UNIVERSITY OF HELSINKI

Lidia Pivovarova

Expansion of an Information Extraction System to the Russian Language

University of Oslo, 6.6.2012

PULS Project

- http://puls.cs.helsinki.fi/puls
- University of Helsinki, Department of Computer Science
- Team:

Project lead: Roman Yangarber

Mian Du

Peter von Etter

Silja Huttunen

Lidia Pivovarova, St. Petersburg State University, Russia

Visitors and Past Members:

Mikhail Novikov (2011)

Esben Alfort, Copenhagen Business School, Denmark (Summer 2011)

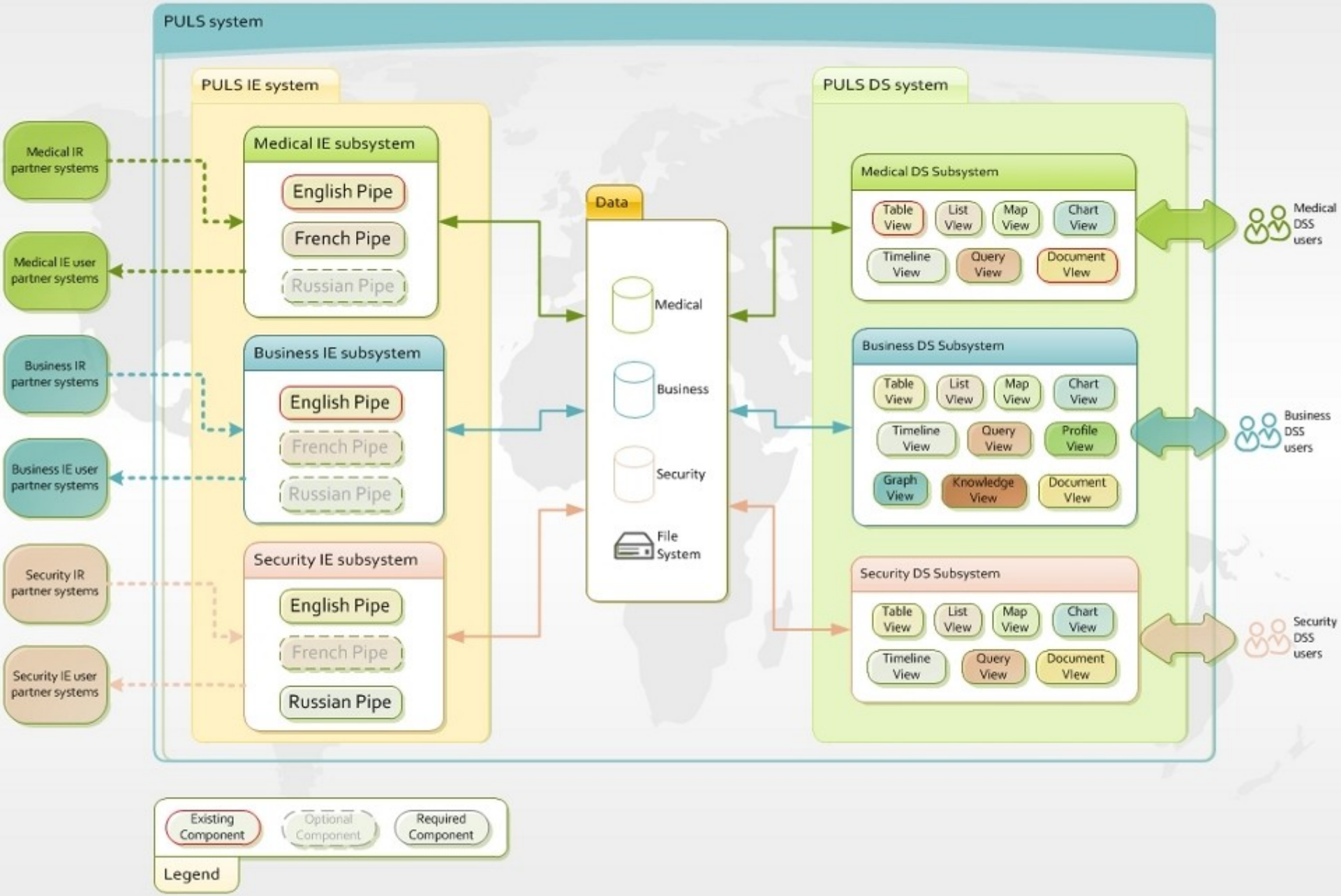
Lauri Jokipii (2006)

Gaël Lejeune, University of Caen, France (2009)

Heikki Manninen (2009)

Natalia Tarbeeva, State University of Perm', Russia (2011)

Arto Vihavainen (2010-2011)


PULS system
PULS IE system
PULS DS system
Medical IR partner systems
Medical IE user partner systems
Business IR partner systems
Business IE user partner systems
Security IR partner systems
Security IE user partner systems
Medical IE subsystem
English Pipe
French Pipe
Russian Pipe
Business IE subsystem
English Pipe
French Pipe
Russian Pipe
Security IE subsystem
English Pipe
French Pipe
Russian Pipe
Data
Medical
Business
Security
File System
Medical DS Subsystem
Table View
List View
Map View
Chart View
Timeline View
Query View
Document View
Business DS Subsystem
Table View
List View
Map View
Chart View
Timeline View
Query View
Profile View
Graph View
Knowledge View
Document View
Security DS Subsystem
Table View
List View
Map View
Chart View
Timeline View
Query View
Document View
Medical DSS users
Business DSS users
Security DSS users
Existing Component
Optional Component
Required Component
Legend

SA treats 68 cholera patients on Zimbabwe border

South Africa has treated 68 cholera patients since the weekend in a town by the border with Zimbabwe, where the disease has killed dozens of people in recent weeks, a health official said today. "Since Saturday, we have received and treated a total of 68 cholera patients from Zimbabwe," said Phuti Selobi, spokesman for the health department in the town of Musina said. "Sixty-six of them are Zimbabweans while two others are South Africans engaged in cross-border business," Selobi told AFP. "Only 14 of them are still in the hospital," he added, noting that no one has died of cholera in South Africa.

Data Collection

News items arrive continuously in real-time, as **plain text**. **PULS** currently receives raw text from the partner system, EC-JRC's MedISys. **PULS** processes news from specialized sites, such as ProMED-Mail, and can process data from other news aggregators.

Linguistic Analysis

Plain text is analysed for indicators of epidemic-related cases. The analysis is fully automatic, employing state-of-the-art language technology and machine learning.

Understanding

Each detected case is transformed into **structured information**:

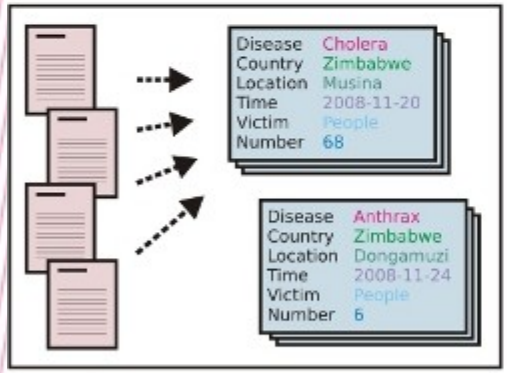
- **what** disease
- **where** — country, location,
- **when** — date
- **who** — victims, number, ...

and added to a database of events

Disease	Cholera
Country	Zimbabwe
Location	Musina
Time	2008-11-20
Victim	People
Number	68

Learning

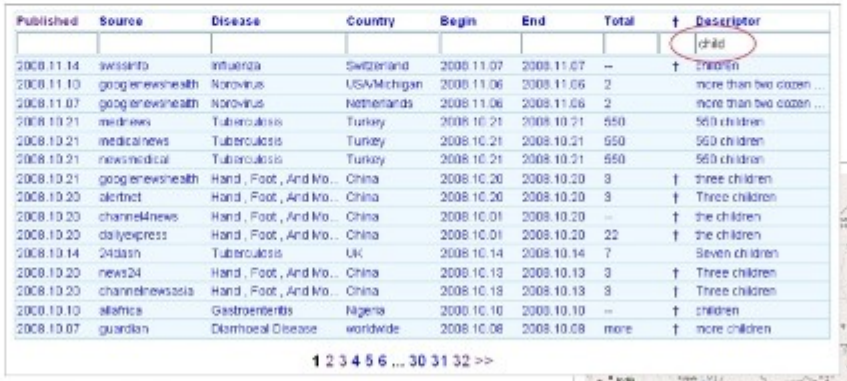
PULS gives authorized users the option to provide feedback to the system, e.g., by correcting errors in the automatic analysis. The system adapts to improve the analysis on subsequent reports.



Aggregation

The same report is frequently encountered in multiple news sources, possibly with slight variations. For epidemic surveillance, the user does not want to read the same story over and over again, multiple times. **PULS** tries to reduce redundancy by:

- Gathering multiple reports on the same story, across time and multiple sources, and presenting them to the user as a **unified record**. Links to original documents are preserved
- Assigning a **relevance** score to the unified record, based on criteria that determine the urgency and importance of the event
- Linking related stories across time into one evolving event/thread

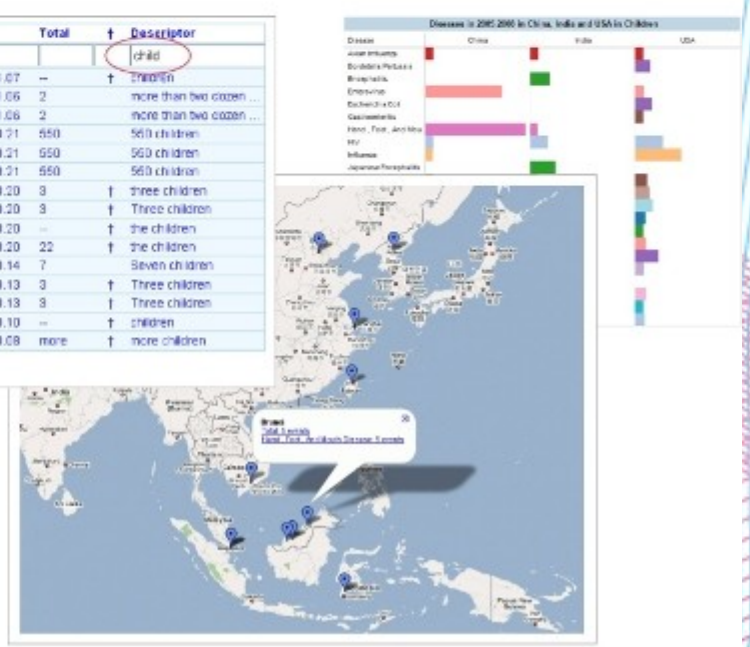


Published	Source	Disease	Country	Begin	End	Total	†	Descriptor
								child
2008.11.14	swissinfo	Influenza	Switzerland	2008.11.07	2008.11.07	--	†	Children
2008.11.10	googlenewshealth	Norovirus	USA/Michigan	2008.11.06	2008.11.06	2		more than two dozen ...
2008.11.07	googlenewshealth	Norovirus	Netherlands	2008.11.06	2008.11.06	2		more than two dozen ...
2008.10.21	medinews	Tuberculosis	Turkey	2008.10.21	2008.10.21	550		560 children
2008.10.21	medicalnews	Tuberculosis	Turkey	2008.10.21	2008.10.21	550		560 children
2008.10.21	newsmedical	Tuberculosis	Turkey	2008.10.21	2008.10.21	550		560 children
2008.10.21	googlenewshealth	Hand , Foot , And Mo...	China	2008.10.20	2008.10.20	3	†	three children
2008.10.20	alertnet	Hand , Foot , And Mo...	China	2008.10.20	2008.10.20	3	†	Three children
2008.10.20	channel4news	Hand , Foot , And Mo...	China	2008.10.01	2008.10.20	--	†	the children
2008.10.20	dailyexpress	Hand , Foot , And Mo...	China	2008.10.01	2008.10.20	22	†	the children
2008.10.14	24dash	Tuberculosis	UK	2008.10.14	2008.10.14	7		Seven children
2008.10.20	news24	Hand , Foot , And Mo...	China	2008.10.13	2008.10.13	3	†	Three children
2008.10.20	channelnewsasia	Hand , Foot , And Mo...	China	2008.10.13	2008.10.13	3	†	Three children
2008.10.10	allafrica	Gastroenteritis	Nigeria	2008.10.10	2008.10.10	--	†	children
2008.10.07	guardian	Diarrhoeal Disease	worldwide	2008.10.06	2008.10.06	more	†	more children

1 2 3 4 5 6 ... 30 31 32 >>

Decision support

PULS stores all events in structured form in a permanent database. To support the work of the Public Health Authorities, the database provides the ability to search and query for specific diseases, countries, victims (adults/children/human/animal, ...), time period, and other attributes. The database feeds into down-stream analysis and visualization tools: maps, charts, etc.



Papers

- Ralph Grishman, Silja Huttunen, Roman Yangarber. Real-Time Event Extraction for Infectious Disease Outbreaks In Proceedings of the 3rd Annual Human Language Technology Conference HLT-2002 (2002) San Diego, CA
- M Atkinson, J Piskorski, H Tanev, E van der Goot, R Yangarber, V Zavarella. Automated event extraction in the domain of Border Security In Proceedings of MINUCS-2009: Workshop on Mining User-Generated Content for Security, at the UCMedia-2009: ICST Conference on User-Centric Media (2009) Venice, Italy
- Silja Huttunen, Arto Vihavainen, Peter von Etter, Roman Yangarber. Relevance prediction in information extraction using discourse and lexical features Nodalida-2011: Nordic Conference on Computational Linguistics (2011) Riga, Latvia
- Mian Du, Peter von Etter, Mikhail Kopotev, Mikhail Novikov, Natalia Tarbeeva, Roman Yangarber. Building support tools for Russian-language information extraction BSNLP-2011: Balto-Slavonic Natural Language Processing (2011) Plzeň, Czech Republic

Russian: task definition

- Russian news analysis for Border Security and Medical scenario
- Results representation in a unified form (common for Russian and English texts)
- Usage of existing (made for English) tools – as much as it possible

SA treats 68 cholera patients on Zimbabwe border

South Africa has treated 68 cholera patients since the weekend in a town by the border with Zimbabwe, where the disease has killed dozens of people in recent weeks, a health official said today. "Since Saturday, we have received and treated a total of 68 cholera patients from Zimbabwe," said Phuti Selobi, spokesman for the health department in the town of Musina said. "Sixty-six of them are Zimbabweans while two others are South Africans engaged in cross-border business." Selobi told AFP. "Only 14 of them are still in the hospital," he added, noting that no one has died of cholera in South Africa.

Data Collection

News items arrive continuously in real-time, as **plain text**. **PULS** currently receives raw text from the partner system, EC-JRC's MedISys. **PULS** processes news from specialized sites, such as ProMED-Mail, and can process data from other news aggregators.

SA treats 68 cholera patients on Zimbabwe border

South Africa has treated **68 cholera patients** since the weekend in a town by the border with Zimbabwe, where the **disease has killed dozens** of people in recent weeks, a health official said today. "Since Saturday, we have received and treated a total of **68 cholera patients from Zimbabwe**," said Phuti Selobi, spokesman for the health department in the town of Musina said. "Sixty-six of them are Zimbabweans while two others are South Africans engaged in cross-border business," Selobi told AFP. "Only 14 of them are still in the hospital," he added, noting that no one has died of cholera in South Africa.

Linguistic Analysis

Plain text is analysed for indicators of epidemic-related cases. The analysis is fully automatic, employing state-of-the-art language technology and machine learning.

SA treats 68 cholera patients on Zimbabwe border

South Africa has treated **68 cholera patients** since the weekend in a town by the border with **Zimbabwe**, where the **disease has killed dozens** of people **in recent weeks**, a health official said today. "**Since Saturday**, we have received and treated a total of **68 cholera patients from Zimbabwe**," said Phuti Selobi, spokesman for the health department in the town of Musina said. "Sixty-six of them are Zimbabweans while two others are South Africans engaged in cross-border business," Selobi told AFP. "Only 14 of them are still in the hospital," he added, noting that no one has died of cholera in South Africa.

Understanding

Each detected case is transformed into **structured information**:

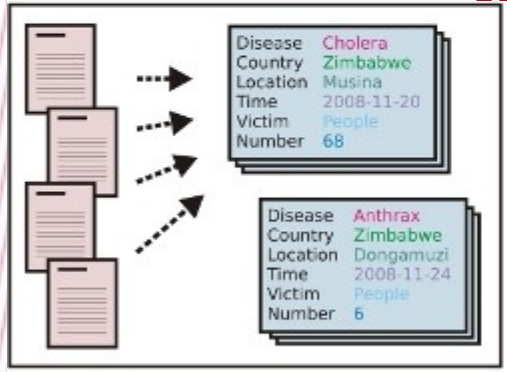
- **what** disease
- **where** — country, location,
- **when** — date
- **who** — victims, number, ...

and added to a database of events

Disease	Cholera
Country	Zimbabwe
Location	Musina
Time	2008-11-20
Victim	People
Number	68

Learning

PULS gives authorized users the option to provide feedback to the system, e.g., by correcting errors in the automatic analysis. The system adapts to improve the analysis on subsequent reports.



Aggregation

The same report is frequently encountered in multiple news sources, possibly with slight variations. For epidemic surveillance, the user does not want to read the same story over and over again, multiple times. **PULS** tries to reduce redundancy by:

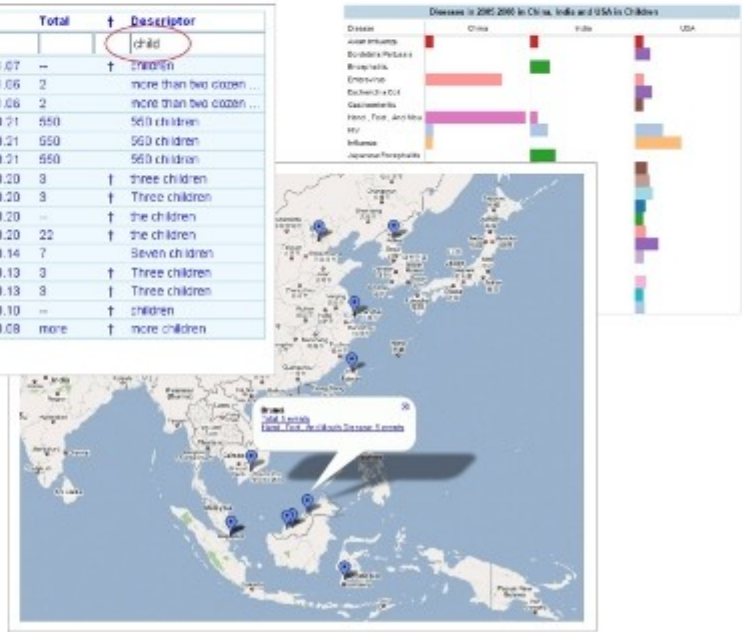
- Gathering multiple reports on the same story, across time and multiple sources, and presenting them to the user as a **unified record**. Links to original documents are preserved
- Assigning a **relevance** score to the unified record, based on criteria that determine the urgency and importance of the event
- Linking related stories across time into one evolving event/thread

Published	Source	Disease	Country	Begin	End	Total	†	Descriptor
								child
2008.11.14	swissinfo	Influenza	Switzerland	2008.11.07	2008.11.07	--	†	Children
2008.11.10	googlenewshealth	Norovirus	USA/Michigan	2008.11.06	2008.11.06	2		more than two dozen ...
2008.11.07	googlenewshealth	Norovirus	Netherlands	2008.11.06	2008.11.06	2		more than two dozen ...
2008.10.21	medinews	Tuberculosis	Turkey	2008.10.21	2008.10.21	550		560 children
2008.10.21	medicalnews	Tuberculosis	Turkey	2008.10.21	2008.10.21	550		560 children
2008.10.21	newsmedical	Tuberculosis	Turkey	2008.10.21	2008.10.21	550		560 children
2008.10.21	googlenewshealth	Hand , Foot , And Mo...	China	2008.10.20	2008.10.20	3	†	three children
2008.10.20	alertnet	Hand , Foot , And Mo...	China	2008.10.20	2008.10.20	3	†	Three children
2008.10.20	channel4news	Hand , Foot , And Mo...	China	2008.10.01	2008.10.20	--	†	the children
2008.10.20	dailyexpress	Hand , Foot , And Mo...	China	2008.10.01	2008.10.20	22	†	the children
2008.10.14	24dash	Tuberculosis	UK	2008.10.14	2008.10.14	7		Seven children
2008.10.20	news24	Hand , Foot , And Mo...	China	2008.10.13	2008.10.13	3	†	Three children
2008.10.20	channelnewsasia	Hand , Foot , And Mo...	China	2008.10.13	2008.10.13	3	†	Three children
2008.10.10	allafrica	Gastroenteritis	Nigeria	2008.10.10	2008.10.10	--	†	children
2008.10.07	guardian	Diarrhoeal Disease	worldwide	2008.10.06	2008.10.06	more	†	more children

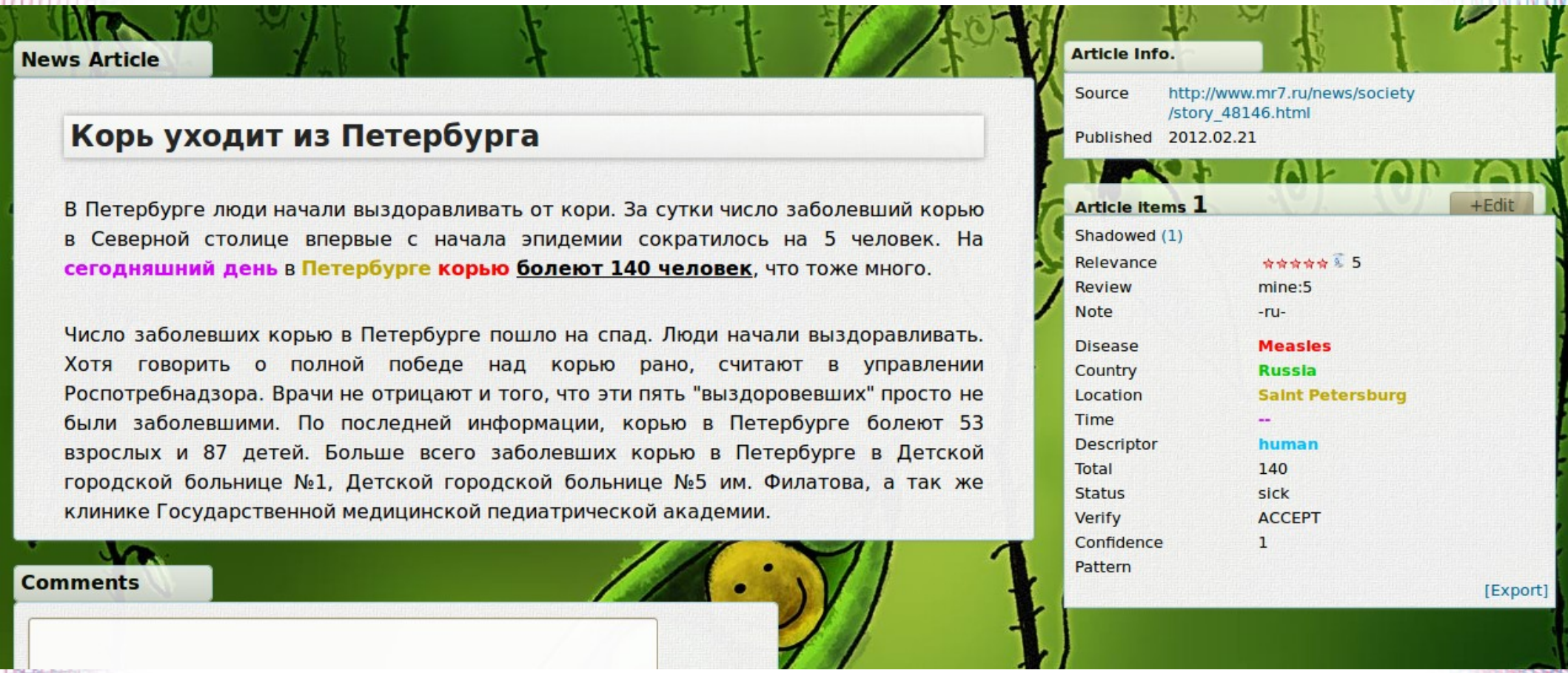
1 2 3 4 5 6 ... 30 31 32 >>

Decision support

PULS stores all events in structured form in a permanent database. To support the work of the Public Health Authorities, the database provides the ability to search and query for specific diseases, countries, victims (adults/children/human/animal, ...), time period, and other attributes.
The database feeds into down-stream analysis and visualization tools: maps, charts, etc.



Scenario: Medical



PULS **Table** | List | Map | Chart | Timeline | Surveillance | Complete

select a saved query ▼ all | reviewed | my group: puls

	Published	Source	Disease	Country	Date	Total	†	Descriptor	Note	Rel
			measles			.				
[18] +	2012.02.21	AHN	Measles	Republic of Congo	2010.12-2011.06	32	†	32 fatalities		2
[18] +	2012.02.21	AHN	Measles	Republic of Congo	2010.12-2011.06	800		800 cases		2
[18] +	2012.02.21	AHN	Measles	Republic of Congo	--	2	†	two fatalities		4
[18] +	2012.02.21	AHN	Measles	Republic of Congo	--	200		200 cases		4
[64] +	2012.02.21	mr_spb	Measles	Russia	--	140		human	-ru-	5
[18] +	2012.02.21	allafrica	Measles				†	32 fatalities		
[18] +	2012.02.21	allafrica	Measles	Republic of Congo	2010.12-2011.06	800		800 cases		
[18] +	2012.02.21	allafrica	Measles	Republic of Congo	--	2	†	two fatalities		
[18] +	2012.02.21	allafrica	Measles	Republic of Congo	--	200		200 cases		
[18] +	2012.02.21	irinnews	Measles	Republic of Congo	2010.12-2011.06	32	†	32 fatalities		
[18] +	2012.02.21	irinnews	Measles	Republic of Congo	2010.12-2011.06	800		800 cases		
[18] +	2012.02.21	irinnews	Measles	Republic of Congo	--	2	†	two fatalities		
[18] +	2012.02.21	irinnews	Measles	Republic of Congo	--	200		200 cases		
[64] +	2012.02.21	vesti	Measles	Russia	--	--		outbreak	-ru-	2
[64] +	2012.02.21	rian	Measles	Russia	--	--		human	-ru-	4
[64] +	2012.02.20	lenta	Measles	Russia	--	--		outbreak	-ru-	0
[64] +	2012.02.20	tatar_inform	Measles	Russia	--	--		outbreak	-ru-	4
[1206] +	2012.02.20	hpa	Measles	UK	--	--		those		
[1206] +	2012.02.20	hpa	Measles	UK	2012	22		22 cases		
[1206] +	2012.02.20	hpa	Measles	UK	2012	--		cases		

На сегодняшний день в Петербурге корью болеют 140 человек, что тоже много. Число заболевших корью в Петербурге пошло на спад.

1 2 3 4 5 6 ... 99 100 101 >>

Viewing 2000 items in 12352719 documents

Legend: reviewed, high-relevance | reviewed, lower-relevance | non-reviewed, high-relevance

Disease measles Country Search Reset all reviewed my group: puls

select a saved query show snippets

1 2 3 4 5 6 ... 148 149 150 >>

3 **Measles - Republic of Congo** 2012-02-21

Cholera 'continues spreading' in Congo 2012-02-21 18:24 www.allheadlinenews.com

Meanwhile, in Brazzaville, 200 cases of measles, including two fatalities, have been recorded over the past two weeks, according to the director-general in the health ministry, Alexis Elira Dokekias, who explained that not all children had been vaccinated against the disease.

Between December 2010 and June 2011, 800 cases of measles, including 32 fatalities, were recorded in the southern Pointe-Noire region, leading to a stepped-up immunization campaign.

6 **Measles - Russia** 2012-02-20 ~ 2012-02-21

Корь уходит из Петербурга 2012-02-21 16:03 www.mr7.ru

На сегодняшний день в Петербурге корью болеют 140 человек, что тоже много. Число заболевших корью в Петербурге пошло на спад. **mine: 5** **Note: -ru-**

18 **Measles - UK** 2012-02-20

Don't take a chance with measles. Arrange MMR vaccination 2012-02-20 16:22 www.hpa.org.uk

We now have 75 reported cases of measles in the area, of which 20 have been confirmed by laboratory testing.

Measles - Ukraine 2012-02-20

25 курсантов черниговского колледжа госпитализированы с подозрением на корь 2012-02-20 15:07 podrobnosti.ua

Госпитализировали еще 15 курсантов. **Note: -ru-**

Представитель добавил, что у госпитализированных ранее 3 курсантов подозрения кори сняты, у 3 подтверждены лабораторно, у 5 - подтверждены клинически. **Note: -ru-**

2 **Measles - Europe** 2012-02-20

Parents urged to vaccinate their children ahead of school holidays 2012-02-20 14:51 www.hpa.org.uk

The European Centre for Disease Control (ECDC) also issued a warning in April 2008 stating that Europe may be about to experience a significant outbreak after an increase in measles in several European countries.

Measles - Ireland 2012-02-20

Measles cases rise in Ireland and over 30,000 Measles cases in Europe in 2011 2012-02-20 12:36 en

In Ireland 278 cases of measles have been reported since January 2011- many of these occurred during an outbreak in Dublin, primarily in North Dublin City.

Scenario: Border Security

- Monitoring of:
  - Illegal migration
  - Criminal activity related to border crossing (e.g. smuggling)
  - Criminal activity in general
- Motivation
  - News may be the only information source for an event
  - Or the fastest source
  - Or provide with an alternative point of view / extra details

Scenario: Border Security

News Article

Из Башкирии выдворены 12 нелегальных мигрантов

Из Башкирии отправлены на Родину 12 нелегальных мигрантов. Об этом 13 марта корреспонденту ИА REGNUM сообщили в пресс-службе УФССП по Башкирии. Так, недавно сотрудники Управления ФССП России по РБ препроводили до пункта пропуска через Государственную границу России международного аэропорта Уфы очередного нелегального мигранта - гражданина Республики Узбекистан. С начала 2012 года по решению судов в Центр содержания иностранных граждан при Управлении МВД России по Уфе были помещены 54 человека. Большинство нарушителей закона - граждане Узбекистана. Также в числе выдворенных - граждане Таджикистана, Азербайджана и Армении. Все эти люди находились на территории РФ, не имея разрешительных документов, либо срок их пребывания истек. Всех нелегальных мигрантов отправили на Родину, и в ближайшие пять лет доступ на территорию России для нарушителей миграционного режима будет закрыт. Отметим, что с 1 января 2012 года на Федеральную службу судебных приставов Федеральным законом от 06.12.2011 № 410-ФЗ возложены функции по исполнению постановлений судов об административном выдворении иностранных граждан или лиц без гражданства за пределы Российской Федерации в форме принудительного перемещения через Государственную границу РФ

Article Info.

Source	www.regnum.ru
Published	2012.03.13

Article events 1 +Edit

Shadowed (1)	
Type	migration-illegal-stay
Relevance	☆☆☆☆☆ 4
Reviewed	puls:4
Note	
Country	Russia
Location	Russia
Country2	Uzbekistan
Location2	Uzbekistan
Time	
Suspect	Illegal-Migrant
Suspect-Count	12
Acting-authority	
Means	
Verify	ACCEPT
Pattern	(Expel: Выдворить Crisis-Anchor Ir-Sec-Ev->Add-Test

Scenario: Border Security

News Article

В Донецкой области девушку пытались продать в сексуальное рабство

В начале февраля на железнодорожной станции г. Краматорска милиция задержала парня и женщину, которые намеревались вывезти за границу 31-летнюю горловчанку с целью продажи сутенерам для последующей сексуальной эксплуатации. Как сообщили 'у в пресс-службе Горловского ГУ ГУМВД Украины в Донецкой области, установлено, что в течение последних двух лет парочка, 23-летний парень и 34-летняя женщина, занималась вербовкой и отправкой молодых девушек из малообеспеченных семей для дальнейшей их сексуальной эксплуатации за рубежом. В данном случае их ожидали в турецком городе Анталия, где вербовщица свое время тоже зарабатывала на жизнь жрицей любви. За свою «работу» вербовщики получали в среднем 1000 долларов за каждую девушку. По словам начальника СБПТЛ Горловского ГУ майора милиции Сергея Евсюкова, в текущем году это первый факт документирования подобного вида преступления. На сегодняшний день установлено пять пострадавших, трое из которых – жительницы Горловки из малообеспеченных семей. По данному факту возбуждено уголовное дело по ч. 2 ст. 149 УК Украины. Торговля людьми или другое незаконное соглашение относительно передачи человека, совершенное по предварительному сговору группой лиц, наказывается лишением свободы на срок от 5 до 12 лет с конфискацией имущества или без таковой. В отношении задержанных избрана мера пресечения – арест.

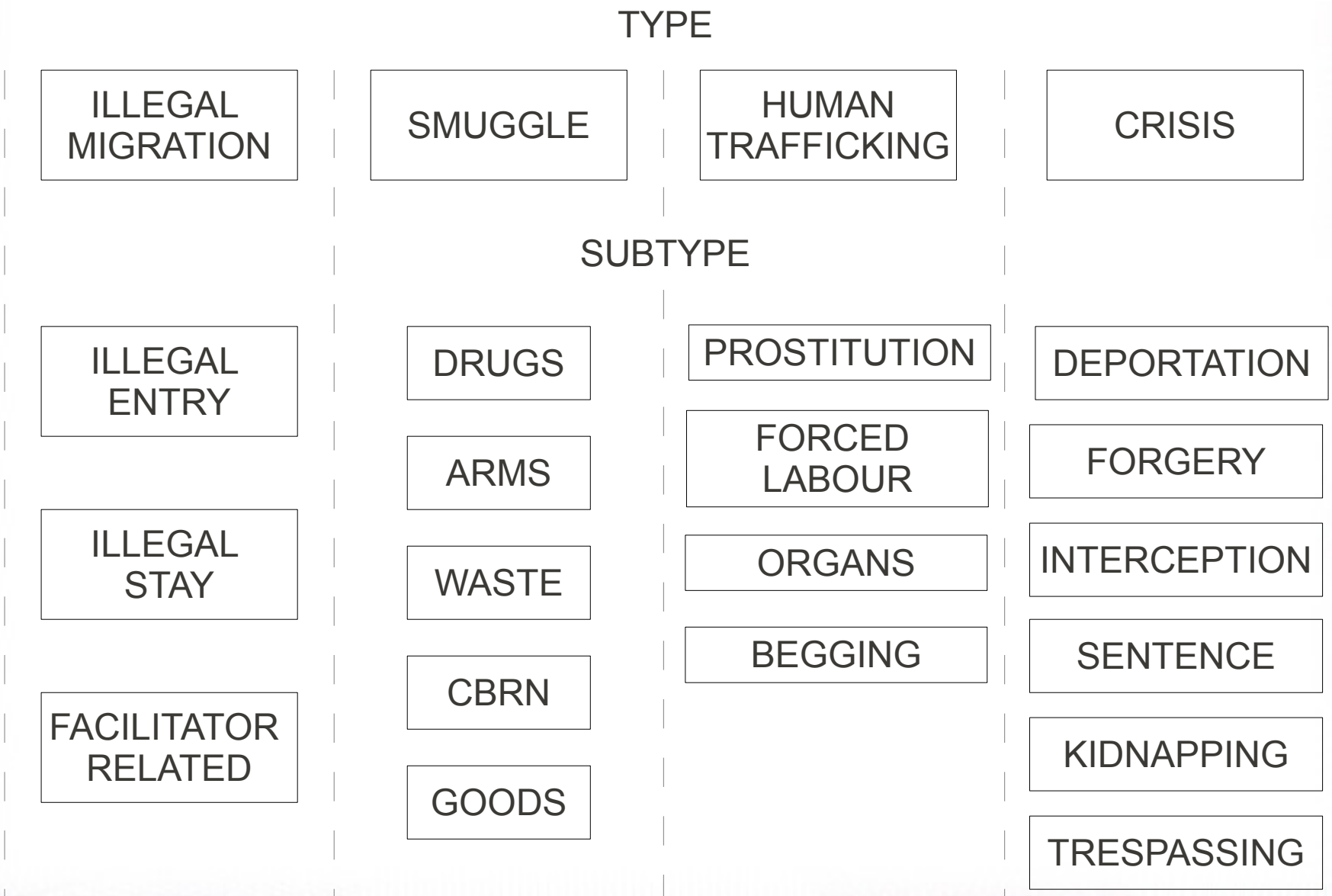
Article Info.

Source	for-ua.com
Published	2012.02.20

Article events 1 2 +Edit

Shadowed	
Type	humtraf-prostitution
Relevance	☆☆☆☆☆ 4
Reviewed	
Note	
Country	Ukraine
Location	Kramatorsk
Country2	Ukraine
Location2	Donetsk Oblast
Time	
Suspect	Chap
Suspect-Count	
Victim	Sexual-Exploitation
Victim-Count	
Acting-authority	
Means	
Verify	
Pattern	(Detain: Задержать Crisis-Anchor Ir-Crisis->Known-T

Scenario: Border Security

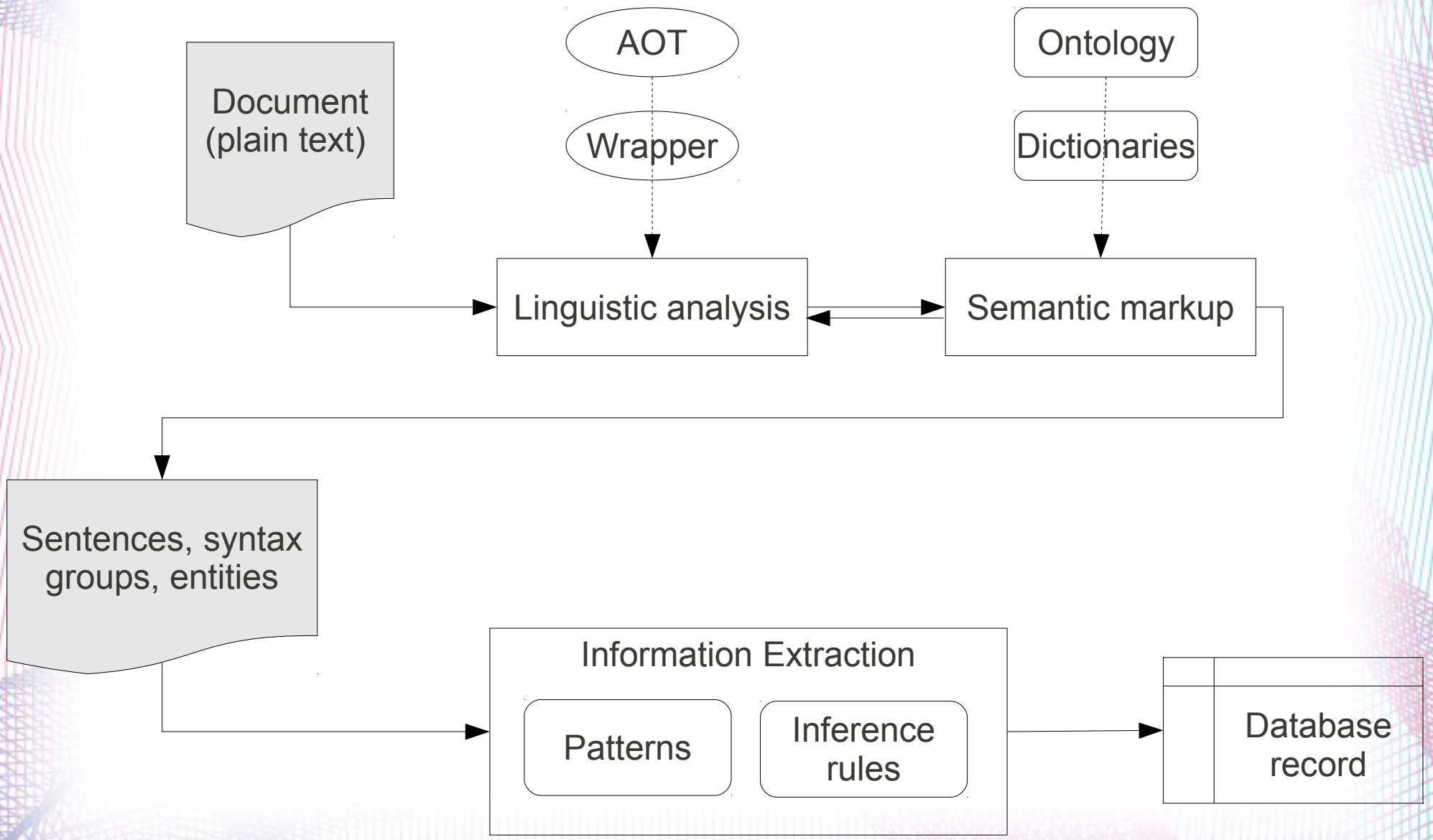


Materials

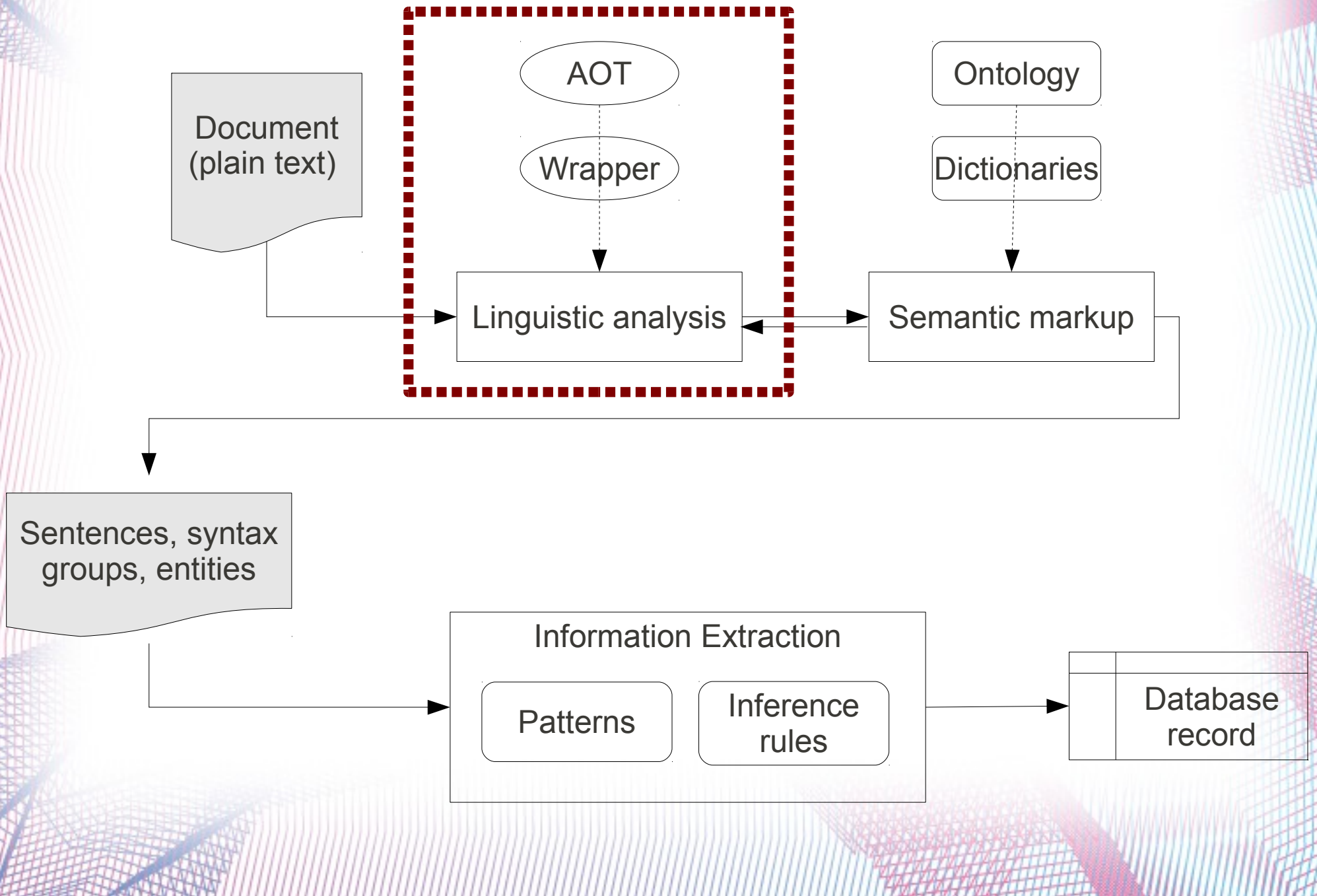
	abs	%		abs	%
http://rus.ruvr.ru/	1709	10.44	http://news.open.by/	251	1.53
http://www.vz.ru/	820	5.01	http://www.dzd.ee/	251	1.53
http://www.ria.ru/	658	4.02	http://www.centrasia.ru/	247	1.51
http://www.nakanune.ru/	629	3.84	http://www.kommersant.ru/	241	1.47
http://www.regnum.ru/	492	3.01	http://www.svobodanews.ru/	238	1.45
http://www.rg.ru/	423	2.58	http://ru.euronews.net/	234	1.43
http://www.regions.ru/	386	2.36	http://www.vesti.ru/	222	1.36
http://korrespondent.net/	369	2.25	http://www.fontanka.ru/	211	1.29
http://www.newsru.com/	328	2.00	http://kp.ru/	202	1.23
http://www.nr2.ru/	295	1.80	http://www.dw-world.de/	199	1.22
http://lenta.ru/	292	1.78	http://www.belta.by/	184	1.12
http://naviny.by/	282	1.72	http://podrobnosti.ua/	174	1.06
http://www.rosbalt.ru/	272	1.66	http://for-ua.com/	169	1.03
http://www.fms.gov.ru/	271	1.66	http://news.online.ua/	167	1.02
http://www.bbc.co.uk/	263	1.61	http://top.rbc.ru/	164	1.00

1.06.2012. Total amount of Russian documents: ~16300

General Scheme



General Scheme



AOT

- http://seman.sourceforge.net/
- open source toolkit for Russian linguistic analysis
- libraries for morphological, syntactic, and semantic analysis, language generation, tools for working with dictionaries, and GUIs for visualizing the analysis
- we use only the morphological and syntactic analyzers, called Lemm and Synan

AOT: LEMM

На берегу пограничной реки задержаны двадцать семь нелегальных мигрантов.

```
                0 0 BEG DOC
На              0 2 RLE Aa NAM? EXPR1 EXPR2 EXPR_NO277 +?? НА яв 44907 0
берегу          3 6 RLE aa +Ун БЕРЕЧЬ кб 133825 0
 берегу         3 6 RLE aa +Фа БЕРЕГ авЭх 153063 0
пограничной     10 11 RLE aa +?? ПОГРАНИЧНЫЙ йзййкйл 167378 0
реки            22 4 RLE aa +Фа РЕКА гбгжгй 150782 0
задержаны       27 9 RLE aa +Ул ЗАДЕРЖАТЬ сэ 144652 0
двадцать        37 8 RLE aa +?? ДВАДЦАТЬ эаэг 145038 0
семь            46 4 RLE aa +?? СЕМЬ эаэг 145046 0
нелегальных     51 11 RLE aa +Уе НЕЛЕГАЛЬНЫЙ йуйхйч 170468 0
мигрантов       63 9 RLE aa CS? SENT_END +Фб МИГРАНТ азай 87080 0
```

AOT: LEMM

На берегу пограничной реки задержаны двадцать семь нелегальных мигрантов.

Twenty seven illegal migrants have been detained on the bank of the borderline river

Byte	Surface	Lemma	POS	Morphological tags
0	На	на	Prep	—
3	берегу	беречь	Finverb	Impf Transv Act Pres 1p Sg
3	берегу	берег	Noun	Inan Masc Sg {Dat|Loc}
10	пограничной	пограничный	Adj	Fem Sg Anim Inan {Gen|Acc|Inst|Loc}
22	реки	река	Noun	Inan Fem {Sg Gen|Pl Nom|Pl Acc}
27	задержано	задержать	SParticip	Perf Transv Anim Inan Past Pass Sg Neut
36	двадцать	двадцать	Card	{Nom|Acc}
45	семь	семь	Card	{Nom|Acc}
50	нелегалов	нелегал	Noun	Anim Masc Pl {Gen|Acc}

AOT: SYNAN

<output>
<chunk>
<input>На берегу пограничной реки задержаны двадцать семь нелегальных мигрантов</input>
<sent>
<synvar>
<clause type="КР_ПРЧ">На[0] берегу[1] пограничной[2] реки[3] задержаны[4] двадцать[5] семь[6] нелегальных[7] мигрантов[8]</clause>
<group type="ПРИЛ_СУЩ">пограничной[2] реки[3]</group>
<group type="ГЕНИТ_ИГ">берегу[1] пограничной[2] реки[3]</group>
<group type="ПГ">На[0] берегу[1] пограничной[2] реки[3]</group>
<group type="КОЛИЧ">двадцать[5] семь[6]</group>
<group type="ПРИЛ_СУЩ">нелегальных[7] мигрантов[8]</group>
<group type="ЧИСЛ_СУЩ">двадцать[5] семь[6] нелегальных[7] мигрантов[8]</group>
</synvar>
<rel name="ПРИЛ_СУЩ" gramrel="жр,рд,ед," lemmprnt="РЕКА" grmprnt="но,жр,рд,ед," lemmchld="ПОГРАНИЧНЫЙ" grmchld="но,од,жр,рд,ед," noprnt="3" nochld="2" > реки -> пограничной </rel>
<rel name="ПРИЛ_СУЩ" gramrel="им,мн," lemmprnt="МИГРАНТ" grmprnt="од,мр,рд,мн," lemmchld="НЕЛЕГАЛЬНЫЙ" grmchld="кач,но,од,рд,мн," noprnt="8" nochld="7" > мигрантов -> нелегальных </rel>
<rel name="ЧИСЛ_СУЩ" gramrel="им,мн," lemmprnt="МИГРАНТ" grmprnt="од,мр,рд,мн," lemmchld="" grmchld="" noprnt="8" nochld="6" > мигрантов -> двадцать[5] семь[6] </rel>
<rel name="ГЕНИТ_ИГ" gramrel="2,но,мр,пр,ед," lemmprnt="БЕРЕГ" grmprnt="2,но,мр,пр,ед," lemmchld="РЕКА" grmchld="но,жр,рд,ед," noprnt="1" nochld="3" > берегу -> реки </rel>
<rel name="ПГ" gramrel="пр," lemmprnt="НА" grmprnt="" lemmchld="БЕРЕГ" grmchld="2,но,мр,пр,ед," noprnt="0" nochld="1" > На -> берегу </rel>
<rel name="ПОДЛ" gramrel="" lemmprnt="ЗАДЕРЖАТЬ" grmprnt="стр,пе,св,но,од,прш,мн," lemmchld="МИГРАНТ" grmchld="од,мр,рд,мн," noprnt="4" nochld="8" > задержаны -> мигрантов </rel>
</sent>
</chunk>
</output>

AOT: SYNAN

На берегу пограничной реки задержаны двадцать семь нелегальных мигрантов.

Twenty seven illegal migrants have been detained on the bank of the borderline river

Relations						
Type	Parent			Child		
	ID	Surface	Lemma	ID	Surface	Lemma
Num-Noun	7	нелегалов	НЕЛЕГАЛ	5	двадцать семь	—
Adj-Noun	3	реки	РЕКА	2	пограничной	ПОГРАНИЧНЫЙ
Gen-Nom-Group	1	берегу	БЕРЕГ	3	реки	РЕКА
Prep-Group	0	На	НА	1	берегу	БЕРЕГ

Groups	
Type	Members
Cardinal-Ordinal-Group	двадцать(5) семь(6)

WRAPPER

- Lemm: does not disambiguate
- Synan: does not contain all the words, only words which are used in grammar relations
- Wrapper: combines results of Lemm and Synan

  + elements of semantic analysis (e.g. proper names)

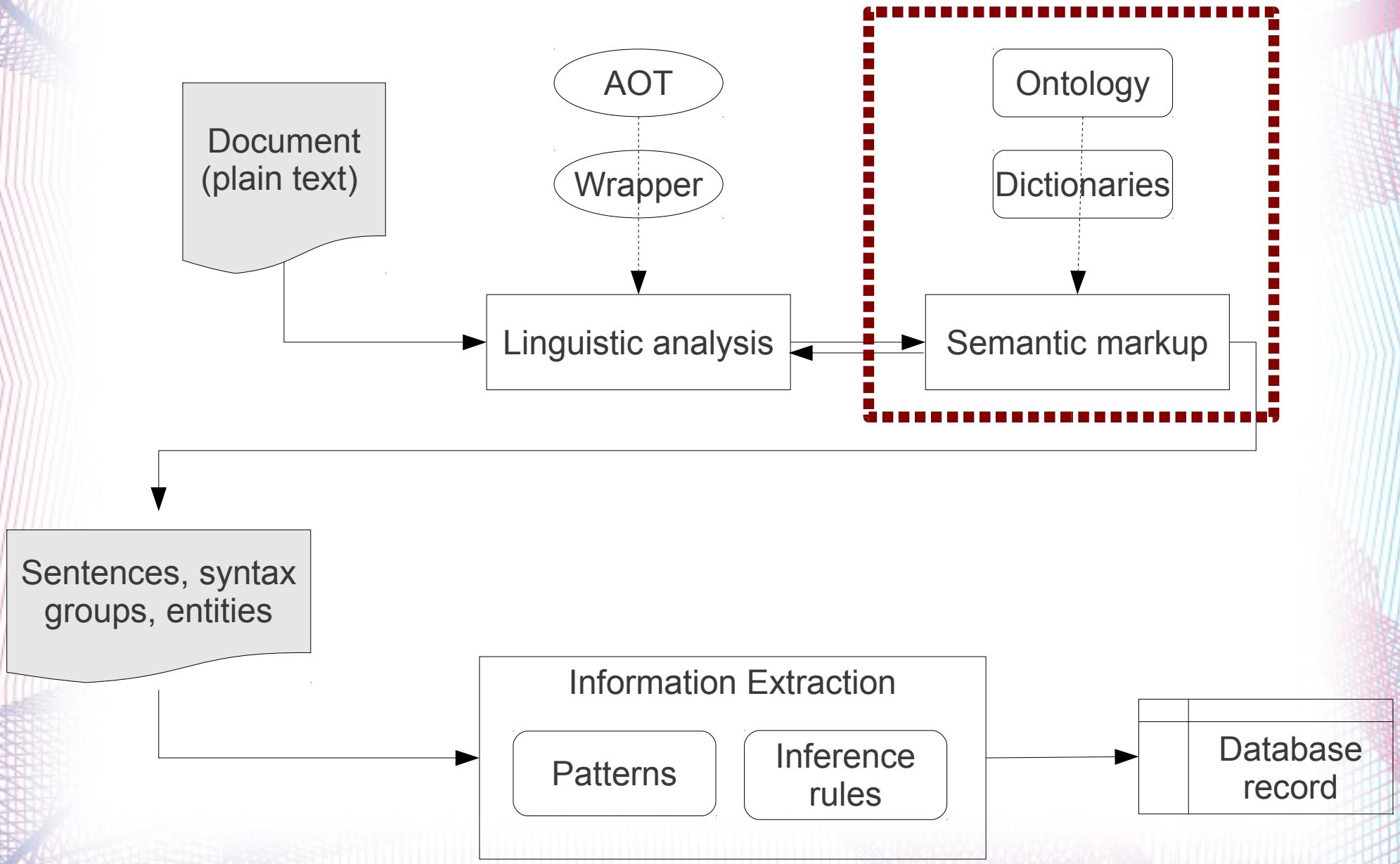
WRAPPER

- Grammar tags are mapped into common English tags
- For every binary relation in Synan output, we take the corresponding parent and child analyses from Lemm
  - all other analyses are removed
  - If the lemma for parent or child was null (e.g. a group) we infer information from Lemm
- If AOT produces two parents for a node (e.g. conjunctions) the wrapper adjusts the links so that they form a proper tree structure
- Some groups are converted into relations
- If a word does not participate in any relation, its analysis is taken entirely from Lemm output, passing along any unresolved ambiguity

WRAPPER

0	0	На	НА	NIL:>NIL	PREP	NIL
1	3	берегу	БЕРЕГ	PREPGR:>0	NOUN	(2GENL INAN MASC LOC SG)
2	10	пограничной	ПОГРАНИЧНЫЙ	ADJ_NOUN:>3	ADJ	(INAN ANIM FEM GEN SG)
3	22	реки	РЕКА	GEN_NOMGR:>1	NOUN	(INAN FEM GEN SG)
4	27	задержаны	ЗАДЕРЖАТЬ	NIL:>NIL	SPARTICIP	(PASS TRV PERF INAN ANIM PAST PL)
5	37	двадцать	ДВАДЦАТЬ	CARD_ORD_GR:>6	CARD	(NOM)
6	46	семь	СЕМЬ	NUM_NOUN:>8	CARD	(NOM)
7	51	нелегальных	НЕЛЕГАЛЬНЫЙ	ADJ_NOUN:>8	ADJ	(QADJ INAN ANIM GEN PL)
8	63	мигрантов	МИГРАНТ	SUBJ:>4	NOUN	(ANIM MASC GEN PL)

General Scheme



Ontology structure

CONCEPT HIERACHY

- IS-A relation

- multiple inheritance

Ontology structure

CONCEPT HIERACHY

- IS-A relation

- multiple inheritance

ENGLISH LEXICON

1. Implicit:
   - If a concept name is the only word it is concidered to be a word which can be found in a text
   - it as also possible to add single word synonyms (aliases)

2. Explicit
   - Multiword English lexicon

Example: transport

```
(DEFCONCEPT C-ART_AIR :TYPEOF (C-ARTIFACT C-TRANSPORT-RELATED))
(DEFCONCEPT C-ART_LAND :TYPEOF (C-ARTIFACT C-TRANSPORT-RELATED))
(DEFCONCEPT C-ART_WATER :TYPEOF (C-ARTIFACT C-TRANSPORT-RELATED))

(DEFCONCEPT A-FLIGHT  :TYPEOF (C-ART_AIR C-MISSION))
(DEFCONCEPT FLIGHT :TYPEOF (A-FLIGHT))
(DEFCONCEPT C-PLANE :TYPEOF (C-ART_AIR))
(DEFCONCEPT CARRIER :TYPEOF (C-ART_WATER C-ART_LAND C-ART_AIR))
(DEFCONCEPT GUNSHIP :TYPEOF (C-ART_AIR))
(DEFCONCEPT LAUNCHER :TYPEOF (C-ART_AIR C-VEHICLE))
(DEFCONCEPT ROCKET :TYPEOF (S-ARMS C-VEHICLE))
(DEFCONCEPT SHUTTLE :TYPEOF (C-ART_AIR C-VEHICLE))
(DEFCONCEPT VEHICLE :TYPEOF (C-ART_LAND C-ART_AIR C-VEHICLE))

(DEFCONCEPT A-PLANE  :TYPEOF (C-PLANE))
(DEFCONCEPT PLANE :ALIAS (JET AIRPLANE AIRLINER AIRCRAFT AEROPLANE HELICOPTER CHOPPER) :TYPEOF (A-PLANE))

(DEFCONCEPT BUS :ALIAS (MINIBUS) :TYPEOF (C-ART_LAND))
(DEFCONCEPT CAR :ALIAS (SUV LIMOUSINE) :TYPEOF (C-ART_LAND))
(DEFCONCEPT CRUISER :TYPEOF (C-ART_LAND C-ART_WATER))
(DEFCONCEPT MOTORBIKE :ALIAS (MOTORCYCLE) :TYPEOF (C-ART_LAND))
(DEFCONCEPT PATHFINDER :TYPEOF (C-ART_LAND))
(DEFCONCEPT SUBWAY :TYPEOF (C-ART_LAND))
(DEFCONCEPT TANK :TYPEOF (C-ART_LAND))
(DEFCONCEPT TRAILER :ALIAS (MINIVAN) :TYPEOF (C-ART_LAND))
(DEFCONCEPT TRUCK :ALIAS (LORRY) :TYPEOF (C-ART_LAND))

(DEFCONCEPT A-SHIP :TYPEOF (C-ART_WATER))
(DEFCONCEPT BOAT :ALIAS (SPEEDBOAT) :TYPEOF (C-ART_WATER))
(DEFCONCEPT FERRY :TYPEOF (C-ART_WATER))
(DEFCONCEPT FLEET :TYPEOF (C-ART_WATER))
(DEFCONCEPT FRIGATE :TYPEOF (C-ART_WATER))
(DEFCONCEPT LIFEBOAT :TYPEOF (C-ART_WATER))
(DEFCONCEPT SHIP :ALIAS (YACHT) :TYPEOF (C-ART_WATER))
(DEFCONCEPT SUBMARINE :TYPEOF (C-ART_WATER))
```

Ontology structure

CONCEPT HIERACHY

- IS-A relation

- multiple inheritance

ENGLISH LEXICON

1. Implicit:
   - If a concept name is the only word it is concidered to be a word which can be found in a text
   - it as also possible to add single word synonyms (aliases)

2. Explicit
   - Multiword English lexicon

Ontology structure

CONCEPT HIERACHY

- IS-A relation

- multiple inheritance

ENGLISH LEXICON

1. Implicit:
   - If a concept name is the only word it is concidered to be a word which can be found in a text
   - it as also possible to add single word synonyms (aliases)

2. Explicit
   - Multiword English lexicon

Russian lexicon

- Words
- Multiword expressions (in a form of low-level patterns)

Ontology structure

CONCEPT HIERACHY

- IS-A relation

- multiple inheritance

ENGLISH LEXICON

1. Implicit:
   - If a concept name is the only word it is concidered to be a word which can be found in a text
   - it as also possible to add single word synonyms (aliases)

2. Explicit
   - Multiword English lexicon

DICTIONARIES
- INSTANCE-OF relation
- locations
- diseases
- companies
- persons
- etc...

Russian lexicon

- Words
- Multiword expressions (in a form of low-level patterns)

Ontology structure

CONCEPT HIERACHY

- IS-A relation

- multiple inheritance

ENGLISH LEXICON

1. Implicit:
   - If a concept name is the only word it is concidered to be a word which can be found in a text
   - it as also possible to add single word synonyms (aliases)

2. Explicit
   - Multiword English lexicon

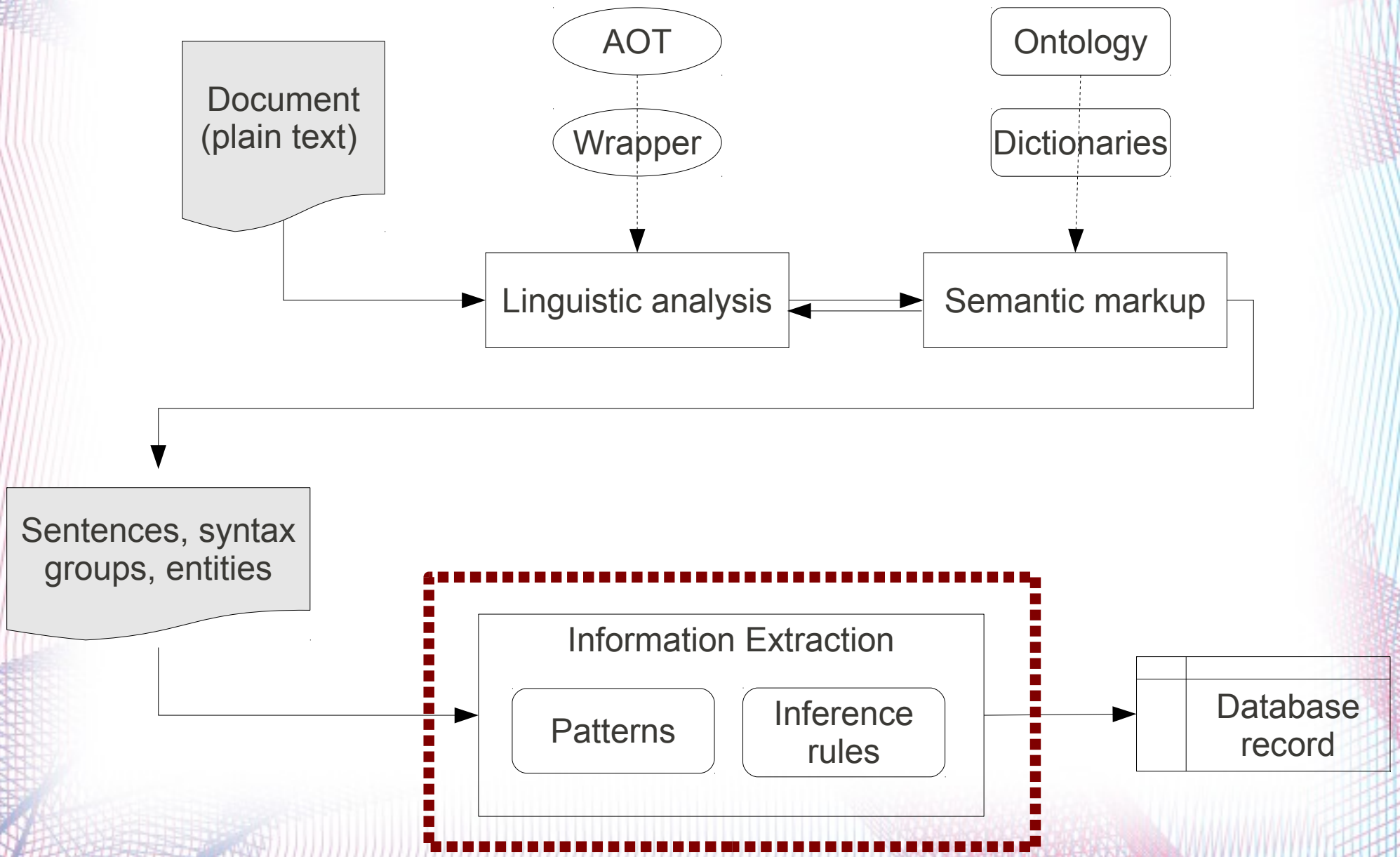
Russian lexicon

- Words
- Multiword expressions (in a form of low-level patterns)

DICTIONARIES
- INSTANCE-OF relation
- locations
- diseases
- companies
- persons
- etc...

Dictionaries mapping to Russian

General Scheme



Patterns

```
;; example
;; Житель Дагестана пересекал белорусско-украинскую границу с пистолетом на ремне.
;; A citizen of Dagestan crossed Russian-Belorussian border with a gun attached to his belt
;; Definition
(defpattern rus-cross-border-act
                " mix* noun-group(c-person) mix*  finv-group(cross-border) mix* noun-group(border) mix*
item-with? sa* location-pp-loc?
                        :
                suspect=2.attributes, anchor=4.attributes, border=6.attributes"
 )

;; Constraints: here we check syntax properties but it also possible to check any other
(lexical, semantic) features as well as their combination
(defun constraint-rus-cross-border-act ()
  (with-bindings (suspect anchor border)
    (active-clause-check-p suspect anchor border)))

;; Action: event is created; it is also possible to create entities (for low-level patterns)
(defun whenpattern-rus-cross-border ()
  (with-bindings (suspect anchor item-with location)
   (event (assert-event :predicate 'SECURITY_EVENT
      :type      'ILLEGAL-MIGRATION
      :subtype   'ILLEGAL-ENTRY
      :suspect   suspect
      :anchor    verb-head
      :location  location
      :item      item-with-entity))))
```

Patterns

- Fixed word order
- Semantic classes verification
- Verification of grammatical features (it may could be any features, the most common are POS)
- Some elements may be optional (?) or multiple (*)
- It is possible to use sub-patterns

This pattern language is already developed for English.

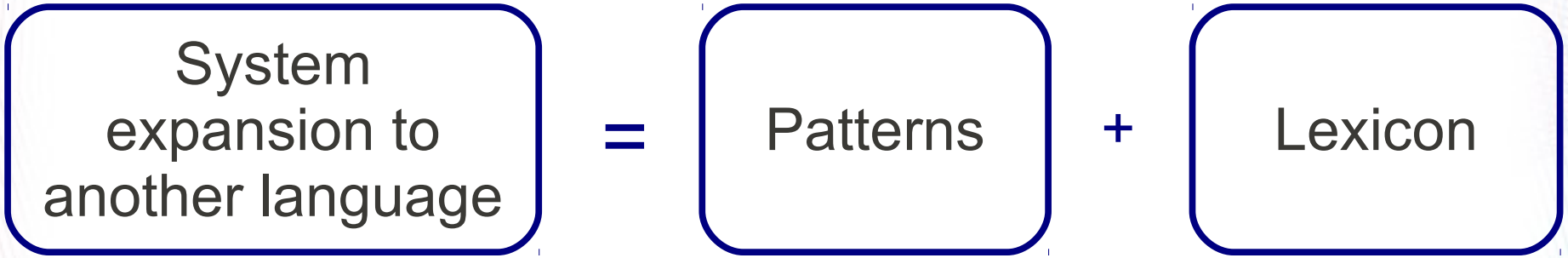
Inference rules

```
;; Area (here: text, plus-minus one sentence)
(def-kb-infrule IR-CRISIS->TYPE-ON-SUSPECT (:pool :discourse :distance 1)

;; Event, found by patterns
  (?crisis-event   (event :predicate SECURITY_EVENT
                   :TYPE      CRISIS
                                        :suspect    ?suspect
                   :SUBTYPE   ?subtype))

;; Some other property form text
 (?perpetrator-entity   (entity :CLASS       (isa C-PERPETRATOR)))

  -->
;; Action (here: event type is changed)
(new-event-type    (case-isa perpetrator-class
       (C-HUMAN-TRAFFICKER  'HUMAN-TRAFFICKING)
       (C-SMUGGLER          'SMUGGLE)
       (C-MIGRANT          'ILLEGAL-MIGRATION)
       (C-TERRORIST         'CRISIS)
       (C-KIDNAPPING        'CRISIS)
       (TRANSPLANTOLOGIST   'HUMAN-TRAFFICKING)))
```

Inference rules

- Work on semantic level
- Do not check any “physical” features, except for distance
- As a consequence cover much more language phenomena than patterns (including stylistic variations)
- Do not depend on language (sic!)
- Cannot be used without patterns (do not precise enough)

Patterns



- All other parts (at least theoretically) can be borrowed from the existing system
- Key question: what is the optimal from for Russian patterns?

First idea: just copy English patterns

- *Subject – Verb – Object*

- Check all the grammar agreement (to distinguish subject and object)

- Word order alterations need additional patterns

  - though they may share the same constraints or actions

Reasons

- Russian has a flexible word order
- However, some word orders are more preferable
- News use a standard language: many clichés, common constructions, bureaucratic collocations etc.
- That is why we can apply Information Extraction

Reasons

- Russian has a flexible word order
- However, some word orders are more preferable
- News use a standard language: many clichés, common constructions, bureaucratic collocations etc.
- That is why we can apply Information Extraction

Well, it was too optimistic

All variants are possible

Полиция арестовала преступника

Полиция преступника арестовала

Арестовала преступника полиция

Преступника полиция арестовала

Преступника арестовала полиция

Арестовала полиция преступника

A police arrested a perpetrator

All variant are possible

Полиция арестовала преступника

Полиция преступника арестовала, а не оштрафовала

Арестовала преступника полиция, а не таможня

Преступника полиция арестовала в тот момент, когда он пытался пересечь границу

Преступника, который пять лет скрывался от закона, в конце концов арестовала полиция

Арестовала наша доблестная полиция преступника только после того, как поступил звонок “сверху”

All variant are possible

Even in news word order depends on:

- information structure (topic – focus)

- relative clauses added to subject (object) may change word order

- stylistic features, e.g. irony

Patterns have to catch other types of clauses:

- passive

- relative clauses (...*perpetrator, who was arrested by police...*)

- participle clause (*...perpetrator arrested by police...*)

For English we have a paraphrase module, which produces all these forms from an active clause. It would be quite useful to make such a system for Russian.

However, it is clear that fixed word order in patterns leads to unnecessary growth of the pattern base.

Preliminary results

- Word order is non-informative
- However, the existing pattern search algorithm based on a fixed order and it would be too painful to change it
- Another solution: patterns as triggers that create events
- Inference rules responsible for specification and extra slots filling
- And no need to change logic – only knowledge bases!

Preliminary results

- Word order
- However, th
based on a
painful to cl
- Another sol
create events
- Inference rules responsible for specification and extra slots filling

Similar to:

Tanev, H., Zavarella, V., etc.: Exploiting Machine Learning Techniques to Build an Event Extraction System for Portuguese and Spanish. LINGUAMÁTICA Journal 2, 55–66 (2009)

and other papers by the authors

Experiment

- Single-word (collocation) pattern
- No grammar is checked
- Pattern checks only semantic class:

  class-or(c-illegal-activity, c-authority-activity)

Single-word pattern

class-or(c-illegal-activity, c-authority-activity)

- Semantic:
  - c-authority-activity includes c-report (announce) → too general
  - the majority of texts devoted to illegal activity are reviews or new about some general events (conferences, government programs etc.)
  - we need events related to arrest, or sentence, or deportation – it was not clear a priori, before the experiment

Single-word pattern

class-or(c-illegal-activity, c-authority-activity)

- Ambiguity:
  - same verbs are used in legal and common context
  - *to accuse smb. of telling lies - to accuse smb. as a thief*
- In some cases syntax determines event type:
  - *A policeman caught a perpetrator* → ARREST
  - *A policeman was caught by a perpetrator* → KIDNAPING
  - In Russian only cases are different

    *Полицейск**ий** поймал преступник**а***

    *Полицейск**ого** поймал преступник*

Syntax is unavoidable.

Final pattern form

- Trigger (verb, or participle, or nominalisation) + object
- Two words is much better than three – less variants, less permutations
- For now we analyze the following constructions:

VERB + NOUN

NOUN + VERB *(<policeman> arrested migrant)*

PARTICIPLE + NOUN

NOUN + PARTICIPLE (*migrant is arrested)*

NOUN + NOUN (*arrest of a migrant*)

Implementation

- Two simple patterns :

  “class-or(p-arrest-or-charge, p-sentence-or-jail, c-deport) sa* noun-group(c-person)”
  “noun-group(c-person) sa* class-or(p-arrest-or-charge, p-sentence-or-jail, c-deport)”

- Reasonable amount of constraints

```
(or
    (in-relation-p anchor person :VERB+DIRECT_OBJ) ;; verb + object
     (member (object-role anchor :category) '(:noun)))
    (and
     (member (object-role anchor :category) '(:finv :inf))  ;; if AOT didn't find :VERB+DIRECT_OBJ
     (object-case-check-p person))
    (and
     (member (object-role anchor :category) '(:particip :sparticip)) ;; passive voice
     (subject-case-check-p person))))))
```

- Pattern fills as many slots as possible:

```
(type (cond
        ((isa suspect-class 'C-HUMAN-TRAFFICKER)  'HUMAN-TRAFFICKING)
        ((isa suspect-class 'C-SMUGGLER)          'SMUGGLE)
        ((isa suspect-class 'CUSTOMS-OFFICER)     'SMUGGLE)
        ((isa suspect-class 'C-MIGRANT)   'ILLEGAL-MIGRATION)
        (T 'CRISIS)))
(subtype (cond
        ((isa suspect-class 'C-HUMAN-TRAFFICKER)  'UNSPECIFIED)
        ((isa suspect-class 'C-SMUGGLER)          'UNSPECIFIED)
                        ((isa suspect-class 'C-ILLEGAL-MIGRANT)    'ILLEGAL-STAY)
        ((isa suspect-class 'C-MIGRANT)           'UNSPECIFIED)
        ((isa suspect-class 'C-TERRORIST)         'TERRORISM)
        ((isa suspect-class 'C-KIDNAPPING)        'VIOLENCE)
        ((isa anchor-head 'C-DEPORT) 'DEPORTATION)
        ((isa anchor-head 'C-ARREST) 'ARREST-INTERSEPTION)
        ((isa anchor-head 'C-CHARGE) 'CHARGE)
        ((isa anchor-head 'P-SENTENCE-OR-JAIL) 'SENTENCE)))
```

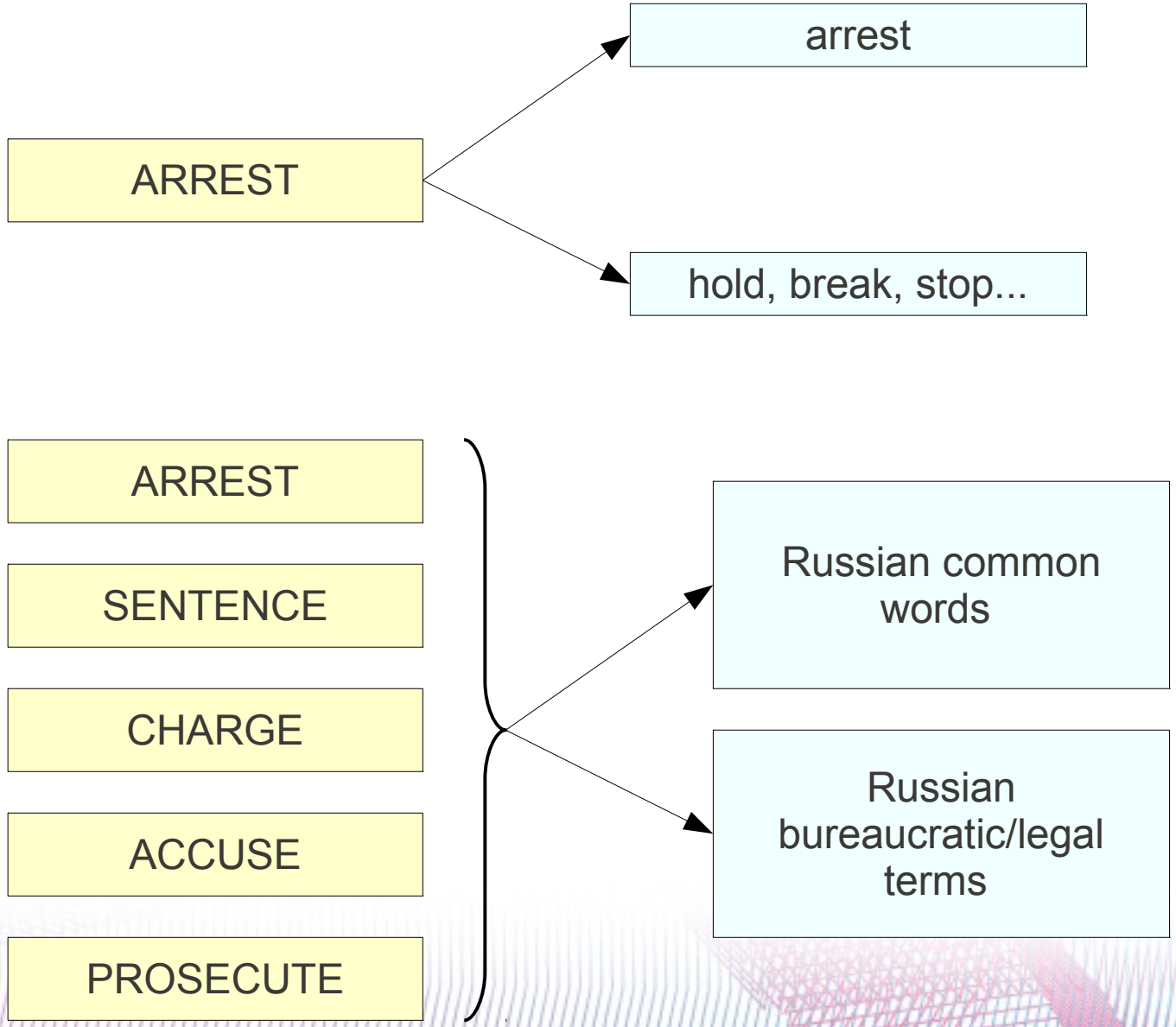
Inference rules

- In most of the cases pattern creates CRISIS event
- Inference rules specify event type, fill additional slots (at least locations)
- Working on the Russian system we succeed to use without major differences inference rules developed previously for the English system
- Furthermore, we added several rules, which now work for both Russian and English:
  - **transplantation** → HUMAN-TRAFFICKING-ORGANS;
  - **border guards** → MIGRATION-ILLEGAL-ENTRY;
  - **customs officers** → SMUGGLING.

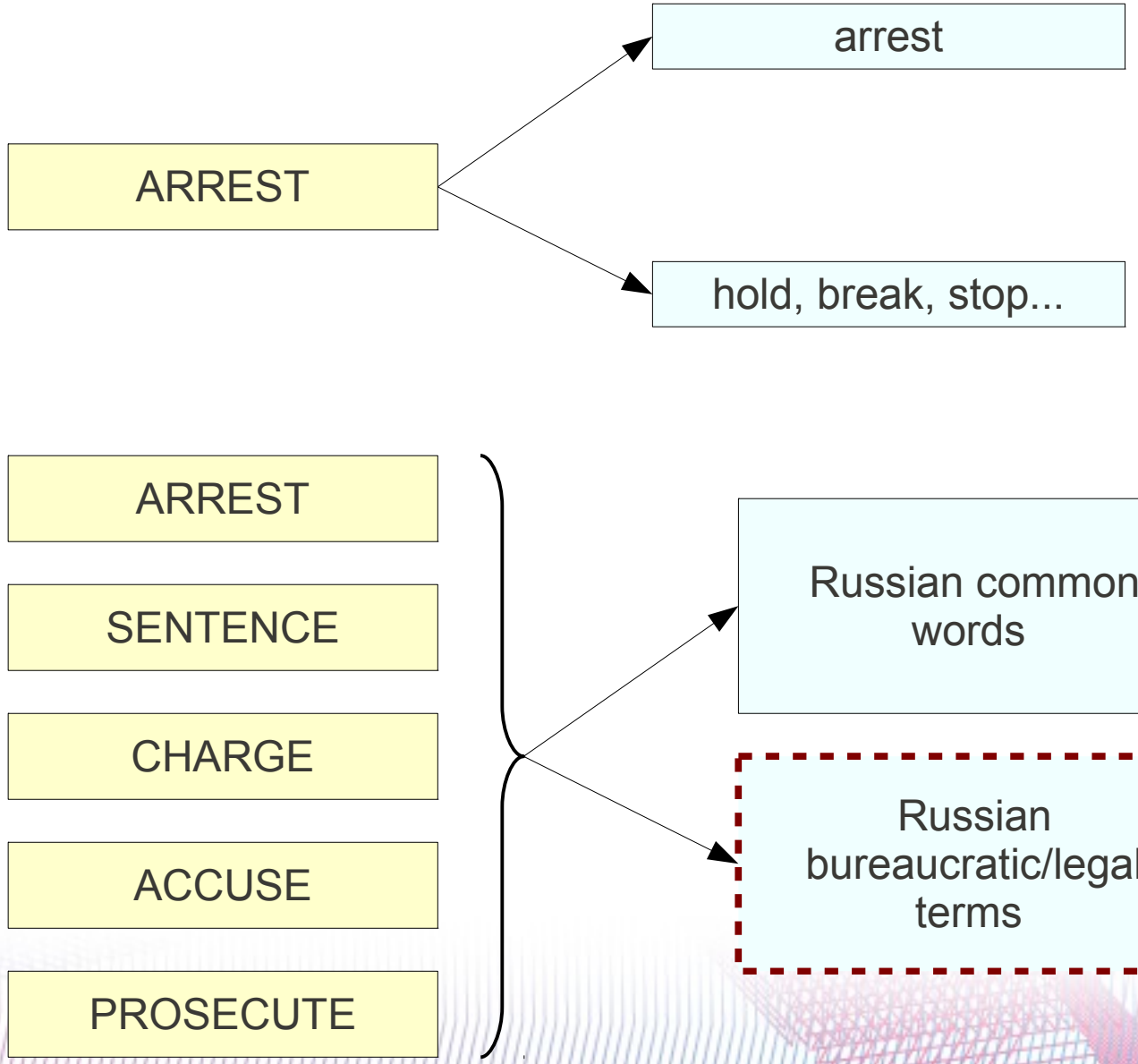
Additional patterns

- I'd like to put all the Information Extraction into one “clever” pattern
- For different reasons it is not always possible:
  - Different semantic of object
    - DETERMINE-CONTRABAND (*police arrested a contraband)*
  - Different semantic of verb
    - CROSS-BORDER (*a perpetrator crossed a border*)
  - Stylistic features
    - RUS-PROSECUTE (Russian collocation for “prosecute” used in legal context only and may be used alone)

Stylistic features



Stylistic features



Is it possible to use stylistic features in more clever way?

- All the Russian words are **dictionary units**
- Any grouping is made with **ontology units** only
- Attempts to group words lead to unmotivated ontology increase with artificial concepts
- Working with Russian system we cleaned up the ontology – all English related concepts were moved to the lexicon
- For now patterns do not distinguish common and legal words

Patterns vs. inference rules

- Patterns: need an exact ontology
  - *A person arrested on a border* → ILLEGAL-ENTRY
  - *Goods arrested on a border* → SMUGGLING
- Rules: need a thesaurus
  - *Border, border-guard, illegal entry* → ILLEGAL-ENTRY
  - *Customs, customs-officer, contraband* → SMUGGLING
- Concept base:
  - balance between exactness and completeness;
  - contradictions between patterns and rules and also between languages

Evaluation

- 64 documents
- Part of them marked up in advanced before system development started
- Another part based on early prototype results (made by St.Petersburg State University students)
- 65 event
- One third of documents contains events

Evaluation

- 64 documents
- Part of them marked up in advanced before system development started
- Another part based on early prototype results (made by St.Petersburg State University students)
- 65 event
- One third of documents contains events

	Recall	Precision	F-measure
Russian system	47	34	39.1
English system	48	45	46.15

Evaluation

	Russian system		English system	
	Recall	Precision	Recall	Precision
TYPE	63	51	64	63
SUBTYPE	39	30	28	29
COUNTRIES	47	41	54	57
LOCATION	35	25	27	14
SUSPECT	57	48	44	46
TOTAL	42	35	32	90
OPERATIONAL ACTIVITY	82	17	58	62
ACTING AUTHORITY	0	0	25	25
TIME	0	0	47	45
ALL SLOTS	**46**	**33**	**48**	**45**
F-MEASURE	**39.01**		**46.15**	

Evaluation

	Russian system		English system	
	Recall	Precision	Recall	Precision
TYPE	63	51	64	63
SUBTYP				29
COUNTRI				57
LOCATIO				14
SUSPEC				46
TOTAL				90
OPERATIO ACTIVIT				62
ACTING AUTHORITY	0	0	25	25
TIME	0	0	47	45
ALL SLOTS	**46**	**33**	**48**	**45**
F-MEASURE	**39.01**		**46.15**	

The numbers reflect not the system performance only but also peculiarity of the keys themselves.

A correct and well-balanced test suite development is a challenging task itself.

The test suite is regularly specified and amplified.

Inference rules contribution

RUSSIAN SYSTEM

	PATTERNS + RULES		PATTERNS ONLY	
	Recall	Precision	Recall	Precision
TYPE	63	51	40	33
SUBTYPE	39	30	10	8
ALL SLOTS	46	33	28	34
F-MEASURE	38.62		30.87	

ENGLISH SYSTEM

	PATTERNS + RULES		PATTERNS ONLY	
	Recall	Precision	Recall	Precision
TYPE	64	63	31	32
SUBTYPE	28	29	16	17
ALL SLOTS	48	45	34	43
F-MEASURE	46.15		37.85	

Further work

- Patterns:
  - Patterns specification, in particular filling of missing slots
  - Expansion to different syntax structures (related clauses, participle clauses etc.)
  - Implementation of date and time pattern set for Russian
  - Implementation of Named Entity Recognition pattern set (in addition to what is done by AOT)
- Ontology:
  - Other relation types
  - Inference rules improvement using new relations
  - Futher lexical sources development and normalization