Intelligent Information Systems 9999
ISBN 666-666-666, pages 1–14

Constructing an Electronic Dictionary of Polish Urban Proper Names

Małgorzata Marciniak[1], Joanna Rabiega-Wiśniewska[1], Agata Savary[3], Marcin Woliński[1], and Celina Heliasz[2]

[1] Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
[2] Institute of Polish Language, University of Warsaw, Poland
[3] Université François Rabelais Tours, France

Abstract

We present an ongoing project of an electronic dictionary of Polish urban proper names. We show . . .

1 Introduction

In the paper, we present an ongoing project[1] of creating a computer dictionary of Polish urban proper names. Our goal is to create a dictionary that can be used for the recognition and generation of proper names in written texts as well as in dialogues concerning a large agglomeration.

The idea of developing such a dictionary is one of the outcomes of the LUNA EU 6th Framework Programme in which ICS PAS is involved. The main goal of this project is to create a robust and effective spoken language understanding module, which can be used in developing automatic telecom services. For Polish, the city transportation domain was chosen, and tools for automatic recognition of the meaning of utterances concerning this domain are under development. One of the very important milestones of the project was the creation of an annotated corpus of dialogues concerning public transportation (Mykowiecka *et al.*, 2007). The corpus was collected at the Warsaw Transport Authority Information Center where operators provide information on tram and bus connections, schedules, routes, fares, reductions etc. During the dialogues annotation the problem of lemmatization (assigning a canonical form) of Polish proper names arose. This problem is not straightforward because of rich inflection and relaxed word order in Polish. Moreover, rather little work on this subject has been undertaken from the computational point of view (Piskorski *et al.*, 2007; Piskorski and Sydow, 2007). In Marciniak *et al.* (2008), the problem of lemmatization of urban proper names is described and a practical but not universal solution is presented.

The long-term aim of LUNA project is improving dialogue systems. The first approach to a dialogue system concerning Warsaw transportation, focused mainly

[1] The project is partially financed by the MNiSW decision number 567/6. PR UE/2008/7.

on speech recognition, is described in Marasek *et al.* (2009). The thoroughly designed dictionary of urban proper names is essential for developing such a dialogue system. The dictionary should enable:

- Recognition of different grammatical forms and variants of the same proper name. For example, the full official name of the street *ulica Bitwy Warszawskiej 1920 r.* is abbreviated in practice into *ulica Bitwy Warszawskiej.* For speech recognition and generation the year should be represented by words *tysiąc dziewięćset dwudziestego roku* 'nineteen twenty'. Our dictionary ought to include all viable possibilities.
- Representation of former and informal names of the city's objects and connection of different names with the same object, e.g., for the street *ulica ks. Jerzego Popiełuszki* the previous name was *ulica Stołeczna.*
- Lemmatization of a proper name — assigning its base form. Names are lemmatized in different ways, for example, for a street name in locative case *ulicy*$_{Loc}$ *Marszałkowskiej*$_{Loc}$ the lemma is *ulica*$_{Nom}$ *Marszałkowska*$_{Nom}$, while for the street *ulicy*$_{Loc}$ *Calineczki*$_{Gen}$ it is *ulica*$_{Nom}$ *Calineczki*$_{Gen}$.
- For a given base form, generation of a desired inflected form — function opposite to lemmatization.
- If several objects have the same name, we should be able to distinguish them. For example, streets *ulica Sportowa* exist in many towns of the Warsaw agglomeration. We want to represent all of them, so we have to introduce several objects with the same name assigned to them.

To account for all these phenomena we have devised the following structure for the dictionary. Each lexical entry describes one name with all its grammatical forms and variants. Lexical entries are linked with city objects, which represent the actual or historical objects referenced by the names. Several names (former, informal, etc.) may be linked with a single city object, and many objects with a single name.

The organization of the paper is as follows. Section 2 provides the description of linguistic and pragmatic features especially concerning variants of the same name. For realization of the project we created our own tool *Toposław* which cooperates with *Morfeusz*, a morphological analyser and generator for Polish words, and *Multiflex*, a cross-language morpho-syntactic generator of multi-word units (section 3). Finally, in section 4 we present the data collected within the project.

2 Linguistic and Pragmatic Phenomena

In this section we present linguistic and pragmatic phenomena which were taken into account while designing the dictionary. Some linguistic observations such as the description of abbreviations, acronyms and initialisms, numerals and name variations based on Warsaw toponyms have been already presented in Savary *et al.* (2009). Here, we focus on these features of proper names that are reflected in the lexicon as name variants and their syntactic structure.

The description of a multi-word lexical unit in the dictionary consists of:

- an entry name,
- an inflection description of name components,
- a graph representation of all possible name variants,
- pragmatic characteristics of the name variants (“official”, “neutral”, “neutral spoken”),
- stylistic characteristics of the name (“former”, “common”, and “marked”),
- a link between a name and an object described by a hierarchy (e.g. area, building, etc.).

We limit the range of the data in our lexicon to proper names of the transportation system and public places in Warsaw. Thus, we consider the following types of places: Warsaw administrative units, traffic routes, stopping places, parks and gardens, cemeteries, public institutions and facilities, mansions, monuments, commercial centres, and business establishments (examples 1-3).

(1) Nowy Dwór Mazowiecki (town in Warsaw agglomeration)

(2) Pomnik Stefana Starzyńskiego (monument)

(3) Urząd Miasta Stołecznego Warszawy (office)

Significant part of toponyms contain other names, e.g. person names, see example 2. To describe a formal structure of names, one should take into account those embedded names, too. In the dictionary we have decided to include names of persons as separate lexical units (examples 4-6).

(4) Stefan Starzyński

(5) Jan Rodowicz „Anoda”

(6) Król Jan I Olbracht

2.1 Proper Names Variants

In the dictionary we have introduced different variants of a name. We mark three of them as they are pragmatically important from application point of view:

- an “official” name variant represent a variant used in official lists of city names,
- a “neutral” name variant is a shortened variant used in written texts,
- a “neutral spoken” name variant is a shortened variant used in spoken language.

Usually, proper name dictionaries (Grzenia, 2003; Rzetelska-Feleszko, 2005; Cieślikowa, 2008) and official lists of public places contain names in their full form that is not commonly used, see examples 7 and 8. We have decided to mark them as “official” names in our lexicon. We describe names in an extensive way by gathering together all their possible pragmatic variants. Among them, we choose one variant, short enough for efficient communication and object identification, which is marked as “neutral” and “neutral spoken”.

(7) official name: Rezerwat przyrody Las Kabacki im. Stefana Starzyńskiego (nature reserve)
neutral name: Las Kabacki

(8) official name: Aleja Jana Rodowicza „Anody" (avenue)
neutral name: Aleja Rodowicza

We also decided to recognize orthographic variants of proper names which usually concerns punctuation marks, such as a hyphen and a quotation mark. Example 9 shows three popular ways of writing a name of the Polish general inside a name of a street. Variants refer to his pseudonym *„Grot"*.

(9) ulica Grota-Roweckiego, ulica „Grota" Roweckiego, ulica Grota Roweckiego

Thanks to our formalism we define name variants which differ in number and order of their components. Frequently one (or more) proper name component is omitted, especially in spoken language. We accept a shortened name if it is usually recognized by city residents, see example 10. As an order variant we accept such a change of component order that holds true to Polish grammar or stylistics rules. Example 11 shows two correct positions of the pseudonym *„Anoda"* according to Polish rules of usage.

(10) official name: Pałac Kultury i Nauki,
shortenedshort^{MW} variant: Pałac Kultury

(11) official name: Aleja Jana Rodowicza „Anody",
order variant: Aleja Jana „Anody" Rodowicza

2.2 Stylistic labels

In the history of every city there are places whose names have changed, sometimes more than once. The history of Warsaw street names is a very good example. During the XX century a lot of street were re-named as a consequence of wars and political system changes. There are examples of names that re-appeared in the present, as in example 12, but a lot of names – especially connected to the previous socialist period – were eliminated. Some of the former names are still in use, see example 13. At present some places are also renamed, as in example 14. In those cases, if we consider a proper name which is no longer on the official list, it is labelled as "former".

(12) ulica Tadeusza Hołówki (1933-1951, and since 1991)[2] (street),
past names: ulica Karpia (1951-1958), ulica Wczasowa (1958-1991)

(13) Aleja Solidarności (avenue),
past name: Aleja Karola Świerczewskiego

(14) Aleja Jana Rodowicza „Anody" (avenue),
a former part of the street: ulica Jana Rosoła

Warsaw citizens sometimes invent their own popular names for places or buildings. If these names are commonly known and recognized by city residents we represent them in the dictionary as different names of the same object. We distinguish two types of such proper names: "common" and "marked" names. Common

[2] According to Handke (1998)

names usually bring some associations of place shape, colour or physical feature, as in example 15. Sometimes, residents give an ironic name to the place or object, see example 16, which expresses their disapproval or funny association connected to it.

(15) official name: Pomnik Bohaterów Warszawy 1939-45 (monument),
common name: Warszawska Nike

(16) official name: Pomnik Józefa Piłsudskiego (monument),
marked name: Dziadek Parkingowy

2.3 Hierarchy of Concepts

We decided to include some semantic information in the dictionary. For this purpose we defined a very simple hierarchy of concepts. All city objects are connected to exactly one concept represented by leafs of a hierarchy presented below. The hierarchy is not crucial for conducting dialogues in the city transportation call center. Thus, we made it possible to introduce a hierarchy of concepts into our dictionary, but we have not developed a complex one.

- PLACE
  - AREA:
    - ADMINISTRATIVE AREA: concept representing administrative division of an agglomeration. It represents: towns e.g. *Nowy Dwór Mazowiecki*, city districts and parts of districts. It is possible that one area includes others such as: city district *Mokotów* is divided into *Górny Mokotów* and *Dolny Mokotów* and the last one contains *Czerniaków* — all these names are connected to this area concept.
    - PUBLIC AREA: areas that can be visited by citizens like: parks, cemeteries, nature reserves.
    - CLOSED AREA: areas that are not accessible to all citizens like: military training areas, bus depots.
  - COMMUNICATION POINT
    - STOP: stops of all means of transport, including different types of buses (municipal, private, long-distance), trams, metro.
    - RAILWAY STATION: e.g.: *Dworzec Centralny*
    - AIRPORT: e.g.: *Port Lotniczy im. Fryderyka Chopina w Warszawie*
  - ROAD:
    - STREET: includes avenues, roads and highways. It can also refer to a named route like *Wisłostrada* which consists of about a dozen streets.
    - SQUARE: includes also roundabouts.
    - BRIDGE, VIADUCT, TUNNEL.
  - FACILITY: buildings, their parts and groups of buildings such as: hospitals, universities, theaters, museums, shopping centers, stadiums, industrial plants, etc. For example, this concept refers to *Pałac Kultury i Nauki* — an exhibition centre and office building, in which a theater *Teatr Dramatyczny* is located and both names are connected to the facility concept.

  - HYDRONYM: it applies to all bodies of water like rivers, brooks, lakes, ponds, e.g.: *Rzeka Wisła, Jeziorko Czerniakowskie*
  - MONUMENT: *Pomnik Adama Mickiewicza*
- PERSONAGE: refers to people like *Adam Mickiewicz*, fictitious characters — a literary hero *Michał Wołodyjowski*, and religious characters e.g.: *Jan Chrzciciel* (John the Baptist).

3 Tools

The dictionary is built using a dedicated graphical interface *Toposław*. The formalism used in the description of compounds is implemented by *Multiflex*, which in turn utilises *Morfeusz* for inflecting components of names.

3.1 Morfeusz

Morfeusz SGJP is a morphological analyser for single words based on the data of the *Grammatical Dictionary of Polish* (Saloni *et al.*, 2007). The interface and the tagset of *Morfeusz SGJP* is compatible with the previous version called *Morfeusz SIaT* (Woliński, 2006), but features a much improved dictionary (ca. 245,000 lexemes, ca. 4,000,000 different textual words).

Morfeusz SGJP has two modules. The first one, given any textual word, provides all possible interpretations of this word as a form of a Polish lexeme. The second module generates all possible forms of a lexeme when given the lemma and the part of speech.

The IPI PAN tagset used by *Morfeusz* is based on a set of morphological, morphosyntactic and syntactic criteria (cf. Przepiórkowski and Woliński, 2003; Woliński, 2003). It operates with more detailed grammatical classes than traditional parts of speech (POS). Some of these classes, however, correspond directly to the traditional POS, e.g., noun, adjective, adverb, preposition, conjunction. Grammatical categories assumed in the tagset include well established ones such as number, case, person, degree, aspect, gender, as well as more restricted categories first introduced in the work of Jan Tokarski and Zygmunt Saloni (Tokarski, 2002).

The following example presents the analysis of the phrase *Jana Rodowicza „Anody”*:

(17)	1	Jana	Jan	subst:sg:gen.acc:m1
	2			sp
	3	Rodowicza	Rodowicz	subst:sg:gen.acc:m1
	4			sp
	5	„	„	interp
	6	Anody	anoda	subst:sg:gen:f|subst:pl:nom.acc.voc:f
	7	”	”	interp

The phrase is segmented as required by *Multiflex*, in particular segments 2 and 4 are spaces. Lemmas *Jan* (a first name) and *Rodowicz* (a last name) are capitalised, *anoda* ‘anode’, being a common noun, is put in lowercase. The tags in this example consist of the following components: *subst* — noun, *sg* — singular, *pl* — plural, *nom* — nominative, *gen* — genitive, *acc* — accusative, *voc* — vocative, *m1* — masculine personal, *f* — feminine, *interp* — punctuation, *sp* — blank. A vertical bar denotes alternative tags, a dot denotes alternative values of a given grammatical category.

3.2 Multiflex

Multiflex is a cross-language morpho-syntactic generator of multi-word units (MWUs), Savary (2005a), Savary (2005b). It allows us to exhaustively and precisely describe the inflection paradigm and variation of a MWU via a 'two-layerponieważ „two-layer” się kojarzy, proponuję two-tier. Teraz jest moda na multi-tier applications^{MW} approach'. First, an underlying morphological module such as *Morfeusz* allows us to tokenize the MWU lemma, to annotate its components, and to generate inflected forms of simple words on demand. Then, each inflected multi-word form is seen as a particular combination of the inflected forms of its components. For instance, Fig. 1 shows, in its upper part, the segmentation of the person name from example (5) into seven tokens, four of which are punctuation marks (spaces I strongly object to calling the space "a punctuation"^{MW} and quotes). Each token which can be inflected, is annotated by a *Morfeusz*-like taga czemu like?^{MM} (cf section 3.1). The inflection of the whole compound is described by a graph. Each path in the graph describes one or more inflected forms and variants.

Morphological descriptions appearing inside the boxes refer to single constituents. For instance, in the graph in Fig. 1 each non empty box refers to one of the seven previously annotated constituents ($1 through $7). If a constituent number appears in a box with no extra information then the token should be left unchanged. For instance the boxes containing $\langle \$2 \rangle$, $\langle \$4 \rangle$, $\langle \$5 \rangle$, and $\langle \$7 \rangle$ indicate that each space and quote should be recopied as such. If however a constituent number is followed by a set of category-value equations then the constituent should be inflected into the desired form. For instance, the box $\langle \$1 : Case = \$c \rangle$ indicates that the first constituent (*Jan*) should be inflected for case, while keeping its other categories unchanged (*m1* and *sing*sg^{MW}). The *unification variable $c*, which is common to components $1, $3 and $6, allows us to express their gendercase!!!^{MW} agreement.

Morphological descriptions appearing under the boxes describe the inflectional features of the whole compound. Here, $\langle Gen = \$3.Gen; Nb = \$3.Nb; Case = \$c \rangle$ says that the gender and the number of the whole compound is inherited from the third constituent, as it appears in the compound lemma (for this entry *m1* and *sing*sg^{MW}), while the case is determined by the current value of variable *$c*.

The full exploration of this graph results in the generation of all inflected forms and variants of the compound, in particular all the variants in example (18) inflected for all cases. Note that numbering of the constituents allows the graph to represent their omissions, insertions and order change.

(18) Jan Rodowicz „Anoda”, Jan Rodowicz Anoda,
Jan „Anoda” Rodowicz, Jan Anoda Rodowicz
Jan Rodowicz, Rodowicz „Anoda”, „Anoda” Rodowicz, Rodowicz

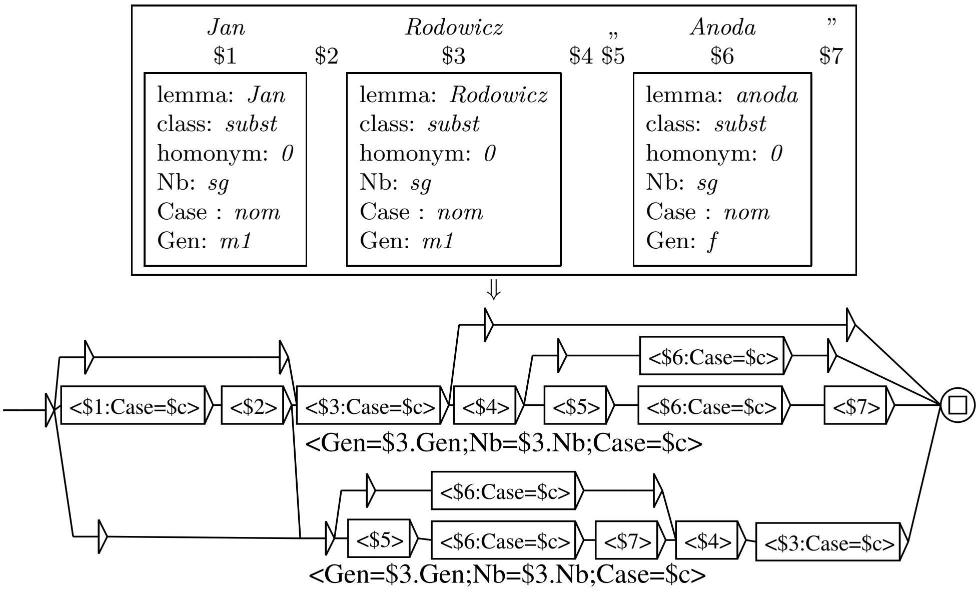

FIGURE 1: Lemma annotation and inflection graph for the patronym *Jan Rodowicz „Anoda”* containing elliptical variants and inversions

Embedded compounding, frequent in urban toponyms, as discussed in section 2, can be easily expressed in *Multiflex.* Given the description in Fig. 1 the name of the avenue dedicated to this person can be annotated as a three-component compound in which the third component is a compound itself, as shown in Fig. 2. The inflection graph in this case is rather trivial. It says that the first constituent, *Aleja*, inflects for case only, and that it can be omitted. The third constituent can take any variant of its genitive form. Thus, all sequences in example (19) can be generated. The whole compound inherits its number and gender from the first constituent, as it appears in the lemma, and its case from the first constituent as it appears in the particular inflected form. Note that omitting the head noun *Aleja* does not provoke a head shifting: the third constituent remains always in genitive.

(19) Aleja Jana Rodowicza „Anody”, Jana Rodowicza „Anody”,
Aleja Jana Rodowicza Anody, Jana Rodowicza Anody,
Aleja Jana „Anody” Rodowicza, Jana „Anody” Rodowicza
Aleja Jana Rodowicza, Jana Rodowicza, Aleja Rodowicza, Rodowicza, . . .

Due to the possibility of embedding descriptions, the graph in Fig. 2 can be applied to most street names having a genitive government structure Noun Noun$_{Gen}$ (cf section **??**), independently of the complexity of their genitive complement. In particular, examples (12), (13), and (9) are described with the same graph.

Savary *et al.* (2009) gives a more detailed discussion on using *Multiflex* with *Morfeusz* and their application to the present study of Polish urban toponyms.

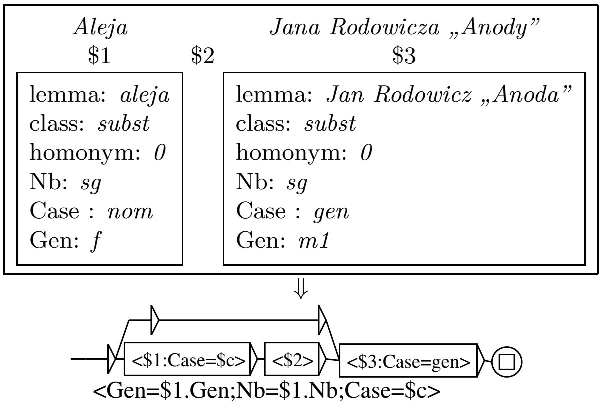

FIGURE 2: Embedded compounding in *Aleja Jana Rodowicza „Anody”*

3.3 A Dictionary Editor *Toposław*

To facilitate the creation of the dictionary a Java application named *Toposław* has been implemented. The list of names to be described is loaded to the application's database. The task of the operator is to describe inflection and provide some features of the object referenced by each name. To describe inflection one needs to construct the labelled lemma as needed by Multiflex and then to assign an inflection graph to the name.

As explained in section 3.2 the lemma of a compound has to be labelled with morphological features of the words constituting the compound. For that purpose names get analysed by Morfeusz. If any of the words has multiple interpretations, the lexicographer has to select the one appropriate for the lemma of the compound. Obviously, the form need not be the lemma of the word in question. For example, *Aleja Jana Rodowicza „Anody”* is a compound lemma where *Aleja* is in *nominative*, but the three other tokens represent genitive forms of their respective lexemes.

The operator can also mark a fragment of the name as a sub-compound to be described separately. As noted in section 2 we use this mechanism for compoundnie tylko, ulica Agatki^{JR}przecież dla Agatki nie będziemy tworzyć podgrafu!^{MW} names of persons, which tend to occur in several urban toponyms.

The tool keeps a pool of inflection graphs, which can be assigned to names. A Unitex-like editor is used to create and edit the graphs (Paumier, 2002). New graphs can be created from scratch or based on any existing one. Since inflectional behaviour is often shared by large groups of names, the tool has a mechanism for selecting a group of names and assigning the same graph to all of them simultaneously.

The pragmatic labels mentioned in section 2.1 (“official”, “neutral”, “neutral spoken”) are attached to particular paths within a graph, which means they apply only to particular variants of the name.

Each name can be linked to a city object referenced by it. As stated in section 2.1, some city objects have several names. Such names (as opposed to variants of one name) are described separately and only then linked to one city object.

TABLE 1: Types of city objects represented by proper names

Areas	380
administrative areas, e.g. districts, city areas	220
public areas, e.g. parks, national parks, cemeteries	142
closed areas, e.g. depots, military training grounds	18
Roads	4998
streets, boulevards, avenues	4856
squares, roundabouts	133
bridges, tunnels	9
Communication points	1901
bus/tram stops, metro stations	1788
railway stations	112
airports	3
Buildings, e.g. theatres, cinemas, churches	473
Hydronyms, rivers, lakes, canals	37
Monuments	110
TOTAL	7897

Each of the links is labelled with the type of the name: “common”, “marked”, or “former”.

In the dictionary, we include names which are still used by people for orientation, but which denote places that no longer exist (e.g., *Kino Moskwa* — a famous Warsaw cinema). For that reason it is possible that the only name linked with a city object is tagged as “former”.

Sometimes, the same name is used for several city objects. This should not be the case within the city proper, but can happen when we include the names from the satellite towns forming the Warsaw agglomeration. For example, most Polish towns have a street named after *Józef Piłsudski*, a landmark figure in Polish history.

City objects are also categorised using the hierarchy described in section 2.3.może wyrzucić to zdanie bo taki krótki akapit wyszedł?MM

4 Description of the Collected Data

The collected proper names come from the city hall, offices and websites, web pages of *Zarząd Dróg Miejskich* (the creator of the City Information System) and directly from *Biuro Geodezji i Katastru.*

At present the database of city proper names includes of 7897 records. There are proper names of different city objects, e.g. streets, squares, roundabouts, bridges, parks, cemeteries, urban transport stops and buildings, e.g. offices, departments, hospitals, theaters, museums, and churches. The typology of these names and their current number in the database are shown in Tab. 1.

The collected urban proper names have various linear and syntactic features. The names consists of 1 to 23 components when counted according to the rules of Multiflex. Compare two examples: *Belweder* (the palace name) and *Biblioteka Publiczna im. Juliana Ursyna Niemcewicza w Dzielnicy Ursynów m.st. Warszawy* (the library name). The data contain not only words of POS classes which inflect, namely nouns, adjectives and numerals but also those which do not change their forms, such as prepositions and conjunctions. Each of the mentioned POS classes can be found, e.g. in the set of cultural institutions names:

(20) Kino$_{\text{N}}$ Kultura$_{\text{N}}$ (cinema),

(21) Filharmonia$_{\text{N}}$ Narodowa$_{\text{ADJ}}$ (concert hall),

(22) Muzeum$_{\text{N}}$ X$_{\text{NUM}}$ Pawilonu$_{\text{N}}$ Cytadeli$_{\text{N}}$ Warszawskiej$_{\text{ADJ}}$ (museum),

(23) Teatr$_{\text{N}}$ Na$_{\text{PREP}}$ Woli$_{\text{N}}$ (theater),

(24) Muzeum$_{\text{N}}$ Azji$_{\text{N}}$ i$_{\text{CONJ}}$ Pacyfiku$_{\text{N}}$ (museum).

In particular names also contain acronyms, abbreviations and digits which represent numbers. The significant number of those can be found in square names: *Skwer im. Grupy AK „Granat”*, *Skwer 1. Dywizji Grenadierów – Francja 1940*.

According to the data, compound proper names consists of nominal phrases based on grammatical agreement or government. We assume here that an agreement between components of a phrase occurs if all subordinate complements inflect in the same way as their head, see example 25. If only the head of a phrase undergoes inflection but the grammatical form of complements depends on it, then the relation between the head and the subordinate components of a name is defined as a government. Example 26 illustrates a government relation between the head *Aleja* and the subordinate clause *Jana Rodowicza „Anody”*. Governed phrases can be in *genitive*, *dative*, *instrumental* and *locative*.

Some multi-word proper names have specific inflection patterns rarely observed in common nominal phrases.niemożliwa to za mocno, raczej skrajnie rzadka$^{\text{JR}}$a dasz przykład (nie w artykule), bo ja nie umiałem wymyślić.$^{\text{MW}}$ If a name contains another proper name whose form is in *nominative* case, as in example 27, the inner-name remains unflected whereas the head of this phrase takes the forms of appropriate cases. The next example 28 shows that the name previously “frozen” does inflect if it is used independently (without the head *Kino*).

(25) Jan Rodowicz „Anoda” (nom.), Jan*a* Rodowicz*a* „Anod*y*” (gen.), Jan*em* Rodowicz*em* „Anod*ą*” (inst.), etc.

(26) Alej*a* Jana Rodowicza „Anody” (nom.), Alei Jana Rodowicza „Anody” (gen.), Alej*ą* Jana Rodowicza „Anody” (inst.), etc.

(27) Kin*o* Femina (nom.), Kin*a* Femina (gen.), Kin*em* Femina (inst.), etc

(28) Femin*a* (nom.), Femin*y* (gen.), Femin*ą* (inst.), etc.

Table 2 shows name distribution of the concept ROAD from the collected database. As components we count not only words but also spaces, punctuation marks, e.g. quotation marks, Arabic and Roman numerals, e.g. 1920, IX, acronyms, e.g. ZUS. The table shows that structures based on agreement are the most numerous in this group of proper names (2661), nonetheless they are mostly formed only by three components (2649). Phrases based on government are a smaller set (2354), they are formed by three to fourteen components. The variety of name subordinate structures will be reflected in visual graphs describing

Table 2: Dependency between the number of name components and frequency of the syntactic realization (agreement and government)

№ Comp.	Agreement structure	№ Agr.	Government structure	№ Gover.	Total
3	ulica Zielona	2649	Plac Zbawiciela	1002	3651
5	ulica I Poprzeczna	12	Aleja 3 Maja	986	998
6	–	0	ulica św. Teresy	12	12
7	–	0	ulica Jana Sebastiana Bacha	130	130
8	–	0	Aleja ks. Józefa Stanka	94	94
9	–	0	Rondo Ligi Morskiej i Rzecznej	85	85
10	–	0	Most gen. Stefana Grota-Roweckiego	23	23
11	–	0	Aleja Franciszka Żwirki i Stanisława Wigury	8	8
12	–	0	Ulica mjr. Henryka Dobrzańskiego „Hubala”	8	8
13	–	0	Skwer 7 Pułku Piechoty AK „Garłuch”	4	4
14	–	0	Rondo gen. Augusta Emila Fieldorfa „Nila”	2	2
		2661		2354	5015

their inflection. We assume that the number of graphs representing government relations will exceed the number of graphs for agreement relations several times.

5 Summary

We have presented the ongoing project of creating an electronic dictionary of Polish proper names. The methodological and computational prerequisites of the lexicographic work have been presented. We have developed a computing platform allowing us to describe proper names with respect to their types/subtypes, as well as their inflection and variability. *Morfeusz SGJP* manages the inflection of simple words while *Multiflex* offers a graph-based description of inflectional and syntactic variants of compound proper names. A graphical interface *Toposław* cooperating with both tools supports the lexicographic work by automating dictionary lookup, graph management, generation of inflected forms, encoding of blocs of entries, and the embedding descriptions.

We gathered almost 8,000 toponyms to be described, and performed their quantitative nad qualitative analysis. The names referring to the ROAD concept (streets, avenues, squares,. . .) are by far the most numerous: they constitute over 60% of all entries. Communication points (stops and stations) is the second largest category (24%). As far as the syntactic structure of the names is concerned they are mostly nominal phrases based on an agreement or a government structure. Over 73% of all entries contain three constituents, while about 93% of them contain up to five constituents (including separators and punctuation).

At present, systematic lexicographic work has started. The resulting linguistic resource can be used for natural language processing applications such as information extraction, dialogue systems, or geographical information systems with natural language access.

References

A. Cieślikowa, editor (2008), *Mały słownik odmiany nazw własnych*, Rytm, Warszawa.

J. Grzenia (2003), *Słownik nazw własnych*, Wydawnictwo naukowe PWN.

K. Handke (1998), *Słownik nazewnictwa Warszawy*, Slawistyczny Ośrodek Wydawniczy, Warszawa.

Krzysztof Marasek, Łukasz Brocki, Danijel Koržinek, Krzysztof Szklanny, and Ryszard Gubrynowicz (2009), User Centered Design for a Voice Portal, *Lecture Notes in Artificial Intelligence*, 5070:??–??

Małgorzta Marciniak, Joanna Rabiega-Wiśniewska, and Agnieszka Mykowiecka (2008), Proper Names in Dialogs from the Warsaw Transportation Call Center, in *Intelligent Information Systems XVI*, EXIT.

A. Mykowiecka, K. Marasek, M. Marciniak, R. Gubrynowicz, and J. Rabiega-Wiśniewska (2007), Annotation of Polish spoken dialogs in LUNA project, in *Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of 3rd Language & Technology Conference. October 5-7, 2007, Poznań, Poland.*

Sébastien Paumier (2002), Manuel d'utilisation du logiciel Unitex, http://www-igm.univ-mlv.fr/unitex/manuelunitex.ps.

J. Piskorski and M. Sydow (2007), Usability of String Distance Metrics for Name Matching Tasks in Polish, in *Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of 3rd Language & Technology Conference. October 5-7, 2007, Poznań, Poland.*

J. Piskorski, M. Sydow, and A. Kupść (2007), Lemmatization of Polish Person Names, in *ACL 2007. Proceedings of the Workshop on Balto-Slavonic Natural Language Processing 2007 Special Theme: Information Extraction and Enabling Technologies.*

Adam Przepiórkowski and Marcin Woliński (2003), A Flexemic Tagset for Polish, in *Proceedings of the Workshop on Morphological Processing of Slavic Languages, EACL 2003*, pp. 33–40.

Ewa Rzetelska-Feleszko, editor (2005), *Polskie nazwy własne*, Instytut Języka Polskiego Polskiej Akademii Nauk, Kraków.

Zygmunt Saloni, Włodzimierz Gruszczyński, Marcin Woliński, and Robert Wołosz (2007), *Słownik gramatyczny języka polskiego*, Wiedza Powszechna, Warszawa.

Agata Savary (2005a), A formalism for the computational morphology of multi-word units, *Archives of Control Sciences*, 15(3):437–449.

Agata Savary (2005b), MULTIFLEX. User's Manual and Technical Documentation. Version 1.0, Technical Report 285, LI-François Rabelais University of Tours, France.

Agata Savary, Joanna Rabiega-Wiśniewska, and Marcin Woliński (2009), Inflection of Polish Multi-Word Proper Names with Morfeusz and Multiflex, *Lecture Notes in Artificial Intelligence*, 5070:??–??

Jan Tokarski (2002), *Schematyczny indeks a tergo polskich form wyrazowych*, ed. Zygmunt Saloni, Wydawnictwo Naukowe PWN, Warszawa, 2 edition.

Marcin Woliński (2003), System znaczników morfosyntaktycznych w korpusie IPI PAN, *Polonica*, XXII–XXIII:39–55.

Marcin Woliński (2006), Morfeusz — a Practical Tool for the Morphological Analysis of Polish, in Mieczysław Kłopotek, Sławomir Wierzchoń, and Krzysztof Trojanowski, editors, *Intelligent Information Processing and Web Mining, IIS:IIPWM'06 Proceedings*, pp. 503–512, Springer.