Urszula Paprocka-Piotrowska
Instytut Filologii Romańskiej, KUL

TRANSKRYPCJA, KODOWANIE I ANALIZA DANYCH JĘZYKOWYCH W SYSTEMIE CHILDES

Transcription, coding and analysis of data in the CHILDES system

The Childes Language Data Exchange System (CHILDES), composed of Codes for Human Analysis of Transcripts (CHAT) and Computerized Language Analysis (CLAN), offers the possibility of transcription, grammatical coding and analysis of transcripts of language data, which can be used in psycholinguistics research. The CHILDES system enables us to code language data and provides easy and fast access to automatically-generated sets, such as, for example, statistical information concerning the frequency of words, the contexts in which they appear, syntactic properties, combination of keywords as well as type/token ratio. The paper briefly presents the CHILDES, with a particular emphasis on those of its functions which can be employed in research on language acquisition.

1. Wstęp

Opracowany i rozpowszechniony przez MacWhinneya (MacWhinney 2000; por. Champaud 2001) system CHILDES (ang. *Child Language Data Exchange System*) składa się z trzech komponentów:

a) zasad transkrypcji i kodowania danych CHAT (ang. *Codes for the Human Analysis of Transcipts*);
b) edytora tekstu CLAN (ang. *Computerized Language Analysis*), w którym niektóre procedury transkrypcji oraz kodowania danych są zautomatyzowane i pozwalają na połączenie transkrybowanych plików z plikami video oraz plikami dźwiękowymi zapisanymi w wersji cyfrowej;

c) banku danych (ang. *Database*): zarejestrowane oraz transkrybowane wypowiedzi dzieci (najczęściej spontaniczne) pochodzące z około 30 języków; korpus zawiera także dane zebrane u dzieci z zaburzeniami rozwoju mowy, u osób dwujęzycznych, u dorosłych dotkniętych afazją oraz u uczących się języków obcych.

Elementy te uzupełnia baza danych bibliograficznych dotyczących akwizycji języka, zawierająca ponad 30 000 haseł. Pełna informacja o systemie CHILDES oraz jego komponentach znajduje się na stronach www: http://childes.psy.cmu.edu (strona główna systemu) oraz http://www.cnts.ua.ac.be/childes (europejska strona CHILDES, moderowana przez Uniwersytet w Antwerpii). Obie strony proponują też wpis na listę mailingową *Info-CHILDES*. Dane zapisane w banku danych są regularnie aktualizowane, a dostęp do nich jest nieograniczony, wszystkie skopiować można ze strony internetowej CHILDES, zachowując zasadę *data sharing*, wedle której zachowana jest anonimowość informatorów, a obowiązkowo cytowane są nazwiska badaczy, którzy dane zebrali, udostępnili oraz wyrazili zgodę na ich użycie (np. jeśli stanowią one materiał opracowany w publikacjach).

2. CHAT: Codes for the Human Analysis of Transcripts

CHAT to metoda zapisu i transkrypcji danych językowych opierająca się na trzech zasadach:

1. Każde wypowiedzenie (enuncjacja) zapisane jest w systemie jako odrębne hasło. W wypadku, gdy lokutor produkuje (wypowiada) kilka następujących po sobie wypowiedzeń (tak dzieje się na przykład w wypadku nagrywania i transkrybowania narracji), każde z nich odpowiada jednej tak zwanej *linii głównej* (ang. *main line*); ważne jest to ze względu na możliwe do zapisania informacje dodatkowe oraz kody odnoszące się zawsze i jedynie do danej linii głównej. Każda linia główna zaczyna się od gwiazdki (*), po niej następuje trójliterowy symbol (zapisany drukowanymi literami) oznaczający lokutora, następnie dwukropek, tabulacja, wypowiedzenie zakończone obowiązkowo znakiem interpunkcyjnym. Wewnątrz wypowiedzenia (wewnątrz linii głównej) nie należy stosować żadnych znaków interpunkcyjnych. Przykładem może być następujący zapis:

   *PIO: była zima.

2. Informacje dodatkowe, ewentualne komentarze oraz linie kodów wprowadzane są za pomocą *linii zależnych* (ang. *dependent tiers*), zapisywane tuż pod linią główną, której bezpośrednio dotyczą. Linie zależne rozpoczynają się symbolem %, po nim następuje trójliterowy symbol kodu linii (zapisany małymi literami), dwukropek, tabulacja oraz zawartość linii. System CHILDES zawiera już pewną ilość predefiniowanych linii zależnych, i tak na przykład linia oznaczona symbolem *%mor* zawierać będzie kody do-

tyczące morfologii i składni, linia *%com* – komentarz transkrybującego, a linia *%act* – komentuje działania mówiącego. Poniżej podajemy przykład takiego zapisu:

*PIO: była zima.
%mor: 3SF V | być-PAST | zima
%com: z wahaniem
%act: wskazuje palcem na obrazek.

Użytkownicy systemu mogą tworzyć na potrzeby swoich badań własne linie zależne, kodując tym samym interesujące dla nich zjawiska, jak poniżej:

*PIO: była zima.
%cli: exi_być

W powyższym przykładzie, stworzona przez użytkownika (tu: autor tekstu) linia nosząca kod *%cli* odnosi się to typu użytego czasownika (czasownik fazowy, czasownik ruchu, czasownik posiłkowy, czasownik egzystencjalny itp.), a kod *exi_xxx* oznacza zdanie egzystencjalne (inne praktyczne zastosowania tworzonych linii zależnych – patrz poniżej).

3. Celem linii głównej (*) jest zapis nagranej wypowiedzi w sposób najbardziej zbliżony do formy fonicznej, jaką nadał jej mówiący. Oczywistym jest także, że formy wypowiadane przez dzieci oraz uczących się języka obcego (zwłaszcza w stadium początkowym) różnią się znacznie od standardowej formy języka, CHILDES pozwala na zapis tych zmian (form niestandardowych) i umożliwia informatyczną obróbkę danych. Zmiany (zwłaszcza fonetyczne czy morfologiczne) zapisane być mogą już w linii głównej, tym niemniej do komentarzy i kodów wszelkiego typu zaleca się stosowanie odpowiednich linii zależnych, w przeciwnym wypadku, linia główna szybko staje się nieczytelna i trudna do analizy.

Obok linii głównych oraz linii zależnych, trzecim istotnym elementem plików CHAT są *nagłówki* (ang. *headers*); niektóre z nich zawsze umieszczane są na początku pliku będącego zapisem wypowiedzi i zawierają stałe informacje pozwalające na identyfikację pliku, informatora, zadania językowego i innych. Wszystkie nagłówki (headers) rozpoczynają się od symbolu @. Wszystkie pliki zapisane w systemie CHAT muszą zaczynać się od nagłówka *@Begin* oraz kończyć się zapisem *@End*. Każdy z nich musi też w drugiej linii zawierać nagłówek *@Participants*. Zasady te ilustruje następujący przykład:

@Begin
@Participants
@End

@Begin, *@Participants* i *@End* to nagłówki obowiązkowe dla wszystkich plików (ang. *obligatory headers*); ich brak nie pozwala na sprawne funkcjonowanie systemu.

Użytkownik wprowadzić może także dodatkowe, tak zwane *nagłówki stałe* (ang. *constant headers*), które nie są obowiązkowe, ale zawierają za to często przydatne informacje dotyczące badanych, nagrania czy interakcji. Wymienić można tu przykładowo:

@Birth of Learner:	23-JAN-1997
@Age of Learner:	9;7.27
@Time spent in Poland:	1;0.24
@Date:	19-AUG-2006
@Filename:	ffe4lucf.cha
@Transciber:	UPP
@Warning:	transcript has not been double-checked

Wiek i datę urodzenia badanych zwyczajowo notuje się w wypadku badań przeprowadzanych z dziećmi, w wypadku badań dorosłych wystarcza adnotacja *adult* (dorosły).

Możliwe są również *nagłówki zmienne* (ang. *changeable headers*), umieścić można je na początku pliku razem z nagłówkami stałymi, bądź też w treści pliku, wewnątrz samej transkrypcji, przykładowo:

@Situation:	opowiedzieć film o Reksiu
@Activities:	uzupełnić układankę.

Wśród nagłówków zmiennych warto zwrócić uwagę na oznaczenia *@Bg: reading/ picture story* oraz *@Eg: reading/ picture story*. Oznaczają one odpowiednio początek i koniec tak zwanych *fragmentów specjalnych* (ang. *special passages)* poświęconych na czytanie tekstu bądź opowiadanie historyjki obrazkowej:

*PIO:	nie wiem co powiedzieć.
@Bg:	reading
*PIO:	dawno dawno temu.
*PIO:	za górami, za lasami.
*PIO:	żyła w samotnej wieży piękna królewna.
@Eg:	reading

Pliki transkrybowane w formacie CHAT muszą posiadać rozszerzenie *.cha*. Są one kompatybilne z formatem *txt* (pliki tekstowe), mogą więc być edytowane (drukowane, opracowywane) w formacie Word. Przykładowa nazwa pliku w formacie CHAT (stworzona dla konkretnego projektu) może mieć postać: *ffe4lucf.cha* (tu: w nazwie zakodowane jest: *f* – frankofon, *f* – nagranie w języku francuskim, *e* – dziecko [*enfant*], *4* – czterolatek, *luc* – trzy pierwsze litery imienia identyfikujące dziecko, *f* – opowiadanie filmu).

Szczegółowy i pełny opis zasad transkrypcji w systemie CHAT dostępny jest w wersji elektronicznej (podręcznik w formacie PDF) na stronie www: http://childes.psy.cmu.edu (*The CHILDES Project*, Part 1: The CHAT Transcription Format). Przykładowe symbole transkrypcji prezentuje lista poniżej:

Jednostki leksykalne (w linii głównej)

@	niestandardowa forma językowa (*special form*)
xx	jednostka niezrozumiała, prawdopodobnie jedno słowo
xxx	jednostka niezrozumiała, prawdopodobnie grupa słów
www	materiał nietranskrybowany
&	początek zapisu w wersji fonologicznej
[?]	jednostka niezrozumiała – transkrypcja najbardziej prawdopodobna
( )	brak części jednostki leksykalnej

Morfemy

-	sufiks
#	prefiks

Interpunkcja

.	kropka, koniec wypowiedzenia
?	pytanie
!	wykrzyknik
+…	wypowiedzenie niedokończone
+/.	wypowiedzenie przerwane

Pauzy

#	pauza krótka (mniej niż 2 sekundy)
##	pauza długa (2-8 sekund)
eh@fp	pauzy wypełnione

Ton wypowiedzi

↑	ton wznoszący
↓	ton opadający

Symbole „zasięgu”

[=texte]	texte: wyjaśnienie znaczenia
[=? texte]	texte: transkrypcja alternatywna lub najbardziej prawdopodobna
[/]	powtórzenie bez poprawek
[//]	powtórzenie z poprawkami

Transkrypcja danych zgodna z systemem CHAT gwarantuje sprawne funkcjonowanie programu CLAN. Zależnie od typu przeprowadzanej analizy niektóre zasady transkrypcji mogą zostać uproszczone (pominięte); należy jednak pamiętać, iż w takim wypadku nie wszystkie komendy programu CLAN – a więc nie wszystkie analizy wykonywane na danych automatycznie – będą dostępne.

3. CLAN: Computerized Language Analysis (Informatyczna Analiza Języka)

Komendy programu CLAN skonstruowane są tak, by umożliwić analizę danych językowych transkrybowanych w formacie CHILDES. Sam CLAN zawiera *edytor tekstu* (ang. *editor*) przystosowany do pracy z plikami zakodowanymi w systemie zapisu CHAT. Zależnie od potrzeb, może on funkcjonować w trybie umożliwiającym *zapis i transkrypcję danych* (ang. *CHAT mode*), *kodowanie danych* (ang. *coder mode*), *łącznie plików transkrypcji z odpowiadającymi im plikami audio* (ang. *sonic mode*) i video (ang. *video mode*).

W czasie transkrypcji plików w systemie CHAT w celu ich automatycznej analizy przy użyciu komend programu CLAN, niezwykle ważna jest weryfikacja zapisu i jego zgodności z zasadami transkrypcji CHAT. Program wykonuje ją automatycznie przy użyciu funkcji *CHECK* (wybrać *mode* w menu okna komend programu >> z listy wybrać funkcję *check open file*) i wskazuje na ewentualne błędy w transkrypcji (najczęstsze to brak znaków końcowych zamykających linię zapisu, opuszczenia nawiasów, niedozwolone zasadami transkrypcji znaki, nieopisane linie zapisu). Plik zapisany w systemie CHAT zgodnie z wymogami programu CLAN pozwala automatycznie generować analizy i informacje dotyczące jednego pliku bądź grupy plików; dane będące wynikami analiz można osobno edytować, formatować, drukować, przenosić do innych partii tekstu.

Przykładowe komendy aplikacji CLAN to *CHECK*, *DATES*, *FREQ*, *COMBO*, *MLU*. Komenda *CHECK*, opisana powyżej, pozwala na weryfikację transkrypcji danych, wykrycie i poprawienie błędów w zapisie; trzeba jednak podkreślić, że *CHECK* nie jest korektorem ortograficznym. *DATES*, pozwala policzyć wiek badanego z dokładnością do ilości dni, biorąc pod uwagę datę urodzenia (dziecka) oraz datę nagrania. *FREQ* pozwala zbadać frekwencję danej jednostki leksykalnej, kilku lub wszystkich jednostek, w jednym pliku bądź w serii plików. Komenda ta jest o tyle istotna, że oprócz danych typowo liczbowych (ilość użyć) dostarcza informacji na temat tak zwanych *types* (ilość różnych mobilizowanych jednostek leksykalnych, np. ilość różnych czasowników) oraz *tokens* (ilość wszystkich jednostek leksykalnych mobilizowanych w ogóle, np. ilość wszystkich form czasownika w danym pliku). *Stosunek type/token* (ang. *types/token ratio*) pozwala zaś oznaczyć bogactwo leksykalne wypowiedzi czyli tzw. *wskaźnik Guirauda* (ang. *index of Guiraud,* fr. *indice de richesse*, cf. Broeder i in. 1993). *COMBO*, pozwala na wyszukiwanie występowania jednostek leksykalnych lub ich części w całych liniach tekstu bądź w opcjonalnie zdefiniowanym kontekście. W końcu *MLU* (ang. *Mean Length of Utterance*) pozwala policzyć średnią długość wypowiedzenia (ilość jednostek leksykalnych lub morfemów) dla danego lokutora, w jednym pliku bądź w serii plików, oraz jego odchylenie standardowe. MLU może podać także całkowitą ilość wypowiedzeń lub całkowitą ilość jednostek leksykalnych (lub morfemów) w danej partii danych (analizowanej części korpusu danych). Szczegółowy opis funkcji dostępnych przy pomocy komend

programu *CLAN* dostępny jest w wersji elektronicznej (podręcznik w formacie PDF) na stronie www: http://childes.psy.cmu.edu. (*The CHILDES Project*, Part 2: The CLAN Programs).

Komendy programu CLAN wpisywane są w pojawiające się okno komend (*Commands*) i posiadają ścisłą składnię: trzy kolejne elementy wpisywanej komendy rozdzielone spacjami. Element pierwszy to nazwa używanej komendy (skrót: *FREQ*, *CIMBO*, etc.); element drugi uściśla linię (a w zasadzie linie główne, ewentualnie linie zależne), która objęta jest analizą (*+t* = linia główna) oraz specyfikuje uczestnika nagrania (**PIO*), którego analiza dotyczy (zazwyczaj: badane dziecko, uczeń, etc.); element ostatni wskazuje na plik (grupę plików), których analiza dotyczy (*ffe4luc.cha* – jeden wybrany plik o szczegółowo podanej nazwie, **cha* – grupa plików z określonego katalogu posiadająca rozszerzenie .cha). W ten sposób podana komenda: *freq +t*LUC ffe4lucf.cha* zbada frekwencję słów (ilość *types*, ilość *token* oraz *type/token ratio*) w wypowiedzi czteroletniego Luca.

4. Przykład pracy z plikiem zapisanym w systemie CHILDES (format CHAT)

Poniższy plik, zapisany w systemie CHILDES pochodzi z korpusu badań nad akwizycją języka u dzieci (języka ojczystego) i dorosłych (języka obcego); jest on przykładem praktycznego zastosowania systemu oraz jego funkcji do badań językoznawczych wykorzystujących elementy statystyki (Demagny i Paprocka-Piotrowska 2004; Paprocka-Piotrowska 2008). Prezentowany plik *pp4efilf.cha* jest zapisem wypowiedzi 4-letniego Filipa, który po polsku opowiadał fragment obejrzanych przygód Reksia (pięciominutowy film-rekwizyt był częścią protokołu badań). Transkrypcji dokonano bez zachowania polskich znaków diakrytycznych. Dodatkowa aplikacja systemu umożliwia ich zastosowanie, niemniej jednak, program funkcjonuje sprawniej nie obciążony systemem polskich znaków.

@Begin
@Participants: SBJ – Subject INV – Investigator (Anna, Katarzyna)
@Name of SBJ: Filip – FIL
@Birth of SBJ: 02.03.1997
@Sex of SBJ: male
@Language: polish
@Texte type: narration (film)
@Date: 23-JAN-2002
@Location: school
@Transcriber: Magda
@Filename: ppe4filf
@Dependent: cli, adv

*SBJ: pamietam jak Reksio sie slizgal pod budka.
%cli: aut_pamietac, dep_slizgac_sie

*SBJ: jak go wyciagal z wody # tego chlopczyka.
%cli: dep_wyciagnac
*SBJ: jak ten chlopczyk go ubieral.
%cli: aut_ubierac
*SBJ: ## jak Reksio mu pokazywal ile jest stopni na termometrze.
%cli: aut_pokazywac
*SBJ: ## jak wpadl w siano.
%cli: dep_wpasc
*SBJ: ## wiecej juz chyba nic nie pamietam.
%cli: aut_pamietac
%adv: cav_juz
*INV: cos chyba jeszcze bylo # dalej cos jeszcze bylo czy juz nic # mowiles ze ten chlopczy wpadl do wody tak.
*SBJ: potem chyba on sie owinal w cieply koc.
%cli: aut_owinac_sie
%adv: paav_potem
*SBJ: i poszedl do domu # jakos tak bylo # w kazdym razie.
%cli: dep_pojsc, exi_byc
*INV: no dobra.
@End

Transkrypcja zaczyna się od obowiązkowego nagłówka *@Begin* i kończy na również obowiązkowym *@End*. Wśród zapisanych nagłówków wpisano również obowiązkowy *@Participants*: *SBJ – Subject*, *INV – Investigator*; o ile jednak imiona ankieterów są specyfikowane (istotne jest bowiem, kto przeprowadzał wywiad z dzieckiem), to samo dziecko kodowane jest jako *SBJ* (ang. *subject*), co pozwala na automatyczne przeszukiwanie plików bez podawania w każdej komendzie konkretnego imienia dziecka ('ant', 'luc', 'fil' itp.) w składni. W nagłówkach stałych pominięto wiek dziecka (*@Age of SBJ*) – protokół badań zakładał bowiem wywiady w grupie czterolatków, podobnie poziom wykształcenia (*@Grade-level*). Czas nagrania (*@Time duration*) nie odgrywał roli w badaniach, wiec również został pominięty, nie było też dodatkowych komentarzy (*@Coment*). Pozostałe nagłówki specyfikują kolejno: imię dziecka (w tym wypadku jest to imię autentyczne, w wypadku dorosłych często imię-identyfikator zostaje zmienione dla zachowania większej anonimowości badań), datę urodzenia, płeć, język wypowiedzi (badania przeprowadzono na dzieciach polskich, francuskich, brytyjskich, włoskich, greckich), typ wypowiedzi (dzieci poddawane były kolejno trzem zadaniom językowym: opowiadaniu filmu animowanego z cyklu *Reksio*, opowiadaniu historyjki obrazkowej – rekwizyt stworzony na potrzeby protokołu badań, oraz opisywały obraz przedstawiający plac w małym mieście), datę i miejsce nagrania (w wypadku dzieci młodszych istnieje duża różnica pomiędzy nagraniem przeprowadzonym w przedszkolu a w domu), osobę, która dokonała transkrypcji (niezwykle istotne zwłaszcza przy pewnych niejasnościach w transkrypcji) i nazwę pliku (tu: bez rozszerzenia *.cha* ponieważ i tak jest ono obowiązkowe). Nagłówek @Dependent: *cli, adv* wskazuje zaś, że plik przygotowany jest do prze-

prowadzenia analizy nie automatycznie generowanych funkcji zapisanych w systemie (morfologia, składnia, fonetyka), a funkcji stworzonych przez użytkownika na potrzeby jego własnych badań: typu użytego czasownika (*%cli*) oraz typu użytego przysłówka (*%adv*).

W linii zależnej oznaczonej symbolem *%cli* kodowany jest typ użytego czasownika: *%cli: cop_** (ang. *copula*), *aux_** (czasownik posiłkowy – ang. *auxiliary verb*), *exi_** (czasownik egzystencjalny), *spp_** (czasownik posiłkowy – ang. *support verb*), *dep_** (czasownik ruchu), *vdd_** (*verbum dicendi*), *vdp_** (czasownik percepcji), *vdf_** (czasownik frazowy), *aut_** (inne). Ta lina kodów została wprowadzona i wykorzystana przez użytkownika; wskazują na to jasno linie *%cli* następujące po każdym zakończonym wypowiedzeniu – po każdej linii oznaczonej symbolem **SBJ:* i zakończonej kropką, jak poniżej:

*SBJ: pamietam jak Reksio sie slizgal pod budka.
%cli: aut_pamietac, dep_slizgac_sie

W linii zależnej oznaczonej symbolem *%adv* zakodowany został typ przysłówka (ang. *adverb*), który występuje w danym wypowiedzeniu, przykładowo:

%adv: cav_* (przysłówek kontrastu w pozycji preverbalnej); paav_* (przysłówek pozycji, anaforyczny, w pozycji prewerbalnej).

Ta linia kodów następuje bezpośrednio po linii oznaczonej symbolem *%cli* i odnosi się do tej samej linii wypowiedzenia, oznaczonej odpowiednio symbolem **SBJ*:

*SBJ: potem chyba on sie owinal w cieply koc.
%cli: aut_owinac_sie
%adv: paav_potem

Kategorie zakodowanych czasowników i przysłówków wynikają z potrzeb zastosowanego modelu analizy (por. Damagny i Paprocka-Piotrowska 2004).

Analiza pliku przy pomocy komendy FREQ (frekwencja słów, *types/token ratio*) zastosowana została do zbadania bogactwa słownictwa, a konkretnie do prześledzenia repertuaru mobilizowanych czasowników, jakimi dysponują badani na danym etapie akwizycji (tu: J1).

```
freq +s"aux*" +s"sau*" +s"cop*" +s"pre*" +s"exi*" +s"spp*" +s"dep*"
+s"vdd*" +s"vdp*" +s"vdf*" +s"aut*" +t%cli ppe4filf.cha
Mon Jan 16 05:33:29 2006
freq (02-Dec-2002) is conducting analyses on:
ALL speaker tiers
  and those speakers' ONLY dependent tiers matching: %CLI;
****************************************
From file <ppe4filf.cha>
```

```
   1 aut_owinac_sie
   2 aut_pamietac
   1 aut_pokazywac
   1 aut_ubierac
   1 dep_pojsc
   1 dep_slizgac_sie
   1 dep_wpasc
   1 dep_wyciagnac
   1 exi_byc
------------------------------
 9 Total number of different word types used
10 Total number of words (tokens)
0,900 Type/Token ratio
```

W swojej wypowiedzi Filip użył 4 czasowników ruchu (*dep_**), jednego zdania egzystencjalnego (*exi_**) oraz 4 czasowników zaklasyfikowanych jako *inne* (*aut_**). Na 10 form czasownika zakodowanych w linii *%cli* (*total number of words*), 9 – to różne formy podstawowe (*total number of different word types used*), stąd też wskaźnik bogactwa leksykalnego (*types/tokes ratio*) jest wysoki i zbliża się do wartości 1.

Ta sama komenda FREQ, użyta do analizy danych wszystkich badanych 4-latków, podaje kompletną listę całości i ich repertuaru językowego wraz ze wskaźnikiem *types/token ratio* obliczonym dla całej grupy wiekowej, przykładowo:

```
freq +u +s"aux*" +s"sau*" +s"cop*" +s"pre*" +s"exi*" +s"spp*" +s"dep*"
+s"vdd*" +s"vdp*" +s"vdf*" +s"aut*" +t%cli *.cha
Mon Jan 16 05:29:57 2006
freq (02-Dec-2002) is conducting analyses on:
ALL speaker tiers
  and those speakers' ONLY dependent tiers matching: %CLI;
****************************************
From file <*.cha>
------------------------------
 121 Total number of different word types used
 312 Total number of words (tokens)
0,388 Type/Token ratio
```

Cała grupa 4-latków ma więc wskaźnik bogactwa leksykalnego (jeśli idzie o zmobilizowany repertuar czasowników) na poziomie 0,338; w sumie, dzieci użyły 312 form czasownikowych, w tym 121 różnych czasowników.

Ta sama komenda FREQ pozwala również pytać o pewien typ danych, tym samym dostarcza więc szczegółowych informacji o pewnym typie mobilizowanych jednostek leksykalnych. Na przykład, polecenie: *freq +u +s"dep*" +t%cli *.cha* zastosowane do tej samej grupy badanych 4-latków pozwala ustalić listę czasowników ruchu (*dep*_*), które pojawiły się w wypowiedziach dzieci:

```
freq +u +s"dep*" +t%cli *.cha
Mon Jan 16 05:41:16 2006
freq (02-Dec-2002) is conducting analyses on:
ALL speaker tiers
  and those speakers' ONLY dependent tiers matching: %CLI;
****************************************
From file <*.cha>
  2 dep_ciagnac
  3 dep_isc
  1 dep_isc_id=wejsc
  1 dep_isc_id=wpasc
  3 dep_jechac
 10 dep_jezdzic
  1 dep_odjechac
  2 dep_odwrocic_sie
  1 dep_pedzic
  1 dep_pociagnac
  2 dep_pojechac_id=pojezdzic
  6 dep_pojezdzic
 35 dep_pojsc
  2 dep_pojsc_id=przejsc
  1 dep_pojsc_id=wejsc
  1 dep_polozyc
  1 dep_przesliznac_sie_id=posliznac_sie
  1 dep_przewracac_sie
  7 dep_przewrocic_sie
  1 dep_przycisnac
  1 dep_przyjechac
  4 dep_przyjsc
  1 dep_przyniesc
  1 dep_przysunac
  1 dep_siasc
  1 dep_skakac
  1 dep_skoczyc
  3 dep_slizgac_sie
  2 dep_spasc
  1 dep_spasc_id=upasc
  1 dep_spasc_id=wpasc
  1 dep_wdrapywac_sie
  6 dep_wejsc
  1 dep_wejsc_id=przejsc
 12 dep_wpasc
  2 dep_wracac
  1 dep_wrocic
  1 dep_wsadzic
  1 dep_wychodzic
 11 dep_wyciagnac
```

```
  5 dep_wyjac
 12 dep_wyjsc
  1 dep_wylozyc_id=wyjac
  1 dep_wyslizgac_sie_id=poslizgnac_sie
  1 dep_wywalic_sie
  2 dep_zaprowadzic
------------------------------
 46 Total number of different word types used
157 Total number of words (tokens)
0,293 Type/Token ratio
```

Przeprowadzona analiza pozwala stwierdzić, że cała grupa (20 dzieci) użyła w sumie 157 czasowników ruchu (przy ogólnej liczbie 312 wyprodukowanych form, cf. supra), i że było to 46 różnych czasowników (przy ogólnej liczbie 121 różnych czasowników wyprodukowanych w ogóle, cf. supra). *Types/token ratio* dla klasy czasowników ruchu (0,293) jest więc niższy niż ogólny wskaźnik bogactwa słownikowego mierzony ogólnie dla całej grupy badanych. Szczegółowa analiza liczbowa poszczególnych klas czasowników mobilizowancyh przez dzieci pozwala zaś stwierdzić, jaki jest procentowy udział w wypowiedziach czasowników ruchu, czasowników fazowych, posiłkowych, etc., a także jakie czasowniki używane są najczęściej/najrzadziej w obrębie poszczególnych klas, por. *12 dep_wyjść vs. 1dep_wywalić_się*.

Na uwagę zasługuje również możliwość szybkiego wyszukiwania form idiosynkratycznych (agramatycznych, niedostosowanych do kontekstu, neologizmów, tworów językowych i innych). System CHILDES (kody CHAT) pozwala na dodatkowe ich oznaczanie, tu: dopiskiem *_id=[wyjaśnienie]*. Lista poniżej prezentuje formy idiosynkratyczne wyprodukowane przez 4-latki w czasie wypowiedzi na temat przygód Reksia:

```
1 aut_obwinac_id=owinac
1 aut_ogryzc_id=odgryzc
2 aut_polamac_sie_id=zalamac_sie
1 aut_przystyknac_id=creation_lex
1 aut_zalamac_sie_id=lod_sie_zalamal
1 aut_zdjac_id=zabrac
1 aut_zlamac_sie_id=rozerwac_sie
1 aut_zlamac_sie_id=zalamac_sie
1 dep_isc_id=wejsc
1 dep_isc_id=wpasc
2 dep_pojechac_id=pojezdzic
2 dep_pojsc_id=przejsc
1 dep_pojsc_id=wejsc
1 dep_przesliznac_sie_id=posliznac_sie
1 dep_spasc_id=upasc
1 dep_spasc_id=wpasc
1 dep_wejsc_id=przejsc
1 dep_wylozyc_id=wyjac
```

```
  1 dep_wyslizgac_sie_id=posliznac_sie
------------------------------
   19 Total number of different word types used
   22 Total number of words (tokens)
0.864 Type/Token ratio
```

Tak skompilowana lista – będąca w gruncie rzeczy listą jednostek leksykalnych i form czasownikowych, które dla dzieci okazały się najtrudniejsze – pozwala stwierdzić niezbicie, że perfektywne czasowniki ruchu budowane z prefiksem są największą przeszkodą dla 4-latków w biegłym posługiwaniu się językiem polskim (ojczystym).

BIBLIOGRAFIA

Broeder, P., Extra, G. i van Hout, R. 1993. „Richeness and variety in the developing lexicon", w: Perdue, C. (red.). 1993. 145-163.

Champaud, C. 2001. „Une introduction au système CHILDES en français". (http:///childes.psy.cmu.edu/intro/french.pdf [maj 2009]).

Demagny, A. C. i Paprocka-Piotrowska U. 2004. „L'acquisition du lexique verbal et des connecteurs temporels dans les récits de fiction en français L1 et L2". *Langages* 155. 52-75.

Mac Whinney, B. 2000. *The CHILDES Project: Tools for Analyzing Talk*. NJ: Lawrence Erlbaum Associates. (http:///childes.psy.cmu.edu [maj 2009]).

Paprocka-Piotrowska, U. 2008. *Conter au risque de tout changer. Complexité conceptuelle et référence aux procès dans l'acquisition du français L2 et du polonais L2*. Lublin: Towarzystwo Naukowe KUL.

Perdue, C. (red.). 1993. *Adult language acquisition: Cross-linguistics perspectives*. Cambridge: Cambridge University Press.