Modelowanie treści sumeryjskich tekstów gospodarczych z epoki Ur III.

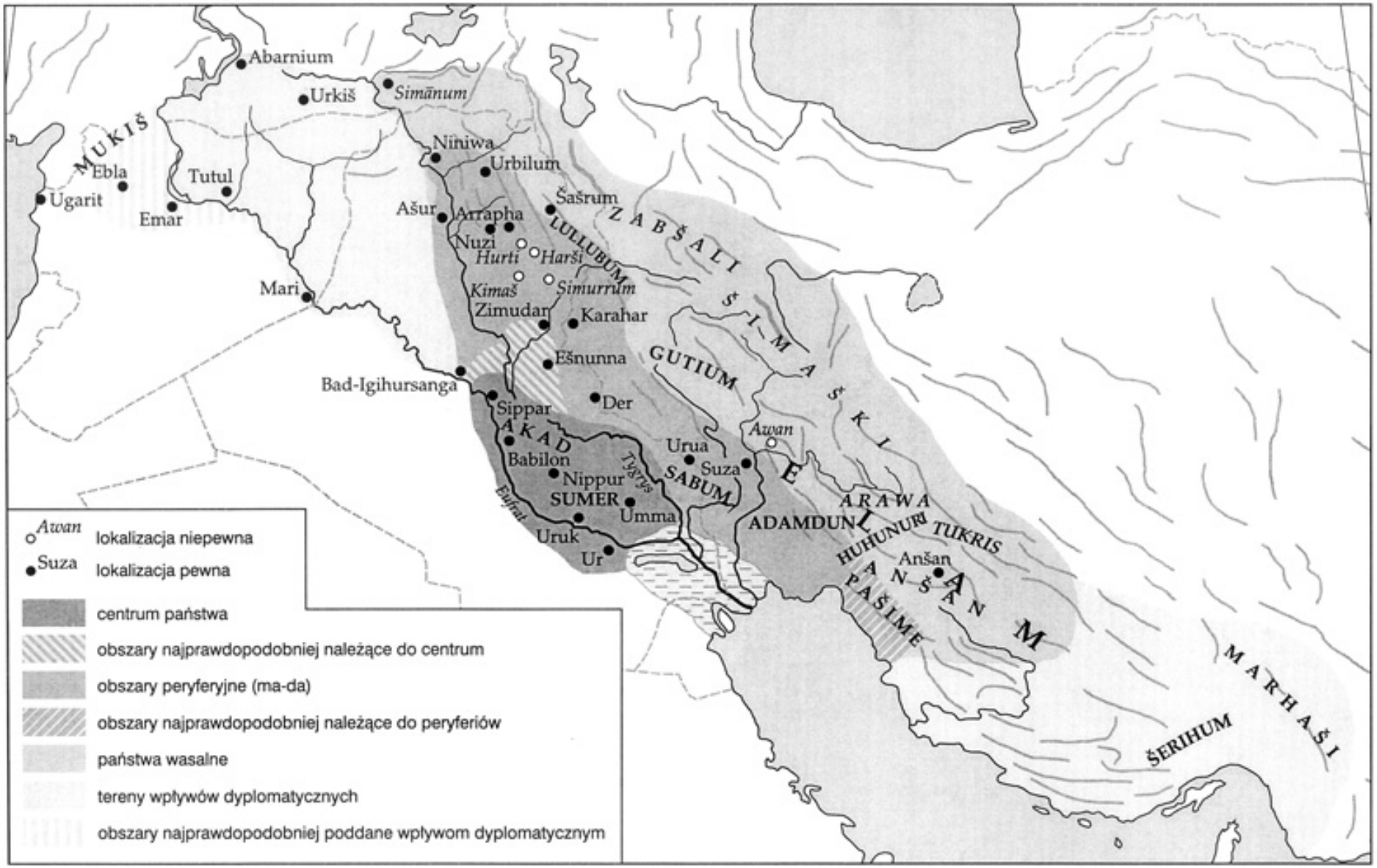
Wojciech Jaworski
Uniwersytet Warszawski

19 listopada 2007

Plan referatu

1. Język
2. Parsowanie
3. Zastosowania
4. Rozszerzenia

Starożytny Sumer w okresie Ur III



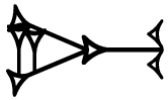
Korpus tekstów

- Archeolodzy odkopali ok. 100000 tabliczek pochodzących z czasów panowania trzeciej dynastii z Ur (2100p.n.e.-2000p.n.e.).
- Większość z nich zostało napisanych w przeciągu 40 lat i pochodzi z 4 miast.
- Urzędnicy królewscy dokumentowali na glinianych tabliczkach przekazywanie dóbr, składanie ofiar, opłacanie pracowników, bilanse produkcji w zakładach przemysłowych, wykonaną pracę, aprowizowanie wojska, diety podróżne itp.
- Około 43000 tekstów dostępnych jest w wersji elektronicznej w bazie danych Cuneiform Digital Library Initiative (CDLI) prowadzonej przez Uniwersytet Kalifornijski w Los Angeles i Instytut Historyczny im. Maxa Plancka w Berlinie.
- Tabliczki są przechowywane w bazie w postaci tekstowej w formie transliteracji w alfabecie łacińskim.
- Na potrzeby badań wybrałem korpus ok. 17000 dokumentów dotyczących dystrybucji zwierząt hodowlanych.

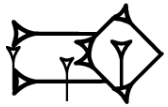
Tekst gospodarczy

Pismo: homonimia

- Jeden dźwięk może być zapisywany na różne sposoby, przyjmując różne znaczenia.
- Na przykład dźwięk *gu* miał czternaście odpowiadających mu znaków graficznych, m.in.

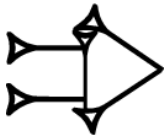


gu len



gu2 szyja

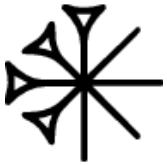
gu3 głos



gu4 wół

Pismo: polifonia

Znak może być różnie czytany w zależności od kontekstu.



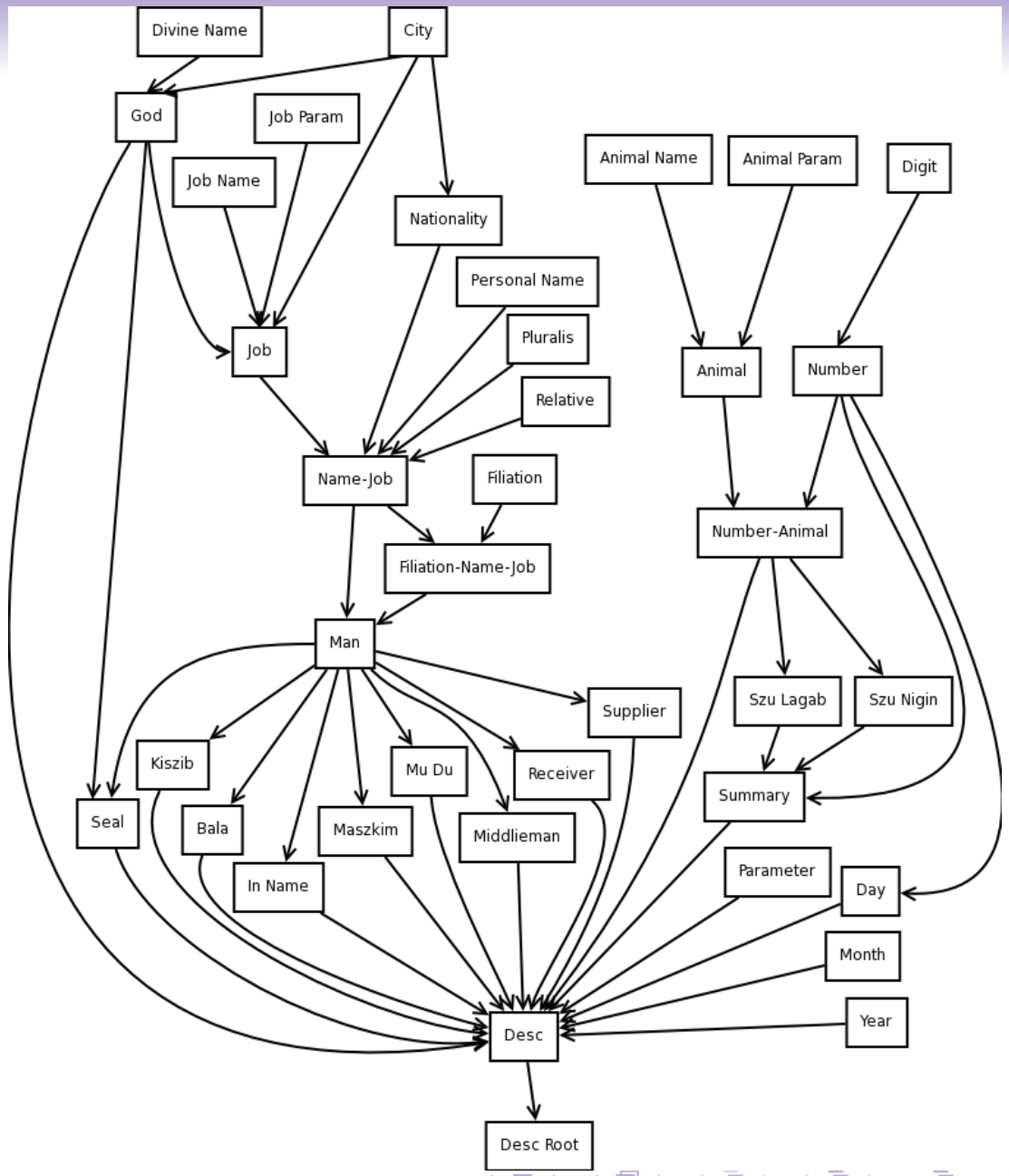
- główne odczyty: `an`, `dingir`, `d`
- pozostałe odczyty: `a11?`, `am6`, `ana2`, `anu`, `dalhamunx`, `dibur`, `digir`, `dimer`, `/dimer/`, `dimir`, `dimmer`, `diner`, `dingira`, `dinir`, `dir3`, `el3`, `eszszu2`, `il3`, `ila`, `ili`, `ilu`, `le4`, `sa8`, `szubul`, `s,a3`, `usanx`, `za5`
- `dingir` oznacza *bóg*
- `d` (skrót od `dingir`) jest determinatywem oznaczającym, że następne słowo jest imieniem bóstwa
- `an` to imię boga niebios (jest to jedyny bóg, którego imię jest pisane bez determinatywu)
- `an` oznacza *niebo*, *niebiański*
- `an` może być czytany fonetycznie jako sylaba *an*, wykorzystywany w zapisie afiksów gramatycznych i oraz słów (nazw własnych) pochodzenia pozasumeryjskiego.

Transliteracja

```
&P123831 = OIP 121, 101
@tablet
@obverse
1.  1(disz) sila4 ur-mes ensi2
2.  1(disz)# sila4 da-da dumu
lugal
3.  1(disz)# sila4 id-da-a
@reverse
1.  u4 2(u) 3(asz@t)-kam
$ 1 line blank
3.  mu-DU
4.  ab-ba-sa6-ga i3-dab5
5.  iti sze-KIN-ku5
6.  mu en dinanna ba-hun
@left
1.  3(disz)
```

Ontologia

- Diagram komponentów znaczeniowych.
- Relacje bycia częścią, bycia cechą.
- Ontologia jest domknięta za pomocą specjalnej kategorii zawierającej wszystkie obiekty nie mieszczące się w kategoriach przedstawionych na diagramie.



Język reprezentacji znaczenia

- Język reprezentacji znaczenia składa się ze stałych oraz predykatów, które są łączone ze sobą za pomocą alternatywy i koniunkcji.
- Nazwy predykatów są kategoriami ontologicznymi.
- Pierwszym argumentem predykatu jest stała reprezentująca obiekt należący do kategorii wyznaczonej przez predykat.
- Pozostałe argumenty są stałymi reprezentującymi części (cechy) obiektu.

Fraza

`ur-mes ensi2`

zapisana w języku reprezentacji znaczenia będzie miała postać

$$\texttt{Name-Job}(a_1, a_2, a_3) \wedge \texttt{Name}(a_2, a_4, a_5) \wedge \texttt{Job}(a_3, a_6) \wedge \texttt{JobName}(a_6, a_7) \wedge$$

$$\wedge\ \texttt{ur}(a_4) \wedge \texttt{mes}(a_5) \wedge \texttt{ensi2}(a_7).$$

Semantyka języka reprezentacji znaczenia

$$\texttt{Name-Job}(a_1, a_2, a_3) \wedge \texttt{Name}(a_2, a_4, a_5) \wedge \texttt{Job}(a_3, a_6) \wedge \texttt{JobName}(a_6, a_7) \wedge$$

$$\wedge\ \texttt{ur}(a_4) \wedge \texttt{mes}(a_5) \wedge \texttt{ensi2}(a_7)$$

- Predykaty definiują kategorię ontologiczną swojego pierwszego argumentu oraz wiążą pierwszy argument z pozostałymi za pomocą relacji bycia częścią, bądź opisem.
- Interpretacjami predykatów `ur`, `mes`, `ensi2` są sylaby lub wyrazy — elementy języka.
- Interpretacją predykatu `Name` jest zbiór imion rozumianych jako elementy języka.
- Interpretacją predykatu `JobName` jest zbiór nazw zawodów rozumianych jako elementy języka.
- Interpretacjami stałych są konkretne obiekty.

Plan referatu

1. Język
2. Parsowanie
3. Zastosowania
4. Rozszerzenia

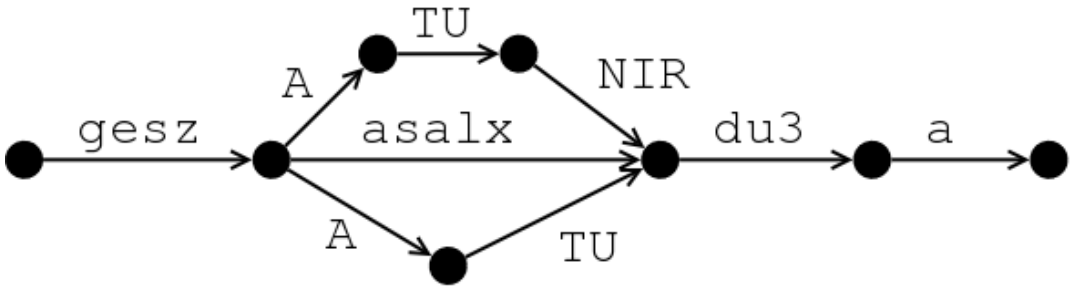
Segmentacja

- Podział tekstu na znaki.
- Odróżnienie tekstu sumeryjskiego od oznaczeń edytorskich,
- Brak jednolitej konwencji oznaczania struktury i uszkodzeń tekstów. CDLI stopniowo ujednolica teksty.
- W wyniku segmentacji tekst może utracić strukturę listy

Tekst

```
{gesz}asalx(|A.TU.NIR|)!(A.TU) du3-a
```

po segmentacji ma postać grafu



Składnia dokumentów gospodarczych

- Niemal wszystkie słowa są rzeczownikami albo liczebnikami.
- Afiksy deklinacyjne i słowotwórcze niemalże nie występują, a sens tekstu jest wyznaczony przez kolejność fraz.
- Badania empiryczne wykazały, że język sumeryjskich tekstów gospodarczych jest regularny.
- Dokument gospodarczy jest pojedynczym zdaniem w szyku OSV, z bardzo rozbudowaną częścią O; zaopatrzonym w datę.

Schemat prostego pokwitowania

((`Number Animal`)+ `Man`)+ `Day Receiver Month Year Summary`.

Schemat miesięcznego zestawienia

(((`Number Animal`) + `Man`) + `Summary Day`) + `SzuLagab`

`Receiver Month Year`.

Gramatyka

- reguły gramatyczne mają postać

$$A \to \alpha_1, \ldots, \alpha_n \qquad A \to \beta+$$

  gdzie $\alpha_1, \ldots, \alpha_n, \beta$ są symbolami.
- Reguła $A \to \beta+$ jest równoważna zbiorowi reguł

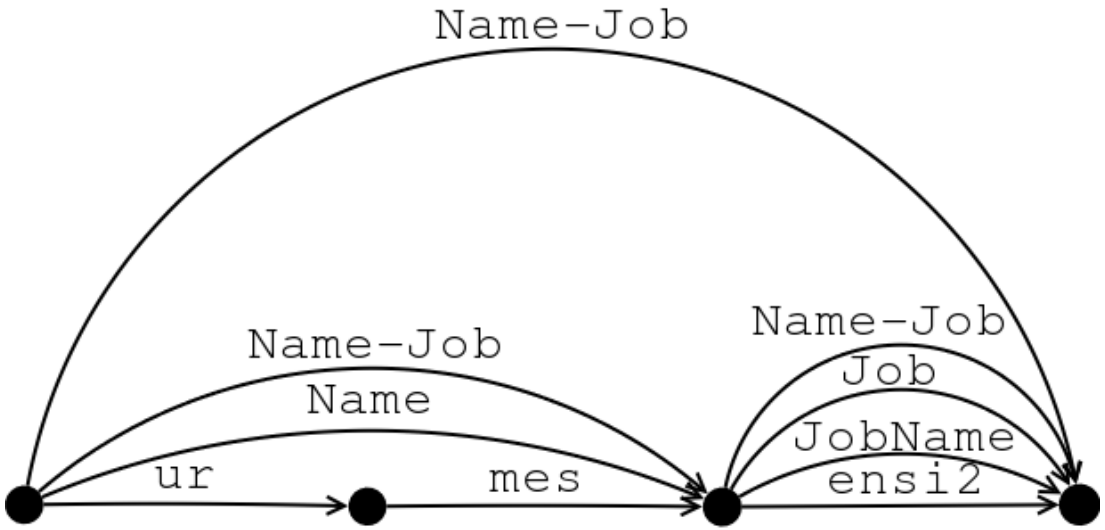
$$A \to \beta, A \to \beta\beta, A \to \beta\beta\beta, \ldots.$$

- w gramatyce nie ma rekurencji
- Gramatyka jest semantyczna — większości symboli z gramatyki odpowiadają kategorie ontologiczne.
- Przykładowe reguły

```
Name                ::=  ur mes
Name-Job            ::=  Name Job
Name-Job            ::=  Name
Name-Job            ::=  Job
Number-Animal       ::=  Number Animal
Number-AnimalList   ::=  NumberAnimal +
```

Analizator składniowy: reprezentacja danych

- Częściowo zinterpretowany tekst reprezentujemy za pomocą skierowanego grafu acyklicznego, którego. krawędzie etykietujemy symbolami z gramatyki.
- Graf ten będziemy nazywać *grafem rozbioru składniowego*.
- Parsowanie przebiega z dołu do góry.
- Aplikacja reguły polega na znalezieniu ścieżki w grafie i dodaniu do niego nowej krawędzi etykietowanej produkcją reguły.
- Gramatyka nie musi być kompletna.



Wieloznaczność

- Wieloznaczność wynikająca z niewiedzy:

  `zabar-dab5.`

- Wieloznaczność wynikająca ze skrótowego zapisu:

  `mu en {d}inanna.`

- Wieloznaczność wynikająca z braku nawiasowania:

  `NAList Man NAList Man Middleman NAList Man Middleman.`

  Możliwe są dwie interpretacje:

  `(NAList Man NAList Man) Middleman NAList Man Middleman,`

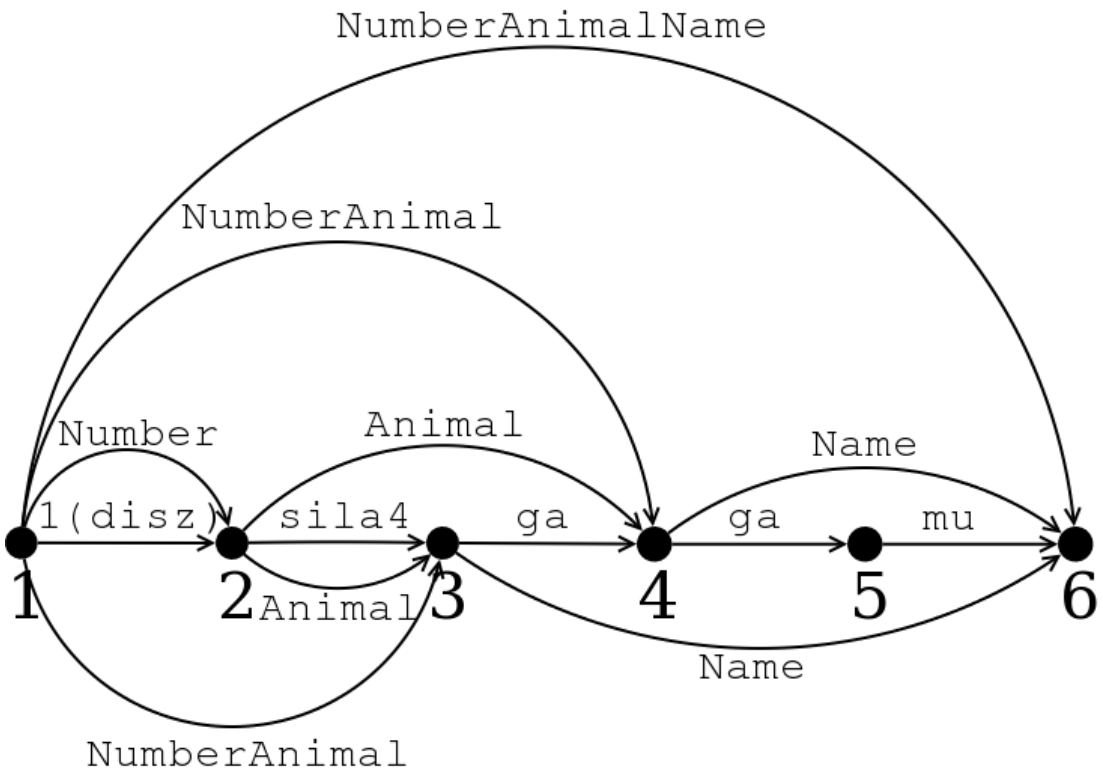
  `NAList Man (NAList Man Middleman) NAList Man Middleman.`

Wieloznaczność

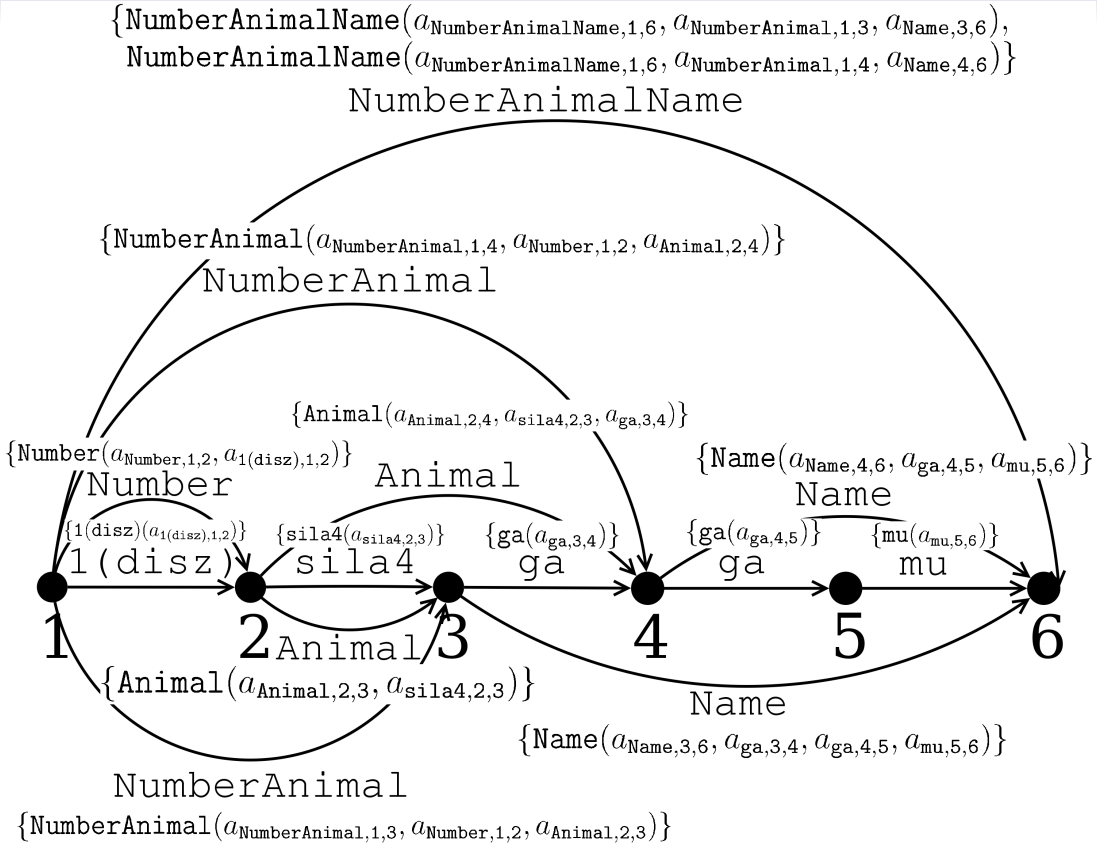
- Wieloznaczność wynikająca z braku znaków rozdzielających słowa

  `1(disz) sila4 ga ga mu`

  `ga-ga-mu` i `ga-mu` to imiona, a `ga` jest określeniem zwierzęcia. `NumberAnimalName` można uzyskać na dwa sposoby.

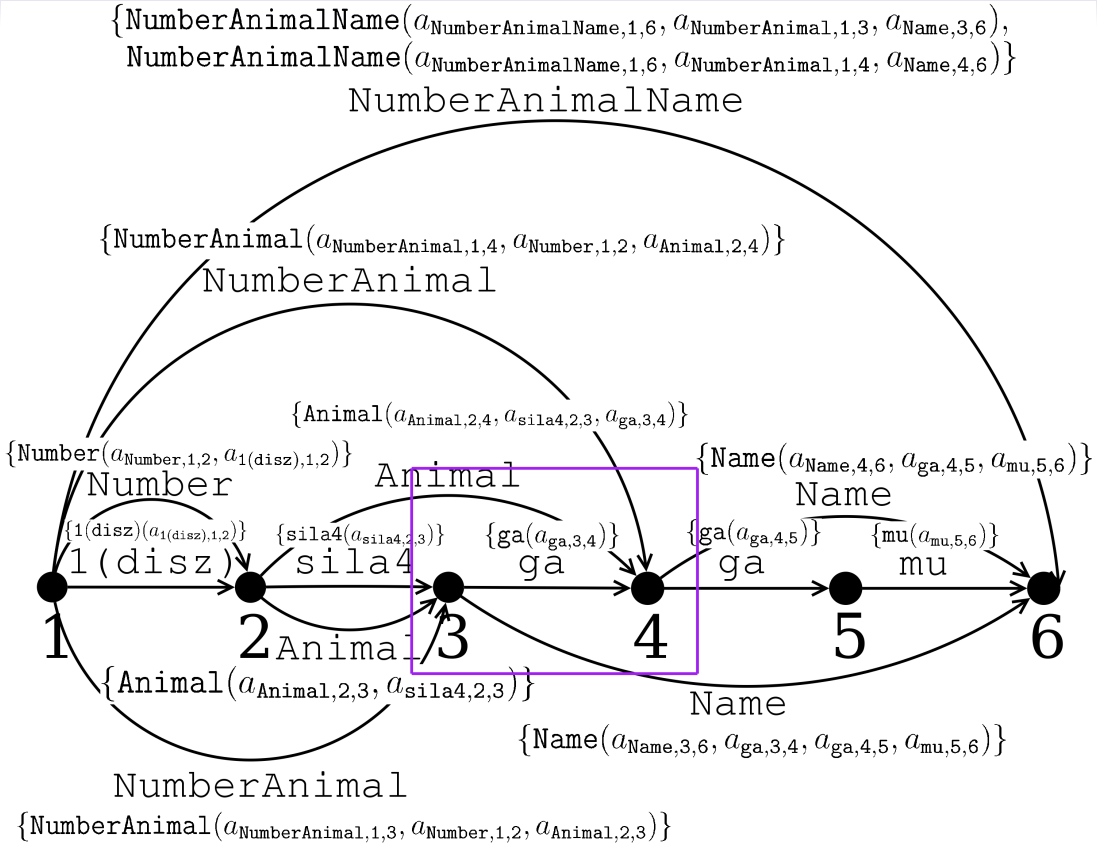


Analiza semantyczna



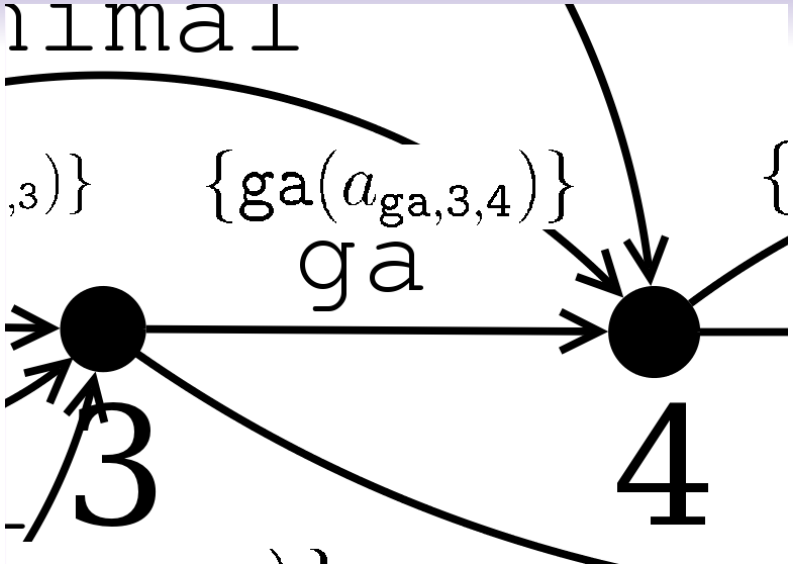
- Formuły języka reprezentacji znaczenia powiązane są z grafem rozbioru składniowego.
- Pozwala to w zwarty sposób zapamiętać treść niejednoznacznych dokumentów.

Analiza semantyczna



- Formuły języka reprezentacji znaczenia powiązane są z grafem rozbioru składniowego.
- Pozwala to w zwarty sposób zapamiętać treść niejednoznacznych dokumentów.

Analiza semantyczna

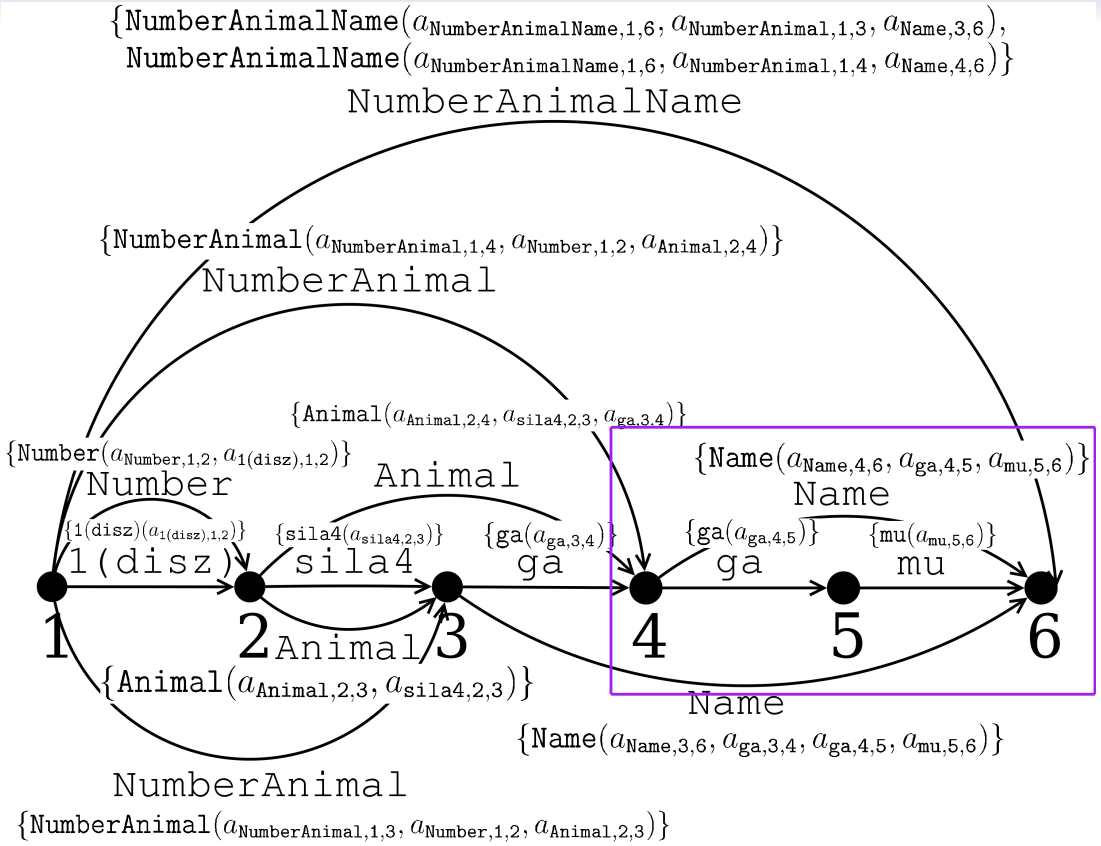


- Z każdym symbolem gramatyki wiążemy zbiór predykatów.
- Zbiór przypisany do symbolu utworzonego podczas segmentacji zawiera jeden jednoargumentowy predykat.
- Nazwa predykatu jest tożsama z nazwą symbolu.
- Semantyka krawędzi grafu

$$[\![\mathtt{ga}]\!]_{3,4} = \mathtt{ga}(a_{\mathtt{ga},3,4})$$

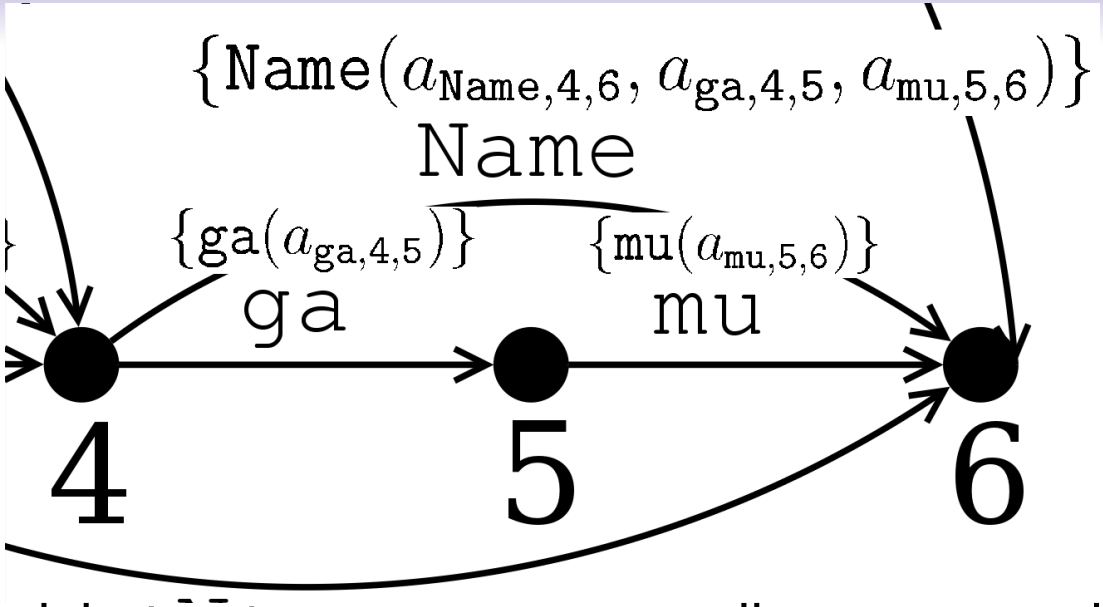
- Argument predykatu jest stałą, której indeksy identyfikują krawędź grafu.

Analiza semantyczna



- Formuły języka reprezentacji znaczenia powiązane są z grafem rozbioru składniowego.
- Pozwala to w zwarty sposób zapamiętać treść niejednoznacznych dokumentów.

Analiza semantyczna



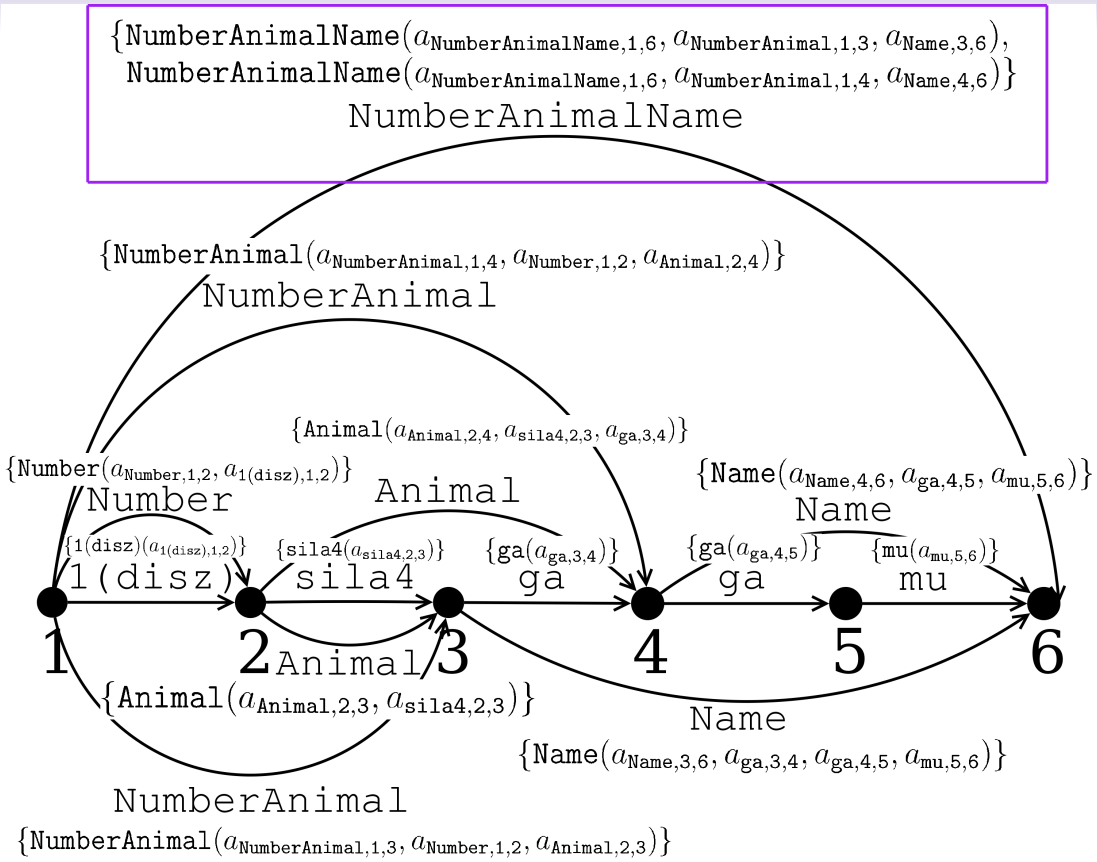
- Postać predykatu utworzonego w wyniku parsowania określa reguła semantyczną przypisaną do reguły gramatycznej.
- Indeksy stałych wiążą ze sobą predykaty

$$[\![\mathtt{Name}]\!]_{4,6} = \mathtt{Name}(a_{\mathtt{Name},4,6}, a_{\mathtt{ga},4,5}, a_{\mathtt{mu},5,6}) \wedge [\![\mathtt{ga}]\!]_{4,5} \wedge [\![\mathtt{mu}]\!]_{5,6} =$$

$$= \mathtt{Name}(a_{\mathtt{Name},4,6}, a_{\mathtt{ga},4,5}, a_{\mathtt{mu},5,6}) \wedge \mathtt{ga}(a_{\mathtt{ga},4,5}) \wedge \mathtt{mu}(a_{\mathtt{mu},5,6})$$

- Powiązanie to odpowiada relacji bycia częścią lub bycia opisem.
- Odniesienie przedmiotowe stałej jest tożsame z odniesieniem przedmiotowym frazy wskazywanej przez indeksy stałej.

Analiza semantyczna



- Formuły języka reprezentacji znaczenia powiązane są z grafem rozbioru składniowego.
- Pozwala to w zwarty sposób zapamiętać treść niejednoznacznych dokumentów.

Analiza semantyczna

$$\{\texttt{NumberAnimalName}(a_{\texttt{NumberAnimalName},1,6}, a_{\texttt{NumberAnimal},1,3}, a_{\texttt{Name},3,6}),$$
$$\texttt{NumberAnimalName}(a_{\texttt{NumberAnimalName},1,6}, a_{\texttt{NumberAnimal},1,4}, a_{\texttt{Name},4,6})\}$$

- Semantyka krawędzi mającej dwa znaczenia

$$[\![\texttt{NumberAnimalName}]\!]_{1,6} = \Big(\texttt{NumberAnimalName}(a_{\texttt{NumberAnimalName},1,6}, a_{\texttt{NumberAnimal},1,3}, a_{\texttt{Name},3,6}) \wedge$$
$$\wedge\ [\![\texttt{NumberAnimal}]\!]_{1,3} \wedge [\![\texttt{Name}]\!]_{3,6} \Big) \vee \Big(\texttt{NumberAnimalName}(a_{\texttt{NumberAnimalName},1,6}, a_{\texttt{NumberAnimal},1,4}, a_{\texttt{Name},4,6}) \wedge$$
$$\wedge\ [\![\texttt{NumberAnimal}]\!]_{1,4} \wedge [\![\texttt{Name}]\!]_{4,6} \Big)$$

- Stała $a_{\texttt{NumberAnimalName},1,6}$ identyfikuje oba znaczenia.

Gramatyka semantyczna korpusu tekstów gospodarczych

- Semantyka jest konstruowana w sposób *kompozycjonalny*, tzn. reguły semantyczne są związane z regułami składniowymi.
- Aplikacja reguły składniowej pociąga za sobą aplikację przypisanej do niej reguły semantycznej, która z semantyk symboli rozpoznanych przez regułę konstruuje semantykę produkcji reguły.
- Gramatyka zawiera 3427 reguł (3 akumulacyjne).
- Z 3055 tabliczek parser wygenerował formułę logiczną mającą 271499 stałych (Liczba stałych jest równa sumie ilości krawędzi grafów rozbioru składniowego).

Typy reguł

- Reguły pełniące rolę leksykonu (1724)

$$\texttt{Name} \quad ::= \quad \texttt{ur mes} \quad \texttt{Name}(var_{\texttt{Name}}, var_{\texttt{ur}}, var_{\texttt{mes}})$$

- Reguły opisujące strukturę tekstu (241)

$$\begin{array}{lcll} \texttt{Name-Job} & ::= & \texttt{Name Job} & \texttt{Name-Job}(var_{\texttt{Name-Job}}, var_{\texttt{Name}}, var_{\texttt{Job}}) \\ \texttt{Name-Job} & ::= & \texttt{Name} & \texttt{Name-Job}(var_{\texttt{Name-Job}}, var_{\texttt{Name}}) \\ \texttt{Name-Job} & ::= & \texttt{Job} & \texttt{Name-Job}(var_{\texttt{Name-Job}}, var_{\texttt{Job}}) \end{array}$$

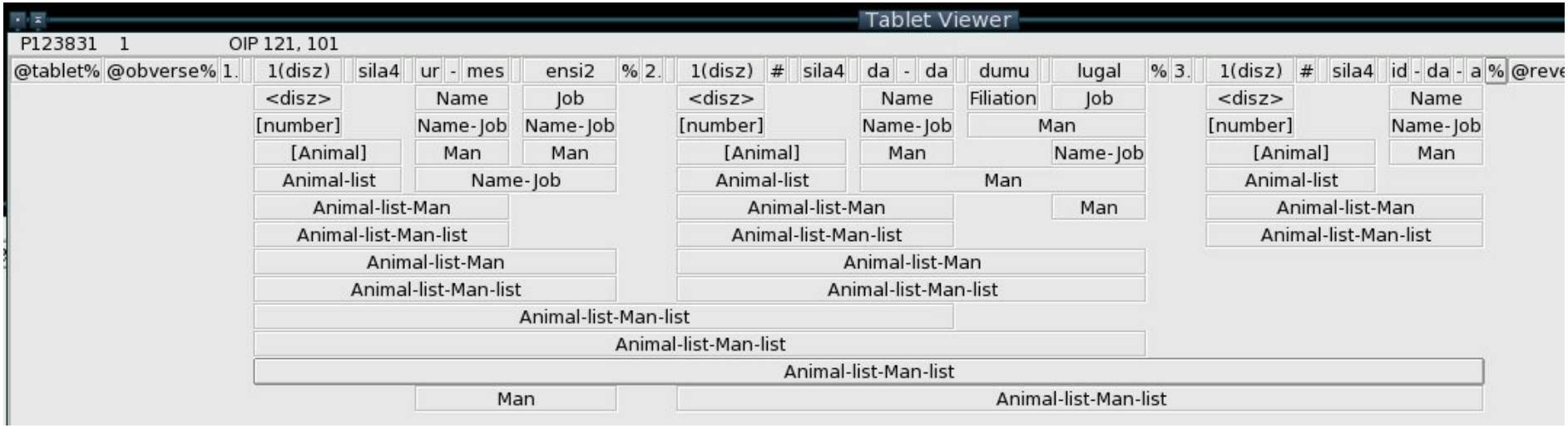
- Reguły wprowadzające wiedzę dziedzinową (liczby, określenia roczne itp) w tych regułach są zawarte algorytmy (1459)

$$\begin{array}{lcl} \texttt{YearGod} & ::= & \texttt{d inanna} \\ & & \{AS04(var_{\texttt{YearGod}}), AS05(var_{\texttt{YearGod}}))\} \\ \texttt{Year} & ::= & \texttt{mu en YearGod Verb} \\ & & \{p(var_{\texttt{Year}}) : p \in val_{\texttt{YearGod}} \cap val_{\texttt{Verb}}\} \end{array}$$

- $var_{\texttt{Name}}$ jest zmienną, której indeksem jest krawędź grafu związana z symbolem `Name`
- $val_{\texttt{YearGod}}$ jest zbiorem nazw predykatów związanych z symbolem `YearGod`.

Generowanie zbioru reguł

- Tworzenie gramatyki, semantyki i ontologii odbywa się jednocześnie.
- Zastosowana jest metodologia *active learning*.



- System posiada interfejs wizualizujący grafy rozbioru składniowego i ułatwiający pisanie reguł.
- System pozwala użytkownikowi odszukiwać często powtarzające się ciągi symboli i tworzyć reguły na ich podstawie.
- Pozwala to szybko stworzyć gramatykę obejmującą typowe fragmenty tekstów.

Plan referatu

1. Język
2. Parsowanie
3. Zastosowania
4. Rozszerzenia

Wyszukiwanie tekstów w korpusie

- Wyszukiwanie tekstów zawierających dany ciąg odczytów znaków.
- Wyszukiwanie tekstów zawierających dany ciąg znaków.
- Wyszukiwanie tekstów wg. wzorca semantycznego: znajdowanie tekstów, które posiadają frazę o zadanym znaczeniu.

Model gospodarki sumeryjskiej

- Podział tekstów na transakcje, zbudowanie na ich podstawie modelu gospodarki sumeryjskiej

Data	Ilość	Towar	Dostawca	Odbiorca
23-11-AS05	1	sila4	ur-mes ensi2	ab-ba-sa6-ga
23-11-AS05	1	sila4	da-da dumu lugal	ab-ba-sa6-ga
23-11-AS05	1	sila4	id-da-a	ab-ba-sa6-ga

- Prosopographia, czyli ustalenie tożsamości osobowej urzędników sumeryjskich.
- Sklasyfikowanie typów dokumentów generowanych przez administrację sumeryjską.

Model gospodarki sumeryjskiej

P122592	1 sila4 niga	ab-ba-sa6-ga	lu2-dingir-ra	2	4 AS02
P124900	2 ab2 mu-2	ab-ba-sa6-ga	xxx	2	4 AS02
P124900	3 sila4	ab-ba-sa6-ga	xxx	2	4 AS02
P102183	4 udu niga	ab-ba-sa6-ga	na-lu5	4	4 AS02
P131562	4 udu niga	ab-ba-sa6-ga	na-lu5	4	4 AS02
P128138	2 amar masz-da3	ab-ba-sa6-ga	lu2-dingir-ra dumu ARAD2-hul3-la	6	4 AS02
P130395	2 gu4	ab-ba-sa6-ga	xxx	8	4 AS02
P130395	3 ab2	ab-ba-sa6-ga	xxx	8	4 AS02
P103081	1 udu niga	ab-ba-sa6-ga	lu2-dingir-ra	9	4 AS02
P128809	2 munus-asz2-gar3	ab-ba-sa6-ga	na-lu5	11	4 AS02
P131562	2 munus-asz2-gar3	ab-ba-sa6-ga	na-lu5	11	4 AS02
P128809	4 udu niga	ab-ba-sa6-ga	na-lu5	11	4 AS02
P131562	4 udu niga	ab-ba-sa6-ga	na-lu5	11	4 AS02
P131562	3 udu	ab-ba-sa6-ga	na-lu5	12	4 AS02
P131562	16 udu niga	ab-ba-sa6-ga	na-lu5	12	4 AS02
P131562	1 munus-asz2-gar3	ab-ba-sa6-ga	na-lu5	13	4 AS02
P131562	1 sila4	ab-ba-sa6-ga	na-lu5	13	4 AS02
P131562	1 u8 niga	ab-ba-sa6-ga	na-lu5	21	4 AS02
P131562	1 sila4	ab-ba-sa6-ga	na-lu5	21	4 AS02
P131562	1 munus-asz2-gar3	ab-ba-sa6-ga	na-lu5	21	4 AS02
P131562	3 masz2	ab-ba-sa6-ga	na-lu5	21	4 AS02
P131562	3 sila4	ab-ba-sa6-ga	na-lu5	21	4 AS02
P103866	4 udu niga	ab-ba-sa6-ga	na-lu5	21	4 AS02
P131562	6 udu niga	ab-ba-sa6-ga	na-lu5	21	4 AS02
P131562	4 masz2	ab-ba-sa6-ga	na-lu5	22	4 AS02
P131562	1 sila4 niga	ab-ba-sa6-ga	na-lu5	22	4 AS02
P131562	1 sila4	ab-ba-sa6-ga	na-lu5	22	4 AS02
P130511	1 masz2	|PU3.SZA|-ha-ni lu2-kas4	ab-ba-sa6-ga	23	4 AS02
P136299	4 masz2	ab-ba-sa6-ga	xxx	28	4 AS02
P131562	5 gukkal	ab-ba-sa6-ga	na-lu5	29	4 AS02
P103020	5 gukkal	ab-ba-sa6-ga	na-lu5	29	4 AS02
P103020	1 u8 a-lum	ab-ba-sa6-ga	na-lu5	29	4 AS02
P131562	1 u8 a-lum-masz2	ab-ba-sa6-ga	na-lu5	29	4 AS02
P103020	2 udu a-lum	ab-ba-sa6-ga	na-lu5	29	4 AS02
P131562	2 udu a-lum-udu	ab-ba-sa6-ga	na-lu5	29	4 AS02
P103020	1 munus-asz2-gar3 niga	ab-ba-sa6-ga	na-lu5	29	4 AS02
P131562	1 munus-asz2-gar3 niga	ab-ba-sa6-ga	na-lu5	29	4 AS02
P103020	2 udu	ab-ba-sa6-ga	na-lu5	29	4 AS02
P103020	2 masz2	ab-ba-sa6-ga	na-lu5	29	4 AS02
P128097	2 masz-da3	ab-ba-sa6-ga	lu2-dingir-ra dumu ARAD2-hul3	29	4 AS02
P106179	200 udu	na-ra-am-i3-li2	ur-{d}-lamma ensi2 gir2-su-{ki}	xxx	4 AS02
P106179	8 gu4	na-ra-am-i3-li2	ur-{d}-lamma ensi2 gir2-su-{ki}	xxx	4 AS02
P106179	2 ab2	na-ra-am-i3-li2	ur-{d}-lamma ensi2 gir2-su-{ki}	xxx	4 AS02
P126015	6 udu szimaszki niga	ab-ba-sa6-ga	xxx	xxx	4 AS02

Znajdowanie par powiązanych tekstów

&P131562 = TCL 2, 4683
@tablet
@obverse
1. 4(disz) udu niga u4 4(disz)-kam
2. 4(disz) udu niga 2(disz) {munus}asz2-gar3
3. u4 1(u) 1(disz)-kam
4. 1(u) 6(disz) udu niga 3(disz) udu
5. u4 1(u) 2(disz)-kam
6. 5(disz) udu 4(disz) masz2
7. 1(disz) {munus}asz2-gar3 1(disz) sila4
8. u4 1(u) 3(disz)-kam
9. 3(disz) sila4 2(disz) masz2
10. 1(disz) masz2-gaba u4 1(u) 5(disz)-kam
11. 1(disz) udu 2(disz) sila4
12. 1(disz) masz2 u4 1(u) 6(disz)-kam
13. 2(disz) sila4 1(disz) masz2 u4 1(u) 7(disz)-kam
14. 1(u) 5(disz) masz2 gal u4 1(u) 8(disz)-kam
15. 2(disz) udu niga 1(disz) masz2 gal niga
16. 1(disz) sila4 2(disz) masz2
17. u4 2(u) la2 1(disz)-kam
18. 2(disz) udu niga 1(disz) u8 niga
19. 1(disz) udu 3(disz) sila4
20. 3(disz) masz2 1(disz) {munus}asz2-gar3
21. 6(disz) udu niga 1(disz) sila4
22. 1(disz) masz2 u4 2(u) 1(disz)-kam
@reverse
1. 1(disz) sila4 niga 1(disz) sila4 4(disz) masz2
2. u4 2(u) 2(disz)-kam
3. 2(disz) udu 3(disz) [masz2 u4] 2(u) 3(disz)-kam
4. 3(disz) sila4 3(disz) masz2 [u4] 2(u) 4(disz)-kam
5. 1(disz) {munus}asz2-gar3 niga 5(disz) gukkal
6. 2(disz) udu a-lum udu
7. 1(disz) u8 a-lum masz2
8. u4 3(u) la2 1(disz)-kam
9. |SZU+LAGAB| 3(u) 4(disz) udu niga 1(disz) sila4 niga
10. |SZU+LAGAB| 1(disz) u8 niga 1(disz) masz2 gal niga
11. |SZU+LAGAB| 1(disz) {munus}asz2-gar3 niga 1(u) 4(disz) udu
12. |SZU+LAGAB| 5(disz) gukkal 2(disz) udu a-lum
13. |SZU+LAGAB| 1(u) 7(disz) sila4 1(disz) u8 a-lum
14. |SZU+LAGAB| 1(u) 5(disz) masz2 gal 2(u) 6(disz) masz2
15. |SZU+LAGAB| 1(disz) masz2-gaba 4(disz) {munus}asz2-gar3
16. 2(gesz2) 3(disz)
17. mu-DU lugal
18. ki ab-ba-sa6-ga-ta
19. na-lu5 i3-dab5
20. iti ki-siki {d}nin-a-zu
21. mu {d}amar-{d}suen lugal-e ur-bi2-lum{ki} mu-hul

&P102183 = ASJ 07, 122 05
@tablet
@obverse
1. 4(disz) udu niga
2. u4 4(disz)-kam
3. ki ab-ba-sa6-ga-ta
4. na-lu5
@reverse
1. i3-dab5
2. iti ki-siki {d}nin-a-zu
3. mu {d}amar-{d}suen lugal-e ur-bi2-lum{ki} mu-hul
@left
1. 4(disz)

&P103866 = AUCT 2, 048
@tablet
@obverse
1. 4(disz) udu niga
2. u4 2(u) 1(disz)-kam
3. mu-DU lugal
4. ki ab-ba-sa6-ga-ta
@reverse
1. na-lu5 i3-dab5
$ (blank line)
2. iti ki-siki {d}nin-a-zu
3. mu {d}amar-{d}suen lugal-e ur-bi2-lum{ki} mu-hul
@left
1. 4(disz)

&P128809 = SACT 1, 054
@tablet
@obverse
1. 4(disz) udu niga
2. 2(disz) {munus}asz2-gar3
3. u4 1(u) 1(disz)-kam
4. mu-DU lugal
5. ki ab-ba-sa6-ga-ta
@reverse
1. na-lu5
2. i3-dab5
$ (blank line)
3. iti ki-siki {d}nin-a-zu
4. mu {d}amar-{d}suen* lugal-e ur-bi2-lum{ki} mu-hul
@left
1. 6(disz)

&P103020 = AUCT 1, 174
@tablet
@obverse
1. 1(disz) {munus}asz2-gar3 niga
2. 5(disz) gukkal
3. 2(disz) [udu] a-lum
4. 2(disz) udu
5. 1(disz) u8 a-lum
6. 2(disz) masz2
7. u4 3(u) la2 1(disz)-kam
@reverse
1. mu-DU lugal
2. ki ab-ba-sa6-ga-ta
3. na-lu5 i3-dab5
4. iti ki-siki {d}nin-a-zu
5. mu {d}amar-{d}suen lugal-e ur-bi2-lum{ki} mu-hul
@left
1. 1(u) 3(disz)

Plan referatu

1. Język
2. Parsowanie
3. Zastosowania
4. Rozszerzenia

Niewiedza

- Bogactwo języka naturalnego powoduje, że treść pewne liczby fraz nie daje się wyrazić za pomocą powszechnie występujących kategorii ontologicznych.
- Frazy te są rzadkie, nieregularne i pozbawione charakterystycznych kontekstów.
- Jest ich zbyt dużo, by je przejrzeć i dodać do ontologii pojęcia w nich zawarte.
- Aby parsować częściowo niezrozumiałe teksty domykam gramatykę i ontologię wprowadzając kategorię 'nieznane'.
- Początek i koniec nieznanej frazy określam uznając, że każda nieznana fraza zajmuje jeden wers (Koniec wersu jest końcem słowa, zwykle też końcem frazy).
- Otrzymuję system zdolny sparsować tekst niezależnie swojej znajomości języka.
- Stan zrozumienia decyduje o ilości informacji z niego wydobytych.

Akumulacja

- Semantyka produkcji reguły akumulacyjnej jest grafem, którego wierzchołkami są zmienne wskazujące na semantyki symboli rozpoznanych przez tą regułę, a każda jego ścieżka jest możliwą semantyką produkcji reguły.
- Formułą dla grafu struktury logicznej symbolu akumulacyjnego jest alternatywa po wszystkich ścieżkach grafu predykatów, których argumentami są zmienne zawarte w danej ścieżce.

Uszkodzenia

Techniki analizy uszkodzeń:

- Uznawanie uszkodzonych wersów jako nieznanych.
- Interpretacja poszczególnych nieczytelnych znaków jako symboli, do których parser może dopasować dowolny symbol.
- Zastąpienie uszkodzonego wersu przez listę wersów typowych.

Analizator składniowy

- Analizator składniowy jest podobny do parserów tablicowych.
- Podobnie jak one reprezentuje tekst za pomocą grafu skierowanego acyklicznego.
- Różnica polega na dostosowaniu mojego parsera do języków regularnych (z rozszerzeniem na bezkontekstowe mogą być problemy wydajnościowe) oraz dostosowaniu do korzystania z niekompletnych gramatyk, co ułatwia uczenie.