

# On Description of Polish Proper Names in City Transportation System

Joanna Rabięga-Wiśniewska  
jrabięga@ipipan.waw.pl

Institute of Computer Science  
Polish Academy of Sciences

Proper Names in Spoken Language, nomina2007  
22nd November 2007



# Plan of the Talk

- Introduction
- LUNA project
- Proper Names in LUNA
- Description of Proper Names
- Plans and Summary

# Motivation

This work is inspired by goals of LUNA project. There are natural spoken dialogs (telephone conversations) collected, where a lot of Proper Names of Places appear.

The aim is to present the methodology of transcribing and interpreting Proper Names at morphological level.

Aims in LUNA project:

- to collect PN from recorded dialogs for domain description,
- to transcribe PN in a way allowing the later analysis,
- to extend a lexicon of morphological analyzer.

# Longterm Aims

- Preparation of an electronic lexicon of Polish Proper Names,
- Overview of morphological and syntactic structure of Polish PN,
- Elaboration of tools allowing recognition (extraction) Polish PN in texts.

# LUNA Project

LUNA – spoken Language Understanding in multilingual communication systems, <http://www.ist-luna.eu/>

## Aims of LUNA

The aim of the project is to elaborate a set of Spoken Language Understanding tools that can be used in dialog services of several languages. The long-term aim is human/machine communication.

# Goals of Project

Projects goals:

- Collecting (natural) dialogs corpora for French, Italian, Polish
- Semantic annotation of dialogs
- Collecting the data for training of:
  - automatic speech analysis and synthesis
  - dialog systems

# Polish Corpus of Dialogs

- To collect and annotate a dialog corpus in two steps:
  - recording 500 human/human dialogs
  - morphosyntactic and semantic annotation
  - recording 500 human/machine dialogs (Wizard-of-Oz)
  - morphosyntactic and semantic annotation
- Domain: public transport in Warsaw

# Warsaw City Transportation Center

Tel. 94-84

Warsaw residents call City Transportation Center approximately 200-300 times per day and they receive information on routes between the given points in the city, time schedules of public transport, trip durations and others.

Topics of conversations:

- How can one travel from point A to B?
- What time does next/last/low bus come?
- At what stop should one get off?
- Where should one transfer to get to A?
- Who has discounts in public transportation?
- Why is the tram late?

# Prototype Dialog in LUNA

operator: xxx dobry wieczór

*operator – greeting*

user: dobry wieczór chciałabym jutro tramwaj jutro rano tramwaj  
dwudziesta mijanka **Grody** w stronę **Koła** tak gdzieś ko~ po siódma  
siódma trzydzieści

*user – question about timetable of a tram 22 from place X in direction Y*

operator: siódma trzydzieści trzy jest

*operator – answer*

user: trzydzieści trzy dobra dziękuję

*user – repeating, thanks*

1. user: bardzo

2. operator: proszę

*overlapping speech*

# Transportation System in Warsaw

- Streets, Alleys, Traffic-Circles:
  - ulica Broniewskiego Broniewski's Street,
  - Aleja Na Skarpie At Scarp Alley,
  - Rondo Gen. Charles'a de Gaulle'a General Ch. de Gaulle Circle.
- Squares:
  - Plac Konstytucji Constitution Square,
  - Plac Wilsona Wilson's Square.
- Stations, Stops:
  - Dworzec Centralny The Central Station,
  - przystanek Nowy Świat New World bus-stop.
- Buildings, Shopping Centers:
  - Błękitny Wieżowiec Blue High-Rise,
  - Galeria Mokotów Mokotów Gallery.

# Administrative Units

- Quarters:
  - Wola,
  - Śródmieście.
- Settlements:
  - Rakowiec,
  - Za Żelazną Bramą.
- Township:
  - Ursynów,
  - Wesoła.

# Background for the Lexicon

## Resources:

- public lists of names of Warsaw streets, alleys, squares etc.,
- web-sites, e. g. <http://www.ewarszawa.com/przewodnik/>.

## Usage in practice:

- a corpus of dialogs that were recorded in the city call center.

# Proper Names in the Analyzer

## AMOR analyzer

Collected data is added to the morphological lexicon of AMOR analyzer. The morphological analysis concerns single words. The analysis of names give an information how many inflectional patterns are needed for given proper names.

At the time the lexicon contains:

- 107 inflectional patterns for nominal PN (900 nouns),
- 8 inflectional patterns for adjectival PN (300 adjectives),
- 10 inflectional patterns for numeral PN (20 numerals),
- single rules for acronyms, prepositions (100).

# Proper Names – Statistics

	number
Analyzed dialogs	125
PN word-forms	1226
PN identified lemmata	325 (10)
PN POS	6
PN inflection tags	44

# Proper Names – Statistics

POS	number
Np	799
ADJp	395
PN incorrect	10
Inflection tags	number
nom.sg.masc	139
nom.sg.fem	120
gen.sg.masc	232
gen.sg.fem	158
-	32
voc.sg.m1	2

# Introduction

A description of Proper Names in LUNA is meant to serve the whole possible information about names used by Warsaw residents at least in two ways.

- 1 Annotation at transcription level gives an opportunity to look for unusual pronunciation, mistakes and distortion in speech.
- 2 Annotation at morphological level gives quantitative information about variability of morphological means.

# Speech to Text

## Rules of transcription in respect to Proper Names.

- Words are capitalized following the standards of the language.
- Acronyms are transcribed with capital letters.
- Spellings are transcribed using capital letters separated by spaces and tagged with a symbol pron=SPELLED.
- Syllabified words are transcribed using capital letters separated by spaces and tagged with a symbol pron=SYL.
- Foreign words or acronyms are transcribed in their original orthographic form and tagged with a symbol lang= and the name of the language.
- Possible truncation are marked by lex=∼.
- Mispronunciation are transcribed with the correct form and a tag pron=\*.

# Examples

Proper Name	Transcription
name	Park Żeromskiego
acronym	STOCER
spellings	[pron=SPELLED-] F S O [-pron=SPELLED]
syllabified names	[pron=SYL-] Grenadierów [-pron=SYL]
foreign names	[lang=English-] Blue City [-lang=English]
foreign spellings	[lang=English-] [pron=SPELLED-] C D [-pron=SPELLED] ROM [-lang=English]
truncation	Wol[lex=~] Wolska, Woł[lex=~-]w[-lex=~]skie

# Text to Interpretation

Description of PN required modification and extension of AMOR lexicon.

- PN obtain specific POS characteristics, e.g. ADJp = proper adjective (comp. ADJc = common adjective).
- Some of acronyms can undergo declension, e.g. ZUS-u, that is why they are kept in the lexicon as nouns as well.
- False PN, users mistakes get a partial POS characteristics, PN = proper name.
- Truncated names remain without interpretation (-).

# Examples

Proper Name	Interpretation
word='Berestecka'	lemma='Berestecki' POS='ADJp' morph='nom.sg.fem'
word='SKM'	lemma='SKM' POS='acronym' morph='-'
word='STOCERze'	lemma='STOCER' POS='Np' morph='loc.sg.m3'
word='Blizna'	lemma='- ' POS='PN' morph='- '
word='Ra~'	lemma='- ' POS='- ' morph='- '

# Problems and peculiarities

For many names there are more than one variant, e. g. names of persons in the street names are often omitted:

- Zygmunta Krasińskiego → Krasińskiego

But sometimes full names are more frequent: *Emilii Plater*.

Complicated names are simplified:

- Bitwy Warszawskiej 1920 r. → Bitwy
- Plac Powstańców Warszawy → Plac Powstańców
- Generała Antoniego Józefa Madalińskiego → Madalińskiego

# Problems and peculiarities

## Inflection of foreign Proper Names

- Rondo (im. Charles'a) de Gaulle'a
- Transcription:  
Rondo [lang=French-] de Gaulle [-lang=French]a
- Plac (Thomasa Woodrowa) Wilsona
- Transcription:  
Plac [lang=English-] Wilson [-lang=English]a

# Problems and peculiarities

Warsaw citizens give also “their” names some places or buildings. Sometimes they are better known (recognized) than the real ones:

- aleja Prymasa Tysiąclecia – aleja Wyszyńskiego,
- Pomnik Braterstwa Broni – Pomnik Czterech Śpiących,
- al. Jana Pawła II – ul. Marchlewskiego,
- Rondo Zgrupowania Radość – Rondo Babka.

# Plans and Summary

- This is the first attempt to Polish Proper Names description collected from spoken dialogs.
- The data referred to public places and transportation system in Warsaw is collected as the base for future proper names recognition system.
- The analysis of the given data will allow to describe syntactic patterns of the typical proper names of places.
- It is planned to build a dictionary of proper names using the available corpora and tools elaborated for the purpose.

Thank you for your attention!