Stworzenie obszernego zasobu tekstów migowych, jakim jest KPJM, procentuje zatem już teraz podnoszeniem kompetencji komunikacyjnych przyszłych absolwentów FPJM, a perspektywy wykorzystywania go w kolejnych latach są jeszcze bardziej obiecujące.

Stworzenie KPJM ma olbrzymią wagę nie tylko dla przyszłych analiz językoznawczych, ale także dla dokumentacji współczesnego PJM dla kolejnych pokoleń polskich głuchych. Jest to pierwszy tego typu zbiór danych w historii tej społeczności (jak już była o tym mowa powyżej, języki migowe nie wykształciły własnych systemów zapisu, w związku z czym nie są utrwalane w usystematyzowany sposób). Starania o sprawiedliwą politykę językową wobec głuchych obywateli oraz o ich pełne uczestnictwo w społeczeństwie (np. poprzez zapewnienie odpowiednich programów edukacyjnych opartych na języku migowym) byłyby skazane na niepowodzenie bez wnikliwego rozpoznania podstawowych mechanizmów rządzących miganiem. Wydaje się to wyjątkowo ważne w krajach takich jak Polska, w których status języków wizualno-przestrzennych jako pełnoprawnych i pełnowartościowych systemów komunikacyjnych był w nieodległej przeszłości bardzo często kwestionowany.

Bibliografia

M. Czajkowska-Kisil, 2014, *Głusi, ich język i kultura – zarys problematyki* [w:] P. Rutkowski, S. Łozińska (red.), *Lingwistyka przestrzeni i ruchu. Komunikacja migowa a metody korpusowe*, Warszawa, s. 17–35.

Ł. Degórski, A. Przepiórkowski, 2012, *Ręcznie znakowany milionowy podkorpus NKJP* [w:] A. Przepiórkowski, M. Bańko, R.L. Górski, B. Lewandowska-Tomaszczyk (red.), *Narodowy Korpus Języka Polskiego*, Warszawa, s. 51–58.

W. Dziewanowska, 2022, *Wariantywność predykacji klasyfikatorowych w polskim języku migowym – badanie pilotażowe*, „Poradnik Językowy" (zeszyt 3, s. 42).

D.M. Eberhard, G.F. Simons, C.D. Fennig (red.), 2021, *Ethnologue: Languages of the World. Twenty-fourth edition*, Dallas; http://www.ethnologue.com

S. Fabisiak, 2010, *Języki migowe a lingwistyka korpusowa*, „Język Polski" nr XC (4–5), s. 338–345.

D. Janus, A. Przepiórkowski, 2007, *Poliqarp 1.0: Some technical aspects of a linguistic search engine for large corpora* [w:] J. Waliński, K. Kredens, S. Goźdź-Roszkowski (red.), *The proceedings of Practical Applications in Language and Computers PALC 2005*, Frankfurt nad Menem.STRONY

A. Kuder, 2019, *Wykładniki negacji w polskim języku migowym (PJM)*, rozprawa doktorska, Uniwersytet Warszawski.

I. Kurcz, A. Lewicki, J. Sambor, K. Szafran, J. Woronczak, 1990, *Słownik frekwencyjny polszczyzny współczesnej*, t. 1–2, Kraków.

B. Lewandowska-Tomaszczyk, 2012, *Parameter variability in translational approximation* [w:] B. Lewandowska-Tomaszczyk, M. Thelen (red.), *Translation and Meaning*, vol. 10, Maastricht, s. 27–37.

J. Łacheta, M. Czajkowska-Kisil, J. Linde-Usiekniewicz, P. Rutkowski (red.), 2016, *Korpusowy słownik polskiego języka migowego*, Warszawa; https://www.slownikpjm.uw.edu.pl/.

S. Łozińska, 2014a, *Czasownik w polskim języku migowym. Studium semantyczno-gramatyczne*, rozprawa doktorska, Uniwersytet Warszawski.

S. Łozińska, 2014b, *Języki migowe jako przedmiot badań* [w:] P. Rutkowski, S. Łozińska (red.), *Lingwistyka przestrzeni i ruchu. Komunikacja migowa a metody korpusowe*, Warszawa, s. 37–46.

P. Mostowski, A. Kuder, J. Filipczak, P. Rutkowski, 2018, *Workflow management and quality control in the development of the PJM Corpus: The use of an issue-tracking system* [w:] M. Bono, E. Efthimiou, S.-E. Fotinea, T. Hanke, J. Hochgesang, J. Kristoffersen, J. Mesch, Y. Osugi (red.), *8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community. Proceedings*, Paryż, s. 133–138.

M. Oliviéri, 2010, *Syntax and corpora*, „Corpus" nr 9, s. 7–20.

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

P. Pęzik, 2012b, *Wyszukiwarka PELCRA dla danych NKJP* [w:] A. Przepiórkowski, M. Bańko, R.L. Górski, B. Lewandowska-Tomaszczyk (red.), *Narodowy Korpus Języka Polskiego*, Warszawa, s. 253–273.

A. Przepiórkowski, M. Bańko, R.L. Górski, B. Lewandowska-Tomaszczyk (red.), 2012, *Narodowy Korpus Języka Polskiego*, Warszawa.

M. Rudolf, 2004, *Metody automatycznej analizy korpusu tekstów polskich*, Warszawa.
P. Rutkowski, M. Czajkowska-Kisil, 2010, *O kategorii zaimka osobowego w Polskim Języku Migowym (PJM)*, „LingVaria" nr 1 (9), s. 65–77.
P. Rutkowski, A. Kuder, J. Filipczak, P. Mostowski, J. Łacheta, S. Łozińska, 2017, *The design and compilation of the Polish Sign Language (PJM) Corpus* [w:] P. Rutkowski (red.), *Different faces of sign language research*, Warszawa, s. 125–151.
P. Rutkowski, S. Łozińska (red.), 2014, *Lingwistyka przestrzeni i ruchu. Komunikacja migowa a metody korpusowe*, Warszawa.
P. Rutkowski, S. Łozińska, 2011, *O niedookreśloności semantycznej migowych predykatów klasyfikatorowych* [w:] M. Bańko, D. Kopcińska (red.), *Różne formy, różne treści*, Warszawa, s. 211–223.
P. Rutkowski, S. Łozińska, J. Filipczak, J. Łacheta, P. Mostowski, 2013, *Jak powstaje korpus polskiego języka migowego (PJM)?*, „Polonica" nr 33, s. 297–308.
P. Rutkowski, S. Łozińska, J. Filipczak, J. Łacheta, P. Mostowski, 2014, *Korpus polskiego języka migowego (PJM): założenia – procedury – metodologia* [w:] M. Sak (red.), *Deaf Studies w Polsce*, t. I, Łódź, s. 219–226.
P. Rutkowski, J. Wójcicka, P. Mostowski, A. Kuder, 2021, *Korpus Polskiego Języka Migowego jako źródło danych frekwencyjnych do badań nad leksyką migową*, referat, „Czwartkowe spotkania lingwistyczne" – seminarium naukowe Instytutu Języka Polskiego Wydziału Polonistyki UW, 14 stycznia 2021 r.
M. Świdziński, 2005, *Języki migowe* [w:] T. Gałkowski, E. Szeląg, G. Jastrzębowska (red.), *Podstawy neurologopedii*, Opole, s. 679–692.
M. Świdziński, 2015, *Testowanie biegłości językowej w zakresie języków wizualno-przestrzennych* [w:] J. Sujecka-Zając (red.), *Ewaluacja biegłości językowej. Od pomiaru do sztuki pomiaru*, Warszawa, s. 69–82.
M. Świdziński, P. Rutkowski, 2014, *Ikoniczność nieleksykalna: reprezentacja referencjalna jako składnik tekstu w językach wizualno-przestrzennych* [w:] P. Rutkowski, S. Łozińska (red.), *Lingwistyka przestrzeni i ruchu. Komunikacja migowa a metody korpusowe*, Warszawa, s. 145–154.
J. Wójcicka, A. Kuder, P. Mostowski, P. Rutkowski (red.), 2020, *Otwarte Repozytorium Korpusu Polskiego Języka Migowego*, Warszawa; https://www.korpuspjm.uw.edu.pl

A general corpus as a model of a natural language: corpora of phonic languages and the Corpus of Polish Sign Language

Summary

The aim of this paper is to discuss the major differences and similarities between the Corpus of Polish Sign Language (KPJM), which has been developed for a decade by the team of the Section for Sign Linguistics, Faculty of Polish Studies, University of Warsaw, and corpora of phonic languages (and in particular the National Corpus of Polish (NKJP)). The KPJM is a general corpus with an ambition to represent the whole language, used by the Polish Deaf. Unlike the corpora of phonic languages, which are collections of existing texts, the material of the KPJM was generated purposefully by recording and annotating an extensive set of videos. The paper shows that the sign language corpus should be viewed as analogous to spoken language corpora rather than to written language corpora. The KPJM can be perceived as a model of Polish Sign Language.

Keywords: PJM (Polish Sign Language) – Corpus of Polish Sign Language (KPJM) – National Corpus of Polish (NKJP) – general corpus – sign linguistics – corpus linguistics

Trans. Monika Czarnecka