



University of Latvia
Institute of Mathematics
and Computer Science

Teksta semantiska marķēšana ar likumos bāzētu mašīnmācīšanās metožu palīdzību

Didzis Goško

Vadītājs: Prof. Guntis Bārzdīņš

LUMII AILab, 2015

Semantika

Semantiku definē kā:

1. Valodas vienības nozīme (jēga), nozīmju kopums
2. Valodniecības nozara, kas pētī valodas vienību nozīmes; arī semasioloģija.
3. Semiotikas nozare, kas zīmju sistēmas pētī kā jēgas izteikšanas līdzekļus.

Teksta semantiska marķēšana – nozīmes (jēgas) izgūšana no teksta; citiem vārdiem – faktu atrašana tekstā.

Situāciju semantika

«Freimu» semantika (situāciju semantika) – lingvistikas teorija, kas apgalvo, ka gandrīz visas ikdienā sastopamās situācijas var aprakstīt ar galīgu skaitu situāciju šabloniem – «freimiem».

Semantiskais «freims» ir fakts un tā parametru kopa, kas raksturo noteiktu situāciju.

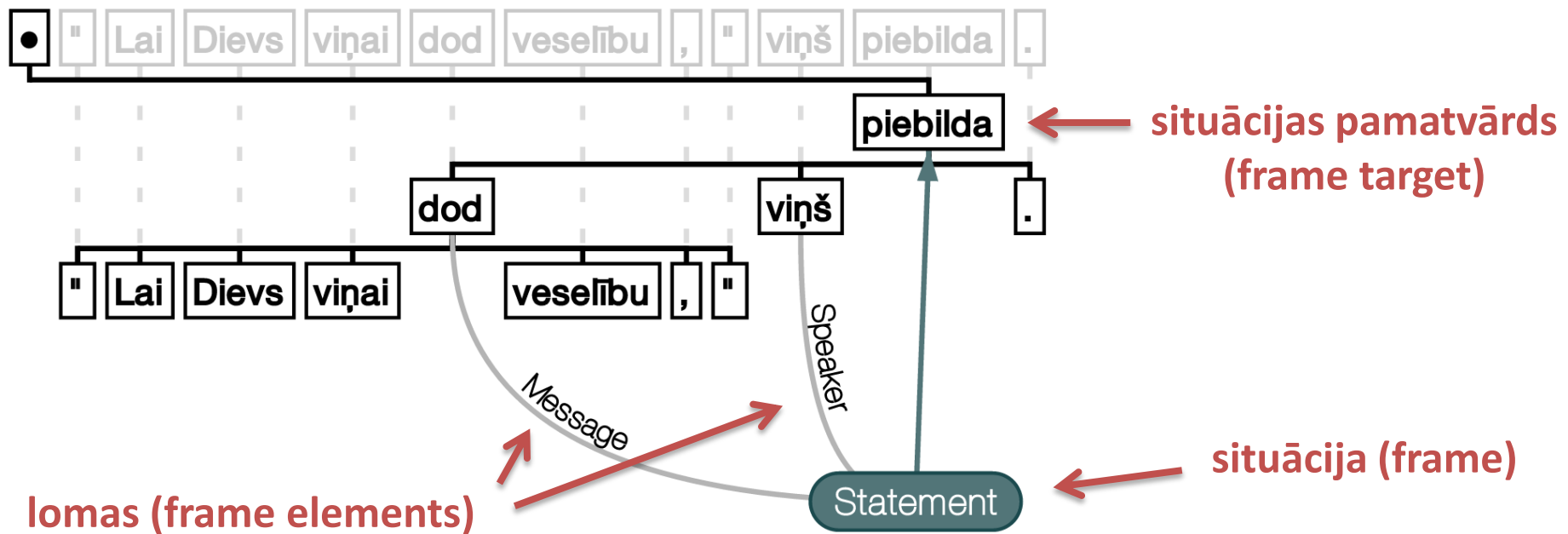
FrameNet^[1] – projekts, kura ietvaros ir izveidots elektroniskais resurss freimu semantikas aprakstam; tas ietver freimu un to lomu (elementu) sadalījumu, kā arī ar roku anotētus piemēra teikumus angļu valodai. Angļu FrameNet 1.5 ir 1019 freimi.

[1] Christopher R. Johnson, Miriam R. L. Petruck, Collin F. Baker, Michael Ellsworth, Josef Ruppenhofer, Charles J. Fillmore. FrameNet: Theory and Practice.

<http://nats-www.informatik.uni-hamburg.de/pub/CDG/FrameNet/book.pdf>

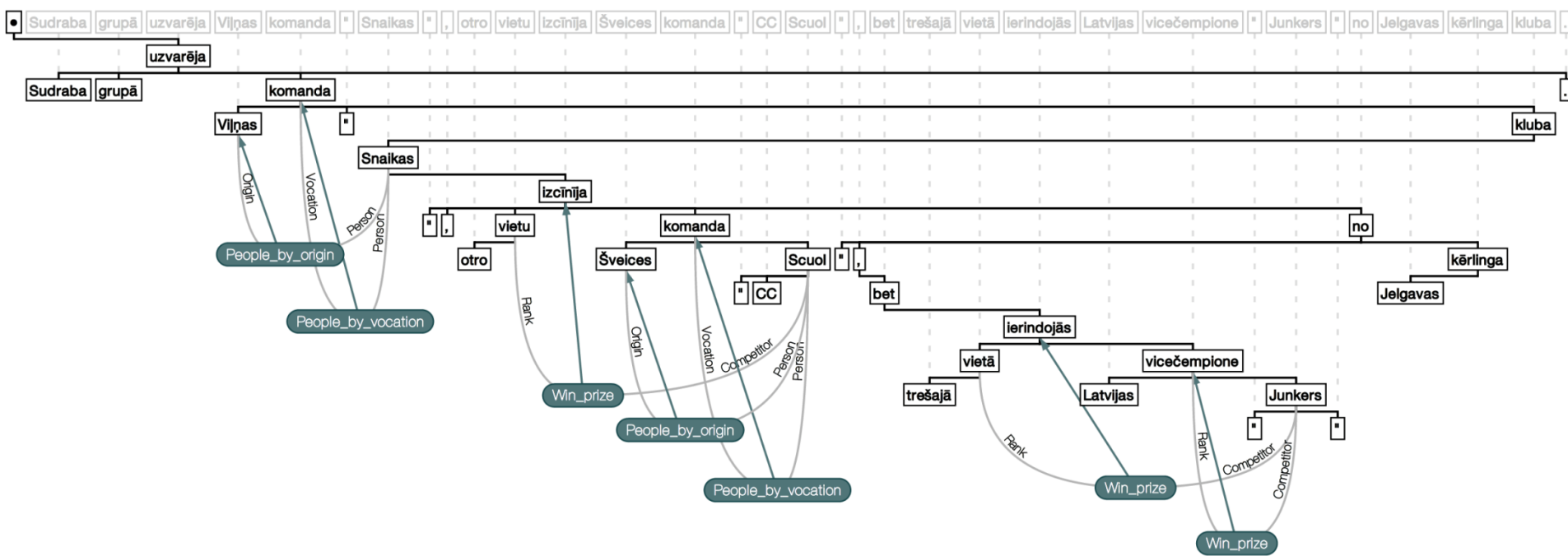
FrameNet

Teksta semantiskā marķēšana – FrameNet
situāciju (**frame**) un to lomu (**frame element**)
atrašana dotajā tekstā



...

Mērķis – veikt šādu marķēšanu automātiski!



Metodika

Pamatideja: semantisko situāciju un to lomu atrašanai izmantot likumos bāzētu klasifikātoru.

Klasifikātorā sekmīgai apmācībai ir nepieciešama teksta bagātināšana (anotēšana) ar papildinformāciju, ko ir iespējams noteikt automātiski.

Teksta anotēšana

Kādu papildinformāciju no teksta ir iespējams izgūt automātiski ar relatīvi augstu precizitāti (virs 75%) ?

Ir iespējams izgūt automātiski:

- katra vārda pamatformu (lemmu)
- katra vārda morfoloģisko marķējumu – vāršķiru, locījumu, skaitli utt. (angliski: Part-Of-Speech jeb POS marķējums)
- sintakses koku – raksturo sintaktisko struktūru hierarhiju
- nosaukto vienību atpazīšana (Named Entity Recognition) – laiks, vieta, persona, organizācija, ...
- koreferenču atrašana – tekstā viena un tā pati persona/vienība minēta dažādos veidos (piemēram, ar vietniekvārdu palīdzību)

LUMII AILab semantisko rīku komplekts^[1]

LUMII AILab ir izstrādāts semantisko rīku komplekts latviešu valodas teksta anotēšanai:

- morfoloģiskās informācijas ģenerators (pamatforma, vārdšķiras, ...)
- nosaukto vienību atpazīnējs (Named Entity Recognition) – tekstā atpazīst un klasificē personas, organizācijas, vietas, kā arī laikus un daudzumus.
- sintakses koka ģenerators
- koreferenču risinātājs – konkrēto vienību pieminējumu savilkšana neskatoties uz pieminējuma formu un veidu

[1] Normunds Grūzītis, Pēteris Paikens, Guntis Bārzdīņš. 3rd Workshop on Controlled Natural Language (CNL), 29-31 August 2012, Zurich

http://attempto.ifi.uzh.ch/site/cnl2012/slides/gruzitisetal_framenet.pdf

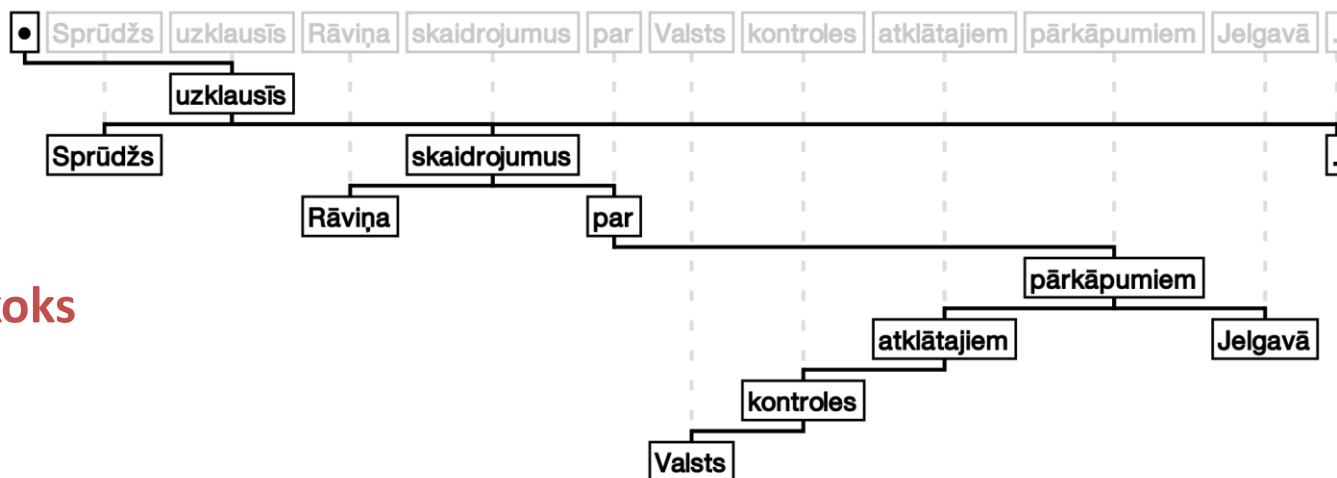
Sprūdžs uzklausīs Rāviņa skaidrojumus par Valsts kontroles atklātajiem pārkāpumiem Jelgavā.

Index	Form (Word)	Lemma	POS	Tag	Features	Parent	Named Entity Type	ID Type	Named Entity ID
1	Sprūdžs	sprūdžs	n	n_msn1	locijums=Nominatīvs dzimt...	2	O		
2	uzklausīs	uzklausīt	v	v_nif_330an	avota_pamatforma=uzklau...	0	O		
3	Rāviņa	Rāviņa	n	n_msg_	locijums=Ģenitīvs dzimte=...	4	person		
4	skaidrojums	skaidrojums	n	n_mpa1	avota_pamatforma=skaidr...	2	O		0
5	par	par	s	sppdn	avota_pamatforma=par re...	4	O		
6	Valsts	valsts	n	ncfsg6	avota_pamatforma=valsts ...	7	organization		
7	kontroles	kontrole	n	ncfsg5	avota_pamatforma=kontro...	8	organization	organization	1
8	atklātajiem	atklāt	v	vmnpdmpdpsy	transitivitate=Pārejošs var...	9	O		
9	pārkāpumiem	pārkāpums	n	ncmpd1	avota_pamatforma=pārkā...	5	O		
10	Jelgavā	Jelgava	n	npfsl4	avota_pamatforma=jelgav...	9	location	location	2
11	.	.	z	zs	avota_pamatforma=. vard...	2	O		

morfologija
sintakse
nosaukumi
anaforas

nosaukumi

Sprūdžs uzklausīs **Rāviņa** skaidrojumus par **Valsts** **kontroles** atklātajiem pārkāpumiem **Jelgavā**.



atkarību sintakses koks

Latvian FrameNet

Vietējās (Latvijas) ziņu aģentūras vajadzībām tika izvēlētas 26 būtiskākās situācijas («freimi»), ar kurām tika sagatavota relatīvi maza treniņdatu kopa – treniņkorpuss.

Treniņkorpusa marķēšana tika veikta ar roku izmantojot speciāli šim mērķim izstrādātu programmatūru ar grafisku saskarni.

Angļu FrameNet 1.5 satur piemēru kopu, kuru var izmantot kā treniņkorpusu, kā arī ir pieejama SemEval 2007 marķētu datu kopa.

Latvian FrameNet

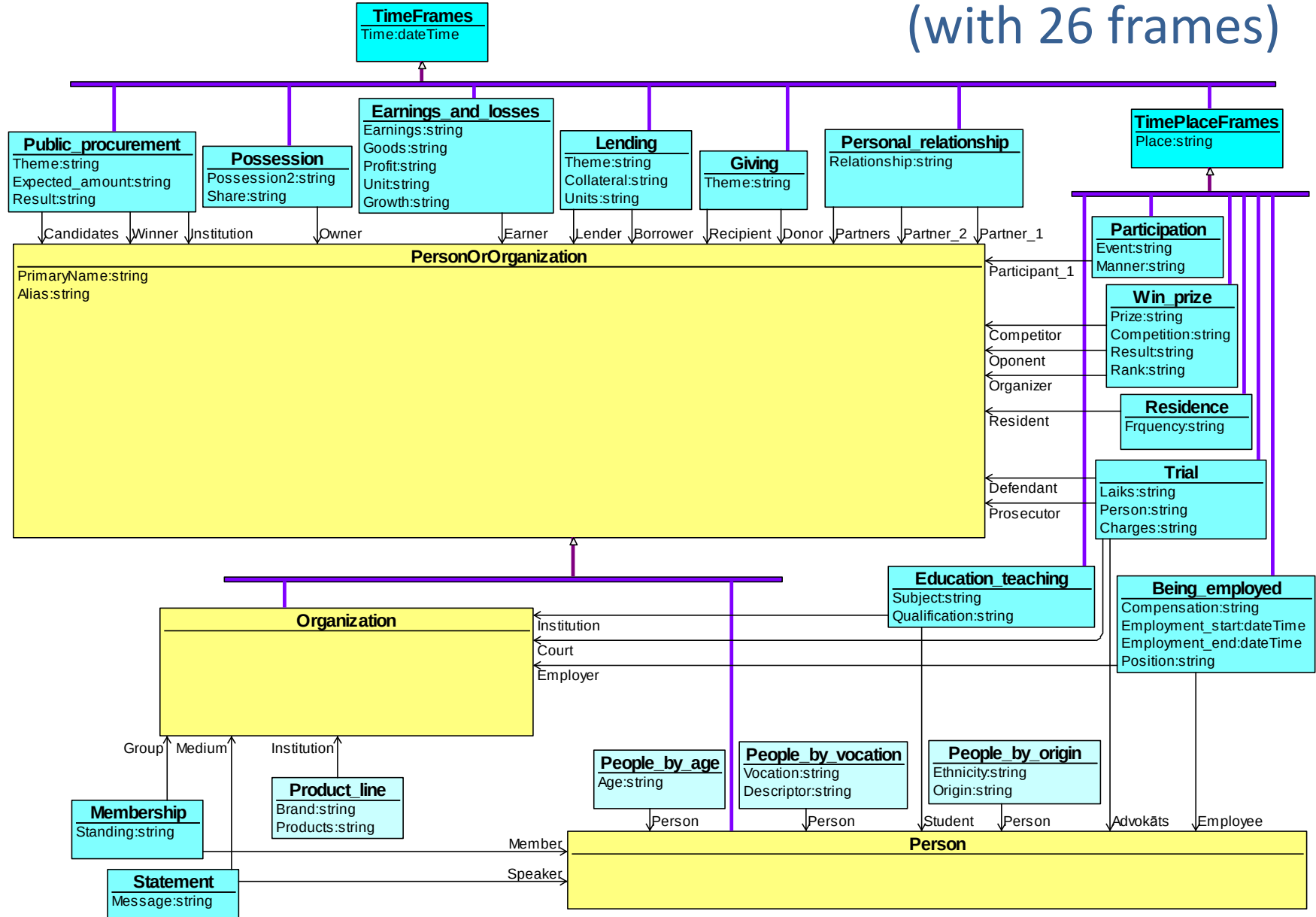
	FrameNet 1.3 (English)	FrameNet 1.5 (English)	FrameNet LV (Latvian)
Frame types	665 (795)	877 (1019)	26
FrameElement types	720	1068	80
Training sentences with full annotation	2198	3256	4079
Training sentences with single frame annotation	139439	154607	–
Test sentences with full annotation	120	2420	844



SemEval-2007,
Task19 dataset

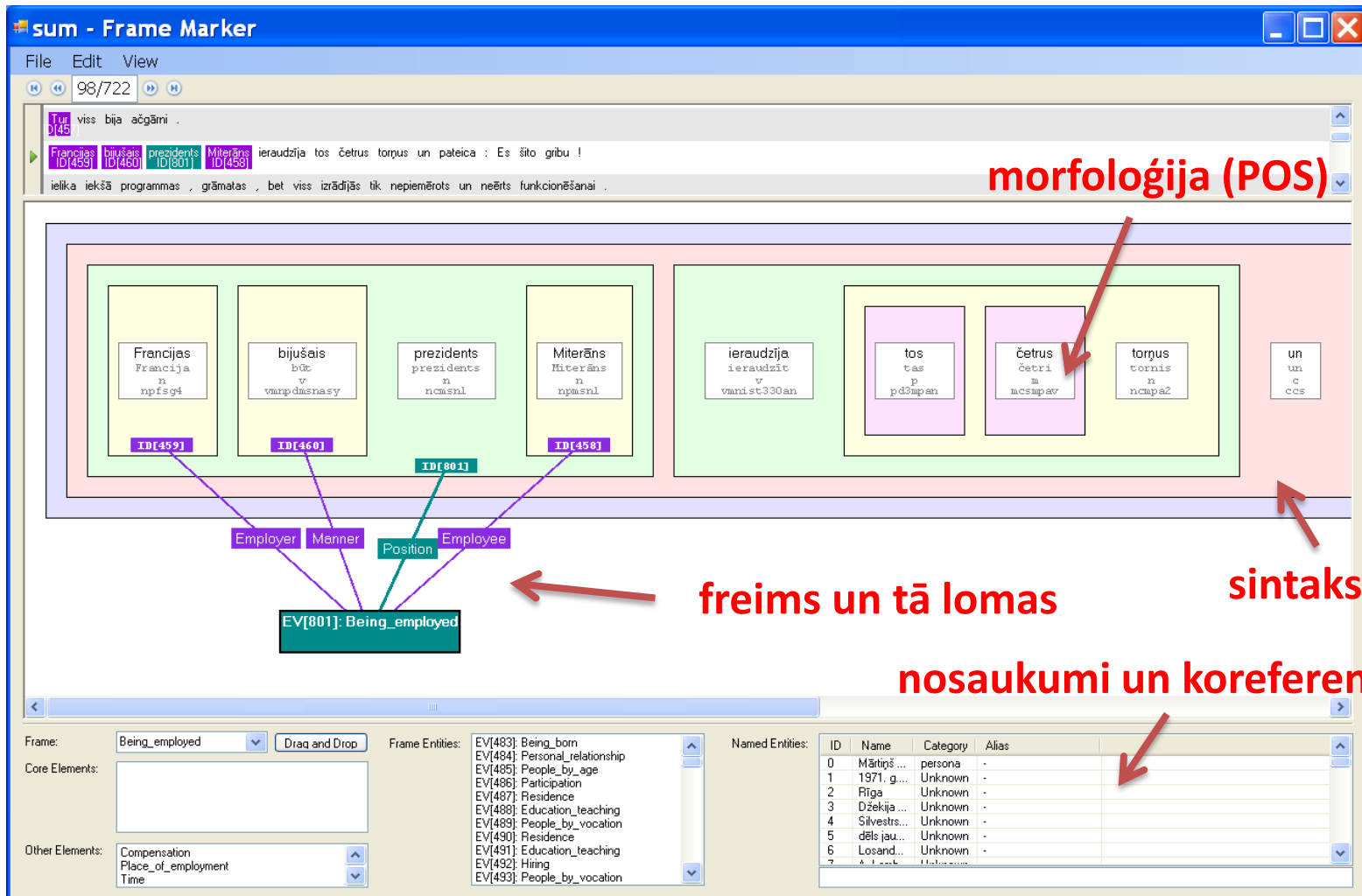
Abstract Knowledge Representation

(with 26 frames)



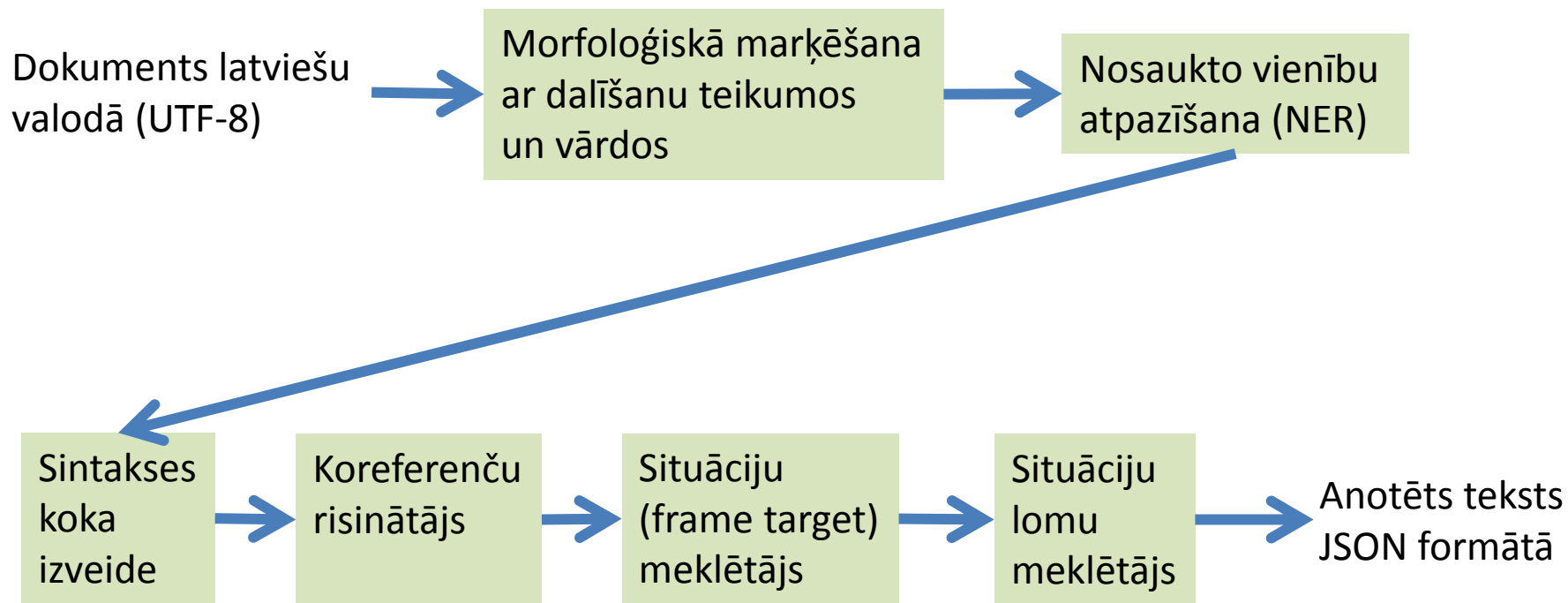
Latviešu FrameNet treniņkorpora marķēšanas rīks

Treniņdatu sagatavošanas nolūkos tika izstrādāts freimu redaktors^[1]:



[1] D. Brēdiks, FrameNET semantiskās anotēšanas redaktors ar koreferenšu identifikāciju, maģistra darbs, Latvijas Universitāte, 2013.

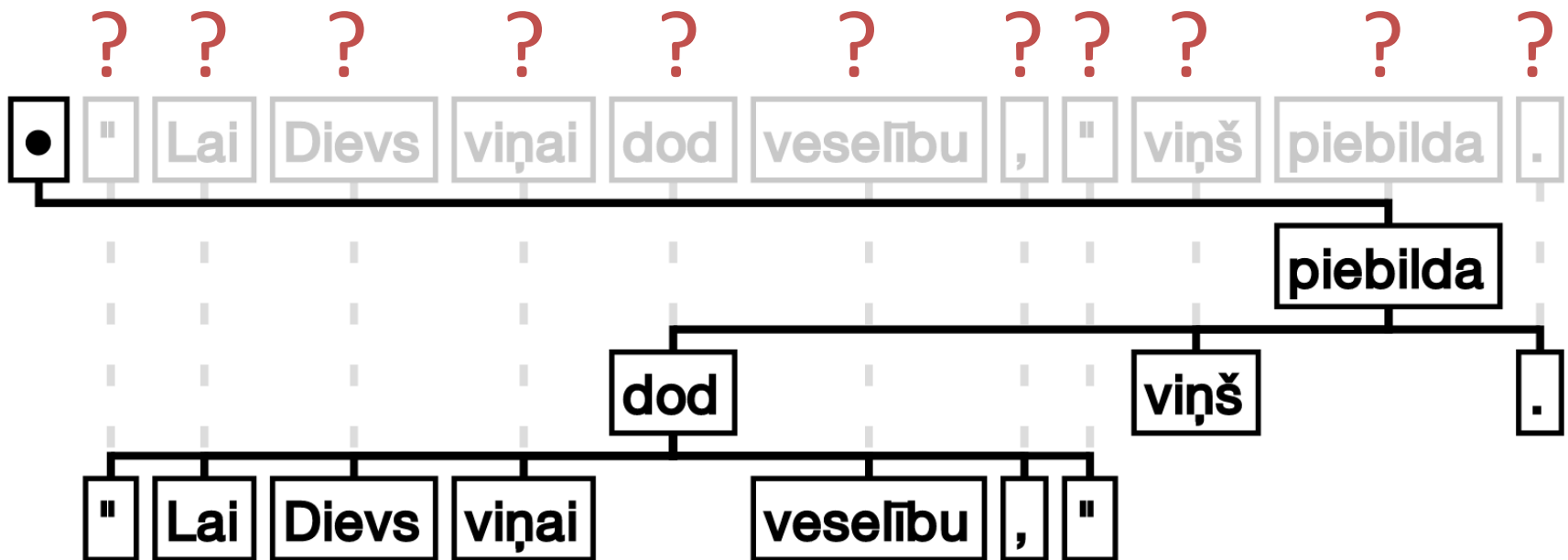
LUMII AILab FrameNet anotēšanas shēma latviešu valodai



Anotētais teksts nonāk datubāzē, kur glabājas citas, arī identiskas, situācijas no citiem teksta avotiem; identiskās situācijas tiek grupētas kopā.

Klasifikācija

Vai kāds no šiem «vārdiem» raksturo kādu situāciju ?
Varbūt kādas situācijas kādu no lomām ?



Klasifikatora apmācība

Klasifikatora uzdevums ir datiem (to fragmentiem) piekārtot kādu no definētajām klasēm balstoties uz likumsakarībām šajos datos.

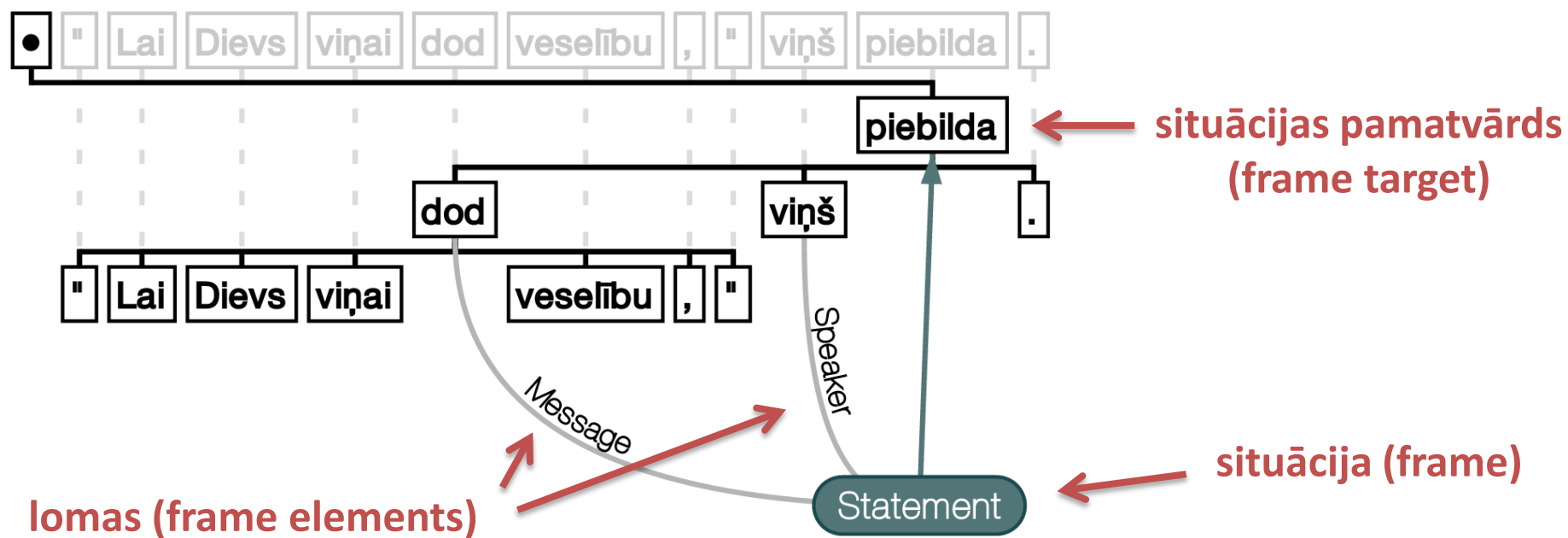
Likumsakarības izsaka kā kombinācijas no pazīmēm, kas izgūtas no datiem, tās var būt gan pašu datu fragmenti, gan to pārveidojumi un to dažādas kombinācijas.

Tipisks pazīmju komplekts teksta analīzē vienam vārdam:

- vārds tādā formā, kādā tas ir sastopams tekstā
- vārda pamatforma (lemma)
- vārda morfoloģiskā informācija: vārdšķira, locījums utt.
- vārda konteksts – blakus esošie vārdi tuvākajā apkārtnē, šo vārdu pazīmes
- vieta sintakses kokā – tuvumā esošo vārdu pazīmes
- nosauktās vienības (entity) klase
- u.c.

Vēlamais rezultāts

- Situācijas (kādas?) pamatvārda (ierosinātājevārda) noteikšana.
- Atbilstošās situācijas lomu noteikšana.



Likumos bāzēti klasifikātori

Izmantotie likumu bāzētie klasifikātori:

- C4.5/C5.0^[1] klasifikātors Ross Quinlan (decision-trees, rules) ar manuālu likumu korekciju
- Pašdarināts pilnās pārlases binārais klasifikātors (aka C6.0^[2])
- Pašdarināts optimizēts daļējas pārlases klasifikātors

1. <http://www.rulequest.com/see5-info.html>

2. <http://c60.ailab.lv>

Quinlan C4.5/C5.0 klasifikātors

Izmantojot Quinlan C4.5/C5.0^[1] decision-tree bāzētu klasifikātoru:

- iespēja decision koka vietā ģenerēt vienkāršus likumus;
- decision-tree/likumu ģenerēšana darbojas gandrīz momentāni;
- ātra iegūto (satrenēto) likumu izpilde (produkcijas apstrāde);
- laba sākotnējā precizitāte izmantojot relatīvi mazu treniņkorpusu;
- iegūtie (satrenētie) likumi ir viegli pārskatāmi, interpretējami un labojami;
- ar relatīvi nelielu manuālu likumu labošanu var būtiski uzlabot rezultātus;
- likumi pēc savas būtības parāda, kuras pazīmes ir būtiskas, kuras nav.

[1] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

Pilnās pārlases binārais klasifikātors (C6.0^[1])

1. Pašdarināts klasifikātors, kas veido likumus balstoties uz katra treniņdatos sastopamā pozitīvā datu piemēra visu pazīmju relaksāciju.
2. Iegūtie likumi tiek mērīti ar Laplasa mēru attiecībā pret visu treniņkorpusu dodot skaitu, uz cik piemēriem likums ir nostrādājis pareizi un uz cik piemēriem – nepareizi.
3. Pēc iegūtā mēra likumi tiek kārtoti un atmesti tie, kas tiek uzskatīti par nepareiziem – pārāk vispārīgiem.
4. Pozitīvais: var iegūt labākus rezultātus kā ar C5.0 klasifikātoru, rezultātā (uzstādītajam mērķim) var iztikt bez manuālas likumu labošanas
5. Ierobežojumi:
 - a. nepieciešamā aprēķinu jauda ir proporcionāla datu apjomam
 - b. tikai binārs klasifikātors
 - c. likumi darbojas ar precīzām simbolu virkņu pazīmēm
 - d. neaizstāj C5.0 klasifikātoru (C5.0 ir daudzklešu vispārīga pielietojuma klasifikātors)

Iegūto likumu paraugi

Laplasa mērs $(n-m+1)/(n+2)$ palīdz novērst iegūto likumu pārāk lielu specializāciju vai vispārīgumu (overfitting/underfitting)

(n – kopējais precedentu skaits, m – negatīvo precedentu skaits)

– reālistisks likumu kvalitātes mērs uz vēl neredzētiem datiem

FRAME TARGET: Revenge

	<i>n</i>	<i>m</i>	<i>Laplace ratio</i>
[_, _, _, _, {retaliation.n.1, punish.v.1, revengeful.s.1}, _, _, _, _, _]	193	9	95%
[_, _, _, {avenger, retaliated, retaliate, avenged}, _, _, _, _, _]	49	0	98%
[_, MD, _, get, _, _, _, _, RB, _]	23	3	84%
[_, JJ, _, sanction, _, _, _, _, _]	4	0	83%
[_, _, _, sanction, _, NNS, _, _, IN, _]	5	1	71%
[_, _, #NONE#, sanction, _, _, _, ', ', _, _]	2	0	75%

Pazīmes: PLEMMMA, PPOS, PNER, LEMMA, HYPERNYM, POS, NER, NLEMMMA, NPOS, NNER

Validācija uz treniņdatiem: Correct YES = 206; Correct NO = 56473; False YES = 10; False NO = 8

Priekšrocības

Pēc trenēšanas tiek iegūts:

- uz likumiem bāzēts (rule based) klasifikātors – vienkāršs un ātrdarbīgs
 - likumus var lasīt un labot cilvēks, piemēram:
 - LEMMA = iet
 - PLEMMA = opera AND LEMMA = darbinieks
 - POS = lietvārds AND TYPE = profession
- utt.

Rezultātu precizitāte (ar C5.0)

Latviešu valodas FrameNet marķēšanas precizitātes novērtējums un salīdzinājums ar state-of-art angļu valodai^[1].

	<i>Latvian FrameNet</i>	<i>English FrameNet 1.3</i>
Treniņkorpusa izmērs teikumos	1682	139439
Freimu skaits	26	665
Lomu (elementu) skaits	80	720
<i>Testa kopas izmērs teikumos</i>	199	120

<i>Frame Target recognition</i>	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
C5.0 manual boosting (Latvian FrameNet data)	55,6	43,9	49,1
LTH (English SemEval'07 data)	66,2	50,6	57,3
CMU/Google (English SemEval'07 data)	69,7	54,9	61,4

<i>Frame Element recognition</i>	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
C5.0 manual boosting (Latvian FrameNet data)	59,4	43,3	50,1
LTH (English SemEval'07 data)	51,6	35,4	42,0
CMU/Google (English SemEval'07 data)	58,1	38,8	46,5

[1] Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider. Frame-Semantic Parsing. Computational Linguistics, 2013.

Rezultāti ar pilnās pārlases klasifikātoru (latviešu, angļu valodām)

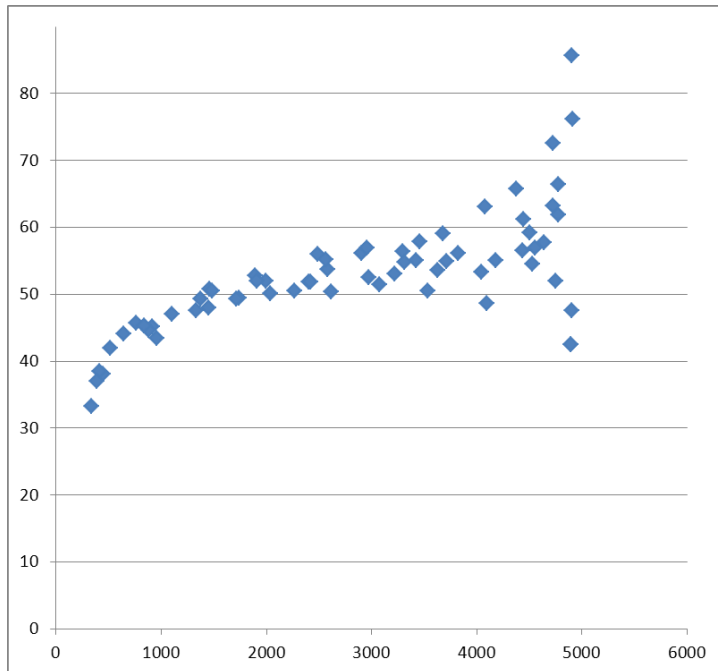
<i>English FrameNet 1.3 SemEval-2007 dataset</i>	<i>Frame Target identification</i>			<i>Frame Element identification</i>		
	<i>Precision</i>	<i>Recall</i>	<i>F1</i>	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
LTH ¹⁾	68.9	53.6	60.3	51.6	35.4	42.0
SEMAFOR/Google ²⁾	69.7	54.9	61.4	58.1	38.8	46.5
C6.0 RuleSet EN	77.1	53.7	63.3	47.3	47.0	47.1
C6.0 RuleSet LV	63.5	62.7	63.1	65.9	76.8	70.9

- 1) Johansson, R., Nugues, P. **(2007)**. LTH: semantic structure extraction using nonprojective dependency trees. In *Proceedings of SemEval-2007: 4th International Workshop on Semantic Evaluations*. Prague, pp. 227--230.
- 2) Das, D., Chen, D., Martins, A.F.T, Schneider, N., Smith, N.A. **(2014)**. Frame-Semantic Parsing, *Computational Linguistics*, 40(1), pp. 9--56.

Precizitāte

C6.0 uzbūvē latviešu FrameNet likumu kopu apmēram 5 min.
(26 freimi, 5000 anotēti teikumi)

— var veikt inkrementālu mācīšanos



Freimu target identifikācijas F1 mērs
atkarībā no teikumu skaita treniņkorpusā

Freimu targetu
identifikācija F1
mērs dažiem
freimiem

Frame	F1
Being born	100
Earnings and losses	89
Death	80
Education teaching	71
Being employed	70
Change of leadership	67
Intentionally create	67
Residence	67
Statement	67
Hiring	62
Membership	50
Possession	48
People by vocation	46
Win prize	45
Participation	40
Employment end	33
Product line	33
Lending	29
Personal relationship	25
Trial	18
People by origin	16

Anglu FrameNet demo interface^[1]

FrameNet 1.3 parser demo (trained with C60 classifier)

Text Frames About

Tip: select sentence at left pane, drag to pan and scroll to zoom at right pane. To download output press [Download](#)

Sentences:

1. Islamic State crisis : ' More than 3,000 ' Europeans join IS The number of Europeans joining Islamist fighters in Syria and Iraq has risen to more than 3,000 , the EU 's anti-terrorism chief has told the BBC .
2. Gilles de Kerchove also warned that Western air strikes would increase the risk of retaliatory attacks in Europe .
3. US-led forces have launched nearly 200 air strikes against Islamic State (IS) militants in Iraq since August and on Monday began targeting IS in Syria .
4. The UK parliament is due to vote on possible air strikes in Iraq on Friday .
5. IS has seized large parts of Iraq and Syria in recent months .
6. Mr de Kerchove said the number of 3,000 included all those who have been to the region , including those who have returned and those who have been killed there .
7. Earlier this year , experts put the figure at about 2,000 .
8. Mr de Kerchove said that Islamic State 's declaration of a Caliphate in June may have played a role in drawing more support from Europe . "
9. If you believe in this , probably you want to be part of it as early as possible , " he said .
10. He warned that air strikes by the US and its Western allies had increased the risk of a violent response from militant Islamists against European targets . "
11. I think we have to acknowledge that it will , " he said . "
12. That was clear with the French because three days ago Isil [IS] issued a statement saying there would be retaliation against the coalition .

Diagram illustrating the FrameNet structure for the sentence: "IS has seized large parts of Iraq and Syria in recent months ."

The diagram shows the sentence tokens: IS, has, seized, large, parts, of, Iraq, and, Syria, in, recent, months, .

Key frames and their relationships:

- seized** (Verb frame) is the central event.
- IS** (Noun frame) is the subject of the event.
- has** (Auxiliary frame) is the auxiliary verb.
- large** (Adjective frame) is the dimension of the event.
- parts** (Noun frame) is the object of the event.
- of** (Preposition frame) is the preposition.
- Iraq and Syria** (Noun frame) is the theme of the event.
- in** (Preposition frame) is the preposition.
- recent** (Adjective frame) is the time frame.
- months** (Noun frame) is the duration of the event.

Relationships between frames:

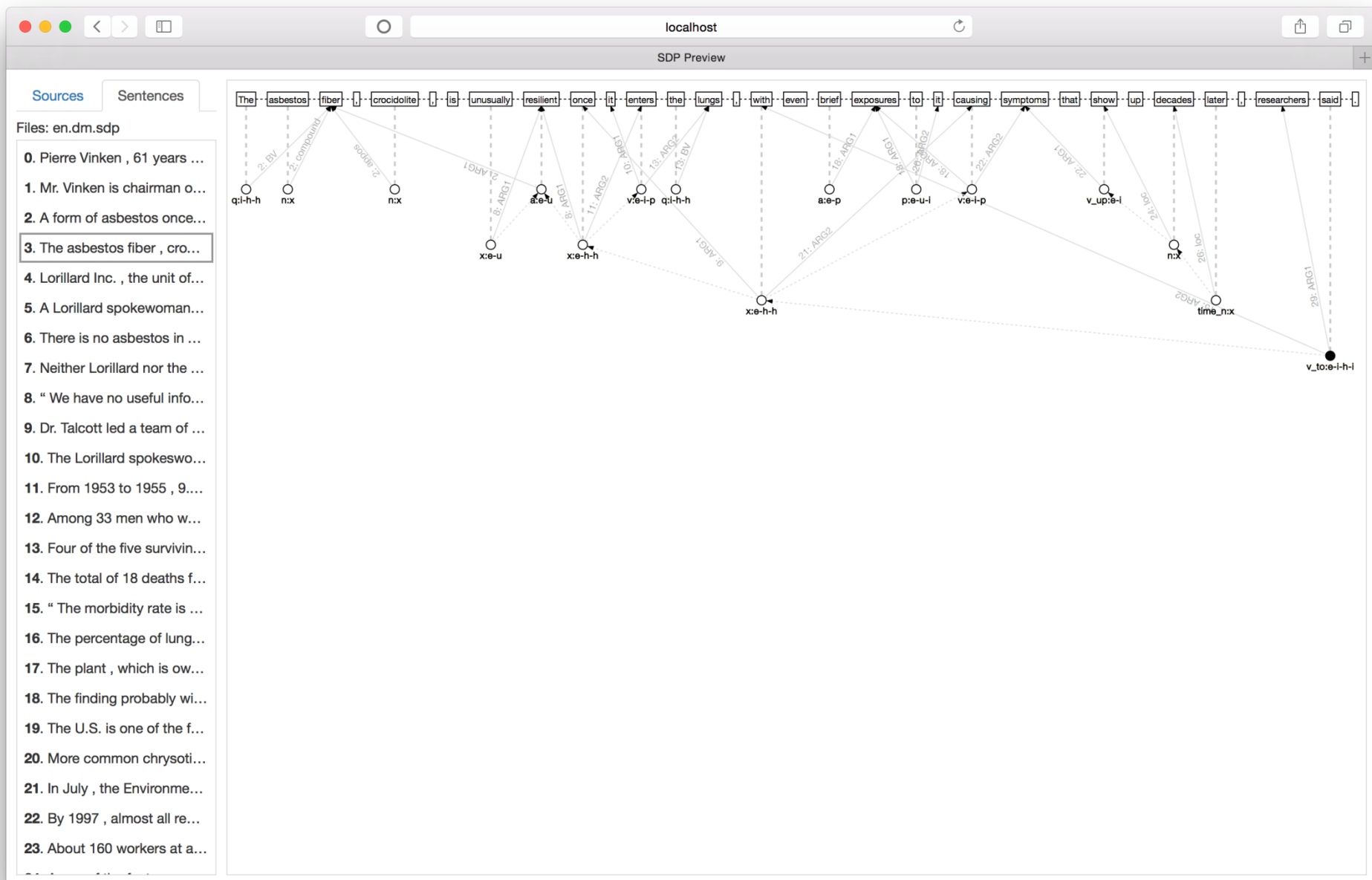
- Dimension** (Frame) is related to **large** (Frame).
- Object** (Frame) is related to **parts** (Frame).
- Theme** (Frame) is related to **Iraq and Syria** (Frame).
- Containing event** (Frame) is related to **in** (Frame).
- Taking** (Frame) is related to **seized** (Frame).

SemEval 2015

Piedalāmies SemEval 2015 semantiskās marķēšanas uzdevuma sacensībās:

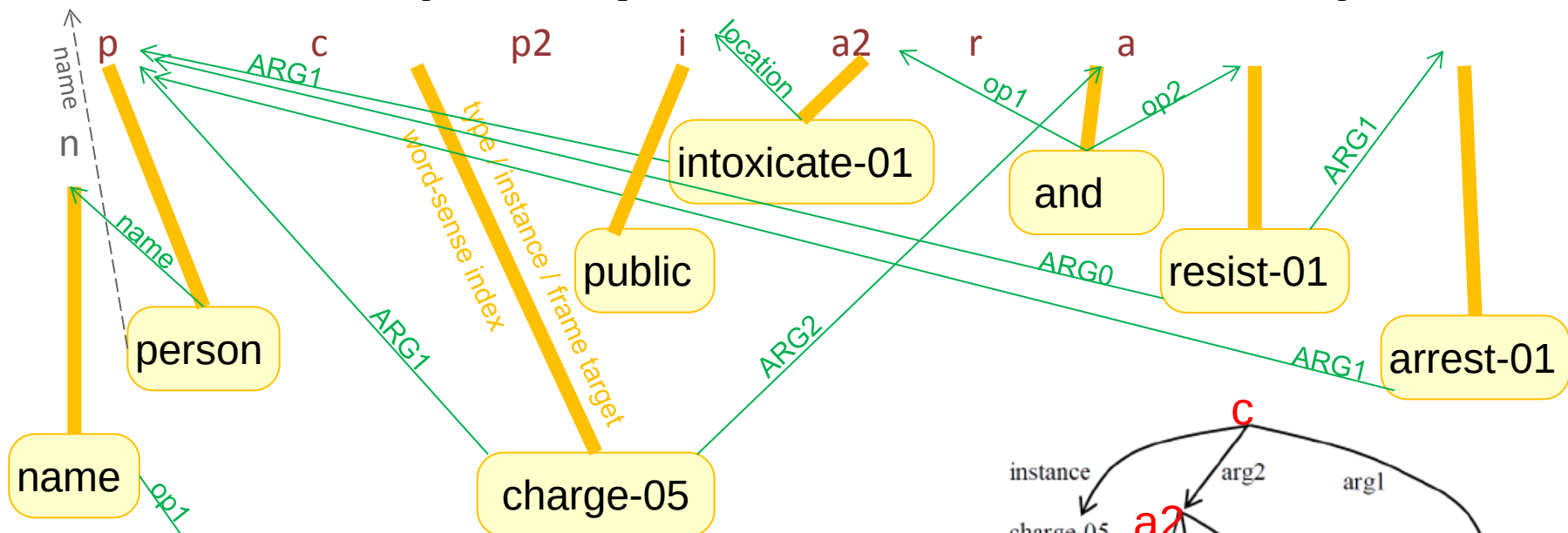
- Semantiskā marķējuma formāts atšķiras no FrameNet (nav īsti cilvēklasāms)
- Nav sintakses, bet ir semantisks grafs
- Grafu meklējam ar modificētu sintakses parseri
- Mezglu punktu mārķējumu noteikšana neatkarīgi no grafa meklēšanas

SemEval 2015 treninđati

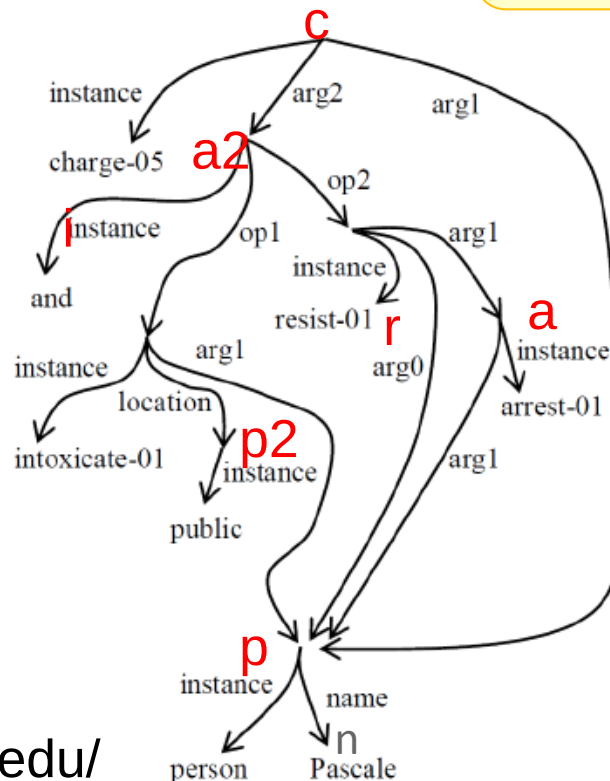


AMR (Abstract Meaning Representation)

Pascale was charged with public intoxication and resisting arrest.



```
(c / charge-05
  :ARG1 (p / person
    :name (n / name :op1 "Pascale"))
  :ARG2 (a / and
    :op1 (i / intoxicate-01
      :ARG1 p
      :location (p2 / public))
    :op2 (r / resist-01
      :ARG0 p
      :ARG1 (a / arrest-01
        :ARG1 p))))
```



Secinājumi

- Iegūtie rezultāti ir salīdzināmi ar citām state-of-art metodēm pasaulē
- Vietējai ziņu aģentūrai tika uzbūvēta automātiskas informācijas izguves sistēma no ziņu tekstiem
- Gala precizitāte ir atkarīga no visu komponentu precizitāšu «reizinājuma»

Nākotnē:

- jaunu papildinformācijas veidu/avotu pieslēgšana
 - jaunu/labāku pazīmju definēšana (no pieejamās papildinformācijas)
 - situāciju un to lomu atbilstības heuristikas uzlabošana
 - iegūto klasifikatora likumu manuāla pārskate/labošana
 - citu klasifikatoru izmēģināšana
-
- pāriešana no FrameNet uz AMR

Paldies!