Automatyczne wspomaganie tworzenia słowników fleksyjnych jednostek wieloczłonowych

Piotr Sikora

IPI PAN

26 stycznia 2015

Przegląd treści

Problem podstawowy

Odmiana jednostek wieloczłonowych: (np.: *ulica Króla Jana II Kazimierza Wazy*)

- omijanie członów,
- np.: *ulica Jana Kazimierza*,
- nietypowe formy: skróty, inicjały,
- np.: *ul. Jana Kazimierza*,
- uzgodnienia lub wręcz przeciwnie,
- np.: *ulicą Jana Kazimierza, ulicy Jana Kazimierza*.

Podejście do rozwiązania

Toposław - narzędzie do tworzenia elektronicznych słowników fleksyjnych jednostek wieloczłonowych.

- Wykorzystuje *Morfeusza SGJP* do analizy morfologicznej i generacji form pojedynczych wyrazów.
- Oparte na formalizmie *Multiflex* Agaty Savary, operującym grafami odmiany jednostek wieloczłonowych.
- Zintegrowany z Toposławem jest edytor ww. grafów z platformy *Unitex*.

Toposław technicznie

- Java.
- Otwarty kod źródłowy.
- Dostępny na licencji GPL 3.
- Dystrybucja na systemy Windows i Linux, 32- i 64-bitowe.

Krótka historia Toposława

Przy wykorzystaniu Toposława powstały:

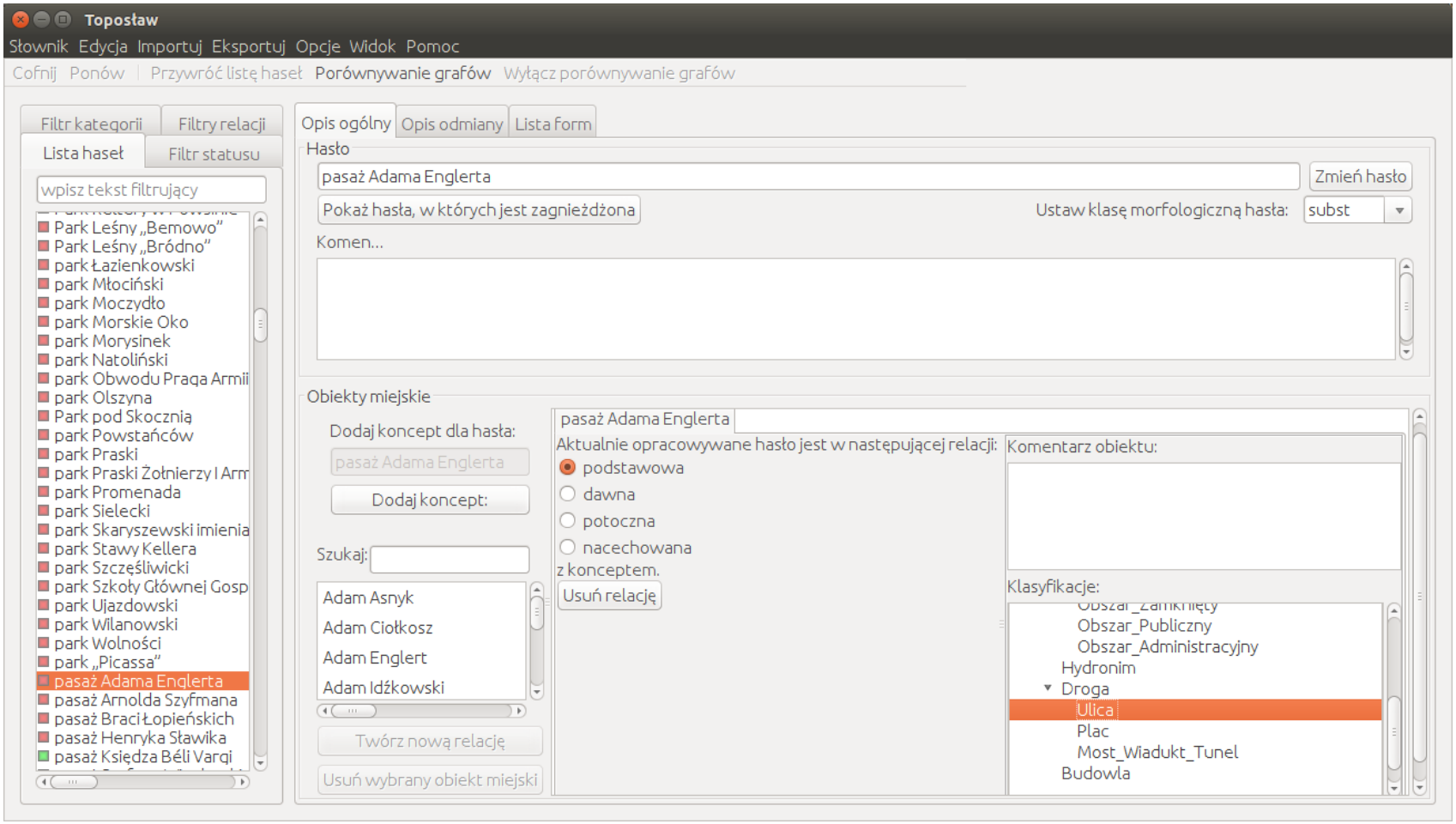
- *SAWA — słownik toponimów warszawskich (ponad 9 tys. haseł),*
- *SEJF — Słownik Elektroniczny Jednostek Frazeologicznych (3200 haseł),*
- *SEJFEK — SEJF z EKonomii (11 212 haseł),*

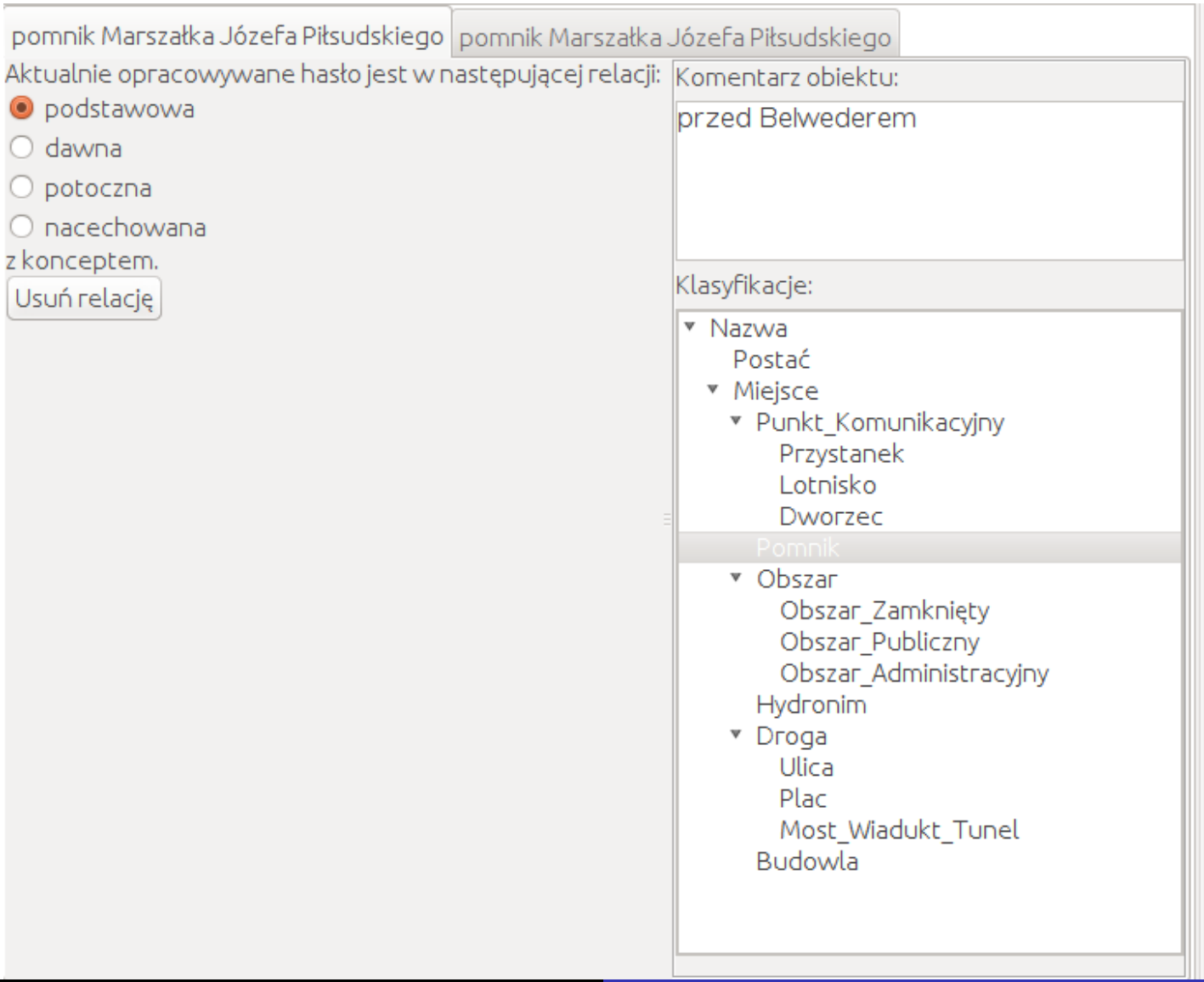
Problemy szczególne Toposława

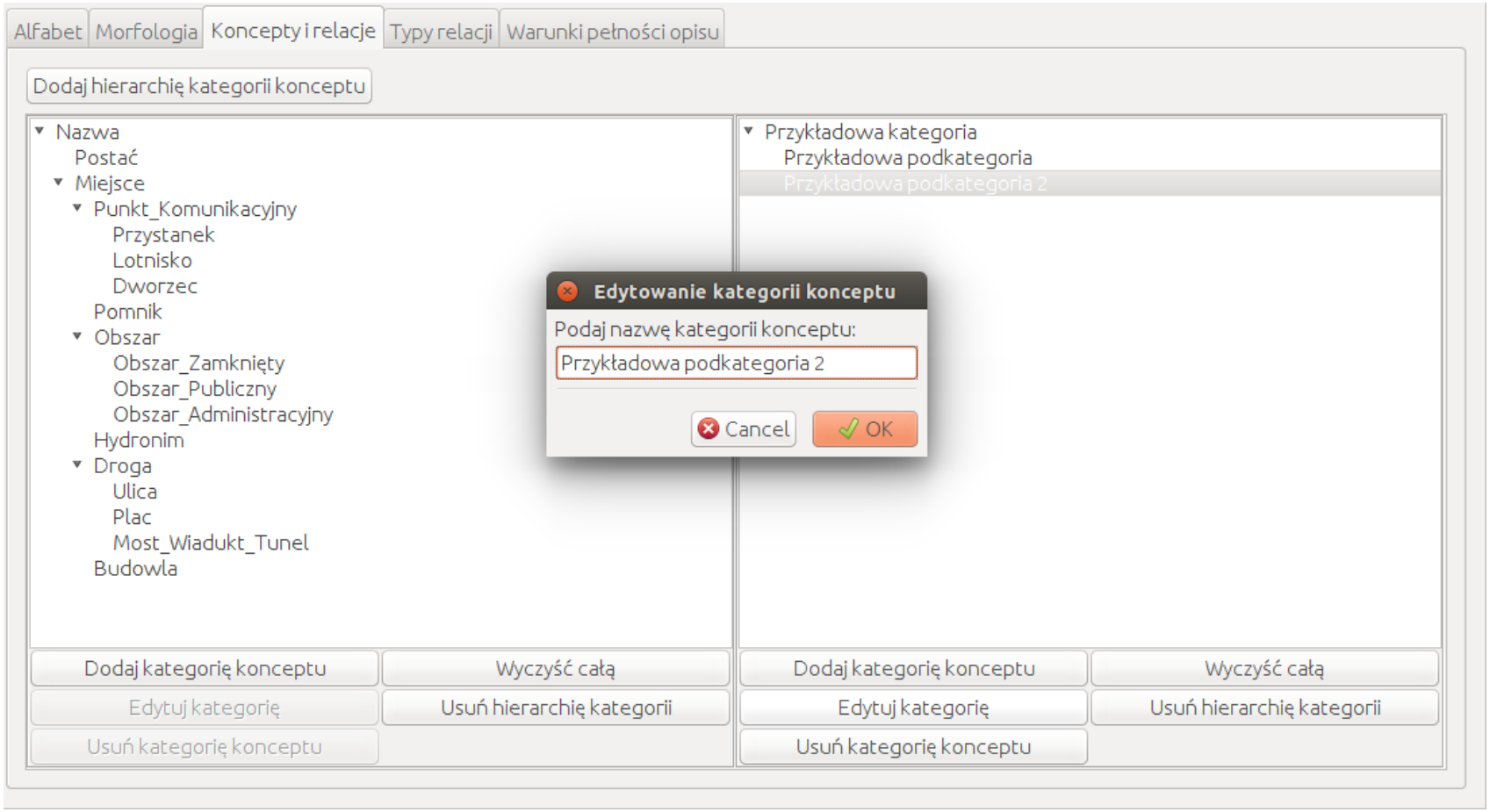
- Silne powiązanie z tematem toponimii.
- Praca leksykografa jest żmudna.
- Wymaga wkładu leksykografa w powstanie słownika - zamiast nadzoru.

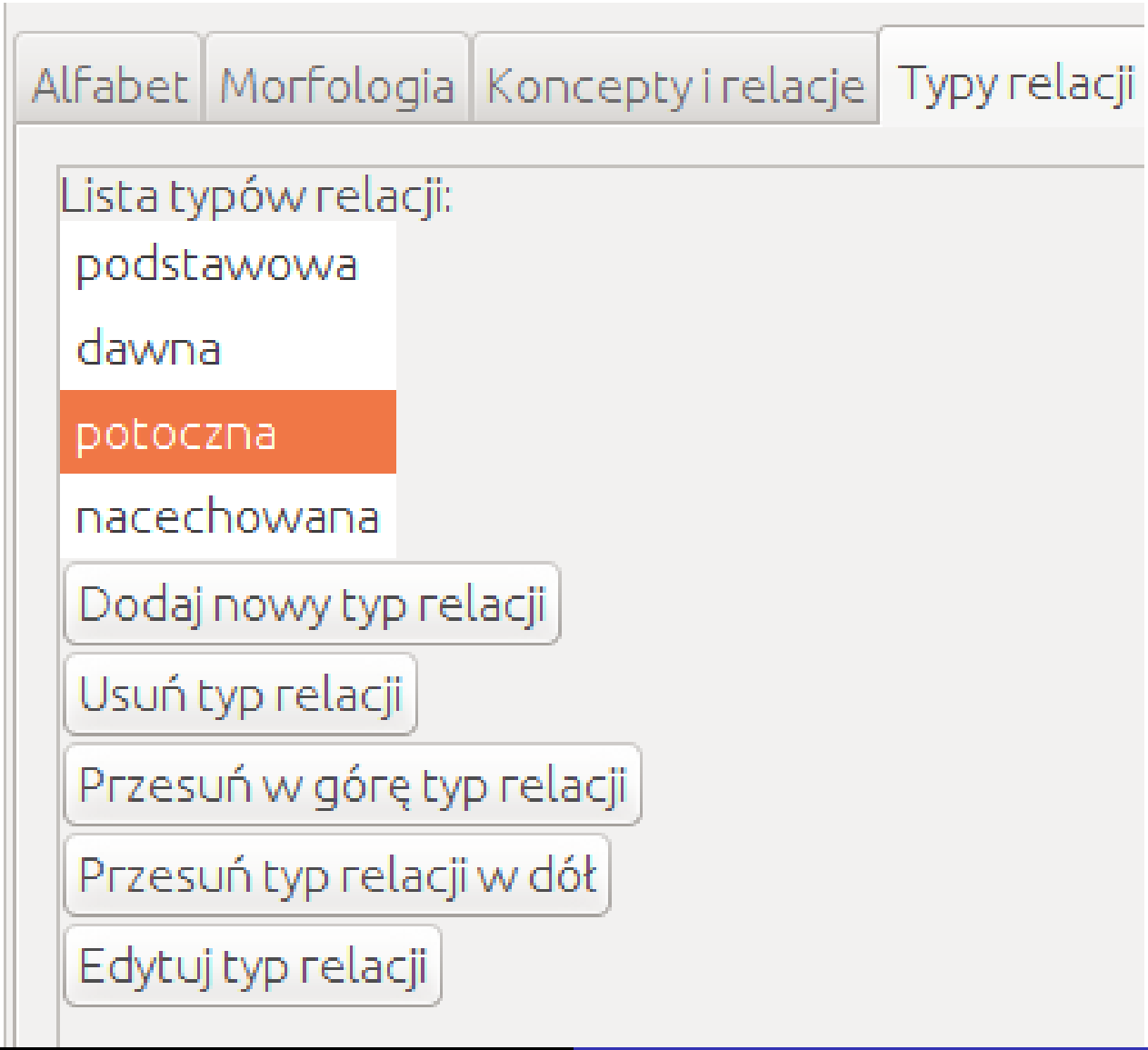
Zarys ogólny

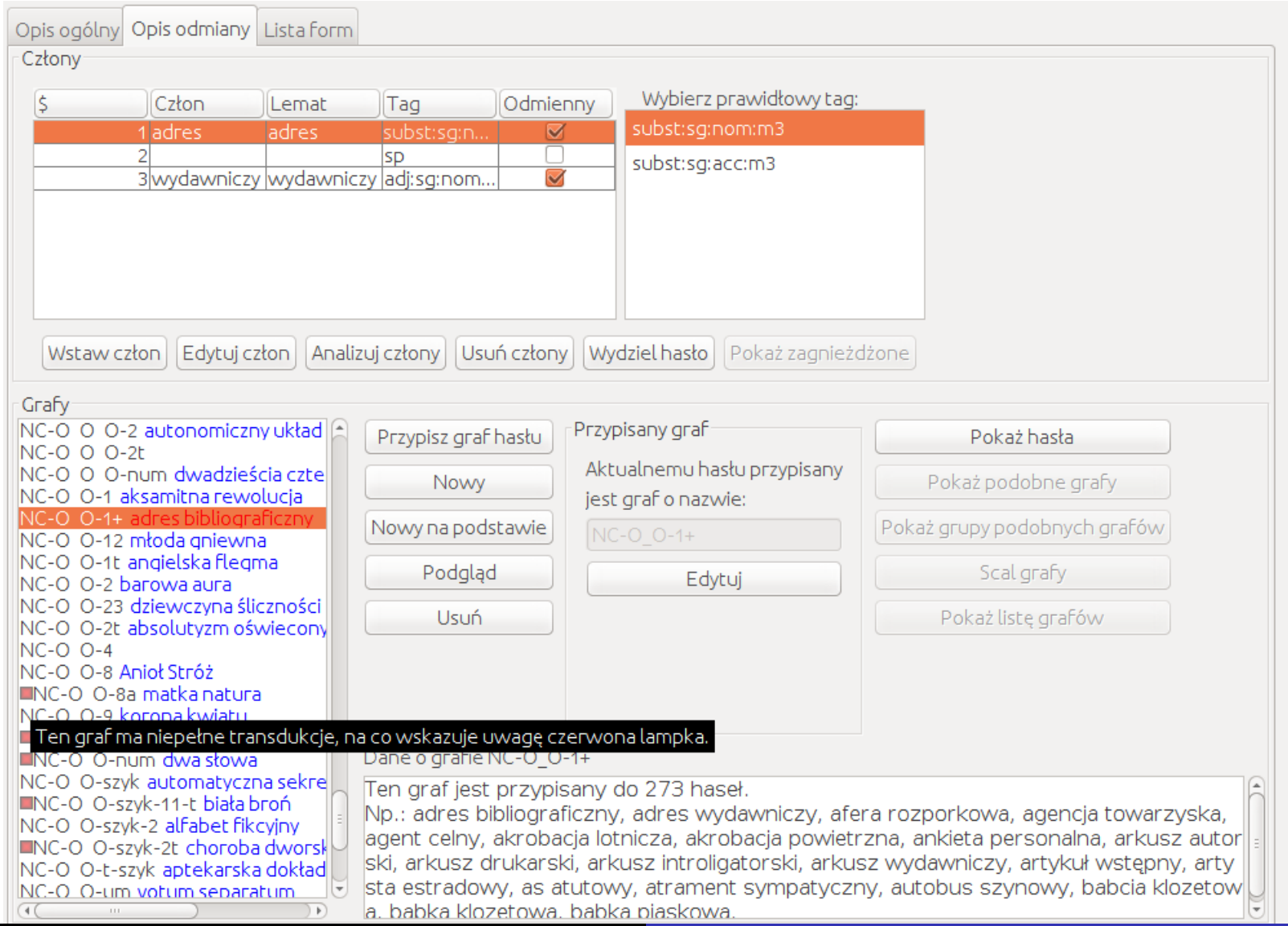
- Słowniki w binarnych plikach.
- Różnorakie inne formaty dla części lub całości danych.
- Warstwa semantyczna.
- Parametryzacja pozwalająca na generalizację lub specjalizację w nowych dziedzinach.

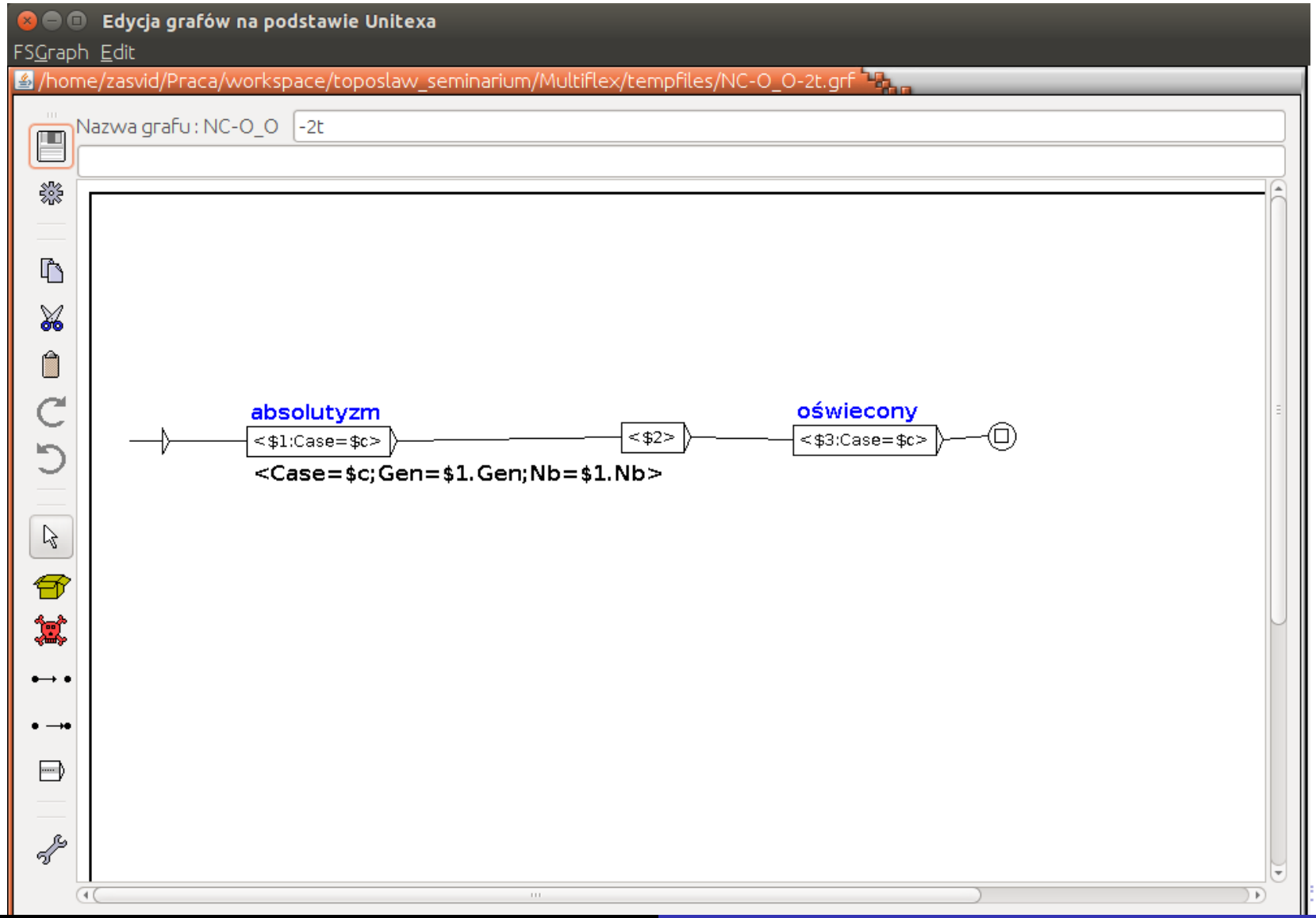

Toposław
Słownik Edycja Importuj Eksportuj Opcje Widok Pomoc
Cofnij Ponów Przywróć listę haseł Porównywanie grafów Wyłącz porównywanie grafów
Filtr kategorii
Filtry relacji
Lista haseł
Filtr statusu
wpisz tekst filtrujący
Park Leśny „Bemowo"
Park Leśny „Bródno"
park Łazienkowski
park Młociński
park Moczydło
park Morskie Oko
park Morysinek
park Natoliński
park Obwodu Praga Armii
park Olszyna
Park pod Skocznią
park Powstańców
park Praski
park Praski Żołnierzy I Arm
park Promenada
park Sielecki
park Skaryszewski imienia
park Stawy Kellera
park Szczęśliwicki
park Szkoły Głównej Gosp
park Ujazdowski
park Wilanowski
park Wolności
park „Picassa"
pasaż Adama Englerta
pasaż Arnolda Szyfmana
pasaż Braci Łopieńskich
pasaż Henryka Sławika
pasaż Księdza Béli Vargi
Opis ogólny
Opis odmiany
Lista form
Hasło
pasaż Adama Englerta
Zmień hasło
Pokaż hasła, w których jest zagnieżdżona
Ustaw klasę morfologiczną hasła:
subst
Komen...
Obiekty miejskie
Dodaj koncept dla hasła:
pasaż Adama Englerta
Dodaj koncept:
Szukaj:
Adam Asnyk
Adam Ciołkosz
Adam Englert
Adam Idźkowski
Twórz nową relację
Usuń wybrany obiekt miejski
pasaż Adama Englerta
Aktualnie opracowywane hasło jest w następującej relacji:
podstawowa
dawna
potoczna
nacechowana
z konceptem.
Usuń relację
Komentarz obiektu:
Klasyfikacje:
Obszar_Publiczny
Obszar_Administracyjny
Hydronim
Droga
Ulica
Plac
Most_Wiadukt_Tunel
Budowla


pomnik Marszałka Józefa Piłsudskiego
pomnik Marszałka Józefa Piłsudskiego
Aktualnie opracowywane hasło jest w następującej relacji:
podstawowa
dawna
potoczna
nacechowana
z konceptem.
Usuń relację
Komentarz obiektu:
przed Belwederem
Klasyfikacje:
Nazwa
Postać
Miejsce
Punkt_Komunikacyjny
Przystanek
Lotnisko
Dworzec
Pomnik
Obszar
Obszar_Zamknięty
Obszar_Publiczny
Obszar_Administracyjny
Hydronim
Droga
Ulica
Plac
Most_Wiadukt_Tunel
Budowla


Alfabet
Morfologia
Koncepty i relacje
Typy relacji
Warunki pełności opisu
Dodaj hierarchię kategorii konceptu
Nazwa
Postać
Miejsce
Punkt_Komunikacyjny
Przystanek
Lotnisko
Dworzec
Pomnik
Obszar
Obszar_Zamknięty
Obszar_Publiczny
Obszar_Administracyjny
Hydronim
Droga
Ulica
Plac
Most_Wiadukt_Tunel
Budowla
Przykładowa kategoria
Przykładowa podkategoria
Przykładowa podkategoria 2
Edytowanie kategorii konceptu
Podaj nazwę kategorii konceptu:
Przykładowa podkategoria 2
Cancel
OK
Dodaj kategorię konceptu
Wyczyść całą
Edytuj kategorię
Usuń hierarchię kategorii
Usuń kategorię konceptu
Dodaj kategorię konceptu
Wyczyść całą
Edytuj kategorię
Usuń hierarchię kategorii
Usuń kategorię konceptu

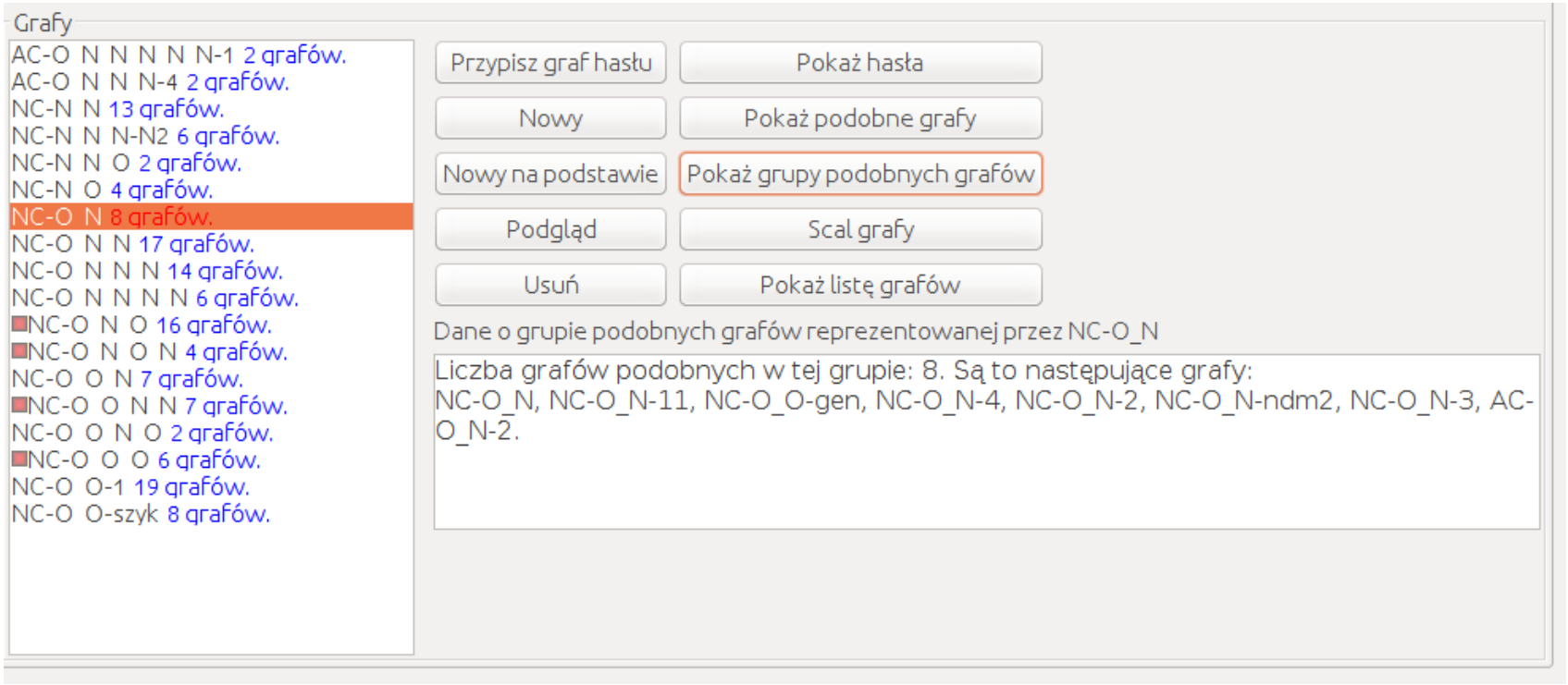

Alfabet
Morfologia
Koncepty i relacje
Typy relacji
Lista typów relacji:
podstawowa
dawna
potoczna
nacechowana
Dodaj nowy typ relacji
Usuń typ relacji
Przesuń w górę typ relacji
Przesuń typ relacji w dół
Edytuj typ relacji

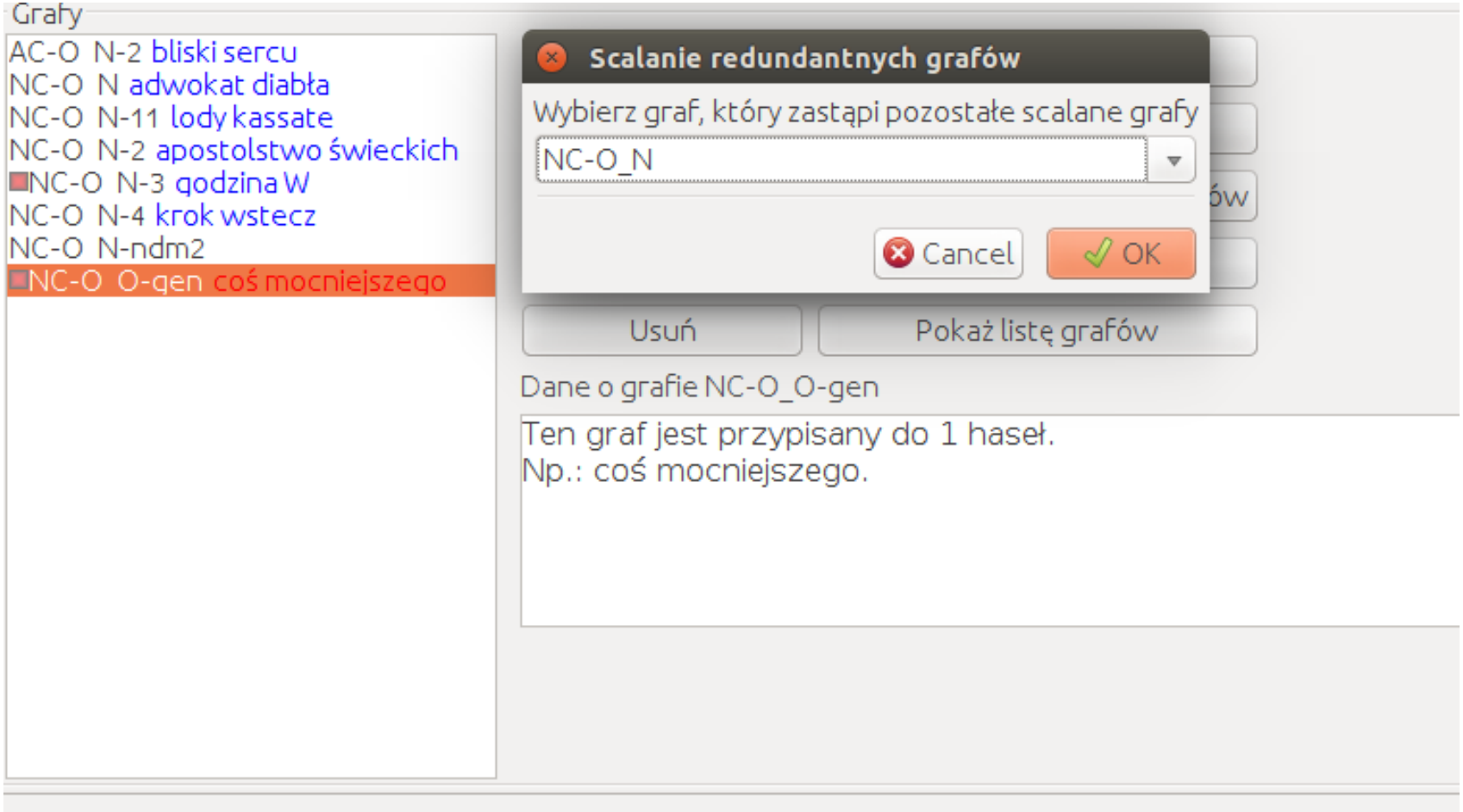

Opis ogólny
Opis odmiany
Lista form
Człony
$
Człon
Lemat
Tag
Odmienny
1 adres adres subst:sg:n...
2 sp
3 wydawniczy wydawniczy adj:sg:nom...
Wybierz prawidłowy tag:
subst:sg:nom:m3
subst:sg:acc:m3
Wstaw człon
Edytuj człon
Analizuj człony
Usuń człony
Wydziel hasło
Pokaż zagnieżdżone
Grafy
NC-O O O-2 autonomiczny układ
NC-O O O-2t
NC-O O O-num dwadzieścia czte
NC-O O-1 aksamitna rewolucja
NC-O O-1+ adres bibliograficzny
NC-O O-12 młoda gniewna
NC-O O-1t angielska flegma
NC-O O-2 barowa aura
NC-O O-23 dziewczyna śliczności
NC-O O-2t absolutyzm oświecony
NC-O O-4
NC-O O-8 Anioł Stróż
NC-O O-8a matka natura
Ten graf ma niepełne transdukcje, na co wskazuje uwagę czerwona lampka.
NC-O O-num dwa słowa
NC-O O-szyk automatyczna sekre
NC-O O-szyk-11-t biała broń
NC-O O-szyk-2 alfabet fikcyjny
NC-O O-szyk-2t choroba dworsk
NC-O O-t-szyk aptekarska dokład
Przypisz graf hasłu
Nowy
Nowy na podstawie
Podgląd
Usuń
Przypisany graf
Aktualnemu hasłu przypisany jest graf o nazwie:
NC-O_O-1+
Edytuj
Pokaż hasła
Pokaż podobne grafy
Pokaż grupy podobnych grafów
Scal grafy
Pokaż listę grafów
Dane o grafie NC-O_O-1+
Ten graf jest przypisany do 273 haseł.
Np.: adres bibliograficzny, adres wydawniczy, afera rozporkowa, agencja towarzyska, agent celny, akrobacja lotnicza, akrobacja powietrzna, ankieta personalna, arkusz autorski, arkusz drukarski, arkusz introligatorski, arkusz wydawniczy, artykuł wstępny, artysta estradowy, as atutowy, atrament sympatyczny, autobus szynowy, babcia klozetowa, babka klozetowa, babka piaskowa.


Edycja grafów na podstawie Unitexa
FSGraph Edit
/home/zasvid/Praca/workspace/toposlaw_seminarium/Multiflex/tempfiles/NC-O_O-2t.grf
Nazwa grafu : NC-O_O
-2t
absolutyzm
<$1:Case=$c>
<$2>
oświecony
<$3:Case=$c>
<Case=$c;Gen=$1.Gen;Nb=$1.Nb>

Porównywanie grafów

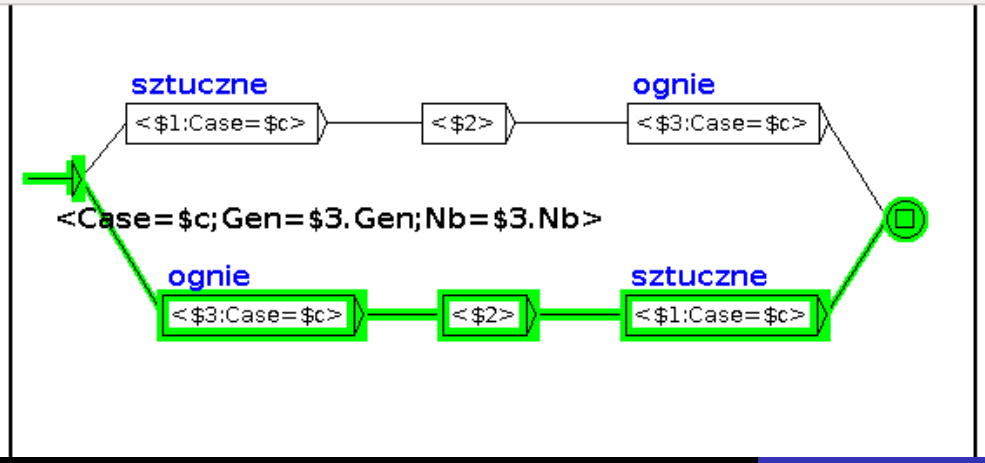
- Ujednolicenie reprezentacji grafów.
- Badanie izomorficzności grafów.
- Badanie ekwiwalencji etykiet wierzchołków.
- Scalanie grafów.
- Porównywanie wyników zastosowania grafów do odmiany haseł.

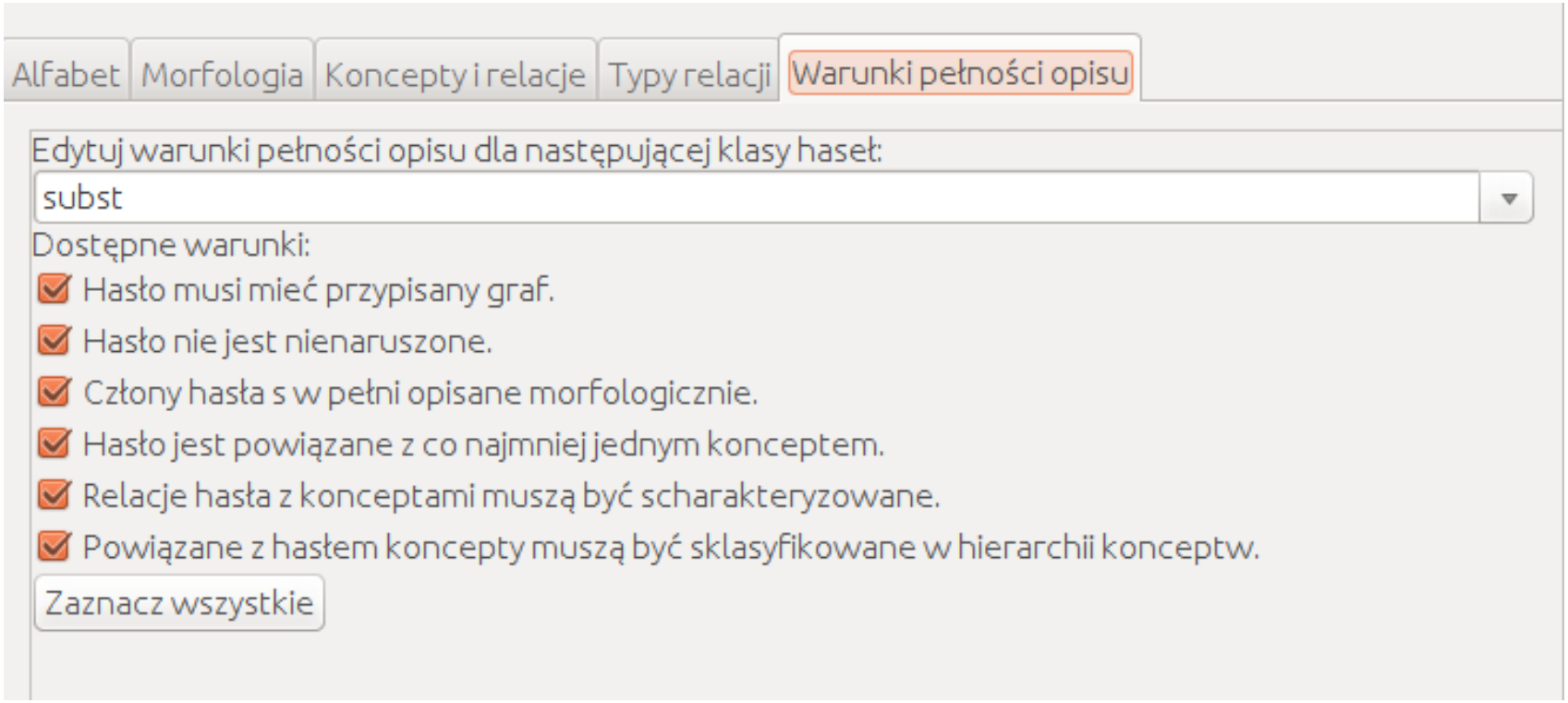

Grafy
AC-O N N N N N-1 2 grafów.
AC-O N N N-4 2 grafów.
NC-N N 13 grafów.
NC-N N N-N2 6 grafów.
NC-N N O 2 grafów.
NC-N O 4 grafów.
NC-O N 8 grafów.
NC-O N N 17 grafów.
NC-O N N N 14 grafów.
NC-O N N N N 6 grafów.
NC-O N O 16 grafów.
NC-O N O N 4 grafów.
NC-O O N 7 grafów.
NC-O O N N 7 grafów.
NC-O O N O 2 grafów.
NC-O O O 6 grafów.
NC-O O-1 19 grafów.
NC-O O-szyk 8 grafów.
Przypisz graf hasłu
Pokaż hasła
Nowy
Pokaż podobne grafy
Nowy na podstawie
Pokaż grupy podobnych grafów
Podgląd
Scal grafy
Usuń
Pokaż listę grafów
Dane o grupie podobnych grafów reprezentowanej przez NC-O_N
Liczba grafów podobnych w tej grupie: 8. Są to następujące grafy:
NC-O_N, NC-O_N-11, NC-O_O-gen, NC-O_N-4, NC-O_N-2, NC-O_N-ndm2, NC-O_N-3, AC-O_N-2.


Grafy
AC-O N-2 bliski sercu
NC-O N adwokat diabła
NC-O N-11 lody kassate
NC-O N-2 apostolstwo świeckich
NC-O N-3 godzina W
NC-O N-4 krok wstecz
NC-O N-ndm2
NC-O O-gen coś mocniejszego
Scalanie redundantnych grafów
Wybierz graf, który zastąpi pozostałe scalane grafy
NC-O_N
Cancel
OK
Usuń
Pokaż listę grafów
Dane o grafie NC-O_O-gen
Ten graf jest przypisany do 1 haseł.
Np.: coś mocniejszego.

Generuj | Zatwierdź opracowanie hasła

sztuczne ognie ◂ sztuczne ognie:subst:pl:nom:m3
sztucznych ogni ◂ sztuczne ognie:subst:pl:gen:m3
sztucznych ogniów ◂ sztuczne ognie:subst:pl:gen:m3
sztucznym ogniom ◂ sztuczne ognie:subst:pl:dat:m3
sztuczne ognie ◂ sztuczne ognie:subst:pl:acc:m3
sztucznymi ogniami ◂ sztuczne ognie:subst:pl:inst:m3
sztucznych ogniach ◂ sztuczne ognie:subst:pl:loc:m3
sztuczne ognie ◂ sztuczne ognie:subst:pl:voc:m3
ognie sztuczne ◂ sztuczne ognie:subst:pl:nom:m3
ogni sztucznych ◂ sztuczne ognie:subst:pl:gen:m3
ogniów sztucznych ◂ sztuczne ognie:subst:pl:gen:m3
ogniom sztucznym ◂ sztuczne ognie:subst:pl:dat:m3




Alfabet
Morfologia
Koncepty i relacje
Typy relacji
Warunki pełności opisu
Edytuj warunki pełności opisu dla następującej klasy haseł:
subst
Dostępne warunki:
Hasło musi mieć przypisany graf.
Hasło nie jest nienaruszone.
Człony hasła s w pełni opisane morfologicznie.
Hasło jest powiązane z co najmniej jednym konceptem.
Relacje hasła z konceptami muszą być scharakteryzowane.
Powiązane z hasłem koncepty muszą być sklasyfikowane w hierarchii konceptw.
Zaznacz wszystkie

- Zintegrowanie z zewnętrznym ekstraktorem złożeń.
- Automatyczna selekcja grafów odmiany.
- Automatyczna konstrukcja grafów odmiany.

- Zintegrowanie z zewnętrznym ekstraktorem złożeń.
- Automatyczna selekcja grafów odmiany.
- Automatyczna konstrukcja grafów odmiany.

- Zintegrowanie z zewnętrznym ekstraktorem złożeń.
- Automatyczna selekcja grafów odmiany.
- Automatyczna konstrukcja grafów odmiany.

Automatyczna selekcja grafów odmiany

- Reguły przypisywania grafów hasłom — jak w *LeXimirze*.
- Automatyczna ekstrakcja reguł z istniejących słowników.
- Dynamiczne tworzenie reguł w miarę postępów leksykografa.

Problemy

- Wiele fleksemów czasownikowych o różnych właściwościach odmiany wymaga wielu haseł w słowniku i wielu grafów odmiany.
- Wielość form trudna do sprawdzenia.
- Wątpliwości co do przyporządkowania np. *niech czytają, będę zbijał, zbijałbym*.

Proponowane rozwiązania

- Superklasy morfologiczne trybów: przypuszczający, oznajmujący, rozkazujący.
- Superkategorie morfologiczne czasów: przeszły, teraźniejszy, przyszły.
- Implementacja rozwiązań w ramach konceptów i relacji.
- Modyfikacja morfologii w słowniku.

Alfabet | Morfologia | Koncepty i relacje | Typy relacji | Warunki pełności opisu

Zapisz | Importuj

```
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, com, sup
Asp: imperf, perf
Neg: aff, neg
Accent: akc, nakc
Postprep : praep, npraep
Accom: congr, rec
Agglt: nagl, agl
Vocal: wok, nwok
<EXTRA_CATEGORIES>
Usage: <E>,offic, neut, spok
<GRAPHICAL_CATEGORIES>
LetterCase: same, all_lower, all_upper, first_upper,first_upper_each_word,no_letter_case,other
Init:<E>,dot,no_dot,dot2,no_dot2,dot3,no_dot3,dot4,no_dot4,dot5,no_dot5
Dot : pun , npun
<CLASSES>
subst: (Nb,<var>),(Case,<var>),(Gen,<fixed>),(Usage,<var>)
depr: (Nb,<fixed>),(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:
```

Bibliografia

- SAVARY, A., RABIEGA-WIŚNIEWSKA, J., WOLIŃSKI, M. (2009): Inflection of Polish Multi-Word Proper Names with Morfeusz and Multiflex, in MARCINIAK, M., MYKOWIECKA, A. (eds.) Ąspects of Natural Language Processing", Lecture Notes in Computer Science 5070, Springer Verlag, pp. 111–141.
- MARCINIAK, M., RABIEGA-WIŚNIEWSKA, J., SAVARY, A., WOLIŃSKI, M., HELIASZ, C. (2009): Constructing an Electronic Dictionary of Polish Urban Proper Names, in Recent Advances in Intelligent Information Systems (Proceedings of the Balto-Slavonic Natural Language Processing Workshop, Kraków), Academic Publishing House EXIT, Warsaw, pp. 743–749.
- GRALIŃSKI, F., SAVARY, A., CZEREPOWICKA, M., MAKOWIECKI, F. (2010): Computational Lexicography of Multi-Word Units: How Efficient Can It Be?, in Proceedings of Multiword Expressions: from Theory to Applications (MWE 2010), Workshop at COLING 2010, Beijing, China, August 28.
- CZEREPOWICKA, M., KOSEK, I. (2011): Problemy opisu związków frazeologicznych w formalizmie „Multifleks” (na przykładzie rodzaju wyrażeń frazeologicznych), in Kopcińska, D., Bańko, M. (eds.) Żóżne formy, różne treści", pp. 117–126, Warszawa 2011.
- CZEREPOWICKA, M. (2011): „Toposław” jako narzędzie znakowania jednostek wieloczłonowych, in Matusiak-Kempa, I., Przybyszewski, S. (eds.) Nowe zjawiska w języku, tekście, komunikacji. Kontekst a komunikacja, Olsztyn, pp. 28–35.
- CZEREPOWICKA, M. (2014): Jednostki obce w słowniku języka polskiego na przykładzie „Słownika elektronicznego jednostek frazeologicznych” (SEJF), in LingVaria (IX), vol. 1 (17), pp. 59-68 [doi: 10.12797/LV.09.2014.17.04].
- CZEREPOWICKA, M. (2014): SEJF - Słownik elektroniczny jednostek frazeologicznych, in Język Polski (XCIV), v. 2, pp. 116-129.

Bibliografia c.d.

- Małgorzata Marciniak, Agata Savary, Piotr Sikora, and Marcin Woliński. Toposław – a lexicographic framework for multi-word units. In Zygmunt Vetulani, editor, Human Language Technology. Challenges for Computer Science and Linguistics: 4th Language and Technology Conference, LTC 2009, Poznań, Poland, November 6–8, 2009, Revised Selected Papers, volume 6562 of Lecture Notes in Artificial Intelligence, pages 139–150. Springer-Verlag, Berlin, 2011.
- Piotr Sikora. Narzędzia do tworzenia elektronicznych słowników fleksyjnych jednostek wieloczłonowych, Uniwersytet Warszawski, Warszawa 2011.
- Piotr Sikora and Marcin Woliński. Toposław — a dictionary creation tool. In Mieczysław A. Kłopotek, Adam Przepiórkowski, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, Recent Advances in Intelligent Information Systems, pages 743–749. Akademicka Oficyna Wydawnicza EXIT, Warsaw, 2009.
- Marcin Woliński, Agata Savary, Piotr Sikora, and Małgorzata Marciniak. Usability improvements in the lexicographic framework Toposław. In Zygmunt Vetulani, editor, Proceedings of the 4th Language and Technology Conference, pages 321–325, Poznań, Poland, 2009.
- Małgorzata Marciniak, Joanna Rabiega-Wisniewska, Agata Savary, Marcin Wolinski, Celina Heliasz. Constructing an Electronic Dictionary of Polish Urban Proper Names, In Recent Advances in Intelligent Information Systems, pages 233–246, Exit, Warszawa 2009.
- Cvetana Krstev, Ranka Stankovic, Ivan Obradovic, Dusko Vitas, Milos Utvic. Automatic Construction of a Morphological Dictionary of Multi-Word Units. Advances in Natural Language Processing, 7th International Conference on NLP, IceTAL 2010, Reykjavik, Iceland, August 16-18, 2010; 01/2010.
- Monika Czerepowicka, Agata Savary. Kodowanie czasowników w Toposławie. Raport techniczny, instrukcja leksykografa.