Table of Contents

Inflection of Polish Multi-Word Proper Names with Morfeusz and Multiflex

Inflection of Polish Multi-Word Proper Names with Morfeusz and Multiflex

Agata Savary[1], Joanna Rabiega-Wiśniewska[2], and Marcin Woliński[2]

[1] Université François Rabelais de Tours, France,
agata.savary@univ-tours.fr,
[2] Institute of Computer Science, Polish Academy of Sciences, Poland,
{joanna.rabiega,marcin.wolinski}@ipipan.waw.pl

Abstract. We discuss morphological properties of Polish multi-word proper names. We present a cooperating framework of two morphological tools: *Morfeusz*, a morphological analyser and generator for Polish simple words, and *Multiflex*, a cross-language morpho-syntactic generator of multi-word units. We discuss interface constraints required for the interoperability of these tools, and we show how the resulting platform allows one to describe the morpho-syntactic behaviour of some interesting examples of Warsaw multi-word toponyms.

Key words: computational morphology, Polish proper names, multi-word units, inflection and variability of compounds

1 Introduction

Proper names and other named entities are of crucial quantitative and qualitative importance for natural language processing (NLP) due to their frequent occurrence in corpora and their rich semantic content. Despite continual efforts in the NLP community aiming at named entity extraction and modelling, the formal linguistic description of proper names remains a challenge.

In many application domains, such as technical sublanguages, most of the proper names are multi-word units. As in the case of single words they are subject to inflection. In highly inflected languages, such as Polish, each multi-word nominal typically yields seven or fourteen inflected forms (depending on whether it does or does not inflect for number), whose behaviour may be idiosyncratic. Moreover, even if compound proper names have the reputation of institutionalized phrases, they often show a varying degree of flexibility on the orthographic, lexical, syntactic, and/or semantic level. For example, they may allow for some component inflections, omissions, substitutions, insertions, and embedding. This results in a number of their possible textual realisations which should possibly be recognized as variants of the same base form. For instance, it should be indicated that all of the following sequences, along with their corresponding inflected forms, refer to the same roundabout in Warsaw: *Rondo Zgrupowania AK "Radosław", Rondo Zgrupowania "Radosław", Rondo Radosława*, and *Rondo Babka*.

We are interested in the morphological properties of Polish multi-word proper names in general. Our approach is lexical: we aim at explicit description of all grammatically correct inflected forms of a multi-word proper name and their variants.

In this paper we present a cooperating framework of two morphological tools: *Morfeusz*, a morphological analyser and generator for Polish simple words, and *Multiflex*, a cross-language morpho-syntactic generator of multi-word units. We discuss interface requirements for the interoperability of these tools. The resulting platform is characterized by: a rich flexemic tagset of Polish, a graph-based description of multi-word paradigms, and compact inflection patterns due to unification. We show how the platform allows one to describe the morpho-syntactic behaviour of some interesting examples of Warsaw multi-word toponyms.

Finally, we address some remaining problems such as:

- conflation of derivational and semantic variants,
- the necessity of distinguishing variants possessing the same inflection features,
- "extreme" elliptical variants transforming multi-word units into inflectionally intriguing single words (e.g. *ulica Kazimerza Pułaskiego* > *Pułaskiego* in all cases), and
- morpho-graphical problems caused by abbreviations and acronyms.

The results presented in this paper are a part of a larger project aiming at modelling Warsaw topography (streets, places, institutions, monuments, etc.) and transportation system (bus- and tram-stops, underground stations, etc.), as well as a large-scale lexical description of Warsaw toponyms, in view of information extraction and dialogue systems. We also wish to examine the feasibility of integrating Polish toponyms within a multilingual ontology of proper names, considering machine translation, multilingual information extraction and text alignment.

2 State of the art

Onomastic studies on the origin, history and regional particularities of Polish proper names have been performed for decades [1, 2]. However, relatively few efforts have been made in view of a large-scale formal description of their grammatical, in particular inflectional, behaviour. The printed dictionary [3] is dedicated to spelling, pronunciation, formation and inflection of proper names in Polish. Among its 11,000 entries, numerous foreign names are described with respect to their declension, which copes with inflectional suffixes atypical for Polish inflection paradigms. The lexicon [4] is dedicated to formal and systematic description of inflection paradigms of single and compound proper names in Polish. The study aims to describe almost all paradigms that serve inflection of the most important categories of proper names: person names (Polish and loan-words), geographical names (Polish and loan-words) and names of significant chains of shopping centres and supermarkets in Poland. At the basis of this research, a database of about 4,500 proper names and their inflection patterns was created, which, according to our knowledge, is not freely available at the moment. The Polish electronic inflectional dictionary [5], used by *Morfeusz* in our study, provides full inflectional paradigms for almost 245,000 Polish uni-word lemmas. It contains about 3,000 first names and 2,000 most popular Polish family names, as well as a sufficient number of toponyms and relational adjectives stemming from proper names.

In the context of natural language processing the quantitative and qualitative importance of proper names and other named entities is unquestionable. They appear in up to 83% of user queries on the Web [6]. They are semantically rich and prove good clues for NLP applications such as document classification [7, 8] or information retrieval [9, 10]. With MUC [11] and CoNLL [12] campaigns, named entity extraction has become one of the crucial challenges in the NLP community.

Polish initially remained a less studied language within these mainstream efforts. However, some interesting issues have been investigated recently within the Polish NLP community. Automatic lemmatization of proper names using a string distance metric is proposed in [13] and [14]. In [15] this method is applied to automatic recognition of proper names extracted from the corpus of real-life dialogues. The same paper also presents a method of automatic recognition of proper names during a rule-based semantic annotation of dialogue texts with concept names. This approach has been proposed within the European project LUNA (`http://ist-luna.eu`). One of its focal issues is the behaviour of proper names in Polish spoken dialogues. The syntactic behaviour of names, in particular their variability, has been shown in [16]. It constitutes the main obstacle for direct application of the existing official lists of proper names to their identification in texts. This conclusion is one of the main motivations for the work presented below.

One of the interesting particularities of Slavic languages, such as Polish, is their rich inflectional morphology. As argued in [17], taking them into account in designing language models enhances the universality of those models. Thus, for instance, a multilingual lexical model for proper names proposed in [18] is based in particular on studies of the morphological complexity of Serbian, Polish and Bulgarian [19, 20].

Compound proper names belong to a larger class of multi-word units (MWUs). As shown in [21] and [22], the linguistic (orthographic, morphological, syntactic and semantic) variability of MWUs is their essential property in natural language corpora. Moreover, their morpho-syntactic non-compositionality and idiosyncrasy [23] calls for lexicalized models in which compound units are listed and explicitly described. In [17], a comparative study of eleven such approaches (e.g. [24, 21, 25, 26] and [27]) in seven languages was performed. We can expect the *Multiflex* formalism assumed in this paper to be sufficiently expressive for a large-scale description of inflection and variation of compounds.

As shown below, the prerequisite for automatic processing of compound proper names is the morphological treatment of their components. These contain both one-word proper names and common words. The existing morphological analyzers and generators for Polish single words use stochastic or rule-based methods [28]. *Morfeusz* belongs to the latter category.

3 Morpho-syntactic behaviour of Polish multi-word proper names

3.1 The data

We limit the scope of our research to Polish 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, business establishments.

Administrative units. The city of Warsaw is divided into eighteen districts. The names of the districts consist of one word, however there are two exceptions:

(1) *Praga Północ* 'North Praga', *Praga Południe* 'South Praga'

Although the district names describe wide-spread areas, they may appear in names of bus- and tram-stops:

(2) *przystanek Bielany, przystanek Śródmieście, przystanek Ursynów*
'Bielany stop, Śródmieście stop, Ursynów stop'

The districts are further subdivided into smaller areas such as housing estates. Those are usually given traditional Polish one-word or multi-word names (3), however recently foreign names have been given to some modern residential quarters as in (4).

(3) *Rakowiec* 'Rakowiec quarter', *Za Żelazną Bramą* 'Behind the Iron Gate'

(4) *Villa Nova*

Traffic routes. The major part of the lexicon of Warsaw transportation system refers to traffic routes. There are specific lexemes which belong to these types of names, such as: *ulica* 'street', *Aleja* 'avenue', *Rondo* 'roundabout', *Plac* 'square, plaza', *Skwer* 'square', *Trasa* 'route, line', *Trakt* 'route, way', *Most* 'bridge', and *Wał* 'wall'. Only the first one – *ulica* – is not capitalised in proper nouns and is usually omitted in spoken language:

(5) *ulica Bracka, Bracka* 'Bracka Street'

(6) *Aleja Zjednoczenia, *Zjednoczenia* 'Union Avenue'

The routes are named after historical and fictional persons, historical events, fauna, flora and other places or names of different types (e.g. the names of military units). These embedded names may each consist of one or more words, as in:

(7) *ulica* ⟨*Fryderyka Chopina*⟩ 'Frederic Chopin Street'

(8) *Aleja* ⟨*Na Skrapie*⟩ 'At Scarp Avenue'

(9) *Rondo* ⟨*Generała Charlesa de Gaulle'a*⟩ 'General Charles de Gaulle Roundabout'

(10) *Plac* ⟨*J. W. Wilsona*⟩ 'J. W. Wilson Square'

(11) *Trasa* ⟨*Toruńska*⟩ 'Toruń Way', Toruń is the name of a city

(12) *Most* ⟨*Siekierkowski*⟩ 'Siekierki Bridge', Siekierki is the name of a historical district

Stopping places. Stopping places in the public transportation system, such as bus- and tram-stops, underground and railway stations, and airports, are particularly significant from both quantitative and qualitative points of view, in the context of applications such as information extraction and dialogue systems (cf. section 7). The key-lexemes attributed to this type contain: *Dworzec* 'Station', *przystanek* 'stop' and *Port Lotniczy* 'Airport', as for instance in:

(13) *Dworzec Centralny* 'Central Station'

(14) *Dworzec PKS Warszawa Zachodnia* 'West-Warsaw Bus Station'

(15) *Port Lotniczy im. F. Chopina* 'Frederic Chopin Warsaw Airport'

Most names of stops have three kinds of origin. They can:

- be given after the names of the streets they are located at, as in example (16)
- refer to public or private institutions they are close to, as in example (17)
- combine a name of a street and a name of an institution, as in example (18)

(16) *przystanek Nowy Świat* 'New World bus-stop'

(17) *przystanek CH Arkadia* 'Arkadia Shopping Center bus-stop'

(18) *przystanek Płocka-Szpital* 'Płocka Str.-Hospital bus-stop'

Other places and objects. Other names which we consider in our study describe places in the city. Our analysis includes urban places such as: names of parks and gardens (19), cemeteries (20), public institutions and facilities (21), mansions (22), monuments (23), commercial centres (24) and business establishments. For the last ten years a lot of new financial and business centers have been built in Warsaw. Some of them were given foreign — almost exclusively English – names as in example (25).

(19) *Park Skaryszewski im. I. J. Paderewskiego* 'Skaryszew Park of I. J. Paderewski'
Pola Mokotowskie 'Mokotów Fields'

(20) *Komunalny Cmentarz Północny* 'North Municipal Cemetery'
Muzułmański Cmentarz Tatarski 'Muslim Tatar Cemetery'

(21) *Muzeum Azji i Pacyfiku* 'Museum of Asia and Pacific'
Mazowiecki Teatr Muzyczny Operetka Mazovia 'Music Theater Operetta'

(22) *Pałac na Wodzie* ‘Palace on the Isle’
Zamek Królewski ‘Royal Castle’

(23) *Pomnik Bitwy o Monte Cassino* ‘Battle of Monte Cassino Monument’
Pomnik Braterstwa Broni ‘Monument to Brotherhood in Arms’
Kolumna Zygmunta III Wazy ‘Column of Sigismund III Vasa’

(24) *Galeria Mokotów* ‘Mokotów Gallery’
Złote Tarasy ‘Golden Terraces’

(25) *Blue Point* (office building)
Millenium Plaza (banking center)

3.2 Linguistic analysis of the data

Major proper names have been listed and described for decades within traditional dictionaries and encyclopedias meant for human readers. These data implicitly assume the reader’s linguistic and extra-linguistic knowledge. Thus, they are rarely applicable as such to automatic processing of corpora. Construction of machine-usable lexicons requires a formal, disciplined approach in which all relevant properties are explicitly described.

In this section we present an overview of the morpho-syntactic features of urban proper names, as well as usage manners of Polish speakers with respect to these names. Our study is based on the spoken language corpus resulting from the LUNA project (cf. section 2), as well as on our own Warsaw native speakers’ life-long experience.

Abbreviations. Polish proper names frequently contain abbreviations, as in examples (26) through (31). They are written in upper or lower case. According to Polish spelling dictionaries the letter case may determine the precise meaning of the abbreviation, e.g. in (26) and (27) it marks the singular or the plural number of ‘avenue’.

(26) *al. Lotników* ‘Avenue of Pilots’, *al.* = *aleja* ’avenue’ (sing.)

(27) *Al. Jerozolimskie* ‘Avenue of Jerusalem’, *Al.* = *Aleje*, ’avenue’ (pl.)

(28) *ul. I. J. Paderewskiego* ‘I. J. Paderewski Street’, *ul.* = *ulica* ’street’, *I. J.* = initials for *Ignacy Jan*

(29) *Port Lotniczy im. F. Chopina* ‘Frederic Chopin Warsaw Airport’, *im.* = *imienia* ’dedicated to’

(30) *Rondo Gen. Charles’a De Gaulle’a* ‘General Charles De Gaulle Roundabout, *Gen.* = *Generała* ’general’ (gen.)

(31) *ul. Prof. Janusza Groszkowskiego* ‘Professor Janusz Groszkowski Street’, *Prof.* = *Profesora* ’professor’ (gen.)

Acronyms and initialisms. Acronyms and initialisms are abbreviations formed from initials of the given proper name's constituents. To distinguish between the two terms, we restrict acronyms to pronounceable words formed from the initial letters or syllables, see examples (32)-(34). If each letter is pronounced individually, we speak of initialisms, as in examples (35) and (36). Foreign acronyms and initialisms are sometimes pronounced according to rules of the language of their origin, as in example (37). There are also acronym-initialism hybrids (38).

(32) *ZUS* /zus/ = ***Z****akład* ***U****bezpieczeń* ***S****połecznych* 'Social Security Office'

(33) *Polfa* /'polfa/ = ***Pol****ska* ***Fa****rmacja* 'Polish Pharmacy'

(34) *Polmos* /'polmos/ = ***Pol****ski* ***Mo****nopol* ***S****pirytusowy* 'Polish Spirits Monopoly'

(35) *PKP* /peka'pe/ = ***P****olskie* ***K****oleje* ***P****aństwowe* 'Polish State Railways'

(36) *ONZ* /oen'zet/ = ***O****rganizacja* ***N****arodów* ***Z****jednoczonych* 'United Nations Organization'

(37) *BBC* /bibi'si/ = **B**ritish **B**roadcasting **C**orporation

(38) *SGPiS* /esgie'pis/ = ***S****zkoła* ***G****łówna* ***P****lanowania* ***i*** ***S****tatystyki* 'Main School of Planning and Statistics'

The distinction between these types of abbreviations is relevant to inflection. The majority of acronyms and acronym-initialism hybrids in Polish inflect according to inanimate nominal patterns, see example (39). Initialisms may or may not inflect, depending on how well their ultima fits into the system of nominal morphological endings in Polish. For instance, the ultima of example (35) is /-pe/, which is a rather unusual Polish nominal ending, thus the whole initialism remains uninflected as presented in example (40). On the contrary, the ultima of example (36) is /-et/, which is common for many masculine inanimate nouns, e.g. *bilet* 'ticket'. Therefore, it undergoes inflection according to the same pattern, see example (41).

(39) *ZUS* (nom.), *ZUS-u* (gen.), *ZUS-owi* (dat.), etc.

(40) *PKP* (nom., gen., dat., etc.)

(41) *ONZ* (nom.), *ONZ-u* (gen.), *ONZ-owi* (dat.), etc.
bilet (nom.), *biletu* (gen.), *biletowi* (dat.), etc.

Numerals. Numbers in Polish proper names, such as shown in examples (42) through (45), are usually represented with Arabic or Roman digits. Thus, a particular description is needed in order to link digital representations of numbers with the numerals they refer to.

Morphologically, numbers are divided into cardinal and ordinal numerals. These two classes show essential differences in morphological and syntactic behaviour in Polish. Cardinal numbers constitute a morphological class on their own, with rather complex morpho-syntactic properties as discussed in [29], [30] and [31]. Ordinal numbers show adjectival inflection, and there are no morphological hints allowing their distinction from regular adjectives.

(42) *ul. II Armii Wojska Polskiego* 'The Second Polish Army Street'

(43) *ul. IX Poprzeczna* 'The Ninth Transversal Street'

(44) *ul. 36 Pułku Piechoty Legii Akademickiej* 'The Thirty Sixth Regiment of Academic Infantry Street'

(45) *ul. Posag 7 Panien* 'Dowry of Seven Maids Str.'

Synonyms and variations. Similarly to other multi-word units, compound proper names are subject to inflectional, syntactic and semantic variation. In this sense synonyms are defined here as different textual sequences referring to the same name.

Inflectional variants rely most often on inflectional patterns of their constituents, as well as on syntagmatic agreement and government rules. For instance, in order to generate all possible inflected forms of example (15) we need to be able to inflect its two first lexemes *port* and *lotniczy*, and to know that they agree in gender, number and case, while the other constituents do not vary:

(46) *Port**u** Lotnicz**ego** im. F. Chopina, Porci**e** Lotnicz**ym** im. F. Chopina*, etc. (gen., loc., etc.)

Syntactic variants result most often from ellipses. It is due to the fact that full official names may be rather long and impractical in everyday language. Components that are most frequently omitted are:

- the *ulica* 'street' keyword, systematically missing in spoken language, as in example (47).
- "non-essential" complements, as in examples (48) and (49). Proper understanding of such shortnames implies one's extra-linguistic knowledge of the city.
- first names, as in example (50).
- personal titles such as *generał*, *profesor*, etc., as in example (51).

(47) *ulica Chmielna, Chmielna* 'Chmielna Str.'
ulica Humanistów, Humanistów 'Humanistów Str.'

(48) *ulica Bitwy Warszawskiej 1920 r.* 'Warsaw Battle 1920 Str.'
ulica Bitwy Warszawskiej 'Warsaw Battle Str.'
ulica Bitwy 'Battle Str.'

(49) *Rondo Zgrupowania AK "Radosław"* '*Radosław* Division of the Domestic Army Roundabout'
Rondo Zgrupowania "Radosław" '*Radosław* Division Roundabout'
Rondo Radosława 'Radosław's Roundabout'

(50) *ulica Władysława Broniewskiego* 'Władysław Broniewski Str.'
ulica Broniewskiego
Broniewskiego

(51) *ulica Gen. Antoniego Józefa Madalińskiego* 'General A. J. Madaliński Str.'
ulica Antoniego Józefa Madalińskiego
ulica Antoniego Madalińskiego
ulica Madalińskiego
Madalińskiego

The phenomenon of systematic *semantic variation* can be presently observed within Warsaw urban toponyms. In the early 90's many names of places referring to the previous socialist period of the Polish history have been eliminated. Although the old names frequently suffer from negative connotations, they continue to be used, in particular by elderly people who have problems with remembering their recent substitutes. Thus, many diachronic synonyms co-exist, as in:

(52) *Al. Jana Pawła II* 'John-Paul II Avenue'
previously: *Al. Juliana Marchlewskiego* 'Julian Marchlewski Avenue'

(53) *Al. Solidarności* 'Avenue of Solidarity'
previously: *Al. Karola Świerczewskiego* 'Karol Świerczewski Avenue'

In present days renaming places in Warsaw continues, even if the old names bear no negative connotation. Warsaw residents prove reluctant to integrate such new names into their every day vocabulary. For instance, in example (54) the former – official – name is less frequently used than the later – traditional one. In (55) the former name refers to the street whose part has been recently given the later name.

(54) *Rondo Zgrupowania AK "Radosław"* 'Radosław Division of the Domestic Army Roundabout'
previously: *Rondo Babka* 'Babka Roundabout'

(55) *ul. Jana Rosoła* 'Jan Rosół Str.'
al. Jana "Anody" Rodowicza 'Jan "Anoda" Rodowicz Avenue'

Warsaw citizens sometimes invent their own popular names for places or buildings. It so happens that those informal names are better known than the official ones. The possible sources of such semantic synonymy include the following phenomena:

- If the official name is a periphrasis, it can be substituted by its base word if no confusion arises, as in example (56).
- If the shape or physical feature of a monument or building brings some associations, it may be given a descriptive synonym. For example, the monument mentioned in example (57) represents four statues of soldiers standing still with their heads facing the ground. Consequently, they are usually called the 'Four Sleeping Ones'.
- Colour can play a role in naming. Example (58) mentions a high-rise building in Warsaw which should have been of the golden colour according to the initial architectural plans. The final effect, however, is of the shiny glass-blue colour. Thus, the former official name was quickly changed but three versions of the name co-exist.
- Irony may be a source of linguistic creativity. A recent language joke concerns the roundabout mentioned in example (30), which was decorated several years ago with an artificial 15-meter-high palm tree, rather unnatural for the Warsaw landscape. Warsaw residents reacted to this search for exotic effects by inventing the ironic name, shown in example (59).

(56) *aleja Prymasa Tysiąclecia* ‘Avenue of the Millenium Cardinal’
informally: *aleja Wyszyńskiego* ‘Wyszyński Ave.’

(57) *Pomnik Braterstwa Broni* ‘Monument to Brotherhood in Arms’
informally: *Pomnik Czterech Śpiących* ‘Monument of the Four Sleeping’
rarely: *Pomnik Czterech Smutnych* ‘Monument of the Four Sad’

(58) *Błękitny Wieżowiec* ‘Blue Building’
initially: *Złoty Wieżowiec* ‘Golden Building’
informally: *Srebrny Wieżowiec* ‘Silver Building’

(59) *Rondo De Gaulle’a* ‘De Gaulle Roundabout’
informally: *Rondo de Palma* ‘Palm-tree Roundabout’

4 Formalisms and tools

High-quality automatic processing of urban multi-word proper names requires the precise and exhaustive description of their morpho-syntactic properties addressed in the previous sections. Our response to this need is fully lexicalized. Proper names are listed, and their inflected forms and variants are explicitly described on a two-layer basis. First, the inflectional morphology of single words is treated by *Morfeusz*, a general-purpose morphological analyser and generator, containing a large-scale lexicon of Polish. Second, a cross-language multi-word morpho-syntactic generator *Multiflex* is used to express the behaviour of compounds in the light of their constituents’ properties, as well as of some syntagmatic patterns and idiosyncratic rules. The tools have been developed independently. They have been recently integrated within a common framework, whose first experimental application is the proper name description project addressed in this paper.

In this section we present the formalisms implemented in the resulting platform, and show how they can be applied to proper names. We also discuss the interoperability facilities which were necessary for both tools to cooperate.

4.1 Morfeusz

Morfeusz is a morphological analyser of Polish available at `http://nlp.ipipan.waw.pl/~wolinski/morfeusz`. The program (or to be more precise its current version called *Morfeusz SIaT*) is based on Jan Tokarski’s lemmatization rules for Polish [32]. The set of lemmatisation rules is large (c.a. 20,000) but the rules are generic in the sense that they capture all grammatical forms possible in the language, but do not give information which of the possibilities is actually realized for any given lexeme [33]. This causes the analyser to overgenerate interpretations.

A new version of the program, named *Morfeusz SGJP*, is under development. This new version is based on an inflectional dictionary of Polish [5]. The dictionary contains exact information on inflection for almost 245,000 Polish lexemes, including several thousand proper names, which correspond to about 4,000,000 different textual words.

For the present work, a generating module has been added to *Morfeusz SGJP* which is able to provide all inflected forms for any given lexeme. The interface between this module and Multiflex is discussed in section 4.3.

The tagset used by *Morfeusz* was designed at the Institute of Computer Science, Polish Academy of Sciences (pol. IPI PAN) [34, 35]. The design and the repertoire of morphosyntactic classes and categories assumed in the IPI PAN tagset is based on the rich body of work on Polish morphosyntax by Zygmunt Saloni and his colleagues (e.g., [36, 37, 29, 38, 39, 32, 40, 5]).

The IPI PAN tagset used by *Morfeusz* is based on a homogeneous set of morphological, morphosyntactic and (as secondary) syntactic criteria. The main criteria for delimiting grammatical classes are morphological (how a given class inflects; e.g., nouns inflect for case, but not for gender) and morphosyntactic (in which categories it agrees with other classes; e.g., Polish nouns agree in gender with adjectives and verbs). Such an approach leads to more detailed classes than traditional parts of speech (POS); we call such classes *flexemic classes.* Some of these flexemic classes correspond rather directly to the traditional parts of speech, e.g., noun, adjective, adverb, preposition, conjunction, particle; others are more fine grained, e.g., various verbal classes such as infinitive, four classes for the four participial forms (two adjectival and two adverbial), non-past verb, impersonal *-no/-to* form, imperative, l-participle (word-forms like *przysz-l-i*), gerund. The traditional classes of numerals and pronouns, usually defined on the basis of their semantics, are redefined in purely morphological and morphosyntactic terms: the flexemic class of numerals only consists of cardinal numerals, and three specific classes for nominal pronouns are distinguished: non-3rd person pronoun, 3rd person pronoun, and the pronoun *siebie.* Other flexemic classes are: depreciative noun, ad-adjectival adjective, post-prepositional adjective, future *być*, agglutinative *być* (forms like *-śmy*), *winien*-like verb, predicative, alien nominal/other. Grammatical categories assumed in the tagset include traditional categories such as number, case, person, degree, aspect, gender, as well as more restricted categories. Some of them were first introduced in the work of Jan Tokarski and Zygmunt Saloni [32], such as accentability, post-prepositionality, acommodability, agglutination, vocability and negativity.

The following examples should give the impression of the granularity and the positional nature of the tagset:

- *ulicami*: `ulica` ‘street’, `subst:pl:inst:f` — noun, plural number, instrumental case, feminine gender;
- *przyszedł*: `przyjść` ‘come’, `praet:sg:m1.m2.m3:perf` — l-participle, singular number, one of the three masculine genders (i.e., in fact, this is an abbreviation for three different tags), perfective aspect;
- *czterech*: `cztery` ‘four’, `num:pl:nom.voc:m1:rec|num:pl:gen.loc:m1.m2.m3.n2.f:congr|num:pl:acc:m1:congr` — three separate tags are possible: the first one nominative or vocative in masculine personal; the second one genitive or locative in any of masculine subgenders, neuter impersonal, or feminine; the third one accusative in masculine personal. The first tag bears

the value `rec` of accomodability while the other two have the value `congr` (see section 5.1 for the meaning of these values).

4.2 Multiflex

Multiflex [41], [42] is a cross-language morpho-syntactic generator of multi-word units. It relies on a graph-based formalism in which all inflected forms of a compound and their variants are described within one graph. Compounds having the same morpho-syntactic behaviour are assigned to the same graph. Each path in the graph describes one or more inflected forms and variants. A unification mechanism allows one to account for agreement within constituents, and to represent huge inflection paradigms compactly.

Description on the language level. Multiflex formalism adapts to the given natural language by encoding its morphological model, i.e. the list of all inflectional categories (number, gender, case, person, etc.) and their corresponding values (e.g. singular and plural for number, feminine, masculine and neuter for gender, etc.), as well as the existing inflectional classes (noun, verb, adjective, etc.). Each class needs information on the categories it inflects for and those that are relevant to it but have constant values. For instance, Polish nouns usually inflect for number and case, and they have a fixed gender.

Fig. 1 shows the encoding of the Polish morphological model according to the IPI PAN tagset, as explained in section 4.1. As it can be easily seen, this model admits two numbers (*Nb*), seven cases (*Case*), nine genders (*Gen*), three persons (*Pers*), three degrees (*Deg*), etc. Adjectives (*adj*) are supposed to inflect for number, case, gender and degree. Cardinal numerals (*num*) should have a fixed number and inflect for case, gender and accomodability (agreeing as in *dwaj*, or governing as in *dwóch*). The whole model includes 12 inflectional categories, 38 values and 32 classes.

Segmentation of a compound lemma. The morpho-syntactic description implemented in *Multiflex* relies on the assumption that it is possible to detect boundaries between individual components within a compound. How a single component (token) is defined depends on the underlying morphological module for the single words *Multiflex* works with. In this study, the *Morfeusz*-like token definition is admitted. A lexical token is one of the following:

- a contiguous sequence of the alphabet characters between two non alphabet characters (e.g. *drzewo* is one token)
- a sub-sequence appearing within a contiguous sequence of the alphabet characters, in a closed list of well defined cases (e.g. *kiedy|śmy* contains two tokens)
- several contiguous sequences of alphabet and non alphabet characters, in a closed list of well defined cases (e.g. *ping-pong*, *d'Arc*)
- a single non alphabet character such as a digit or a punctuation mark (space, hyphen, apostrophe, dot, etc.)

```
Polish
⟨CATEGORIES⟩
Nb:          sg , pl
Case:        nom, gen, dat, acc, inst, loc, voc
Gen:         m1, m2, m3, f, n1, n2, p1, p2, p3
Pers:        pri, sec, ter
Deg:         pos, comp, sup
Asp:         imperf, perf
. . .
⟨CLASSES⟩
subst:       (Nb,⟨var⟩),(Case,⟨var⟩),(Gen,⟨fixed⟩)
num:         (Nb,⟨fixed⟩),(Case,⟨var⟩),(Gen,⟨var⟩),(Accom,⟨var⟩)
numcol:      (Nb,⟨fixed⟩),(Case,⟨var⟩),(Gen,⟨fixed⟩),(Accom,⟨var⟩)
adj:         (Nb,⟨var⟩),(Case,⟨var⟩),(Gen,⟨var⟩),(Deg,⟨var⟩)
adja:
. . .
```

Fig. 1. Morphological model of Polish in *Multiflex*

Components of a compound lemma are numbered, which allows for referring to them while describing the compound inflected forms and their variants. For instance, Fig. 2 shows the division of compound proper names (29) and (49) into single components many of which are spaces, dots and quotes.

Port	⟨blank⟩	Lotniczy	⟨blank⟩	im	.	⟨blank⟩	F	.	⟨blank⟩	Chopina
$1	$2	$3	$4	$5	$6	$7	$8	$9	$10	$11

Rondo	⟨blank⟩	Zgrupowania	⟨blank⟩	AK	⟨blank⟩	“	Radosław	”
$1	$2	$3	$4	$5	$6	$7	$8	$9

Fig. 2. Numbering single components within compounds in *Multiflex*

Annotating components within a compound. Naturally enough, the inflection paradigm of a compound lemma, according to *Multiflex*, consists mostly of pointing at which inflected form of a compound requires which transformations of its single components. For instance, in order to obtain the locative forms of the following compound toponyms in Warsaw:

(60) *Most Grota-Roweckiego* ‘Grot-Rowecki bridge’

(61) *Aleje Jerozolimskie* ‘Avenue of Jerusalem’, pl.

we have to inflect their head nouns *Most* ‘bridge’ and *Aleje* ‘avenues’ for the locative while keeping their number unchanged (singular and plural, respectively). All remaining constituents in example (60) do not change. In example (61) the adjective *Jerozolimskie* is subject to agreement rules with the head noun *Aleje*, i.e. it gets inflected for the locative with no change in number. The results are:

(62) *Moście Grota-Roweckiego*

(63) *Alejach Jerozolimskich*

Thus, in order to inflect compounds we need to be able to analyse single words morphologically, as well as to synthesize desired forms. That is where *Multiflex* depends on a morphological module for simple words, e.g. *Morfeusz*. It is necessary to annotate each single component of a compound that is subject to inflection.[3] Note that components of a compound lemma do not necessarily represent lemmas themselves (e.g. *Aleje* is not a lemma in example (61)), therefore they need to be annotated with their own lemmas, accompanied by their inflectional values as in a *Morfeusz* tag (cf. [15]). Whenever a lemma is morphologically ambiguous, a cardinal number allows the indication of the desired homonym. For instance, the following *Morfeusz* annotation is attributed to the component *Aleje* within the compound (61):

(64) aleja.0:subst:pl:nom:f
'lemma *aleja*, first homonym, noun, plural, nominative, feminine'

which can be graphically represented as in Fig. 3.

Aleje ⇒
lemma: *aleja*
homonym: *0*
class: *subst*
Nb: *pl*
Case: *nom*
Gen: *f*

Fig. 3. A possible annotation of the single component *Aleje*

Similarly, processing all possibly inflected components in the compound lemmas (60) and (61) results in annotations shown in figures 4 and 5.

The form and meaning of the inflection codes *NC-CXXXX* and *NC-CXC* attached to the compound entries are explained in the following subsection.

Inflection graphs for compounds. The *Multiflex* graph-based formalism is meant for precise and exhaustive description of the morphological behaviour of compounds, i.e. each correct form must be described and no incorrect form can be admitted. Each compound is assigned a graph identifier, as *NC-CXXXX* and *NC-CXC*[4] in figures 4 and 5. Each path in a graph contains zero or more boxes,

[3] Such a 'two-layer' description of compounds, first on the level of single words, then on the level of their sequences, is the most commonly assumed in related works, however some approaches see compounds as plain sequences of characters – cf. [17].

[4] Names of the inflection graphs are arbitrary and follow a convention allowing easy navigation in already existing graphs. Here *NC* stands for *Noun-Compound*, *X* stands for an uninflected unit, and *C* for a possibly inflected one.

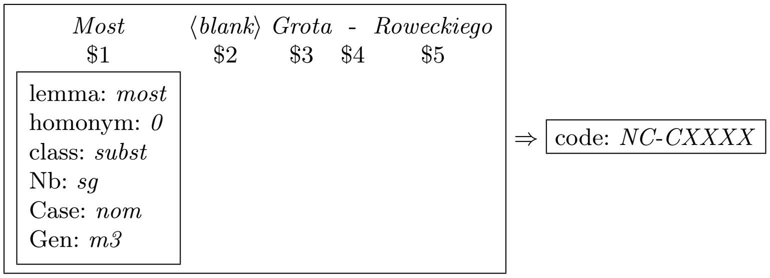

Fig. 4. Lemma annotation for *Most Grota-Roweckiego*

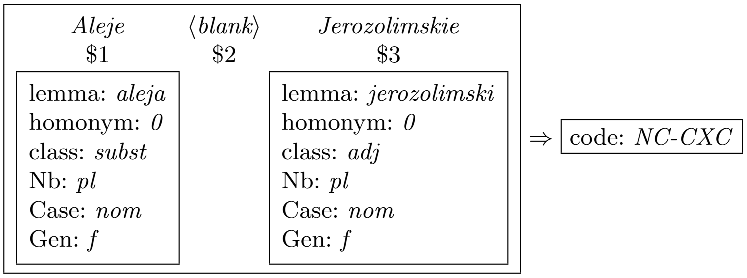

Fig. 5. Lemma annotation for *Aleje Jerozolimskie*

and describes one or more inflected forms or variants of a compound. The path begins with the leftmost arrow and ends with the rightmost encircled box. Morphological descriptions appearing inside the boxes refer to single constituents, while those appearing under the boxes refer to the whole compound.

For instance, Fig. 6 represents the inflection graph *NC-CXXXX* for compound (60). It contains a unique path referring to five constituents of the compound lemma. If a constituent number appears in a box with no morphological description, as in the case of *$2* through *$5*, then it should be left unchanged. However, if it is accompanied by one or more category-value equations, then the constituent (here *$1*) must be inflected for the desired form.

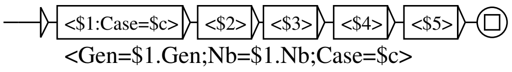

Fig. 6. Inflection graph *NC-CXXXX* for compound inflecting like *Most Grota-Roweckiego*

The right-hand side of a category-value equation may have one of the three forms:

- A *constant value*. It is taken from the domain of the corresponding category, as specified in the language model in Fig. 1, as in *Nb=pl; Gen=f; Case=acc.*

It indicates that the current component has to be inflected for the corresponding form.

- A *unification variable.* It is introduced by the *$* sign, as in *Nb=$n; Gen=$g; Case=$c*. It may take any value from the corresponding category domain, unless it is limited by unification constraints with respect to other constituents (see Fig. 8). For instance, in *Case=$c* the variable *$c* may potentially take values *nom* through *voc*.
- An *inherited value.* It is introduced by a constituent number followed by a category, e.g. *Nb=$1.Nb; Gen=$2.Gen; Case=$5.Case*. It is fixed and equal to the value which the corresponding constituent has in the compound lemma. For instance, *Gen=$1.Gen* means that the gender must be the same as the gender of the first constituent in the lemma.

For instance, when the graph in Fig. 6, is applied to the annotated lemma in Fig. 4 the first constituent *$1*, *Most*, inflects freely for case (see *Case=$c* in the first box). The whole compound obtains the same case value as the first constituent takes in each particular form during inflection (see *Case=$c* under the path). Gender and number values of the compound are equal to those possessed by the first constituent in the compound lemma (see *Gen=$1.Gen; Nb=$1.Nb* under the path), here *m3* and *sg*, as in the annotation in Fig. 4.

Complete exploration of a graph consists in following each path as many times as there are possible values for all the unification variables present on the path. Here, the unique path is explored seven times (variable *$c* may take seven values), which results in seven compound inflected forms, annotated with their lemma and *Morfeusz*-like tags, as shown in Fig. 7.

Most Grota-Roweckiego, Most Grota-Roweckiego:subst:m3:sg:nom
Mostu Grota-Roweckiego, Most Grota-Roweckiego:subst:m3:sg:gen
Mostowi Grota-Roweckiego, Most Grota-Roweckiego:subst:m3:sg:dat
Most Grota-Roweckiego, Most Grota-Roweckiego:subst:m3:sg:acc
Mostem Grota-Roweckiego, Most Grota-Roweckiego:subst:m3:sg:inst
Moście Grota-Roweckiego, Most Grota-Roweckiego:subst:m3:sg:loc
Moście Grota-Roweckiego, Most Grota-Roweckiego:subst:m3:sg:voc

Fig. 7. Annotated inflected forms of *Most Grota-Roweckiego*

Unification variables not only allow to freely inflect a component for a certain category, but to express agreement and government rules as well. For instance, the graph in Fig. 8 contains the same unification variable *$c* for both the first and the third constituent. The variable may take any of the seven cases provided that each time it is identical for these two constituents. Thus, instead of $7 \times 7 = 49$ different form combinations we obtain only seven correct forms, one of which has two variants, which results in eight forms, as shown in Fig. 9.

Note that the inherited values allow the graph to be independent of the inflectional features of its head. For instance, in Fig. 8 the equations *Gen=$1.Gen; Nb=$1.Nb* allow the graph to be assigned not only to feminine plural compounds as in (61), but also to masculine/neuter and singular ones, as in Fig. 10.

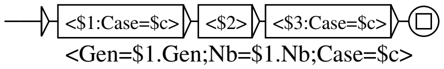

Fig. 8. Inflection graph *NC-CXC* for compounds inflecting like *Aleje Jerozolimskie*

Aleje Jerozolimskie, Aleje Jerozolimskie:subst:f:pl:nom
Alei Jerozolimskich, Aleje Jerozolimskie:subst:f:pl:gen
Alej Jerozolimskich, Aleje Jerozolimskie:subst:f:pl:gen
Alejom Jerozolimskim, Aleje Jerozolimskie:subst:f:pl:dat
Aleje Jerozolimskie, Aleje Jerozolimskie:subst:f:pl:acc
Alejami Jerozolimskimi, Aleje Jerozolimskie:subst:f:pl:inst
Alejach Jerozolimskich, Aleje Jerozolimskie:subst:f:pl:loc
Aleje Jerozolimskie, Aleje Jerozolimskie:subst:f:pl:voc

Fig. 9. Annotated inflected forms of *Aleje Jerozolimskie*

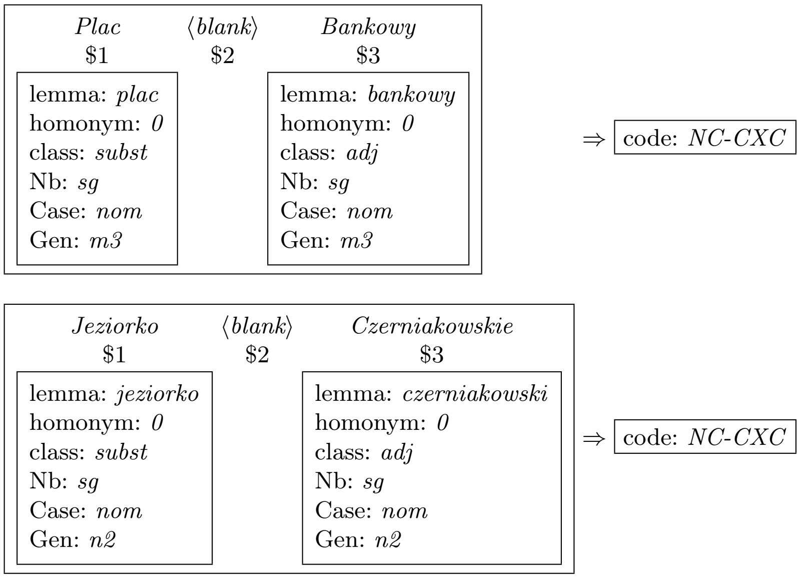

Fig. 10. Lemma annotations for *Plac Bankowy* 'Bank Square' and *Jeziorko Czerniakowskie* 'Czerniakowskie Lake'

Bypassing components. As mentioned in section 3.2 urban toponyms show a high degree of variability. They tend in particular to be more frequently used in their abbreviated forms than in their full official forms. Omitting a component is easy to express in a graph by bypassing the appropriate box.[5]

For instance, recall that example (49) admits at least three morpho-syntactic variants, all of which can be inflected for case. The corresponding annotation and

[5] Inserting a new component is also natural enough to be expressed within a graph, however we have not run across a toponym example requiring this feature.

inflection graph are shown in Fig. 11. The lowest path describes the official full version of the toponym, *Rondo Zgrupowania AK "Radosław"*, where all nine components are present, and where only the initial head noun *Rondo* inflects for case.[6] The middle path corresponds to the slightly abbreviated variant, *Rondo Zgrupowania "Radosław"*, with the fifth component, i.e. the *AK* acronym, and the following blank space missing. The uppermost path triggers the shortest (and most commonly used) version, *Rondo Radosława*, where *Radosław* shifts into the head's complement position and takes the genitive case.

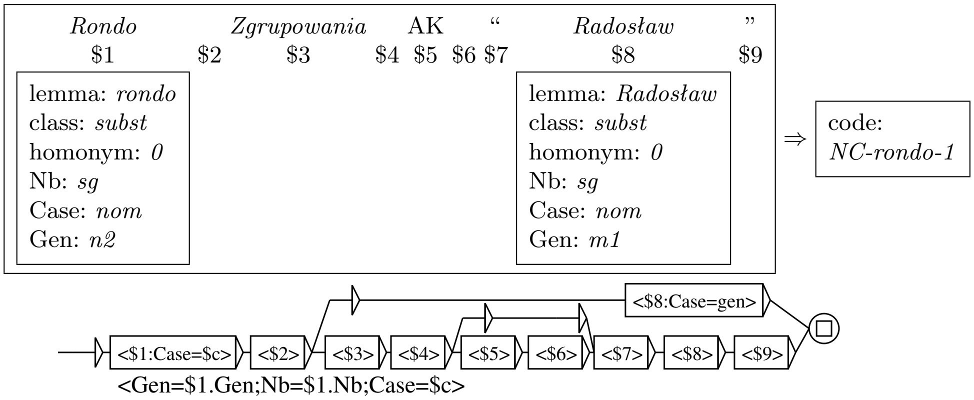

Fig. 11. Lemma annotation and inflection graph for *Rondo Zgrupowania AK "Radosław"*

4.3 Interoperability and collaborative framework

As shown in the two previous sections, the morpho-syntactic description of compounds in *Multiflex* is based on a 'two-layer' approach. Single words are described first, then each inflected compound form is seen roughly as a particular combination of the inflected forms of its components. If the combination is regular (i.e. respects "typical" syntagm patterns and agreement rules), the corresponding graph is shared by a large class of compounds (as *NC-CXC* in Fig. 5 and Fig. 10. Conversely, if it is idiosyncratic, the graph applies to isolated compounds (as *NC-rondo-1* in Fig. 11).

Multiflex was initially developed within a large-scale multilingual lexicographic project based on the volume [43] and supported by the *Unitex* system [44]. However, numerous other morphological systems for single words have been developed for many languages. Therefore, for the reasons of interoperability, *Multiflex* does not impose its own model but is meant to cooperate with any other

[6] Inflection for number is a controversial question for toponyms as they usually name unique objects. However, an exceptional plural usage is acceptable as in: *Twierdzisz, że w Warszawie istnieją dwa Ronda Radosława?* 'Do you claim that there are two Radoslaw Roundabouts in Warsaw?'

external morphological module for single words, hereafter called the *underlying module*, as long as the following interface constraints are observed.

Common morphological model. The underlying module and *Multiflex* must share the same morphological model. The present model for Polish, common for *Morfeusz* and *Multiflex*, is described in Fig. 1. It represents the inflectional morphology, both of single and compound words, as a group of categories (number, gender, case, etc.), each of which admits a list of values (singular, plural, feminine, etc.). *Multiflex* interprets this model in the sense of a Cartesian product, i.e. if an inflectional class (noun, adjective, etc.) inflects for a number of categories, then, implicitly, each value of one category combines with each value of another category. For instance, if a noun inflects for number and case, it admits at least $2 \times 7 = 14$ inflected forms. Exceptions to this rule must either be controlled by the underlying module (no form is generated for a particular combination of features), or be explicitly mentioned in the corresponding inflection graphs.

However, some inflectional phenomena, particularly in Slavic languages, are not necessarily of the nature of a Cartesian product. For instance, in the model for Serbian in [45], some particular value combinations are systematically prohibited, others are modelled by ‘no-care’ values. With *Multiflex* applied to this model in [46], such constraints cannot be expressed at the most general language level, but have to be mentioned for each individual inflection graph concerned.

Recognition of token boundaries. The underlying module should provide a clear-cut definition of token boundaries. As shown in [17, sec. 4.1], the distinction between single words and compounds on the purely graphical level is controversial. The definition of an indivisible graphical item may or may not:

- allow token boundaries to occur within contiguous sequences of letters; for instance, *Białystok* ‘literally: White Slope = a city name’ contains an adjective and a noun with no separating space, however, both components inflect as in a typical *Adjective Noun* syntagm, *Białego|stoku, Białym|stokiem, etc.*,
- admit separators within a token, e.g. *aujourd’hui* ‘today’ in French may be seen either as a single word or as two words separated by an apostrophe,
- view sequences of non-alphabet characters as single tokens or as sequences of tokens; e.g. a numeral such as *2008* may be seen as one or four tokens.

Here again, *Multiflex* does not impose its own view on a token boundary but adapts to the underlying module via an interface function. With *Morfeusz*, an indivisible token is defined as in section 4.2 where segmentation of a compound lemma is discussed.

Generation of particular inflected forms on demand. Given a lemma and a morphological tag, the underlying module should generate all inflected forms of the lemma corresponding to the tag.

For a given lemma, a homonym number and an inflectional class, the generating module of *Morfeusz* produces the set of all inflected forms of the word together with their annotations, as in Fig. 12. Choosing the desired form within this set is done by a matching function which solves possible factorizations and alternatives, as in the case of the form *czterech* explained in section 4.1.

lemma: *cztery*
homonym: *0*
class: *num*

⇒

czworgu	num:pl:dat.loc:n1.p1.p2:congr.rec
czterem	num:pl:dat:m1.m2.m3.n2.f:congr
czworga	num:pl:gen:n1.p1.p2:rec
czterema	num:pl:inst:m1.m2.m3.n2.f:congr
czworgiem	num:pl:inst:n1.p1.p2:rec
cztery	num:pl:nom.acc.voc:m2.m3.n2.f:congr
czworo	num:pl:nom.acc.voc:n1.p1.p2:rec
czterej	num:pl:nom.voc:m1:congr
czterech	num:pl:nom.voc:m1:rec| num:pl:gen.loc:m1.m2.m3.n2.f:congr| num:pl:acc:m1:congr

Fig. 12. Annotated inflected forms of *cztery*

As a result, we obtain a process which is roughly opposite to annotation. For instance, for the same lemma and morphological tag as in Fig. 3, the generated form is *aleje*, as shown on the left-hand side of Fig. 13. Note, however, that the form generated for a particular morphological tag is not necessarily unique because of variants, as seen in the middle of the same figure.

lemma: *aleja*
homonym: *0*
class: *subst*
Nb: *pl*
Case: *nom*
Gen: *f*

⇒ *aleje*

lemma: *aleja*
homonym: *0*
class: *subst*
Nb: *pl*
Case: *gen*
Gen: *m1*

⇒ *alei*
alej

lemma: *cztery*
homonym: *0*
class: *num*
Nb: *pl*
Case: *nom*
Gen: *f*
Accom: *rec*

⇒ czterech

Fig. 13. Generating inflected forms of *aleja* and *cztery*

5 Some interesting urban toponyms

5.1 Numerals within Polish compounds

Polish ordinal numerals show rather complex morpho-syntactic properties, as discussed largely in [29] and [31]. In particular, in some genders they inflect for accomodability, i.e. in the agreeing form (*cong*) they agree with the nouns they modify, while in the governing form (*rec*) they require the noun to appear in genitive. Governing forms admit all seven cases in neutral, but only two cases in masculine human. Moreover, numerals *jeden* 'one' through *cztery* 'four' behave different from all others.

Before this behaviour finds an elegant model at the language level, it needs to be described within inflection graphs for compounds containing numerals. Recall, for instance, example (57) mentioning the popular synonym of the Monument to Brotherhood in Arms:

(65) [*Pomnik*$_{nom}$ *Czterech*$_{gen}$ *Śpiących*$_{gen}$]$_{nom}$
'Monument of Four Sleeping'

It belongs to the class of type *Noun [Num Noun]*$_{gen}$ whose first noun *Pomnik* 'monument' holds the head position and is the only one to inflect. This fact is represented by the uppermost path in the graph in Fig. 14. The same compound admits an elliptical variant which may appear either in the agreeing or in the governing form:

(66) [*Czterej*$_{nom.congr}$ *Śpiący*$_{nom}$]$_{nom}$
'[The] Four Sleeping'

[*Czterech*$_{nom.rec}$ *Śpiących*$_{gen}$]$_{nom}$
'[The] Four Sleeping'

In the former case the numeral and the noun agree in case, which is represented by the middle path in Fig. 14, where the unification variable *$c* is common for the third and the fifth[7] component (*Czterech* and *Śpiących*). In the latter case the numeral imposes the genitive gender on the noun, that is why in the lowermost path in Fig. 14 the fifth component gender is assigned the constant *gen*.

5.2 Embedded compounds

The *Multiflex* development project, at its present stage, is dedicated to the optimal handling of embedded compounds, whose quantitative importance is particularly significant within urban and institutional proper names. Note that in previous examples, compounds are assigned to flat representations, i.e. their internal syntactic structures are neglected. In compounds containing numerous components this approach may lead to considerable complexity in inflectional graphs and to some degree of redundancy within the pattern description. Consider examples (67) and (68) allowing for a considerable number of elliptical variants (each of them inflected), represented by graphs in Fig. 15 and Fig. 16, respectively.

(67) *ulica Marii Skłodowskiej-Curie* 'Maria Skłodowska-Curie Street';
ulica Marii Skłodowskiej; *ulica Marii Curie*; *ulica Skłodowskiej-Curie*;
ulica Skłodowskiej; *Marii Skłodowskiej-Curie*; *Marii Skłodowskiej*, etc.

[7] Recall that the components are numbered with respect to their position in the compound lemma, which is not necessarily the same as in the inflected form.

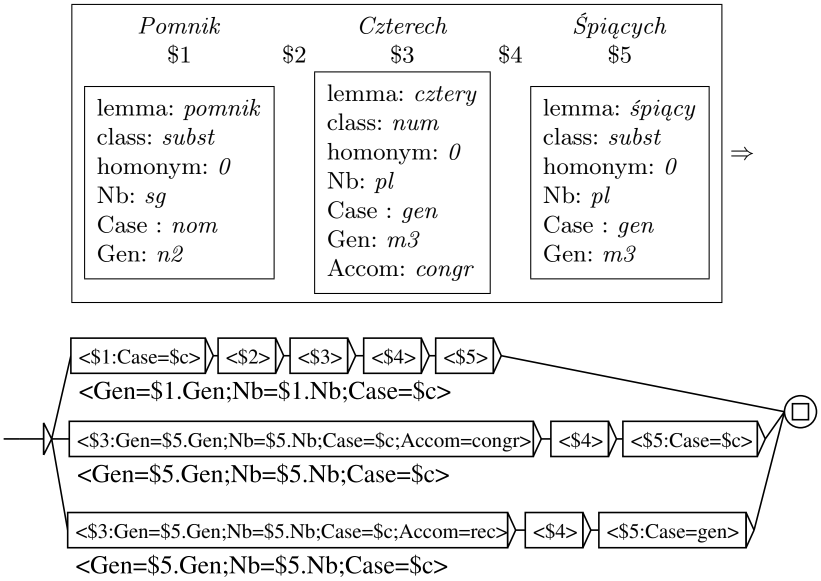

Fig. 14. Lemma annotation and inflection graph for *Pomnik Czterech Śpiących*

(68) *XXIII Liceum Ogólnokształcące im. Marii Skłodowskiej-Curie*
'Maria Skłodowska-Curie 23rd High School'
XXIII Liceum Ogólnokształcące im. Marii Skłodowskiej
XXIII Liceum Ogólnokształcące im. Marii Curie
XXIII Liceum Ogólnokształcące im. Skłodowskiej-Curie
XXIII Liceum Ogólnokształcące im. Skłodowskiej
XXIII Liceum Ogólnokształcące
XXIII Liceum im. Marii Skłodowskiej-Curie
Liceum Ogólnokształcące im. Marii Skłodowskiej-Curie
Liceum im. Marii Skłodowskiej-Curie
Liceum Marii Skłodowskiej-Curie, etc.

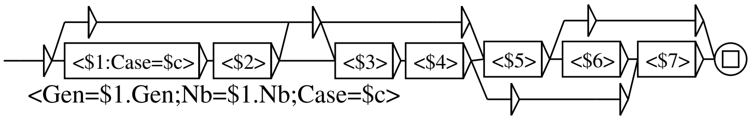

Fig. 15. A flat description of *ulica Marii Skłodowskiej-Curie*

The high number of these variants results in particular from the fact that the underlying person name *Maria Skłodowska-Curie* admits five variants on its own. In graphs in Fig. 15 and Fig. 16 these variants have to be described independently (by combinations of the five rightmost boxes in each graph). Thus, the morphosyntactic behaviour of *Maria Skłodowska-Curie* is represented redundantly each time a compound contains this person name.

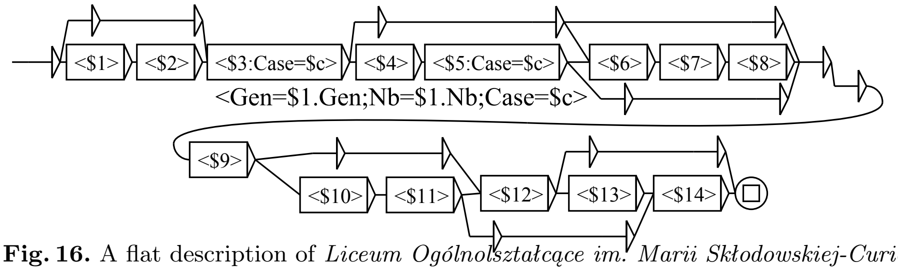

Fig. 16. A flat description of *Liceum Ogólnokształcące im. Marii Skłodowskiej-Curie*

Rather than describing such compounds as flat structures, it is more elegant to have a unique representation of each person and institution name, such as in Fig. 17 and Fig. 18. Then, compounds (67) and (68) are described as syntactically simpler structures of types *Noun* $Noun_{gen}$ and *Noun im.* $Noun_{gen}$, in which the first and the last components may be compounds on their own, as shown in Fig. 19 and Fig. 20. Thus, the same graph can be applied to compounds with a different number of components. For instance, the graph in Fig. 19 describes not only example (67), but (50) as well, while the graph in Fig. 20 applies both to example (68) and to (29).

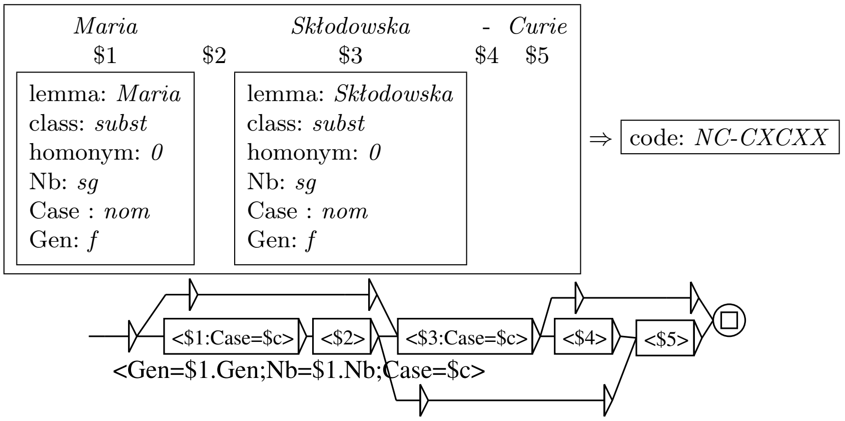

Fig. 17. Lemma annotation and inflection graph for *Maria Skłodowska-Curie*

5.3 Extreme elliptical variants

The wide-spread tendency of urban toponyms to appear in elliptical variants has been discussed in the preceding sections. When the headword of a compound is omitted, as in example (66), the remaining complement usually shifts to the head position, possibly changing its gender and number, as in:

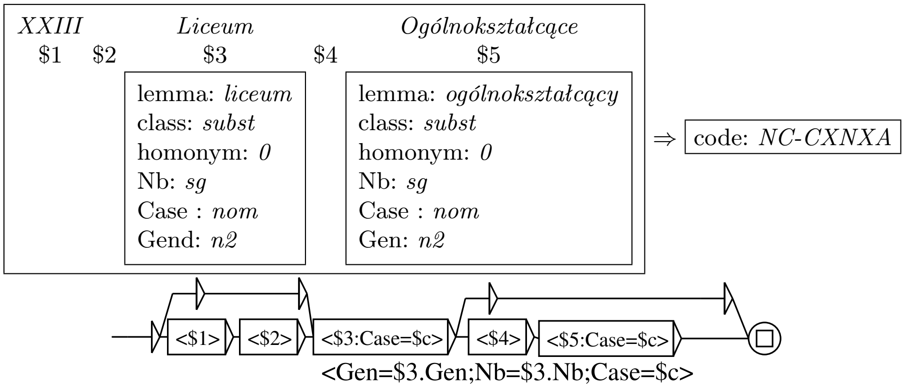

Fig. 18. Lemma annotation and inflection graph for *XXIII Liceum Ogólnokształcące*

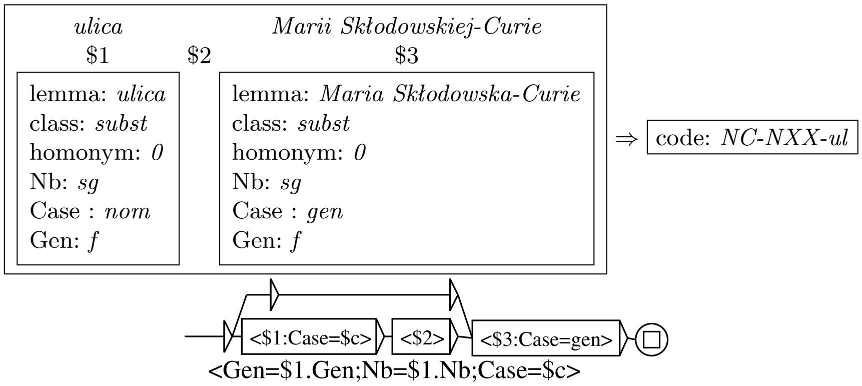

Fig. 19. Embedded lemma annotation and inflection graph for *ulica Marii Skłodowskiej-Curie*

(69) *Pomnik*$_{nom.sg}$ *[Czterech Śpiących]*$_{gen.pl}$ *[znajduje się]*$_{sg}$ *na Placu Wileńskim.*
'The monument of the Four Sleeping is located in Vilnius Square.'
[Czterej Śpiący]$_{nom.pl}$ *[znajdują się]*$_{pl}$ *na Placu Wileńskim.*
'The Four Sleeping are located in Vilnius Square.'

As shown in the inflection graph in Fig. 19 this 'shifting rule' is not respected in street names of type *ulica Noun*$_{gen}$. The headword *ulica* 'street' is rarely used, particularly in spoken language, however, the remaining complement keeps its original genitive form, as in the following example:

(70) *Ulica*$_{nom.fem}$ *[Kazimierza Pułaskiego]*$_{gen.masc}$ *jest zatłoczona*$_{fem}$.
'Kazimierz Pułaski Street is crowded.'
Kazimierza Pułaskiego jest zatłoczona$_{fem}$.
Pułaskiego jest zatłoczona$_{fem}$.

(71) *Pułaskiego, Pułaski:subst:m1:sg:gen*
Pułaskiego, ulica Kazimierza Pułaskiego:subst:f:sg:nom

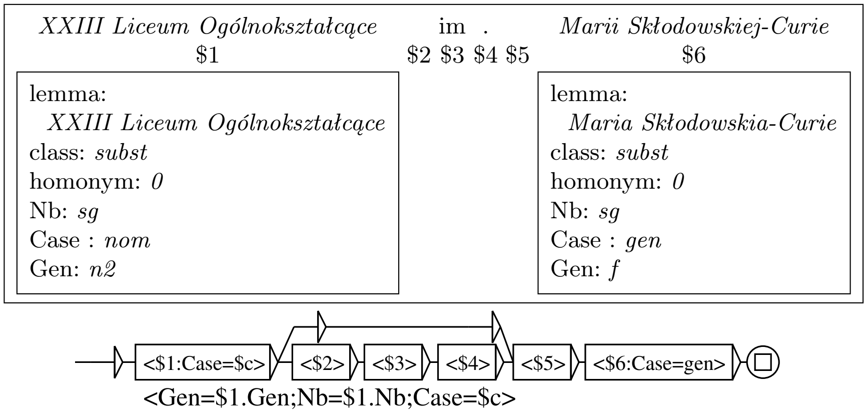

Fig. 20. Embedded lemma annotation and inflection graph for *XXIII Liceum Ogólnokształcące im. Marii Skłodowskiej-Curie*

This produces the surprising effect of a double morphological interpretation in the last two sentences. The sequences *Kazimierza Pułaskiego* and *Pułaskiego* as such can only be analysed as genitive singular masculine forms. However, as subjects in a sentence, they are necessarily in nominative singular feminine. In other words, depending on which lemma is assigned to the form *Pułaskiego*, it may have different morphological features, as shown in example 71.

6 Remaining problems

We believe that the morphosyntactic framework presented above allows, in its present state, to initiate a large-scale description of Polish proper names, in particular urban toponyms. Nevertheless, we are aware of the fact that some remaining problems still await solutions.

6.1 Conflating derivational and semantic variants

For applications such as information retrieval, it is crucial to be able to conflate all kinds of variants that a proper name may admit. In other words it is necessary to recognize different textual representations as variants of the same proper name. As it was shown above, conflating inflectional and syntactic (in particular elliptical) variants is possible at present. However, we still have no satisfying method of handling derivational and semantic variants.

For instance, the two following names are equivalent:

(72) *Park Łazienki*, *Park Łazienkowski* 'Łazienki Park'

but we cannot describe them as variants because in the model depicted in Fig. 1 we lack morphological features allowing to say that the adjective *Łazienkowski* is a derivative of the noun *Łazienki*.

Many semantic variants, as those discussed in section 3.2, are even more difficult to recognize since they frequently contain components with no or few morpho-syntactic relations to the original compound lemma. Recall, for instance, the following synonymic pairs:

(73) *Rondo Zgrupowania AK "Radosław" = Rondo Babka*

(74) *aleja Jana Pawła II = aleja Juliana Marchlewskiego*

Only additional external links may solve such problems of synonymy.

6.2 Distinguishing variants

Inflectional and syntactic variants of an inflected form frequently share the same morphological tag. Therefore, they are hard to distinguish within embedded constructions. Recall example (67) for which we suggest first to describe the person name separately as in Fig. 17, and then to represent the street name as a structure of type *Noun* $Noun_{gen}$ in which the third constituent may be an embedded compound, as in Fig. 19. The latter graph allows the third constituent to appear in any variant of its genitive form, here: *Marii Skłodowskiej-Curie, Marii Skłodowskiej, Marii Curie, Skłodowskiej*, or *Skłodowskiej-Curie*. In particular, applying one of the two last variants, where the first name is omitted, gives *ulica Skłodowskiej-Curie* and *ulica Skłodowskiej*. Similarly, all the forms in example (50) can be described in the same way.

Applying the same graph to similar examples fails in some cases. If the surname is morphologically a masculine noun it remains uninflected in female person names, as in example (75). If such a name appears in a street name, the first name should not be omitted (see example (76)).

(75) *Emilia Plater, Emilii Plater, Emilię Plater*, etc.

(76) *ulica Emilii Plater, Emilii Plater, *ulica Plater, *Plater* 'Emilia Plater street'

This fact cannot, however, be accounted for if the generic graph in Fig. 19 is used. Since the elliptic form *Plater* has an identical morphological description as the full form *Emilii Plater* (see example (77)), no morphological feature can be used in the graph to restrict the use of the elliptic variant.

(77) *Emilii Plater, Emilia Plater:subst:f:sg:gen*
Plater, Emilia Plater:subst:f:sg:gen

6.3 Morpho-graphical problems

In many examples shown above, compound proper names contain capitalised common words, such as *Most* 'Bridge', *Aleje* 'Avenue', *Plac* 'Square', *Czterech* 'Four', etc. Currently, these words are formally described in *Morfeusz* in the lower-case forms only. Thus, when a compound like (61) is annotated as in Fig. 5, the single-word lemmas appear in lower case. While exploring the inflection graphs *Multiflex* requires of *Morfeusz* to generate particular inflected forms of these lemmas. The resulting forms are in the lower case, which produces *aleje jerozolimskie, alei jerozolimskich*, etc. Additional *ad hoc* filters are needed in order to provide the desired capitalised effect shown in Fig. 9.

In order to solve this problem more elegantly we are considering two alternative solutions. First, the lower or the upper case spelling of a word may be seen as a morphological value. Thus, an extra inflectional category *LetterCase* can be introduced to the model in Fig. 1 with possible values *low, upp, allupp* (lower, upper, all upper), etc. One problem here would be to handle examples where the intial and some following letters (but not all) are capitalized as in *PeKaO*. Second, we can admit that the letter case is a problem on a different level and is not to be handled within morphosyntactic description. If we decide so, we still need to handle possible dependencies between morphological and case variants.

A similar problem concerns abbreviations, acronyms and initialisms addressed in examples (26) through (38). It is necessary to express the fact that *al.* is a possible variant of *aleja* or *aleje*, and that *ZUS* stands for *Zakład Ubezbieczeń Społecznych* 'Social Security Office'. Then again, creating a short form can be seen as a morphological category and incorporated into the model in Fig. 1. Alternatively, it can be handled at a separate level as a semi-derivational process.

7 Perspectives

The *Multiflex-Morfeusz* platform presented in this paper allows for the treatment of inflectional morphology and variation not only in proper names, but also in various kinds of Polish multi-word units, such as compounds, complex terms, contiguous collocations, etc. At present, the tool is generation-oriented: for a given compound it allows one to generate some or all of its inflected forms and variants. Naturally enough, we also consider its integration into a fully-fledged morphological analyser in which compounding would be largely taken into account.

The particular application of our platform to the description of Polish proper names was motivated by the outcome of the LUNA project (see section 2). In the long run, we aim at a large linguistic description of Warsaw toponyms, that could serve as a part of an intelligent information system containing a model of Warsaw topography (streets, places, institutions, monuments, rivers, etc.) and transport (bus- and tram-stops, underground stations, etc.). It could be applied to a dialogue system in which a voice server answers natural language questions, e.g. how to get by public transport from one place in the city to another.

Each lexical approach such as ours is labour intensive. Proper names need to be listed and described explicitly. Maintaining such human-controlled lists becomes harder with their growing size. For better efficiency and reliability, automated ergonomic procedures are needed, thanks to which manual work can be minimized. We are presently working on a graphical encoding aid module supporting the process of annotation of compound lemmas. It will specifically include an automatic morphological look-up of simple constituents, and a simulation of the exploration of inflectional graphs.

Another interesting perspective is integration of Polish names into the multilingual ontology of proper names *Prolexbase* [47], meant in particular for machine translation. If links between the Prolex conceptual proper names and the Polish lexical items are introduced, the Polish module will be able to benefit from the language-independent relations already present in the database. Such relations (synonymy, meronymy, etc.) are useful e.g. in uni- or multilanguage information extraction and text alignment.

8 Conclusions

We have presented ongoing work on the formal description of Polish compound proper names. We have performed an analysis of their linguistic properties. We have shown how two separate morphological tools, *Morfeusz* and *Multiflex*, were brought into a collaborative framework allowing for a two-layer model of inflection and variation of compounds. Thus, graph-based rules in the latter tool describe combinations of particular inflected forms of single units controlled by the former tool. The resulting framework allows one to address not only the inflectional morphology, but also other linguistic phenomena such as ellipsis, accomodability of numerals, and embedded compounding. The inflected paradigms are represented compactly thanks to unification. Some lacking facilities include proper treatment of derivational and semantic variants, distinguishing variants which share the same morphological tag, as well as modelling letter cases and abbreviations. At its present stage the platform allows the undertaking of a project of describing urban toponyms for a model of transportation system in Warsaw, in view of dialogue systems and other natural language applications.

Acknowledgements. The work presented has been supported by the French and Polish Ministries of Research and Higher Education within the Hubert Curien Partnership POLONIUM programme.

References

1. Handke, K.: Słownik nazewnictwa Warszawy. Slawistyczny Ośrodek Wydawniczy, Warszawa (1998)
2. Rzetelska-Feleszko, E., ed.: Polskie nazwy własne. Instytut Języka Polskiego Polskiej Akademii Nauk, Kraków (2005)
3. Grzenia, J.: Słownik nazw własnych. Wydawnictwo naukowe PWN (2003)
4. Cieślikowa, A., ed.: Mały słownik odmiany nazw własnych. Rytm, Warszawa (2008)
5. Saloni, Z., Gruszczyński, W., Woliński, M., Wołosz, R.: Słownik gramatyczny języka polskiego. Wiedza Powszechna, Warszawa (2007)
6. Thompson, P., Dozier, C.: Name Searching and Information Retrieval. In: Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, Providence, Rhode Island. (1997)
7. Hatzivassiloglou, V., Klavans, J.L., Eskin, E.: Detecting Text Similarity over Short Passages: Exploring Linguistic Feature Combinations via Machine Learning. In: Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. (1999) 203–212

8. Friburger, N., Maurel, D.: Textual similarity based on proper names. In: Proceedings of the Workshop on Mathematical Formal Information Retrieval (MFIR'2002), SIGIR'2002, Tampere, Finland. (2002) 155–167
9. Chena, H.H., Huang, S.J., Ding, Y.W., Tsai, S.C.: Proper Name Translation in Cross-Language Information Retrieval. In: Proceedings of COLING-ACL-1998. (1998) 232–236
10. Pašca, M., Lin, D., Bigham, J., Lifchits, A., Jain, A.: Names and Similarities on the Web: Fact Extraction in the Fast Lane. In: Proceedings of COLING-ACL-2006, Sydney, Australia. (2006) 809–816
11. Grishman, R., Sundheim, B.: Message Understanding Conference - 6: A Brief History. In: Proceedings of COLING-96. (1996) 466–471
12. Sang, E.F.T.K., de Meulder, F.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings of CoNLL-2003, Edmonton, Canada. (2003) 142–147
13. Piskorski, J., Sydow, M., Kupść, A.: 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. (2007)
14. Piskorski, J., Sydow, M.: 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. (2007)
15. Mykowiecka, A., Marciniak, M., Rabiega-Wiśniewska, J.: Proper Names in Polish Dialogs. In: Proceedings of the IIS 2008 Workshop on Spoken Language Understanding and Dialogue Systems, Zakopane, Poland, Springer Verlag (2008)
16. Rabiega-Wiśniewska, J.: Polish proper names in the corpus od spoken dialogues. the syntactic experiment. Beiträge der Europäischen Slavistischen Linguistik (Polyslav). Band 12 (to appear)
17. Savary, A.: Computational Inflection of Multi-Word Units. A contrastive study of lexical approaches. Linguistic Issues in Language Technology **1**(2) (2008) 1–53
18. Tran, M., Maurel, D.: Prolexbase: Un dictionnaire relationnel multilingue de noms propres. Traitement Automatiques des Langues **47**(3) (2006) 115–139
19. Agafonov, C., Grass, T., Maurel, D., Rossi-Gensane, N., Savary, A.: La traduction multilingue des noms propres dans PROLEX. Mϵta **51**(4) (2006) 622–636 Les Presses de l'Université de Montréal.
20. Maurel, D., sko Vitas, D., Krstev, C., Koeva, S.: Prolex: a lexical model for translation of proper names. Application to French, Serbian and Bulgarian. BULAG **32** (2007) 55–72
21. Jacquemin, C.: Spotting and Discovering Terms through Natural Language Processing. MIT Press (2001)
22. Savary, A., Jacquemin, C.: Reducing Information Variation in Text. Lecture Notes in Artificial Intelligence **2705** (2003) 145–181 Springer.
23. Savary, A., Krstev, C., Vitas, D.: Inflectional Non Compositionality and Variation of Compounds in French, Polish and Serbian, and Their Automatic Processing. BULAG **32** (2007) 73–93
24. Karttunen, L., Kaplan, R.M., Zaenen, A.: Two-Level Morphology with Composition. In: Proceedings of COLING-92, Nantes. (1992) 141–148
25. Oflazer, K., Çetonoğlu, Ö., Say, B.: Integrating Morphology with Multi-word Expression Processing in Turkish. In: Second ACL Workshop on Multiword Expressions, July 2004. (2004) 64–71

26. Alegria, I., Ansa, O., Artola, X., Ezeiza, N., Gojenola, K., Urizar, R.: Representation and Treatment of Multiword Expressions in Basque. In: Second ACL Workshop on Multiword Expressions, July 2004. (2004) 48–55
27. Silberztein, M.: NooJ's dictionaries. In: Proceedings of LTC'05, Poznań, Wydawnictwo Poznańskie (2005) 291–295
28. Hajnicz, E., Kupść, A.: Przegląd analizatorów morfologicznych dla języka polskiego. Wydawnictwo IPI PAN, Warszawa (2001)
29. Saloni, Z.: Kategorie gramatyczne liczebników we współczesnym języku polskim. Studia gramatyczne **I** (1977) 145–173 Wrocław.
30. Saloni, Z., Świdziński, M.: Składnia współczesnego języka polskiego. PWN, Warszawa (1998)
31. Świdziński, M., Derwojedowa, M., Rudolf, M.: A Computational Account of Multi-Word Numeral Phrases in Polish. In Kosta, P., Błaszczak, J., Frasek, J., Geist, L., Żygis, M., eds.: Investigations into Formal Slavic Linguistics. Volume 10, part I., Peter Lang (2003) 405–415
32. Tokarski, J.: Schematyczny indeks a tergo polskich form wyrazowych, ed. Zygmunt Saloni. 2 edn. Wydawnictwo Naukowe PWN, Warszawa (2002)
33. Woliński, M.: Morfeusz — a Practical Tool for the Morphological Analysis of Polish. In Kłopotek, M., Wierzchoń, S., Trojanowski, K., eds.: Intelligent Information Processing and Web Mining, IIS:IIPWM'06 Proceedings, Springer (2006) 503–512
34. Woliński, M.: System znaczników morfosyntaktycznych w korpusie IPI PAN. Polonica **XXII–XXIII** (2003) 39–55
35. Przepiórkowski, A., Woliński, M.: The Unbearable Lightness of Tagging: A Case Study in Morphosyntactic Tagging of Polish. In: Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora, EACL 2003. (2003)
36. Saloni, Z.: Klasyfikacja gramatyczna leksemów polskich. Język Polski **LIV** (1974) z.1, 3–13, z.2, 93–101
37. Saloni, Z.: Kategoria rodzaju we współczesnym języku polskim. In: Kategorie gramatyczne grup imiennych we współczesnym języku polskim. Ossolineum, Wrocław (1976) 41–75
38. Bień, J.S., Saloni, Z.: Pojęcie wyrazu morfologicznego i jego zastosowanie do opisu fleksji polskiej (wersja wstępna). Prace Filologiczne **XXXI** (1982) 31–45
39. Bień, J.S.: Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji. Rozprawy Uniwersytetu Warszawskiego t. 383. Wydawnictwa Uniwersytetu Warszawskiego, Warszawa (1991)
40. Saloni, Z.: Czasownik polski. Odmiana, słownik. Wiedza Powszechna, Warszawa (2007)
41. Savary, A.: A formalism for the computational morphology of multi-word units. Archives of Control Sciences **15**(3) (2005) 437–449
42. Savary, A.: MULTIFLEX. User's Manual and Technical Documentation. Version 1.0. Technical Report 285, LI-François Rabelais University of Tours, France (2005)
43. Courtois, B., Silberztein, M., eds.: Les dictionnaires électroniques du français. Larousse, Langue française, vol. 87 (1990)
44. Paumier, S.: Manuel d'utilisation du logiciel Unitex. http://www-igm.univ-mlv.fr/unitex/manuelunitex.ps (2002)
45. Krstev, C.: Processing of Serbian: Automata, Texts and Electronic Dictionaries. Faculty of Philology, University of Belgrade (2009)
46. Krstev, C., sko Vitas, D., Savary, A.: Prerequisites for a Comprehensive Dictionary of Serbian Compounds. LNCS **4139** (2006) 552–563
47. Maurel, D.: Prolexbase. A multilungual relational lexical database of proper names. In: Proceedings of LREC'08, Marrakech, Marocco. (2008)