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 a project of constructing an electronic dictionary of Polish urban proper names. The dictionary describes inflection and variants of proper names and links them with city objects. Moreover, we classify proper names according to their former, common, official, or spoken usage. We present tools which we use for the creation of the dictionary, and an analysis of Warsaw toponyms gathered for description.

Keywords: Warsaw toponyms, electronic dictionary, Polish, inflection and variability of proper names

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 the Institute of Computer Science PAS is involved. The main goal of the LUNA 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 the 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

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

subject has been undertaken from the computational point of view, see Piskorski and Sydow (2007), Piskorski *et al.* (2009). 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 the LUNA project is improving dialogue systems. The first approach to a dialogue system concerning Warsaw transportation, focused mainly 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, with particular focus on transliterated spoken variants. For example, the full official name of *ulica Bitwy Warszawskiej 1920 r.* (The Battle of Warsaw 1920 Street) is abbreviated in practice into *ulica Bitwy Warszawskiej* (The Battle of Warsaw Street). For speech recognition and generation the year should be represented by words *tysiąc dziewięćset dwudziestego roku* (the year nineteen twenty). Our dictionary ought to include all possible forms.
- Representation of former and common names of the city's objects and connection of different names with the same object, e.g., for the street *ul. ks. Jerzego Popiełuszki* (Fr. Jerzy Popiełuszko Str.) the previous name was *ul. Stołeczna* (Capital Str.).
- 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*$_{Loc}$ (Marshal Street) the lemma is *ulica*$_{Nom}$ *Marszałkowska*$_{Nom}$, while for the street *ulicy*$_{Loc}$ *Calineczki*$_{Gen}$ (Thumbelina Street) it is *ulica*$_{Nom}$ *Calineczki*$_{Gen}$.
- Generation of a desired inflected form for a given base form — function opposite to lemmatization.
- If several objects have the same name, we should be able to distinguish them. For example, a street *ulica Sportowa* (Sports Street) exists 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 requirements 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 actual or historical objects referenced by the names. Several names (former, common, etc.) may be linked with a single city object, and many objects with a single name.

A similar distinction between the conceptual and the lexical level can be found in *Prolexbase*, Krstev *et al.* (2005) and Tran and Maurel (2006). It contains a language-independent ontology consisting of: (i) 4 supertypes (anthroponyms, ergonyms, pragmonyms, toponymes), (ii) 30 types (association, celebrity, catastrophe, product, region, etc.), (iii) a large number of objects called *conceptual proper names*. Objects are associated with *points of view* which reflect the variability of a proper name in time, usage and social status. Relations occurring between conceptual proper names and/or types include meronymy (*France* vs.

Europe), synonymy (*Zaire* vs. *Democratic Republic of the Congo*), hyperonymy (*Tchernobyl* vs. *catastrophe*) and accessibility (*Bangkok – capital of – Thailand*). The language-dependent hierarchy attached to this ontology contains lexemes (e.g. *United Nations*), their components (e.g. *Nation*), aliases (e.g. *UN*) and derivatives, all of them decorated by links to external inflection paradigms and some language-dependent relations. The implementation of this rich model contains 55,000 French prolexemes and 54,000 relations. Toponyms are numerous (almost 50,000) but only 14 of them refer to city objects (streets). A Serbian and a Korean module contain 606 and 113 prolexemes, respectively.

In Japanese, Sekine (2008) developed a fine-grained 3-level *Extended Named Entity* hierarchy of about 200 categories and their accompanying attributes (e.g. *source location*, *outflow* and *length* for a *river*). City-related categories include postal addresses, facilities (buildings, parks, stations, airports, etc.), and lines (roads, canals, bridges, etc.). No figures are given as to either the size of the vocabulary classified, nor its morphological properties.

In Polish, Abramowicz *et al.* (2006) present a geo-referencing cadastral system, with a three-level hierarchy of several dozen supertypes, types and subtypes. The database contains 164,000 Polish and 19,000 English names, most of which are cities, counties, countries and given names. Their morphological variants are processed by semi-automatic ad hoc methods. Urban objects such as routes, transportation points, facilities, etc. are mentioned in the hierarchy of types but only a small number of them appears in the data (less than 100). Street names and organizations are considered as inaccurate for explicit listing and instead are covered by recursively embedded grammars. Some coverage gain is obtained by string edit-distance methods. A graphical application allows the user to navigate within the hierarchy and view occurrences of the selected names in cadastral corpora, whereas anaphora resolution is used when the same name may refer to different objects.

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 Pragmatic and Stylistic Phenomena

In this section we present pragmatic and stylistic 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 already been presented by Savary *et al.* (2009). Here, we focus on how most of these features are reflected in the lexicon. Some previously unsolved problems, such as derivational and semantic variants, are solved through stylistic links introduced in the database. Moreover the hierarchy of concepts, the qualitative and quantitative analysis of the data, and an encoding interface enhance the descriptive framework created earlier.

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 the name and a concept 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, palaces, monuments, commercial centres, and business establishments, see examples (1-3).

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

(2) Pomnik Stefana Starzyńskiego (Monument of Stefan Starzyński)

(3) Urząd Miasta Stołecznego Warszawy (Town Hall of the Capital City of Warsaw)

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, see examples (4-6).

(4) Stefan Starzyński

(5) Jan Rodowicz „Anoda"

(6) Król Jan I Olbracht (King Jan Olbracht the First)

2.1 Proper Names Variants

In the dictionary we represent several variants of names. We mark three of them which are pragmatically important for potential applications.

- the "official" name variant represents 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)-(10). 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 (Stefan Starzyński Nature Reserve of the Kabaty Forest),
neutral/neutral spoken name: Las Kabacki (Kabaty Forest)

(8) official name: Aleja Jana Rodowicza „Anody" (Jan Rodowicz "Anoda" Avenue),
neutral/neutral spoken name: Aleja Rodowicza (Rodowicz Avenue)

(9) official name: Szkoła Podstawowa nr 211 im. Korczaka w Warszawie (Korczak Primary School no. 211 in Warsaw),
neutral name: Szkoła Podstawowa nr 211 (Primary School no. 211),
neutral spoken: Szkoła Podstawowa numer dwieście jedenaście (Primary School number one hundred and eleven)

(10) official/neutral name: Plac 1831 r. (The Year 1831 Square),
neutral spoken: Plac Tysiąc Osiemset Trzydziestego Pierwszego Roku (The Year Eighteen Thirty One Square)

We recognize orthographic variation of proper names which usually concerns punctuation marks, such as a hyphen or a quotation mark. Example (11) shows three popular ways of writing a name of the Polish general Rowecki inside a name of a street. Variants refer to his pseudonym *„Grot"*.

(11) 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) component is omitted, especially in spoken language. We accept a shortened name if it is usually recognized by city residents, see example (12). As an order variant we accept such a change of component order that is consistent with Polish grammar or stylistics rules. Example (13) shows two correct positions of the pseudonym *„Anoda"* according to Polish rules of usage.

(12) official name: Pałac Kultury i Nauki (Palace of Culture and Science),
shortened variant: Pałac Kultury (Palace of Culture)

(13) 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 20^{th} century a lot of streets 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 (14), 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 (15). At present some places are also renamed, as in example (16). In those cases, if we consider a proper name which is no longer on the official list, it is labelled as "former" (cf. the diachronic viewpoint in Tran and Maurel (2006)).

(14) ul. Tadeusza Hołówki (Tadeusz Hołówko Str., 1933-1951, and since 1991)[2], former names: ul. Karpia (Carp Str.), 1951-1958), ulica Wczasowa (Vacation Str., 1958-1991)

(15) Aleja Solidarności (Solidarity Avenue),
former name: Aleja Karola Świerczewskiego (Karol Świerczewski Avenue)

(16) Aleja Jana Rodowicza „Anody”,
a former part of the street: ul. Jana Rosoła (Jan Rosół Str.)

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 names usually bring some associations of place shape, colour or physical feature, as in example (17). Sometimes, residents give an ironic name to the place or object, see example (18), which expresses their disapproval or funny association connected to it.

(17) official name: Pomnik Bohaterów Warszawy 1939-45 (Monument of Warsaw Heroes 1939-45),
common name: Warszawska Nike (Warsaw Nike)

(18) official name: Pomnik Józefa Piłsudskiego (Monument of Józef Piłsudski),
marked name: Dziadek Parkingowy (Parking Grandpa)

2.3 Hierarchy of Concepts

Our dictionary enables the introduction of some semantic information about represented objects. This information is not crucial to conducting dialogues in the city transportation call center, so we introduced a very simple hierarchy of concepts. It can be developed or substituted by a more sophisticated hierarchy, see projects described in Teller *et al.* (2007). In the dictionary, all city objects are connected to exactly one concept represented by leafs of a hierarchy presented below.

- 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* (Upper and Lower 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.

[2] According to Handke (1998).

  - 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* (Central Station)
    * AIRPORT: e.g.: *Port Lotniczy im. Fryderyka Chopina w Warszawie* (Frederic Chopin Airport Warsaw)
  - 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* (Drama Theatre) 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* (Vistula River), *Jeziorko Czerniakowskie* (Czerniaków Lake)
  - MONUMENT: *Pomnik Adama Mickiewicza* (Adam Mickiewicz Monument)
- 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 in *Multiflex*, which in turn utilises *Morfeusz* for inflecting components of names.

3.1 Morfeusz

Morfeusz SGJP is a morphological analyser and generator 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”*:

(19)	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 (Savary (2005a), Savary (2005b)) is a cross-language morpho-syntactic generator of multi-word units (MWUs). It allows us to exhaustively and precisely describe the inflection paradigm and variation of a MWU via a ‘two-tier approach’. First, an underlying morphological module such as *Morfeusz* allows us to tokenize the MWU lemma, to annotate its components, and to generate infected 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 spaces and quotes. Each token which can be inflected, is annotated by a *Morfeusz* tag (cf section 3.1). The inflection of the whole compound is described by a graph containing: (i) inflection operators to be applied to constituents (inside boxes), e.g. $\langle \$1 : Case = \$c \rangle$, (ii) features of the generated forms (under boxes), e.g. $\langle Gen = \$1.Gen; Nb = \$1.Nb; Case = \$c \rangle$. Unification variables (here $\$c$) allow to express inflection and agreement. Embedded compounding is easily expressed by delimiting and describing substructures, e.g. the official name in examples (8) and (13) is seen as a compound containing three components, the third of which is a compound itself (*Jana Rodowicza „Anody”*) described as on figure 1.

The full exploration of a graph results in the generation of all inflected forms and variants of the compound, in particular all the variants in example (20) inflected for all cases[3]. Note that numbering of the constituents allows the graph to represent their omissions, insertions and order change. See Savary *et al.* (2009) for a more detailed discussion on using *Multiflex* with *Morfeusz* and their application to the study of Polish urban toponyms.

(20) Jan Rodowicz „Anoda", Jan Rodowicz Anoda,
Jan „Anoda" Rodowicz, Jan Anoda Rodowicz
Jan Rodowicz, Rodowicz „Anoda", „Anoda" Rodowicz, Rodowicz

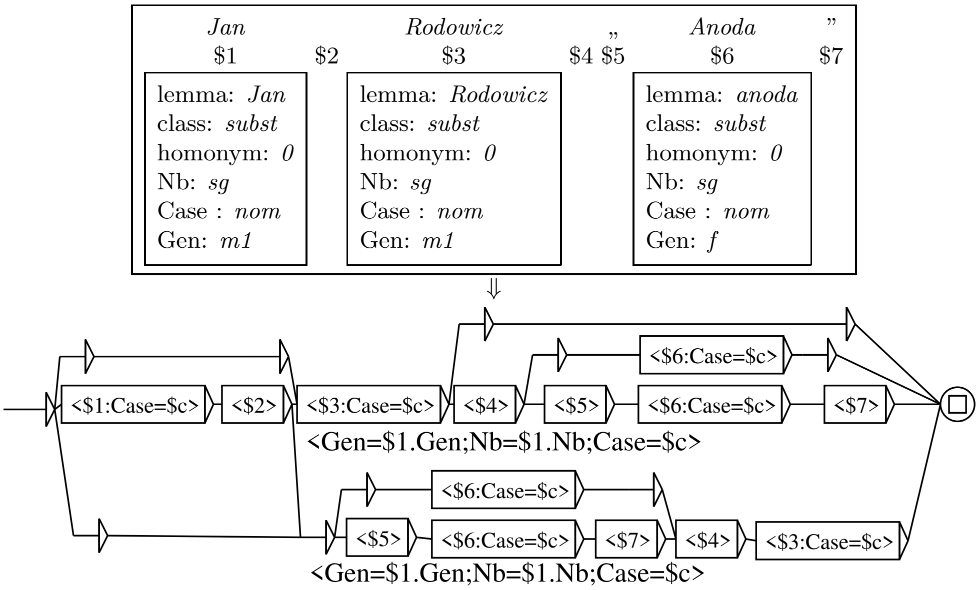

FIGURE 1: Lemma annotation and inflection graph for the patronym *Jan Rodowicz „Anoda"* containing elliptical variants and inversions

3.3 A Dictionary Editor *Toposław*

The creation of high-quality linguistic resources, particularly for morphologically rich languages, is labor-intensive. It calls for ergonomic tools in which encoding may be performed in a semi-automatic way. Krstev *et al.* (2006) describe a workstation in which heterogenous resources such as lexicons for single and compounds words, aligned corpora, and wordnets, can be looked-up, developed and maintained within a unique framework. It uses *Multiflex* and *Unitex* modules (Paumier (2002)), and it proved efficient for Serbian, whose complexity on the

[3]For the sake of simplicity of this presentation the graph discussed here is shown in its simplified version containing no pragmatic features "official", "neutral" and "neutral spoken".

morphological level is comparable to Polish. Tran *et al.* (2005) present a collaborative Web interface for *Prolexbase* described in section 1. Its important features are processing data in blocks, and importing external table-based descriptions.

To facilitate the creation of our dictionary a Java application named *Toposław* was developed. 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 compound 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 a city object. The links can be labelled with the type of the name: "common", "marked", or "former". City objects are also categorised using the hierarchy described in section 2.3.

4 Description of the Collected Data

The collected proper names come from the city hall, offices and websites, web pages of *Zarząd Dróg Miejskich* (Town Roads Authority) and directly from *Biuro Geodezji i Katastru* (Geodesic and Cadastral Office).

At present the database of city proper names includes 7898 records. The typology of these names and their current frequency (according to the classification presented in section 2.3) is shown in Tab. 1.

The collected urban proper names have various linear and syntactic features. The names consist of 1 to 25 components when counted according to the rules of *Multiflex* (not only words but also all dots, spaces and punctuation marks are counted as a component). Compare two examples: *Belweder* (the palace name)

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	4999
streets, boulevards, avenues	4857
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	7898

and *Biblioteka Publiczna im. Juliana Ursyna Niemcewicza w Dzielnicy Ursynów m.st. Warszawy* (Julian Ursyn Niemcewicz Public Library of the Ursynów District in the Capital City of Warsaw). 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:

(21) Kino_{N} $\text{Kultura}_{\text{N}}$ (“Culture” Cinema),

(22) $\text{Filharmonia}_{\text{N}}$ $\text{Narodowa}_{\text{ADJ}}$ (National Filharmony),

(23) Muzeum_{N} X_{NUM} $\text{Pawilonu}_{\text{N}}$ $\text{Cytadeli}_{\text{N}}$ $\text{Warszawskiej}_{\text{ADJ}}$ (Museum of the 10th Pavilion of the Warsaw Citadel),

(24) Teatr_{N} Na_{PREP} Woli_{N} (“In Wola” Theater),

(25) Muzeum_{N} Azji_{N} i_{CONJ} $\text{Pacyfiku}_{\text{N}}$ (Asia and Pacific 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”* (Square of the “Grenade” Group of the Interior Army), *Skwer 1. Dywizji Grenadierów – Francja 1940* (Square of the 1st Grenadiers Division – France 1940).

According to the data, compound proper names consist 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 (26). 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 (27) illustrates a government relation between the head *Aleja* (Avenue) 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. If a name contains another proper name whose form is in *nominative* case, as in example (28), the inner-name remains uninflected whereas the head of this phrase takes the forms of appropriate cases. The next example (29) shows that the name previously "frozen" does inflect if it is used independently (without the head *Kino*).

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

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

(28) Kin*o* Femina (nom.), Kin*a* Femina (gen.), Kin*em* Femina (inst.), etc. ("Femina" Cinema)

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

Table 2 shows name distribution of the concept ROAD (streets, avenues,...) from the collected database[4]. 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 3 to 14 components. The variety of name subordinate structures will be reflected in visual graphs describing 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 blocks of entries, and the embedding descriptions.

We gathered almost 8,000 toponyms to be described, and performed their quantitative and qualitative analysis. The names referring to the ROAD concept are by far the most numerous: they constitute over 60% of all entries. COMMUNICATION POINT is the second largest category (24%). As far as the syntactic

[4]Translations of the examples from top to bottom and from left to right: *Green Str., 1st Crosswise Str., Saviour's Sq., 3rd May Ave., St Theresa Str., Johann Sebastian Bach Str., Fr. Józef Stanek Ave., Sea and River League Roundabout, Gen. Stefan Grot-Rowecki Bridge, Franciszek Żwirko and Stanisław Wigura Str., Maj. Henryk Dobrzański „Hubal" Str., Interior Army's 7th Regiment „Garłuch" Sq., Gen. August Emil Fieldorf "Nil" Roundabout*

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	1001	3650
5	ulica I Poprzeczna	12	Aleja 3 Maja	979	991
6	–	0	ulica św. Teresy	12	12
7	–	0	ulica Jana Sebastiana Bacha	129	129
8	–	0	Aleja ks. Józefa Stanka	93	93
9	–	0	Rondo Ligi Morskiej i Rzecznej	84	84
10	–	0	Most gen. Stefana Grota-Roweckiego	21	21
11	–	0	Aleja Franciszka Żwirki i Stanisława Wigury	7	7
12	–	0	Ulica mjr. Henryka Dobrzańskiego „Hubala”	7	7
13	–	0	Skwer 7 Pułku Piechoty AK „Garłuch”	3	3
14	–	0	Rondo gen. Augusta Emila Fieldorfa „Nila”	2	2
		2661		2338	4999

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

Witold ABRAMOWICZ, Agata FILIPOWSKA, Jakub PISKORSKI, Krzysztof WĘCEL, and Karol WIELOCH (2006), Linguistic Suite for Polish Cadastral System, in *Proceedings of LREC'06, Genoa, Italy*, pp. 2518–2523.

Aleksandra CIEŚLIKOWA, editor (2008), *Mały słownik odmiany nazw własnych*, Rytm, Warszawa.

Jan GRZENIA (2003), *Słownik nazw własnych*, Wydawnictwo naukowe PWN.

Kwiryna HANDKE (1998), *Słownik nazewnictwa Warszawy*, Slawistyczny Ośrodek Wydawniczy, Warszawa.

Cvetana KRSTEV, Ranka STANKOVIĆ, Duško VITAS, and Ivan OBRADOVIĆ (2006), Workstation for Lexical Resources - WS4LR, in *Proceedings of LREC'06*, pp. 1692–1697.

Cvetana KRSTEV, Duško VITAS, Denis MAUREL, and Mickaël TRAN (2005), Multilingual Ontology of Proper Names, in *Proceedings of LTC-05, Poznań, Poland*.

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.

Agnieszka MYKOWIECKA, Krzysztof MARASEK, Małgorzata MARCINIAK, Ryszard GUBRYNOWICZ, and Joanna RABIEGA-WIŚNIEWSKA (2007), Annotation of Polish spoken dialogs in LUNA project, in *Proceedings of LTC-07, Poznań, Poland*.

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

Jakub PISKORSKI and Marcin SYDOW (2007), Usability of String Distance Metrics for Name Matching Tasks in Polish, in *Proceedings of LTC-07, Poznań, Poland*.

Jakub PISKORSKI, Marcin SYDOW, and Karol WIELOCH (2009), On knowledge-poor methods for person name matching and lemmatization for high inflection languages, *Information Retrieval*, 12:275–299.

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.

Satoshi SEKINE (2008), Extended Named Entity Ontology with Attribute Information, in *Proceedings of LREC'08*.

Jacques TELLER, John LEE, and Catherine ROUSSEY (2007), *Ontologies for Urban Development*, Springer.

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

Mickael TRAN and Denis MAUREL (2006), Prolexbase: Un dictionnaire relationnel multilingue de noms propres, *Traitement Automatiques des Langues*, 47(3):115–139.

Mickael TRAN, Denis MAUREL, Duško VITAS, and Cvetana KRSTEV (2005), A French-Serbian Web Collaborative Work on a Multilingual Dictionary of Proper Names, in *Workshop on Multilingual Lexical Databases, Chiang Rai, Thailand*, pp. 67–71.

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.