INVESTIGATIONES LINGUISTICAE, VOL. XVI; POZNAŃ, DECEMBER 2008

Rejestracja, transkrypcja i tagowanie mowy oraz gestów w narracji dzieci i dorosłych[1]

Recording and transcription of speech and gesture in the narration of Polish adults and children

Maciej Karpiński[a, d], *Ewa Jarmołowicz-Nowikow*[a, d],
Zofia Malisz[b, d] *Michał Szczyszek*[c, d], *Konrad Juszczyk*[a, d]

maciej.karpinski@amu.edu.pl, ewa@jarmolowicz.art.pl,
zmalisz@gmail.com, mszczyszek@gmail.com, juszczyk@amu.edu.pl

[a]Instytut Językoznawstwa, Uniwersytet im. Adama Mickiewicza
al. Niepodległości 4, 61-874 Poznań
[b]School of English, Uniwersytet im. Adama Mickiewicza
al. Niepodległości 4, 61-874 Poznań
[c]Institute of Polish Philology, Uniwersytet im. Adama Mickiewicza
ul. Fredry 10, 61-701 Poznań
[d]Center for Speech and Language Processing, Uniwersytet im. Adama Mickiewicza
al. Niepodległości 4, 61-874 Poznań

Abstract

In the present paper, the experimental procedure, the details of sound and video recording set-up as well as the system for speech and gesture transciption and coding used in the Polish Cartoon Narration Corpus (PCNC) project are described. The audio-visual data come from a cartoon narration task performed by both children and adults. The recordings are transcribed orthographically and phonemically, and labelled for selected phenomena on a number of levels, including gesture, lexicon, prosody, and dialogue acts.

1. Cele projektu

W niniejszym tekście przedstawiono system opisu danych multimodalnych (obrazu filmowego oraz dźwięku). Celem projektu jest analiza zachowań komunikacyjnych dzieci i dorosłych realizujących zadanie polegające na opowiadaniu obejrzanego wcześniej filmu animowanego. Procedura przeprowadzenia nagrań została w dużej mierze oparta na stosowanej w podobnych

[1] Projekt jest realizowany z funduszy grantu specjalnego Ministerstwa Nauki i Szkolnictwa Wyższego w ramach Europejskiego Działania COST2102.

badaniach, prowadzonych od wielu lat w McNeill Laboratory. Na podstawie anotacji dziesięciu sesji z udziałem dziewięcioletnich dzieci oraz dziesięciu sesji z udziałem dorosłych zespół zamierza zrealizować porównawcze badania nad wkładem gestu i mowy w wypowiedzi multimodalne, skupiając się na realizacjach wybranych aktów dialogowych. Projekt ten jest realizowany jako część inicjatywy Multilingual and Multicultural Database of Cartoon Narration Recordings, prowadzonej w ramach Działania COST2102 przez Annę Esposito (2nd University of Naples). Dzięki standardowym procedurom rejestracji korpusu, ujednoliceniu formatu oraz zadeklarowaniu sposobu opisu danych, możliwa będzie ich wymiana z innymi zespołami, a w konsekwencji – porównawcze badania interkulturowe.

2. Przebieg eksperymentu i rejestracja dźwięku i obrazu.

2.1. Przebieg nagrań

Nagrania dzieci przeprowadzono w sali lekcyjnej. Przed nagraniami grupa dzieci była zapraszana do pomieszczenia i zapoznawała się tam z prowadzącymi badania oraz z przygotowaną aparaturą. Następnie kolejne dzieci zapraszano osobno, wyjaśniano im ich rolę i zadanie, starając się przy tym stworzyć możliwie przyjazną i swobodną atmosferę. Sesje z udziałem dorosłych przeprowadzono w cichym, ale nie zaadaptowanym akustycznie pomieszczeniu. Tutaj również przywiązywano dużą wagę do zapewnienia sprzyjającej odprężeniu, swobodnej atmosfery.

Każda sesja składała się z dwóch etapów. Uczestnik oglądał film animowany, a następnie był proszony o jego opowiedzenie osobie przygotowanej przez autorów korpusu. Rozmowa trwała od pięciu do dziesięciu minut.

2.2. Rejestracja dźwięku i obrazu

Nagrania wykonano za pomocą profesjonalnego sprzętu przenośnego. Do rejestracji dźwięku wykorzystano nagrywarkę Fostex VF80EX z wewnętrznym dyskiem twardym i mikrofony pojemnościowe Behringer B-1. Nagrania zostały zarchiwizowane na płytach CD za pomocą nagrywarki Tascam CD-RW700 i poddane normalizacji z wykorzystaniem odpowiedniego oprogramowania. Mikrofony zostały umieszczone na statywach w odległości umożliwiającej mówcy swobodną gestykulację. Jakość fonii na nagraniach zrealizowanych w szkole jest stosunkowo niska ze względu na własności akustyczne pomieszczenia oraz hałas. Do rejestracji obrazu zastosowano kamery Sony HVR-HD1000 i HDR-SV12. Nagrania zostały przeniesione na dysk twardy komputera, skompresowane (aby zapewnić łatwiejszą przenośność plików) oraz przekonwertowane do formatu MPEG1 (aby umożliwić ich późniejszą anotację w programie ELAN). Wykonane odrębnie nagrania fonii zostały w programie ELAN zsynchronizowane z plikami wideo.

3. Segmentacja i transkrypcja

Transkrypcję nagrań wykonano w programach Praat (dźwięk) i Elan (obraz). Programy te umożliwiają opis mowy i gestów w wielu oddzielnych warstwach. Dla każdej warstwy przyjęto odpowiednią segmentację na jednostki wypowiedzi:

- Transkrypcja ortograficzna – wyrazy tekstowe zgodnie z zasadami współczesnej ortografii języka polskiego;
- Transkrypcja fonematyczna – zapis fonematyczny, odzwierciedlający niektóre aspekty realizacji wypowiedzi, które pomija transkrypcja ortograficzna;

- Transkrypcja gestów – zapis wybranych elementów ruchu ramion.

Podczas gdy transkrypcja mowy stanowi w pewnym stopniu jej interpretację, to w przypadku gestów sytuacja ta jest jeszcze bardziej wyrazista, gdyż ich zapis w przyjętym tutaj systemie wymaga szeregu działań analizujących oraz interpretujących. Nie jest to bowiem szczegółowy zapis trajektorii gestów, lecz jedynie przyporządkowanie ich do pewnych kategorii.

3.1. Transkrypcja ortograficzna

Nagrania są transkrybowane ortograficznie. W tej warstwie anotacji wyodrębniane są wyrazy tekstowe użyte przez nagrywane osoby. Zapis jest prowadzony zgodnie z zasadami współczesnej polskiej ortografii wyłożonymi w najnowszym słowniku ortograficznym języka polskiego (Polański 2006), pomijane są natomiast kwestie interpunkcyjne czy dotyczące segmentacji na wypowiedzenia (segmentacja syntaktyczna jest wykonywana w odrębnej warstwie anotacji). Każdy wyodrębniony wyraz jest zapisany dokładnie w takiej formie morfologicznej, w jakiej faktycznie został użyty przez mówiącego. W wypadku struktur słabo słyszalnych bądź takich, co do których pojawiają się wątpliwości „co zostało powiedziane" stosuje się procedurę polegającą na zapisywaniu słyszalnej części struktury; jeśli możliwa jest rekonstrukcja części niesłyszalnej, dopisuje się ją w nawiasie kwadratowym.

3.2. Transkrypcja fonologiczna

Zastosowano transkrypcję fonologiczną opartą na systemie wypracowanym w ramach projektu Pol'n'Asia. Jest to zmodyfikowana wersja znanego systemu SAMPA (Wells 1997), który w miejsce znaków transkrypcji IPA wykorzystuje znaki lub sekwencje znaków z podstawowego zestawu ASCII. Ułatwia to obróbkę i transfer danych, gdyż wiele programów i systemów nie obsługuje nadal w wystarczającym zakresie Unicode (systemu kodowania, w którym uwzględniono znaki transkrypcji fonetycznej IPA). Należy podkreślić, że systemy IPA i SAMPA są ze sobą zgodne pod tym względem, iż istnieje możliwość automatycznego przekonwertowania transkrypcji SAMPA na IPA lub w przeciwnym kierunku. Nie determinują one oparcia się na konkretnych modelach fonologicznych dla danego języka.

Przyjęty tutaj system transkrypcji określamy roboczo jako "fonologiczny", lecz trafniejsza wydaje się (niekiedy utożsamiana z nią) nazwa „szeroka transkrypcja fonetyczna", gdyż w zapisie transkrybujący starają się oddać w przybliżeniu słyszane segmenty wypowiedzi i unikać w miarę możliwości sugerowania się własną teoretyczną kompetencją fonologiczną mówcy języka polskiego. W konsekwencji, nie wyklucza się możliwości pojawienia się zapisów, które nie stanowią odzwierciedlenia poprawnej wymowy wyrazów języka polskiego, ani nawet odpowiadających jego systemowi fonologicznemu i fonotaktyce logatomów. Jest to możliwe szczególnie w przypadku różnego typu zaburzeń płynności wypowiedzi (zająknięcia, pauzy wypełnione).

Przyjętą metodą segmentacji sygnału jest metoda opisana w (Turk et al. 2006) przystosowana do potrzeb analizy prozodycznej. Najważniejszym kryterium takiej segmentacji jest wyznaczanie granic odpowiadających artykulacyjnym granicom samogłosek i spółgłosek. Przykładowo, jakikolwiek odcinek bezdźwięczności przypadający po spółgłosce bezdźwięcznej wybuchowej jest zaliczany do iloczasu następującej samogłoski. W częściej stosowanych metodach segmentacji śledzących granice akustyczne (stosowanych np. na potrzeby syntezy i rozpoznawania mowy) taki odcinek zostałby włączony do iloczasu spółgłoski.

4. Tagowanie

Przetranskrybowane ortograficznie i fonematycznie wypowiedzi mówione oraz zapis towarzyszących im gestów stanowił wyjściowy materiał do wielopoziomowych analiz językowych i komunikacyjnych. Fragmenty zapisu zostały szczegółowo opisane pod kątem leksyki, konstrukcji składniowych, prozodii oraz gestykulacji. Analizy te, w zależności od potrzeb, mogły różnić się stopniem szczegółowości. W większości przypadków skupiały się one na fragmentach wypowiedzi, w których pojawiała się gestykulacja. Poniżej przedstawiono systemy opisu dla poszczególnych podsystemów.

4.1. Leksyka, słowotwórstwo i składnia

Na potrzeby analiz zebranego korpusu nagrań dzieci i dorosłych wyodrębniono następujące poziomy opisu w warstwie leksykalno-składniowej:

- wypowiedzenia (podział toku narracyjnego na wypowiedzenia);
- schematy syntaktyczne zdań niezłożonych;
- schematy syntaktyczne zdań złożonych;
- części zdania;
- predykaty;
- argumenty;
- części mowy;
- wyrazy słowotwórczo podzielne;
- wyzyskane typy formantów.

Anotacje przeprowadzane są w programie Praat, który umożliwia wyodrębnianie wielu warstw anotacji, a także obserwowanie wszystkich poziomów anotacji jednocześnie (na jednym ekranie), co jest bardzo ważne podczas analiz porównawczych warstw leksykalno-składniowych z innymi warstwami anotacji: prozodyczną, intonacyjną.

4.1.1. Warstwa leksykalno-słowotwórcza

Na potrzeby analiz zjawisk z poziomu leksykalnego sporządzany jest korpus leksemów wykorzystanych w nagrywanych aktach komunikacji międzyludzkiej. Umożliwi to podjęcie kolejnego kroku badawczego – analizę systemu słowotwórczego i próbę odpowiedzi na pytanie czy sytuacja stresogenna (wynikająca z uczestniczenia w eksperymencie odbywającym się w wygłuszonym pomieszczeniu i przy sztucznie ograniczonym czasie wykonania zadania) sprzyja inwencji słowotwórczej, czy raczej ją gasi, a nagrywane osoby wyzyskują znane sobie leksemy. Wnioski z tych obserwacji pozwolą sformułować hipotezę dotyczącą rozwoju systemu słowotwórczego: czy równomiernie rozwija się on w odmianie pisanej i mówionej języka polskiego, a jeśli są różnice między tymi odmianami, to w jakim stopniu te warstwy języka różnią się od siebie w zakresie realizacji tego systemu. Ponadto interesujące jest także i to, czy pojawiające się w zebranym materiale językowym konstrukcje słowotwórcze powstają w wyniku analogii słowotwórczej, czy są rezultatem procesu derywacyjnego.

W warstwie leksykalnej, opisującej wyrazy tekstowe, czyli rzeczywiste realizacje leksemów w nagraniach, (punkt 7.) klasyfikuje się te wyrazy do odpowiednich klas części mowy i określa się wartości fleksyjne tych wyrazów. System znaczników wykorzystanych dla anotacji morfologicznej jest oparty na koncepcji tagowania tekstów polskich zaproponowanej przez A. Przepiórkowskiego i M. Wolińskiego (Przepiórkowski, Woliński 2003; Przepiórkowski, Woliński 2001; Woliński 2003). Jednakże ujęcie to wymaga modyfikacji,

ponieważ na potrzeby opisu materiału uznaje się fakty przemawiające za nie fleksyjnym, a słowotwórczym charakterem stopniowania przysłówków i przymiotników oraz także za nie fleksyjnym, a słowotwórczym charakterem aspektu czasownika; w wypadku i kategorii strony czasowników znów uznaje się jej niefleksyjność na rzecz składniowego charakteru zjawiska (za: Bańko 2002).

Sposób anotacji wygląda następująco: część mowy, do której należy dany leksem : informacje gramatyczne (np. przypadek, rodzaj, liczba itp.).

W słowotwórczej warstwie anotacji (punkty: 8. i 9.) wyróżnia się konstrukcje słowotwórcze — tu oznacza się wyzyskane techniki słowotwórcze i kategorie słowotwórcze, do których należą opisywane derywaty. Znakowanie konstrukcji słowotwórczych jest wykonywanie zgodnie z zasadami formalizacji opisu przyjętymi w opracowaniach gniazd słowotwórczych języka polskiego (Jadacka 1995; Jadacka 1988; Jadacka 2001-2005) oraz Mirosława Skarżyńskiego (Skarżyński 1989; Skarżyński 2003). Na potrzeby anotacji korpusu przyjęto zasadę rekonstrukcji jednego, bezpośrednio poprzedzającego taktu derywacyjnego, dla danego, odnotowanego w korpusie wyrazu podzielnego słowotwórczo. Włączanie derywatów do poszczególnych kategorii słowotwórczych zostało przeprowadzane zgodnie z przyjętymi dla języka polskiego opracowaniami teoretycznymi (Grzegorczykowa, Laskowski, Wróbel 1998; Grzegorczykowa 1984).

Sposób anotacji polega na wpisywaniu informacji słowotwórczych: kategoria słowotwórcza; funkcja formantu. W odrębnej warstwie anotacji rekonstruowana jest postac formantu słowotwórczego

4.1.2. Warstwa syntaktyczna

Na potrzeby analiz zjawisk z poziomu składniowego istotne jest zrekonstruowanie schematów składniowych (zarówno zdań pojedynczych, jak i złożonych), którymi posługują się badane osoby – a więc schematów mowy potocznej. Elementem łączącym analizę słowotwórczą i składniową jest rekonstrukcja struktury argumentowo-predykatowej opisywanych aktów komunikacji językowej, co jest konsekwencją przyjętego złożenia, że za operacją nazwotwórczą wyrażającą się czy to pojawieniem się neologizmu – derywatu, czy też wypowiedzenia (np. zdania) stoją te same struktury myślowe i procesy poznawczo-nazwotwórcze (Doroszewski 1962).

Zatem w syntaktycznej warstwie anotacji (punkty 1. – 6.) uwagę zwraca się na takie elementy, jak: wypowiedzenia i ich rodzaje. Oznacza sie poszczególne części zdania i elementy struktury predykatowo-argumentowej. Punktem wyjścia do anotacji na poziomie składniowym są opracowania teoretyczne Zygmunta Saloniego i Marka Świdzińskiego (Saloni, Świdziński 1998), a także kontynuującej ich myśl teoretyczną Renaty Grzegorczykowej (Grzegorczykowa 1994). Zebrany materiał empiryczny pozwoli podjąć próbę weryfikacji wypracowanego przez tych badaczy – na podstawie obserwacji ogólnego języka pisanego (dialektu kulturalnego) – modelu teoretycznego składni języka polskiego w odniesieniu do języka mówionego. Wypowiedzenia pojawiające się w nagranym materiale wyodrębnia się dzięki intonacji zdaniowej (a: Grzegorczykowa 1994).

Sposób anotacji polega na segmentowaniu ciągu wyrazów na wypowiedzenia i określaniu typów wypowiedzeń. Następnie oznacza się części zdań (wypowiedzeń) oraz rekonstruuje strukturę argumentowo-predykatową. System znaczników wykorzystanych do anotacji syntaktycznej również jest oparty na koncepcji A. Przepiórkowskiego i M. Wolińskiego (Przepiórkowski, Woliński 2003; Przepiórkowski, Woliński 2001; Woliński 2003).

Rezultatem obserwacji będzie pełny leksykalno-składniowy opis instruktażowego aktu komunikacyjnego. Pozwoli to określić sposoby wykorzystywania słownika współczesnej polszczyzny w sytuacji stresogennej (czy wyzyskuje się znane leksemy, czy raczej pojawiają się neologizmy słowotwórcze). Ponadto rezultatem może być stworzenie słownika minimum dla dialogu instruktażowego. Wnioski sformułowane na podstawie analiz językowych (leksykalno-składniowych) skonfrontowane z wnioskami wynikającymi z analiz gestowo-prozodycznych przeprowadzonych na zebranym materiale audiowizualnym umożliwią taki opis aktu dialogowego, w którym wyszczególnione zostaną wszystkie istotne komponenty tego aktu; możliwe będzie pokazanie ich synchronizacji, wzajemnej korelacji i hierarchii.

4.2. Prozodia

4.2.1. Podział na sylaby i frazy

Podział ciągłych wypowiedzi spontanicznych na sylaby napotyka liczne trudności. Z jednej strony są one związane z brakiem w pełni przekonywującego i spójnego systemu reguł sylabifikacji dla języka polskiego, z drugiej - z własnościami mowy, szczególnie w wymowie niestarannej. W wymowie niestarannej i niepoprawnej pojawiają się dodatkowe problemy związane przede wszystkim ze zjawiskami o charakterze lenicyjnym, prowadzącymi niekiedy nawet do „zaniku" całych sylab (np. /zatSasnow/ → /st^Sasnow/), rzadziej – fortycyjnym, mogącymi prowadzić do zmiany struktury sylabicznej, a nawet do pojawienia się nowych sylab (np. epenteza /brvi/ → /byryvi/). W bieżącym projekcie ogólny profil anotacji mowy charakteryzuje fonetyczna interpretacja sygnału. Wymienione zjawiska maja więc swoje odzwierciedlenie w transkrypcji i podziale sylab.

Granice sylab wyznaczane są według reguł przyjmujących model sylaby oparty na sonorności (Szpyra-Kozłowska 1998). Nie jest brany pod uwagę model który, przykładowo, nakazuje maksymalizować części nagłosowe w podziale na sylaby. Przyjęte zasady pozwalają na zamykanie sylab w przypadku środkowej zbitki zawierającej dwie spółgłoski, jak w słowach: „miasto" → mias.to, „mokry" → mok.ry. W słowach zawierających bardziej skomplikowane zbitki, zasadniczo prawidłowe są dwa podziały: „karczma" → karcz.ma lub kar.czma, „kartka" → kart.ka lub kar.tka. Przypadki takie jak „z okna", „w wodzie" traktowane są jako „zok.na" i „wwo.dzie", natomiast symetryczna sytuacja: „oko" jako o.ko.

Podział na frazy intonacyjne jest w jeszcze większym stopniu arbitralny. Mimo podjęcia licznych prób ich zdefiniowania (dla języka polskiego zob. np. (Demenko, Jassem 1997; Demenko 2000)), pozostają one nadal jednostkami o stosunkowo niepewnym statusie. Tutaj przyjęto reguły podziału na frazy przedstawione w (Karpiński 2006), które opierają się na wstępnym podziale tekstu na tzw. jednostki robocze. W ich wyodrębnieniu stosowane są kryteria „techniczne", które pozwalają objąć segmentacją również wypowiedzi o zaburzonej płynności i strukturze intonacyjnej. Uwzględnia się zarówno kryteria dotyczące struktury wewnętrznej, jak i te związane z charakterem granic międzyfrazowych. Frazy intonacyjne to poprawnie zrealizowane (well-formed) jednostki robocze. Dodatkowo, uwzględnia się dwupoziomowość systemu frazowego, zgodnie z sugestiami zawartymi w (Wagner 2008). Przyjmuje się, że obok fraz odpowiadających pojedynczym jednostkom roboczym, mogą istnieć również frazy obejmujące kilka takich jednostek - a zatem mamy tutaj do czynienia z prostą strukturą hierarchiczną. Frazy wyższego poziomu (Major Intonational Phrase) częściej odpowiadają „skończonym myślom" – wypowiedziom, które płynnie wyrażają pewną pojedynczą intencję. Frazy niższego poziomu (Minor Intonational Phrase) charakteryzują się

słabszymi granicami, prostszą budową wewnętrzną oraz tym, że często nie odpowiadają całej frazie syntaktycznej lub aktowi mowy, a jedynie realizacji pewnej części jednostki.

4.2.2. Prominencje i struktura rytmiczna

Jednym z głównych celów projektu jest dogłębna analiza struktury prozodycznej mowy dzieci i dorosłych. Analiza zależności iloczasowych oraz struktury rytmicznej opiera się na zintegrowaniu informacji fonetycznych i fonologicznych zawartych na trzech poziomach opisu: sylaby i frazy, intonacja oraz prominencje i struktura rytmiczna. W rezultacie, po kolejnym opracowaniu opisu wymienionych poziomów, otrzymujemy warstwy anotacyjne zawierające:

- długości sylab i fraz;
- kluczowe przebiegi intonacyjne;
- prominencje i granice metryczne.

W ten sposób możliwy jest opis i analiza w pierwszej kolejności zjawisk lokalnych, np.: liczby sylab lub pauz w jednostce czasu, a następnie zjawisk globalnych, np.: rozmieszczenia i zmienności iloczasu sylab w języku polskim (Ramus et al. 1999; Low et al. 2001; Dellwo 2006). Uzyskanie takich parametrów umożliwia opis struktury rytmicznej mowy zgodnie z aktualnymi metodami. Porównanie rytmu mowy z „prozodią gestu" – jego iloczasem, intensywnością oraz typowymi strukturami, jest dosyć trudne ze względu na fundamentalne różnice w sposobie „artykułowania" jednostek oraz ich funkcji. Typ gestu, którego analiza jest możliwa przy użyciu metod podobnych do tych stosowanych w analizie prozodii mowy, to gesty zwane przez McNeilla „uderzeniami" („beats") (McNeill 1995). Uderzenia podkreślają zwykle istotne prozodycznie sylaby i często korelują z akcentem. Posiadają także powtarzalną i dosyć periodyczna strukturę. Dzięki anotacji zmiany ruchu rąk (patrz sekcja 4.3.1) ściślejsze porównanie takich rytmów w obu modalnościach jest możliwe. Analiza współwystępowania gestów innych niż uderzenia, tj. gestów niosących znaczenie czy też ogólnie, analiza komunikacyjnie istotnej koordynacji ruchów rak i głowy ze strukturami rytmicznymi, jest dokonywana obecnie w sposób opisowy, bez wykorzystywania parametrów statystycznych.

Sposób anotacji warstw zawierających informacje fonetyczne oparto na bezpośredniej analizie sygnału akustycznego. W takim wypadku często stosuje się automatycznie realizowane algorytmy, jednak w bieżącym projekcie warstwa pierwsza opisu jest tworzona ręcznie, według wskazówek dla anotacji segmentalnej na potrzeby analizy prozodycznej oraz reguł sylabifikacji ustalonych dla polskiego (patrz. 4.2.1). Warstwa druga, szczegółowo opisana w rozdziale 4.2.3, zawiera reprezentację przebiegów intonacyjnych.

W warstwie 3, zawierającej większość informacji fonologicznej, wyróżniamy dwie kategorie zjawisk i odpowiadających im grup symboli: prominencje i granice metryczne. Warstwa ta jest opracowywana na podstawie wielokrotnego odsłuchu materiału w programie Praat przy częstym użyciu ułatwień technicznych pozwalających na wybieranie i porównywanie szerszych lub węższych fragmentów narracji. Opisujący polega w większości na własnej percepcji mowy, ale konsultacja decyzji dotyczących opisu istotnych granic i prominencji z anotacją zjawisk intonacyjnych jest zalecana, czasem konieczna, w kontrowersyjnych przypadkach. Nie określono dotąd jednoznacznie korelatów akcentu w języku polskim, wiadomo jednak, że intonacja gra o wiele większa rolę w kształtowaniu struktury prominencji we frazie niż np. iloczas akcentowanej sylaby (Dukiewicz, Sawicka 1995). Często jednak, opisujący może napotkać trudności z interpretacją prominencji szczególnie w pobliżu granic prozodycznych. Na przykład, we współczesnej polszczyźnie i w

mowie dzieci, koniec frazy często charakteryzuje nagle wznosząca się, niepytająca intonacja padająca na ostatnia sylabę. Często taka sylaba jest nawet do 5 razy dłuższa od poprzedzającej. Takie wydłużenie oraz kształt przebiegu intonacyjnego może prowadzić do konkurencji między sylabą naznaczoną potencjalnie jako prominentna, tj. przedostatnią w danym słowie, ale percepcyjnie słabszą od ostatniej. Poniższy opis tagów wykorzystywanych w anotacji warstwy nr 3 wyjaśnia m.in. sposób anotacji takich przypadków.

System tagowania jest oparty na systemie Rhythm and Pitch (Dilley 2005). Granice fraz metrycznych definiowane są według dwóch kategorii:

1. „)" jeden nawias prawy oznacza wyraźnie słyszalną przerwę w płynnym toku mowy. Takie przerwy mogą być interpretowane jako zawahania, jako umyślne, stylistyczne modyfikowanie struktury rytmicznej, jako przerwa na oddech itp. Często funkcja granicy „)" nie jest jasna, dlatego opis tego typu granic jest niezależny od ich interpretacji; w tej kategorii ekspert anotator polega wyłącznie na swojej percepcji ciągu mowy. Granice typu „)" nie zawsze korelują z granicami fraz intonacyjnych.
2. „))" dwa nawiasy prawe oznaczają istotną, wyraźnie słyszalną pauzę, która pozwala jednoznacznie na interpretację jej funkcji i jest dłuższa od sąsiadujących granic, oznaczonych jako „)". Takie granice fraz metrycznych często korelują z granicami fraz intonacyjnych.

Prominencje definiowane sa zasadniczo dwojako:

1. „X" oznacza sylabę o percepcyjnie najistotniejszej prominencji w ramach frazy, najczęściej frazy metrycznej typu „)". Często koreluje z intonacyjnym akcentem frazowym.
2. „x" oznacza sylabę o słabszej od „X" prominencji w ramach frazy metrycznej. Często pokrywa się z potencjalnym akcentem leksykalnym. Jednak przypadek wyjątkowy, konkurencji sylaby ostatniej w słowie (i często, we frazie) z przedostatnią, opisany powyżej, jest interpretowany najczęściej na korzyść sylaby ostatniej.

Sylaby najsłabsze, nieprominentne, nie są oznaczane. W razie wątpliwości co do istotności danej prominencji używamy znaku „x?" lub „X?".

4.2.3. Intonacja

Analiza i opis intonacji jest realizowany, zależnie od potrzeb badawczych, dwoma grupami technik, tj. instrumentalnie i percepcyjnie.

a. Instrumentalna ekstrakcja częstotliwości podstawowej, wyznaczanie jej istotnych wartości w określonych punktach i przedziałach czasowych.
b. Instrumentalna ekstrakcja częstotliwości podstawowej i stylizacja konturów intonacyjnych na podstawie modelowania percepcji tonalnej.
c. Odsłuch, który - w zależności od celu analizy - może przybrać charakter "odsłuchu wnikliwego", z wykorzystaniem ułatwień technicznych (odsłuchiwanie po fragmencie, w zwolnionym tempie, itd.) oraz "odsłuchu naturalnego", w warunkach zbliżonych do normalnych warunków komunikacyjnych i bez technicznych ułatwień.

W przypadku techniki (a) wykorzystuje się program Praat (Boersma, Wenink 2008), w przypadku (b) Praat ze skryptem Prosogram (Mertens 2004). W przypadku (c) Praat służy jako odtwarzacz sygnałów lub ich fragmentów. Program umożliwia zestawianie wybranych sylab w celu percepcyjnego wychwycenia kontrastów intonacyjnych, spowolnione oraz przyspieszone odtwarzanie sygnału. Może posłużyć zarówno do odsłuchu eksperckiego, jak i przygotowania materiału do badań odsłuchowych z udziałem niespecjalistów.

Pierwszy system wykorzystuje się m.in. tam, gdzie w analizach istotne są kluczowe punkty przebiegów intonacyjnych (a więc głównie lokalne ekstrema przebiegu fo). Drugi pozwala uzyskać stylizowane reprezentacje przebiegów intonacyjnych, możliwe do wykorzystania w resyntezie i łatwiejsze w dalszej analizie niż "surowe" przebiegi fo, a przede wszystkim - lepiej oddające te elementy, które są istotne percepcyjnie dla słuchacza. Trzecia technika pozwala w pierwszym wariancie dokonywać opisu wypowiedzi w kategoriach percepcyjnych, tj. reprezentować "melodię wypowiedzi" jako szereg spadków i wzrostów. Zmienność przebiegu fo opisuje się tutaj globalnie (np. "kontur opadający") lub lokalnie, z reguły w domenie sylaby. W drugim przypadku możliwe jest stosowanie jednoetykietowych opisów, odzwierciedlających przebieg melodii: wznosząca, opadająca, opadająco-wznosząca, wznosząco-opadająca, równa, (równa) niska, (równa) wysoka. Taki opis jest częściowo kompatybilny z większą liczbą systemów, m.in. opartych na szkole brytyjskiej, a w pewnym stopniu pozwala również na zestawienia z materiałem opisanym w ToBI. Opis ten operuje jednak na poziomie percepcyjno-fonetycznym. Z jednej strony jest zdeterminowany własnościami percepcji ludzkiej, z drugiej zaś kompetencją (przygotowaniem teoretycznym oraz treningiem) anotującego.

4.3. Transkrypcja zachowań niewerbalnych

Różnice w systemach transkrypcji zachowań niewerbalnych wynikają zazwyczaj z potrzeby dostosowania ich do wymogów analizy danych. Różnice te dotyczą m. in. stopnia precyzji opisu (np. transkrypcja trzech sekund materiału filmowego w systemie FORM (Martell 2005) jako jednym z najbardziej pracochłonnych i czasochłonnych trwa około jednej godziny), podejścia do kwestii wyodrębniania podstawowych jednostek opisu (podejście funkcjonalne (McNeill 1995; Kendon 2005), formalne (Gut 2002, Yassinik et al. 2001), sposobu kodowania danych oraz zapisu (w postaci tekstu (McNeill 1995) lub na poszczególnych ścieżkach za pomocą odpowiedniego oprogramowania).

System transkrypcji zachowań niewerbalnych wykorzystywany na potrzeby projektu PCNC bazuje na systemie opisu gestykulacji opracowanym przez McNeilla (1995). Został on jednak dostosowany oraz rozbudowany zgodnie z potrzebami wynikającymi z charakteru prowadzonych badań; w sposób możliwie jak najbardziej precyzyjny (film oglądany jest klatka po klatce) wyznaczane są czasowe granice każdego ruchu ręki, Fraz gestykulacyjnych oraz składających się na nie Faz, a do opisu danych wykorzystany został program ELAN (powstały w MPI w Nijmegen) pozwalający na opisanie poszczególnych cech zachowań niewerbalnych na osobnych ścieżkach. Uwzględniane elementy opisu to:

- Ruch prawej i lewej ręki;
- Fazy gestu;
- Frazy gestykulacyjne;
- Relacja gestu względem mowy;
- Rodzaje gestów;
- Punkt widzenia realizowanych gestów;
- Strefa gestykulacyjna.

W przypadku, gdy obie ręce realizują równocześnie różne gesty, wtedy Fraza gestykulacyjna odpowiednia dla każdej z rąk zapisywana jest na osobnej ścieżce. Cechy gestów realizowanych przez obie ręce jednocześnie (punkt 4, 5, 6, 7) zapisywane są na jednej ścieżce (pierwszy składnik zapisu dotyczy lewej ręki, drugi prawej).

System stosowany na potrzeby projektu PCNC nie pozwala na odtworzenie gestów na podstawie samego zapisu, ponieważ pod uwagę brane są tylko wybrane elementy charakteryzujące formę gestu jak i jego trajektorię.

W opisie danych stosowana jest anglojęzyczna terminologia zaczerpnięta z systemu opisu zachowań niewerbalnych McNeilla (1995).

4.3.1. Ruch prawej i lewej ręki

Pierwszym etapem transkrypcji jest opisanie na osobnych ścieżkach ruchów prawej i lewej ręki, które wyznaczane są w większości przypadków przez zmiany ich kierunku. Istnieją jednak sytuacje, w których ruch ręki w tym samym kierunku rozdzielony jest poprzez pauzę. Wyróżnione zostają wtedy dwa oddzielne ruchy. Szczególnym przypadkiem jest również ruch kolisty, który – ponieważ zmiana kierunku ma charakter ciągły – jest traktowany jako jeden ruch.

Zazwyczaj aktywna jest jedna ręka, a w przypadku jednoczesnej aktywności obu rąk, ich gestykulacja jest najczęściej symetryczna lub równoległa. W takich sytuacjach opisywany jest ruch jednej z rąk, a na ścieżce przeznaczonej do opisu drugiej ręki zaznaczony zostaje odcinek, w ramach którego granice ruchu są identyczne jak dla ręki opisywanej. Rzadko zdarza się (chociaż odnotowano takie przypadki w materiałach bieżącego projektu), aby dwie dłonie realizowały jednocześnie dwie powiązane ze sobą treściowo, lecz różniące się formą Frazy gestykulacyjne. W takich przypadkach ruchy każdej ręki opisywane są w niezależnych warstwach anotacji.

4.3.2. Fraza gestykulacyjna oraz jej elementy

Za podstawową jednostkę opisu gestykulacji przyjęto Frazę gestykulacyjną składającą się z odpowiednich Faz. Termin Fraza gestykulacyjna oraz Gest stosowane są wymiennie. Gest rozumiany jest jako forma ruchu opisywana przez Frazę gestykulacyjną pozostająca w relacji znaczeniowej i pragmatycznej z wypowiadanymi słowami. Opisywana na podstawie narracji dzieci i dorosłych struktura Frazy gestykulacyjnej opiera się na modelu Kendona (2005), uwzględniającym trzy podstawowe Fazy: Przygotowanie, Uderzenie [2] oraz Cofnięcie. Struktura Frazy Kendona poszerzona została o Wstrzymanie przed uderzeniem i Wstrzymanie po uderzeniu (Kita 1998) oraz Uderzenie wstrzymane (McNeill 2007). Warunkiem istnienia Frazy gestykulacyjnej jest Uderzenie, wokół którego skupione są pozostałe Fazy będące opcjonalnymi elementami składającymi się na Frazę. Na każdą Fazę z wyjątkiem Uderzenia, które może składać się ze zwielokrotnionego ruchu, składa się zazwyczaj jeden ruch (z obserwacji własnych wynika, że jedynie w pojedynczych przypadkach pojawiają się wątpliwości dotyczące zakwalifikowania kilku ruchów jako Fazy przygotowania lub cofnięcia). Według Kippa (2004) pomocne w klasyfikacji poszczególnych Faz jest uwzględnienie dwóch warunków determinujących wyznaczenie ich granic. Są to: nagła zmiana kierunku ruchu oraz zmiana szybkości ruchu zarówno przed jak i po zmianie kierunku ruchu. Jeśli zachodzi tylko pierwszy warunek mamy do czynienia z Uderzeniem wielokrotnie złożonym.

[2] ang. *beat*

Rys. I przedstawia model Frazy gestykulacyjnej uwzględniany w systemie DiaGest.

4.3.3. Rodzaje gestów

Na podstawie klasyfikacji zachowań niewerbalnych McNeilla (2007) wyodrębniono gesty ikoniczne, metaforyczne, bity oraz deiktyczne. Nie uwzględniono w opisie zachowań będących formą adaptatorów. W przypadkach, w których gest nie występował w czystej postaci, opisywano cechy poszczególnych kategorii zawarte w jednym geście (np. iconic/deictic).

4.3.4. Związki między gestem a mową

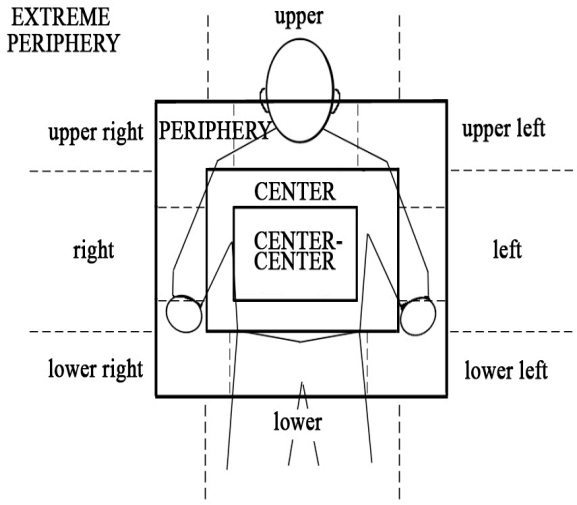
Uwzględnione zostały trzy rodzaje relacji zachodzących między gestem a werbalną warstwą wypowiedzi: komplementarność, koekspresywność oraz redundancja (McNeill 2007). Dotyczą one tego, w jakim stopniu treści wnoszone przez gest pokrywają się z treściami wyrażanymi przez słowa. Wskazanie wyraźnej granicy między tymi trzema rodzajami relacji jest niemożliwe, gdyż praktycznie każdy gest wnosi pewne treści nie wypowiedziane przez słowa. Umownie przyjmuje się jednak, że koekspresywność dotyczy sytuacji, w których zachowania niewerbalne odnoszą się do wydarzeń, obiektów opisywanych przez słowa, a więc do przynajmniej częściowego pokrywania się treściowego, np. osoba opowiadająca bajkę w trakcie wypowiadania słów „wszedł do rury" wykonuje ruch naśladujący wejście do rury. Komplementarność natomiast zachodzi wtedy, gdy gest odnosi się do kwestii, które nie zostały w żaden sposób uwzględnione przez słowa np. dziecko opowiadające bajkę mówi: on rozhuśtał się na linie i tak … (tu realizowany jest gest pokazujący co wydarzyło się w konsekwencji huśtania się). Natomiast redundancja dotyczy pokrywania się treściowego w warstwie werbalnej i niewerbalnej.

4.3.5. Punkt widzenia zachowań niewerbalnych

W opisie gestykulacji brany jest pod uwagę punkt widzenia realizacji gestów. Za przykładem systemu McNeilla wyróżniono dwa punkty widzenia przedstawianych w sferze niewerbalnej sytuacji: Punkt Widzenia Bohatera (*Character Viewpoint*, CVPT) i Punkt Widzenia Obserwatora (*Observer Viewpoint*, OVPT). Określają one dystans między narratorem a prowadzoną przez niego narracją. W przypadku zachowań realizowanych z Punktu Widzenia Bohatera osoba relacjonująca wydarzenia wciela się w ich bohatera, a tym samym zaangażowane w gestykulację części ciała narratora reprezentują odpowiednie części ciała bohatera zdarzeń. Szczególny rodzaj gestów realizowanych z Punktu Widzenia Bohatera nazwanych Ciało Jako Punkt Odniesienia (*Body as Reference Point*, *BARP-gestures*) został wyodrębniony przez Holler i Beattie (2002). Narrator wskazując na część swojego ciała wskazuje tym samym na część ciała bohatera. Kiedy gest realizowany jest z Punktu Widzenia Obserwatora zwiększa się dystans do relacjonowanych wydarzeń, a dłoń narratora reprezentuje najczęściej całą postać bohatera.

4.3.6. Strefa Gestykulacyjna

Jeden z poziomów transkrypcji zachowań niewerbalnych koncentruje się na przestrzeni, w której realizowane są gesty. Do opisania właściwego poszczególnym zachowaniom niewerbalnym miejsca ich realizacji wykorzystany został schemat McNeilla (1995) składający się z koncentrycznych kwadratów wyznaczających sfery gestykulacyjne. Ponieważ *Fraza gestykulacyjna* stanowi odcinek rozciągnięty w czasie, dlatego dla określenia miejsca realizacji gestu pod uwagę brana jest najistotniejsza z semantycznego oraz strukturalnego punktu widzenia jej część, czyli *Uderzenie*. Ponieważ samo *Uderzenie* również nie jest punktem w czasie, charakteryzuje go pewien przebieg, dlatego uwzględniany jest zarówno jego początek jak i koniec, co pozwala na opisanie amplitudy *Uderzenia*.



Rys. II przedstawia schemat określający przestrzeń gestykulacyjną.
Rysunek bazuje na schemacie McNeilla (1995).

5. Akty dialogowe

Podział materiału na realizacje multimodalnych aktów dialogowych przebiega na podstawie zasad opisanych m.in. w (Karpiński 2009) dla systemu DiaGest. System ten powstał na podstawie doświadczeń z wcześniejszych projektów badawczych (np. Pol'n'Asia, (Karpiński 2006)) oraz nowych opracowań z tego zakresu, szczególnie zaś publikacji Bunta, Popescu-Belisa, Trauma i Allena (Bunt 2006, 2008; Popescu-Belis 2008; Traum 2000). Akt dialogowy rozumie się tutaj jako twór wielowarstwowy. Pojedyncza wypowiedź może stanowić realizację wielu funkcji dialogowych (lub wielu aktów - jeśli przyjąć taką terminologię). Stąd koncepcja wymiarów aktu dialogowego jako możliwych do wyodrębnienia obszarów jego oddziaływania. W systemie DiaGest przyjęto, że funkcje wypowiedzi będą rozpatrywane w czterech wymiarach:

1. Task Action Control (TAC)
2. Dialogue Flow Control (DFC)
3. Information Transfer Management (ITM)
4. Approach Expression Marker (AEM)

System jest przeznaczony do opisu komunikacji skupionej na zadaniach, dlatego też wymiar TAC jest bezpośrednio związany z realizacją zadania. Obejmuje on przede wszystkim polecenia, ich objaśnienia, potwierdzenia ich zrozumienia oraz realizacji. DFC to wymiar odnoszący się do przebiegu dialogu i związany głównie z mechanizmem zabierania głosu. W wymiarze ITM jest opisywany kierunek i charakter przepływu informacji. Wymiar AEM poświęcono natomiast opisowi możliwych do zaobserwowania przejawów nastawienia nadawcy do tematu wypowiedzi oraz do rozmówcy. W każdym wymiarze, obok zbioru specyficznych dla niego wartości, wprowadzono wartości "irrelevant" (dla sytuacji, gdy dana wypowiedź nie oddziałuje w danym wymiarze) oraz "other" (gdy wypowiedź oddziałuje w danym wymiarze, lecz oddziaływaniu temu nie odpowiada żaden z dostępnych tagów). W zestawieniu ze złożonym systemem DIT++ (Bunt 2008) zrezygnowano tutaj z bardziej drobnoziarnistego rozgraniczania wymiarów (tam: grup funkcji), jak i z wyróżnienia funkcji specyficznych dla wymiarów oraz funkcji "ogólnego zastosowania". Podział ten jest tutaj konceptualnie akceptowany, lecz z przyczyn praktycznych właściwe wydało się jego uproszczenie. Jest to tym bardziej uzasadnione, że w przedstawianym projekcie badania skupiają się na ograniczonym inwentarzu kategorii aktów dialogowych, które są najistotniejsze dla badanego typu wypowiedzi (wspomagana przez słuchacza narracja) i dostatecznie często pojawiają się w zgromadzonym materiale. Wprawdzie system Pol'n'Asia został zaprojektowany do opisu dialogów zadaniowych, jednak użycie jego zmodyfikowanej wersji w odniesieniu do analizowanych tutaj sesji narracyjnych jest możliwe z dwóch przyczyn. Po pierwsze, nie mamy do czynienia z narracją w pełni monologową, lecz narracją w dużej mierze interaktywną. Po drugie, system zawiera pewne ogólne elementy semantyczno-pragmatycznego opisu wypowiedzi, które można wykorzystać rownież w odniesieniu do wypowiedzi o charakterze monologowym. W tym przypadku niektóre wymiary systemu będą miały mniejsze znaczenie lub pozostaną nierelewantne, lecz pozostałe wymiary mogą dostarczyć wystarczającej dozy informacji o intencjonalnej warstwie wypowiedzi.

6. Zgromadzony materiał i jego zastosowania

Transkrypcja oraz stosunkowo szczegółowe, wielowarstwowe otagowanie wybranych partii korpusu stanowi zadanie niezwykle pracochłonne. Jednak uzyskany w ten sposób materiał

pozwoli na przeprowadzenie wielu analiz w sposób częściowo lub całkowicie zautomatyzowany. Dotyczy to szczególnie współwystępowania pewnych zjawisk w poszczególnych warstwach przekazu, tj. obecności danych kategorii gestów, jednostek leksykalnych, zjawisk prozodycznych, konstrukcji składniowych w obrębie realizacji poszczególnych aktów dialogowych. Możliwe będzie również prowadzenie częściowo zautomatyzowanych analiz sekwencji występowania pewnych jednostek (np. w kategoriach bi-, trigramów lub skupień). Włączenie do opisu warstwy pragmatyczno-semantycznej (akty dialogowe) umożliwi zbadanie wkładu poszczególnych modalności i kanałów w ostateczny przekaz danej wypowiedzi.

Bibliography:

Bańko M., 2002. Wykłady z polskiej fleksji, Warszawa.

Boersma, Wenink 2008. Praat: Doing Phonetics by Computer. A Computer Programme. (wersja 3.0 i późniejsze; pobrane z http://www.praat.org)

Bunt, H. 2006. Dimensions in dialogue act annotation. (w:) *Proceedings of the Fifth International Conference on Language Resources and Evaluation LREC 2006*.

Bunt, H. 2008. DIT++ Taxonomy of dialogue acts. Release 3, version 2, February 8, 2008. Dostępne na stronie internetowej http://dit.uvt.nl)

Dellwo,V. (2006). Rhythm and Speech Rate: A Variation Coefficient for deltaC. in Karnowski,P., Szigeti,I. (ed.) Language and language-processing. Frankfurt am Main: Peter Lang, 231–241.

Demenko, G. 2000. Automatyczna analiza granicy frazy w języku polskim (w:) W. Jassem, Cz.Basztura, W. Jassem (red.) Speech and Language Technology, vo. IV, Poznań, str. 13–22.

Demenko, G., Jassem, W. 1997. Phonetic and syntactic coherence of the phrase (w:) W. Jassem, Cz. Basztura (red.) *Speech and Language Technology*, vol. I, Wrocław, str. 125–139.

Dilley, L., Brown, M. (2005) The RaP (Rhythm and Pitch) Labelling System. http://tedlab.mit.edu/tedlab_website/RaP%20System/RaP_Labeling_Guide_v1.0.pdf

Doroszewski W., 1962. *Kategorie słowotwórcze*, w: *Studia i szkice językoznawcze*, Warszawa.

Dukiewicz, L, Sawicka, I. (1995) Fonetyka i fonologia. Gramatyka współczesnego języka polskiego. Kraków: Wydawnictwo Instytutu Języka Polskiego PAN.

Grzegorczykowa R., 1984, *Zarys słowotwórstwa polskiego. Słowotwórstwo opisowe*, Warszawa

Grzegorczykowa R., 1994, *Wykłady z polskiej składni*, Warszawa.

Grzegorczykowa R., Laskowski R., Wróbel H., (red.), 1998, *Gramatyka współczesnego języka polskiego*, t. II: *Morfologia*, Warszawa.

Gut, U., Looks, K., Thies, A., Trippel, T., Gibbon, D., 2002. CoGesT. Conversational Gesture Cranscription System. Version 1.0. Technical Report. University Bielefeld.

Holler, J., Beattie, G., 2002. A Micro-Analytic Investigation on How Iconic Gestures and Speech Represent Core Semantic Features in Speech. Semiotica 142 – ¼, 31–69

Jadacka H., (red.), 2001-2005, *Słownik gniazd słowotwórczych współczesnego języka ogólnopolskiego*, t. I-IV, Kraków.

Jadacka H., 1988, *Zeszyt próbny słownika gniazd słowotwórczych współczesnego języka polskiego*, Warszawa.

Jadacka H., 1995, *Rzeczownik polski jako baza derywacyjna. Opis gniazdowy*, Warszawa.

Karpiński, M. 2006. Struktura i intonacja polskiego dialogu zadaniowego. Poznań: Wydawnictwo Naukowe UAM.

Karpiński, M. 2009. From Speech and Gestures to Dialogue Acts. (In:) A. Esposito, A. Hussain, M. Marinaro, R. Martone (red.) *Multimodal Signals: Cognitive and Algorythmic Issues*. Seria LNAI 5398, Berlin - Heidelberg - New York: Springer Verlag, str. 164–169.

Kendon, A. 2005, Gesture: Visible Action as Utterance. Cambridge University Press.

Kipp, M.: Gesture Generation by Imitation. From Human Behavior to Computer Character Animation. Dissertation.com, Boca Raton (2004)

Kita, S., Gijn van, I., Hulst van der, H.: Movements phases in signs and co-speech gestures, and their transcription by human coders. In: Wachsmuth, I., Fröhlich, M. (eds.) Gesture and Sign Language in Human-Computer Interaction. Springer Verlag, Berlin (1998) 23–35

Low, E.L., Grabe, E. and Nolan, F. (2001). Quantitative characterisations of speech rhythm: Syllable-timing in Singapore English. Language and Speech 43 (4), 377–401.

Martell, C. 2005. FORM: An Experiment in the Annotation of the Kinematics of Gesture. Dissertation.

Mertens, P. 2004. The Prosogram: Semi-Automatic Transcription of Prosody based on a Tonal Perception Model. In: B. Bel, I. Marlien (red.) Proceedings of Speech Prosody 2004, Nara, Japan.

McNeill, D. 1995. Hand and Mind: What Gestures Reveal about Thought. The University of Chicago Press.

McNeill, D. 2007. Gesture and Thought. The University of Chicago Press, Chicago London.

Mertens, P. & Alessandro, Ch. d' 1995. Pitch contour stylization using a tonal perception model. Proc. Int. Congr. Phonetic Sciences 13, 4, str. 228-231.

Popescu-Belis, A. 2008. Dimensionality of Dialogue Act Tagsets: An Empirical Analysis of Large Corpora. Language Resources and Evaluation 42(1), str. 99-107.

Przepiórkowski A., Woliński M., 2001, *Projekt anotacji morfosyntaktycznej korpusu języka polskiego*, Warszawa.

Przepiórkowski A., Woliński M., 2003, *A Flexemic Tagset for Polish*, w: *The Proceedings of the Workshop on Morphological Processing of Slavic Languages, EACL.*

Ramus, F., Nespor, M., & Mehler, J. "Correlates of linguistic rhythm in the speech signal". Cognition, 73(3), 265–292, 1999.

Saloni Z., Świdziński M., 1998, *Składnia współczesnego języka polskiego*, wyd. czwarte, zmienione, Warszawa.

Skarżyński M., (red.), 2003, *Słowotwórstwo gniazdowe. Historia. Metoda. Zastosowania*, Kraków.

Skarżyński M., 1989, *Mały słownik słowotwórczy języka polskiego dla cudzoziemców*, Kraków.

Polański E., (red.), 2006, *Wielki słownik ortograficzny PWN z zasadami pisowni i interpunkcji*, Warszawa.

Szpyra-Kozłowska, Jolanta. 1998. The sonority scale and phonetic syllabification in Polish. Biuletyn Polskiego Towarzystwa Językoznawczego.Zeszyt LIV, str. 63–79.

Traum, D. 2000. Twenty Questions on Dialogue Act Taxonomies. Journal of Semantics 17(1), 7–30.

Turk A., Nakai S., Sugahara M. 2006 Acoustic Segment durations in prosodic research: a practical guide. Methods in Empirical Prosody Research, Berlin: Mouton de Gruyter.

Wagner, A. 2008. A comprehensive model of intonation for application in speech synthesis. Unpublished PhD thesis, Adam Mickiewicz University, Poznan, Poland.

Wells, J. C. 1997. SAMPA computer readable phonetic alphabet. (w:) Gibbon, D., Moore, R. i Winski, R. (red.) *Handbook of Standards and Resources for Spoken Language Systems*. Berlin and New York: Mouton de Gruyter (część IV, sekcja B).

Woliński M., 2003, *System znaczników morfosyntaktycznych w korpusie IPI PAN*, „Polonica" XXII-XXIII, s. 39–55.

Yassinik, Y., Renwick, M., Shattuck-Hufnagel, S., 2004. The timing of speech-accompanying gestures with respect to prosody. Acoustical Society of America Journal 115 (5), pp. 2397-2397