

**Narzędzia do rozpoznawania
i wizualizacji struktur informacyjnych**

30.05.2016

dr inż. Michał Marcińczuk

michal.marcinczuk@pwr.edu.pl

1. Wstęp

Niniejszy dokument opisuje następujące narzędzia:

1. **g419-spatial** — narzędzie do przetwarzania tekstów w celu rozpoznania i oznaczenia struktur informacyjnych wyrażen przestrzennych. Docelowo narzędzie zostanie zintegrowane z systemem *nlprest*, który obsługuje zdalne przetwarzanie tekstów poprzez wywołania REST. Narzędzie zostało zaimplementowane w taki sposób, aby możliwa była integracja z systemem *nlprest*.
2. **ieviewer** — webowe narzędzie do wizualizacji struktur informacyjnych wyrażen przestrzennych. Jego zadaniem jest przetworzenie

2. Narzędzie g419-spatial

2.1. Realizacja

Narzędzie *g419-spatial* zostało wykonane jako dwa niezależne moduły systemu Liner2: *g419-spatial-core* i *g419-spatial-cli*. Pierwszy moduł, tj. *g419-spatial-core* zawiera niezbędne struktury danych, klasy pomocnicze oraz klasę *SpatialRelationRecognizer* realizującą zadanie rozpoznawania wyrażen przestrzennych (zob. 2.3). Drugi moduł, tj. *g419-spatial-cli*, jest modułem pomocniczym pozwalającym na uruchomienie narzędzia bezpośrednio z linii poleceń (zob. 2.2).

2.2. Uruchamianie narzędzia z linii komend

Do uruchomienia narzędzia w konsoli służy skrypt `spatial-cli`.

```
./spatial-cli pipe
```

Po uruchomieniu skryptu bez parametrów zostanie wyświetlony opis dostępnych trybów:

```
*-----*
* Tools for spatial expression recognition and related
*
* Authors: Michał Marcińczuk (2015)
* Contact: michal.marcinczuk@pwr.wroc.pl
*
*           G4.19 Research Group, Wrocław University of Technology
*
*-----*

[Option error] Missing required options: M, w

usage: ./liner2-cli pipe [options]
  -f,--input_file <filename>      read input from file
  -i,--input_format <format>      input format [iob, ccl, plain, plain:maca,
plain:wcrft, tei,
                                   batch:{format}]
  -M,--malt <filename>            path to maltparser model
  -o,--output_format <filename>  output format [iob, ccl, arff, tokens,
tuples, tei,
                                   batch:{format}]
  -t,--output_file <filename>    path to an output file
  -v,--verbose                    print help
  -w,--wordnet <path>            path to a folder with a wordnet in Princeton
format
```

Przykład wywołania narzędzia w celu przetworzenia dokumentu:

```
./spatial-cli pipe -i tei -o json-frames -f sample_input_tei
```

2.3. Przetworzenie dokumentu z poziomu kodu

Poniższy kod prezentuje wykorzystanie klasy `SpatialRelationRecognizer` w celu przetworzenia dokumentu.

```
package g419.spatial.action;

import org.apache.commons.cli.CommandLine;
import org.apache.commons.cli.DefaultParser;
import g419.corpus.io.reader.AbstractDocumentReader;
import g419.corpus.io.reader.ReaderFactory;
import g419.corpus.io.writer.AbstractDocumentWriter;
import g419.corpus.io.writer.WriterFactory;
import g419.corpus.structure.Document;
```

```
import g419.lib.cli.Action;
import g419.lib.cli.CommonOptions;
import g419.liner2.api.tools.parser.MaltParser;
import g419.spatial.tools.SpatialRelationRecognizer;
import g419.toolbox.wordnet.Wordnet3;

public class ActionPipe extends Action {

    private String inputFilename = null;
    private String inputFormat = null;
    private String outputFilename = null;
    private String outputFormat = null;
    private String maltparserModel = null;
    private String wordnetPath = null;

    public ActionPipe() {
        super("pipe");
        this.setDescription("recognize spatial expressions");
        this.options.addOption(CommonOptions.getInputFileNameOption());
        this.options.addOption(CommonOptions.getInputFileFormatOption());
        this.options.addOption(CommonOptions.getOutputFileFormatOption());
        this.options.addOption(CommonOptions.getOutputFileNameOption());
        this.options.addOption(CommonOptions.getMaltparserModelFileOption());
        this.options.addOption(CommonOptions.getWordnetOption(true));
    }

    @Override
    public void parseOptions(String[] args) throws Exception {
        CommandLine line = new DefaultParser().parse(this.options, args);
        parseDefault(line);
        this.inputFilename = line.getOptionValue(CommonOptions.OPTION_INPUT_FILE);
        this.inputFormat = line.getOptionValue(CommonOptions.OPTION_INPUT_FORMAT);
        this.outputFilename = line.getOptionValue(CommonOptions.OPTION_OUTPUT_FILE);
        this.outputFormat = line.getOptionValue(CommonOptions.OPTION_OUTPUT_FORMAT);
        this.maltparserModel = line.getOptionValue(CommonOptions.OPTION_MALT);
        this.wordnetPath = line.getOptionValue(CommonOptions.OPTION_WORDNET);
    }

    @Override
    public void run() throws Exception {
        Wordnet3 wordnet = new Wordnet3(this.wordnetPath);
        MaltParser malt = new MaltParser(this.maltparserModel);
        SpatialRelationRecognizer recognizer = new SpatialRelationRecognizer(malt, wordnet);

        AbstractDocumentReader reader = ReaderFactory.get().getStreamReader(this.inputFilename,
this.inputFormat);
        AbstractDocumentWriter writer = null;

        if ( this.outputFilename == null ){
            writer = WriterFactory.get().getStreamWriter(System.out, this.outputFormat);
        }
        else{
            writer = WriterFactory.get().getStreamWriter(this.outputFilename, this.outputFormat);
        }

        Document document = null;
        while ( ( document = reader.nextDocument() ) != null ){
            System.out.println("=====");
            System.out.println("Document: " + document.getName());
            System.out.println("=====");
            recognizer.recognizeInPlace(document);
            writer.writeDocument(document);
        }
        writer.close();
        reader.close();
    }
}
```

2.4. Format wejściowy TEI

Dokument wejściowy powinien zawierać następujące dane:

1. Analizę morfologiczną — dane mogą być uzyskane przy pomocy narzędzia *wcrft*¹.
2. Jednostki identyfikacyjne — dane mogą być uzyskane przy pomocy narzędzia *Liner2*².
3. Grupy nominalne — dane mogą być pozyskane przy pomocy narzędzia *Spejd*³ i *gramatyki NKJP*⁴.
4. Frazy składniowe — dane mogą być pozyskane przy pomocy narzędzia *Iobber*⁵.

Powyższe dane mogą być zapisane w formacie TEI. Struktura tego formatu jest opisana na stronie <http://nlp.ipipan.waw.pl/TEI4NKJP/>.

2.5. Format wyjściowy json-frames

Rozpoznane wyrażenia przestrzenne zapisywane są dodawane do dokumentu w postaci struktur *frames*. Struktura *frame* zawiera zbiór atrybutów oraz zbiór nazwanych anotacji. Wymagane atrybuty to identyfikator (*id*) i typ ramy (*type*). Wyrażenia przestrzenne mają przypisany typ "spatial". Zbiór nazwanych anotacji zawiera: "trajector", "spatial_indicator", "landmark" i "region" (opcjonalnie). Każda anotacja jest reprezentowana jako identyfikator anotacji (*id*) oraz zbiór atrybutów. Dla anotacji "trajector" i "landmark" jest to atrybut "sumo", która zawiera listę konceptów SUMO⁶ przypisanych do głowy anotacji.

Format *json-frames* zapisany jest w formacie JSON i zawiera następujące dane:

1 <http://nlp.pwr.wroc.pl/redmine/projects/wcrft/wiki>
2 <https://clarin-pl.eu/dspace/handle/11321/231>
3 <http://zil.ipipan.waw.pl/Spejd>
4 http://clip.ipipan.waw.pl/LRT?action=AttachFile&do=view&target=gramatyka_Spejd_NKJP_1.0.zip
5 <http://nlp.pwr.wroc.pl/redmine/projects/iobber/wiki>
6 <http://www.adampeace.org/OP/>

- lista tokenów — każdy token reprezentowany jest jako tablica następujących elementów: identyfikator, forma ortograficzna, forma bazowa, tag morfologiczny, informacja o istnieniu spacji po tokenie.
- lista anotacji — każda anotacja opisana jest następującymi atrybutami: identyfikator, lista identyfikatorów tokenów wchodzących w skład anotacji, forma tekstowa anotacji.
- lista ram — każda rama opisana jest następującymi atrybutami: identyfikator, typ, lista nazwanych anotacji (para nazwa-anotacja).

Poniżej znajduje się przykładowy dokument w formacie *json-frames*.

```
{ "frames":
  [ { "slots": {
      "debug": {
        "attributes": {
          "schema": "w,we#w1",
          "pattern": "\u003cFirstNG|...|PrePNG\u003e"}},
      "trajector": {
        "attributes": {
          "sumo": "PsychologicalAttribute, CompoundSubstance, CorpuscularObject,
Artifact"},
          "id": "a28"},
      "landmark": {
        "attributes": {
          "sumo": "GroupOfPeople, City"},
          "id": "a24"},
      "region": {},
      "spatial_indicator": { "id": "a14" },
      "id": "x",
      "type": "spatial" } },
    "annotations": [
      { "tokens": ["t1"],
        "id": "a1",
        "text": "Toronto",
        "type": "Noun",
        "category": "word"
      },
      { "tokens": ["t2"],
        "id": "a2",
        "text": "Dominion",
        "type": "Noun",
        "category": "word"
      },
      (...)],
    "tokens": [
      ["t1", "Toronto", "Toronto", "subst:sg:nom:n", "0"],
      ["t2", "Dominion", "dominion", "subst:sg:nom:n", "0"],
      ["t3", "Centre", "centre", "subst:sg:nom:n", "0"],
      ["t4", "Toronto", "Toronto", "subst:sg:nom:n", "0"],
      (...)]
  ]
}
```

3. Narzędzie ieviewer

3.1. Opis

Narzędzie do wizualizacji zostało zaimplementowane w technologii PHP z wykorzystaniem dynamicznych elementów napisanych w JavaScript. Do uruchomienia aplikacji wymagany jest serwer www Apache2, biblioteka PEAR z modułem HTTP_Session2.

3.2. Dostęp

Narzędzie ieviewer dostępne jest pod adresem <http://inforex.clarin-pl.eu/ieviewer>

3.3. Funkcje ieviewer

Narzędzie ievierwer udostępnia następujące funkcje:

1. Widok dokumentu z możliwością wyświetlenia informacji dla każdego tokenu. Informacja o tokenie zawiera: formę ortograficzną, formę bazową, tag morfologiczny, listę anotacji zawierających danych token.
2. Lista ram z możliwością podświetlenia całej ramy i jej składowych w dokumencie.
3. Lista anotacji podzielonych na kategorie (word, ne, group, chunk, mention i undefined).

3.4. Przykładowe zrzuty ekranu

Information Structure Viewer

The screenshot displays the Information Structure Viewer interface. On the left, a sidebar titled "2 Spatial expressions" lists two items:

- grę w USA**: A table with columns for Trajectory, Indicator, and Landmark. The values are "grę", "w", and "USA" respectively. Below the table, it shows the pattern "<FirstNGI>PrepNG" and schema "w.we#w1".
- zapow robot**: A table with columns for Trajectory, Indicator, and Landmark. The values are "TR znajduje się w obrębie granic przestrzeni LM", "#Organization, #Organism, #SocialRole, #GroupOfPeople, #GroupOfAnimals, #Device, #Substance", and "Trajectory: #region," respectively. Below the table, it shows the pattern "MALT" and schema "za#1".

A tooltip is shown over the "zapow robot" entry, displaying the text: "TR znajduje się w obrębie granic przestrzeni LM Landmark: #Artifact, #GeographicArea, #region, #Organization, #Organism, #SocialRole, #GroupOfPeople, #GroupOfAnimals, #Device, #Substance Trajectory: #region,".

On the right, a text document titled "Text (blog/00100598.txt)" contains the following text:

RoboRally czy Wysokie napięcie ? Ponieważ nie mamy ostatnio czasu grać w żadne gry, czas kupić nową. Zawsze jest to jakaś dodatkowa motywacja do ściągnięcia znajomych. Roboty mają kilkanaście lat i pochodzą z USA, Wysokie napięcie jest dużo młodsze, powstało w Niemczech. Do robotów zachęca demo, w którym na próbę można sobie robota zaprogramować, do napięcia dodatkowa plansza z Europą środkową. Dodatkowa oznacza dodatkowy wydatek. Można go uniknąć decydując się na grę w USA lub w Niemczech (dwa największe rynki gier planszowych...). W sumie skoro już zbudowaliśmy w Stanach linie kolejowe , może czas na elektryfikację? Za robotami przemawia zapowiedź kompletnego chaosu, co po dopracowanych do ostatniego szczegółu i uporządkowanych grach niemieckich może być miłą odmianą. m napięciu szans na robienie drugiemu co tobie niemiłe jakby mniej, szczęście też chyba mniejszą rolę. Zamiast laserów atmosferę podgrzewają aukcje surowców, choć są elektrownie, problem zaopatrzenia nie dotyczy.

Ilustracja 1: Tooltip z opisem dopasowanego schematu semantycznego do wyrażenia przestrzennego.

2 Spatial expressions

Move mouse cursor over landmark, region or trajector to see additional information. Click expressions to highlight it in the text.

- **figury w świątyniach**

Trajector	figury
Indicator	w
Landmark	świątyniach

Pattern: <NG|Pact|PrepNG>
 Schema: w.we#1; w.we#w1;
- **koń w Radogoszczu**

Trajector	koń
Indicator	w
Landmark	Radogoszczu

Pattern: <NG|Pact|PrepNG>
 Schema: w.we#w1;
- **pucharze przez posąg Świątowita**

Trajector	pucharze
Indicator	przez
Landmark	posąg Świątowita

Pattern: <FirstNG|_|PrepNG>
 Schema: przez#3a;
- **koń w Arkonie**

Trajector	koń
Indicator	w
Landmark	Arkonie

Pattern: MALT
 Schema: w.we#w1;
- **lasach na polach**

Trajector	lasach
Indicator	na
Landmark	polach

Pattern: <FirstNG|_|PrepNG>
 Schema: na#2; na#na1;
- **posąg Świątowita w Arkonie**

Trajector	posąg Świątowita
Indicator	w

Annotations

Text (popularno_naukowe_i_podr?czniki/00100632.txt)

Click token to see its attributes and annotations.

Długa była droga przeciętnego członka społeczności plemiennej do miejsca kultu czy świątyni. Wynosiła ona w dzisiejszych miarach wiele kilometrów uciążliwej drogi, oderwania od normalnych zajęć. Na to nie każdy i nie na co dzień mógł sobie pozwolić. A z bogami mimo takich czy innych świąt trzeba było mieć codzienny niemal kontakt, trzeba było ciągle się z nimi stykać. Trzeba było składać ofiary, aby ich ubłagać lub im podziękować, trzeba było zasięgnąć wróżby. Wreszcie bogowie to nie tylko figury tkwiące w świątyniach. Ci bogowie tkwili niemal wszędzie. Byli w lasach i na polach, w jeziorach i rzekach, w źródłach, które czasami miały moc leczniczą. Jeśli nie było w pobliżu kapłanów, miejsce ich zajmowali starcy z rodu lub po prostu ojcowie rodzin. Oni składali ofiary i sprawowali wróżby. Wróżby odgrywały w życiu Słowian dużą rolę. Nie tylko bowiem sprawy osobiste jednostek, ale także sprawy państwowe, jak wyprawa wojenna lub pokój, zależały od wyniku wróżb. Dawało to kapłanom jako wróżbitom znaczną władzę do ręki. Najwięcej uwagi badaczy pociągały wróżby z zachowania się koni. Wiadomo mianowicie, że w Szczecinie przy świątyni Trzygłowa hodowano wyłącznie do wróżb konia czarnego, w Arkonie przy świątyni Świątowita – białego. Wyraźnie poświadczony jest też koń wróżący w Radogoszczu. Były one otoczone szczególną czcią i opieką, miały nawet osobnych kapłanów do obsługi (w Arkonie mógł czasem ten kapłan konia dosiadać). Na tych koniach jeździł niekiedy bóg (Świątowit na nocne walki), ale najważniejszą ich funkcją było wróżenie. Kapłan wbijał najpierw w ziemię kilka włóczni parami, oczywiście w specjalny sposób, raczej niejasno opisany przez kronikarzy, a następnie przeprowadzał przez nie konia. Wróżba uchodziła za pomyślną, jeżeli koń przekroczył wszystkie te przeszkody prawą nogą (tak w Arkonie), albo jeżeli żadnej nie potrafił (tak znów w Radogoszczu i Szczecinie). W Radogoszczu wróżba była ważna dopiero wtedy, gdy się udała po raz drugi, a w Arkonie i Szczecinie musiała się udać trzy razy z rzędu. Za te natchnione wysiłki konia pobierała świątynia dziesięcinę z łupów wojennych; za wróżenie prywatne musiała oczywiście zapłacić z własnej kieszeni osoba zapytująca o radę. Trudno przypuścić, aby kapłani zostawili konia bez tresury i polegali tylko na bożym natchnieniu, wypuszczając z ręki tak ważny instrument polityczny. To była, zdaje się, najwyższa klasa wróżb, zwłaszcza zaś wyprawy morskie od nich zależały. Oprócz tego wróżono z napotkanego zwierzęcia (coś jak my z kota), liczone, czy do pary wypadły kreski robione bezmyślnie na popiele rzucono jakieś drewnianka z jednej strony białe, z drugiej czarne (w Arkonie i Szczecinie), jakaś kombinacja zakopywania i rzucania losów wraz z wróżbą z konia była uprawiana w Radogoszczu. Za prognostyk służyło ubywanie, większe lub mniejsze, wina w pucharze trzymanym przez posąg Świątowita w Arkonie, takim samym też prognostykiem urodzaju był kołacz składany mu w ofierze na święcie żniw. Jezioro Głomaczów złowróźnie czasem zmieniało barwę, a z jeziora w Radogoszczu wychodził przed nadchodzącym nieszczęściem ogromny dzik. W innych stronach Słowiańszczyzny nie miały, zdaje się, wróżby tak wielkiego znaczenia, jakkolwiek było ich sporo, sądząc po częstych wzmiankach w źródłach. Wymienia się przeróżne wróżby na Rusi, w Czechach i w Polsce. Synod wrocławski w XIV w. potępia wróżenie z ręki, wosku i ołowiu, ognia i wody. O wiek późniejsze kazanie potępia wróżenie ze snu, z rzucanych losów, z głosów i lotu ptaków, z obserwacji dni, z przyjścia i odejścia ludzi, z ubywania soli.

Ilustracja 2: Widok listy wyrażen przestrzennych (po lewej) i widoku dokumentu (po prawej).

The screenshot shows a software interface with two main panels. The left panel, titled 'Annotations', contains a list of categories under the prefix 'ne'. The categories are: 'Słowian [nam_org_nation]', 'Szczecinie [nam_loc_gpe_city]', 'Trzygłowa [nam_liv_god]', 'Arkonie [nam_loc_gpe_city]', 'Świętowita [nam_liv_god]', 'Radogoszczu [nam_loc_gpe_city]', and 'Arkonie [nam_loc_gpe_city]'. The 'Słowian' category is selected and highlighted in blue. Above the list, there is a yellow instruction box: 'Click a category name to highlight all annotations of the category.' The right panel, titled 'Text (popularno_naukowe_i_podr?czniki/00100632.txt)', contains a yellow instruction box: 'Click token to see its attributes and annotations.' Below this, there is a large block of text in Polish, which is a historical account of the Slavic people and their customs, mentioning various locations like Szczecinie, Trzygłowa, Arkonie, and Radogoszczu, and figures like Świętowit.

Ilustracja 3: Widok listy anotacji podzielonych na kategorie (po lewej).

