# Quotation Retrieval System for Bulgarian Media Content

**Ivelina Stoyanova, Martin Yalamov, Svetla Koeva**
Department of Computational Linguistics
Institute for Bulgarian Language, Bulgarian Academy of Sciences
{iva,martin,svetla}@dcl.bas.bg

## Abstract

This paper presents a method for automatic retrieval and attribution of quotations from media texts in Bulgarian. It involves recognition of report verbs (including their analytical forms) and syntactic patterns introducing quotations, as well as source attribution of the quote by identification of personal names, descriptors, and anaphora.

The method is implemented in a fully-functional online system which offers a live service processing media content and extracting quotations on a daily basis. The system collects and processes written news texts from six Bulgarian media websites. The results are presented in a structured way with description, as well as sorting and filtering functionalities which facilitate the monitoring and analysis of media content.

The method has been applied to extract quotations from English texts as well and can be adapted to work with other languages, provided that the respective language specific resources are supplied.

## 1. Introduction

In the age of digital technologies the daily amount of information made available on the internet has increased significantly. That is why information extraction, media monitoring, and opinion mining have become the focus of active research in NLP.

Retrieval of quotes from media content and identifying their author can be important for analysing the behaviour of various actors in the political or social life. This can help provide context for actions, events or statements, clarify the standing of certain figures regarding topics or issues, make a comparison of opinions. Research in this area has applications in social sciences, political sciences, journalism, etc.

There are three types of quotes – direct (literal presentation of someone's words), indirect (paraphrased speech) and mixed (where part of the statement is presented directly, while another part is paraphrased). They exhibit different features – word order, punctuation, grammatical dependencies (e.g., use of particular verb tense, voice and evidentiality forms), some of which are language specific (e.g., use of punctuation for subordinate clauses).

Although it may look like a trivial task, simple approaches for quotation retrieval do not perform particularly well and need improving. Actually, the task of quotation retrieval involves subtasks that still pose challenges to NLP, such as named entity (NE) recognition, including multiword NEs, anaphora resolution, syntactic parsing. Information retrieval and structured presentation of extracted information is also essential to ensure applicability of results for various research purposes.

The paper presents a system for automatic quotation retrieval from media content in Bulgarian. Our purpose is three-fold: (a) to elaborate on the practical aspects of quotation retrieval and attribution; (b) to offer a meaningful, structured representation of quotes which facilitates analysis of media content; and (c) to offer a live service which processes media content and extracts quotations.

In section 2. we discuss related work in the field. Section 3. presents in detail the features of quotation description and the method for quotation retrieval and attribution. The following section 4. is focused on the implementation of the online service for quotation retrieval and the structured representation of results. Section 5. shows some directions for extending the description of quotations by information extraction. The paper concludes by outlining some directions for future work.

## 2. Related Work

Approaches to quotation recognition vary in terms of: (a) coverage (direct, indirect and/or mixed quotations); (b) techniques (syntactic patterns, heuristics, machine learning); (c) applications (whether they were theoretical or have been implemented in a fully functional system).

Pouliquen et al. (2007) present the system `NewsExplorer` that extracts quotations from multilingual news, their author, as well as named entities occurring in the quote. The system also recognises variants of personal names.

Sagot et al. (2010) describes a corpus-based approach to quotation extraction based on the study of quotation verbs, their features and sentential categorisation frames. Krestel et al. (2008) developed a quotation extraction and attribution system that combines a lexicon of reporting verbs and a manually constructed grammar to detect specific constructions satisfying lexical constraints. Similarly, de La Clergerie et al. (2011) employ syntactic patterns to identify quotes in French news texts.

Sarmento and Nunes (2009) present an online service `verbatim` working on data from the Portuguese mainstream media. The authors outline some generic tasks: data acquisition and parsing, quotes extraction, removal of duplicates, topic distillation and classification, and interface design for presentation and navigation. They apply syntactic patterns on text to identify and attribute quotes to a speaker.

Schneider et al. (2010) present a system called `PICTOR`, that queries a large news corpus for topical quotations and then visualises them over time. Alongside identification of quotes and speakers in an article, authors select quotes relevant to a user query, scoring quote similarity in order to filter and cluster related quotes, and present a graph-based visualisation for plotting relevant quotes over time.

Atteveldt (2013) uses syntactic analysis and topic models to identify quotations from politicians. His method relies substantially on lexical resources. The author uses a dictionary to identify the sources (the person who is being quoted), and a list of verbs (e.g., *say, state*) and attribution phrases (e.g., *according to*).

Pareti et al. (2013) note the low portion of direct quotes (30-52% in the corpora they use) and focus on extraction of indirect and mixed quotes as well. They report on the results and evaluation of the extraction and attribution of direct, indirect and mixed quotations over two large news corpora.

Machine learning approaches for quotation retrieval have been suggested by Fernandes et al. (2011), O'Keefe et al. (2012), Pareti et al. (2013).

More often quotation retrieval is implemented as part of a more complex task, such as opinion mining and sentiment analysis (O'Keefe et al., 2013), or comparative analysis of political statements (Atteveldt, 2013).

Based on the review of related works we note several possible directions in which quotation retrieval can be further extended: (a) to develop fully functional retrieval systems on media content rather than applications for purely research purposes; (b) to perform analysis on dynamic media content on a daily basis rather than a fixed text corpus; and (c) to provide efficient description of quotations with filtering and sorting functionalities. Still, not many quotation retrieval systems are available online and on live media content. To the best of our knowledge, no system for quotation retrieval exists for Bulgarian.

## 3. Quotation Retrieval

### 3.1. Outline of the Task

The main task includes identification of the quotations and their attribution to a source. Here we cover direct and indirect quotes, while the (direct and indirect) components of mixed quotes are handled as separate entities. Since in the presentation of results quotations from the same news text and attributed

to the same person are grouped together, treatment of mixed quotes in this way does not affect their information representation.

Pareti (2012) and Pareti et al. (2013) define a quotation attribution relation by four components: a source span (the entity the content is attributed to); a cue span (the lexical anchor of the relation, e.g. a report verb); a content span (the quoted text); and a supplement span (any additional elements relevant to the interpretation of the attribution relation).

Taking into account the features above, we extend and organise the description of a quotation into the following sets of features: structural features used to extract the quotation, informational content features which characterise the content of the quotation, and external, or metadata-based, features which provide editorial information about the text the quotation appears in.

1. **Structural features**

   - **Span.** The quote can be contained within a sentence (Example 1a), or span over several sentences (Example 1b).
   - **Syntactic patterns and lexical elements.** In most cases the quotation is introduced by a reporting verb and a specific syntactic pattern. Indirect speech is also marked by subordinate conjunctions and linking words.
   - **Punctuation.** Punctuation is an essential feature for quotation retrieval. Direct quotes in Bulgarian (and many other languages) are introduced by colons and/or surrounded by quotation marks, or (rarely in news) introduced by a dash on a new line. Indirect speech is usually expressed as a subordinate (object) clause within the sentence without any distinctive punctuation.
   - **Source.** A quote is attributed to a speaker, who can be represented in the text by his name (e.g., *Boyko Borisov, Borisov*, Examples 2a and 2b), a descriptor (e.g., *the prime minister*, Example 2c), or an anaphora (e.g., *he*). In some cases the source can be an organisation or group presented by its name (e.g., *Bulgarian Socialist Party*) or an abbreviation (e.g., *BSP*, Example 2d).

2. **Informational content features**

   The content features include essential elements characterising the informational content and the topic of the quoted text. See Section 5. for more details on the techniques used for information extraction. Since in many cases the quoted text is short, it is not always possible to detect a particular topic in it. The content features include:

   - **Named entities of persons, places, and organisations within the quoted text.**
   - **Temporal expressions in the quoted text.** We identify dates and times which can be used to describe the topic of the quotation and to find relations with other quotations.
   - **Keywords** which relate to the overall content of the news text or are significant for describing the content of the quoted text.
   - **Opinion and sentiment features.** The reporting verb often reflects evaluation of the quote's content made by the author of the text, i.e. external for the quote itself, which is crucial for the analysis of its content. This can be expressed lexically, for example using verbs for negation (*The prime minister denied that ...*) or modal verbs (*The Bulgarian Socialist Party should state that it wants to get the power*), or morphologically using negative forms (*The prime minister did not state that ...*) or conditional mood (*The prime minister could have said: "..."*).

3. **External (metadata-based) features**

   - **Publication time of the news text.** This is the date (and possibly time) of the publication. It is useful in order to enable filtering or sorting by time, or creating a timeline for the quotations on a given topic.

- **Media source.** The media source is an essential part of the description of the quote. It allows to filter by source or to follow how various topics or events are presented across different media.
- **URL to the publication.** The publication can provide context of the quotation and it is a way to avoid copyright issues by (a) linking to the source, and (b) not publishing the whole text or large excerpts from it.

**Example 1.**
(a)
*Gunlaugson tvardi,* [*che zakonite ne sa narusheni i saprugata mu ne se e oblagodetelstvala finansovo.*]
*Gunlaugston states* [*that laws have not been broken and his wife has not benefited financially.*]
(b)
[*"Tova beshe edin dostoen mach za final. Tryabva da prodalzhim da se razvivame kato tehnika",*] *zayavi trenyorat Miroslav Zhivkov sled dvuboya.*
[*"This was a decent match for the final. We should continue to develop our technique",*] *said coach Miroslav Zhivkov after the game.*

**Example 2.**
(a)
*"Tolkova parkove i gradini se napraviha v Sofia", kaza oshte* ***Boyko Borisov****.*
*"So many parks and gardens have been built in Sofia", added* ***Boyko Borisov****.*
(b)
***Borisov*** *potvarzhdava, che Balgariya shte podkrepya evropeyskata perspektiva na Sarbiya.*
***Borisov*** *confirms that Bulgaria will support the European prospects of Serbia.*
(c)
*V profila si vav Facebook* ***premierat*** *napisa: "Edna naistina otlichna vecher za balgarskiya sport."*
*On Facebook the* ***prime minister*** *has posted: "One really excellent evening for Bulgarian sport."*
(d)
*Ot* ***BSP*** *zayaviha, che partiyata ne e saglasna s proekta za koalitsionno sporazumenie.*
*From* ***BSP*** *announced that the party does not agree with the proposal for a coalition agreement.*

### 3.2. Method for Quotation Retrieval and Attribution

The method for quotation retrieval relies on the following language specific resources for Bulgarian: dictionary of verbs used for reporting speech; list of patterns defining the analytical verb forms in order to identify the form of the reported verb and its tense, voice and mood (Leseva et al., 2015); dictionary of correspondences between names and titles or descriptors (e.g., *Boyko Borisov* and *prime minister*).

Initially, texts are annotated with POS and lemma. Taking into account the free word order in Bulgarian, the implementation of rigid syntactic patterns is not efficient. Instead, similarly to Pouliquen et al. (2007), we identify each quote as a triple of quoted text, reporting verb, and source (person) with the restriction that the verb and the source are both either on the left or on the right of the quoted text.

We perform pattern matching to identify the quoted text, as well as the source and the reporting verb. In direct quotations, the quoted text is introduced by punctuation (quotation marks, colons, dash, new line) and can span over several sentences. Indirect quotations are found within a sentence and are identified as subordinate clauses introduced by a report verb, subordinate conjunction and/or punctuation.

In Bulgarian, most NEs and more specifically, names of persons and organisation, are tagged by the POS and grammatical tagger and lemmatiser. Additional rules for identification of NEs were manually crafted. A sequence of single word personal NEs (e.g., first name followed by a surname) are combined and annotated as one NE. Special check is performed in a dictionary of categorised NEs from Wikipedia in order to separate geographic or organisational NEs from adjacent personal name.

A dictionary of correspondences between names and descriptors (such as titles, job posts, etc.) in Bulgarian has been automatically compiled from Wikipedia. Currently, it includes 31,446 personal names with a corresponding set of descriptors. The names include popular Bulgarian and foreign politicians, artists, sportsmen and public figures. All names and descriptors are matched to a canonical form,

usually the full name (e.g., *Borisov* is matched to *Boyko Borisov*). To avoid mismatches, the canonical form should occur at least once in the same text.

We apply a set of simple rules for anaphora resolution which cover only a selected number of cases in order to improve the recall of the method. We first identify third person singular pronouns in nominative (*toy – he*, and *tya – she*) which immediately precede the reporting verb. The attempt to resolve them includes looking backwards in the current and the previous sentence for a noun, including personal names, which agrees in gender and number with the anaphora. It is resolved only if the first agreeing noun is a NE. If the anaphora is matched with a common noun before reaching a NE, the anaphora is regarded as unresolved.

A dictionary of 114 reporting verbs in Bulgarian is used to identify the quotations in the text. The dictionary is extracted from the Bulgarian wordnet[1] by exploiting the semantic relations of synonymity and hypernymity – all synonyms and hyponyms of the synsets containing *govorya* (*speak*).

Based on the distances measured in number of tokens between any pair of the triple (quoted text Q – report verb V – potential source S), we evaluate a simple confidence measure for the validity of the retrieved quotation where the confidence ($C$) is reduced for any extra position separating any pair of the three components:

$$C(Q,S,V) = 0.99 - \frac{d(Q,V) + d(Q,S) + d(S,V) - 1}{3} \times 0.07$$

A set of 'penalties' is also introduced to adjust the score in some specific cases. They are applied in the following order:

- If the identified source (NE, descriptor, anaphora) and the reporting verb are on different sides of the quoted text, the score is reduced by 70%. This effectively excludes such cases.
- If the reporting verb and the source precede the quoted text, and the reporting verb is in active voice and precedes the source, the score is reduced by 30%.
- If the reporting verb is in active voice and the source (including any adjectives in front of it if it is a descriptor noun) is preceded by a preposition, the score is reduced by 20%.
- If the reporting verb is in passive voice and the source (including any adjectives in front of it) is not preceded by the preposition *ot* (*by*), the score is reduced by 20%.

The score is used for filtering out direct quotations attributed to the wrong source. The score is also applied to rank possible triples from the same sentence and select the most reliable from conflicting quotations. Example 3 shows the scores for three possible attributions of the indirect quotation in the sentence, the first attribution is disregarded as it is below the threshold of $0.5$, and the attribution with the higher score (*Emil Radev*) is selected.

**Example 3.**
*Po povod kandidata na GERB i dumite na Boyko Borisov evrodeputatat Emil Radev* $_V$*komentira,* $[_Q$ *che tryabva da se promenyat pravilata za izdigane i izbor na prezident.*$]$
*With respect to the candidate of GERB and the words of Boyko Borisov, the European MP Emil Radev* $_V$*commented* $[_Q$ *that the rules for president nominations and elections should be changed.*$]$

$$C(Q, \text{GERB, V}) = 0.4000, \quad C(Q, \text{Boyko Borisov, V}) = 0.5867, \quad C(Q, \text{Emil Radev, V}) = 0.9207$$

The method is applied on Bulgarian media content collected from six major news websites. On average, daily about 3,200 potential quotations are identified, which are further filtered based on: (a) attribution to a named source – we exclude quotations that cannot be matched to named entities directly (a name is identified in the sentence) or indirectly (a descriptor or anaphora is identified in the sentence

[1] http://dcl.bas.bg/bulnet/

which is matched to a name); and (b) confidence score – we set a threshold of 0.5 for both direct and indirect quotes. Further, in the presentation of results we combine separate quotations, both direct and indirect, attributed to the same source within a single text (see Section 4.2.).

### 3.3. Evaluation

The evaluation of the method is based on a manually verified set of 200 quotations (79 direct and 121 indirect). We evaluate the precision and recall of discovering the full quotations (both boundaries) or only the start of the quotation. The evaluation of source attribution is performed on all identified quotations and includes NEs, descriptors and anaphoras. Only fully recognised names and matches to NEs are considered as correct. We perform experiments with different confidence thresholds, the results of which are presented in Table 1.

| **Type** | **Confidence threshold** | **Full quotation** | | **Start of quotation** | | **Source attribution** | |
|---|---|---|---|---|---|---|---|
| | | **Precision** | **Recall** | **Precision** | **Recall** | **Precision** | **Recall** |
| Direct | 0.3 | 0.97 | 0.77 | 1.00 | 0.80 | 0.94 | 0.73 |
| | 0.5 | 1.00 | 0.63 | 1.00 | 0.63 | 0.97 | 0.65 |
| | 0.7 | 1.00 | 0.53 | 1.00 | 0.53 | 1.00 | 0.60 |
| Indirect | 0.3 | 0.81 | 0.58 | 0.89 | 0.66 | 0.82 | 0.68 |
| | 0.5 | 0.88 | 0.55 | 0.89 | 0.56 | 0.83 | 0.62 |
| | 0.7 | 0.90 | 0.50 | 0.92 | 0.51 | 0.87 | 0.61 |

Table 1: Evaluation of the results (precision and recall) in terms of: (a) the full quotation, (b) the identification of the start of the quotation, and (c) source attribution.

## 4. Online System for Quotation Retrieval

### 4.1. Workflow

The online quotation retrieval system is part of a complex system for collection and analysis of media content in Bulgarian. The results are available at `http://dcl.bas.bg/quotations/` (Figure 1). The workflow includes the following components:

1. **Download of texts from several news agencies**. Two approaches were implemented: (a) monitoring of RSS feeds; or (b) focused crawling with pre-crawl data mining. Metadata are extracted from the original webpage and stored separately from the text according to the principles of the Bulgarian National Corpus (Koeva et al., 2012).

2. **Processing and linguistic annotation** on Bulgarian texts was performed using the Bulgarian Language Processing Chain (Koeva and Genov, 2011) through a RESTful service. Downloaded and processed texts are added to the Bulgarian National Corpus and can be used for other applications, such as neologism detection[2].

3. **Quotation retrieval and text analysis** to describe quotation features as outlined in Section 3. The application for quotation retrieval is implemented in Java 7.

4. **Presentation of results**. The results are represented online in a structured manner and with a search and sorting functionality.

5. **Update routine**. Results are automatically updated on regular intervals throughout the day after newly downloaded data have been processed.

[2]http://dcl.bas.bg/neologisms/

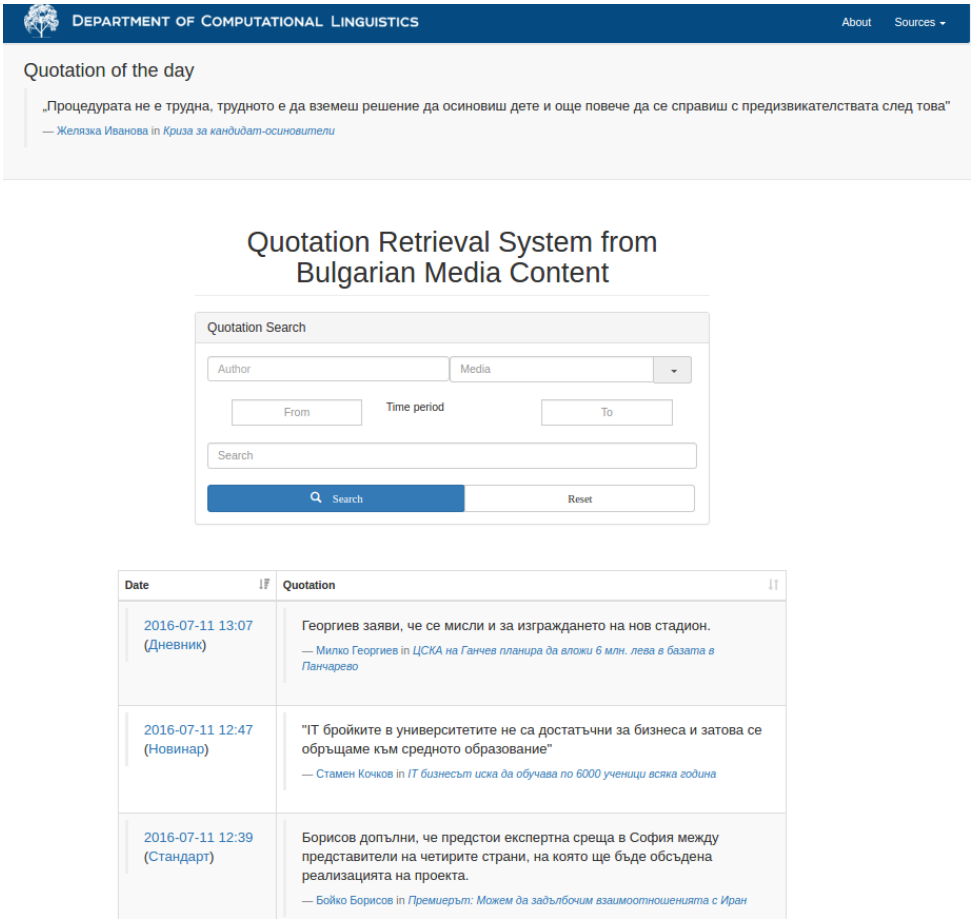


Figure 1: Results displayed online

### 4.2. Structured presentation of results

Retrieved quotations are put into a database and are presented online in a structured manner to facilitate their viewing and analysis. Each quote is presented with the following information: quoted text, source and link to the original news article. Quotations attributed to the same source within a text are combined together.

Quotations can be sorted by date of the news articles. There is also a searching and filtering functionality based on: (a) source – name of the person; (b) period of time; (c) media; and (d) query words within the quoted text. Each field has an autocomplete dropdown list which shows possible values and updates upon typing.

Further, we offer a 'Quote of the day' on a selected popular topic (e.g., on 5 April the most frequent topic was Panama papers). The selected quote needs to satisfy the following conditions: (i) to contain as many as possible of the top 10 most frequent keywords discovered within all texts of the day; and (ii) to have high confidence measure (above 0.9, or the highest available) in order to ensure that it is correctly identified and attributed.

## 5. Towards Topic Detection

In recent years topic modelling is gaining popularity as a way to discover and represent the abstract topics in a collection of documents, including in conversational texts such as emails and social media posts (Carenini et al., 2011) and news (Blei, 2012). Various well developed approaches have been applied, such as Latent Semantic Analysis or Latent Dirichlet Allocation. Recently, neural networks have been employed for the task of topic modelling (Mikolov and Zweig, 2012). Toolkits for topic modelling have also been developed, e.g. MALLET (Graham et al., 2012) or Stanford Topic Modelling Toolkit (Ramage

et al., 2009).

Here we perform the first steps towards topic modelling by identifying significant components within the set of quotes by the same source within a single text document. The following elements are extracted from the quoted text: (a) named entities of persons, places, and organisations; (b) temporal expressions; and (c) a set of keywords. Basically, we answer the set of questions *who, what, when, where* and define the topic in a very narrow sense. The list of identified elements can include proper names and temporal words (e.g., *Theresa May, Boris Johnson, London, Brexit, Great Britain, Wednesday*), as well as concrete and abstract nouns (*borders, minister, foreign affairs, politics*).

For NE recognition we use the same module applied for quotation attribution (see Section 3.2.). While in attribution we are only interested in named entities of persons or organisations (to whom quotations can be attributed), here we also identify geographical entities and event names. Categorisation of NEs is performed using a dictionary of NEs derived from Wikipedia and other sources divided into semantic categories – personal names, organisations, places, and events (Koeva et al., 2016).

Temporal expressions include dates (*14 July*, *14/07/2016*, etc.), time (e.g., *18:00*), concrete or relative temporal expressions (e.g., *on Tuesday, in April, yesterday, last year*). Temporal relations can also be expressed morphologically (e.g., by the verb form). So far we only consider explicit dates and time.

Keywords extraction on the quoted text is based on: (i) predefined dictionary of 139 domain-specific words which point to a domain (e.g., budget – Economy; parliament – Politics); and (ii) frequency analysis (words, except stop words, with frequency above a threshold are identified as keywords). The dictionary in (i) is applicable in the cases of short texts where frequency analysis is not informative.

The topic detection is essential for providing more functionalities in the online system for quotation retrieval in terms of grouping of results, finding quotations about the same or similar topic, or discovering relations between quotation from different media sources.

## 6. Future Work

The method for quotation retrieval has been also applied on English news texts collected through the BBC RSS feed and annotated using Stanford CoreNLP (Manning et al., 2014). We compile an English dictionary of reporting verbs containing 43 unique verbs derived from the Princeton WordNet in a similar way to that of the Bulgarian reporting verbs (see Section 3.2.). Essentially, the methods for NE recognition are the same with the use of some language specific resources such as the dictionary of English NEs from Wikipedia. The same patterns for matching quotations are applied.

In the future our efforts will be focused on improving the method for quotation retrieval and its results. At present, quotation attribution is performed only for named sources, i.e. either labelled as or matched directly to NEs. However, these depend on the quality of the modules for anaphora resolution and the coverage of the dictionaries matching descriptors to NEs. Moreover, we are looking into ways to establish more matches (e.g., based on previous occurrences in media texts) and to increase significantly the recall of the system. Machine learning methods also look promising for the purposes of quotation identification and source attribution.

Furthermore, we could use information about whether the report verb is a marker of opinionated content and of what polarity (Esuli and Sebastiani, 2006). Some of these verbs are neutral (e.g., *say, tell, explain*) while others express opinion about particular features of the quotation such as its truth value (e.g., *deny*) or importance and validity (e.g., *emphasise, hint*). SentiWordNet (Baccianella et al., 2010) is used to obtain the positivity and the negativity scores of verbs for the purposes of opinion mining and sentiment analysis. The analysis based on the verb semantic features falls outside of the scope of the present study. Here we use the reporting verb purely as a lexical marker introducing the quotation.

The work on the online system for quotation retrieval is ongoing. Our aim is to cover more web sources and possibly extend the data beyond news and media domain. Finally, improvement in the presentation of results is also among our future tasks – including more information in the quote description, filtering on more features, etc. User feedback will also be valuable in this respect.

## References


Atteveldt, W. V. (2013). Quotes as Data: Extracting Political Statements from Dutch Newspapers. In *New Directions in Analyzing Text as Data Workshop*.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, pages 2200–2204, Valletta, Malta. European Language Resources Association (ELRA).

Blei, D. (2012). Topic Modeling and Digital Humanities. *Journal of Digital Humanities*, 2(1).

Carenini, G., Ng, R., and Murray, G. (2011). *Methods for Mining and Summarizing Text Conversations*. Syntesis Lectures on Data Management.

de La Clergerie, E., Sagot, B., Stern, R., Denis, P., Recource, G., and Mignot, V. (2011). Extracting and Visualizing Quotations from News Wires. In Vetulani, Z., Ed., *Human Language Technology. Challenges for Computer Science and Linguistics*, pages 522–532.

Esuli, A. and Sebastiani, F. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06*, pages 417–422.

Fernandes, W. P. D., Motta, E., and Milidiu, R. L. (2011). Quotation Extraction for Portuguese. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology (STIL 2011)*, pages 204–208, Cuiaba.

Graham, S., Weingart, S., and Milligan, I. (2012). Getting Started with Topic Modeling and MALLET. Programming Historian (02 September 2012), http://programminghistorian.org/lessons/topic-modeling-and-mallet.

Koeva, S. and Genov, A. (2011). Bulgarian Language Processing Chain. In *Proceeding to The Integration of multilingual resources and tools in Web applications Workshop in conjunction with GSCL 2011*. University of Hamburg.

Koeva, S., Stoyanova, I., Leseva, S., Dekova, R., Dimitrova, T., and Tarpomanova, E. (2012). The Bulgarian National Corpus: Theory and Practice in Corpus Design. *Journal of Language Modelling*, 0(1):65–110.

Koeva, S., Stoyanova, I., Todorova, M., and Leseva, S. (2016). Semi-automatic Compilation of the Dictionary of Bulgarian Multiword Expressions. In *Proceedings of the GLOBALEX 2016 Workshop: Lexicographic Resources for Human Language Technology, LREC*, pages 86–95.

Krestel, R., Bergler, S., and Witte, R. (2008). Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles. In *Proceedings of the Sixth International Language Resources and Evaluation*.

Leseva, S., Stoyanova, I., and Koeva, S. (2015). Automatic Recognition of Verb Forms in Bulgarian. In *Paisievi Cheteniya*, Plovdiv.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Mikolov, T. and Zweig, G. (2012). Context Dependent Recurrent Neural Network Language Model. In *Proceedings of the 2012 IEEE Workshop on Spoken Language Technologies*, pages 234–239, Miami, USA, December.

O'Keefe, T., Pareti, S., Curran, J. R., Koprinska, I., and Honnibal, M. (2012). A Sequence Labelling Approach to Quote Attribution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 790–799.

O'Keefe, T., Curran, J. R., Ashwell, P., and Koprinska, I. (2013). An Annotated Corpus of Quoted Opinions in News Articles. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 516–520. ACL.

Pareti, S., O'Keefe, T., Konstas, I., Curran, J. R., and Koprinska, I. (2013). Automatically Detecting and Attributing Indirect Quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 989–999. ACL.

Pareti, S. (2012). A Database of Attribution Relations. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, pages 3213–3217.

Pouliquen, B., Steinberger, R., and Best, C. (2007). Automatic Detection of Quotations in Multilingual News. In *Proceedings of Recent Advances in Natural Language Processing*, pages 487–492.

Ramage, D., Rosen, E., Chuang, J., Manning, C. D., and McFarland, D. A. (2009). Topic Modeling for the Social Sciences. In *Neural Information Processing Systems (NIPS) Workshop on Applications for Topic Models: Text and Beyond*, Whistler, Canada, December.

Sagot, B., Danlos, L., and Stern, R. (2010). A Lexicon of French Quotation Verbs for Automatic Quotation Extraction. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., Eds., *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, Valletta, Malta. European Language Resources Association (ELRA).

Sarmento, L. and Nunes, S. (2009). Automatic Extraction of Quotes and Topics from News Feeds. In *4th Doctoral Symposium on Informatics Engineering*.

Schneider, N., Hwa, R., Gianfortoni, P., Das, D., Heilman, M., Black, A. W., Crabbe, F. L., and Smith, N. A. (2010). *Visualizing Topical Quotations Over Time to Understand News Discourse*. Technical report. T.R. CMU-LTI-10-013, Carnegie Mellon University, Pittsburgh, PA.