Ekstrakcja informacji z tekstów mammograficznych

Anna Kupść, Małgorzata Marciniak

(aniak, mm)@ipipan.waw.pl

Instytut Podstaw Informatyki PAN

Zespół Inżynierii Lingwistycznej

Plan seminarium

- Przedstawienie celu pracy.
- Omówienie architektury systemu.
- Kilka informacji na temat systemu SProUT.
- Omówienie elementów systemu.
- Przykład działania.
- Prezentacja programu.

O dziedzinie

- Projekt we współpracy z p. **Teresą Podsiadły-Marczykowską**, IBIB
- Dane: około 2000 raportów z 3 źródeł
- Na podstawie szczegółowej ontologii (IBIB) powstała ontologia uproszczona. Posiada 57 atrybutów w tym:
  - ogólne — 14
  - opis budowy (utkanie) — 12
  - opis zmiany — 26
  - lokalizacja — 5

Przykładowe teksty

- 285
Sutki o resztkowym utkaniu gruczołowym z przewagą tkanki tłuszczowej. Zmian ogniskowych podejrzanych o złośliwość nie wykazano. W sutku lewym wewnątrzsutkowe węzły chłonne (prawidłowe).
- 18690
Sutki o utkaniu z przewagą tkanki tłuszczowej. W sutku lewym w KGZ zagęszczenie odpowiadające skupisku resztkowej tkanki gruczołowej. Zmian podejrzanych o złośliwość nie wykazano. Doły pachowe w normie. Kontrolna mammografia za rok.

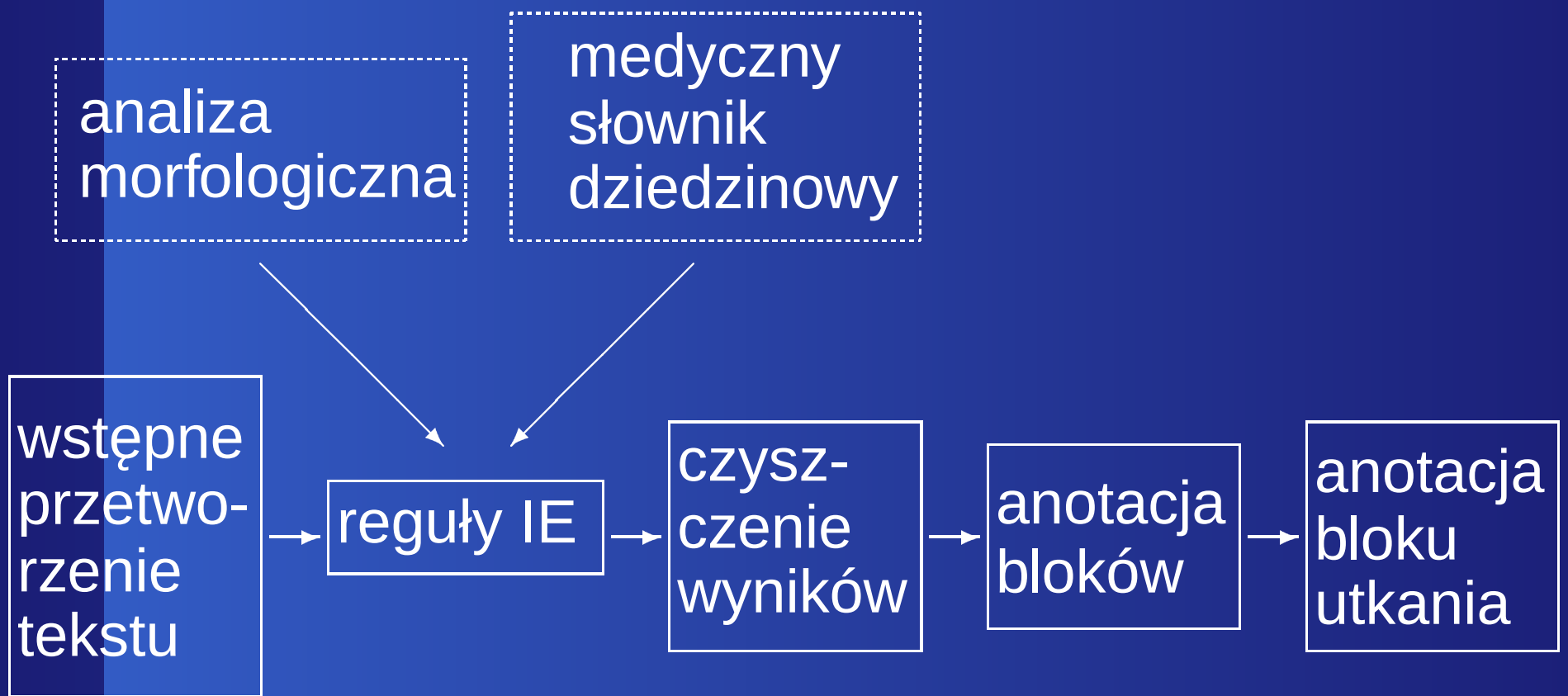
Przykładowe teksty

- Badanie: MAMMOGRAFIA.
  Identyfikator badania: 26321
  Identyfikator pacjenta: 38499
  Data badania: 2001-12-07
  Rozpoznanie:
  Opis:
  Sutki o resztkowym utkaniu gruczołowym z przewagą tkanki tłuszczowej. W sutku prawym w KGZ widoczne owalne, słabo wysycone, dobrze ograniczone zagęszczenie o wym. 15x10mm odpowiadające zmianie łagodnej - najpewniej torbieli. Zmian podejrzanych o złośliwość nie wykazano. Doły pachowe wolne. Kontrolna mammografia za rok.

Teksty

- Specyficzne dla dziedziny słownictwo.
- Wiele skrótów:
  - *ww, w.- chł* — węzły chłonne
  - *o śr. ok.10mm*
  - *kl. piersiowej*
- Wiele błędów:
  - Nagminne pomijanie znaków diakrytycznych.
  - Niestandardowe użycia znaków przestankowych:
    - *z dnia 31,01,2000r.*
    - *trzy zagęszczenia o śr. 15,10 i 5mm*
  - i takie teksty: *o śr IOmm*

Architektura Systemu



SProUT

- **S**hallow Text **Pro**cessing with **U**nification and **T**yped Feature Structure, opracowany w DFKI Saarbrücken,
- dostosowany do współpracy z 11 językami w tym z polskim, wykorzystuje analizator morfologiczny Morfeusz,
- połączenie technik automatów skończonych z formalizmem unifikacyjnym,
- ścisła kontrola typów.

SProUT cd

Umożliwia:

- odwoływanie się do innych reguł gramatycznych: operator `@seek`,
- koreferencję zmiennych: `#z1`,
- tworzenie słownika dziedzinowego (*gazetteer*),
- definiowanie własnych funkcji dołączanych bezpośrednio do SProUT'a.

Nasza gramatyka ma 167 reguł.

Przykład reguły

(1)

```
t_majority:>
 (@seek(loc) & [LOC #loc])?
 (morph & [POS prep, SURFACE ‘o’,
           INFL infl_prep & [CASE_PREP #c1]]
  @seek(utk_tkan)& [C #c1])?
 (morph & [POS prep, SURFACE ‘z’])?
 (morph & [STEM ‘przewaga’] |
  morph & [STEM ‘przeważać’])
 (@seek(utk_tkan)& [C gen])?
 (gazetteer & [GTYPE gaz_med_utkanie,
               G_CONCEPT #typ])
-> btiss_str & [BTISSUE gl & #typ, LOC #loc].
```

Funkcje

- Wykorzystujemy do przetwarzania wszelkich napisów nierozpoznawanych przez Morfeusza np: 7x10mm
- struktura wejściowa:

token
SURFACE *'7x10mm'*
TYPE *other_symbol*

- `xvalues :/ token & [TYPE other_symbol, SURFACE #s] -> #fs, where #fs=FindX(#s).`
- rozpoznane wymiary:

size_str
NUM1 *'7'*
NUM2 *'10'*
NUM3 *string*
DIM *mm*

Słownik dziedzinowy

```
resztkowej | GTYPE:gaz_med_ilosc| G_CONCEPT:rem |
    G_CASE:loc | G_NUMBER:singular | G_GENDER:fem

resztkową | GTYPE:gaz_med_ilosc | G_CONCEPT:rem |
    G_CASE:ins | G_NUMBER:singular | G_GENDER:fem
torbielowato-włóknista |GTYPE:gaz_med_budowa_piersi |
    G_CONCEPT: dyspl-cyst-fib | G_CASE:nom | G_NU...
włóknista |GTYPE:gaz_med_budowa_piersi |
  G_CONCEPT: mastopat-fibr| G_CASE:nom | G_NU...
włóknistej|GTYPE:gaz_med_budowa_piersi |
  G_CONCEPT: mastopat-fibr | G_CASE:gen | G_NU...
mastopatyczne |GTYPE:gaz_med_budowa_piersi |
  G_CONCEPT: mastopat | G_CASE:nom | G_NUMB...
mastopatycznych |GTYPE:gaz_med_budowa_piersi |
  G_CONCEPT: mastopat | G_CASE:gen_loc |...
```

Zmiany

- ANAT_CHANGE: guz, guzek, guz dysplastyczny, guzek dysplastyczny, zacienienie, zagęszczenie, zwapnienia: mikro, makro, zwapnienie, zaburzenie architektury.
- INTERPRETATION: wewnątrzsutkowy węzeł chłonny, torbiel: olejowa, tłuszczowa, rak: neo-mal, meta, tłuszczak, gruczolako-włókniak, włókniak, blizna, struktura promienista, skupisko tkanki . . .
- SHAPE: okrągły, okrągławy, owalny, miseczkowaty, nieregularny, muszelkowaty, linijny, . . .
- CONTOUR: gładki, zatarty, nieco zatarty, spikularny. . . .

Atrybuty lokalizacji

BODY_PART	część ciała, np. pierś, dół pachowy
L_R	lateralizacja: lewa lub prawa
LOC_A	anatomiczna część, np. brodawka sutkowa, węzeł chłonny
LOC_CONV	lokalizacja konwencjonalna: KDW, kwadranty górne-wewnętrzne itd.
LOC_CONV1	lokalizacja konwencjonalna wskazująca głębokość np. przy klatce piersiowej

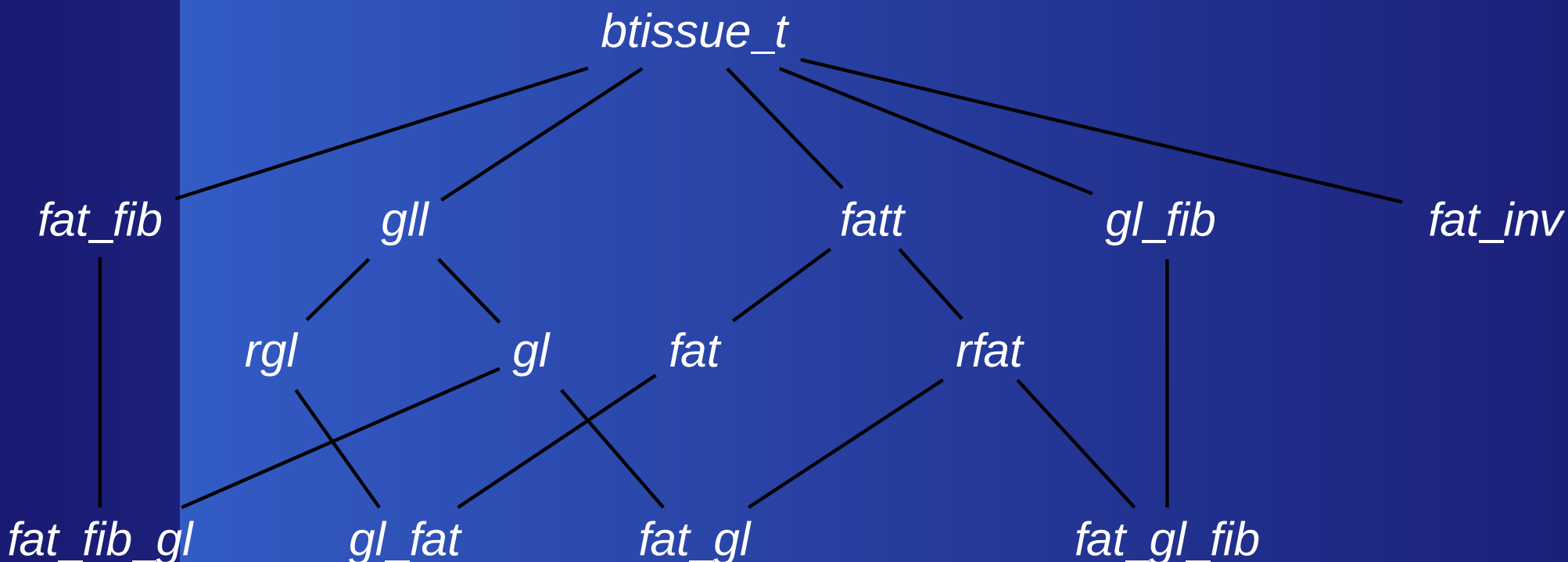
Lokalizacja

- Następującym tekstom:
  - Sutek prawy - w KDW
  - W kdw sutka prawego
  - w sutku prawym, w kwadrancie dolnym-wewnętrznym
- Odpowiada struktura:

```
LOC|BODY_PART:breast <--pierś
   |LOC_CONV:liq <--kwadrant dolny wewnętrzny
                      lower inner quadrant
   |L_R:right <--prawa strona
```

Hierarchia typów utkania

Wartość atrybutu BTISSUE



Przykłady fraz określających utkanie

- Sutki o przewadze tkanki tłuszczowej.
- Sutki z przewagą utkania tłuszczowego.
- Przewaga tkanki tłuszczowo-włóknistej.
- Sutki o utkaniu z przewagą tłuszczowego.
- W sutkach przeważa utkanie tłuszczowo-włókniste.
- Sutki o budowie z przeważającą tkanką tłuszczową.
- . . .

Działanie reguły

(2)

```
t_majority:>
 (@seek(loc) & [LOC #loc])?
 (morph & [POS prep, SURFACE ‘o’,
           INFL infl_prep & [CASE_PREP #c1]]
  @seek(utk_tkan)& [C #c1])?
 (morph & [POS prep, SURFACE ‘z’])?
 (morph & [STEM ‘przewaga’] |
  morph & [STEM ‘przeważać’])
 (@seek(utk_tkan)& [C gen])?
 (gazetteer & [GTYPE gaz_med_utkanie,
               G_CONCEPT #typ])
-> btiss_str & [BTISSUE gl & #typ, LOC #loc].
```

Wyznaczanie bloków

- Etykietowanie istotnych linii: ANAT_CHANGE — a_ch, INTERPRETATION — i_ch, BTISSUE — ut, diagnozy lub zalecenia — rp.
- Idąc od góry szukamy a_ch, i_ch lub ut,
- Idąc w góre staramy się przyłączyć linie zawierające odpowiednie rodzaje atrybutów.
- Robimy to samo idąc w dół
- Kończymy wyznaczanie bloku jeśli pojawi się dloc lub atrybut unikalny w bloku (dla zmiany są to np. lokalizacja, kształt, wysycenie, . . .).
- Po podziale na bloki sprawdzamy czy nie zostały lokalizacje bez przydziału i ew. poprawiamy podział.

Przykład wyznaczania bloków

Sutek prawy – w kwadrancie górnym zagęszczenie dobrze wysycone o średnicy około 20 mm i zatartych granicach. Wymaga dalszej diagnostyki – konieczne wykonanie badania USG i PCI. Wewnątrzsutkowy węzeł chłonny w kwadrancie górno-zewnętrznym sutka lewego.

```
zp
 LOC|BODY_PART:breast||LOC|LOC_CONV:uq||LOC|L_R:right
 ANAT_CHANGE:density||GRAM_MULT:singular
 DIM:mm||NUM1:20||NUM2:20
 CONTOUR:obscured
 REC:rec_list||REC|FIRST:pci||REC|REST|FIRST:usg
 INTERPRETATION:intr_lymph_node <--wstępnie źle
zk              |
zp              V
 INTERPRETATION:intr_lymph_node <--po poprawie granic
 LOC|BODY_PART:breast||LOC|LOC_CONV:uoq||LOC|L_R:left
zk
```

Jak dzielimy blok utkania

- Etykietowanie bloku:
  - etykiety lokalizacji `loc, log, lsz`,
  - jeśli atrybuty GLAND to etykieta `gland`,
  - etykiety utkania `tog`, `tsz`.
- Blok dzielony jest od końca
  - dla `lsz` szukamy najbliższego `log` i kopiujemy je — rodzaj uproszczonej unifikacji lokalizacji,
  - jeśli `tog` to szukamy najbliższego `loc` lub `log`,
  - jeśli `tsz` lub `gland` to tworzymy blok opisujący tkankę gruczołową i szukamy najbliższej lokalizacji, którą dołączmy lub kopiujemy,
  - jeśli lokalizacja to łączymy ją z najbliższą informacją oznaczoną: `gland, tog, tsz`.

Przykład bloku utkania

Sutki o resztkowym utkaniu gruczołowym w kwadrantach górno-zewnętrznych. Przewaga tkanki tłuszczowej.

```
up
log    LOC|BODY_PART:breast||LOC|L_R:left-right
gland  GLAND|QUANT:rem
tsz    BTISSUE:gll
lsz    LOC|LOC_CONV:uoq
tog    BTISSUE:fat_gl
uk
```

Podział bloku utkania

```
up
log    LOC|BODY_PART:breast||LOC|L_R:left-right
utp
gland GLAND|QUANT:rem
tsz    BTISSUE:gll
+      LOC|BODY_PART:breast||LOC|L_R:left-right
lsz    LOC|LOC_CONV:uoq
utk
utp
+      LOC|BODY_PART:breast||LOC|L_R:left-right
tog    BTISSUE:fat_gl
uk
```

Końcowa ocena raportu

- MMG_REL — czy mammografia jest wiarygodna:
  - *reliable* — tkanka z przewagą tłuszczowej,
  - *avg_reliable* — tkanka z przewagą gruczołowej,
  - *unreliable*:
    - tkanka bardzo gęsta gruczołowa,
    - tkanka o charakterze dysplastycznym,
    - jawnie napisane, że mmg trudne do oceny.
- REPORT_CLASS — wybierana jest z raportu najgorsza diagnoza i dodatkowo uwzględniane są zalecenia. Jeśli zalecane wykonanie biopsji to zmiana podejrzana, jeśli konsultacja onkologa to sugestia zmiany złośliwej.
- REPORT_WITH_FINDINGS — czy wykryto zmiany.

Przykładowe badanie

Badanie: MAMMOGRAFIA.
Identyfikator badania: 26321
Identyfikator pacjenta: 38499
Data badania: 2001-12-07
Rozpoznanie:
Opis:
Sutki o resztkowym utkaniu gruczołowym z przewagą tkanki tłuszczowej. W sutku prawym w KGZ widoczne owalne, słabo wysycone, dobrze ograniczone zagęszczenie o wym. 15x10mm odpowiadające zmianie łagodnej - najpewniej torbieli. Zmian podejrzanych o złośliwość nie wykazano. Doły pachowe wolne. Kontrolna mammografia za rok.

Wyniki po anotacji bloków

```
bp
EXAM_ID:26321||PATIENT_ID:38499
up
LOC|BODY_PART:breast||LOC|L_R:left-right
BTISSUE:fat_gl||GLAND|QUANT:rem
uk
zp
LOC|BODY_PART:breast||LOC|LOC_CONV:uoq||
                                LOC|L_R:right
ANAT_CHANGE:density||CONTOUR:circumscribed||
  GRAM_MULT:singular||SATURATION:low||SHAPE:oval
DIM:mm||NUM1:15||NUM2:10
DIAGNOSIS_RTG:benign
INTERPRETATION:cyst
DIAGNOSIS_RTG:no_susp
zk
DIAGNOSIS_RTG:no_susp||LOC_D|BODY_PART:armpit||
                              LOC_D|L_R:left-right
rp
RECOMMENDATION|FIRST:mmg||TIME|DESCRIPTOR:rok
bk
========================
```

```
bp
    EXAM_ID:26321||PATIENT_ID:38499
up
log       LOC|BODY_PART:breast||LOC|L_R:left-right
utp
log       LOC|BODY_PART:breast||LOC|L_R:left-right
tog       BTISSUE:fat_gl
utk
utp
log       LOC|BODY_PART:breast||LOC|L_R:left-right
gland     GLAND|QUANT:rem
utk
uk
zp
    LOC|BODY_PART:breast||LOC|LOC_CONV:uoq||
                                LOC|L_R:right
    ANAT_CHANGE:density||CONTOUR:circumscribed||
      GRAM_MULT:singular||SATURATION:low||SHAPE:oval
    DIM:mm||NUM1:15||NUM2:10
    DIAGNOSIS_RTG:benign
    INTERPRETATION:cyst
    DIAGNOSIS_RTG:no_susp
zk
    DIAGNOSIS_RTG:no_susp||LOC_D|BODY_PART:armpit||
                              LOC_D|L_R:left-right
rp
    RECOMMENDATION|FIRST:mmg||TIME|DESCRIPTOR:rok
    MMG_REL:reliable
    REPORT_CLASS:diag_benign
    REPORT_WITH_FINDINGS:yes
bk  ==========================
```

Wstępne wyniki

	liczba	%
liczba badań	361 (448)	
FINDINGS	496	100.00
poprawnie wstawione początki	416	83.87
nierozpoznana zmiana	13	2.62
błędnie rozpoznana zmiana	17	3.42
źle wstawione początki	67	13.50
przykładowe atrybuty:		
SATURATION	185	100.00
poprawnie rozpoznane	182	98.38
WITH_CALCIF	40	100.00
poprawnie rozpoznane	35	87.50

Przyczyny błędów

- negacja:

  Opisywane [w badaniu poprzednim z dnia 18.10.99r] [zagęszczenie] [w sutku prawym] obecnie nie jest widoczne.

- koordynacja:

  [pojedyncze makro] i [mikrozwapnienia]

- porównania:

  [W obu sutkach] widoczne [plamiste zagęszczenia] o [charakterze łagodnym] . Największe i [najlepiej wysycone] zlokalizowane [w sutku prawym w KDW].

Zapraszamy na prezentację.