RU-EVAL-2012: Evaluating dependency parsers for Russian

Anastasia Gareyshina[1], Maxim Ionov[1], Olga Lyashevskaya[2], Dmitry Privoznov[1], Elena Sokolova[3], Svetlana Toldova[1,3]
(1) MOSCOW STATE UNIVERSITY, Philological Faculty, Dept. of Theoretical and Applied Linguistics, Leninskie gory, GSP-1, 119991 Moscow, Russia
(2) NATIONAL RESEARCH UNIVERSITY HIGHER SCHOOL OF ECONOMICS, Faculty of Philology, Myasnitskaya 20, 101000 Moscow, Russia
(3) RUSSIAN STATE UNIVERSITY FOR THE HUMANITIES, Institute of Linguistics, Miusskaya pl. 6, GSP-3, 125993 Moscow, Russia
a.r.gare@gmail.com, max.ionov@gmail.com, olesar@gmail.com, dprivoznov@gmail.com, minegot@rambler.ru, toldova@yandex.ru

ABSTRACT

The paper reports on the recent forum RU-EVAL – a new initiative for evaluation of Russian NLP resources, methods and toolkits. It started in 2010 with evaluation of morphological parsers, and the second event RU-EVAL 2012 (2011-2012) focused on syntactic parsing. Eight participating IT companies and academic institutions submitted their results for corpus parsing. We discuss the results of this evaluation and describe the so-called "soft" evaluation principles that allowed us to compare output dependency trees, which varied greatly depending on theoretical approaches, parsing methods, tag sets, and dependency orientations principles, adopted by the participants.

TITLE AND ABSTRACT IN RUSSIAN

RU-EVAL-2012: Оценка парсеров грамматики зависимостей для русского языка

RU-EVAL – это форум по оценке русскоязычных ресурсов, методов и инструментов автоматической обработки текста. Первый этап форума состоялся в 2010 году и был посвящен оценке морфологических парсеров (Lyashevskaya et al. 2010), второй цикл (2011-2012) связан с оценкой синтаксического анализа текста (Toldova et al. 2012). На синтаксическом форуме результаты разметки тестового корпуса в формате синтаксиса зависимостей прислали 8 участников из коммерческих компаний и академических учреждений. В статье описываются принципы «мягкой» оценки, позволившие сравнивать ответы, которые весьма значительно различались как теоретическими подходами и методами парсинга, так и по конкретному составу тегов и направлению зависимостей. Обсуждаются результаты, сложные для оценки случаи, а также некоторые проблемные точки в работе русских синтаксических парсеров, которые выявила экспертиза результатов.

KEYWORDS : Parsing evaluation, dependency grammar, Russian, Russian treebank
KEYWORDS IN RUSSIAN : Оценка синтаксических парсеров, грамматика зависимостей, русский язык, русский трибанк

Proceedings of COLING 2012: Posters, pages 349–360,
COLING 2012, Mumbai, December 2012.

1 RU-EVAL-2012: Оценка парсеров грамматики зависимостей для русского языка

Статья посвящена первому опыту проведения в России форума по оценке методов автоматического синтаксического анализа текстов на русском языке. В задачи форума входило оценить общее положение дел в этой области: какие парсеры русского языка существуют, какие теоретические подходы представлены, каковы средние и максимальные показатели существующих разработок. В статье излагаются основные принципы и проблемы подготовки форума: создание тестовой коллекции и Золотого стандарта (ЗС), проработка заданий и мер оценки, подводятся итоги форума, анализируются результаты сравнения работы синтаксических парсеров, представленных на форуме. Тестовой коллекцией служил корпус из отдельных предложений и последовательностей предложений из художественной и научно-публицистической литературы, а также новостных сообщений общим объемом 1 млн. токенов.

В соревновании участвовали системы: SyntAutom, DictaScope Syntax, SemSin, ЭТАП-3, синтактико-семантический парсер SemanticAnalyzer Group, AotSoft, ABBYY Compreno (DIALOGUE 2012). Один участник, Russian Malt (С.Шаров, Лидс, Великобритания), участвовал вне конкурса, в то время как участник Link Grammar Parser (С. Протасов, Москва) не смог конвертировать результаты в адекватный формат грамматики зависимостей и отказался от участия в соревновании.

Предварительная оценка известных открытых систем синтаксического анализа показала, что большинство парсеров для русского языка базируются на грамматике зависимостей. Анализ пробного разбора 100 предложений, представленного разработчиками – потенциальными участниками форума 2011–2012, показал, что в России системы синтаксического анализа развивались автономно, без использования какого бы то ни было корпуса в качестве эталона. Поскольку расхождения между системами по составу тегов и по принципам установления связей оказались значительными, было принято решение о том, что на данном этапе оцениваться должно только правильное определение системами синтаксически связанных пар словоформ и установление «главного» элемента в паре. Оценивалась правильность приписывания вершины зависимой словоформе (однако, правильность разметки всего предложения не оценивалась).

Результаты, полученные от участников, сравнивались на корпусе ЗС: 800 предложений, случайным образом выбранных из тестовой коллекции и размеченных вручную. Принципы и средства синтаксической разметки, использованные при аннотировании ЗС были сформулированы в (Sokolova 2011; ср. также Hovy and Lavid 2010). Разметка производилась параллельно тремя аннотаторами. Была предпринята попытка свести результаты анализа к общему формату автоматически, однако, большáя вариативность в сложных случаях не позволила обойтись без ручной проверки.

Использовалось так называемое «мягкое» оценивание: допускались отклонения от ЗС в ответах систем, обусловленные спецификой теоретических или производственных решений, если такие решения проводятся последовательно на всем тестовом корпусе. Для классификации расхождений с ЗС использовалась шкала оценок, включающая как «допустимые» расхождения (расхождения объясняются расхождением в принципиальных решениях системы и ЗС), так и семантически допустимую синтаксическую омонимию.

Sentence 1819

GS (Золотой стандарт)

id	token	type	head a
1	**Каких** ← *результатов*	amod	3
2	**именно** ← *Каких*	spec	1
3	**результатов** ← *ждать*	obj	5
4	**можно**	pred	
5	**ждать** ← *можно*	comp	4
6	**от** ← *ждать*	comp	5
7	**совместных** ← *усилий*	amod	8
8	**усилий** ← *от*	pcomp	6
9	**членов** ← *усилий*	mod	8
10	**группы** ← *членов*	mod	9
11	.		

Вариант участника

id	token	type	head	mark
1	**Каких** ← *результатов*	Какой	3	0
2	**именно** ← *результатов*	Частица	3	4
3	**результатов** ← *ждать*	Род	5	0
4	**можно**			
5	**ждать** ← *можно*	Сост_сказ	4	0
6	**от** ← *ждать*	Откуда,Ото	5	0
7	**совместных** ← *усилий*	Какой	8	0
8	**усилий** ← *от*	Род	6	0
9	**членов** ← *группы*	Род	10	1
10	**группы** ← *ждать*	Вин	5	1
11	.			

группа (но,мн,С,жр,вн)

Рис.1. Сопоставление разметки ЗС и ответа системы с градуальной оценкой (mark).

Результаты оценивались с использованием стандартных мер: точность (P), полнота (R) и F-мера. Точность оценивалась как отношение количества допустимых ответов системы. Результаты Unlabeled Attachment Score составили: Pmax – 0.952, F-мера – 0.967, Pmin – 0.789, F-мера = 0.872, средний результат по всем системам: Pav – 0.88.

Наилучшие результаты достигнуты системами, «обогащенными» семантическими и другими экспертными лингвистическими знаниями. Эти системы создавались большими коллективами высокопрофессиональных лингвистов в течение длительного периода времени. Третья по точности – система Russian Malt, основанная на машинном обучении (MALT). Обучение происходило на трибанке SynTagRus (http://ruscorpora.ru), который, таким образом, обеспечивает машинное обучение с высокими результатами по точности и полноте. Как свидетельствуют остальные результаты, менее дорогие и ресурсозатратные решения также имеют неплохую точность и полноту.

В ходе подготовки и проведения форума были выработаны принципы и методы оценки работы зависимостных парсеров, основанных на разных теоретических принципах. Также были созданы важные ресурсы: (а) ЗС объемом 800 предложений, размеченных вручную; (б) Паралелльный Трибанк, в котором представлена параллельная аннотация тестового корпуса (1 млн. токенов) четырьмя системами с визуализацией и возможностью поиска (оба ресурса представлены в свободном доступе на сайте http://testsynt.soiza.com).

Опыт проведения форума показал, что автоматический синтаксический анализ для языков типа русского имеет целый ряд особенностей, связанных с развитой морфологией и богатой омонимией на уровне форм, а также со свободным порядком слов. Эти обстоятельства существенным образом влияют не только на специфику разработки, но и на специфику проведения сравнения между системами. На сегодняшний день наиболее распространённые и успешные методы преодоления данных трудностей и методы борьбы с синтаксической омонимией – это учет ограничений на лексическую сочетаемость и усиление статистическими процедурами лингвистических компонентов, основанных на правилах.

2 Introduction

The NLP Evaluation forum RU–EVAL started in 2010 as a new initiative aimed at independent evaluation of NLP systems for Russian. The second evaluation campaign (2011–2012) is focused on syntactic parsing. It is open both to academic institutions and industrial companies, and its general objective is to assess the current state-of-the-art in the field and promote the development of syntactic technologies. The forum has also an educational component: the expert group includes students who plan to work in the field of computational linguistics. The forum provides a good opportunity for them to have a hands-on experience of how the NLP tools work, and to see their strong and weak sides.

The first NLP Evaluation forum focused on morphological taggers (see http://ru-eval.ru, Lyashevskaya et al. 2010), bringing together 15 participants from Moscow, Saint-Petersburg, Yekaterinburg, Ukraine, Belarus and UK. In 2011-2012, syntactic parsing technologies were evaluated (Toldova et al. 2012). It was the first time such evaluation was held in Russia. This task turned out to be much more complicated than morphological taggers evaluation.

The main features for Russian parsers are the following: they are mostly based on the dependency trees representation, they are rule-based, and there is no uniform annotation scheme for such systems. The controversial issues we faced while working out the evaluation routine for Russian parsers could be explained first of all by some peculiarities of Slavic languages: Russian is a morphologically rich language with a rather free word order. In fact, word order is mostly triggered by information flow (e.g. topic-focus hierarchy, prominence of participants in a profiled frame, emphasis etc.), though there exist some 'neutral word order' patterns, grounded in certain discourse registers (question, beginning of narrative, etc.) and individual morphosyntactic structures (such as Dative construction). Since frame relations are mainly encoded by grammatical case and prepositions, the role of word order in the recognition of semantic-syntactic relations shrinks dramatically. So, it is not surprising that a wide variety of formalisms and principles of syntactic structure representation are used for parsing Russian texts. There are considerable differences in parsing outputs, depending mainly on the end task of the NLP system.

Since the majority of potential participants develop the dependency parsers, only dependency trees were evaluated. The overall procedure was organized as follows: participants received a tokenized text collection, processed it in their systems and sent the result back in a unified format. Precision and recall was assessed by comparing the result against the manually tagged Gold Standard (GS). The expertise of the task output was performed semi-automatically with subsequent double manual check.

Section 3 presents possible approaches to evaluating Russian syntactic parsers and critical points that should be taken into consideration. Section 4 reports on track design, the board of participants, datasets for the training, task and test collections, evaluation measures and results. In Section 5, we discuss most systematic cases of variation in the output as well as some crucial points that still pose a problem for many Russian parsers.

3 Approaches to evaluating Russian syntactic parsing

A preliminary study on the current state of syntactic parsing for Russian has shown that most of the systems use the dependency grammar representation. Given this, dependency trees were

chosen as an output format, and those participants who used mixed dependency-constituency representation or other formalisms, were asked to convert their results.

The general practice suggests that the organizers provide a syntactic treebank ready to use as a GS; this provides also a standard tag set, namely, names and types of relations. Moreover, most developers use these corpora for building their systems, especially if the system is ML-aided. For example, in EVALITA, Turin University Treebank (TUT) is used, that is tagged with respect to both formalisms: dependency grammar and phrase structure grammar. Using the sentences from such treebanks as a test corpus also simplifies the procedure of automatic assessment.

During the organization we relied upon similar evaluation events (EVALITA and other mentioned in Section 2). However, we could not simply use the main principles of EVALITA per se for the reasons mentioned below. We did not take into account morphological and syntactical tags (despite the fact that we included them into the output to make the manual evaluation easier).

For the dependency tree parsing tracks, participants got the text corpus split into sentences and tokens. The task was to mark the syntactic head and the type of syntactic relation for each word.

The analysis of the 100-sentence test sample, parsed by potential participants of the forum 2011–2012, showed that in Russia, syntactic parsing systems developed autonomously, without using any corpus as a GS. As a result, differences between the systems in both tag sets and principles of tagging were so significant that on several issues there could not be proposed any single solution for data output format. Therefore, at this point, we decided to assess only syntactic pairs detection and detection of their syntactic heads. In addition, we decided that theoretically motivated divergences should not be evaluated as errors.

The main assumption of the expertise was the following: there is no single 'correct' answer to complicated questions, and there is no 'correct' parsing algorithm. We tried to mark as wrong only those parses that were motivated neither by theoretical nor by practical decisions. In many cases, the solution to a complicated syntax problem depends on the end goal of the system. There were also some problematic cases which did not have a single solution. After a comparison of results, produced by different parsers, the list of problematic cases for syntactic analysis and methods for their processing were specified.

4 Participants, data sets and results

4.1 Participants

Eleven NLP groups from Moscow, St. Petersburg, Nizhniy Novgorod (Russia), Donetsk (Ukraine), and Leeds (UK) expressed their interest in participating in both tracks. These were systems that use dependency parsing, phrase structure parsing, link grammar and mixed approaches. The answers were submitted by eight groups: SyntAutom (A.Antonova and A.Misyurev, Moscow), DictaScope Syntax (Dictum, Nizhniy Novgorod), SemSin (K.Boyarsky, E.Kanevsky, St.Petersburg), ETAP-3 (Kharkevich Institute for Information Transmission Problems RAS, Moscow), syntactical-semantic analyzer from the SemanticAnalyzer Group (D.Kan, St.Petersburg), AotSoft (V.Vasilyev, Moscow), Compreno (ABBYY, Moscow), Russian Malt (S.Sharoff, Leeds; participated out of competition). One of the participants (namely, Link Grammar Parser (S.Protasov, Moscow)) had not succeeded in converting results to output format, thus there were seven participants involved in the final assessment.

4.2 Test collections and tasks

Evaluation corpus consisted of untagged texts of different types. Corpus for the main track consisted of fiction, news, non-fiction and texts from social networks (5%). There were both separate sentences (0.2 MW from the open collection of the Russian National Corpus) and text fragments. Corpus for news track consisted of text fragments from the ROMIP news collection. These were sequences of three sentences picked randomly. All sentences were tokenized and indexed.

Participants were to markup a syntactic head for each token. Correctness of parsing the whole sentence was irrelevant, only correctness of choosing the head was evaluated. Assessment was conducted on a GS subcorpus which included about 800 randomly selected sentences (500 for the main collection and 300 for the news collection) that had been manually tagged (see section 3.5).

4.3 Input and output format

Input data were in two different formats: plain-text without any markup, and XML with numeration and detailed tokenization. Tokenization and numeration allowed us to simplify the assessment procedure, making it semi-automatic. Plain-text was provided to the participants who take plain-text as the input.

Output format was also specified. Sentence and token numeration should match numeration in the input file, for each token there should be: a number of syntactic head token, relation type and, optionally, morphological tags (provided for experts so that they could analyze reasons of mismatches with GS).

4.4 Gold Standard

Before the assessment, the GS was tagged manually using the tagging tool created by Maxim Ionov. Each sentence was independently tagged by two experts, then divergences were discussed, if any, and the common decision was made. Then the result was checked by the third expert. Such procedure allowed us to achieve three aims. First, it helped to minimize the fraction of arguably tagged tokens. Second, organizers wanted to avoid 'overfitting': getting the experts used to common error of the specific system and omitting errors by not noticing them. Third, tagging was supposed to give the experts the basic knowledge about difficult cases and to help them form criteria for evaluating mismatches.

The group of experts developed the tag set and principles of manual annotation based on (Sokolova 2011; see also Hovy and Lavid 2010). Since uniformity of tagging performed by several people was the main concern, the annotators were asked to choose, among other possible decisions, the most natural one which would correspond to the most popular understanding of the sentence in a possible context. The simpler and clearer decisions are, the more inter-annotator agreement score they provide.

4.5 Evaluation measures

The common evaluation strategy is to compare the output of the parsers to the GS test set (cf. CONLL and EVALITA, Buchholz and Marsi 2006, Nivre et al. 2007, Bosco and Mazzei 2012). The test sets usually are based on a treebank used for the development of the parsers. As it was mentioned above, there is no comparable generally accepted treebank for Russian. Moreover,

there is a great variation in labelling syntactic relations and even in the head-modifier relations through parsers. For these reasons the only Unlabeled Attachment Score measure was taken into account. In the dependency parsing each token in the sentence is assigned the only one Head ID. Thus, the precision could be measured as the percentage of tokens with correct or "admissible" heads. As noted above, we considered certain kinds of mismatches between the GS variant and the system response as acceptable.

The assessment assumed comparing the ID number on the tagged head for each token with the corresponding number in the GS. The match was automatically marked as 0. Mismatches (along with matches, however) were given to the experts for further examination. They could mark a mismatch as:

1 — system error
2 — GS error
3 — acceptable mismatch (theoretical difference between the system and the GS)
4 — acceptable mismatch (in case of homonymy)
5 — the response matches the GS, but they are both wrong
6 — syntactic head is not specified for the token, but it should be specified
7 — syntactic head is not specified for the token, and could be omitted
8 — uncertain
9 — other (cases that do not fall into categories 1–8).

There were a significant number of mismatches in choices of syntactic relation directions among parsers. These were not mistakes but decisions made during the systems' development, so they could not be qualified as errors. For the purpose of simplification of assessment, the participants agreed to unify relation directions in some cases, such as: (1) preposition – noun; (2) auxiliary verb – lexical verb; (3) relations in coordinating constructions.

However, the other dependencies had to be consistent with the decisions concerning such relation directions. For example, if auxiliary verbs were taken to be heads, then subjects had to be dependents of auxiliary verbs, whereas if main verbs were considered heads, then subjects had to be dependents of main verbs; in the case of coordination, it was the phrase heads established by a system that had to be conjoined, e.g. if noun (noun phrase) was considered a head of a prepositional phrase, then only nouns (noun phrases) could be conjoined, for prepositional phrases to form a coordinated structure. We did not penalize such decisions, should they be not unified, – again, provided that they (and the "outer" dependencies) were consistent throughout the output. When relation directions were unified and converted to the GS format, but there were still "old" mismatches with "outer" dependencies (i.e. relation directions were updated so as to fit the GS format, but their dependencies were not), such cases were treated as artifacts – conversion errors.

The number of such cases was significant, so they required further detailed assessment. After the developers got access to their intermediate scores, they sent some comments, which proved to be of great help in improving the assessment design. But even then we could not fully eliminate 'false positives', where penalty was assigned by mistake (see Section 5 for the discussion of some difficulties that we had to face).

4.6 Results

The results of the main track are shown in table 2. According to the "soft" evaluation measures the best result has been achieved by ABBYY Compreno (precision 0,952, F-measure 0,967). The results of the ETAP–3 system are slightly lower. The average precision was 88,8.

Mask name	P	R	F1	System Name
Trieste	0,952	0,983	0,967	Compreno
Marceille	0,933	0,981	0,956	ETAP–3
Barcelona	0,895	0,980	0,935	SyntAutom
Toulon	0,889	0,947	0,917	SemSyn
Brega	0,863	0,980	0,917	Dictum
Nice	0,856	0,860	0,858	SemanticAnalyzer Group
Napoli	0,789	0,975	0,872	AotSoft

TABLE 1 – Dependency parsing, main track evaluation.

The best results have been achieved by two systems that developed their parsers on the basis of manual rule-based approach, enriched with a thoroughly elaborated semantic component by teams of linguist experts. However, low-time-consuming systems, such as SyntAutom, have also proved to be reliable. One of the systems, Russian Malt, was based on the machine-learning technology. It used the SynTagRus Treebank (http://ruscorpora.ru) as a learning corpus and achieved the third-highest results (the results are not shown in the chart since the system participated out of competition). In the next section we will discuss in detail some questions touched upon during RU-EVAL 2012 and the difficulties that we had to face.

5 Discussion

5.1 Variation in parsing

As it has been mentioned above, the systems vary significantly with respect to tag sets and dependency assignment rules. It is only in the simplest cases (e.g. attributes that agree with nouns) that there is hardly any variation at all. More often, the systems process a particular construction in several different ways. For instance, while in some parsers simple clauses can be connected with each other by means of establishing a syntactic relation between their verbal heads, other analyzers parse a complex sentence by linking its simple clauses with a subordinating conjunction.

What is more important, there can be cases where there is no uniform theoretical decision within dependency formalism. Sometimes it generally remains unclear which one of the units of the syntactic relation is the head or dependent (Iomdin 1990, Gladkij 1973). Such ambiguities emerge when different criterions on head-dependent distinction yield different results (Testelets 2001), or not a single criterion is applicable.

(1)	(*pol'zovat'sya*)	velikimi(Adj1)	i	*udivitel'nymi*(Adj2)	blagami(N)
	(use)	wonderful	and	great	amenities

A. N → i; i → Adj1; i → Adj2;

B. N → Adj1; Adj1 → i → Adj2;

C. N → Adj1; Adj1 → i; Adj1 → Adj2; D. N → Adj1; Adj1 → Adj2; Adj2 → i.

The coordinated structure in (1) can be parsed in several ways: a conjunction can be treated as a head itself (A); the coordinated group can form a linear dependent-head chain (B); it can be treated as a dependent on any element in the coordinated group (C and D). For this example, no parsing result can be argued to be a system error, as long as the whole coordinated structure is successfully parsed in a consistent way.

Further steps should be taken to reduce variation in dependency relations labels so that tag assignment evaluation could be performed. There is a considerable variation in classifications of dependency relations: some of them are based on morphological properties of the head or of the dependent while others rely upon general syntactic functions of a given word form. For example, one system has the tag 'card' (cardinal) for encoding the numeral-noun dependency; in other systems, quantifier is just an instance of a noun modifier. Merging different classifications is still a goal to be achieved.

5.2 Qualitative output analysis: some problem cases

After we had analyzed all systems' answers, we came to the conclusion that there were no 'universal problem cases' – cases that cannot be properly parsed with all systems, and that conclusion is a pleasant fact indeed. A special case example here would be prepositional dependents (it can have verb as a head irrespective of weather it is an argument or an adjunct or a noun as a dependent in an NP). If there are several head candidates in a sentence, the parsers choose either the first noun preceding prepositional dependent, or verbal head, or the closest finite verb in a tree. Yet many thus generated variants are not semantically admissible, compare acceptable examples (2A-C), (3A-B) to inacceptable ones (2D), (3C):

(2)	Google	prodolzhaet	*ukrepljat'*	pozicii	na	rynke
	GoogleNOM.SG	continues	strengthen.INF	position.ACC.PL	on	market.LOC.SG
	prilozhenij	dlja	sovmestnoj	raboty.		
	application.GEN.PL	for	collaborative.GEN.SG	work.GEN.SG		

"Google continues strengthening positions on the market of applications for collaborative work".

A. Ok pozicii 'position.ACC.PL' → na rynke 'on market.LOC.SG'
B. Ok ukrepljat' 'strengthen.INF' → na rynke 'on market.LOC.SG'
C. Ok prilozhenij 'application.GEN.PL' → dlja sovmestnoj raboty 'for collaborative.GEN.SG work.GEN.SG'
D. * *ukrepljat'* → dlja sovmestnoj raboty 'for collaborative.GEN.SG work.GEN.SG'

(3)	chto	mozhet	*dobit'sja*	svojej	celi	*lish'*	pri
	that	can	achieve.INF	REFL.POSS.GEN.SG	goal.GEN.SG	only	at
	odnom	uslovii...					
	one.LOC.SG	condition.LOC.SG					

'…that [he] can achieve his goal only on a single condition…'

A. Ok *dobit'sj*a 'achieve.INF' → pri uslovii 'on condition.LOC.SG'
B. Ok mozhet 'can' → pri uslovii 'on condition.LOC.SG'
C. * celi 'goal.GEN.SG' → pri uslovii 'on condition.LOC.SG'

There are certainly much more errors in complex sentences. Among the most typical problem cases is establishing the simple (dependent) clause head in a clause that precedes the dependent one. Similarly, nominal and copular heads may not be regarded as possible candidates for being a clause head. Finally, quite often are the cases when a distant dependent is connected to a hypothetical head across the clause boundary and the cases when heads remain undefined for words absent from the system dictionary (words and abbreviations like "OC", "Intel" etc.).

Conclusion

The RU-EVAL 2012 has brought together a considerable number of IT companies and academic groups that work on Russian syntax parsing, and made it possible to assess the state-of-the-art in the field (so far, mostly in Russia). The forum has shown that the majority of parsers for Russian are based on dependency formalism. They are rule-based.

The event has the following practical outcomes:

- A manually tagged standard set, consisting of 800 sentences, is made available through testsynt.soiza.com; the guidelines for tagging according to GS principles have been compiled and tested.
- Variations in theoretical and practical decisions between existing parsers have been registered.
- The treebank with parallel annotation (1 mln. tokens, annotated by four participants) is made available at http://testsynt.soiza.com; it is presumed that the treebank can enable reliable machine learning for parsing.

The RU-EVAL 2012 has shown that there are three basic approaches to parsing for Russian:

1. systems, manually enriched with expert linguistic knowledge (Compreno, ETAP-3);
2. automata-based systems (SyntAutom);
3. machine-learning systems.

The manually enriched with rules systems have shown the best results. However, low-time-consuming systems, such as SyntAutom, have also proved to be reliable. The results have also demonstrated that there exists at least one Russian treebank that enables reliable machine learning for parsing Russian (the Russian Malt system). Although Russian is a free-word order language with a rich morphology, the quality of syntactic parsing is quite high. The majority of Russian parsers override the difficulties due to lack of word order constraints by developing semantic components and integrating statistical approaches into the rule-based systems. The best result has been demonstrated by the system that heavily depended on semantic components and took into consideration the semantic constraints on lexeme co-occurrence.

Acknowledgments

The work was partly supported by Corpus Linguistics Program of the Presidium of Russian Academy of Sciences. We would like to thank Irina Astaf'eva, Anastasia Bonch-Osmolovskaya, Julia Grishina, Anna Koroleva, Pavel Koval', Anna Lityagina, Natalia Men'shikova, Alexandra Semenovskaya, Eugenia Sidorova, Lyubov' Tupikina who took part in all stages of evaluation routine as experts and annotators. We are also most grateful to the participants of the forum.

References

Bosco, C. and Mazzei, A. (2012). The EVALITA 2011 parsing task: the dependency track. In *EVALITA'11 Working Notes*, Roma.

Buchholz, S., Marsi E. (2006). CoNLL-X Shared Task on Multilingual Dependency Parsing. In Proceedings of the CoNLL-X. New York, NY, pages 149-164.

DIALOGUE (2012). Computational linguistics and intellectual technologies. Proceedings of the International Workshop Dialogue'2012. Vol. 11 (18). Part 2. Moscow, pages 77-131.

Gladkij, A.V. (1973). *Formal'nye grammatiki i jazyki* [Formal Grammars and Languages], Moscow, Nauka.

Hovy, E. and Lavid, Ju. (2010). Towards a 'science' of corpus annotation: a new methodological challenge for corpus linguistics. International journal of translation, 22 (1): 1–25.

Iomdin, L.L. (1990). Avtomaticheskaja obrabotka teksta na estestvennom jazyke: *model'soglasovanija* [Natural Language Processing: a Model of Agreement]. Moscow, Nauka.

Lyashevskaya, O., Astaf'eva, I., Bonch-Osmolovskaya, A., Garejshina, A., Grishina, Ju., D'jachkov, V., Ionov, M., Koroleva, A., Kudrinsky, M., Lityagina, A., Luchina, E., Sidorova, E., Toldova, S., Savchuk, S., and Koval', S. (2010). Ocenka metodov avtomaticheskogo analiza teksta: morfologicheskije parsery russkogo jazyka [NLP evaluation: Russian morphological parsers], in Computational linguistics and intellectual technologies. Proceedings of the International Workshop Dialogue'2010. Vol. 9 (16). Moscow, pages 318–326.

Toldova, S., Sokolova E., Astaf'eva I., Gareyshina A., Koroleva A., Privoznov D., Sidorova E., Tupikina L., and Lyashevskaya O. (2012). Ocenka metodov avtomaticheskogo analiza teksta 2011-2012: sintaksicheskie parsery russkogo jazyka [NLP evaluation 2011-2012: Russian syntactic parsers]. In Computational linguistics and intellectual technologies. Proceedings of the International Workshop Dialogue'2012. Vol. 11 (18). Moscow, pages 797-809.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., and Yuret, D. (2007). The CoNLL 2007 Shared Task on Dependency Parsing. In Proceedings of EMNLP-CoNLL. Prauge, Czech Republic, pages 915-932.

ROMIP. (2009). Rossijskij seminar po ocenke metodov informacionnogo poiska. Trudy ROMIP 2009, Petrozavodsk, 16 sentjabrja 2009 [Russian Information Retrieval Evaluation Seminar. Proceedings of ROMIP 2009, Petrozavodsk, September 16, 2009]. Saint–Petersburg, NU CSI.

Sokolova, E. (2011). Sintaksicheskaja razmetka v terminax grammatiki zavisimostej i sintaksicheskix funkcij [Syntactic annotation in terms of dependency grammar and syntactic functions], Moscow, RGGU, available at: http://elib.lib.rsuh.ru/elib/000003603.pdf

Testelets Ja.G. (2001). Vvedenie v obshchij sintaksis [Introduction to general syntax], Moscow, RGGU.