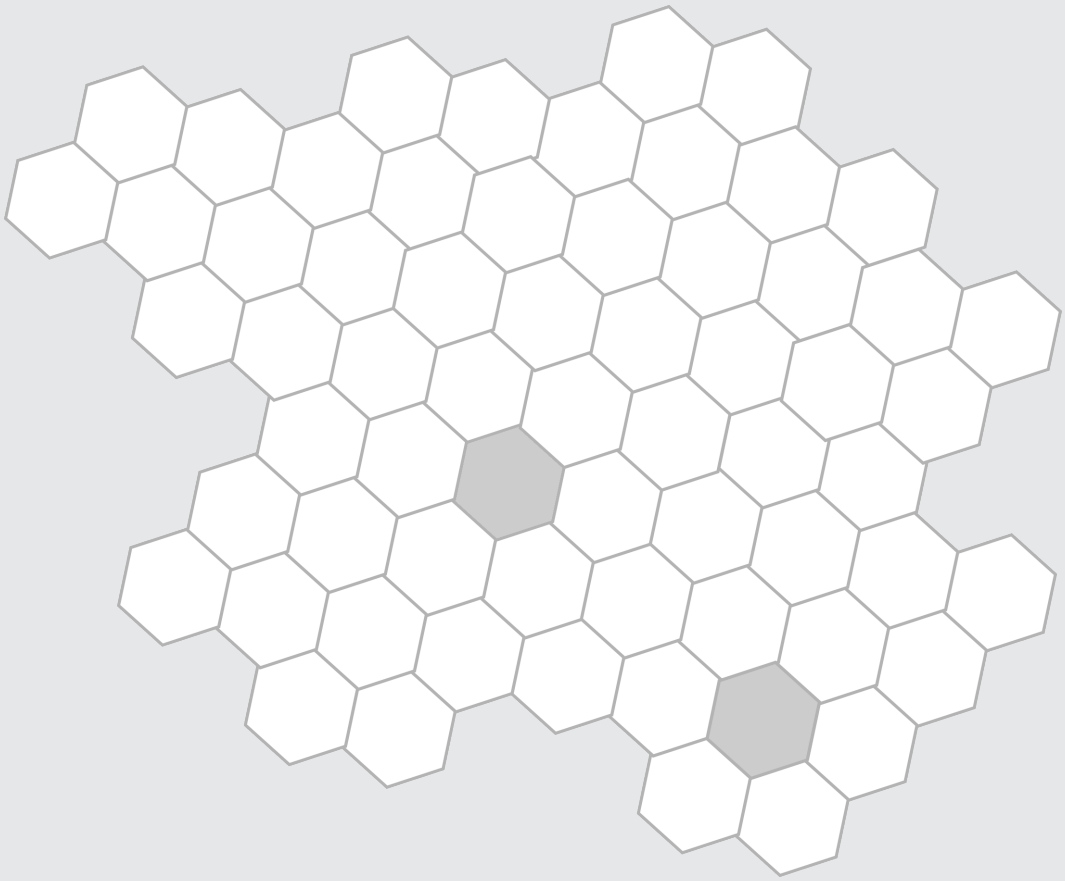
International Journal on

SOCIAL MEDIA

MMM: Monitoring, Measurement, and Mining



II, 2011, 1–2
ISSN 1804-5251

...OFFPRINT...

International Journal on SOCIAL MEDIA

MMM: Monitoring, Measurement, and Mining

II, 2011, 1–2

Editor-in-Chief: Jan Žižka

Publisher's website: www.konvoj.cz
E-mail: konvoj@konvoj.cz, SoNet.RC@gmail.com
ISSN 1804-5251

Language-specific Features in Multilingual Sentiment Analysis

Taras Zagibalov
Brandwatch, United Kingdom
e-mail: taras8055@gmail.com

Katerina Belyatskaya
Siberian Federal University, Russia
e-mail: e.o.belyatskaya@gmail.com

John Carroll
University of Sussex, United Kingdom
e-mail: J.A.Carroll@sussex.ac.uk

Abstract

We present newly-produced comparable corpora of book reviews in English and Russian. The corpora are comparable in terms of domain, style and size. We are using them for cross-lingual experiments in document-level sentiment classification. Quantitative analyses of the corpora and the language differences they exhibit highlight a number of issues that must be considered when developing systems for automatic sentiment classification. We describe experiments applying a supervised sentiment classification technique to the corpora. The results of the experiments suggest that differences in the basic characteristics of the two languages and the ways in which sentiment is expressed in the languages lead to significant differences in sentiment classification accuracy.

Key words Sentiment analysis; comparable sentiment corpora

Introduction

We investigate the effect of language-specific features on automated analysis of sentiment in English and Russian. Sentiment analysis is concerned not with the topic or factual content in a document, but rather with the opinion expressed in it. Sentiment analysis has often been broken down into a set of subtasks, including subjectivity classification, opinion classification (or sentiment classification), opinion holder and opinion target extraction, and feature-based opinion mining.

Sentiment classification is usually framed as a two-way classification into positive and negative sentiment, and has been applied at various levels: phrases, sentences, documents and collections of documents. An opinion may have a holder (a person or a group that expresses an opinion) and a target (an object which is being discussed or evaluated). Feature-based opinion mining tries to find opinions about particular features of a product or service (as opposed to an overall opinion about something). Automatic classification of document sentiment (and more generally extraction of opinion from text) has recently attracted much interest. One of the

main reasons for this is the importance of such information to companies, other organisations and individuals.

Applications include computer-based tools that help a company see market or media reaction towards their brands, products or services. Another type of application is a search engine that helps potential purchasers make an informed choice of a product they want to buy. Such search engines may include a sentiment classification subsystem that not only presents to a customer the overall sentiment about a product, but may also select positive or negative reviews to illustrate the perceived strengths and weaknesses of a product.

Automated sentiment analysis provides a range of possibilities for researchers in humanities whose studies involve analysis of large amounts of human-generated data. For example, in media studies one might be interested to see whether sentiment regarding the same events is shared in the mainstream media and in social media. Analysis of user-generated content may be very helpful in political studies. For example, the monitoring of political debates in social media may help to estimate the prospects of political candidates in elections or evaluate the effectiveness of political campaigns. The study of 'the language of hatred' contributes to efforts against political and religious extremism and intolerance. Many aspects of social studies may benefit from automatic analysis of sentiments expressed by people in ever-growing social networks. This approach offers unobtrusive and fast access to large amounts of information.

A major current challenge is to be able to automatically extract sentiment information from a variety of documents in different languages. In a recent white paper addressing the role of sentiment analysis in organisations, Grimes (2010) noted 'one axiom of full-circle sentiment analysis is ability to use all relevant sentiment sources'. This obviously includes sources representing different genres and styles, and written in different languages.

The most widely used approach to opinion and subjectivity classification is based on supervised machine learning, in which a system learns from human-annotated training data how to classify documents (e.g. Pang et al., 2002). However, a major obstacle for automatic classification of sentiment and subjectivity is the cost of collecting annotated training data. The rapid growth in the number of languages represented on the Internet and the emergence of new forms of social media makes it increasingly difficult to create and maintain suitable annotated corpora. Rule-based or dictionary-based approaches to sentiment analysis have similar limitations since they rely on large sets of manually created resources that need to match the language data being processed.

There are a number of publicly available sentiment-annotated corpora, such as MPQA (Wiebe et al., 2005), and the Pang and Lee (2004) Movie Review corpus. However, most of these corpora consist only of English text. There are some corpora designed for cross-lingual evaluations, but these seem not to be publicly available, for example the NTCIR MOAT corpora of English, Japanese and Chinese (Seki et al., 2008).

There has been little previous work on applying sentiment analysis to languages with scarce relevant language resources. A notable exception is the work towards producing cross-lingual subjectivity analysis resources from English data by Mihalcea et al. (2007). They use a parallel corpus to adjust a subjectivity lexicon

Table1: Case forms of the Russian adjective *хороший* (good).

Cases	Masculine singular	Feminine singular	Neuter singular	Plural
Nominative	*хороший*	*хорошая*	*хорошее*	*хорошие*
Genitive	*хорошего*	*хорошую*	*хорошего*	*хороших*
Dative	*хорошему*	*хорошей*	*хорошему*	*хорошим*
Accusative	*хорошего / хороший*	*хорошую*	*хорошее*	*хороших / хорошие*
Ablative	*хорошим*	*хорошей*	*хорошим*	*хорошими*
Prepositional	*хорошем*	*хорошей*	*хорошем*	*хороших*

translated from English to Romanian. Other multilingual opinion mining work (in English, Japanese and Chinese) was carried out by Zagibalov and Carroll (2008, 2009), using techniques requiring limited manual input to classify newswire documents with respect to subjectivity and to extract opinion holders and targets.

A related issue that has also received little attention to date is the impact on the design and performance of sentiment analysis systems across languages, stemming from differences in the characteristics of the languages and the means commonly used to express sentiment in them. To address this issue, we have designed and built comparable corpora of book reviews in English and Russian. The corpora are comparable in terms of domain, style and size. The Russian corpus is probably the first sentiment-annotated resource in that language. In the following sections we outline characteristics of the two languages, describe the corpora and quantify their various relevant aspects, and analyse some important language-specific issues that would be likely to impact on automatic sentiment processing. We go on to apply supervised and unsupervised sentiment classification techniques to the corpora to quantify the impact of these language-specific issues on classification accuracy.

Language Characteristics

In this study we focus on English and Russian.

Russian has a relatively complex morphology that comprises gender, case and number forms of adjectives and nouns as well as inclination and tenses, and aspect forms of verbs. For example, the adjective *хороший* (*good*) has the following forms:

- *хороший* – masculine, singular
- *хорошая* – feminine, singular
- *хорошее* – neuter, singular
- *хорошие* – plural (the same for all genders)

Each of these forms may be used with different cases, many of which have different endings (see Table 1).

There are also comparative and superlative forms of the adjective: *лучше* and *наилучший / самый лучший* (the latter is an analytical superlative form). The word can also be used in a short form: *хорош*. The number of forms (16 distinct forms) suggests the need for language-specific lexical processing (for example with a morphological analysis tool) before any application-level processing could take place.

English uses morphological means to express grammatical tense and aspect for verbs, and singular and plural for nouns. Arguably the most important part

of speech for sentiment analysis – adjectives – also have comparative and superlative forms which sometimes are formed irregularly (e.g., *good – better – best* and *bad – worse – worst*). Nevertheless, the variation of grammatical forms in English is not as complex as in Russian.

In a language-based application, such as sentiment analysis, without lexical processing (such as morphological analysis, stemming or lemmatisation) one may have the problem of data sparseness since numerous word forms would 'hide' a single word, even if a large amount of corpus data was available. However, lexical processing of this type is necessarily language-dependent, making it expensive to use this type of approach in a system that covers multiple languages.

The Corpora

Corpora Content

Our English and Russian book review corpora consist of reader reviews of science fiction and fantasy books by popular authors. The reviews were written in 2007, so the language used is current.

The Russian corpus consists of reviews of Russian translations of books by popular science fiction and fantasy authors, such as S. King, S. Lem, J. K. Rowling, T. Pratchett, R. Salvatore, J. R. R. Tolkien as well as by Russian authors of the genre such as S. Lukyanenko, M. Semenova and others. The reviews were published on the website `www.fenzin.org`.

The English corpus comprises reviews of books by the same authors, if available. If some of the authors were not reviewed on the site or did not have enough reviews, they were substituted with other writers of the same genre. As a result, the English corpus contains reviews of books such as: S. Erickson (*Guardians of the Moon*, *Memories of Ice*), S. King (*Christine, Duma Key, Gerald's Game, Different Season* and others), S. Lem (*Solaris, Star Diaries of Iyon Tichy, The Cybriad*), A. Rise (*Interview with the Vampire, The Tale of the Body Thief* and others), J.K. Rowling (*Harry Potter*), J. R. R. Tolkien (*The Hobbit, The Lord of the Rings, The Silmarillion*), S. Lukyanenko (*The Night Watch, The Day Watch, The Twilight Watch, The Last Watch*), and a few others. The reviews were published on the website `www.amazon.co.uk`.

Although both of the sites from which the reviews were collected feature review-ranking systems (e.g., one to ten stars), many reviewers did not use the system or did not use it properly. For this reason all of the reviews were read through and hand-annotated. There were a lot of reoccurring short reviews such as: Хорошо (*Good*); Интересная книга (*Interesting book*); Супер! (*Superb!*); Нудятина!! (*Boring!!*); Ниже среднего (*Below average*); *Awesome!*; *Amazing!*; *The best book I've ever read!*; *Boring*, and so on. These reviews were added to the corpus only once. Also both sites had a number of documents which did not have any direct relation to book reviewing, such as advertisements, announcements and off-topic postings. Such texts were excluded as irrelevant. The documents that were included in the corpora were not edited or altered in any other way.

We annotated each review as 'POS' if positive sentiment prevails or 'NEG' if the review is mostly negative based on the tags assigned by reviewers, but moderated where the tag was obviously incorrect. Each corpus consists of 1,500 reviews,

Table2: Overall quantitative measures of the English and Russian corpora.

	Mean tokens POS	Mean tokens NEG	Total types POS	Total types NEG
English	58	58	7349	8014
Russian	30	38	9290	12309

half of which are positive and half negative. The annotation is simple and encodes only the overall sentiment of a review, for example:

[TEXT = POS]
Hope you love this book as much as I did. I thought it was wonderful!
[/TEXT]

The English reviews contain a mean of 58 words (the mean length for positive and negative reviews being almost the same). Positive Russian reviews have a mean length of only 30 words; negative reviews are slightly longer, at 38 words (see Table 2). It is not possible to compare these figures directly between the languages as they have different grammar structures which makes English more 'wordy', as it has function words (articles, auxiliary verbs) which are almost completely absent in Russian.

As noted above, Russian, being a synthetic language, has many forms of the same lemma. This results in a large number of distinct word forms: the corpus contains a total of 13 472 word forms, with 6 589 (42%) in positive reviews and 8 993 (58%) in negative. The total number of words in the corpus is 50 745, which means that every word form was used a little more than three times on average. The English corpus has only 7 489 distinct word forms in the whole corpus, 4 561 (47%) in positive reviews, and 5 098 (53%) in negative. These figures also suggest that Russian reviewers used a richer vocabulary for expressing *negative* opinions (compared to the number of unique words used in Russian positive reviews) than English reviewers.

Further evidence of the different ways in which people distinguish sentiment polarity in Russian compared with English is the distribution of the lengths of positive and negative reviews. The Russian corpus has a large number of short reviews (less than 50 words) with a median of 15 words for positive reviews and 10 words for negative reviews. Apart from the language-specific differences mentioned above that partly account for the smaller number of words in Russian documents, there is a clear difference from English reviews in terms of length. The English reviews feature a more or less equal number of documents of different lengths (mostly in the range 15 to 75 words). The prevalence of short reviews in the Russian corpus, together with the rich morphological variation, may lead to data sparseness which would be a problem for current sentiment classification techniques.

Ways of Expressing Sentiment

Sentiment can be expressed at different levels in a language: from lexical and phonetic levels up to discourse level. This range is reflected in the corpora (see Tables 3

Figure1: Distribution of documents by number of words

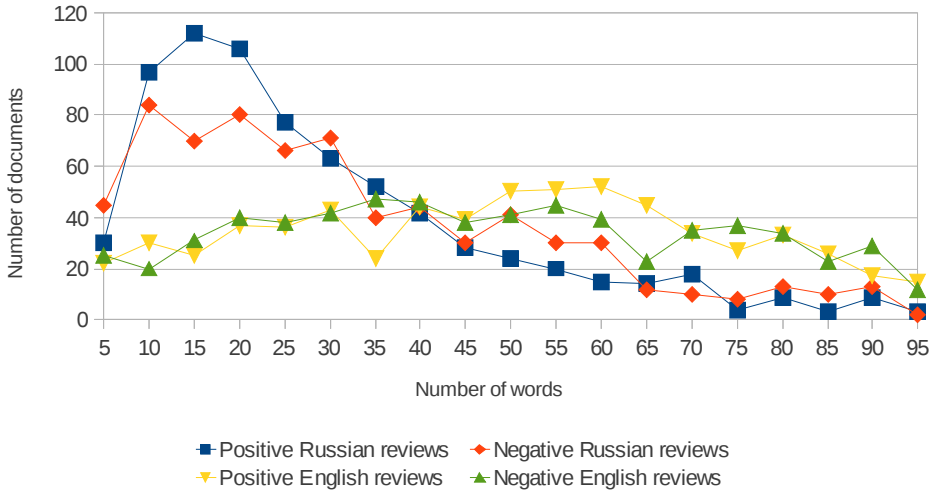


Table3: Ways of expressing sentiment in the English Book Review Corpus (numbers of documents)

	Syntactic	Lexical				Phonetic
		Verb	Adjective	Noun	Other	
Positive	432	312	708	225	325	12
Negative	367	389	652	238	407	16
Total	799	701	1360	463	732	28

and 4)[1]. As the Tables show, the authors of reviews in the two languages express sentiment in slightly different ways. In English they make heavy use of adjectives to express sentiment (this class of words is used to express sentiment in a third of all documents). In contrast, in Russian they use verbs as often as adjectives to express sentiment (both of these classes are used in about a quarter of all reviews) and make more use of nouns (expressing sentiment in 15% of all documents compared to 11% in English). The Russian corpus also demonstrates a tendency to combine different ways of expressing sentiments in a document: the total number of uses of different ways in the English corpus is 4,083 compared to 4,716 in Russian, which means that given an equal number of reviews for each language, Russian reviews tend to have more ways of expressing sentiment per document.

Lexical Level

Adjectives are the most frequent way of expressing opinions in both corpora, closely followed by verbs in the Russian corpus. 1,215 Russian reviews use adjectives to ex-

[1] All the numerical data presented below comes from manual counting and is not represented in the corpus annotation.

Table4: Ways of expressing sentiment in the Russian Book Review Corpus (numbers of documents).

	Syntactic	Lexical				Phonetic
		Verb	Adjective	Noun	Other	
Positive	417	492	648	374	367	27
Negative	475	578	567	334	394	43
Total	892	1070	1215	708	761	70

press sentiment and 1,070 reviews use verbs. In the English corpus there are 1,360 reviews that use adjectives, but only 701 use verbs to express opinion.

Apart from adjectives, which are recognised as the main means of expressing evaluation, other parts of speech are also often used in this function, most notably verbs and nouns. The English reviews also feature adverbials, and both languages also use interjections.

Akimova and Maslennikova (1987) observe that opinions delivered by means of verbs are more expressive compared to opinions expressed in other ways. This is explained by the fact that a verb's denotation is a situation and the semantic structure of the verb reflects linguistically relevant elements of the situation described by the verb. Verbs of appraisal not only name an action, but also express a subject's attitude to an event or fact.

Consider the following examples:

- *I truly loved this book, and I KNOW you will, too!*
- *понравилось, научная фантастика в хорошем исполнении*

  I liked it, it's science fiction in a very good implementation

The English verbs *loved* and *liked* describe an entire situation which is completed by the time of reporting it. This means that a subsequent shift in sentiment polarity is all but impossible:

- **I truly loved this book, but it turned out to be boring.*

However, adjectives usually describe only attributes of certain members of a situation leaving a significant amount of context aside:

- *The story is pretty good but it stretches on and on.*

In the example above a positive sentiment towards the story is shifted to negative. A verb is less usual in such a context:

- (?) *I liked the story but it stretches on and on.*

Nouns can both identify an object and provide some evaluation of it. But nouns are less frequently used for expressing opinion compared to verbs. Nonetheless in the Russian corpus, nouns were used more than in the English corpus. There are 708 Russian reviews that have opinions expressed by nouns, however only 463 English reviews made use of a noun to describe opinion. The most frequent such nouns used in Russian reviews are чудо (*miracle*), классика (*classics*), шедевр (*masterpiece*), гений (*genius*), прелесть (*delight*), бред (*nonsense*), мура (*raspberry*), жвачка (*mind-numbing stuff*), ерунда (*bugger*).

Phonetic Level

Although the corpora consist of written text and do not have any speech-related mark-up, some of the review authors used speech-related methods to express sentiment, for example:

- *This was a sloooow, frail story*
- *A BIG FAT ZEEROOOOOOOOOOOO for MA*
- *i have to say is a good booooooooooooooooook!*
- *Ну что сказать...чепуха...ЧЕ-ПУ-ХА.*
  What should I say... boloney... BO-LO-NEY
- *Ндаааааа.....такую муть давно не видел*
  Weeeeelll..... I haven't seen such a stinkaroo for long
- *абалденная книшкаааааа!!!!!!!!!!!!!!!))) оч давно её люблю))*
  jaw-droppin' booooook!!!!!!!!!!!!!!!))) been lovin' it for long
- *Мозг ломиться от этого несоответствия... и получает ооочень большой кайф!!!*
  My brain is bursting because of this inconstancy... and it enjoys it veeery much!!!
- *Читать ВСЕЕЕЕЕЕЕЕЕМ*
  Read, EVERYBOOOOODY

Another way to express opinion in Russian is based on the use of a sub-culture language, Padonky. This sociolect has distinctive phonetic and lexical features that are distant from 'standard' Russian (both official and colloquial). For example, a phrase usually used to express a negative attitude to an author about his book:

- *Аффтор, выпей ЙАДУ*
  (lit) Autor, drink some POIZON

Padonky is close to some variants of slang (corresponding in English to expressions such as *u woz*, *c u soon* etc.), however it is more consistent and is used quite often on the Web.

Sentence Level

Sentence-level means of expressing sentiment (mostly exclamatory clauses, imperatives or rhetorical questions) is slightly more frequent in the Russian corpus than in the English: 892 and 799 respectively. The distribution of positive and negative sentiments realised at the sentence level is opposite in the two corpora: syntactic means are used more frequently in negative reviews in Russian but they are more frequent in positive reviews in English.

One particularly common sentiment-relevant sentence-level phenomenon is the rhetorical question. This is a question only in form, since it usually expresses a statement. For example:

- *Иоткуда столько восторженных отзывов? Коробит от крутости главных героев*
  Why are there so many appreciative reviews? The 'coolness' of the main characters makes me sick

- *Что же такого пил/принимал/нюхал автор, чтобы написать такое?*
  What did the author drink / eat / sniff to write stuff like that?
- *Интересно, кто-нибудь дотянул хотя бы до середины? Лично я - нет.*
  I wonder if anyone managed to get to the middle? I failed.

Considering imperatives, the review author is telling their audience 'what to do', which is often to read a book or to avoid doing so.

- *Run away! Run away!*
- *Pick up any Pratchett novel with Rincewind and re-read it rather than buying this one*
- *Читать однозначно.*
  Definitely should read.
- *Читать !!!!!!!!!! ВСЕМ*
  Read!!!!!!!! EVERYONE

Another way of expressing sentiment through syntactic structure is by means of exclamatory clauses, which are, by their very nature, affective. This type of sentence is widely represented in both corpora.

- *It certainly leaves you hungering for more!*
- *Buy at your peril. Mine's in the bin!*

Discourse Level

Some means of sentiment expression are quite complex and difficult to analyse automatically:

- *Иэто автор вычислителя и лемминго? ... НЕ ВЕРЮ! Садись, Громов, два.*
  (lit) So this author calculator and lemmings? ... (DO)NOT BELIEVE! sit, gromov, two.
  So is this the author of The Calculator and of The Lemmings? ... Can't believe it! Sit down, Gromov, mark 'D'!

This short review of a new book by Gromov, the author of the popular novels *The Calculator* and *The Lemmings*, consists of a rhetorical question, an exclamatory phrase and an imperative. All of these means of expression are difficult to process. Even the explicit appraisal expressed by utilising a secondary school grade system is problematic as it requires specialised real-word knowledge about the meaning of the numeral 'two'[2] in this context.

The example below also features an imperative sentence that is used to express negative sentiment. This review also lacks any explicit sentiment markers. The negative appraisal is expressed by the verbs *stab* and *burn* which only in this context show a negative attitude.

- *Stab the book and burn it!*

[2] Russian schools use a 5-point marking system, with 5 as the highest mark. Thus a '2' can be considered as equivalent to a 'D'.

Discussion

The reviews in English and in Russian often use different means of expressing sentiment, many of which are difficult (if at all possible) to process automatically. Often opinions are described through adjectives (86% of reviews contain adjectives). The second most frequent way of expressing sentiment is through verbs (59% of reviews have sentiment-bearing verbs). Less frequent is the noun, in 39% of reviews. Sentence-level and discourse-level sentiment phenomena are found in 56% of reviews. 3% of reviews contain phonetic sentiment phenomena.

Issues that may Affect Sentiment Analysis

One of the features of web content not mentioned above is a high level of mistakes and typos. Sometimes authors do not observe the standard rules on purpose (for example using sociolects, as outlined above). For example, in the corpora 52% of all documents contain spelling mistakes in words that have sentiment-related meaning. The English corpus is less affected as authors do not often change spelling on purpose and use contractions that have already become conventional (e.g., *wanna, gonna* and *u*). However, the number of spelling mistakes is still high: 48% of reviews contain mistakes in sentiment-bearing words. The proportion of misspelled words in the Russian corpus is higher, at 58%.

Of course, a spelling error is not always fatal for automatic sentiment classification of a document, since reviews usually have more sentiment indicators than just one word. However, as many as 8% of the reviews in both corpora have all of their sentiment-bearing words misspelled. This would pose severe difficulties for automatic sentiment classification.

Another obstacle that makes sentiment analysis difficult is topic shift, in which the majority of a review describes a different object and compares it to the item under review. The negative review below is an example of this:

- *Дочитала с трудом. Ничего интересного с точки зрения информации. Образец интеллектуального детектива – романы У.Эко. И читать приятно, и глубина философии, и в историческом плане познавательно. А в эстетическом отношении вообще выше всяких похвал.*
  Hardly managed to read to the end. Nothing interesting from the point of view of information. An example of intellectual detective stories are novels by U.Eko. It's a pleasure to read them, and (they have) deep philosophy, and are quite informative from the point of view of history. And as for aesthetics it's just beyond praise.

The novel being reviewed is not the one being described, and all the praise goes to novels by another author. None of the positive vocabulary has anything to do with the overall sentiment of the review's author towards the book under review.

Other reviews that are difficult to classify are those that describe some positive or negative aspects of a reviewed item, but in the end give an overall sentiment of the opposite direction. Consider the following positive review:

- *Сюжет довольно обычен, язык изложения прост до безобразия. Много грязи, много крови и смерти. Слишком реально для сказки коей является фэнтези. Но иногда такие книги читать полезно, ибо они описывают неприглядную реальность.*

The plot is quite usual, the language is wickedly simple. A lot of filth, a lot of blood and death. Too true-to-life for a fairy-tale, which a fantasy genre actually is. But it is useful to read such books from time to time, as they depict ugly reality.

The large number of negative lexical units may mislead an automatic classifier to a conclusion that the review is negative.

The three issues described above are present in approximately one-third of all reviews in the corpora. This suggests that a sentiment classifier using words as features could only correctly classify around 55–60% of all reviews.

This performance may be even worse for the Russian corpus since many of its reviews feature very unexpected ways of expressing opinion. Unlike most of the English reviews, in which a reviewer simply gives a positive or negative appraisal of a book, backing it with some reasoning and probably providing some description and analysis of the plot, Russian reviews often contain irony, jokes, and use non-standard words and phrases, making use of a variety of language tools, as illustrated in the following examples:

- *Скушнаа. дошёл до бегства ГГ в мир Януса, и внезапно понял (я), что гори он (ГГ) хоть синим пламенем*
  Booorin'. got to the (episode of) GG fleeing to the world of Janus, and suddenly (I) realised that let it (GG) burn with blue flames (I do not at all care about GG)
- *Я эту муть не покупал. Shift+del.*
  I didn't buy this garbage. Shift+del.

Since there are more reviews of this kind in the Russian corpus than in the English, it is very likely that a Russian sentiment classifier would have lower accuracy.

Sentiment Classification Experiments

In this section we apply supervised sentiment classification techniques to the English and Russian corpora to quantify the impact on accuracy of the language-specific issues discussed above.

Feature Extraction

Approaches to sentiment classification of documents using machine learning require a set of features to be extracted from each document. Most work on English uses word forms as the features, tokenised by splitting the character stream at whitespace and punctuation characters (e.g., Pang et al., 2002).

An alternative approach is to use 'lexical units' as features, where a lexical unit is any commonly-occurring sequence of characters, which may constitute a part of a word, a complete word or even a phrase. This approach avoids the need for word segmentation, and can also capture some grammatical and syntactic information, because lexical units can incorporate function words and parts of grammatical constructions. We extracted lexical units in a pre-processing step by finding the longest strings occurring at least twice in the corpus.

The English book review corpus produced 7,913 such lexical units. Some of these are word sequences expressing features that are often discussed by reviewers,

Table5: Classification results (10-fold cross-validation, words)

	NBm			SVM		
	Precision	Recall	F-measure	Precision	Recall	F-measure
English	0.85	0.85	0.85	0.83	0.83	0.83
Russian	0.78	0.78	0.78	0.73	0.73	0.73

such as *the plot* or *the characters*, as well as phrases that are used for appraisal such as *good performance* and *best performance*.

The same approach was applied to the Russian corpus; despite the language's complex morphology one might expect the technique to be able to capture more unchangeable (stable) units as well as frequent word forms. This indeed turns out to be the case, since the approach extracts some 'semi-stemmed' forms that comprise the most important part of the word, leaving out affixes denoting minor grammatical features, for example, the lexical unit *бессмыленн* which is a common part of the word forms *бессмыленный, бессмыленная, бессмыленных, бессмыленного* and many others meaning *senseless*. The Russian corpus produced 8,372 lexical units.

Results

We used two machine learning algorithms, Naïve Bayes multinomial (NBm) and Support Vector Machines (SVM)[3], trained and evaluated on the corpora of English and Russian, and the two techniques for feature extraction (word forms and lexical units). The evaluations used 10-fold cross-validation.

With word forms as features, in order to make the resulting lexicons comparable in terms of their elements' frequencies we filtered out all words that occurred less than 10 times. We extracted all words from the corpora but did not process them in any way. 1,075 words were extracted from the Russian corpus and 1,247 words from the English book reviews. The classification results are shown in Table 5. The results for Russian are much worse than for English, which might be expected since the abundance of word forms in Russian makes the data sparse.

We also ran the same machine learning algorithms with lexical units extracted from the two corpora as features. The results are shown in Table 6. It could be expected that the 'semi-stemming' property of lexical units would even out differences in accuracy due to different levels of morphological productivity in the two languages. Indeed, the accuracies for Russian are much improved over using word forms as features. Nevertheless, the accuracies for Russian are still lower than for English; this might be explained by the apparently more diverse means of expressing opinion in the Russian corpus than the English one, as discussed above.

Conclusions

In this paper we presented comparable corpora of English and Russian book reviews, examined language-specific features of the reviews that are relevant to senti-

[3] We used WEKA 3.4.11 (`http://www.cs.waikato.ac.nz/~ml/weka`)

Table6: Classification results (10-fold cross-validation, lexical units).

	NBm			SVM		
	Precision	Recall	F-measure	Precision	Recall	F-measure
English	0.88	0.88	0.88	0.84	0.84	0.84
Russian	0.81	0.81	0.81	0.78	0.78	0.78

ment classification, and showed that sentiment in different languages is expressed in slightly different ways, covering all levels of the language: from phonetic to discourse.

We also considered features of the languages themselves; in particular, the complex morphology of Russian may affect the performance of a supervised classifier that does not use any pre-processing techniques, such as stemming or lemmatisation. However, an approach based on identifying common 'lexical units' in a pre-processing step performed much better on the Russian corpus compared to using words as features.

We also found significant differences in sentiment classification accuracy between English and Russian, despite using comparable corpora for training and testing. We conclude that more work is needed to determine the best approach to sentiment analysis for different languages.

References

Akimova, T., Maslennikova, A. (1987): *Lingvisticheskie issledovanija*. Moscow : [s. n.], pp. 3–33. Chapter: Imperativa i Ocenka (Semantics of Imperatives and Appraisal).

Grimes, S. (2010): The Three Secrets to Successful Sentiment Analysis [on-line]. [cit. 2011-09-13]. Retrieved February 16, 2010. Available at: `http://www.mycustomer.com/topic/customer-intelligence/seth-grimes-how-get-sentiment-analysis-right/103102`.

Mihalcea, R., Banea, C., Wiebe, J. (2007): Learning multilingual subjective language via cross-lingual projections. In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic : Association of Computational Linguistics, pp. 976–983.

Pang, B., Lee, L., Vaithyanathan, S. (2002): Thumbs up?: sentiment classification using machine learning techniques. In: *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing (Vol. 10)*. Stroudsburg (Philadelphia, USA) : Association for Computational Linguistics, pp. 79–86.

Pang, B., Lee, L. (2004): A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: *Proceedings of the 42nd Annual Meeting of the Association of Computational Linguistics*. Stroudsburg (Philadelphia, USA) : Association for Computational Linguistics, pp. 271–278.

Seki, Y. (2008): A multilingual polarity classification method using multi-label classification technique based on corpus analysis. In: *Proceedings of the NTCIR-7 MOAT Workshop Meeting*. Tokyo (Japan) : National Institute of Informatics, pp. 284–291.

Wiebe, J., Wilson, T. A., Cardie, C. (2005): Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2), pp. 165–210.

Zagibalov, T., Carroll, J. (2008): Almost unsupervised cross language opinion analysis at NTCIR-7. In: *Proceedings of the NTCIR-7 MOAT Workshop Meeting*. Tokyo (Japan) : National Institute of Informatics, pp. 204–209.

Zagibalov, T., Carroll, J. (2009): Multilingual opinion holder and target extraction using knowledge-poor techniques [on-line]. [cit. 2011-09-13]. Available at: `http://www.informatics.sussex.ac.uk/research/groups/nlp/carroll/papers/ltc09.pdf`.

Acknowledgement

The first author was supported by the Ford Foundation International Fellowships Program.