Adapting the PULS Event Extraction Framework to Analyze Russian Text

Lidia Pivovarova,[1,2] **Mian Du,**[1] **Roman Yangarber**[1]
[1] Department of Computer Science
University of Helsinki, Finland
[2]St. Petersburg State University, Russia

Abstract

This paper describes a plug-in component to extend the PULS information extraction framework to analyze Russian-language text. PULS is a comprehensive framework for information extraction (IE) that is used for analysis of news in several scenarios from English-language text and is primarily monolingual. Although monolinguality is recognized as a serious limitation, building an IE system for a new language from the bottom up is very labor-intensive. Thus, the objective of the present work is to explore whether the base framework can be extended to cover additional languages with limited effort, and to leverage the pre-existing PULS modules as far as possible, in order to accelerate the development process. The component for Russian analysis is described and its performance is evaluated on two news-analysis scenarios: epidemic surveillance and cross-border security. The approach described in the paper can be generalized to a range of heavily-inflected languages.

1 Introduction

1.1 Problem Statement

PULS[1] is a framework for information extraction from text (IE), designed for decision support in various domains and scenarios. To date, work on PULS has mostly concentrated on English-language text, though some effort has gone into adapting PULS to other languages, (Du et al., 2011). This paper describes a component that is used to extend PULS to analyze Russian-language text, and demonstrates its performance on two IE scenarios: infectious epidemics and cross-border security. The epidemics scenario is built to provide an early warning system for professionals and organizations responsible for tracking epidemic threats around the world. Because information related to outbreaks of infectious disease often appears in news earlier than it does in official sources, text mining from the Web for medical surveillance is a popular research topic, as discussed in, e.g., (Collier et al., 2008; Huttunen et al., 2002; Rortais et al., 2010; Zamite et al., 2010). Similarly, in the security scenario, the system tracks cross-border crime, including illegal migration, smuggling, human trafficking, as well as general criminal activity and crisis events; text mining for this scenario has been previously reported by (Ameyugo et al., 2012; Atkinson et al., 2011). The new component monitors open-source media in Russian, searching for incidents related to the given scenarios. It extracts information from plain, natural-language text into structured database records, which are used by domain specialists for daily event monitoring. The structure of the database records (called *templates*) depends on the scenario. For the epidemics scenario the system extracts the fields: disease name, location of the incident, date, number of victims, etc. In the security domain, the template contains the type of event, date and location, the perpetrator, number of victims (if any), goods smuggled, etc.

Monolinguality is a serious limitation for IE, since end-users are under growing pressure to cover news from multiple languages, (Piskorski et al., 2011). The Russian-language component that we describe here is an experiment in extending PULS to multi-lingual coverage. Our aim is to explore whether a such an extension can be built with limited effort and resources.

1.2 Prior work on IE from Russian

IE in Russian has been the topic of several recent studies. For example, (Piskorski et al., 2011) uses

[1]http://puls.cs.helsinki.fi

Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, pages 100–109,
Sofia, Bulgaria, 8-9 August 2013.

Russian among other languages to study information fusion across languages. Extraction techniques are used for ontology learning in (Bocharov et al., 2010) and (Schumann, 2012). The University of Sheffield's GATE system, (Bontcheva et al., 2003), which supports multi-lingual IE, has been adapted to Russian as part of the MUSE-3 project, (though little is published on functionality available in Russian). HP Labs have recently started adaptation of their information extraction solutions to Russian, (Solovyev et al., 2012).

Much literature devoted to Russian-language information extraction is published only in Russian; a brief review can be found in (Khoroshevsky, 2010). The majority of existing applications for Russian IE, and Natural Language Processing (NLP) in general, are commercially based, and are either published in Russian only, or not at all. One major player in Russian text mining is Yandex, the leading Russian search engine. Yandex uses IE to support its main search service, e.g., to underline addresses and persons in search results, and in a service called "Press Portraits,"[2] which builds profiles for various personalities found in the news. A profile may include the profession, biographical facts, news that s/he is involved in, and related people—using information automatically extracted from on-line Russian media. Yandex also recently unveiled an open-source toolkit Tomita, for developing IE systems based on context-free grammars.

Dictum, a company that builds applications for NLP and sentiment analysis in Russian, provides a toolkit for Russian morphological, syntactic and semantic analysis. Their *Fact Extraction* component[3] finds persons, organizations, locations, etc., and creates simple facts about persons: corporate posts, date of birth, etc.

RCO, a company focused on research and development of text analysis solutions, provides the RCO Fact Extractor tool[4], which performs fact extraction from unstructured text. One common usage scenario is setting up a list of target objects (persons, companies) and extracting all events where these objects are mentioned as participants. The tool also includes a module that allows the user to adjust search patterns.

With the exception of Tomita and AOT (see Section 3), few resources are available in open-source.

[2]http://news.yandex.ru/people

[3]http://dictum.ru/en/object-extraction/blog

[4]http://www.rco.ru/eng/product.asp

2 The Baseline English System

The PULS news-tracking pipeline consists of three main components: a Web-crawler that tries to identify potentially relevant articles using a broad keyword-based Web search; a rule-based Information Extraction system that uses patterns acquired through semi-supervised learning, that determines exactly what happened in the article, creating a structured record that is stored in the database; and a relevance classifier that determines the relevance of the selected articles—and events that they describe—to the particular use-case scenario and the users' needs. This paper will mostly focus on the IE component, as other two components are language-independent.

The IE system contains modules for lower-level—morphological and syntactic—analysis, as well as higher-level—semantic—analysis, and produces filled templates on output, extracted from an input document, (Du et al., 2011).

PULS follows a classic IE processing pipeline:

- Pre-processing,
- Lexical markup,
- Shallow syntactic analysis/chunking,
- Semantic pattern matching
- Reference resolution and logical inference

Pre-processing includes tokenization, part-of-speech tagging, processing of punctuation, numeric expressions, etc.

Lexical markup is tagging of lexical units found in text with semantic information found in a dictionary and/or ontology. PULS uses several domain-independent and domain-specific lexicons and ontologies. The ontology is a network of concepts organized in a hierarchy by several relations, among which the "is-a" relation is the most common. One key factor that enables the addition of new languages efficiently is that the ontology is language-independent. The system uses the lexicons to map words into concepts. A lexicon consists of word-forms and some common multi-word expressions (MWEs), which appear in text and represent some ontology concept. We assume that *within a given domain* each word or

MWE in the lexicon represents exactly one concept, (Yarowsky, 1995). A concept may be represented by more than one word or MWE.[5] Each scenario has its own scenario-specific ontology and lexicons; the Epidemics ontology consists of more than 4000 concepts (which includes some disease names). Diseases are organized in a hierarchy, e.g., "hepatitis" is a parent term for "hepatitis A". The Security ontology consists of 1190 concepts.

The domain-specific lexicon is a collection of terms that are significant for a particular scenario, mapped to their semantic types/concepts. The Security and Epidemics scenarios use a common location lexicon, that contains approximately 2500 names of countries, cities and provinces. Locations are organized according the "part-of" relation: cities are part-of provinces, which are part-of states, etc.

Syntactic analysis is implemented as a cascade of lower-level patterns. PULS uses shallow analysis (chunking), which does not try to build complete syntactic tree for a sentence but recognizes local grammatical structures—in particular, the noun and verb groups. This phase also identifies other common constructions needed for IE, (names, dates, etc.). As a result of the syntactic analysis, each sentence is represented as a set of fragments.

The pattern base is the main component of the IE system, responsible for finding factual information in text. A pattern is a set of semantic, syntactic and morphological constraints designed to match pieces of natural-language text. When a pattern fires it triggers an action, which creates an abstract logical entity based on the text matched by the pattern. The entity is added to an internal pool of entities found in the document so far. Facts produced by the system are based on the entities in this pool. The patterns are arranged in a cascade such that the results produced by one pattern are used by subsequent patterns to form more complex entities.

Patterns operate on a sentence-by-sentence basis. To link information in the surrounding sentences PULS uses concept-based reference resolution and logical inference rules. The reference resolution component is a set of rules for merging mentions of the same object and events.

Inference rules work on a logical level (rather than text), operating on entities found at the preceding stages of analysis. These entities can be used to fill slots in an event description, for example, to find event time and location, or to perform logical inference. For example, if the event type is *human-trafficking* and a concept related to *organ-transplantation* is mentioned in the sentence, an inference rule may specialize the event type to *human-trafficking-organs*.

3 Russian Morphology and Syntax

To speed development, we use pre-existing tools for tokenization, morphological and syntactic analysis in Russian. The range of freely-available, open-source tools for Russian is quite narrow, especially for syntactic analysis. Significant efforts for overcoming this situation have been the focus of the recent "Dialogue" series of conferences[6], which organized workshops on Russian morphology, (Astaf'eva et al., 2010), and syntax, (Toldova et al., 2012). Workshops take the form of competitions, where the participants tackle shared tasks. Eight teams participated in the latest workshop, devoted to syntax. However, only one—AOT[7]—offers their toolkit under the GNU LGPL license.

The AOT toolkit, (Sokirko, 2001) is a collection of modules for NLP, including libraries for morphological, syntactic, and semantic analysis, language generation, tools for working with dictionaries, and GUIs for visualization of the analysis. Due to its open availability and high quality of linguistic analysis, AOT is currently a *de-facto* standard for open-source Russian-language processing.

The AOT morphological analyzer, called "*Lemm*", analyzes text word by word; its output for each word contains: an index, the surface form, the base lemma, part of speech, and morphological tags. Lemm works on the morphological level only, and leaves all morphological ambiguity intact, to be resolved by later phases.

Lemm uses a general-purpose Russian morphological dictionary, which can be edited and extended (e.g., with neologisms, domain-specific terms, etc.). To add a new lemma into the dictionary, one needs to specify its inflectional

[5] By default, words that appear only in the general-purpose dictionary, and do not appear in any domain-specific lexicon, are automatically identified with a concept having an identical name.

[6] Dialogue—International Conference of Computational Linguistics (http://www.dialog-21.ru/en/)

[7] The AOT project ("Automatic Processing of Text" in Russian)—www.aot.ru

paradigm. For Russian IE, we had to add to the dictionary certain words and terms that designate scenario-specific concepts, for example "мигрант" (*migrant*) and "гастарбайтер" (*gastarbaiter*), which have become common usage in recent Russian-language media.

The syntactic analyzer in AOT, "*Synan*", uses a hybrid formalism, a mix of dependency trees and constituent grammars. The output for a sentence contains two types of syntactic units: binary parent-child relations, and "groups", which are token sequences *not* analyzed further but treated as an atomic expression. This approach is theoretically natural, since certain syntactic units may not have a clear root, for example, complex name expressions ("Aleksey Sokirko") or numeric expressions ("forty five"). To make it compatible with the overall PULS structure, we transform all Synan output into dependency-tree form; groups simply become linked chains. Synan attempts to produce a complete, connected parse structure for the entire sentence; in practice, it produces a set of fragments, consisting of relations and groups. In the process, it resolves morphological ambiguity, when possible.

To unify the results of Lemm and Synan, we built a special "wrapper," (Du et al., 2011). The wrapper takes every binary (syntactic) relation in the Synan output, finds the items corresponding to the relation's parent and child in Lemm's output, and resolves their morphological ambiguity (if any) by removing all other morphological readings. If the lemma for parent or child is null—as, e.g., when the corresponding element is a group—we infer information from Lemm output for the element that is missed in Synan. If a word does not participate in any relation identified by Synan, its analysis is based only on Lemm output, *preserving* all unresolved morphological ambiguity—to be potentially resolved at a later stage, typically by scenario-specific patterns. Finally, the wrapper assembles the resulting analysis for all words into a set of tree fragments.

4 Russian Information Extraction

4.1 Ontology and Dictionaries

The ontology, a network of semantic classes, is language-independent, and in Russian IE, we used the pre-existing domain ontologies for the epidemics and security domains, with minor modifications. Most of the changes centered on removing vestiges of English language-specific information, e.g., by making explicit the distinctions among certain concepts that may be confounded in English due to ambiguity of English lexical units. For example, in English, the word "convict" means both the verb and the convicted person (patient nominalization), so it may be tempting to represent them by the same concept. In Russian, as in many other languages, these are different concepts as well as distinct lexemes.

A Russian domain-specific lexicon was added to the system. Russian IE uses a shared lexicon for epidemics and security. The lexicon contains not only translations of the corresponding English words, but also includes MWEs that appear in Russian media and correspond to the domain-specific concepts. The current Russian domain-specific lexicon contains approximately 1000 words and MWEs. Constructing the multiword lexicon for Russian is more complicated than for English because Russian has a rich morphology and complex grammatical agreement. For example, to find a simple *Adjective+Noun* collocation in text, the system needs to check that the adjective agrees with the noun in gender, case, and number. To resolve this problem, we built a special set of low-level patterns, which match MWEs. These patterns are subdivided into several classes, according to their syntactic form: *Adjective+Noun*, *Noun+Noun*, *Verb+Noun*, *Verb+Preposition+Noun*, etc. The grammatical constraints are coded only once for each class of pattern, and apply to all patterns in the class. For example, in the *Noun+Noun* class, the second noun must be in genitive case (a genitive modifier of the head noun), e.g., "цироз печени" (*cirrhosis of the liver*), or in the instrumental case, e.g., "торговля людьми" (*human trafficking*). This simplifies adding new MWEs into the dictionary.

We use the multilingual GeoNames database, (www.geonames.org) as the source of geographic information in Russian. The disease dictionary is mapped into Russian using the International Classification of Diseases.[8] The system also identifies common animal diseases: anthrax, African swine fever, rabies, etc.

4.2 Pattern Bases

The pattern base is the main component of the IE system for extracting higher-level logical objects.

[8] ICD10: http://www.who.int/classifications/icd/en/

	Syntactic variant	Example		Syntactic variant	Example
I	Verb + Object (active clause)	арестовали мигранта *[someone] arrested a migrant*	**II**	Object + Verb (reverse word order)	мигранта арестовали (same meaning)
III	Participle + Object (passive clause)	арестован мигрант *migrant is arrested [by someone]*	**IV**	Object + Participle (reverse word order)	мигрант арестован (same meaning)
V	Noun + Object (nominalization)	арест мигранта *arrest of a migrant*	**VI**	(reverse word order is rare, unlikely in news)	—

Table 1: Examples of syntactic variants for a single pattern Russian

Patterns are language-dependent and domain-dependent, which means that patterns must capture the lexical, syntactic and stylistic features of the analyzed text. It was not possible to directly translate or map the English pattern base into Russian for at least two reasons.

The first reason is technical. PULS's English pattern base has over 150 patterns for the epidemics domain, and over 300 patterns for security.[9] These patterns were added to the system through an elaborate pattern-acquisition process, where semi-supervised pattern acquisition for English text was used, (Greenwood and Stevenson, 2006; Yangarber et al., 2000), to bootstrap many pattern candidates from raw text based on a small set of seed patterns; the candidates were subsequently checked manually and included in the system. Many of these patterns are typically in "base-form", i.e., simple active clauses; the English system takes each active-clause, "subject-verb-object" pattern, and generalizes it to multiple syntactic variants, including passive clauses, relative clauses, etc. Thus we created the Russian domain-specific patterns directly in PULS's pattern-specification language. A pattern consists of a regular expression trigger and action code.

The second reason is theoretical. Unlike English, Russian is a heavily inflected, free word-order language. In English, the active "subject-word-object" clause has only one form, whereas in Russian all six permutations of the three elements are possible, depending on the information structure and pragmatic focus. This means that we would need 6 pattern variants to match a single active clause, and many more to process other clausal types. The free word-order makes it difficult to generate syntactic clausal variants; it also complicates the bootstrapping of patterns from seeds.

Therefore, for Russian we used a different strategy, close to that used by (Tanev et al., 2009) for Romance languages. In this approach, the patterns first create "shallow", incomplete events where only 1–2 slots are filled. Then, the inference rule mechanism attempts to fill the remaining slots and complete the events. The majority of Russian patterns currently consist of two elements (such as verb and object, or verb and subject), so that only two word-order variants are possible. Currently, the Russian patterns match five syntactic constructions. These are listed in Table 1, along with examples from the security scenario. All example phrases have the same meaning ("migrant was arrested") but different syntactic form. The active clause and the passive clause in Russian may have either V–O word order—types I and III—or O–V,—types II and IV. The difference between the active and the passive variants is in the grammatical features only, which are marked by flexions.

Types I, III, and V in the table can be captured by one simple pattern:

class(ARREST) noungroup(MIGRANT)

This pattern matches when a content phrase—belonging to *any* part of speech (noun, verb, or participle)—whose semantic head is the concept *"ARREST"* governs (i.e., in this case, precedes) a noun group headed by the concept *"MIGRANT"*. The pattern primitives—*class*, *noungroup* and others—build on top of the POS, syntactic, and morphological tags that are returned by the AOT wrapper. Types II and IV show variants of the pattern in reverse order. Note that the patterns use general ontology classes—shared with English—rather than literal words.[10]

When a pattern fires, the system checks the constraints on grammatical features (e.g., case and number agreement) on the matched phrases or words. We introduce three types of constraints: accusative object-case agreement for type I and

[9]The difference is partly due to the fact that the security scenario has several event types—illegal migration, human-trafficking, smuggling, general crisis—and sub-types, while epidemics deals with one event type.

[10]NB: in practice, the patterns are more complex because they allow various sentence modifiers to appear between verb and object, which is a standard extension to this basic form of the pattern.

Concept	Event type
organ-transplant	*Human-Trafficking-Organs*
border-guard	*Migration-Illegal-Entry*
customs-officer	*Smuggling*

Table 2: Examples of concepts found in context that trigger rules to specialize the event type

II, for nominative subject-case agreement for type III and IV, and and genitive-case nominalization agreement for type V. If the constraints are satisfied, the event is created—that is, the same event structure for any of the five pattern types.

For the security scenario the system currently has 23 such "basic" patterns. Most of them initially produce an event of a general class *CRISIS* and fire when the text mentions that someone was arrested, sentenced, jailed, etc. If additional security-related concepts are found in text nearby, inference rules will fill additional slots in the event template, and specialize the type of the event. The Russian security scenario uses *exactly the same* set of inference rules as does the English Security Scenario. Example rules are shown in Table 2. For example, when an inference rule finds in the context of an event a semantic concept that is a sub-type of the type given in the left column, the *Type* of the event is specialized to the corresponding value in the right column, Table 2.

For the epidemics scenario, the system currently uses only 7 patterns. Two produce an underspecified event, when the text mentions that someone has become sick. The actual disease name is again found by inference rules from nearby context; if no disease is mentioned, the event is discarded. Two additional patterns work "in reverse": they match in cases when the text mentions an outbreak or case of a disease. Then the inference rules try to find who is suffering from disease and the number of victims. The inference rules are again fully shared between English and Russian. Some of the patterns are "negative"—they match such statements as "there is no threat of epidemic", which appear often in official reports.

In addition, the Russian pattern base contains 41 lower-level patterns, common for the security and epidemics domains. These include, for example, patterns to match date expressions, to analyze collective-partitive noun groups (*"a group of migrants", "a team of doctors"*, and so on), which have general applicability.

Slot	*English system*			*Russian system*		
	rec	pre	F	rec	pre	F
Event Type	67	72	**69.41**	70	57	62.83
Suspect	46	52	**48.81**	52	44	47.67
Total	27	71	**46.47**	44	37	40.20
Countries	56	55	**55.49**	48	40	43.63
Time	29	29	**29.00**	29	22	25.02
All	53	58	**53.31**	55	45	49.09

Table 3: Border Security scenario evaluation

Event type	*English test suite*	*Russian test suite*
CRISIS	19	28
HUMAN-TRAFFICKING	4	4
ILLEGAL-MIGRATION	34	34
SMUGGLE	10	2
Total	67	68

Table 4: Distribution of event types in the test suites for the Security scenario

5 Evaluation

5.1 Security

For evaluation, we used a test corpus of 64 Russian-language documents. Several assessors annotated 65 events, and approximately one third of the documents contained events. We compared the Russian-language IE system with the English-language system. The English test suite consists of 50 documents with 70 events.

Evaluation results for the security domain are presented in table 3, with scores given for the main slots: *Event Type* (one of *Migration*, *Human Trafficking*, *Smuggling*, and *Crisis*), *Suspect*, *Total* (number of suspects), *Countries* (a list of one or more countries involved in event), and *Time* (event date). The table shows that currently the Russian system achieves a lower overall score than the English system—the F-measure for all slots is 4–5% lower, with precision being consistently lower than recall for the Russian system.

Note that the development of a correct and well-balanced test suite is in itself a challenging task, and hence the evaluation numbers may be biased. In the test suites used for these experiments, shown in table 4, the English security scenario includes more events of type *SMUGGLE* than the Russian validation suite, and both validation suites contain few events of type *HUMAN-TRAFFICKING*.

5.2 Epidemic Surveillance

For evaluation, we used a test corpus of 75 Russian documents. We asked several assessors to

Slot name	*English system*			*Russian system*		
	r	p	F	r	p	F
Disease	74	74	74.00	93	81	**86.58**
Country	65	67	65.98	91	86	**88.42**
Total	68	79	**73.09**	30	78	43.33
Time	56	58	**56.98**	38	52	43.91
Status	77	75	75.99	93	81	**86.58**
All Slots	68	69	68.83	70	71	**70.44**

Table 5: Epidemics scenario evaluation.

Event Type	**English**	**Russian**
Epidemic Surveillance		
DISEASE	31	5
HARM	825	412
Total	**856**	**417**
Border Security		
CRISIS	694	476
HUMAN-TRAFFICKING	10	12
ILLEGAL-MIGRATION	32	31
SMUGGLE	7	19
Total	**743**	**538**

Table 6: Number of events found by IE systems in parallel English-Russian news corpus.

correct events found by the system and add missing events in case they were not found by system. Assessors annotated 120 events. We compare the Russian-language IE system with the English-language system. The PULS English validation suite for Epidemics currently consists of 60 documents with 172 events.

Evaluation results are shown in table 5, where the scores are given for the main slots: *Disease*, *Country*, *Total* (number of victims), *Status* ("dead" or "sick") and *Time*. Results for the Russian system are somewhat better than for English. This is due in part to the bias in the process which we used to select documents for the test suite: the assessors marked documents in which the system found events, rather than searching and annotating documents from scratch. (This aspect of the evaluation will be corrected in future work.) The events that the system found could be relevant, spurious, or erroneous; in case the system missed an event, the assessor's job was to add it to the gold-standard answers. Note that in general the amount of irrelevant documents processed by PULS is much larger than the amount of relevant documents (only about 1% of all documents that contain keywords relevant to epidemics contain useful events). Thus it is impractical to ask assessors to read raw documents. As a consequence, the scores for the main slots, such as *Disease* or *Country*, may be overstated: the majority of documents mention only one disease, and since an event was found by the system in most documents selected for the test suite, the *Disease* slot is usually filled correctly. The results for the auxiliary slots, e.g., *Time*, *Total*, are closer to our expectation.

5.3 Comparison of Languages and Scenarios

In general, the epidemics scenario performs much better than security, both in Russian and English. This is due to fact that the task definition for epidemics is simpler, better formalized, and deals with one type of event only. As noted in (Huttunen et al., 2002), event representation in text may have different structure depending on the scenario: the "classic" IE scenarios, such as the MUC Management Succession or Terror Attacks, describe events that occur at a specific point in time, whereas other scenarios, such as *Natural Disasters* or *Disease Outbreaks* describe a process that is spread out in time and space. Consequently, events in the latter ("nature") scenarios are more complex, may have hierarchical structure, and may even overlap in text. From the theoretical point of view it would be interesting to compare how the *events representation*, (Pivovarova et al., 2013), differs in different languages. Moreover, such differences can be important in cross-language information summarization, (Ji et al., 2013).

We use a freely-available *comparable* news corpus, (Klementiev and Roth, 2006), to investigate the difference of event representation in English and Russian. The corpus contains 2327 BBC messages from the time period from 1 January 2001 to 10 May 2005, and their approximate translations from the *Lenta.ru* website; the translations may be quite different from their English sources and are stylistically similar to standard Russian news. We processed the corpora with the security and epidemics IE systems, using the respective language; the results are presented in the Table 6.

The table shows that for both scenarios the English system finds more events than the Russian, which probably means that coverage of the Russian IE is lower. We have yet to conduct a thorough evaluation of the events found. It is also clear from the table that specific events are much more rare than general events; for the security scenario, the majority of events have type CRISIS, which is a general type that indicates some incident related to crime; in the epidemics scenario, the majority of events have type HARM, i.e., which is a gen-

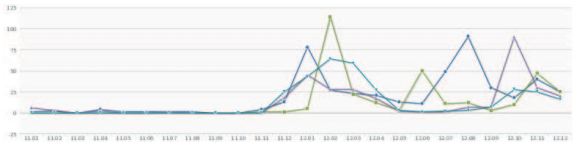

Figure 3: Monthly frequency of events for the four top-reported diseases in Russia

eral type indicating that there are victims (e.g., humans) suffering from some cause, not only harm caused by infections. The distributions of event types are similar in English and Russian corpora, which may hint that a lack of specific events may be a property of the scenarios, irrespective of the language. This agrees with the expectation that the majority of retrieved documents are not relevant.

6 Discussion

The Russian-language processing pipeline presented above is compatible with the working, pre-existing PULS IE system. It is worth noting again, that the output of the Russian-language analysis has the same form as that of the English-language PULS event extraction, that is, all fills for the template slots are output in *English* (except in the case of person names). This is made possible by the shared, language-independent ontology. An important benefit of this sharing is that the end-user is not required to understand Russian in order to determine whether the extracted facts and documents are relevant to her/his need. Thus, the slot fills may be presented in English, as shown in Figure 1. The document text, however, may be presented in Russian; users who can read Russian can see the original article text where event elements are indicated (by highlighting or underlining).

Figure 2 shows a summary-style list of events found from the news stream. The user can see events extracted from documents in a mix of languages (identified by the language tag in the left-most column). The database representation for events is shared and independent of the language; this permits the user get a grasp of current situation in the domain of interest, in more than one language.

We checked the impact of the Russian component on the system's coverage over the geographic area of the former USSR, which includes regions (outside Russia) where Russian may be used as a *lingua franca*, and may be common in press. Figure 3 shows the total number of events found in Russia, using both the Russian- and English-language IE systems for the four most frequently reported diseases. The check was conducted on news streams over 2011–2012. The number of events increases dramatically after deploying the Russian component, at the end of 2011 (near the middle of the timeline).

6.1 Conclusion

We have presented a "plug-in" extension to PULS, an English-language IE system, to cover Russian-language text. We currently handle two scenarios: Security and Epidemic Surveillance. The amount of effort needed to develop the Russian component was modest compared to the time and labour spent on the English-language IE system. The Russian system demonstrates a comparable level of performance to the baseline English IE: F-measure is about 4% lower for the Security scenario and 2% higher for the Epidemic Surveillance. We believe that this success is due to two main factors: first, the re-use of as many existing modules and knowledge bases as possible from the pre-existing English-language system; second, the use of shallow, permissive patterns in Russian in combination with logical inference rules.

In future research, we plan to further expand the pattern sets and lexicons, to analyze more kinds of syntactic and lexical phenomena in Russian. We plan to compare structural differences between the Security and Epidemics scenarios and their representation in Russian and English, to find language-dependent and language-independent features of the event representations. We plan to use cross-lingual analysis to obtain advances in two directions: first, pre-IE automatic pattern and paraphrase acquisition for free-word-order languages; second, post-IE aggregation of extracted information to improve overall quality by use of cross-document context, (Chen and Ji, 2009; Yangarber and Jokipii, 2005; Yangarber, 2006).

Acknowledgements

We thank Peter von Etter and Mikhail Novikov for help with the implementation; students of the Department of Information Systems, St. Petersburg State University, for annotating evaluation data. Work was funded in part by Frontex, Project on Automated Event Extraction, and the ALGODAN Center of Excellence of the Academy of Finland.

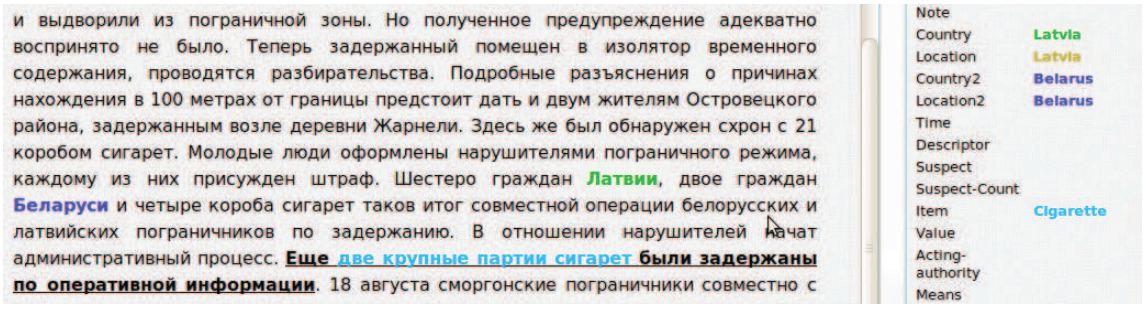

Figure 1: Document view and template view: a Smuggling event from the Security domain

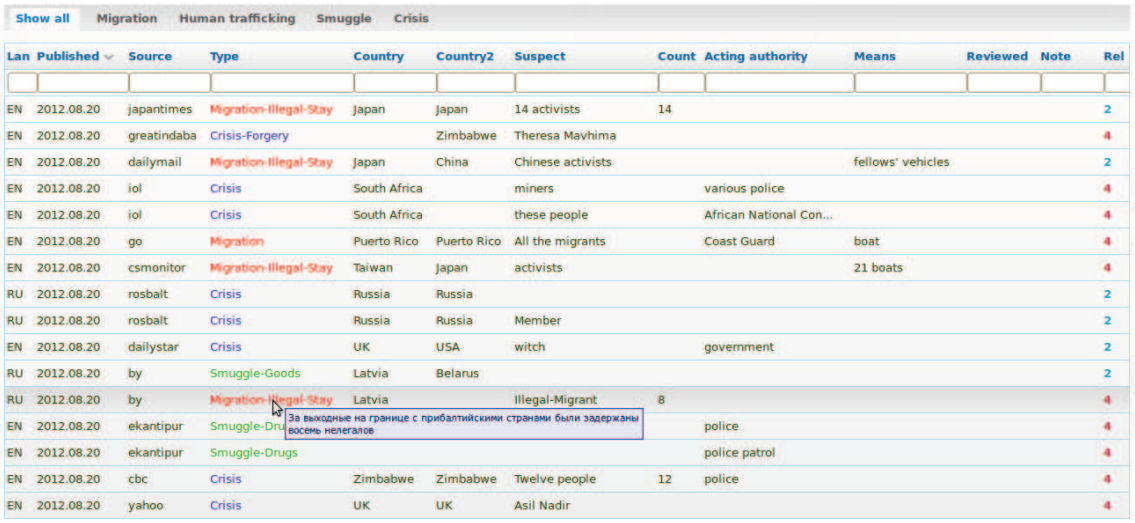

Figure 2: Summary view: a list of events in the Security domain. The tool-tip under the mouse shows a snippet of the original text, from which the event was extracted.

References

Ameyugo, G., Art, M., Esteves, A. S., and Piskorski, J. (2012). Creation of an EU-level information exchange network in the domain of border security. In *European Intelligence and Security Informatics Conference (EISIC)*. IEEE.

Astaf'eva, I., Bonch-Osmolovskaya, A., Garejshina, A., Grishina, J., D'jachkov, V., Ionov, M., Koroleva, A., Kudrinsky, M., Lityagina, A., Luchina, E., Sidorova, E., Toldova, S., Lyashevskaya, O., Savchuk, S., and Koval', S. (2010). NLP evaluation: Russian morphological parsers. In *Proceedings of Dialog Conference*, Moscow, Russia.

Atkinson, M., Piskorski, J., van der Goot, E., and Yangarber, R. (2011). Multilingual real-time event extraction for border security intelligence gathering. In Wiil, U. K., editor, *Counterterrorism and Open Source Intelligence*, pages 355–390. Springer Lecture Notes in Social Networks, Vol. 2.

Bocharov, V., Pivovarova, L., Rubashkin, V., and Chuprin, B. (2010). Ontological parsing of encyclopedia information. *Computational Linguistics and Intelligent Text Processing*.

Bontcheva, K., Maynard, D., Tablan, V., and Cunningham, H. (2003). GATE: A Unicode-based infrastructure supporting multilingual information extraction. In *Proceedings of Workshop on Information Extraction for Slavonic and other Central and Eastern European Languages*, Borovets, Bulgaria.

Chen, Z. and Ji, H. (2009). Can one language bootstrap the other: a case study on event extraction. In *Proceedings of the NAACL-HLT Workshop on Semi-Supervised Learning for Natural Language Processing*.

Collier, N., Doan, S., Kawazoe, A., Goodwin, R. M., Conway, M., Tateno, Y., Ngo, Q.-H., Dien, D., Kawtrakul, A., Takeuchi, K., Shigematsu, M., and Taniguchi, K. (2008). BioCaster: detecting public health rumors with a Web-based text mining system. *Bioinformatics*, 24(24).

Du, M., von Etter, P., Kopotev, M., Novikov, M., Tarbeeva, N., and Yangarber, R. (2011). Building sup-

port tools for Russian-language information extraction. In Habernal, I. and Matoušek, V., editors, *Text, Speech and Dialogue*, volume 6836 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg.

Greenwood, M. and Stevenson, M. (2006). Improving semi-supervised acquisition of relation extraction patterns. In *Proceedings of Workshop on Information Extraction Beyond The Document, COLING-ACL*, volume 3808, pages 29–35. Springer, Lecture Notes in Artificial Intelligence, Sydney, Australia.

Huttunen, S., Yangarber, R., and Grishman, R. (2002). Diversity of scenarios in information extraction. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas de Gran Canaria, Spain.

Ji, H., Favre, B., Lin, W.-P., Gillick, D., Hakkani-Tur, D., and Grishman, R. (2013). Open-domain multi-document summarization via information extraction: Challenges and prospects. In *Multi-source, Multilingual Information Extraction and Summarization*. Springer.

Khoroshevsky, V. F. (2010). Ontology driven multilingual information extraction and intelligent analytics. In *Web Intelligence and Security: Advances in Data and Text Mining Techniques for Detecting and Preventing Terrorist Activities on the Web*. IOS Press.

Klementiev, A. and Roth, D. (2006). Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Sydney, Australia.

Piskorski, J., Belyaeva, J., and Atkinson, M. (2011). Exploring the usefulness of cross-lingual information fusion for refining real-time news event extraction: A preliminary study. In *Proceedings of RANLP: 8th Conference on Recent Advances in Natural Language Processing*, Hissar, Bulgaria.

Pivovarova, L., Huttunen, S., and Yangarber, R. (2013). Event representation across genre. In *Proceedins of the 1st Workshop on Events: Definition, Detection, Coreference, and Representation*, NAACL HLT, Atlanta, Georgia.

Rortais, A., Belyaeva, J., Gemo, M., van der Goot, E., and Linge, J. P. (2010). Medisys: An early-warning system for the detection of (re-)emerging food- and feed-borne hazards. *Food Research International*, 43(5):1553–1556.

Schumann, A.-K. (2012). Towards the automated enrichment of multilingual terminology databases with knowledge-rich contexts–experiments with russian eurotermbank data. In *CHAT 2012: The Second Workshop on Creation, Harmonization and Application of Terminology Resources*, Madrid, Spain.

Sokirko, A. (2001). *Semantic dictionaries in automatic text analysis, based on DIALING system materials*. PhD thesis, Russian State University for the Humanities, Moscow.

Solovyev, V., Ivanov, V., Gareev, R., Serebryakov, S., and Vassilieva, N. (2012). Methodology for building extraction templates for Russian language in knowledge-based IE systems. Technical Report HPL-2012-211, HP Laboratories.

Tanev, H., Zavarella, V., Linge, J., Kabadjov, M., Piskorski, J., Atkinson, M., and Steinberger, R. (2009). Exploiting machine learning techniques to build an event extraction system for Portuguese and Spanish. *Linguamatica*, 2.

Toldova, S. J., Sokolova, E. G., Astaf'eva, I., Gareyshina, A., Koroleva, A., Privoznov, D., Sidorova, E., Tupikina, L., and Lyashevskaya, O. N. (2012). NLP evaluation 2011–2012: Russian syntactic parsers. In *Proceedings of Dialog Conference*, Moscow, Russia.

Yangarber, R. (2006). Verification of facts across document boundaries. In *Proceedings of the International Workshop on Intelligent Information Access (IIIA-2006)*, Helsinki, Finland.

Yangarber, R., Grishman, R., Tapanainen, P., and Huttunen, S. (2000). Automatic acquisition of domain knowledge for information extraction. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, Saarbrücken, Germany.

Yangarber, R. and Jokipii, L. (2005). Redundancy-based correction of automatically extracted facts. In *Proceedings of HLT-EMNLP: Conference on Empirical Methods in Natural Language Processing*, Vancouver, Canada.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge, MA. ACM Press.

Zamite, J., Silva, F., Couto, F., and Silva, M. (2010). MEDCollector: Multisource epidemic data collector. In Khuri, S., Lhotská, L., and Pisanti, N., editors, *Information Technology in Bio- and Medical Informatics, ITBAM 2010*. Springer Berlin.