# Creation of reusable components and language resources for Named Entity Recognition in Russian

**Borislav Popov*, Angel Kirilov*, Diana Maynard†, Dimitar Manov***

*Ontotext Lab, Sirma AI, Bulgaria
{borislav, angel.kirilov, mitac}@sirma.bg

†Dept of Computer Science, University of Sheffield, Sheffield, UK
{diana}@dcs.shef.ac.uk

**Abstract**

This paper describes the development of the RussIE system in which we experimented with the creation of reusable processing components and language resources for a Russian Information Extraction system. The work was done as part of a multilingual project to adapt existing tools and resources for HLT to new domains and languages. The system was developed within the GATE architecture for language processing, and aims to explore the boundaries of language resource reuse and adaptability across languages and language types, rather than to create a full-scale IE system at the very peak of performance. Nevertheless, the systgem achieves a very creditable 71% F-Measure on news texts, and there is much scope for future improvement of this score.

## 1. Introduction

The RUSSIE system is part of the Multilingual MUSE project aimed at adapting a Named Entity (NE) system for English to new languages. It consists of a rule-based approach based on the architectures of GATE (Cunningham et al., 2002a) and MUSE (Maynard et al., 2002; Maynard et al., 2003). One of the primary aims is the creation of new and reusable tools and resources for different languages with minimum effort.

The IE process implemented in RussIE follows the basic steps undertaken by the MUSE equivalent for English. The approach of reusing the IE resources (both language resources and components) rather than creating them from scratch was possible because of the characteristics of the GATE Platform. It was easy to configure a quick-start basic NE system using the GATE tokenizer, followed by the gazetteer component with some sample lists. Initially we also reused the MUSE pattern-matching grammar rules by suppressing some of them that were obviously irrelevant for Russian. The next step was to collect Russian language resources to boost the performance of the existing NE application. The structure of the gazetteer resources (and associated types) remained intact and provided a framework for the Russian-specific gazetteer entries collected.

The problem with the inflectional nature of the language remained, but we were already able to recognize the main forms of many lexical resources and named entities. At the same time, we were building a morphology resource along with a Russian morphological analyzer. Its aim was to generate morpho-syntactic descriptions (MSD) over known words in the text. These MSDs had to be translated to the POS category feature of the tokens, so we could unleash more of the potential of the MUSE NE grammars. The last step was to build an inflectional gazetteer component to boost the identification of context clues and the recognition of entities despite of their inflection.

## 2. Issues in adaptation to Russian

The main problematic issues in adapting the English system are the highly inflectional nature of Russian and the limited amount of existing resources available for Russian. This raised the following issues, amongst others:

- extension of the part-of-speech (POS) annotation set, since the POS variations relevant for NE in Russian are greater than those used typically for Latin languages (e.g. to handle cases).
- collection/preparation of Russian morphology resources, including the inflectional paradigm and other aspects necessary for POS tagging and further processing.
- the pattern matching engine and rules for the semantic tagger needed to be adjusted to handle language with more agreement involved on a morphological level (e.g. for gender)
- preparation of suitable gazetteer lists, including
    1. handling Cyrillic;
    2. inflectional names;
    3. dual usage of English/Latin names together with the Russian ones.
- further transformation of the pattern-matching grammars with respect to the specifics of Russian language phenomena

## 3. System components

The RussIE system consists of the following processing resources in a pipeline architecture: tokeniser, Russian gazetteer, inflectional gazetteer, English gazetteer, sentence splitter, morphological analyser, part-of-speech tagger, and semantic tagger. The default resources for GATE are detailed in (Cunningham et al., 2002b).

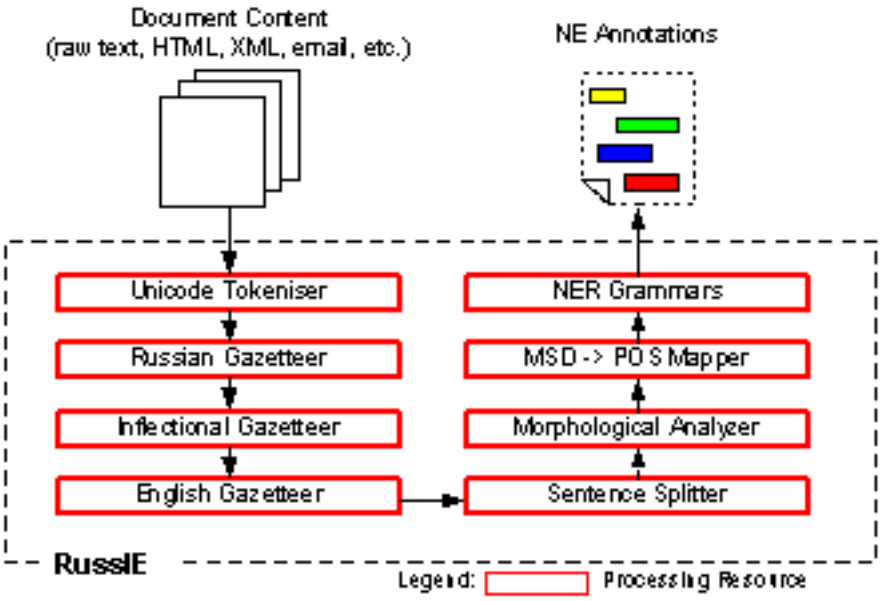



Figure 1: Architecture of the RussIE system

Figure 1 shows a graphical representation of the pipeline. The processing starts with a Unicode Tokenizer, followed by Inflectional, Russian and English gazetteers (the latter engaged in identifying English names in Russian texts). In the next step, the sentence boundaries are identified by the Sentence Splitter. Then morphological analysis is performed. This step identifies the morpho-syntactic types of the words in the text. Next, the MSDs identified are translated into the POS categories used in MUSE. A pattern-matching semantic grammar component with slightly modified and filtered MUSE rules finalizes the recognition of named entities. From the processing resources involved, the Sentence Splitter and the pattern-matching grammars have been directly reused in RussIE, while the other components are either completely new (Morphological Analyzer, Inflectional Gazetteer), either modified (Gazetteer, Tokenizer). The following sections detail the processing resources in more depth.

### 3.1. Tokeniser

The tokeniser is similar to the default English tokeniser in GATE, except that it has been adapted to deal with Cyrillic words as well as Latin ones. It appeared that, with such small changes, the tokenizer performed very well and was able to identify the tokens. It annotates each identified token with information about its case, whether it is a number, etc.

### 3.2. Gazetteer

The gazetteer component looks for items from the gazetteer lists and assigns annotations with type Lookup to the items recognised in the text. There are two gazetteers in the application - one for the Russian lists and one for the Latin English lists. There is a set of lists with items in Cyrillic and another set with items in English (a subset of the MUSE gazetteer lists). The reason for having the English lists is that there are some NEs in the Russian news written in English. Using the English gazetteer we can separate the English lists from the Russian ones without missing any English NEs in the text. The following gazetteer lists have been created for Russian:

- 21,500 large Russian companies
- 113 Federal Government Organizations (Ministries, Departments, etc.)
- 88 largest Russian companies (publicly trading)
- 67 Government Persons (e.g. The President, Ministers, etc)
- 99 famous Persons (e.g. Persons of the year 2001 and 2002, including non-Russians)
- 216 largest Russian cities
- 236 names of Locations (continents and countries in Russian)
- 185 female first names
- 326 male first names
- top 100 surnames (according to the phone directory) – currently unused
- top 107,603 surnames (according to the Moscow phone directory) – currently unused
- month names
- demo lists (equivalents of MUSE lists with just one or two items translated)

#### 3.2.1. Enhancing the gazetteer lists

Once a basic set of gazetteer lists for Russian had been created, along similar lines to the default set for English, our goal was to enrich both their content and the structure. The enrichment of the lists was partly manual and partly automatic (e.g. playing with the capitalization of the already collected lists of companies and people). The changes to the structure involved additions of new gazetteer lists (e.g. for spurious persons and organizations, for general locations such as islands, and for generally famous people). Since part of the list-related work was performed before the inflection gazetteer enhancement, there are some inflection forms that are explicitly listed. This was considered neither necessary nor manageable for large scale lists, so the main focus afterwards was on using a gazetteer component, which, given the main word-forms, could manage the inflection derivates.

### 3.3. Inflectional gazetteer

The inflectional gazetteer is a component populated from an XML input file containing main forms of entities and all their inflectional derivates. It yields more correct identification of the inflections of the entries than the Russian gazetteer, but lacks the scalability of its heuristic-based inflection handling.

#### 3.3.1. Development of a more efficient inflectional gazetteer component

By "more efficient" we mean "easily maintainable"- the xml-input-based inflection gazetteer is hungry for all the inflection word-forms of a word, and we would like to have a gazetteer that is based only on main-forms that are easier to acquire, and to guess automatically the inflection derivates. This has been achieved by modification of the RussIE Gazetteer: now it stems the main word-forms and is ready to deal with most of the inflectional derivates. This does not cover all the possible changes of consonants. However, the implementation allows (through parameters) management of the stemming algorithm, e.g. one can specify which are the expected suffixes, or concentrate only on those suffixes consisting of vowels.

### 3.4. Morphological analyser

The morphological analyser deals with Russian inflectional morphology, through the creation of a full-form lexicon containing approximately 54,000 lexemes. The word forms are grouped in paradigms, and each word-form is connected with associated MSD tags. The lexicon is represented as Prolog clauses, and covers more than half a million word forms. The MSD mapper component then maps the MSD annotations onto the POS categories used later in the semantic tagger.

### 3.5. Sentence splitter

The sentence splitter is taken directly from the default splitter for English, since it is generic across languages that use appropriate punctuation.

### 3.6. POS tagger

The Russian POS tagger is based on a hash gazetteer and uses Russian morphology to create MSD (Morpho-Syntactic Description) annotations. The morphology consists of 50k+ lemmas with more than half a million word-forms. It generates composite morpho-syntactic types (e.g. Nmisn, Nmisg, etc.) on MSD annotations (where the first letter of the type represents the POS information.

| Annotation Type | Precision | Recall | F Measure |
|---|---|---|---|
| Date | 77% | 71.7% | 74.3% |
| Person | 70.5% | 53.9% | 61.1% |
| Organization | 72.5% | 59.8% | 65.5% |
| Location | 91.2% | 68.7% | 78.4% |
| Percent | 87.5% | 87.5% | 87.5% |
| Money | 80.8% | 40.4% | 60.6% |
| Total | 79.9% | 63.7% | 70.9% |

Table 1: Evaluation of RUSSIE

### 3.7. Semantic tagger

The semantic tagger is based on JAPE (Java Annotations Pattern Engine) grammars(Cunningham et al., 2000), as for MUSE, and uses many of the same rules. The default ruleset can be divided into those which are generic across languages (for a given language pair) and those which are language-specific. Some other small changes to the processing are necessary for Russian, for example, the grammar which finds numbers is modified from English in order to use a Cyrillic constant which checks the usage of UTF-8 Cyrillic inside the rules. Very few rules actually needed to be modified from English for Russian, because they are mostly based on POS tags and gazetteer lookup, and given the correct Russian forms for these, the rules function appropriately. Some research into the specifics of Russian were of course necessary to deal with phenomena such as different ordering of co-occurring named entities. For example, English has the pattern "Organisation - Jobtitle - Person" (e.g. UN secretary Kofi Annan), whereas Russian has the pattern "Jobtitle - Organisation - Person". Such patterns are used as context to find new entities.

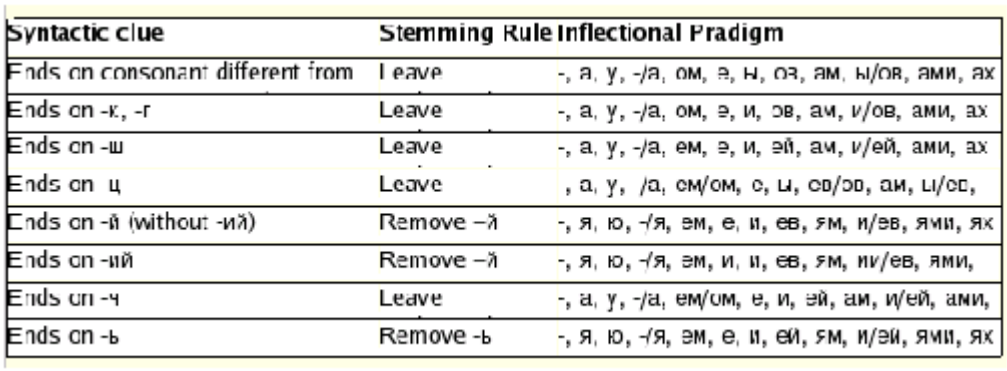
## 4. Evaluation

The system has been evaluated on a corpus of Russian news texts containing 92 articles, consisting of different types of news such as political, local, sport and financial news. These were manually annotated with Person, Location, Organisation, Date, Money and Percent annotations, following the guidelines for MUSE (which are based on MUC guidelines, with some small differences). The system was evaluated according to Precision, Recall and Fmeasure – the results are shown in Table 1.

## 5. Further Work

From Table 1 it is obvious that there is a lot more that could be done to achieve e.g.. F measures above 85%. Below we propose some suggested enhancements to the current RussIE NE application and resources, as work is still ongoing.

1. Usage of the full surnames list. There is a list of over 100,000 surnames extracted from a phone book, which



| Syntactic clue | Stemming Rule | Inflectional Pradigm |
|---|---|---|
| Ends on consonant different from | Leave | -, а, у, -/а, ом, е, ы, ов, ам, ы/ов, ами, ах |
| Ends on -к, -г | Leave | -, а, у, -/а, ом, е, и, ов, ам, и/ов, ами, ах |
| Ends on -ш | Leave | -, а, у, -/а, ем, е, и, ей, ам, и/ей, ами, ах |
| Ends on ц | Leave | , а, у, /а, ом/ом, о, ы, ов/ов, ам, ы/ос, |
| Ends on -й (without -ий) | Remove –й | -, я, ю, -/я, ем, е, и, ев, ям, и/ев, ями, як |
| Ends on -ий | Remove –й | -, я, ю, -/я, ем, и, и, ев, ям, ии/ев, ями, |
| Ends on -ч | Leave | -, а, у, -/а, ем/ом, е, и, ей, ам, и/ей, ами, |
| Ends on -ь | Remove -ь | -, я, ю, -/я, ем, е, и, ей, ям, и/ей, ями, як |

Figure 2: Syntactic Clue - Stemming Rule - Inflectional Paradigm

is not currently used. Its inclusion would increase Recall, but it needs to be carefully handled and further filtered to avoid loss of Precision.

2. Rule-based inflection-focused enrichment of the gazetteer on load/execute time. So far, we have already extracted rules that classify nouns by their syntactic features (e.g. suffixes), and linked them to an inflectional paradigm that gives the possible suffixes for the noun category (displayed in Figure 2). At load time, these rules could be used to enrich the gazetteer in-memory representation by inferring the inflectional forms. Another approach is to use the rules at execute time along with the MSD/POS tag identifying the word as a noun. Then one could try to deduce the main wordform, though this approach could lead to ambiguity.

3. Unleash Unknowns. The NER process would possibly benefit from the generation of Unknown annotations for all Upper annotations (words beginning with a capital letter) that have not been covered by a final NE annotation. At the final phase we have added a grammar that filters the Uppers at the beginning of the sentences, which will make the Unknown generation safer. Given these Unknowns, it is expected (based on experience from English) that the Orthomatcher would be able to match them to an already-identified NE and thus determine their type.

4. Inflectional OrthoMatcher. An orthomatcher which handles inflectional information would be beneficial not only for the correct matching, but also for improving the accuracy of the recognition process, as explained above. Heuristics could be used for initial experiments so to avoid heavy linguistic analysis (e.g. stemming as in the Russie Gazetteer).

5. Use of morphological analyzer in the gazetteer. The gazetteer could query the morphology analyzer on load time and thus enrich its in-memory model with the inflectional forms of its entries.

## 6. Conclusion

In summary, we have created a set of new and reusable tools and resources for Russian Named Entity recognition, in the course of migrating an NE system from one language group to another (highly inflectional) one. The system was built within 6 person-months and achieves very creditable results, largely due to the architecture and design of GATE and MUSE, which is very conducive to the adaptation of systems to new languages. Such resources provide an excellent starting point for other work on Russian.

## 7. References


Cunningham, H., D. Maynard, K. Bontcheva, and V. Tablan, 2002a. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*.

Cunningham, H., D. Maynard, K. Bontcheva, V. Tablan, and C. Ursu, 2002b. *The GATE User Guide*. `http://gate.ac.uk/`.

Cunningham, H., D. Maynard, and V. Tablan, 2000. JAPE: a Java Annotation Patterns Engine (Second Edition). Research Memorandum CS–00–10, Department of Computer Science, University of Sheffield.

Maynard, D., V. Tablan, K. Bontcheva, H. Cunningham, and Y. Wilks, 2003. Muse: a multi-source entity recognition system. *Submitted to Computers and the Humanities*.

Maynard, D., V. Tablan, H. Cunningham, C. Ursu, H. Saggion, K. Bontcheva, and Y. Wilks, 2002. Architectural Elements of Language Engineering Robustness. *Journal of Natural Language Engineering – Special Issue on Robust Methods in Analysis of Natural Language Data*, 8(2/3):257–274.