Rozwinięcie korpusu polskich rozmów telefonicznych LUNA

Aleksandra Wyszyńska

Akademia Górniczo-Hutnicza
im. Stanisława Staszica w Krakowie

Wydział Elektrotechniki, Automatyki, Informatyki i Elektroniki

KATEDRA ELEKTRONIKI

INŻYNIERSKI PROJEKT DYPLOMOWY

ALEKSANDRA WYSZYŃSKA

ROZWINIĘCIE KORPUSU POLSKICH ROZMÓW TELEFONICZNYCH LUNA

PROMOTOR:
dr inż. Bartosz Ziółko

KIERUNEK:
Inżynieria Akustyczna
(studia międzywydziałowe)

Kraków 2012

OŚWIADCZENIE AUTORA PRACY

OŚWIADCZAM, ŚWIADOMY ODPOWIEDZIALNOŚCI KARNEJ ZA POŚWIADCZENIE NIEPRAWDY, ŻE NINIEJSZĄ PRACĘ DYPLOMOWĄ WYKONAŁEM OSOBIŚCIE I SAMODZIELNIE, I NIE KORZYSTAŁEM ZE ŹRÓDEŁ INNYCH NIŻ WYMIENIONE W PRACY.

.....

PODPIS

AGH

University of Science and Technology in Krakow

Faculty of Electrical Engineering, Automatics, Computer Science and Electronics

DEPARTMENT OF ELECTRONICS

ENGINEERING DIPLOMA PROJECT

ALEKSANDRA WYSZYŃSKA

DEVELOPMENT OF POLISH TELEPHONE CONVERSATIONS CORPUS LUNA

SUPERVISOR:
Bartosz Ziółko Ph.D

KIERUNEK:
Inżynieria Akustyczna
(studia międzywydziałowe)

Krakow 2012

Dziękuję Grzegorzowi Wyszyńskiemu,
za pomoc okazaną przy pisaniu pracy.

Spis treści

1. Wprowadzenie

Celem poniższej pracy było rozwinięcie korpusu LUNA poprzez dodanie plików zawierających etykiety czasu początku i końca wyrazów. Prace nad transkrypcjami plików audio będące częścią projektu były bardzo czasochłonne, trwały 118 godzin i 8 minut.

Aby system rozpoznawania mowy mógł działać, potrzebuje danych, na których może "nauczyć się" co oznacza dane słowo i jak ono brzmi. Potrzebny do tego jest zbiór danych, który nazywamy korpusem mowy. Jako że język jest bardzo wieloznaczny i bogaty, trudno stworzyć taką bazę danych. Powinna ona zawierać jak najwięcej informacji oraz jak najwięcej mówców. Najlepiej aby wypowiedzi były spontaniczne [2]. Korpusy są cennym zbiorem informacji zarówno dla informatyków, którzy tworzą systemy rozpoznawania mowy, jak i dla lingwistów, ze względu na wieloznaczność języka i jego interpretacji.

W poniższej pracy opisano jeden z korpusów mowy polskiej, który został zrealizowany dla potrzeb projektu LUNA.

W rodziale pierwszym opisano sam projekt oraz analizę korpusu mowy LUNA, który jest bardzo obszerny i anotowany na różnych poziomach. Dokonano także porównania tego korpusu z innymi, najbardziej popularnymi korpusami mowy polskiej.

W rozdziale drugim szerzej opisano, na czym polegało rozwinięcie korpusu, jakie wiązały się z tym trudności. Przedstawiono, czym są pliki *.mlf i jakie informacje ze sobą niosą oraz zaprezentowano środowisko, w którym wykonano projekt.

Projekt był częściowo realizowany jako udział w grancie Nr OR00001905, MNSiW, “Aplikacje technologii mowy w systemach bezpieczeństwa publicznego”, AGH, Kraków kierowanym przez prof. dr hab. inż Mariusza Ziółkę. Jego efekty są stosowane w systemie rozpoznawania mowy AGH [13].

2. Projekt LUNA

2.1. Czym jest LUNA

Korpus dialogów telefonicznych LUNA został stworzony w ramach projektu LUNA (ang. spoken Language UNderstanding In MultilinguAl Communications systems) stworzonego w latach 2006-2009. Jak pisze Małogorzata Marciniak [8] "Projekt realizowało konsorcjum kierowane przez profesora Renato De Moriego z Uniwersytetu w Avignon. W jego skład wchodziły ośrodki badawcze z czterech krajów. Z Francji Uniwersytet w Avignon i France Telecom, z Włoch: firma Loquendo, Uniwersytet w Trydencie (Trento) i CSI-Piemonte (Piedmont Consorctium for Information Systems), z Niemiec: Uniwersytet Techniczny z Akwizgranu (RWTH Aachen) oraz z Polski: Polsko- Japońska Wyższa Szkoła Technik Komputerowych oraz Instytut Podstaw Informatyki PAN z Warszawy. Badania dotyczyły zagadnień związanych z rozumieniem mowy w językach: francuskim, włoskim oraz polskim."

[8] Celem projektu LUNA było opracowanie narzędzi do dostosowywania oprogramowania rozpoznawania mowy do nowego języka lub nowej tematyki dialogu. Projekt skupiał się na usprawnieniu serwisów telefonicznych, tak by klient nie musiał wysłuchiwać długich list opcji wyboru przypisanych do danych klawiszy. Zarówno listy opcji wyboru jak i ciągłe prośby o powtórzenie wzbudzają w klientach niechęć i zniecierpliwienie. Starano się na tyle poprawić serwisy telefoniczne by rozumiały spontaniczną mowę, w związku z czym skupiono się na działaniu oprogramowania w czasie rzeczywistym oraz na niskiej jakości sygnału telefonicznego.

Zastosowany system na podstawie odpowiedzi na kilka prostych pytań ustaliłby rodzaj problemu i w zależności od jego stopnia zaawansowania, udzieliłby odpowiedzi lub przełączył do odpowiedniego operatora.

W projekcie konieczne było korzystanie z korpusów mowy dla każdego z trzech języków (francuski, włoski i polski). Dla języka francuskiego wykorzystano już istniejący korpus MEDIA. Są to rozmowy człowieka z komputerem i dotyczą rezerwacji hoteli. Dla języka włoskiego i polskiego stworzono nowe korpusy mowy zawierające dialogi człowieka z człowiekiem oraz człowieka z komputerem. W języku włoskim rozmowy dotyczyły rozwiązywania problemów związanych z niewłaściwie działającym sprzętem komputerowym, w języku polskim komunikacji miejskiej. Nagrania wchodzące w skład polskiego korpusu udostępniła infolinia Zarządu Transportu Miejskiego w Warszawie.

Polski korpus dialogów telefonicznych LUNA składa się z dwóch korpusów: korpusu dialogów człowieka z człowiekiem (LUNA.PL) oraz dialogów człowieka z komputerem (LUNA.WOZ). Każdy z korpusów podzielony jest na katalogi tematyczne: JAKDOJECHAC (dialogi dotyczące połączeń ko-

munikacyjnych między dwoma miejscami), KIEDY (rozkłady jazdy oraz czas przejazdów), CZYJEDZIEPRZEZ (trasy przejazdów autobusów), PRZYSTANKI (najbliżej położone przystanki dla danego miejsca oraz przystanki, na których można się przesiąść) i ZNIZKI (opłaty oraz przysługujące zniżki). Korpus LUNA.WOZ zawiera jedynie katalogi JAKDOJECHAC, KIEDY, ZNIZKI. Każdy z tych dwóch korpusów zawiera 500 dialogów, które zostały zanotowane na różnych poziomach. Każdy dialog znajduje się w folderze z nazwą identyczną jak nazwa nagrania. W danym folderze znajdują się także pliki opisujące dany dialog.

Tabela 2.1: Zawartość katalogu "nazwa" opisującego jeden dialog [8]

Nazwa Pliku	Zawartość
Nazwa.wav	Nagranie dialogu
Nazwa.trs	Transkrypcja dialogu
Nazwa_turns.xml	Podział na wypowiedzi
Nazwa_words. Xml	Analiza morfologiczna słów
Nazwa_chunks. Xml	Wyodrębnione proste frazy
Nazwa_attvalue. Xml	Pojęcia
Nazwa_frames. Xml	Predykaty
Nazwa_anaphoras. Xml	Odniesienia anaforyczne
Nazwa_dialacts. Xml	Akty dialogowe

Tabela 2.2: Statystyka dla 500 dialogów korpusu LUNA [8].

Liczba Wypowiedzi	12788
Liczba słów	81049
Rozmiar słownika	7768

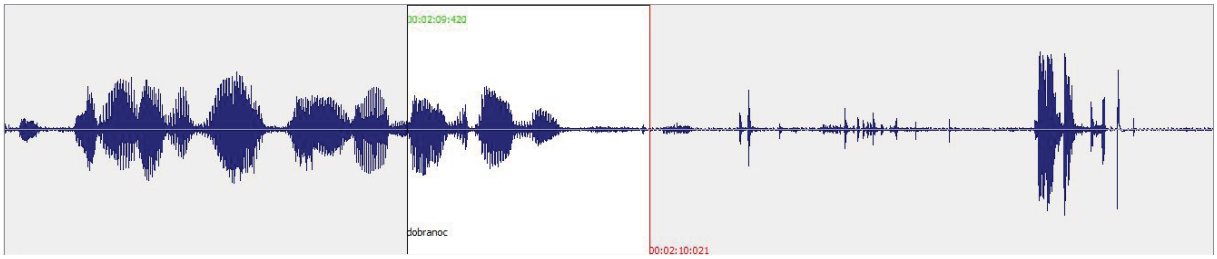
Tabela 2.3: Statystyka rozmów według płci i tematu dialogów [8]

Rodzaj poszukiwanej informacji	Liczba rozmów	Kobiety	Mężczyźni
Trasy komunikacyjne	93	53	40
Połączenia	140	78	62
Rozkłady	112	60	52
Przystanki	55	24	31
Zniżki i bezpłatne przejazdy	101	61	40

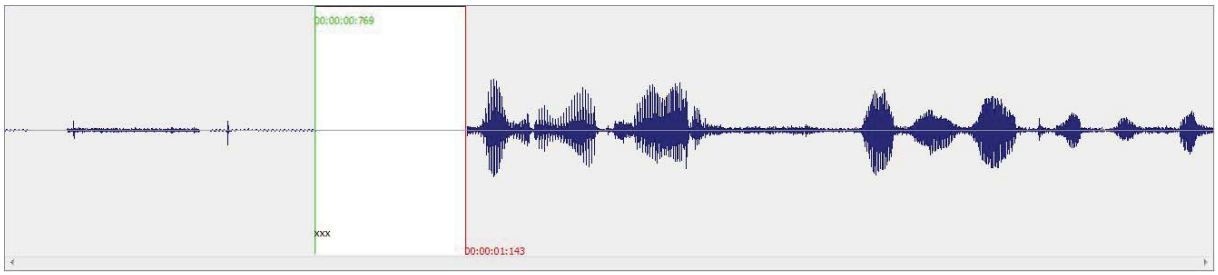
Łączny czas nagrań wynosił 670 minut. Część nagrań można było przyporządkować do różnych kategorii tematycznych. Wiele dialogów dotyczyło kilku tematów (np. zarówno trasy przejazdu, jak i godzin odjazdów). Do której kategorii tematycznej dane nagranie zostało przyporządkowane, zależało od osoby wykonującej transkrypcję dialogu.

Większość rozmów wykonały kobiety (patrz tabela 2.3). Należy zwrócić uwagę na to, że nagrania podzielono na głosy męskie lub żeńskie ze względu na osobę, która dzwoniła do infolinii, nie zaś ze względu na operatora odbierającego połączenie.

Nazwiska osób dzwoniących jak i operatorów usunięto z nagrań, w transkrypcji zapisano je jako xxx 2.2.



Rysunek 2.1: Przebieg czasowy części nagrania z LUNY z zaznaczonym słowem "dobranoc"



Rysunek 2.2: Przebieg czasowy części nagrania z LUNY z zaznaczonym wyciętym nazwiskiem (oznaczonym jako XXX)

2.2. LUNA a inne korpusy mowy polskiej

Najpopularniejszym korpusem mowy polskiej jest CORPORA, stworzona przez Stefana Grocholewskiego na Politechnice Poznańskiej w 1997 roku. Zawiera ona 16.425 wypowiedzi pochodzących od 45 osób (365 od każdej z osób). Słownik bazy składał się z 33 liter alfabetu, 10 cyfr, 200 imion, 8 poleceń sterujących (np. cofnij, kropka) oraz 114 semantycznie niespójnych wypowiedzi (np. Widzą chrzan biały na rzęsach). Wypowiedzi CORPORY zostały stworzone tak, aby uzyskać jak największą róznorodność fonetyczną.

Stefan Grocholewski opisuje korpus następująco [6]: "Do nagrań wykorzystano mikrofony pojemnościowe lub w jednym przypadku mikrofon dynamiczny. Parametry nagrań: częstotliwość próbkowania - 16 kHz, długość słów - 12 bitów. Nagrań dokonano w warunkach «naturalnych» pomieszczeń, w bezpośredniej bliskości pracującego komputera."

Innym znanym korpusem jest korpus mowy prawniczej Jurisdic [5]. Jest to zbiór zawierający 1000 mówców. Nagrania zostały zrobione w sądzie, biurach prawniczych, komisariatach policyjnych oraz uniwesytetach. Nagranie każdego z mówców zawiera około 20-40 minut mowy częściowo spontanicznej oraz około 30 minut czytanego tekstu (170 krótszych i dłuższych zdań). Mowa częściowo spontaniczna

polegała na wykonaniu pewnych zadań, np. dyktowanie tekstu prawniczego, opisanie wakacji, podanie wytłumaczenia/odmowy/zapytania w wyobrażonej sytuacji etc.

Nagrania zostały wykonane dwoma mikrofonami pojemnościowymi. Jeden z nich, pola bliskiego, zamontowany był blisko ust mówcy, drugi stał w odległości około 0,5 m od mówcy. W mikrofonie pola bliskiego uzyskano duży odstęp sygnału od tła, jednakże także wyraźnie słychać było oddechy mówcy oraz "popy". W drugim mikrofonie nie było tych problemów, ale mniejszy był odstęp sygnału od tła. Częstotliwość próbkowania wynosiła 16 kHz.

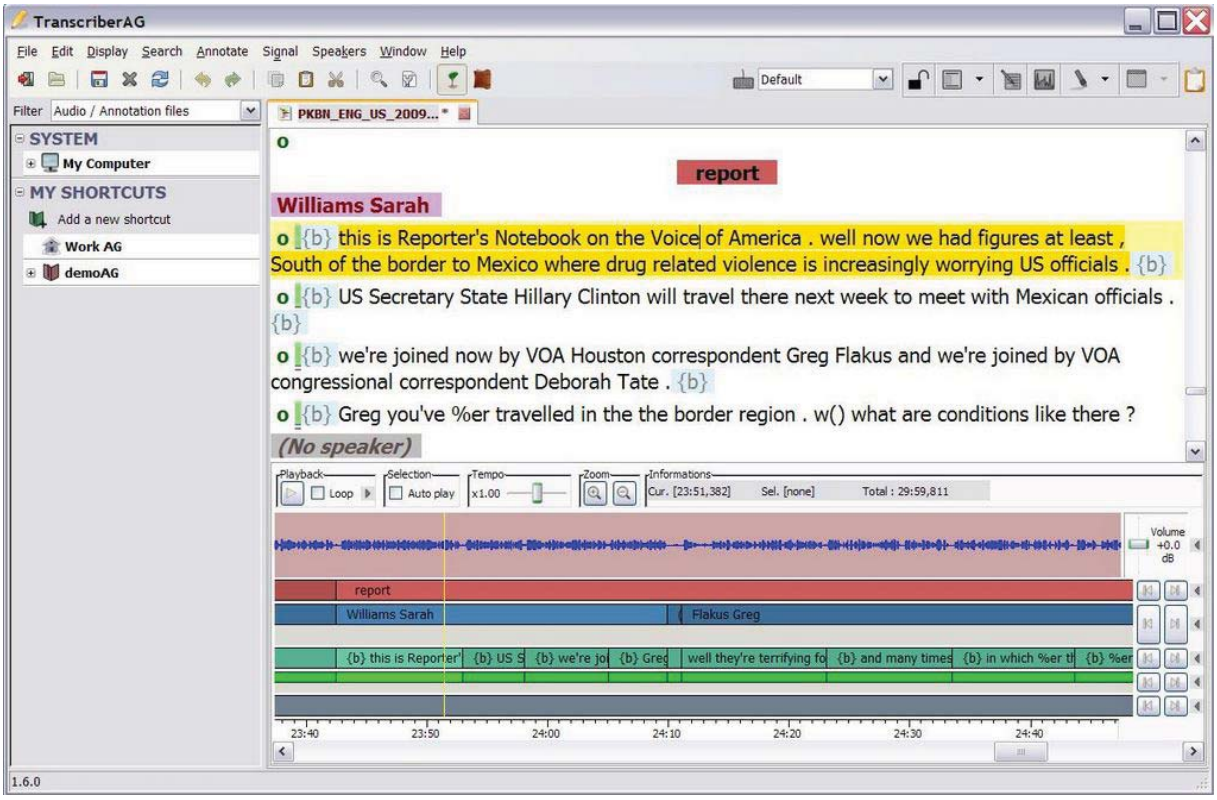
Innym korpusem dla języka polskiego jest SpeechDat(E) [1]. Został stworzony na Politechnice Wrocławskiej przez zespół Piotra Staroniewicza. Zawiera on dane dla 1000 mówców (488 mężczyzn oraz 512 kobiet). Każda z osób ma ściśle określone wypowiedzi (słowa, zdania itp.). Parametry nagrań: częstotliwość próbkowania-8 kHz, długość słowa-8 bitów.

Korpus Europarlmentu składa się z nagrań wystąpień polityków w Parlamencie Europejskim. Nagrania te obejmują wypowiedzi polityków jak także tłumaczy [7]. Korpus zawiera 127 godzin nagrań transkrybowanych automatycznie (978 mówców) oraz 3,5 godziny transkrybowanych ręcznie (46 mówców). Transkrypcja była robiona programem Transcriber (patrz rysunek 2.3).

Korpus Szklannego [9] jest korpusem difonów. Składa się z nagrań jednego mówcy (kobiety) i zawiera nagrania wszystkich połączeń głosek w języku polskim. Nagrano go w profesjonalnym studio, za pomocą mikrofonu pola bliskiego. Został on opracowany do zagadnień związanych z syntezą mowy.

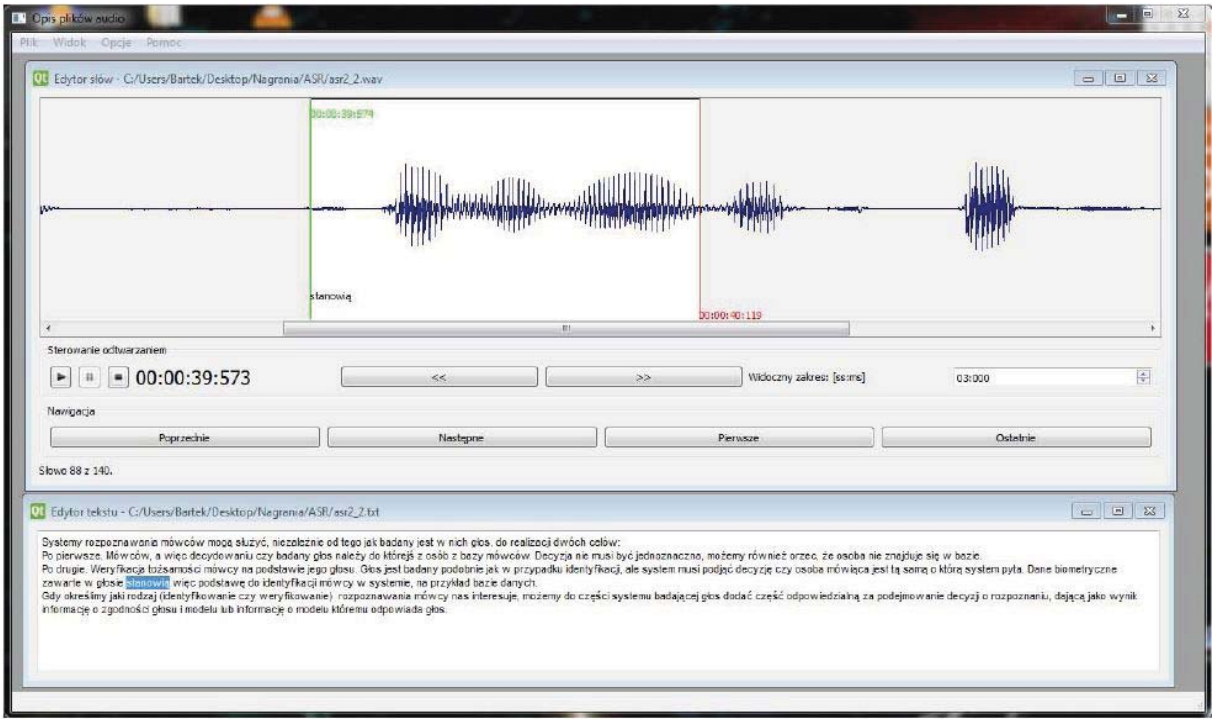
Korpus AGH jest korpusem zawierającym około 9 godzin anotowanych nagrań różnych mówców oraz około pół godziny nagrań testowych bez anotacji. Nagrania jednego z mówców trwają godzinę, pozostałych kilka, lub kilkanaście minut. Mówcy to przeważnie mężczyźni, jednak są także głosy żeńskie. Parametry nagrań: częstotliwość próbkowania - 16 kHz, długość słowa - 8 bitów. Znaczna część wypowiedzi zawartych w korpusie związane jest z technologią mowy [11][3].

Korpus LUNA od powyższych korpusów różni się przede wszystkim tym, że zawiera nagrania spontanicznych dialogów rejestrowanych przez łącze telefoniczne. Większość z powyżej przedstawionych korpusów (najpopularniejszych w Polsce) opiera się na wypowiedziach sztywno zarysowanych lub na mowie częściowo spontanicznej. Korpus Europarlamentu zawiera mowę nieopierającą się na sztywnych wypowiedziach, jednak nie zawiera dialogów między dwoma mówcami (akcja - pytanie; reakcja - odpowiedź). Parametry nagrań LUNY: częstotliwość próbkowania - 16 kHz, rozdzielczość - 16 bitów.



Rysunek 2.3: Program TranscriberAG, nowsza wersja programu, w którym dokonano transkrypcji korpusu Europarlamentu [4]

3. Rozwinięcie korpusu polskich rozmów telefonicznych LUNA



Rysunek 3.1: Program Anotator, w którym wykonywano pliki *.mlf [11]

3.1. Cel projektu

Celem projektu było rozwinięcie korpusu poprzez stworzenie do każdego z nagrań plików o rozszerzeniu *.mlf zawierającego czas nagrania, w którym zaczyna się i kończy każde słowo z nagrania. Częstą praktyką jest tworzenie plików *.mlf, w którym znajdują się znaczniki początku i końca fonemu, jednak w tym przypadku nie było takiej potrzeby.

Fonem jest to najmniejsza rozróżnialna jednostka mowy [12]. Zbiór fonemów jest alfabetem fonetycznym. Wybrane alfabety fonetyczne prezentuje tabela 2.3 wraz z częstością występowania tych fonemów w języku polskim.

Tabela 3.1: Wybrane alfabety fonetyczne dla jęz. polskiego [12]

SAMPA	Grocholewski	AGH	ortogr.	fonetycz.	%	%
#				#	17.10	4.7
e	e	e	test	test	8.11	10.6
a	a	a	pat	pat	7.91	9.7
o	o	o	pot	pot	7.52	8.0
j	j	j	jak	jak	3.46	4.4
n	n	n	nasz	naS	3.39	4.0
t	t	t	test	test	3.39	4.8
i	i	i	PIT	pit	3.39	3.4
I	y	y	typ	tIp	3.37	3.8
r	r	r	ryk	rIk	2.98	3.2
v	w	v	wilk	vilk	2.89	2.9
m	m	m	mysz	mIS	2.76	3.2
p	p	p	pik	pik	2.65	3.0
u	u	u	puk	puk	2.62	2.8
s	s	s	syk	sIk	2.54	2.8
d	d	d	dym	dIm	2.18	2.1
k	k	k	kit	kit	2.09	2.5
w	l_	w	łyk	wIk	2.05	1.8
n’	ni	3	koń	kon’	1.97	2.4
l	l	l	luk	luk	1.92	1.9
z	z	z	zbir	zbir	1.67	1.5
g	g	g	gen	gen	1.38	1.3
b	b	b	bit	bit	1.34	1.5
S	sz	S	szyk	SIk	1.32	1.9
f	f	f	fan	fan	1.19	1.3
s’	si	5	świt	s’vit	1.16	1.6
Z	rz	Z	żyto	ZIto	1.06	1.3
tˆs	c	7	cyk	tˆsIk	1.06	1.2
x	h	x	hymn	xImn	1.01	1.0
tˆS	cz	0	czyn	tˆSIn	0.89	1.2
tˆs’	ci	8	ćma	tˆs’ma	0.83	1.2
dˆz’	dzi	X	dźwig	dˆz’vik	0.68	0.7
o+w∼	a_	2	ciąża	ts’ow∼Za	0.63	0.6
c	k	k	kiedy	cjedy	0.50	0.7
dˆz	dz	6	dzwoń	dˆzvon’	0.24	0.2
z’	zi	4	źle	z’le	0.21	0.2
N	N	N	pęk	peNk	0.21	0.1
J	g	g	giełda	Jjewda	0.14	0.1
e+j∼	e_	1	więź	vjej∼s’	0.06	0.1
dˆZ	drz	9	dżem	dˆZem	0.04	0.1

Przygotowane pliki służą do treningu systemu rozpoznawania mowy na przykład obliczenia parametrów HMM[10] lub modelu kNN[12].

3.2. Wykonanie projektu

3.2.1. Środowisko, w którym wykonano projekt i trudności związane z projektem

Pliki stworzono za pomocą programu Anotator [11]. Zadanie polegało na załadowaniu pliku dźwiękowego *.wav oraz pliku *.txt zawierającego transkrypcję, następnie zaznaczeniu, w oknie zawierającym przebieg czasowy nagrania, miejsca rozpoczęcia i zakończenia danego słowa. W razie potrzeby należało zmodyfikować plik tekstowy tak, by zgadzał się on z nagraniem.

Trudności także wiązały się ze wsłuchaniem się, kiedy kończy się dane słowo, a zaczyna kolejne, gdyż nie zawsze było to wyraźne - człowiek, mówiąc płynnie, przechodzi z jednego słowa do drugiego, łączy fonemy kończące jeden wyraz i rozpoczynający drugi. Często w dialogach występowały słowa wypowiadane równocześnie, co jest zrozumiałe, jako że były to dialogi (rozmówcy sobie przerywali i mówili w tych samych chwilach). W takich sytuacjach kolejność w jakiej anotowano słowa zależała od transkrypcji. Rysunek 3.1 pokazuje środowisko Anotator.

3.2.2. Plik *.mlf i plik *_words.xml

Powstały plik *.mlf (Master Label File) zawiera etykiety początku i końca danego wyrazu w pliku dźwiękowym. Pliki zawierające etykiety są bardzo przydatne, gdy przetwarzamy dużą ilość danych. Plik *.mlf pozwala na przechowywanie zestawu plików w jednym pliku oraz na ich przekierowanie [10].

Dzięki konkretnym plikom *.mlf możemy w jednym nagraniu wyróżnić słowa, nie potrzebujemy posiadać oddzielnych nagrań dla każdego słowa z osobna (co zdecydowanie zmniejsza ilość plików treningowych bez zmniejszania ilości danych).

Przykładowy plik *.mlf:

#!MLF!#
"C:/Users/Ola/Desktop/inż/LUNA.PL/KIEDY/DOBRAJAKOSC/F/1_2007-03-23_14_51_02/1_2007-03-23_14_51_02.wav"
5890000 11230000 xxx
11230000 13000000 dzień
13000000 15820000 dobry
20330000 22060000 dzień
22060000 24660000 dobry
24660000 27910000 proszę
27910000 30690000 pana
30690000 34050000 chciałam
34050000 35120000 się
35120000 41030000 dowiedzieć
41030000 43450000 jak

43450000 46480000 długo
46480000 49480000 jedzie
49480000 57750000 autobus
72160000 75610000 linii
75610000 78010000 sto
78010000 83870000 pięćdziesiąt
83870000 90090000 siedem
97910000 100190000 z
100670000 107740000 ulicy
107740000 118690000 Grójeckiej
118690000 120590000 przy
120590000 124610000 Bitwy
124610000 133890000 Warszawskiej
144210000 146710000 na
146710000 150100000 Plac
150100000 156480000 Wilsona
168420000 169610000 do
169610000 172390000 Placu
172390000 176590000 Wilsona
176590000 180940000 jedzie
180940000 183940000 od
183940000 187440000 dwudziestu
187440000 191390000 sześciu
191390000 192280000 do
192280000 196970000 trzydziestu
196970000 199190000 dwóch
199190000 203120000 minut
208220000 212400000 około
212400000 217790000 trzydziestu
217790000 221110000 minut
221110000 222420000 tak
222420000 225220000 tak
226240000 229690000 tak
231930000 233010000 a
233010000 235740000 jak
235740000 240930000 jedzie
240930000 251070000 autobus
251070000 256190000 linii
256190000 259060000 sto
259060000 267130000 osiemdziesiąt

267130000 272110000 jeden
274260000 276580000 tak
281830000 283120000 z
283120000 286710000 Placu
286710000 291960000 Wilsona
291960000 293610000 na
293610000 299250000 Wólkę
299250000 300520000 do
300520000 303190000 samej
303190000 306280000 pętli
306280000 308090000 tam
308090000 309450000 pod
309450000 313940000 cmentarz
313940000 315950000 pod
318660000 320570000 pod
320570000 323460000 bramę
323460000 326940000 główną
332010000 334290000 pod
334290000 337590000 bramę
337590000 342640000 główną
341010000 344360000 proszę
344360000 346580000 pani
346580000 348890000 jeśli
351420000 353960000 nie
353960000 357050000 będzie
357050000 359260000 miał
359260000 362710000 korka
362710000 364090000 to
364090000 364840000 w
364840000 368720000 dwadzieścia
368720000 371610000 minut
371610000 374880000 powinien
374880000 376580000 się
376580000 382130000 wyrobić
389090000 394210000 dwadzieścia
394210000 398160000 minut
398160000 401590000 dobrze
401590000 405430000 dziękuję
405430000 408030000 panu
408030000 410760000 bardzo

410760000 412250000 do
412250000 416780000 widzenia
406600000 409910000 proszę
413540000 415070000 do
415070000 418380000 widzenia
.

Liczby w pliku oznaczają czas początku i końca słów. Dokładny czas otrzymujemy po wymnożeniu tych liczb przez 100 ns.

Korpus Luna zawiera w sobie już podział nagrań na słowa, jednak jest to anotacja morfologiczna wykonana w celach lingwistycznych i nie jest ona przydatna do treningu systemu rozpoznawania mowy.

Przykładowy plik anotacji morfologicznej słów dołączony do korpusu:

```
<words>
<word id="1" word="xxx" lemma="xxx" POS="Np" morph="case.number.gender" />
<word id="2" word="dzień" lemma="dzień" POS="Nc" morph="nom.sg.m3" />
<word id="3" word="dobry" lemma="dobry" POS="ADJc" morph="nom.sg.masc" />
<word id="4" word="dzień" lemma="dzień" POS="Nc" morph="nom.sg.m3" />
<word id="5" word="dobry" lemma="dobry" POS="ADJc" morph="nom.sg.masc" />
<word id="6" word="proszę" lemma="prosić" POS="VV" morph="1.sg.gender.pres.ind.imperf" />
<word id="7" word="pana" lemma="pan" POS="Nc" morph="gen.sg.m1" />
<word id="8" word="chciałam" lemma="chcieć" POS="VV" morph="1.sg.fem.past.ind.imperf" />
<word id="9" word="się" lemma="się" POS="PART" morph="-" />
<word id="10" word="dowiedzieć" lemma="dowiedzieć" POS="VV" morph="inf.perf" />
<word id="11" word="jak" lemma="jak" POS="PART" morph="-" />
<word id="12" word="długo" lemma="długo" POS="ADV" morph="pos" />
<word id="13" word="jedzie" lemma="jechać" POS="VV" morph="3.sg.gender.pres.ind.imperf" />
<word id="14" word="autobus" lemma="autobus" POS="Nc" morph="nom.sg.m3" />
<word id="15" word="linii" lemma="linia" POS="Nc" morph="gen.sg.fem" />
<word id="16" word="sto" lemma="sto" POS="NUM" morph="nom.nm1" />
<word id="17" word="pięćdziesiąt" lemma="pięćdziesiąt" POS="NUM" morph="nom.nm1" />
<word id="18" word="siedem" lemma="siedem" POS="NUM" morph="nom.nm1" />
<word id="19" word="z" lemma="z" POS="PreP" morph="-" />
<word id="20" word="ulicy" lemma="ulica" POS="Nc" morph="gen.sg.fem" />
<word id="21" word="Grójeckiej" lemma="Grójecki" POS="ADJp" morph="gen.sg.fem" />
<word id="22" word="przy" lemma="przy" POS="PreP" morph="-" />
<word id="23" word="Bitwy" lemma="Bitwa" POS="Np" morph="gen.sg.fem" />
<word id="24" word="Warszawskiej" lemma="Warszawski" POS="ADJp" morph="gen.sg.fem" />
<word id="25" word="na" lemma="na" POS="PreP" morph="-" />
<word id="26" word="Plac" lemma="Plac" POS="Np" morph="acc.sg.m3" />
<word id="27" word="Wilsona" lemma="Wilson" POS="Np" morph="gen.sg.m1" />
```

<word id="28" word="do" lemma="do" POS="PreP" morph="-" />
<word id="29" word="Placu" lemma="Plac" POS="Np" morph="gen.sg.m3" />
<word id="30" word="Wilsona" lemma="Wilson" POS="Np" morph="gen.sg.m1" />
<word id="31" word="jedzie" lemma="jechać" POS="VV" morph="3.sg.gender.pres.ind.imperf" />
<word id="32" word="od" lemma="od" POS="PreP" morph="-" />
<word id="33" word="dwudziestu" lemma="dwadzieścia" POS="NUM" morph="gen.gender" />
<word id="34" word="sześciu" lemma="sześć" POS="NUM" morph="gen.gender" />
<word id="35" word="do" lemma="do" POS="PreP" morph="-" />
<word id="36" word="trzydziestu" lemma="trzydzieści" POS="NUM" morph="gen.gender" />
<word id="37" word="dwóch" lemma="dwa" POS="NUM" morph="gen.gender" />
<word id="38" word="minut" lemma="minuta" POS="Nc" morph="gen.pl.fem" />
<word id="39" word="około" lemma="około" POS="PART" morph="-" />
<word id="40" word="trzydziestu" lemma="trzydzieści" POS="NUM" morph="gen.gender" />
<word id="41" word="minut" lemma="minuta" POS="Nc" morph="gen.pl.fem" />
<word id="42" word="tak" lemma="tak" POS="PART" morph="-" />
<word id="43" word="?" lemma="?" POS="PQ" morph="-" />
<word id="44" word="tak" lemma="tak" POS="PART" morph="-" />
<word id="45" word="tak" lemma="tak" POS="PART" morph="-" />
<word id="46" word="a" lemma="a" POS="PART" morph="-" />
<word id="47" word="jak" lemma="jak" POS="PART" morph="-" />
<word id="48" word="jedzie" lemma="jechać" POS="VV" morph="3.sg.gender.pres.ind.imperf" />
<word id="49" word="autobus" lemma="autobus" POS="Nc" morph="nom.sg.m3" />
<word id="50" word="linii" lemma="linia" POS="Nc" morph="gen.sg.fem" />
<word id="51" word="sto" lemma="sto" POS="NUM" morph="nom.nm1" />
<word id="52" word="osiemdziesiąt" lemma="osiemdziesiąt" POS="NUM" morph="nom.nm1" />
<word id="53" word="jeden" lemma="jeden" POS="NUM" morph="nom.masc" />
<word id="54" word="tak" lemma="tak" POS="PART" morph="-" />
<word id="55" word="z" lemma="z" POS="PreP" morph="-" />
<word id="56" word="Placu" lemma="Plac" POS="Np" morph="gen.sg.m3" />
<word id="57" word="Wilsona" lemma="Wilson" POS="Np" morph="gen.sg.m1" />
<word id="58" word="na" lemma="na" POS="PreP" morph="-" />
<word id="59" word="Wólkę" lemma="Wólka" POS="Np" morph="acc.sg.fem" />
<word id="60" word="do" lemma="do" POS="PreP" morph="-" />
<word id="61" word="samej" lemma="sam" POS="ADJc" morph="gen.sg.fem" />
<word id="62" word="pętli" lemma="pętla" POS="Nc" morph="gen.sg.fem" />
<word id="63" word="i" lemma="i" POS="CC" morph="-" />
<word id="64" word="tam" lemma="tam" POS="PART" morph="-" />
<word id="65" word="pod" lemma="pod" POS="PreP" morph="-" />
<word id="66" word="cmentarz" lemma="cmentarz" POS="Nc" morph="acc.sg.m3" />
<word id="67" word="pod" lemma="pod" POS="PreP" morph="-" />

<word id="68" word="pod" lemma="pod" POS="PreP" morph="-" />
<word id="69" word="bramę" lemma="brama" POS="Nc" morph="acc.sg.fem" />
<word id="70" word="główną" lemma="główny" POS="ADJc" morph="acc.sg.fem" />
<word id="71" word="?" lemma="?" POS="PQ" morph="-" />
<word id="72" word="tak" lemma="tak" POS="PART" morph="-" />
<word id="73" word="pod" lemma="pod" POS="PreP" morph="-" />
<word id="74" word="cmentarz" lemma="cmentarz" POS="Nc" morph="acc.sg.m3" />
<word id="75" word="pod" lemma="pod" POS="PreP" morph="-" />
<word id="76" word="bramę" lemma="brama" POS="Nc" morph="acc.sg.fem" />
<word id="77" word="główną" lemma="główny" POS="ADJc" morph="acc.sg.fem" />
<word id="78" word="proszę" lemma="prosić" POS="VV" morph="1.sg.gender.pres.ind.imperf" />
<word id="79" word="pani" lemma="pani" POS="Nc" morph="gen.sg.fem" />
<word id="80" word="jeśli" lemma="jeśli" POS="CC" morph="-" />
<word id="81" word="nie" lemma="nie" POS="PART" morph="-" />
<word id="82" word="będzie" lemma="być" POS="VA" morph="3.sg.gender.pres.ind.imperf" />
<word id="83" word="miał" lemma="mieć" POS="VV" morph="3.sg.masc.past.ind.imperf" />
<word id="84" word="korka" lemma="korek" POS="Nc" morph="gen.sg.m3" />
<word id="85" word="to" lemma="to" POS="PART" morph="-" />
<word id="86" word="w" lemma="w" POS="PreP" morph="-" />
<word id="87" word="dwadzieścia" lemma="dwadzieścia" POS="NUM" morph="acc.nm1" />
<word id="88" word="minut" lemma="minuta" POS="Nc" morph="gen.pl.fem" />
<word id="89" word="powinien" lemma="powinien" POS="VV" morph="3.sg.masc.pres.ind.imperf" />
<word id="90" word="się" lemma="się" POS="PART" morph="-" />
<word id="91" word="wyrobić" lemma="wyrobić" POS="VV" morph="inf.perf" />
<word id="92" word="dwadzieścia" lemma="dwadzieścia" POS="NUM" morph="nom.nm1" />
<word id="93" word="minut" lemma="minuta" POS="Nc" morph="gen.pl.fem" />
<word id="94" word="dobrze" lemma="dobrze" POS="ADV" morph="pos" />
<word id="95" word="dziękuję" lemma="dziękować" POS="VV" morph="1.sg.gender.pres.ind.imperf" />
<word id="96" word="panu" lemma="pan" POS="Nc" morph="dat.sg.m3" />
<word id="97" word="bardzo" lemma="bardzo" POS="ADV" morph="pos" />
<word id="98" word="do" lemma="do" POS="PreP" morph="-" />
<word id="99" word="widzenia" lemma="widzenie" POS="Nc" morph="gen.sg.neut" />
<word id="100" word="proszę" lemma="prosić" POS="VV" morph="1.sg.gender.pres.ind.imperf" />
<word id="101" word="do" lemma="do" POS="PreP" morph="-" />
<word id="102" word="widzenia" lemma="widzenie" POS="Nc" morph="gen.sg.neut" />
</words>

Pliki *_words.xml zawierają analizę morfologiczną słów. Pliki mają następujący format [8]:

1. "Znacznik otwierający plik: <words>.
2. Znaczniki kolejnych słów, np. <word id="104" word="rogu" lemma="róg" POS="Nc" morph="loc.sg.m3"> Składają się na nie:
   - kolejny numer słowa id="";
   - forma wyrazowa word=""
   - forma hasłowa lemma=""
   - klasa gramatyczna POS=""
   - charakterystyka fleksyjna formy wyrazowej morph="".
3. Znacznik zamykający plik </words>."

Forma wyrazowa mówi o tym, w jakiej postaci słowo wystąpiło w nagraniu, forma hasłowa wyraża, jaka jest forma podstawowa tego wyrazu, klasa gramatyczna określa jaka to część mowy (np. zaimek osobowy, liczebnik własny, czasownik właściwy etc.), charakterystyka fleksjyna charakteryzuje pozostałe cechy tego słowa (np. tryb, formę, czas, osobę etc.).

3.2.3. Czas pracy

Tworzenie plików *.mlf za pomocą programu Anotator było bardzo czasochłonne. Dzięki wbudowanej funkcji program zapisywał czas, w jakim pracowano nad danym plikiem. **Pliki trwające łącznie około 9 godzin (z jedenastu godzin nagrań zawartych w korpusie) anotowano w czasie 118 godzin i 8 minut**. Proporcja czasu pracy do czasu nagrań wynosi 13,1. Zwykle proporcja ta wynosi od 20 do 40, co oznacza, że osiągnięto bardzo dużą efektywność pracy.

3.3. Prace dodatkowe

W ramach pracy dyplomowej wykonano także prace związane z tworzeniem plików *.mlf nagrań nie związanych z korpusem LUNA dla Zespołu Przetwarzania Sygnałów AGH, potrzebnych do Systemu Rozpoznawania Mowy AGH. System ten został zlicencjonowany firmie Stanusch Technologies [3][13]. Inną partię nagrań wykostano do przygotowania prototypu programu Wirtualna Mysz, który zostanie w styczniu 2012 roku przekazany do testów podopiecznej Fundacji Mimo Wszystko Anny Dymnej.

4. Zakończenie

LUNA jest wyjątkowym korpusem mowy polskiej z tego względu, że opiera się na spontanicznych dialogach. Większość najpopularniejszych polskich korpusów zawiera wypowiedzi sztywno zarysowane lub mowę częściowo spontaniczną. Jest on także bogato anotowany na wielu poziomach opisu. Dzięki tej pracy jest on jeszcze obszerniejszy.

Prace nad projektem pochłonęły bardzo dużo czasu. Proporcja czasu poświęconego anotacji do czasu nagrań wynosi 13,1, co mówi o dużej efektywności pracy.

Projekt był częściową realizacją grantu Nr OR00001905, MNSiW, “Aplikacje technologii mowy w systemach bezpieczeństwa publicznego”, AGH, Kraków. W ramach projektu wykonano także anotacje innych nagrań, nie związanych z korpusem LUNA, które wykorzystano w systemie AGH [13].

Bibliografia

[1] Eastern European Speech Databases for Creation of Voice Driven Teleservices. http://www.fee.vutbr.cz/SPEECHDAT-E/.

[2] Korpusy mowy. http://www.korpusy.net/index.php/korpusy-mowy.

[3] Strona internetowa Zesopłu Przetwarzania Syganłów AGH. http://www.dsp.agh.edu.pl.

[4] TranscriberAG. http://transag.sourceforge.net/index.php?content=screenshots.

[5] G. Demenko, S. Grocholewski, K. Klessa, J. Ogórkiewicz, A. Wagner, M. Lange, D. Śledziński, and N. Cylwik. JURISDIC – Polish speech database for taking dictation of legal texts. pages 1280–1287, 2008.

[6] S. Grocholewski. Baza nagrań sygnałów mowy CORPORA. Technical report, Politechnika Poznańska, 1997.

[7] J. Lööf, C. Gollan, and H. Ney. Cross-language bootstrapping for unsupervised acoustic model training: Rapid development of a Polish speech recognition system. *Proceedings of Interspeech, Brighton*, pages 88–91, 2009.

[8] M. Marciniak. *Anotowany korpus dialogów telefonicznych (Eng. Anotated corpus of telephone dialogues)*. Exit, Warszawa, 2010.

[9] K. Szklanny. Przygotowanie bazy difonów języka polskiego dla realizacji syntezy mowy w systemie MBROLA. In *50. Otwarte Seminarium z Akustyki*, pages 391–394, 2003.

[10] S. Young, G. Evermann, M. Gales, Th. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *HTK Book*. Cambridge University Engineering Department, UK, 2005.

[11] B. Ziółko, B. Miga, and T. Jadczyk. Semisupervised production of speech corpora using existing recordings. *Proceedings of International Seminar on Speech Production (ISSP'11), Montreal*, 2011.

[12] B. Ziółko and M. Ziółko. *Przetwarzanie mowy (Eng. Speech processing)*. Wydawnictwa AGH, 2011.

[13] M. Ziółko, J. Gałka, B. Ziółko, T. Jadczyk, D. Skurzok, and M. Mąsior. Automatic speech recognition system dedicated for Polish. *Proceedings of Interspeech, Florence*, 2011.