

# Zaawansowane narzędzie do analizy korpusu w oparciu o reguły

---



Michał Marcińczuk

Politechnika Wrocławska

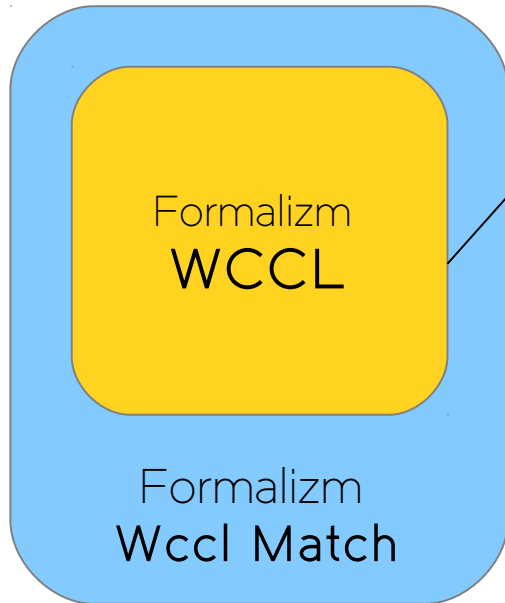
Katedra Inteligencji Obliczeniowej

Grupa Naukowa G4.19

[michal.marcinczuk@pwr.edu.pl](mailto:michal.marcinczuk@pwr.edu.pl)

2015-04-13

# WCCL i Wccl Match



**WCCL** (Wrocław Corpus Constraint Language) to formalizm pozwalający na pisanie **wyrażeń funkcyjnych wartościowanych na tekście oznakowanym morfo-syntaktycznie**. Takie wyrażenie mogą być użyte bezpośrednio jako cechy dla algorytmów maszynowego uczenia przy tworzeniu systemów przetwarzania języka naturalnego. Formalizm powstał z myślą o języku polskim.

**Wccl Match** to formalizm pozwalający na pisanie reguł dopasowania sekwencji tokenów i/lub anotacji w obrębie zdania. Dopasowanie pozwala na naniesienie nowych anotacji na wskazane fragmenty oraz usunięcie istniejących anotacji.

# Składnia reguł Wccl Match

» sekcja match



```
match_rules (  
  apply(  
    match(  
      is("person_nam"),  
      in("w", base[0]),  
      is("city_nam")  
    ),  
    cond(  
      not(  
        annsub(:1, "possessive")  
      )  
    ),  
    actions(  
      mark(M, "person_in_city")  
    )  
  )  
)
```

Wewnątrz `match_rules` znajduje się lista reguł (`apply`). Reguły wykonywane są sekwencyjnie w kolejności wystąpienia w `match_rules`.

# Składnia reguł Wccl Match

## » sekcja match



```
match_rules (  
  apply(  
    match(  
      is("person_nam"),  
      in("w", base[0]),  
      is("city_nam")  
    ),  
    cond(  
      not(  
        annsub(:1, "possessive")  
      )  
    ),  
    actions(  
      mark(M, "person_in_city")  
    )  
  )  
)
```

Sekcja **match** opisuje poszukiwaną sekwencję tokenów i/lub anotacji.

# Operatory dopasowania

---



Operatory w sekcji **match** pozwalają na dopasowanie:

- ♦ **Tokenu**

- ♦ po **wartościach morfo-syntaktycznych tokenu** (forma bazowa, przypadek, liczba, rodzaj, rodzaj, itd.)
- ♦ po **anotacjach semantycznych** (test sprawdzający, czy token jest częścią anotacji, rozpoczyna lub kończy anotację określonego typu)
- ♦ możliwość łączenia warunków przy pomocy **operatorów i/lub** oraz **negacji**.

- ♦ **Anotacji**

- ♦ dopasowanie anotacji określonego typu (operator is)

- ♦ **Wektora dopasowań**

- ♦ powtarzające się sekwencje tokenów/anotacji,
- ♦ opcjonalne dopasowanie,
- ♦ dopasowanie jednej z możliwych sekwencji (pierwsze lub najdłuższe).

# Operatory dopasowania tokenów

---



- ♦ Warunki dla wartości atrybutu
  - ♦ Zgodność wartości (**equal**)
  - ♦ Zawieranie się zbiorów wartości (**in**)
  - ♦ Przecięcie zbiorów wartości (**inter**)
- ♦ Dopasowanie formy tekstowej/bazowej wyrażeniem regularnym (**regex**)
- ♦ Warunki sprawdzające oznakowanie anotacją
  - ♦ Token jest częścią anotacji (**isannpart**)
  - ♦ Token rozpoczyna anotację (**isannbeg**)
  - ♦ Token kończy anotację (**isannend**)

# Atrybuty tokenów

---



- Forma tekstowa (**orth**)
- Forma bazowa (**base**)
- Atrybuty morfologiczne (tagset NKJP)
  - Klasa gramatyczna (**class**)
  - Liczba (**nmb**)
  - Przypadek (**cas**)
  - Rodzaj (**gnd**)
  - Osoba (**per**)
  - Stopień (**deg**)
  - Aspekt (**asp**)
  - Negacja (**ngt**)
  - Akomodacyjność (**acm**)
  - Akcentowość (**acn**)
  - Poprzyikmowość (**ppr**)
  - Agrultynacyjność (**agg**)
  - Wokaliczność (**vc1**)
  - Wymaganie kropki (**dot**)

<http://nkjp.pl/poliqarp/help/ense2.html>

# Łączenie operatorów i negacja



- Koniunkcja operatorów (**and**)

np., rzeczownik w mianowniku:

```
and(  
    equal(class[0], subst),  
    equal(cas[0], nom)  
)
```

- Alternatywa operatorów (**or**)

np., przymiotnik lub dowolny token w mianowniku:

```
or(  
    equal(class[0], adj),  
    equal(cas[0], nom)  
)
```

- Negacja (**not**)

np., dowolny token nie będący rzeczownikiem

```
not(equal(class[0], subst))
```



# Operatory dopasowania wektorów

---



- Dopasowanie sekwencji tokenów o określonej formie tekstowej (**text**)
- Opcjonalne dopasowanie (**optional**)
- Wielokrotne dopasowanie (**repeat**)
- Dopasowanie najdłuższego wektora z listy możliwych (**longest**),
- Dopasowanie pierwszego pasującego wektora (**oneof**)

# Składnia reguł Wccl Match

» sekcja `cond`



```
match_rules (  
  apply(  
    match(  
      is("person_nam"),  
      in("w", base[0]),  
      is("city_nam")  
    ),  
    cond(  
      not(  
        annsub(:1, "possessive")  
      )  
    ),  
    actions(  
      mark(M, "person_in_city")  
    )  
  )  
)
```

Sekcja **cond** zawiera dodatkowe warunki, jakie muszą spełnić elementy dopasowanej sekwencji.

# Dodatkowe warunki

---



Operatory w sekcji **conds** pozwalają na:

- sprawdzenie wartości atrybutu tokenu (**in**, **inter**, **equal**),
- porównanie wartości atrybutów dwóch tokenów (**in**, **inter**, **equal**)
- sprawdzenie oznaczenia tokenu anotacją semantyczną określonej kategorii (**ann**, **annsub**),
- sprawdzenie zgodności przypadku, liczby i rodzaju sekwencji tokenów (**agr**, **agrpp**),
- sprawdzenie czy pozycja znajduje się w obrębie zdania (**inside**, **outside**),
- łączenia i negacji operatorów (**and**, **or**, **not**),
- konstrukcja warunkowa (**if**),
- szukanie tokenu względem określonej pozycji (**skip**).

# Referencja do dopasowania

---



Referencja do dopasowanych elementów w sekcji **cond** i **actions**:

- symbol **M** reprezentuje cały wektor dopasowania,
- symbol **:n** reprezentuje n-ty element dopasowania, np. **:2** to element dopasowany przez drugi operator w **match**,
- dla dopasowania typu wektor możliwe jest odwołanie się do elementów zagnieżdżonych poprzez indeks wewnątrz wektora, np. **:2:1**
- wyłuskanie pozycji tokenu (dla operatorów wymagających pozycji tokenu):
  - **first(dopasowanie)** – pozycja pierwszego tokenu dopasowania, np. **first(M)**,
  - **last(dopasowanie)** – pozycja ostatniego tokenu dopasowania, np. **last(M)**,

# Składnia reguł Wccl Match

» sekcja **actions**



```
match_rules (  
  apply(  
    match(  
      is("person_nam"),  
      in("w", base[0]),  
      is("city_nam")  
    ),  
    cond(  
      not(  
        annsub(:1, "possessive")  
      )  
    ),  
    actions(  
      mark(M, "person_in_city")  
    )  
  )  
)
```

Sekcja **actions** zawiera listę akcji, które zostaną wykonane na dopasowanej sekwencji.

# Akcje

---



Operatory w sekcji **actions** pozwalają na:

- oznaczenie wskazanego fragmentu anotacją określonego typu,
- oznaczenie wskazanego fragmentu anotacją określonego typu z wymuszeniem nadpisania istniejącej anotacji,
- usunięcie dopasowanej anotacji,
- przypisanie wartości do tokenu.

# Zastosowanie Wccl Match

---



- 1) Do wykrywania i kategoryzacji jednostek identyfikacyjnych.
- 2) Do wykrywania i kategoryzacji wyrażeń temporalnych.
- 3) W systemie odpowiedzi na pytania do kategoryzacji pytań i analizy ich struktury.
- 4) Do wykrywania relacji semantycznych między jednostkami identyfikacyjnymi.

# Jednostki identyfikacyjne



```
apply(  
  match(  
    equal(base[0], "stolica"),  
    is("dict_country_name"),  
    equal(base[0], "być"),  
    is("dict_city_name")  
  ),  
  actions(  
    mark(:2,  
      "nam_loc_gpe_country"),  
    mark(:4,  
      "nam_loc_gpe_city")  
  )  
)
```

- Reguła dopasowuje wyrażenie „stolicą [nazwa] jest [nazwa]” i znakuje drugi element jako nazwę państwa oraz czwarty element jako nazwę miasta.



# Analiza pytań



<pre>apply(   match(     optional(in({prep}, class[0])) ,     in(["ile"], base[0]) ,     optional(       and(         in({subst}, class[0]) ,         in({gen}, cas[0])       )     ),   ),   cond(     if(skip(first(M), \$L,             inter(class[\$L],                   {interp, qub, adv, interj})),         -1),     not(inter(class[\$L],              {praet, fin, bedzie,imps, impt,               inf, pant, pcon, aglt, winien})),     True   ),   actions(     mark(M, M, M:2, "qphrase"),     setprop(M:2, "question", "yes")   ) )</pre>	
	Opcjonalny przyimek.
	Opcjonalny rzeczownik w dopełniaczu..
	Sprawdzenie, czy przed dopasowaniem znajduje się token nie będący znakiem interpunkcyjnym, kublikiem, przysłówkiem bądź wykrzyknikiem.
	Jeżeli jest to sprawdzamy jego klasę gramatyczną. Dla określonych klas zwracamy False, a wpp True.
	Jeżeli nie ma, to jest to początek zdania więc zwracamy True.



---

# Narzędzia wykorzystujące język reguł Wccl Match

# Wccl Match Tester

---



Narzędzie do testowania reguł Wccl Match na znakowanym korpusie.

## Wymagania:

- Otagowany i anotowany korpus w formacie ccl,
- Lista kategorii anotacji.

## Funkcje:

- Kolorowanie składni reguł Wccl Match,
- Statystyki liczone w czasie rzeczywistym:
  - Liczba rozpoznanych, nierozpoznanych i błędnie rozpoznanych anotacji dla określonych kategorii.
  - Precyzja, kompletność i średnia harmoniczna.
- Podgląd rozpoznanych i nierozpoznanych anotacji oraz anotacji pomocniczych.
- Możliwość filtrowania anotacji.

## Demo:

Korpus KPWr 1.2.2 znakowany wyrażeniami temporalnymi.

# Wccl Match Tester



Corpora Liner2 CCL Viewer Wccl Match Tester

User: login Inforex

## Rules

```
1 match_rules (
2
3 // 1.1.2014 r.
4 apply(
5   match(
6     inter(class[0], {ign}),
7     inter(base[0], {'.'}),
8     inter(class[0], {ign}),
9     inter(base[0], {'.'}),
10    inter(class[0], {ign}),
11    inter(base[0], {'.'}),
12    inter(base[0], {'.'})
13  ),
14  actions(
15    mark(M, "t3_date")
16  )
17 );
18
19
20 // godz. 10.30
21 apply(
22   match(
23     inter(base[0], {godzina'}),
24     inter(base[0], {'.'}),
25     inter(class[0], {ign}),
26     optional(inter(base[0], {'.'})),
27     optional(inter(class[0], {ign}))
28  ),
29  actions(
30    mark(M, "t3_time")
31  )
32 );
33
34 )
35
```

## Rules evaluation

Typ anotacji	True Positives	False Positives	False Negatives	Precision	Recall	F-measure
t3_time	4	13	104	23.53%	3.70%	6.40%
t3_date	1	21	1469	4.55%	0.07%	0.13%
t3_duration	0	0	237	0.00%	0.00%	0.00%
t3_set	0	0	27	0.00%	0.00%	0.00%
t3_range	0	0	196	0.00%	0.00%	0.00%

### 1. blogi/00100607.xml

Missing annotations:

Vershbow zwraca uwagę na kluczową kwestię – że niezbędne są niekomercyjne, otwarte alternatywy dla działań digitalizacyjnych podejmowanych **dzisiaj** przede wszystkim przez komercyjne firmy.

### 2. blogi/00100608.xml

Missing annotations:

Termin: [5 maja, 10. 00 - 16. 30]

### 3. blogi/00100608.xml

Missing annotations:

[10. 00 - 10. 30] Otwarcie konferencji : dr Alek Tarkowski ( koordynator , Creative Commons Polska , ICM UW )

### 4. blogi/00100608.xml

Missing annotations:

[10. 30 - 12. 45] Sesja [przedpołudniowa] : przegląd zagadnień związanych z otwartą nauką

### 5. blogi/00100608.xml

Missing annotations:

[10. 30 - 11. 00] : dr Ignasi Labastida i Juan ( Uniwersytet Barceloński , Creative Commons Catalonia ) – W stronę otwartości . Doświadczenia biblioteki uniwersyteckiej .

### 6. blogi/00100608.xml

Missing annotations:

[11. 00 - 11. 30] : dr Ahrash Bissell ( Creative Commons Learn , USA ) – Edukacja dla Innowacji – z pomocą Creative Commons .

### 7. blogi/00100608.xml

Missing annotations:

[11. 30 - 11. 40] : przerwa na kawę

### 8. blogi/00100608.xml

Missing annotations:

[11. 40 - 12. 10] : Paweł Szczęsny ( Zakład Bioinformatyki Instytutu Biochemii i Biofizyki PAN , Wydział Biologii UW ) – Nauka 2 . 0

### 9. blogi/00100608.xml

Missing annotations:

[12. 10 - 12. 30] : Dyskusja

### 10. blogi/00100608.xml

Missing annotations:

Corpora: KPWr 1.2.2.TimeML train-all (1-551)

Evaluate

Processed: 551 z 551

Stop

# Wccl Match Tester

## » pole do edycji reguł



Corpora Liner2 CCL Viewer Wccl Match Tester User: login Inforex

### Rules

```

1 match_rules (
2 // 1.1.2014 r.
3 apply(
4   match(
5     inter(class[0], {ign}),
6     inter(base[0], {ign}),
7     inter(class[0], {ign}),
8     inter(base[0], {ign}),
9     inter(class[0], {ign}),
10    inter(base[0], {rok}),
11    inter(base[0], {ign}),
12    inter(base[0], {ign})
13  ),
14  actions(
15    mark(M, "t3_date")
16  )
17 );
18 |
19 // godz. 10.30
20 apply(
21   match(
22     inter(base[0], {godzina}),
23     inter(base[0], {ign}),
24     inter(class[0], {ign}),
25     optional(inter(base[0], {ign})),
26     optional(inter(class[0], {ign}))
27   ),
28   actions(
29     mark(M, "t3_time")
30   )
31 );
32 )
33 )
34 )
35 
```

### Rules evaluation

Typ anotacji	True Positives	False Positives	False Negatives	Precision	Recall	F-measure
t3_time	4	13	104	23.53%	3.70%	6.40%
t3_date	1	21	1469	4.55%	0.07%	0.13%
t3_duration	0	0	237	0.00%	0.00%	0.00%
t3_set	0	0	27	0.00%	0.00%	0.00%
t3_range	0	0	186	0.00%	0.00%	0.00%

- blogi/00100607.xml  
Missing annotations:  
Vershbow zwraca uwagę na kluczową kwestię – że niezbędne są niekomercyjne, otwarte alternatywy dla działań digitalizacyjnych podejmowanych przez komercyjne firmy.
- blogi/00100608.xml  
Missing annotations:  
Termin: [5 maja, 10.00 - 16.30]
- blogi/00100608.xml  
Missing annotations:  
[10.00 - 10.30] Otwarcie konferencji : dr Alek Tarkowski ( koordynator , Creative Commons Polska , ICM UW )
- blogi/00100608.xml  
Missing annotations:  
[10.30 - 12.45] Sesja [przedpołudniowa] : przegląd zagadnień związanych z otwartą nauką
- blogi/00100608.xml  
Missing annotations:  
[10.30 - 11.00] : dr Ignasi Labastida i Juan ( Uniwersytet Barceloński , Creative Commons Catalonia ) – W stronę otwartości . Doświadczenia biblioteki uniwersyteckiej .
- blogi/00100608.xml  
Missing annotations:  
[11.00 - 11.30] : dr Ahrash Bissell ( Creative Commons Learn , USA ) – Edukacja dla Innowacji – z pomocą Creative Commons .
- blogi/00100608.xml  
Missing annotations:  
[11.30 - 11.40] : przerwa na kawę
- blogi/00100608.xml  
Missing annotations:  
[11.40 - 12.10] : Paweł Szczęsny ( Zakład Bioinformatyki Instytutu Biochemii i Biofizyki PAN , Wydział Biologii UW ) – Nauka 2 . 0
- blogi/00100608.xml  
Missing annotations:  
[12.10 - 12.30] : Dyskusja
- blogi/00100608.xml  
Missing annotations:

Corpora: KPWr 1.2.2.TimeML train-all (1-551) Evaluate Processed: 551 z 551 Stop

# Wccl Match Tester

## » wybór korpusu



Corpora Liner2 CCL Viewer Wccl Match Tester
User: login Inforex

### Rules

```

1 match_rules (
2
3 // 1.1.2014 r.
4 apply(
5   match(
6     inter(class[0], {ign}),
7     inter(base[0], {'.'}),
8     inter(class[0], {ign}),
9     inter(base[0], {'.'}),
10    inter(class[0], {ign}),
11    inter(base[0], {'.'}),
12    inter(base[0], {'.'})
13  ),
14  actions(
15    mark(M, "t3_date")
16  )
17 );
18
19 // godz. 10.30
20 apply(
21   match(
22     inter(base[0], {godzina'}),
23     inter(base[0], {'.'}),
24     inter(class[0], {ign}),
25     optional(inter(base[0], {'.'})),
26     optional(inter(class[0], {ign}))
27  ),
28  actions(
29    mark(M, "t3_time")
30  )
31 );
32
33 )
34
35

```

### Rules evaluation

Typ anotacji	True Positives	False Positives	False Negatives	Precision	Recall	F-measure
t3_time	4	13	104	23.53%	3.70%	6.40%
t3_date	1	21	1469	4.55%	0.07%	0.13%
t3_duration	0	0	237	0.00%	0.00%	0.00%
t3_set	0	0	27	0.00%	0.00%	0.00%
t3_range	0	0	186	0.00%	0.00%	0.00%

- blogi/00100607.xml  
Missing annotations:  
Vershbow zwraca uwagę na kluczową kwestię – że niezbędne są niekomercyjne, otwarte alternatywy dla działań digitalizacyjnych podejmowanych przez komercyjne firmy.
- blogi/00100608.xml  
Missing annotations:  
Termin: [5 maja, 10.00 - 16.30]
- blogi/00100608.xml  
Missing annotations:  
[10.00 - 10.30] Otwarcie konferencji : dr Alek Tarkowski ( koordynator , Creative Commons Polska , ICM UW )
- blogi/00100608.xml  
Missing annotations:  
[10.30 - 12.45] Sesja [przedpołudniowa] : przegląd zagadnień związanych z otwartą nauką
- blogi/00100608.xml  
Missing annotations:  
[10.30 - 11.00] : dr Ignasi Labastida i Juan ( Uniwersytet Barceloński , Creative Commons Catalonia ) – W stronę otwartości . Doświadczenia biblioteki uniwersyteckiej .
- blogi/00100608.xml  
Missing annotations:  
[11.00 - 11.30] : dr Ahrash Bissell ( Creative Commons Learn , USA ) – Edukacja dla Innowacji – z pomocą Creative Commons .
- blogi/00100608.xml  
Missing annotations:  
[11.30 - 11.40] : przerwa na kawę
- blogi/00100608.xml  
Missing annotations:  
[11.40 - 12.10] : Paweł Szczepny ( Zakład Bioinformatyki Instytutu Biochemii i Biofizyki PAN , Wydział Biologii UW ) – Nauka 2 . 0
- blogi/00100608.xml  
Missing annotations:  
[12.10 - 12.30] : Dyskusja
- blogi/00100608.xml  
Missing annotations:

Corpora: KPWr 1.2.2.TimeML train-all (1-551) Evaluate

Processed: 551 z 551 Stop

# Wccl Match Tester

## » ocena reguł



Corpora Liner2 CCL Viewer **Wccl Match Tester**
User: login **Inforex**

### Rules

```

1 match_rules (
2
3 // 1.1.2014 r.
4 apply(
5   match(
6     inter(class[0], {ign}),
7     inter(base[0], {'.'}),
8     inter(class[0], {ign}),
9     inter(base[0], {'.'}),
10    inter(class[0], {ign}),
11    inter(base[0], {'.'}),
12    inter(class[0], {'.'}),
13    inter(base[0], {'.'})
14  ),
15  actions(
16    mark(M, "t3_date")
17  );
18 )
19
20 // godz. 10.30
21 apply(
22   match(
23     inter(base[0], {godzina'}),
24     inter(base[0], {'.'}),
25     inter(class[0], {ign}),
26     optional(inter(base[0], {'.'})),
27     optional(inter(class[0], {ign}))
28  ),
29  actions(
30    mark(M, "t3_time")
31  );
32 )
33 )
34 )
35

```

### Rules evaluation

Typ anotacji	True Positives	False Positives	False Negatives	Precision	Recall	F-measure
t3_time	4	13	104	23.53%	3.70%	6.40%
t3_date	1	21	1469	4.55%	0.07%	0.13%
t3_duration	0	0	237	0.00%	0.00%	0.00%
t3_set	0	0	27	0.00%	0.00%	0.00%
t3_range	0	0	196	0.00%	0.00%	0.00%

t.blog/00100608.xml

Missing annotations:  
Vershbow zwraca uwagę na kluczową kwestię – że niezbędne są niekomercyjne, otwarte alternatywy dla działań digitalizacyjnych podejmowanych przez komercyjne firmy.

2. blog/00100608.xml

Missing annotations:  
Termin: [5 maja, 10.00 - 16.30]

3. blog/00100608.xml

Missing annotations:  
[10.00 - 10.30] Otwarcie konferencji : dr Alek Tarkowski ( koordynator, Creative Commons Polska, ICM UW )

4. blog/00100608.xml

Missing annotations:  
[10.30 - 12.45] Sesja przedpołudniowa : przegląd zagadnień związanych z otwartą nauką

5. blog/00100608.xml

Missing annotations:  
[10.30 - 11.00] : dr Ignasi Labastida i Juan ( Uniwersytet Barceloński , Creative Commons Catalonia ) – W stronę otwartości . Doświadczenia biblioteki uniwersyteckiej .

6. blog/00100608.xml

Missing annotations:  
[11.00 - 11.30] : dr Ahrash Bissell ( Creative Commons Learn , USA ) – Edukacja dla Innowacji – z pomocą Creative Commons .

7. blog/00100608.xml

Missing annotations:  
[11.30 - 11.40] : przerwa na kawę

8. blog/00100608.xml

Missing annotations:  
[11.40 - 12.10] : Paweł Szczęsny ( Zakład Bioinformatyki Instytutu Biochemii i Biofizyki PAN , Wydział Biologii UW ) – Nauka 2 . 0

9. blog/00100608.xml

Missing annotations:  
[12.10 - 12.30] : Dyskusja

10. blog/00100608.xml

Missing annotations:

Corpora: KPWr 1.2.2.TimeML train-all (1-551)
**Evaluate**
Processed: 551 z 551
Stop

# Wccl Match Tester

## » podgląd anotacji



Corpora Liner2 CCL Viewer Wccl Match Tester User: login Inforex

### Rules

```
1 match_rules (  
2 // 1.1.2014 r.  
3 apply(  
4   match(  
5     inter(class[0], {ign}),  
6     inter(base[0], {'.'}),  
7     inter(class[0], {ign}),  
8     inter(base[0], {'.'}),  
9     inter(class[0], {ign}),  
10    inter(base[0], {'.'}),  
11    inter(class[0], {rok}),  
12    inter(base[0], {'.'})  
13  ),  
14  actions(  
15    mark(M, "t3_date")  
16  );  
17 )  
18 )  
19 // godz. 10.30  
20 apply(  
21   match(  
22     inter(base[0], {'.'}),  
23     inter(base[0], {'.'}),  
24     inter(class[0], {ign}),  
25     optional(inter(base[0], {'.'})),  
26     optional(inter(class[0], {ign}))  
27  ),  
28  actions(  
29    mark(M, "t3_time")  
30  );  
31 )  
32 )  
33 )  
34 )  
35 )
```

### Rules evaluation

Typ anotacji	True Positives	False Positives	False Negatives	Precision	Recall	F-measure
t3_time	4	13	104	23.53%	3.70%	6.40%
t3_date	1	21	1469	4.55%	0.07%	0.13%
t3_duration	0	0	237	0.00%	0.00%	0.00%
t3_set	0	0	27	0.00%	0.00%	0.00%
t3_range	0	0	186	0.00%	0.00%	0.00%

1. blogi/00100607.xml  
Missing annotations:  
Vershbow zwraca uwagę na kluczową kwestię – że niezbędne są niekomercyjne, otwarte alternatywy dla działań digitalizacyjnych podejmowanych **dzisiaj** przede wszystkim przez komercyjne firmy.

2. blogi/00100608.xml  
Missing annotations:  
Termin: **5 maja, 10.00 - 16.30**

3. blogi/00100608.xml  
Missing annotations:  
**10.00 - 10.30** Otwarcie konferencji : dr Alek Tarkowski ( koordynator , Creative Commons Polska , ICM UW )

4. blogi/00100608.xml  
Missing annotations:  
**10.30 - 12.45** Sesja **przedpołudniowa** : przegląd zagadnień związanych z otwartą nauką

5. blogi/00100608.xml  
Missing annotations:  
**10.30 - 11.00** : dr Ignasi Labastida i Juan ( Uniwersytet Barceloński , Creative Commons Catalonia ) – W stronę otwartości . Doświadczenia biblioteki uniwersyteckiej .

6. blogi/00100608.xml  
Missing annotations:  
**11.00 - 11.30** : dr Ahrash Bissell ( Creative Commons Learn , USA ) – Edukacja dla Innowacji – z pomocą Creative Commons .

7. blogi/00100608.xml  
Missing annotations:  
**11.30 - 11.40** : przerwa na kawę

8. blogi/00100608.xml  
Missing annotations:  
**11.40 - 12.10** : Paweł Szczęsny ( Zakład Bioinformatyki Instytutu Biochemii i Biofizyki PAN , Wydział Biologii UW ) – Nauka 2 . 0

9. blogi/00100608.xml  
Missing annotations:  
**12.10 - 12.30** : Dyskusja

10. blogi/00100608.xml  
Missing annotations:

Corpora: KPWr 1.2.2.TimeML train-all (1-551) Evaluate Processed: 551 z 551 Stop



# Wccl Match (Corpus)

---



Narzędzie do wyszukiwania fraz przy użyciu reguł Wccl Match.

## Wymagania:

Korpus zaimportowany z DSpace.

## Funkcje:

- Przybornik z operatorami Wccl Match,
- Kolorowanie składni reguł Wccl Match,
- Opis anotacji do wyświetlenia.
- Podgląd rozpoznanych anotacji.

## Demo:

Korpus zaimportowany z DSpace.

# Wccl Match (Corpus)



Corpora Liner2 CCL Viewer Wccl Match Tester Administration User: Michał Marciniuk (logout) Inforex

Start Settings Documents Statistics Words frequency Wccl Match

### Toolbox

Rule Match Cond Actions Token attributes

```
apply(  
  // Contains a list of operators matching a sequence of tokens and annotations  
  match(  
    [match_operators]  
  ),  
  // Contains a list of additional conditions to be satisfied to accept a completed match.  
  // This section is optional  
  cond(  
    [cond_operators]  
  ),  
  // Contains a set of actions performed on the matched elements.  
  actions(  
    [action_operators]  
  )  
)
```

### Annotations

```
1 // Enter which annotations should be displayed.  
2 // annotation_name color is required  
3 nam_org green yes // names of organizations  
4 nam_subst red yes // substs which are part of organization names
```

### Rules

```
1 match_rules (  
2  
3   apply(  
4     match(  
5       and(  
6         equal(class[0], subst),  
7         isanpart(0, "nam_org")  
8       ),  
9     ),  
10    actions(  
11      mark(M, "nam_subst")  
12    )  
13  )  
14 )  
15 )
```

Saved Run Status: Done (4.238s) Stop

### Matches

- 1.118146  
1. Peppange Peppange ( luks .
- 1.118150  
1. PZL . 13 PZL . 13 ( PZL - 13 ) – projekt polskiego sześciomiejscowego samolotu pasażerskiego zaprojektowanego w 1931 roku przez inż . Stanisława Praussa w Państwowych Zakładach Lotniczych w Warszawie Historia W drugiej połowie 1930 roku na zamówienie Ministerstwa Komunikacji inż . Stanisław Prauss rozpoczął prace projektowe nad sześciomiejscowym samolotem pasażerskim .  
2. Jednak już na wiosnę 1931 roku Ministerstwo Komunikacji zrezygnowało z tego zamówienia i nie podjęło dalszych prac nad tym projektem .
- 1.118152  
1. W kompleksie znajduje się jedna z najważniejszych galerii sztuki Inuitów – Toronto Dominion Gallery of Inuit Art .
- 1.118153  
1. Dane liczbowe Liczba ludności : ( ♀ 4 958 + ♂ 5 034 ) = 9 992 wiek 0 - 6 : 7 , 2 % wiek 7 - 16 : 13 , 1 % wiek 17 - 66 : 65 , 3 % wiek 67 + : 14 , 4 % zagęszczenie ludności : 68 , 0 osób / km2 bezrobocie : 4 , 8 % osób w wieku 17 - 66 lat cudzoziemcy z UE , Skandynawii i USA : 808 na 10 . 000 osób cudzoziemcy z krajów Trzeciego Świata : 229 na 10 . 000 osób liczba szkół podstawowych : 5 ( liczba klas : 64 )
- 1.118155  
1. Decyzją Rady Miejskiej z 27 października 2004 r . dwadzieścia pięć kwitnących okazów bluszczu uznanych zostało za pomnik przyrody .
- 1.118156  
1. Mariusz Grabowski Mariusz Grabowski ( ur . 2 maja 1966 w Tarnowie ) – polski polityk , poseł na Sejm I , III i IV kadencji .  
2. Życiorys Ukończył w 1991 studia na Wydziale Prawniczym Katolickiego Uniwersytetu Lubelskiego .  
3. W latach 90 . działał w Zjednoczeniu Chryścijańsko - Narodowym .  
4. W latach 1991 – 1993 i 1997 – 2005 zasiadał w Sejmie I , III i IV kadencji .  
5. Był wybierany kolejno z listy Wyborczej Akcji Katolickiej ( 1991 ) , Akcji Wyborczej Solidarność ( 1997 ) i Ligi Polskich Rodzin ( 2001 ) w okręgu tarnowskim .  
6. W III i IV kadencji występował z klubów parlamentarnych AWS i LPR , przystępując do koła poselskiego Porozumienie Polskie .  
7. W 2005 bez powodzenia kandydował do Senatu .

# Wccl Match (Corpus)

## » Przybornik



Corpora Liner2 CCL Viewer Wccl Match Tester Administration User: Michał Marciniuk (logout) Inforex

Start Settings Documents Statistics Words frequency Wccl Match Add document

**Toolbox**

Rule Match Cond Actions Token attributes

```
apply(  
  // Contains a list of operators matching a sequence of tokens and annotations  
  match(  
    [match_operators]  
  )  
  // Contains a list of additional conditions to be satisfied to accept a completed match.  
  // This section is optional  
  cond(  
    [cond_operators]  
  )  
  // Contains a set of actions performed on the matched elements.  
  actions(  
    [action_operators]  
  )  
)
```

**Annotations**

```
1 // Enter which annotations should be displayed.  
2 // annotation_name color is required  
3 nam_org green yes // names of organizations  
4 nam_subst red yes // substs which are part of organization names
```

**Rules**

```
1 match_rules (  
2  
3 apply(  
4 match(  
5   and(  
6     equal(class[0], subst),  
7     isanpart(0, "nam_org")  
8   )  
9 )  
10 actions(  
11   mark(M, "nam_subst")  
12 )  
13 )  
14 )  
15 )
```

**Matches**

- 1.118146  
1. Peppange Peppange ( luks .
- 2.118150  
1. PZL . 13 PZL . 13 ( PZL - 13 ) – projekt polskiego sześciomiejscowego samolotu pasażerskiego zaprojektowanego w 1931 roku przez inż . Stanisława Praussa w Państwowych Zakładach Lotniczych w Warszawie Historia W drugiej połowie 1930 roku na zamówienie Ministerstwa Komunikacji inż . Stanisław Prauss rozpoczął prace projektowe nad sześciomiejscowym samolotem pasażerskim .  
2. Jednak już na wiosnę 1931 roku Ministerstwo Komunikacji zrezygnowało z tego zamówienia i nie podjęło dalszych prac nad tym projektem .
- 3.118152  
1. W kompleksie znajduje się jedna z najważniejszych galerii sztuki Inuitów – Toronto Dominion Gallery of Inuit Art .
- 4.118153  
1. Dane liczbowe Liczba ludności : ( ♀ 4 958 + ♂ 5 034 ) = 9 992 wiek 0 - 6 : 7 , 2 % wiek 7 - 16 : 13 , 1 % wiek 17 - 66 : 65 , 3 % wiek 67 + : 14 , 4 % zagęszczenie ludności : 68 , 0 osób / km2 bezrobocie : 4 , 8 % osób w wieku 17 - 66 lat cudzoziemcy z UE , Skandynawii i USA : 808 na 10 . 000 osób cudzoziemcy z krajów Trzeciego Świata : 229 na 10 . 000 osób liczba szkół podstawowych : 5 ( liczba klas : 64 )
- 5.118155  
1. Decyzją Rady Miejskiej z 27 października 2004 r . dwadzieścia pięć kwitnących okazów bluszczu uznanych zostało za pomnik przyrody .
- 6.118156  
1. Mariusz Grabowski Mariusz Grabowski ( ur . 2 maja 1966 w Tarnowie ) – polski polityk , poseł na Sejm I , III i IV kadencji .  
2. Życiorys Ukończył w 1991 studia na Wydziale Prawniczym Katolickiego Uniwersytetu Lubelskiego .  
3. W latach 90 . działał w Zjednoczeniu Chrześcijańsko - Narodowym .  
4. W latach 1991 – 1993 i 1997 – 2005 zasiadał w Sejmie I , III i IV kadencji .  
5. Był wybierany kolejno z listy Wyborczej Akcji Katolickiej ( 1991 ) , Akcji Wyborczej Solidarność ( 1997 ) i Ligi Polskich Rodzin ( 2001 ) w okręgu tarnowskim .  
6. W III i IV kadencji występował z klubów parlamentarnych AWS i LPR , przystępując do koła poselskiego Porozumienie Polskie .  
7. W 2005 bez powodzenia kandydował do Senatu .

Save Run Status: Done (4.238s) Stop

# Wccl Match (Corpus)



Corpora Liner2 CCL Viewer Wccl Match Tester Administration User: Michał Marciniuk (logout) Inforex

Start Settings Documents Statistics Words frequency Wccl Match Add document

### Toolbox

Rule Match Cond Actions Token attributes

```
apply(  
  // Contains a list of operators matching a sequence of tokens and annotations  
  match(  
    [match_operators]  
  ),  
  // Contains a list of additional conditions to be satisfied to accept a completed match.  
  // This section is optional  
  cond(  
    [cond_operators]  
  ),  
  // Contains a set of actions performed on the matched elements.  
  actions(  
    [action_operators]  
  )  
)
```

### Annotations

```
1 // Enter which annotations should be displayed.  
2 // annotation_name color is required  
3 nam_org green yes // names of organizations  
4 nam_subst red yes // substs which are part of organization names
```

### Rules

```
1 match_rules (  
2  
3   apply(  
4     match(  
5       and(  
6         equal(class[0], subst),  
7         isanpart(0, "nam_org")  
8       ),  
9     ),  
10    actions(  
11      mark(M, "nam_subst")  
12    )  
13  )  
14 )  
15 )
```

Saved Run Status: Done (4.238s) Stop

### Matches

- 1.118146  
1. Peppange Peppange ( luks .
- 2.118150  
1. PZL . 13 PZL . 13 ( PZL - 13 ) – projekt polskiego sześciomiejscowego samolotu pasażerskiego zaprojektowanego w 1931 roku przez inż . Stanisława Praussa w Państwowych Zakładach Lotniczych w Warszawie Historia W drugiej połowie 1930 roku na zamówienie Ministerstwa Komunikacji inż . Stanisław Prauss rozpracował prace projektowe nad sześciomiejscowym samolotem pasażerskim .  
2. Jednak główna komisja w 1931 roku Ministerstwo Komunikacji zrezygnowało z tego zamówienia i nie podjęło dalszych prac nad tym projektem .
- 3.118152  
1. W kompleksie znajduje się jedna z najważniejszych galerii sztuki Inuitów – Toronto Dominion Gallery of Inuit Art .
- 4.118153  
1. Dane liczbowe Liczba ludności : ( ♀ 4 958 + ♂ 5 034 ) = 9 992 wiek 0 - 6 : 7 , 2 % wiek 7 - 16 : 13 , 1 % wiek 17 - 66 : 65 , 3 % wiek 67 + : 14 , 4 % zagęszczenie ludności : 68 , 0 osób / km2 bezrobocie : 4 , 8 % osób w wieku 17 - 66 lat cudzoziemcy z UE , Skandynawii i USA : 808 na 10 . 000 osób cudzoziemcy z krajów Trzeciego Świata : 229 na 10 . 000 osób liczba szkół podstawowych : 5 ( liczba klas : 64 )
- 5.118155  
1. Decyzją Rady Miejskiej z 27 października 2004 r . dwadzieścia pięć kwitnących okazów bluszczu uznanych zostało za pomnik przyrody .
- 6.118156  
1. Mariusz Grabowski Mariusz Grabowski ( ur . 2 maja 1966 w Tarnowie ) – polski polityk , poseł na Sejm I , III i IV kadencji .  
2. Życiorys Ukończył w 1991 studia na Wydziale Prawniczym Katolickiego Uniwersytetu Lubelskiego .  
3. W latach 90 . działał w Zjednoczeniu Chryścijańsko - Narodowym .  
4. W latach 1991 – 1993 i 1997 – 2005 zasiadał w Sejmie I , III i IV kadencji .  
5. Był wybierany kolejno z listy Wyborczej Akcji Katolickiej ( 1991 ) , Akcji Wyborczej Solidarność ( 1997 ) i Ligi Polskich Rodzin ( 2001 ) w okręgu tarnowskim .  
6. W III i IV kadencji występował z klubów parlamentarnych AWS i LPR , przystępując do koła poselskiego Porozumienie Polskie .  
7. W 2005 bez powodzenia kandydował do Senatu .

# Wccl Match (Corpus)



Corpora Liner2 CCL Viewer Wccl Match Tester Administration User: Michał Marciniuk (logout) Inforex

Start Settings Documents Statistics Words frequency Wccl Match Add document

### Toolbox

Rule Match Cond Actions Token attributes

```
apply(  
  // Contains a list of operators matching a sequence of tokens and annotations  
  match(  
    [match_operators]  
  ),  
  // Contains a list of additional conditions to be satisfied to accept a completed match.  
  // This section is optional  
  cond(  
    [cond_operators]  
  ),  
  // Contains a set of actions performed on the matched elements.  
  actions(  
    [action_operators]  
  )  
)
```

### Annotations

```
1 // Enter which annotations should be displayed.  
2 // annotation_name color is required  
3 nam_org green yes // names of organizations  
4 nam_subst red yes // substs which are part of organization names
```

### Rules

```
1 match_rules (  
2  
3 apply(  
4   match(  
5     and(  
6       equal(class[0], subst),  
7       isanpart(0, "nam_org")  
8     ),  
9     actions(  
10      mark(M, "nam_subst")  
11    )  
12  )  
13 )  
14 )  
15 )
```

Saved Run Staus: Done (4.238s) Stop

### Matches

- 1.118146  
1. Peppange Peppange ( luks .
- 2.118150  
1. PZL . 13 PZL . 13 ( PZL - 13 ) – projekt polskiego sześciomiejscowego samolotu pasażerskiego zaprojektowanego w 1931 roku przez inż . Stanisława Praussa w Państwowych Zakładach Lotniczych w Warszawie Historia W drugiej połowie 1930 roku na zamówienie Ministerstwa Komunikacji inż . Stanisław Prauss rozpoczął prace projektowe nad sześciomiejscowym samolotem pasażerskim .  
2. Jednak już na wiosnę 1931 roku Ministerstwo Komunikacji zrezygnowało z tego zamówienia i nie podjęło dalszych prac nad tym projektem .
- 3.118152  
1. W kopercie znajduje się jedna z najważniejszych galerii sztuki Inuitów – Toronto Dominion Gallery of Inuit Art .
- 4.118153  
1. Dane liczbowe Liczba ludności : ( ♀ 4 958 + ♂ 5 034 ) = 9 992 wiek 0 - 6 : 7 , 2 % wiek 7 - 16 : 13 , 1 % wiek 17 - 66 : 65 , 3 % wiek 67 + : 14 , 4 % zagęszczenie ludności : 68 , 0 osób / km2 bezrobocie : 4 , 8 % osób w wieku 17 - 66 lat cudzoziemcy z UE , Skandynawii i USA : 808 na 10 . 000 osób cudzoziemcy z krajów Trzeciego Świata : 229 na 10 . 000 osób liczba szkół podstawowych : 5 ( liczba klas : 64 )
- 5.118155  
1. Decyzją Rady Miejskiej z 27 października 2004 r . dwadzieścia pięć kwitujących okazów bluszczu uznanych zostało za pomnik przyrody .
- 6.118156  
1. Mariusz Grabowski Mariusz Grabowski ( ur . 2 maja 1966 w Tarnowie ) – polski polityk , poseł na Sejm I , III i IV kadencji .  
2. Życiorys Ukończył w 1991 studia na Wydziale Prawniczym Katolickiego Uniwersytetu Lubelskiego .  
3. W latach 90 . działał w Zjednoczeniu Chryścijańsko - Narodowym .  
4. W latach 1991 – 1993 i 1997 – 2005 zasiadał w Sejmie I , III i IV kadencji .  
5. Był wybierany kolejno z listy Wyborczej Akcji Katolickiej ( 1991 ) , Akcji Wyborczej Solidarność ( 1997 ) i Ligi Polskich Rodzin ( 2001 ) w okręgu tarnowskim .  
6. W III i IV kadencji występował z klubów parlamentarnych AWS i LPR , przystępując do koła poselskiego Porozumienie Polskie .  
7. W 2005 bez powodzenia kandydował do Senatu .

# Wccl Match (Corpus)



Corpora Liner2 CCL Viewer Wccl Match Tester Administration User: Michał Marciniuk (logout) Inforex

Start Settings Documents Statistics Words frequency Wccl Match Add document

### Toolbox

Rule Match Cond Actions Token attributes

```
apply(  
  // Contains a list of operators matching a sequence of tokens and annotations  
  match(  
    [match_operators]  
  ),  
  // Contains a list of additional conditions to be satisfied to accept a completed match.  
  // This section is optional  
  cond(  
    [cond_operators]  
  ),  
  // Contains a set of actions performed on the matched elements.  
  actions(  
    [action_operators]  
  )  
)
```

### Annotations

```
1 // Enter which annotations should be displayed.  
2 // annotation_name color is required  
3 nam_org green yes // names of organizations  
4 nam_subst red yes // substs which are part of organization names
```

### Rules

```
1 match_rules (  
2  
3   apply(  
4     match(  
5       and(  
6         equal(class[0], subst),  
7         isanpart(0, "nam_org")  
8       )  
9     ),  
10    actions(  
11      mark(M, "nam_subst")  
12    )  
13  )  
14 )  
15 )
```

Saved Run Status: Done (4.238s) Stop

### Matches

- 1.118146  
1. Peppange Peppange ( luks .
- 2.118150  
1. PZL . 13 PZL . 13 ( PZL - 13 ) – projekt polskiego sześciomiejscowego samolotu pasażerskiego zaprojektowanego w 1931 roku przez inż . Stanisława Praussa w Państwowych Zakładach Lotniczych w Warszawie Historia W drugiej połowie 1930 roku na zamówienie Ministerstwa Komunikacji inż . Stanisław Prauss rozpoczął prace projektowe nad sześciomiejscowym samolotem pasażerskim .  
2. Jednak już na wiosnę 1931 roku Ministerstwo Komunikacji zrezygnowało z tego zamówienia i nie podjęło dalszych prac nad tym projektem .
- 3.118152  
1. W kompleksie znajduje się jedna z najważniejszych galerii sztuki Inuitów – Toronto Dominion Gallery of Inuit Art .
- 4.118153  
1. Dane liczbowe Liczba ludności : ( ♀ 4 958 + ♂ 5 034 ) = 9 992 wiek 0 - 6 : 7 , 2 % wiek 7 - 16 : 13 , 1 % wiek 17 - 66 : 65 , 3 % wiek 67 + : 14 , 4 % zagęszczenie ludności : 68 , 0 osób / km2 bezrobocie : 4 , 8 % osób w wieku 17 - 66 lat cudzoziemcy z UE , Skandynawii i USA : 808 na 10 . 000 osób cudzoziemcy z krajów Trzeciego Świata : 229 na 10 . 000 osób liczba szkół podstawowych : 5 ( liczba klas : 64 )
- 5.118155  
1. Decyzją Rady Miejskiej z 27 października 2004 r . dwadzieścia pięć kwitujących okazów bluszczu uznanych zostało za pomnik przyrody .
- 6.118156  
1. Mariusz Grabowski Mariusz Grabowski ( ur . 2 maja 1966 w Tarnowie ) – polski polityk , poseł na Sejm I , III i IV kadencji .  
2. Życiorys Ukończył w 1991 studia na Wydziale Prawniczym Katolickiego Uniwersytetu Lubelskiego .  
3. W latach 90 . działał w Zjednoczeniu Chryścijańsko - Narodowym .  
4. W latach 1991 – 1993 i 1997 – 2005 zasiadał w Sejmie I , III i IV kadencji .  
5. Był wybierany kolejno z listy Wyborczej Akcji Katolickiej ( 1991 ) , Akcji Wyborczej Solidarność ( 1997 ) i Ligi Polskich Rodzin ( 2001 ) w okręgu tarnowskim .  
6. W III i IV kadencji występował z klubów parlamentarnych AWS i LPR , przystępując do koła poselskiego Porozumienie Polskie .  
7. W 2005 bez powodzenia kandydował do Senatu .

# CLARIN

Common Language Resources and Technology Infrastructure

---



## Pytania i odpowiedzi

---

# CLARIN

Common Language Resources and Technology Infrastructure

---



Dziękuję bardzo za uwagę

---