Intelligent Information Systems 9999
ISBN 666-666-666, pages 1–7

Toposław — a Dictionary Creation Tool

Piotr Sikora and Marcin Woliński

Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

Abstract

Toposław is a Java application for creating inflectional dictionaries of compound toponyms. The tool utilises Multiflex and Morfeusz to describe the inflection. Toposław allows to generate a dictionary of names that are linked with appropriate inflection graphs and objects representing the physical locations. The paper describes the process of linking the names with objects, preparing the lemmas, creating and assigning the inflection graphs and validating the inflection forms.

Keywords: Warsaw toponyms, inflection of compounds, computer dictionary, Multiflex

1 Introduction

We present the program Toposław, a Java application for creating inflectional dictionaries of compound toponyms. Toposław was developed for an ongoing project[1] which aims at creating a computer dictionary of Polish urban proper names. The organisation and the goals of the project have been presented by Marciniak *et al.* 2009. Toposław can be viewed as a general tool for describing inflection of compounds, even though it was created within the mentioned project to create a dictionary toponyms pertaining to the city of Warsaw.

2 The Description of Inflection

The inflection of compound names is modelled in our dictionary using Multiflex (Savary, 2005a,b). Multiflex is a graph-based formalism, which provides tools to describe inflection of compounds by putting restrictions on the possible combinations of inflected constituent words. In the scope of this project we deal with nominal phrases, so our compounds inflect for case and number and some of their constituents agree in gender.

Multiflex requires a cooperating lower-level tool, which handles the inflection of single word constituents. This role is played by Morfeusz SGJP, an analyser and generator of Polish word-forms.

Inflectional patterns are expressed as graphs in Multiflex. The graph nodes correspond to constituents of a compound name's lemma. The paths in the graph represent various inflectional forms of the compound. The nodes are also decorated

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

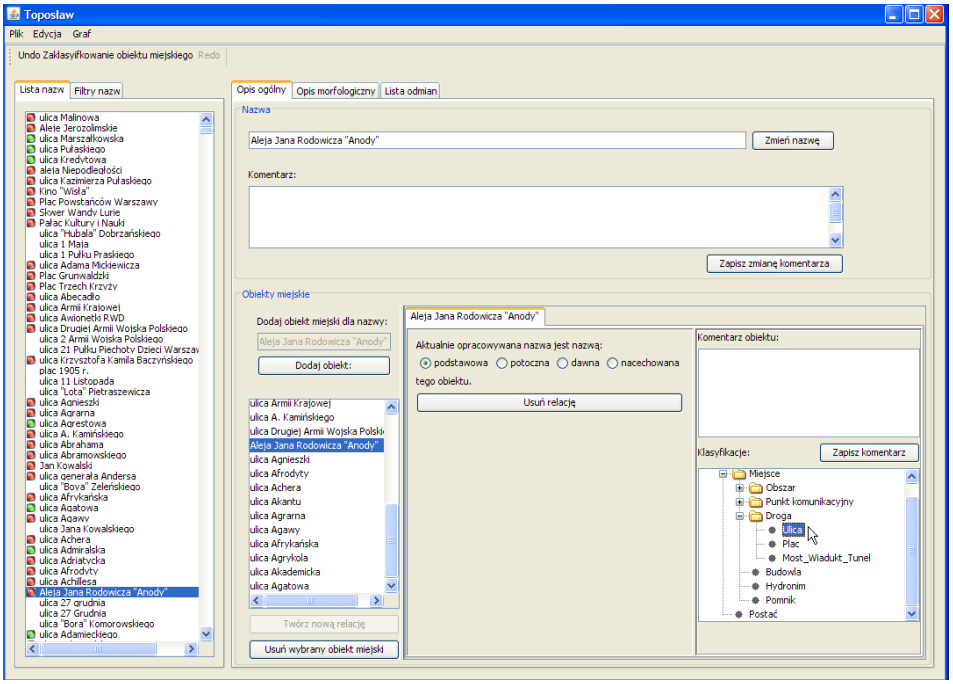

FIGURE 1: The screen for assigning city objects to names.

with equations. These decorations describe the categories for which the given constituent should be inflected.

To describe the inflection of a compound name one needs to label the constituents of its lemma with their morphological features and then to select an inflection graph or to create a new one. A detailed example of this process will be given in section 4.

3 The Structure of the Dictionary

Urban proper names in Polish feature several kinds of variations (cf. Marciniak *et al.*, 2009; Savary *et al.*, 2009), which need to be described. Multi-word names tend to be used in shortened variants, thus long official names are rarely used. We need to encode this information, so the dictionary can be used to generate only the variants that appear in practice.

We should take into account all possible variants of names, so for example what is written as numbers (e.g., *The 43th Regiment*), for applications such as speech recognition needs to be represented by words (*The Forty Third Regiment*). Variants are described within the Multiflex graphs. We label paths in a graph to distinguish the "official", "neutral" (preferred for text generation), and "neutral spoken" (preferred for speech generation) variants.

Another level of complication results from various relationships between names

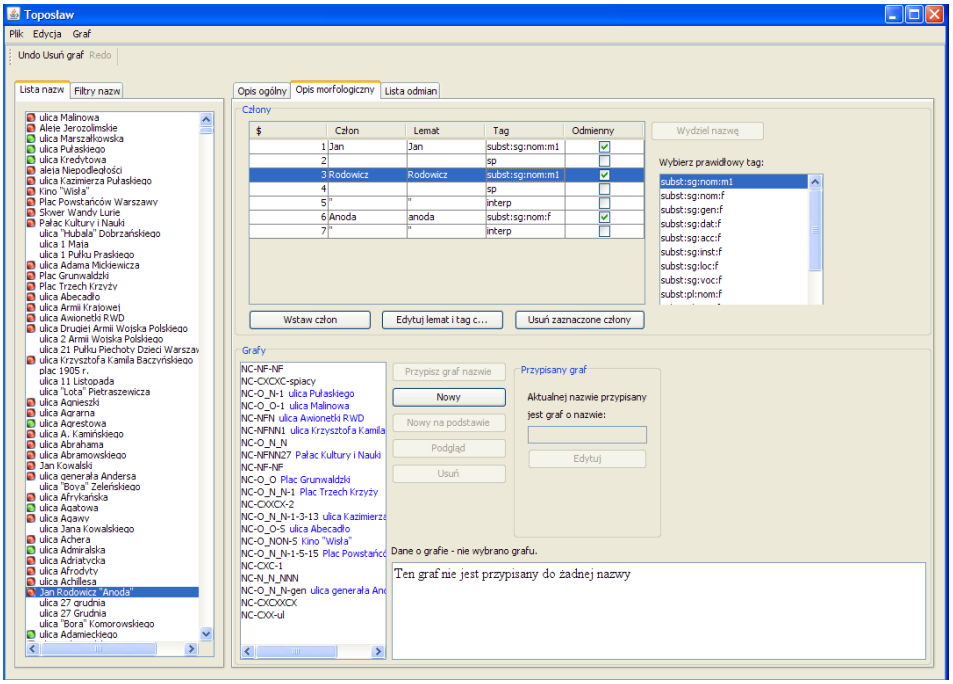

FIGURE 2: A labelled lemma and the list of available graphs

and objects. Names in a city change with time. Some of them have commonly used informal counterparts. In such cases we need to keep track of several names for one place and label them as "former", "common", or "marked". To describe these relations we have introduced the concept of city objects to the dictionary.

Each lexical entry describes one name with all its grammatical forms and variants. Lexical entries are linked with city objects, which represent the real objects referenced by the names. Several names may be linked with a single city object, and many objects with a single name. The latter would apply in a dictionary containing urban names from more than one city.

4 The Process of Describing a Name

Toposław is implemented in Java, which provides powerful tools to create portable, cross-platform, familiar user interfaces and to handle object oriented databases.

The list of names to be described gets loaded to the application's database. The task of the operator is to describe their inflection and provide information about the object referenced by each name.

The list of compound toponyms is presented in the left panel of the program's window (see Fig. 1). From this list, the operator chooses a name or a group of names to describe. A green icon displayed next to a name indicates that the name has a complete description, whereas a red one signals a name which has been

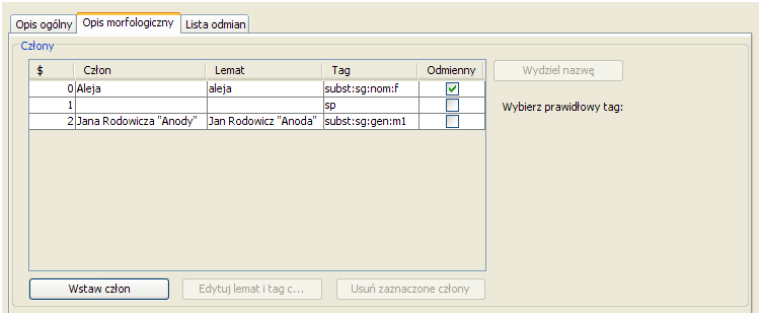

FIGURE 3: The name *Aleja Jana Rodowicza „Anody"* after *Jana Rodowicza „Anody"* has been marked as a sub-compound

described only partially. To make their work easier, the users may filter the list of names according to the stage of their description. This feature is especially convenient when working with a large dictionary. To assign the same inflection graph, it is also possible to select more than one name.

Each name can be linked with a city object referenced by it. Some city objects have several names. Such names (as opposed to variants of one name) are described separately and only then linked with one city object. For example, the names *Plac Thomasa Woodrowa Wilsona (Thomas Woodrow Wilson Square)* and *Plac Komuny Paryskiej (Parisian Commune Square)* name the same object, but the first is its current, official name, while the second is its historical name.

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

City objects are also categorised using a hierarchy, which can be seen in Fig. 1 in the lower right panel, where the example name *Aleja Jana Rodowicza "Anody" (Jan Rodowicz "Anoda" Avenue)* is classified as a PLACE→ROAD→STREET. The full hierarchy has been presented by Marciniak *et al.* (2009).

As explained in section 2 the lemma of the compound has to be labelled with morphological features of the words constituting the compound. This is done on the second screen (cf. Fig. 2). The name is automatically analysed by the morphological analyser Morfeusz. If any of the words has multiple interpretations, the operator has to select the appropriate one as the lemma of the compound.

A compound lemma does not necessarily comprise lemmas of its constituent words. For example *Aleja Jana Rodowicza „Anody"* is a compound lemma where *Aleja* is in a nominative, but the three other tokens represent genitive forms of their respective lexemes. Similarly, names consisting of a nominal head and an adjective will have in their lemma the form of adjective inflected for the gender of the noun.

The operator can also mark a fragment of the name as a sub-compound to be described separately. We use this mechanism for compound names of persons,

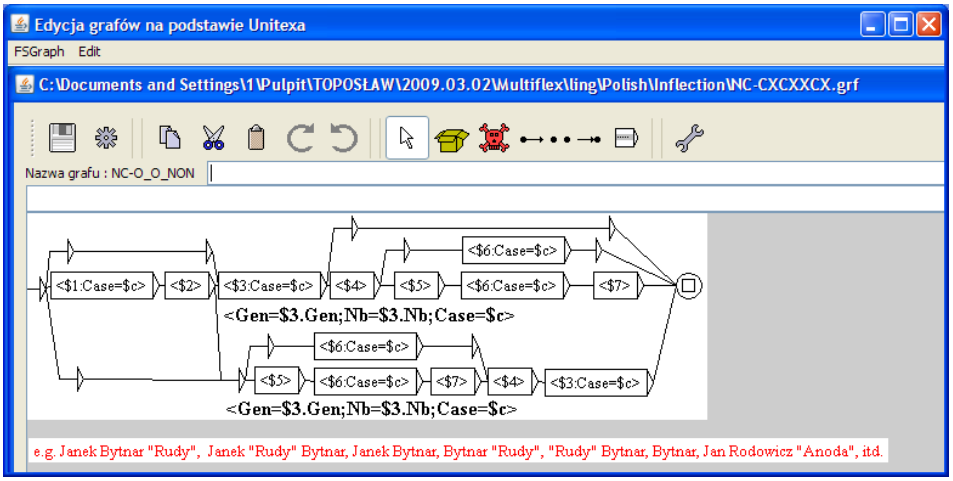

FIGURE 4: Inflection graph for person names of the form *first-name surname „nickname"* (e.g., *Jan Rodowicz „Anoda"*)

FIGURE 5: Inflection graph for names of the form $N\ NP_{gen}$ (e.g., *Aleja Jana Rodowicza „Anody"*)

which tend to occur in several urban names (cf. Fig. 3).

The tool keeps a list of inflection graphs, which can be assigned to names. In Fig. 4 and 5 graphs for the person name *Jan Rodowicz „Anoda"* and the street name *Aleja Jana Rodowicza „Anody"* are presented. The leftmost triangle represents the entry point of the graph, while the circle enclosing a square represents the exit. The numbered boxes correspond to constituents of the name, which can be words, spaces, punctuation or sub-compounds. The arrow-laden lines that connect the boxes represent various paths which can be used while generating the inflected forms of a name. For example paths in Fig. 4 allow for omitting the first name *Jan* and/or the nickname *„Anoda"*, as well as leaving only the nickname.

The formulae inside boxes in Fig. 4 consist of constituents' indexes and equations on morphological variables. These equations impose constraints on the in-

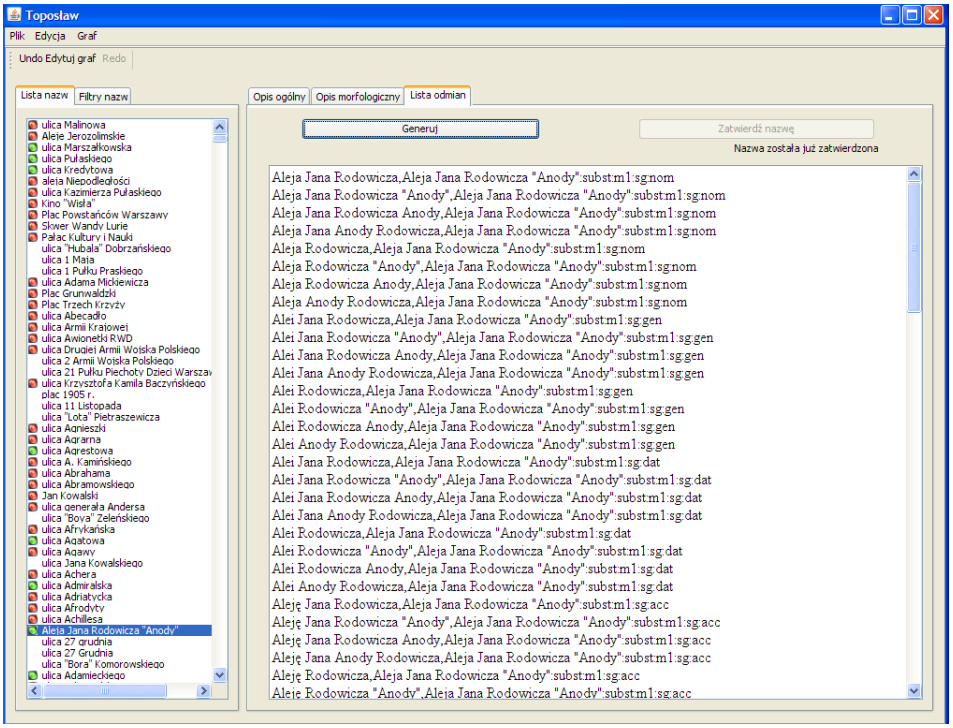

FIGURE 6: The inflected forms of the name *Aleja Jana Rodowicza „Anody"*

flection of constituents. For example, the equation `Case=$c` means that the component inflects for case. The reoccurring variable `$c` means that the respective components must agree in case. The formula out of the boxes determines the features of the inflected forms of the whole compound as a function of the features of its constituents. In Fig. 4 the formula states that the gender and number of the whole compound is inherited from the first component and the case is `$c`. The formula syntax is thoroughly described in (Savary, 2005b).

The editor for graphs is based on Unitex (Paumier, 2006). New graphs can be created from scratch or based on any existing one. The graph editor allows creating, connecting, filling out and deleting boxes. It also generates automatic prefixes to graph names. These prefixes help an operator select the correct graph for a given name. They are based on the operator's decision which constituents are inflected.

Having completed the description, the operator should validate the inflected forms generated by the application. If the validation is positive, the description is marked as completed (cf. Fig. 6).

5 Conclusions and Perspectives

In the paper we describe the first version of Toposław, which has only recently been deployed, so we are currently gathering first reactions from its users. We believe that the tool allows to efficiently build a dictionary of compound inflection taking into account complicated inflection patterns and variants. An important asset for effective work is possibility of marking sub-compounds and describing their inflection and variants once for all compounds in which they occur.

Nonetheless we already see room for improvements. The application should better assist in the process of selecting the right graph for a given compound name. To achieve this goal, it should compare the number of tokens in the lemma with the number of constituents each graph handles. Another improvement would be matching some features of the lemma with those required by appropriate constituents in the graph.

The graph editor could be in a better synergy with the rest of the application. This may be accomplished, e.g., by displaying parts of the lemma next to the corresponding constituents in the graph editor window.

To avoid unnecessary multiplication of graphs, Toposław should detect whether an operator creates redundant graphs to inflect similarly constructed compound names. To perform this task, an algorithm of comparing graphs needs to be devised.

The application could also allow debugging a graph or its assignment to a given name during the review of the name's inflected forms. Colouring the path in the graph corresponding to an inflected form seems a good solution to make this task easier.

References

Małgorzata Marciniak, Joanna Rabiega-Wiśniewska, Agata Savary, Marcin Woliński, and Celina Heliasz (2009), Constructing an Electronic Dictionary of Polish Urban Proper Names, in this volume, pp. ??–??

Sébastien Paumier (2006), Unitex 1.2 User Manual, `http://www-igm.univ-mlv.fr/ unitex`.

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, in *Aspects of Natural Language Processing*, LNCS, Springer.