4

Język mówiony w NKJP

Piotr Pęzik

4.1. Język mówiony a reprezentatywność korpusu

Uwzględnienie języka mówionego w Narodowym Korpusie Języka Polskiego nie wymaga długiego uzasadnienia. Korpus narodowy ma charakter referencyjny i jako taki powinien stanowić źródło informacji o tak ważnej odmianie języka, jaką jest ogół języka mówionego, tym bardziej, że to właśnie język mówiony stanowi największą pod względem ilości naturalnie występujących wypowiedzi odmianę języka. Wydaje się, iż mimo ostatniego wzrostu poziomu piśmiennictwa spowodowanego upowszechnieniem się Internetu jako platformy komunikacji językowej, rodzimi użytkownicy polszczyzny dużo więcej mówią, niż piszą, oraz dużo częściej są odbiorcami języka mówionego niż czytelnikami tekstów[1]. Język mówiony zasługuje więc na szczególne uwzględnienie w korpusie referencyjnym, bez względu na to, czy jego głównym kryterium doboru tekstów jest poziom odbioru, czy też produkcji danej odmiany języka.

Szczególną rolę komponentu danych mówionych w korpusie referencyjnym uzasadniają także wyjątkowe własności w warstwie leksykalnej, frazeologicznej, składniowej oraz pragmatycznej języka mówionego. Kilka przykładów wyjątkowości języka mówionego można znaleźć w dalszej części tego rozdziału, ale warto tu wspomnieć o tym, jak ilościowa analiza języka może zależeć od rodzaju danych, na podstawie których jest przeprowadzana. Jedną z podstawowych technik ilościowej analizy korpusu są listy frekwencyjne wyrazów. Czy na podstawie listy frekwencyjnej wyrazów wygenerowanej z całości danych NKJP można wskazać

[1] Twierdzenie, że więcej języka słyszymy, niż czytamy, wydaje się szczególnie prawdziwe, jeżeli przez pojęcie języka mówionego rozumiemy kanał, a nie typ języka. Trudno bez wiarygodnych danych stwierdzić, czy odbieramy więcej języka mówionego w jego mniej lub bardziej swobodnych odmianach niż czytanego.

najczęstszy wyraz w polszczyźnie? Niemal każdy językoznawca korpusowy zajmujący się angielszczyzną wie, że najczęstszym wyrazem w ogromnej większości korpusów języka angielskiego jest przedimek określony *the*[2]. Zgodnie z prawem Zipfa o rozkładzie frekwencji słów, częstość *the* jest zazwyczaj (co najmniej) dwukrotnie większa od częstości kolejnego słowa na liście. Niestety dużo trudniej wskazać taki wyraz w języku polskim. W języku pisanym jest to zazwyczaj przyimek *w*, ale w podkorpusie danych mówionych konwersacyjnych w NKJP ten sam wyraz plasuje się na piątym miejscu listy frekwencyjnej (tab. 4.1). Ze względu na swoją wieloznaczność oraz funkcje dyskursywne wyrazy *to*, *nie* oraz *no* są dużo częstsze w języku mówionym niż w tekstach pisanych. Jeżeli przyjąć, że więcej mówimy niż piszemy, może to również oznaczać, że przyimek *w* nie jest najczęstszym polskim wyrazem. Przykład ten potwierdza, iż wyników analizy ilościowej przeprowadzanej na podstawie korpusu zawierającego tylko próbki języka pisanego nie można rozciągać na język mówiony, nawet na tak prostym poziomie, jakim są frekwencje pojedynczych leksemów. Dlatego też w NKJP znalazły się próbki języka mówionego zebrane przez grupę PELCRA (z lat 2000–2005) oraz zebrane specjalnie na potrzeby NKJP dane mówione oraz konwersacyjne z lat 2008–2010. Aktualność tych danych widać w wynikach zapytań o leksemy, które w ciągu kilku ostatnich lat nabrały nowego znaczenia w języku mówionym. Przykładem tego mogą być zarejestrowane w NKJP użycia wyrazów *ogarnąć/ogarniać się* oraz *ogarnięty*, które są używane zwłaszcza przez ludzi młodych odpowiednio w znaczeniach *opanować/poradzić sobie z czymś*, *zrozumieć*, *wziąć się w garść* oraz *bystry/zaradny*, np.:

(4.1) Boże dziewczyno *ogarnij się*

(4.2) też to jakoś bo normalnie . muszę to napisać jakoś nie wiem . *ogarnąć* to szybko zrobić żeby mieć z głowy

(4.3) to nie ten majk to jest pikuś przy nim . to Kangur jest *ogarnięty* oni się tak nie biją

(4.4) dla mnie to jest nie do *ogarnięcia*. no rozumiem rozumiem Dorota nie każdy

4.2. Typy języka mówionego w NKJP

Mimo niekwestionowanej potrzeby uwzględnienia języka mówionego w dużym korpusie referencyjnym, w zrównoważonej wersji NKJP różnego rodzaju próbki tego typu danych stanowią jedynie około 10 procent korpusu, czyli 30 milionów

[2] Wyjątkiem mogą tu być korpusy bardzo specyficznych odmian angielszczyzny, np. zbiory tekstów piosenek czy prognoz pogody.

Tabela 4.1. Najczęstsze wyrazy w języku mówionym

Nr	Wyraz	Częstość
1.	to	86 960
2.	nie	82 714
3.	no	61 964
4.	i	54 915
5.	w	47 056
6.	że	40 558
7.	się	40 438
8.	tak	39 677
9.	na	37 979
10.	a	29 358

słów tekstowych. Takie proporcje wynikają przede wszystkim z faktu, że pozyskiwanie, transkrypcja oraz anotacja dużych ilości tego typu danych jest jednym z najbardziej pracochłonnych, a przez to kosztownych etapów budowania dużych korpusów. Podobne proporcje między językiem mówionym a pisanym można znaleźć w takich korpusach referencyjnych jak na przykład Brytyjski Korpus Narodowy. Należy przy tym pamiętać, że nie wszystkie dane mówione w NKJP można uznać za język konwersacyjny. Trzy główne typy danych mówionych w NKJP przedstawia tab. 4.2. Dane *mówione medialne* to transkrypcje audycji ra-

Tabela 4.2. Typy języka mówionego w NKJP

Typ	Liczba segmentów wyrazowych
Mówione medialne	900 000
Mówione konwersacyjne	1 900 000
Inne	27 200 000

diowych oraz programów telewizyjnych nagrywane i anotowane specjalnie na potrzeby NKJP. Dane *mówione konwersacyjne* to zebrane również na potrzeby NKJP transkrypcje rozmów Polaków o dość zróżnicowanym profilu wiekowym, w różnym stopniu wykształconych i pochodzących z różnych regionów kraju. Inne dane mówione to głównie skonwertowane na potrzeby NKJP, dostępne publicznie stenogramy posiedzeń Sejmu oraz sejmowych komisji śledczych. Część danych tej ostatniej kategorii należałoby raczej określić mianem *języka czytanego* niż mówionego[3] (np. liczne wystąpienia posłów), chociaż wśród stenogramów

[3] W angielskiej literaturze korpusowej używa się na określenie takich danych terminu *to-be-spoken*.

z posiedzeń komisji śledczych zdarzają się dość spontaniczne pasaże o charakterze konwersacyjnym, które cechuje duży stopień improwizacji.

Niektóre korpusy referencyjne ze względu na trudności związane z pozyskaniem próbek języka konwersacyjnego poprzestają na gotowych transkrypcjach danych medialnych. Przykładem takiego korpusu jest Korpus Współczesnej Angielszczyzny Amerykańskiej (Corpus of Contemporary American English), który zawiera ok. 80 milionów słów języka mówionego w postaci transkrypcji popularnych programów telewizyjnych oraz radiowych. Oczywiście takie dane są same w sobie niezwykle cenne i interesujące, ale należy podkreślić, że język mówiony programów telewizyjnych i radiowych w dość oczywisty i fundamentalny sposób różni się od nieformalnego języka mniej lub bardziej spontanicznych konwersacji między znajomymi, członkami rodziny, a nawet osobami bliżej sobie nieznanymi, ale rozmawiającymi w nieformalnych okolicznościach. Po pierwsze, wypowiedzi uczestników programów telewizyjnych i radiowych podlegają zazwyczaj pewnym rygorom czasowym i tematycznym. Uczestnicy audycji odpowiadają często na konkretne pytania redaktora prowadzącego, którego zadaniem jest czuwanie nad narzuconą z góry spójnością tematyczną danego programu. Dygresje w wypowiedziach uczestników audycji publicystycznych, jeżeli się zdarzają, to traktowane są z reguły jako odstępstwo od formuły programu. Wyjątkiem od tej reguły są pojedyncze audycje. Na przykład audycję „Rozmowy niedokończone", nadawaną w Radiu Maryja oraz Telewizji Trwam, zgodnie z zasygnalizowaną w jej tytule formułą, często cechuje dość duży stopień swobody w dopuszczalnej długości i spójności tematycznej wypowiedzi uczestników pod warunkiem, że pozostają one w zgodzie z pewnym systemem wartości prezentowanym w audycjach nadawanych na antenie tych stacji. Niemały wpływ na język, którym posługują się uczestnicy programów telewizyjnych, ma także sama świadomość obecności w studio i publicznego charakteru utrwalanej lub transmitowanej przez urządzenia rejestrujące rozmowy. Świadomość ta w oczywisty sposób zawęża nie tylko wybory leksykalne i frazeologiczne rozmówców, ale też stosowane przez nich strategie prowadzenia dyskursu oraz wachlarz technik argumentacyjnych. Z punktu widzenia pragmatyki języka można z kolei zauważyć, że odbiorcami komunikatów wypowiadanych publicznie są, poza bezpośrednimi uczestnikami rozmowy, telewidzowie i radiosłuchacze, co ma niemały wpływ na ich treść oraz formę.

Dane mówione medialne zazwyczaj różnią się też od danych mówionych konwersacyjnych formalnością rejestru. Poza programami typu reality show, w danych medialnych rzadko spotyka się wyrazy, frazeologizmy i zbitki słowne typowe dla nieformalnej polszczyzny konwersacyjnej. Ilustrują to ukazane w tab. 4.3 przykłady *zbitek leksykalnych* charakterystycznych dla języka mówionego. *Zbitkami leksykalnymi* (ang. *lexical bundles*) nazywamy w lingwistyce korpusowej najczęściej

Tabela 4.3. Typowe dla polszczyzny mówionej kombinacje segmentów wyrazowych

Zbitka	Przykład
Rejestr formalny i nieformalny	
na przykład	jak rozumiem nie każda służba **na przykład** . współczesne służby jak na przykład jak CBA . nie każde ich postępowanie kończy się aktem oskarżenia w sądzie a tym się powinny...
po prostu	szuka się haków jako metody **po prostu** pokazania . opinii publicznej że ktoś ma te słabe strony i ukrył je . przed wyborcami .
Rejestr formalny	
jest tak że	**jest tak że** no właśnie bo tak krzyczymy chodźcie z nami a nikt za nami nie idzie
wydaje mi się	ale **wydaje mi się** że co innego pan poseł Grupiński powiedział . bardzo proszę
yy yy	ten okres tutaj powojenny yy przysłużył się tej dewastacji tych obiektów natomiast **yy yy** no w tej chwili tutaj to tylko pokazuje jaki jest potencjał
myślę że	**myślę że** to rzeczywiście jest ułomność . polskiej demokracji nie będzie się ważyć kto bardziej kompetentny .
prawda ?	jeśli są prowokacje to trzeba je ujawnić i powiedzieć że one są nielegalne lub legalne . **prawda ?**
Rejestr nieformalny	
nie ?	kartofli nagotowałam tak najwyżej sobie wiesz nawet krupniku tego ugrzeję włożę kartofli bez . **nie ?**
no ale	tak pani też mnie masowała na młode dziewczyny są młode no . **no ale** wszystko takie tam są .
no i	do Kanady pojechała do syna . **no i** tamci tubylcy . zaczęła tym swoim dzieciom bo dwoje bliźniąt tam miała
no bo	**no bo** to jest wasz ten średni pakiet nie ? nie ten najdroższy tylko ten średni

powtarzające się sekwencje wyrazów w tekstach korpusu[4]. Ta prosta technika korpusowa jest często stosowana w analizie dyskursu mówionego (Biber 2004). Choć wiele z najczęstszych sekwencji segmentów wyrazowych nie tworzy spójnych jednostek składniowych ani nawet frazeologicznych, to wygenerowanie ich listy ukazuje kombinacje wyrazów, które odgrywają ważną rolę w budowie wypowiedzi ustnych. Niektóre z nich (np. *na przykład*, *po prostu*) występują bardzo często

[4] Niektóre z podanych tu przykładów to także kombinacje pojedynczych wyrazów i znaków interpunkcyjnych stosowanych w anotacji danych mówionych w NKJP.

w polszczyźnie mówionej bez względu na poziom formalności rozmowy. Są jednak zbitki leksykalne, których prawdopodobieństwo użycia zależy od rejestru języka mówionego. Za szczególnie charakterystyczne dla formalnego języka mówionego można uznać takie kombinacje segmentów wyrazowych, jak *jest tak że*, *myślę że* czy też *wydaje mi się, że*[5]. W rejestrze formalnym mówionym częściej niż w nieformalnych rozmowach zdarzają się wtrącenia sygnalizujące wahanie mówcy (np. *yy yy*), co można wytłumaczyć presją budowania składniejszych, bardziej wyważonych wypowiedzi na antenie radiowej lub telewizyjnej. Z kolei w swobodnych rozmowach w miejsce *prawda?* pojawia się *nie?*, a kombinacje *no* z innymi wyrazami są dużo częstsze niż w wypowiedziach formalnych, np. *no bo*, *no i*, *no ale*.

4.3. Pozyskiwanie danych konwersacyjnych

W odróżnieniu od nagrań medialnych, których dokonuje się w warunkach studyjnych, pozyskanie autentycznych próbek nieformalnego języka mówionego wymagało zastosowania specjalnej metodologii nagrywania i transkrypcji danych. W pierwszej kolejności, kilkanaście specjalnie przeszkolonych w tym zakresie osób zostało wyposażonych w dobrej klasy dyktafony. Osoby te uzyskały zgodę swoich znajomych oraz członków rodzin na nagrywanie ich rozmów w najbliższej przyszłości. Po dokonaniu nagrania osoby zbierające dane uzyskiwały pisemną zgodę na ich wykorzystanie do celów NKJP. Dzięki temu możliwe było zarejestrowanie wielu spontanicznych rozmów, których uczestnicy nie czuli się skrępowani faktem nagrywania ich wypowiedzi. Uzyskane w ten sposób nagrania zostały przetranskrybowane na format pośredni. W transkrypcjach zmieniono lub usunięto imiona i nazwiska, które mogłyby naruszyć prywatność rozmówców. Osoby nagrywające dane zajmowały się następnie ich transkrypcją, dzięki czemu łatwiej było odtworzyć fragmenty rozmów zarejestrowanych w trudnych warunkach akustycznych.

4.4. Anotacja

Schemat anotacji używany do transkrypcji danych mówionych różni się dość istotnie od anotacji tekstów pisanych. Każda nagrana rozmowa jest traktowana jako jeden tekst, w którym podstawową jednostką strukturalną są kolejne wypowiedzi mówców. Przykład formatu anotacji języka mówionego używanego do transkrypcji rozmów w NKJP przedstawiono na wydr. 4.1. Warto zaznaczyć, że jest to schemat anotacji zgodny ze standardem TEI (Text Encoding Initiative).

[5] Typowość kombinacji dla danego rejestru zmierzono miarą istotności statystycznej przez porównanie częstości względnych w danych mówionych medialnych oraz konwersacyjnych.

Wydruk 4.1. Przykład anotacji strukturalnej z podziałem na wypowiedzi

```
1  <u who="sp3" xml:id="u-1"><incident><desc>pukanie do
   drzwi</desc></incident> proszę. bileciki do kontroli.</u>
2  <u who="sp2" trans="overlap" xml:id="u-2"> co tam ?</u>
3  <u who="sp0" trans="overlap" xml:id="u-3"> cześć cześć
   cześć.</u>
4  <u who="sp6" xml:id="u-4"> czemu wy stoicie ?</u>
5  <u who="sp2" xml:id="u-5"> siadajcie rozgoście się impreza
   się tu przeniosła. czekacie na pocią ?<gap/> jak w
   poczekalni zupełnie. jak na pociąg byśmy czekali.</u>
6  <u who="sp3" xml:id="u-6"> jak na terminalu.</u>
7  <u who="sp2" xml:id="u-7"> na terminalu. ta
   Londyńczycy.<vocal> <desc>śmiech</desc> </vocal></u>
8  <u who="sp4" xml:id="u-8"> dziewczyny a wy już spakowane
   jesteście ?</u>
9  <u who="sp0" xml:id="u-9"> no ta. miałam tyle rzeczy że ho.
   pół szafy zabrałam. co ?</u>
10 <u who="sp2" xml:id="u-10"> nie ma to jak na piecyku nie
   ?</u>
11 <u who="sp1"
   xml:id="u-11"><vocal><desc>ziewając</desc></vocal>
   nie.<vocal><desc>#VOICE_END</desc></vocal> on nie grzeje
   wiesz.</u>
12 <u who="sp0" xml:id="u-12"> Darek. my musimy kluczyki od
   razu oddawać czy to jak to wygląda ? czy do ciebie ?</u>
```

Poszczególne wypowiedzi przypisane są do rozmówców zdefiniowanych w nagłówku pliku transkrypcji. Dzięki temu możliwe jest przeszukiwanie pełnotekstowe wypowiedzi z uwzględnieniem podstawowych kryteriów socjolingwistycznych, takich jak wiek, wykształcenie, płeć czy też miejsce zamieszkania mówców. W transkrypcji języka mówionego wielkie litery używane są tylko w imionach oraz dobrze znanych nazwach własnych. Obce nazwy i wyrazy zapisywane są fonetycznie. Fonetycznie zapisywana jest także wyraźna wymowa gwarowa lub niestandardowa, a także przypadki stylizacji na taką wymowę. Wykrzyknik oraz pytajnik pełnią funkcję podobną do ich odpowiedników w standardowej interpunkcji. Kropki są stosowane do oznakowania pauz w pojedynczych wypowiedziach i nie można ich uważać za granice zdań. Istotne zmiany tonu głosu oraz zdarzenia mogące mieć wpływ na przebieg rozmowy zaznaczono specjalnymi elementami `<vocal>` oraz `<incident>`. Fragmenty rozmów, których ze względu na warunki akustyczne nie udało się przetranskrybować, zostały zastąpione znacznikiem `<gap>`. Nakładanie się na siebie wypowiedzi sygnalizuje

wartość atrybutu `@trans="overlap"`. W najbliższej przyszłości planowane są również prace nad dodaniem do wybranych transkrypcji anotacji czasowej, która umożliwi swobodne odtwarzanie dowolnego fragmentu rozmowy.

Anotacja zdarzeń ma niekiedy szczególne znaczenie ze względu na zubożenie semiotyczne, jakie cechuje zapis ortograficzny autentycznych rozmów, toczonych często w warunkach „polowych". Przykładem takiej sytuacji może być fragment rozmowy ukazany na wydr. 4.2.

Wydruk 4.2. Anotacja zdarzeń w transkrypcjach NKJP

```
1 <u who="sp2" trans="overlap" xml:id="u-105"> to i to i to i
  tak taniej a taki kupić sobie pięcio– sześcioletni rozumiesz
  audi czy to jest samochód do śmierci. to takie tutaj
  pokupowali audiki można powiedzieć dziesięcio–
  piętnastoletnie jak ten Szymanek kupił dizla . też
  sprowadzany też mu sprowadził Niemiec nie ? kurde to ci
  mówię samochód nie do zdarcia. i hak i przyczepę widziałeś
  przecież. a nie byłeś tam<gap/><incident><desc>wszyscy
  spoglądają na kaczkę która je kapselek od
  piwa</desc></incident></u>
2 <u who="sp0" xml:id="u-106"> kapselek je !</u>
3 <u who="sp2" xml:id="u-107"> no tak to wszystko tylko żeby
  nie pożarł. kaczka !<vocal><desc>śmiech</desc></vocal>
  widzisz jakie kaczory to głupki</u>
4 <u who="sp0" xml:id="u-108"> o drugi o</u>
5 <u who="sp2" xml:id="u-109"> kiedyś wyciągłem to ze dwa
  metry sznurka</u>
6 <u who="sp0" trans="overlap" xml:id="u-110"> o Jezu</u>
7 <u who="sp2" xml:id="u-111"> to taki
  duży<incident><desc>pokazuje jaki długi
  sznurek</desc></incident> znalazł sznurek rozumiesz i tak
  połykał połykał połykał i później przestała żreć ta kaczka i
  se latała z tym sznurkiem i czubek było widać mówię ciągnę
  ciągnę<vocal><desc>śmiech</desc></vocal>
```

Tematem rozmowy toczonej upalnego lipcowego dnia w gospodarstwie rolnym w województwie kujawsko-pomorskim są sprowadzane z zagranicy samochody. W pewnej chwili, dość niespodziewanie, w transkrypcji pojawia się wypowiedź *Kapselek je!* Didaskalia dodane przez osobę transkrybującą tekst w znaczniku `<incident>` pomagają zrozumieć tę nagłą zmianę tematu, którą w tym wypadku wywołała kaczka próbująca połknąć leżący na podwórzu kapsel od butelki.

W nagłówkach trankrypcji języka mówionego znajdują się informacje o poszczególnych uczestnikach rozmowy (zob. wydruk 4.3), przy czym w transkrypcjach audycji radiowych i telewizyjnych brakuje czasem informacji o miejscu zamieszkania, wieku oraz wykształceniu niektórych rozmówców. Dodatkowo w nagłówkach zakodowano miejsce i datę dokonania nagrania oraz główne tematy poruszane podczas rozmowy. Co ciekawe, z anotacji tematów rozmów wynika, że różnorodność tematyczna jest jedną z charakterystycznych cech potocznego języka konwersacyjnego i nie zawsze musi być bezpośrednim rezultatem zdarzeń występujących w trakcie samej rozmowy. W ciągu godzinnej rozmowy znajomych często poruszonych zostaje kilkadziesiąt tematów. Niektóre z nich są porzucane tuż po wprowadzeniu, choć zdarzają się także dłuższe okresy stabilizacji tematycznej.

Wydruk 4.3. Nagłówek transkrypcji języka konwersacyjnego

```
<particDesc>
 <person xml:id="sp1" role="speaker">
   <persName>anonim</persName>
   <education xml:lang="pl">wyższe</education>
   <age>25</age>
   <residence>Wieluń</residence>
 </person>
 <person xml:id="sp0" role="speaker">
   <persName>anonim</persName>
   <sex value="1">male</sex>
   <education xml:lang="pl">średnie</education>
   <age>49</age>
   <residence>Wieluń</residence>
 </person>
</particDesc>
```

4.5. Wyszukiwarka

W oparciu o wspomniany powyżej schemat anotacji opracowana została specjalna wyszukiwarka pozwalająca wygodnie przeszukiwać, sortować i wyświetlać konwersacyjną część korpusu NKJP z uwzględnieniem metadanych właściwych dla języka mówionego. Rys. 4.1 ukazuje przykładowy ekran kontekstu wyniku pasującego do zapytania `no i`. Poszczególne wypowiedzi stanowiące kontekst wystąpienia poszukiwanego wyrazu lub frazy ukazane są w kolejnych wierszach tabeli. W ostatnich trzech kolumnach wyświetlane są informacje o płci,

wieku i wykształceniu autora danej wypowiedzi. Wyszukiwarka NKJP dla danych mówionych jest dostępna publicznie pod adresem `http://nkjp.uni.lodz.pl/spoken.jsp`.

Rysunek 4.1. Ekran wyników w wyszukiwarce danych konwersacyjnych

Informacje o nagraniu

Id	Tytuł	Źródło	Data nagrania
31	*Motor*	PELCRA	2008

Kontekst wystąpienia

#	Id	Wypowiedź	Płeć	Wiek	Wykształcenie
1.	3616	czy słono ? normalnie wiesz tak jak	m	34	wy
2.	3617	normalne ceny . ale	m	38	wy
3.	3618	kupił to tam za czterysta tysięcy bodajże	m	34	wy
4.	3619	to faktycznie jak za taką działę	m	38	wy
5.	3620	w tamtym miejscu wiesz chata dwieście osiemdziesiąt metrów czy tam blisko dwustu osiemdziesięciu	m	34	wy
6.	3621	i dwa poziomy ?	m	38	wy
7.	3622	*ogromna działka . dwa poziomy mieszkalne* ***no i*** *piwnica taka wiesz wielka tak że*	m	34	wy
8.	3623	kawał domu	m	38	wy
9.	3624	kawał domu . no i wszystko później mu robili . i kazał wiesz . bo oczywiście robotnicy chcieli już tam schody rozwalać . tu te a schody też fajne drewniane takie pół kręcone jakby na piętro nie ? tu już to . wszystko wiesz . zero rozwalania mają wszystko wiesz	m	34	wy

4.6. Przyszłe prace

Należy podkreślić, że zebranie tak dużej liczby transkrypcji języka potocznego było trudnym zadaniem i wymagało pewnych kompromisów. Z praktyki zbierania tego typu danych wynika, że ze względu na zakłócenia i warunki akustyczne, danej rozmowy często nie może w całości odtworzyć osoba, która nie brała w niej choćby biernego udziału. Jednakże nie było ani możliwe, ani metodologicznie uzasadnione powierzenie zbierania danych konwersacyjnych tylko głęboko uświadomionym językoznawcom lub osobom o nienagannej ortografii, ponieważ z pewnością obniżyłoby to poziom zróżnicowania zebranej próbki. W transkrypcjach danych konwersacyjnych można więc czasem spotkać niejednolity zapis niestandardowej, gwarowej lub stylizowanej wymowy, które mogą wymagać dalszej normalizacji w kontekście całego korpusu. Sporadycznie pojawiają się także błędy ortograficzne, choć często trudno bezpośrednio przekładać normy ortografii języka pisanego na zapis swobodnych rozmów[6]. Zebranie

[6] Co ciekawe, oczywiste błędy ortograficzne występują także w transkrypcjach mówionych Brytyjskiego Korpus Narodowego. Mylone są na przykład formy *their* i *they're*.

tak dużej liczby stosunkowo zróżnicowanych nagrań swobodnych rozmów oraz przetranskrybowanie ich w stosunkowo krótkim czasie rządziło się swoimi prawami. Natomiast dzięki zachowanym nagraniom cyfrowym korekta ortografii i normalizacja transkrypcji może być przedmiotem dalszych prac stosunkowo niewielkiej liczby osób, które nie musiały brać udziału w nagranych rozmowach.

4.7. Podsumowanie

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. Podstawowa funkcja korpusu referencyjnego, jakim jest Narodowy Korpus Języka Polskiego, to reprezentacja poczucia językowego przeciętnego użytkownika polszczyzny. Aby korpus spełniał tę funkcję, konieczne było zebranie możliwie największej liczby próbek języka mówionego, ze szczególnym uwzględnieniem potocznego języka konwersacyjnego. Według wiedzy autora, komponent konwersacyjny NKJP jest największym polskim korpusem języka potocznego o dużym stopniu zróżnicowania sytuacyjnego. Pozostaje mieć nadzieję, że będzie on wykorzystywany w ilościowej oraz jakościowej analizie polszczyzny mówionej.