Business-Media Analysis for Information Extraction

Ekaterina Pronoza, Elena Yagunova

Saint-Petersburg State University, Saint-Petersburg, Russian Federation

{katpronoza,iagounova.elena}@gmail.com

Abstract. In this paper an approach to primary business-media analysis for further information extraction is proposed. We consider business events representation by looking into part of speech (POS) distribution across tagged n-grams. Two Russian business-media corpora, Russian Business Consulting (RBC) and Commersant, are analyzed, and it is shown that they differ not only in style or themes coverage but also in the range of contexts for the words which mark business-events. Purchase, merger and ownership events are given a closer look at, and it is shown that they are mostly represented by noun phrases in both corpora rather than verbal phrases.

Keywords: Information extraction, business media.

1 Introduction

This paper considers the primary stage of news corpus analysis as part of business events extraction. This stage is closely connected to collocation analysis and aims at extracting key named entities and concepts hierarchy as well as investigating business events' representation. The former is a source of information for the local semantic dictionary whereas the latter may suggest specific ways of further business events extraction including part-of-speech patterns construction.

In this paper, two Russian business-media corpora are analyzed and it is shown that they both reveal noun phrases domination in business-events representation.

These Russian business-media sources demonstrate difficulties in Information Extraction (IE) imposed by the language (e.g., free word order characterizes Russian in general and especially in the media and in language specific domains). Our corpora analysis is very important for event extraction rules development taking into account the domain and stylistic features of the sources in question.

2 Related Work

In [5] an approach based on using collocation statistics for information extraction is proposed. Recognition rules are automatically extracted from the collocation database, and collocational context of words (co-occurrences) is treated as features for unknown proper names classification.

According to the approach proposed in [7], filtering techniques are applied to collocation sets as part of concepts extraction from text corpora. An approach for semantic network construction from the knowledge found in text corpora is presented.

In [10] IE algorithm based on local contextual information is presented. They propose a method of recognizing major constituents of a text as the most relevant collocational expressions and an algorithm which models relevant facts extraction.

A system of crime-related IE for the Arabic language is presented in [1]. Collocation analysis is used to obtain the concordance of the key words and also as part of further local syntactic analysis to define the type of crime, nationality, location, etc (indicator nodes) and supply data for the local grammar.

In [8] it is emphasized that for fully unsupervised event extraction, extensive linguistic analysis which increases the importance of text genre style and topic specification for IE (see also [11]) is essential. The idea of statistical comparison between text types and genres goes back at least as far as [2]. In [12] the linguistic cues indicating uncertainty of events in three genres (not only news) are studied; significant difference in lexical usage across genres is demonstrated.

The approach closest to ours is proposed in [11], where event representation across genre is considered. “Subject-verb-object” pattern statistics is analyzed and it is demonstrated that such statistics can differ across different genre/text types because event structure is strongly related to the genre of the corpus.

In this paper we also consider business events representation across the news corpora but in our approach n-grams, collocations (n-grams extracted by an associate measure) and n-gram part of speech (POS) statistics is used. Unlike [11], we do not propose any patterns (like S-V-O triplets) beforehand – on the contrary, the statistics obtained is used for patterns development. Moreover, our approach demands neither syntactic parser nor any special pattern-mining tool.

3 Data

The data used in the experiments consists of two business-media corpora: “РосБизнесКонсалтинг” (Russian Business Consulting, RBC) of 9 905 342 tokens (including 8 316 573 words) and “КоммерсантЪ” (Commersant) of 30 187 316 tokens (24 718 590 words). Both corpora include all the news articles released by these periodicals in 2011.

They were chosen for being main business-media representatives in Russia. We assume that RBC offers short-spoken businesslike news coverage while Commersant has more freedom of style. They also differ thematically: RBC specializes in financial and economic news and in Commersant there is usually a much broader range of events reported (including general political issues, etc.). Therefore it can also be an important task to compare the forms of business events representation in these media.

4 Methods and Instruments

4.1 Methods, Instruments and Main Assumptions

As part of primary analysis the corpora in question were tokenized and then lemmatized using Aot.ru (AOT)[1] morphology tools. Morphological disambiguation problem was solved by adopting the most frequent POS suggested by Aot.ru. Then n-gram frequencies were extracted.

Logarithmized Dice coefficient was chosen to be the measure for collocation extraction [4, 6, 9]:

$$\mathrm{Dice(x,\,y)} = \log_2 \frac{2f(xy)}{f(x)+f(y)}, \tag{1}$$

where $f(\mathrm{x}), f(y)$ refer to the frequencies of the words x and y and $f(xy)$ refers to the frequency of xy word combination. Dice coefficient is considered to be an adequate measure for our problem as it evaluates association degree between collocates based on compatibility and co-occurrence constraints. Moreover, this measure provides robustness of results for both large and small corpora, as shown in [4].

We only analyzed n-grams which consisted of two words not separated by a comma. A lexeme (and not word form) was chosen to constitute unit of analysis because it was important to consider all possible syntactic and semantic roles n-grams play in the sentence. It was also necessary to solve the disambiguation problem which would be unavoidable if we did not take parts of speech (POS feature) into account. For example, the normal form "*продать*" /sell/ is offered by AOT for the participle "*проданный*" /sold/ or "*продающий*" /selling/, the verb "*продать*" /sell/, the infinitive "*продать*" /to sell/ and the gerund "*продав*" /having sold/, and it is important to divide these four cases as they may indicate the form of business events representation. POS features adopted in this paper are noun, adjective, noun pronoun, verb, participle, gerund, infinitive, predicative pronoun, adjective pronoun, cardinal numeral, ordinal numeral, adverb, predicative, preposition, conjunction, interjection, particle, parenthesis and adjective and participle short forms.

We consider Dice coefficient extension for trigrams:

$$\mathrm{Dice(x,\,y,z)} = \log_2 \frac{3f(xyz)}{f(x)+f(y)+f(z)}, \tag{2}$$

where $f(\mathrm{x}), f(y)$ and $f(z)$ refer to the frequencies of the words x, y and z and $f(xyz)$ refers to the frequency of xyz word combination.

[1] An open source project providing tools for morphological, syntactic and semantic analysis for the Russian language: `http://aot.ru`.

The absolute frequency threshold of both bigrams and trigrams is chosen to be equal to 4 as we consider it to be a reasonable value for our task when corpora of several millions words are concerned. Now and later both “bigram” and “collocation” terms are used to refer to bigrams with the only difference that when calculating POS distribution for bigrams we sum frequency values (i.e., the number of occurrences in the corpus) when grouping them by a POS bigram (e.g., <Noun Noun>) while for collocations POS distribution we do not take the number of occurrences of a particular bigram into account which is equivalent to their frequency values being equal to 1.

4.2 Goal and Main Phases of Analysis

Our goal is primary business-media analysis, and the main phases of the media corpora processing may be described as follows:

- n-gram frequencies calculation,
- Dice coefficient calculation for bigrams and trigrams and collocations (extracted by Dice coefficient) analysis,
- tag words list construction for the business events (see Table 1),
- automatic business events tagging (for layoff, appointment and bankruptcy events),
- manual business events tagging (for the rest of the events),
- n-gram statistics calculation for the analyzed business events (including part of speech distribution).

The business events to be extracted from media texts are

- the purchase of assets, shares, etc. made by some company (PURCHASE tag),
- companies merger (MERGER tag),
- ownership of companies (OWNERSHIP tag),
- appointments, retirements or taking up a post (POST tag),
- employee layoff (LAYOFF tag),
- contract signing (CONTRACT tag),
- business investments (INVESTMENT tag),
- bankruptcy (BANKRUPTCY tag).

The tags given in the brackets will be adopted to refer to the event in this paper. In fact, these tags will be used in further IE as database table headers while fulfilling scenario templates.

Let us assume a tag word for a business event to be a lexeme (actually, a collocate) which describes this business event in one way or another. For example, a noun “*продавец*” /*seller*/ and a verb “*продать*” /*sell*/ would be tag words for the PURCHASE business event, while a noun “*банкрот*” /*bankrupt*/ and a verb “*обанкротиться*” /*go bankrupt*/ would be tag words for the BANKRUPTCY business event. It is easy to see that tag words for PURCHASE do not necessarily indicate a purchase made by a company while tag words for BANKRUPTCY describe exactly the event we are analyzing. Therefore our business events can be divided into two classes according to the feature of their tag words described below (for automatically or manually business events detection).

For the events like LAYOFF, POST and BANKRUPTCY tag words always describe the events we are interested in and it is possible to mark all the bigrams which include one of these words with business events tags automatically.

On the other hand, we have business events like PURCHASE, MERGER, OWNERSHIP, CONTRACT and INVESTMENT, and for these events their respectful tag words do not always specify the event we need. Therefore we should look though the list of bigrams containing these tag words and mark only those which indeed refer to the event we are interested in, and this cannot be done automatically. Some of these tag words are outlined in Table 1.

Table 1. Tag words for the business events processed manually

EVENT	TAG WORDS
PURCHASE	Купить /buy, purchase/; покупка /a purchase/; продавать /sell/; продажа /sale/
INVESTMENT	Инвестирование /investment/; инвестировать /invest/; инвестор /investor/
CONTRACT	Договор /contract/; контракт /contract/; подписать /sign/; сделка /bargain/
OWNERSHIP	Главный /chief/; дочерний /subsidiary/; филиал /branch/
MERGER	Объединение /union/; объединяться /unite/; поглощение /takeover/; слиться /merge/

4.3 Approach Restrictions

Trigrams are not considered with respect to the business events because a trigram collocation is usually a bigram collocation extension, and adding the third collocate to a bigram does not change the way it represents a particular business event. For example, a noun phrase "*покупка компания*" /*company purchase*/ cannot become a verbal phrase when a verb is added to the bigram as its action is represented by a verbal noun. It is an example of predicative bigram without any verbal component.

Furthermore, we do not take non-contiguous bigrams into account and therefore are unable to choose MWEs for tag words.

As for business events representation, we consider three following events: PURCHASE, MERGER and OWNERSHIP. INVESTMENT and CONTRACT which are manually marked with tag words are not included into the analysis part because they demonstrate too much ambiguity (e.g., a contract is not necessarily the contract between two companies, and the word "*investment*" can be used in metaphoric sense).

Business events for which bigrams are marked automatically are not considered in this paper because their representation simply depends only on the list of tag words we supply the classification program with. For example, for BANKRUPTCY, such tag words are "*банкрот*" /*bankrupt*/, "*обанкротиться*"/*go bankrupt*/, "*разориться*"

/go bankrupt/, and for POST event they are "*должность*" */post/*, "*пост*" */post/*, "*отставка*" */retirement/*.

5 N-grams and Collocations

5.1 Bigrams and Trigrams. General Characteristics

Top bigram collocations of the RBC corpus sorted in the descending order by Dice coefficient and by their frequency (with frequency threshold value equal to 41) were analyzed (all in all, there are 16698 bigrams).

Top bigram collocations (a few hundred) of this list refer to the three classes of multiword expressions (MWE):

- a Proper Name person (or a part of it), like *Усама Бен /Osama Bin/*, *Пан Ги /Ban Ki/*, *Broco Алексей /Broco Alexey/*, *Фог Расмуссен /Fogh Rasmussen/*, etc.;
- a Proper Name company (organization), e.g., *Middle East*, *IPE Brent*, *Церих Кэпитал /Zerich Capital/*;
- a media cliché and compound function words, like "*сослаться на*" */refer to/*, "*применительно к*" */with reference to/*, "*несмотря на*" */despite/*, etc.

Top trigram collocations for the RBC show the same tendency. Some of the compound phrases partially represented in the bigram collocations list are given here as full phrases, e.g., *Middle East Crude* and *Церих Кэпитал Менеджмент /Zerich Capital Management/*. There is also a political context indication, e.g., in MWE like "*расколоть страна на" /split country into/*, "*экс-кандидат в президент" /ex-presidential candidate/* and "*злободневный вопрос российский" /burning question russian/*. These MWE also prove that financial and economic news dominate in the RBC corpus (see "*фиксинг по золото" /gold fixing/*, "*наличный рынок драгоценный" /cash market precious/* and "*нерыночный актив или" /non-market asset or/*).

As in the RBC case, the Commersant's bigram and trigram collocations also constitute the three main types of MWE: a proper Name person, a Proper Name company and a media cliché and compound function words. The only difference from RBC is poorer financial terminology representation together with social and political terms domination (e.g., "*возбудить уголовный дело*" */launch criminal case/*, "*полномочный представитель президент*" */presidential plenipotentiary/*).
We would also like to emphasize a specialty of person MWE in both RBC's and Commersant's corpora[2].

[2] *Broco Алексей /Broco Alexey/* collocation takes 11th place in top RBC bigrams list (frequency = 154, Dice = 0.99 (out of 1). This is apparently Alexey Matrosov, who is the chief of analytical department of ГК Broco company. At the same time *Алексей Матрос* /Alexey Matrosov/ collocation is less frequent in the corpus (frequency = 122, Dice = -3.04). This tendency can also be illustrated by an example from the Commersant's corpus: we have *Совлинк Ольга /Sovlink Olga/* (Olga Belenkaya, the chief of analytical department of Sovlink company) with frequency = 57 and Dice = 57 and *Ольга Беленький /Olga Belenkiy/* with

Famous politicians' names mostly appear in the top bigram collocations list in the <first name, last name> form, rather than in <last name, post> or <last name, organization/country> combinations. A couple of examples are given in Table 3 (with F1 and D1 for RBC and F2 and D2 for Commersant as frequency and Dice).

Table 2. Some person collocation characteristics

Bigram	F1	D1	F2	D2	Bigram	F1	D1	F2	D2
Барак Обама /Barack Obama/	599	0.44	1381	0.36	*Ангела Меркель /Angela Merkel/*	242	0.39	250	0.26
США Барак /USA Barack/	567	-3.5	487	-3.5	*Германия Ангела /Germany Angela/*	200	-2.4	130	-3.4
президент Барак /president Barack/	80	-6.4	152	-6.8	*ФРГ Ангела /FRG Angela/*	88	-0.7	48	-1.43
президент Обама /president Obama/	< 4	-	320	-5.7	*канцлер Ангела /chancellor Angela/*	27	-2.9	62	-1.4
господин Обама /Mr. Obama/	< 4	-	133	-7.8	*госпожа Меркель /Mrs. Merkel/*	< 4	-	26	-7.13

We can conclude that the famous people are mainly given in the <first name, last name> form, whereas others are represented with reference to a company. This assumption is important for further IE as it shows how POST events should be handled.

5.2 POS Characteristics of Collocations

Since one of the goals of our study is finding out the information necessary for building the patterns for business events extraction system, we consider POS-based statistics. Bigrams POS distribution is estimated (without frequency threshold this time) for both corpora. A list of top 10 POS bigrams is shown in Table 3.

Table 3. POS distribution for RBC and Commersant bigrams

POS1 POS2	RBC		Commersant	
	freq	%	freq	%
Noun Noun	608879	14	1604411	12
Preposition Noun	555211	13	1305335	9.9
Noun Preposition	460552	11	1085355	8.2
Adjective Noun	366324	8.3	826811	6.3
Noun Verb	179283	4.1	444857	3.4
Preposition Adjective	158440	3.6	398119	3
Noun Adjective	148556	3.4	347163	2.6
Verb Preposition	141581	3.2	351963	2.7
Noun Conjunction	117303	2.7	386254	2.9
Verb Noun	101084	2.3	273452	2.1

frequency = 4 and Dice = -8.66 (almost minimal). *Analysis Михаил /Analysis Mikhail/* (Mikhail Korchemkin, the chief director of East European Gas Analysis) and *АвиаПорт Олег /AviaPort Oleg/* (Oleg Panteleev, the chief editor of AviaPort agency) collocations reveal the same tendency.

Thus, <Noun Noun>, <Noun Preposition> and <Preposition Noun> bigrams dominate in both corpora. We shall use non-parametric Mann-Whitney U-test for statistical hypotheses testing as our data is nominal and aggregated (grouped by POS).

RBC and Commersant show similar POS distribution and according to Mann-Whitney U-test the difference is statistically insignificant at the 0.05 level.

In Table 4 top 10 POS collocations with largest Dice value are shown.

Table 4. POS Distribution for RBC and Commersant Collocations

POS1 POS2	**RBC**		**Commersant**	
	amount	**%**	**amount**	**%**
Noun Noun	32604	16	116750	17
Adjective Noun	16886	8.3	51624	7.3
Noun Verb	14742	7.2	51718	7.3
Noun Preposition	13322	6.5	35297	5
Preposition Noun	12223	6	31386	4.4
Noun Adjective	11135	5.5	36246	5.1
Verb Noun	6929	3.4	25444	3.6
Preposition Adjective	5550	2.7	14445	2
Noun Conjunction	5144	2.5	16964	2.4
Conjunction Noun	4448	2.2	16107	2.3

According to Mann-Whitney U-test, the difference between RBC and Commersant POS distribution for the collocations is significant at the 0.05 level. In particular, the RBC corpus is more heterogeneous (with respect to collocations POS distribution) than the Commersant's (with standard deviation value equal to 1.66 against 1.34).

6 Business Events Tagging

As we have previously shown, some tag words can unambiguously define an event. And here we introduce the idea of a lexeme's context variety. If a lexeme (i.e., a tag word) has small context variety, it is a perfect event classification tool (this is what we have for POST, LAYOFF and BANKRUPTCY), but if its context variety is large, an arbitrary bigram containing this lexeme cannot be automatically attributed to an event. We also introduce context determinacy, which is opposite to context variety: the larger context variety is, the smaller context determinacy we have, and vice versa.

We have calculated the latter as a portion of bigrams containing a given tag word and attributed to a particular event out of all the bigrams containing this tag word. Context determinacy values for the tag words (for the business events we are interested in) are shown in Table 5, with L1, R1 for RBC's left and right context determinacy, and L2 and R2 for Commersant's context determinacy respectively.

Table 5. Tag words context determinacy

TAG WORD	L1, %	R1, %	L2, %	R2, %	TAG WORD	L1, %	R1, %	L2, %	R2, %
MERGER									
объединить /to unite/	0	9	0	1	соединяться /to unite/	0	31	-	-
сливаться /to merge/	0	9	-	-	присоединение /adjunct/	0	10	0	2
слияние /merger/	33	51	2	19	объединение /union/	-	-	0	3
объединяться /to unite/	100	0	17	0	объединиться /to unite/	-	-	7	0
поглощение /takeover/	0	31	0	20					
OWNERSHIP									
материнский	66	100	4	61	филиал /branch/	62	84	67	74
дочерний /subsidiary/	62	98	15	96	представительство /agency/	3	48	0	21
PURCHASE									
покупка /purchase/	0	15	0	13	купить /to buy/	3	0	5	5
покупать /to buy/	0	39	8	4	продать /sell/	0	3	2	3
продажа /sale/	0	1	0	5	приобрести /to purchase /	2	6	8	5
продавать /to sell/	4	0	2	2	покупатель /purchaser/	-	-	0	1
приобретение /purchase/	0	27	0	17	приобретать /to purchase/	-	-	11	5

Apparently Commersant has a bit smaller context determinacy than RBC (and larger context variety). According to Chi-Square test, the difference between left determinacy value for RBC and Commersant is significant at the 0.001 level.

RBC also reveals larger standard deviation than Commersant for both left and right context (0.096 and 0.109 for RBC against 0.022 and 0.071 for Commersant). It suggests that the Commersant corpus is more homogeneous than the RBC's when business events representation is concerned but, as we only consider three business events, there is not enough data to prove whether this is statistically significant.

7 Business Events Representation

7.1 PURCHASE

As far as the PURCHASE event is concerned, the RBC corpus shows small representation forms variety. Noun phrases (NPs) like <Noun Noun> dominate throughout all frequency levels, e.g., *покупка компания /company purchase/*, *приобретение актив /asset purchase/*, *приобретение сеть /chain purchase/*, etc. The Commersant corpus has wider constructions variety, and <Noun Noun> phrases dominate among the most frequent ones, e.g., *приобретение доля /share purchase/*, *приобретение актив /asset purchase/*, *покупка ООО /OOO purchase/* (see Table 6).

According to Mann-Whitney U-test, there is no significant difference between the two distributions at the 0.05 level. But if we group all part of speech bigrams into NPs[3] and verbal phrases (VPs)[4] we shall have the total portion of noun phrases equal to 56% against the portion of VPs equal to 44% for RBC and 52% against 48% for Commersant respectively. In other words, noun phrases are slightly more frequent in

[3] <Noun Noun>, <Noun Adjective>, <Participle Noun>, <Noun Participle> and <Noun Participle (short)>

[4] <Infinitive Noun>, <Infinitive Adjective>, <Noun Infinitive>, <Noun Verb>, <Verb Noun> and <Gerund Noun>

both corpora than verbal phrases, and this difference is smaller in Commersant's case which leads to the suggestion that the latter is more "balanced".

Table 6. POS distribution for PURCHASE (RBC, Commersant)

POS1 POS2	Commersant		RBC	
	freq	%	freq	%
Noun Noun	226	35	662	48
Infinitive Adjective	176	27	-	-
Noun Adjective	123	19	9	0.6
Infinitive Noun	78	12	101	7
Noun Verb	26	4.0	321	25
Participle Noun	15	2.3	33	2
Verb Noun	5	0.7	158	11
Noun Infinitive	-	-	42	3
Noun Participle (short)	-	-	13	0.9
Noun Participle	-	-	10	0.7
Gerund Noun	-	-	10	0.7

7.2 MERGER

If we consider MERGER, it is easy to see that in RBC NPs dominate across all frequency levels: *слияние Skype /Skype merger/*, *поглощение концерн /concern takeover/*, *присоединение дочерний /adjunct subsidiary/*, etc. In Commersant NPs are even more frequent: *слияние актив /asset merger/*, *объединение Газпром* /Gazprom union/, *слияние ОАО /OAO merger/*, etc. All these NPs are predicative structures (with verbal nouns with intact verbal valences). POS distribution of the bigrams for MERGER is given in Table 7.

Table 7. POS distribution for MERGER (RBC, Commersant)

POS1 POS2	RBC		Commersant	
	freq	%	freq	%
Noun Noun	177	42	170	45.5
Adjective Noun	140	33	150	40
Participle Noun	40	10	6	1.6
Noun Adjective	33	8	14	3.7
Noun Numeral	15	4	-	-
Verb Noun	12	3	-	-
Noun Verb	4	1	23	6.2
Infinitive Noun	-	-	11	3

Mann-Whitney U-test shows no significant difference between the two distributions at the 0.05 level. Let us again sum portions of NPs and VPs for both corpora. For RBC corpus we have 96% and 4% of NPs and VPs respectively, and for Commersant the results are 91% and 9%.

This shows that MERGER in about 9 cases out of 10 is represented by NP rather than by VP in both corpora and that Commersant has a little more balance when the difference between the portions of NPs and VPs is concerned for the MERGER event.

7.3 OWNERSHIP

OWNERSHIP representation has a certain specialty: in Russian a relatively small number of verbs are appropriate for its description (e.g., the verb "*удочерить*" /*adopt as a daughter*/ is unlikely to be used with respect to a company, and "*представлять компанию*" /*represent a company*/ is referred to the employees and not the subsidiary company). As a result, NPs dominate in both corpora for this event. Some examples presented in Table 8, and in Table 9 OWNERSHIP bigrams POS distribution is given.

Table 8. OWNERSHIP bigrams (examples)

RBC	Commersant
дочерняя компания /subsidiary company/	дочерняя авиакомпания /subsidiary aviacompany/
материнская компания /parent company/	материнская структура /parent structure/
сеть дочерний /subsidiary chain/	дочерний ООО /subsidiary OOO/

Table 9. POS distribution for OWNERSHIP (RBC, Commersant)

POS1 POS2	RBC		Commersant	
	freq	%	freq	%
Adjective Noun	365	52	1022	62
Noun Adjective	126	18	197	12
Noun Noun	124	18	358	22
Adjective Adjective	29	4	43	3
Adjective pron. Adjective	23	3	-	-
Verb Adjective	16	2	4	0.2
Verb Noun	7	1	5	0.3
Preposition Adjective	6	0.9	-	-
Participle Adjective	4	0.6	-	-
Participle Noun	-	-	18	1

As far as the difference between these two distributions is concerned, it is not statistically significant at the 0.05 level according to Mann-Whitney U-test. Summed up portions of NPs and VPs are equal to 97% and 3% for RBC and 99% and 1% for Commersant. The latter shows a wide variety of <Adjective Noun> phrases with the same noun and various adjectives, and this seems to be the reason of the difference. And yet OWNERSHIP is represented almost by NPs only in both corpora.

8 Conclusion and Future Work

Collocation analysis of the RBC and the Commersant corpora shows that these two business media differ in both style and themes. RBC is a finance specialized periodical while Commersant covers a broad range of issues.

There is no statistically significant difference between the two corpora with respect to business events representation. NPs dominate for ***merger*** and ***ownership*** and are slightly more frequent than verbal phrases for ***purchase***. A bigram consisting of two nouns is the most frequent type of ***purchase*** and ***merger*** representation form, while the <Adjective, Noun> pair mostly represents ***ownership*** (for both corpora).

Analysis of tag words context determinacy for ***purchase***, ***merger*** and ***ownership*** reveals that Commersant has significantly wider variety of tag word left context than RBC. In fact, the Commersant corpus is somewhat more balanced with respect to business events representation form.

Our future work plans include development of the Russian business domain ontology (which involves identifying the most significant noun groups and organizing them into lexicon attached to the ontology), identification of the most significant predicate structures (both NPs and VPs) and search patterns implementation (with patterns appropriate for Russian media language) based on them, with NPs prior to VPs (at least for the three business events mentioned above) since, according to the statistics described in the paper, they cover most of business events references in the media corpora.

Acknowledgement. We would like to thank our friends and colleagues, especially Lidia Pivovarova, for their invaluable guidance, support and encouragement.

References

1. Alruily, M.F.: Using Text Mining to Identify Crime. Patterns from Arabic News Report Corpus (2012)
2. Biber, D.: Variation across Speech and Writing. Cambridge University Press (1991)
3. Cowie, J., Wilks, Y.: Information Extraction. In: Dale, R., Moisl, H., Sommers, H. (eds) Handbook of Natural Language Processing, ch. 10, pp. 241−269, Dekker (2000)
4. Daudaravicius, V.: Automatic Identification of Lexical Units. Informatica, 34 (1), pp. 85−91 (2010)
5. Dekang, L.: Using Collocation Statistics in Information Extraction. In: Proceedings of the Seventh Message Understanding Conference (MUC-7) (1998)
6. Dice, L.R.: Measures of the Amount of Ecologic Association between Species, vol. 26 (3), pp. 297−302 (1945)
7. Heyer, G., Quasthoff, U., Wolff, Ch.: Information Extraction from Text Corpora: Using Filters on Collocation Sets. In: Proceedings of the Third Int'l. Conference on Language Resources and Evaluation (LREC 2002), vol. 3, pp. 1103−1107. Las Palmas (2002)
8. Grishman, R.: Structural Linguistics and Unsupervised Information Extraction. In: Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX 2012), pp. 57−61 (2012)
9. Manning, Ch.D., Schutze, H.: Foundations of Statistical Natural Language Processing. The MIT Press, London (1999)
10. Perrin, P., Petry, F.E.: Extraction and Representation of Contextual Information for Knowledge Discovery in Texts. Information Sciences – Informatics and Computer Science: An International Journal, vol. 15, pp. 125−152 (2003)
11. Pivovarova, L., Huttunen, S., Yangarber, R.: Event Representation across Genre. In: Proceedings of NAACL Workshop on Events: Definition, Coreference and Representation (2013)
12. Szarvas, G., Vincze, V., Farkas, R., Mófra, G., Gurevych, I.: Crossgenre and Cross-Domain Detection of Semantic Uncertainty. Computational Linguistics, 38 (2), pp. 335−367 (2012)