Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles

Ralf Krestel,[1] Sabine Bergler,[2] and René Witte[3]

[1]L3S Research Center
Universität Hannover, Germany

[2]Department of Computer Science and Software Engineering
Concordia University, Montréal, Canada

[1]Institut für Programmstrukturen und Datenorganisation (IPD)
Universität Karlsruhe (TH), Germany

Abstract

Reported speech in the form of direct and indirect reported speech is an important indicator of evidentiality in traditional newspaper texts, but also increasingly in the new media that rely heavily on citation and quotation of previous postings, as for instance in blogs or newsgroups. This paper details the basic processing steps for reported speech analysis and reports on performance of an implementation in form of a GATE resource.

1. Introduction

Despite the rapid growth of alternative information sources, newspaper articles continue to be the staple source most readily accessible. Accessibility is, indeed, a major preoccupation for newspaper editors, and most papers are available on-line with "breaking news" features. News aggregators are automatic systems that use on-line information from newsfeeds or on-line newspaper sources and collate overviews across newspapers, worldwide. For example, the European Commission's Joint Research Centre's NewsExplorer[1] clusters news stories by type and country, but also provides additional indexes in form of names, related and associated people, related stories, etc. Names mentioned under *Related People* invoke a screen with links to stories, in which the person plays a role, as well as two special sections: *Quotes from* and *Quotes about* (Pouliquen et al., 2007). This reflects the importance of information attributed to a source.

Newspaper articles typically report information collected from a single (or multiple) source(s). Especially in the North American tradition, this information is usually attributed to the source explicitly in form of *quoted* or *indirect* speech. We will refer to both forms here as *reported speech*.[2]

Reported speech is an important indicator of evidentiality (Bergler, 1992). The reliability of the information conveyed has to be assessed differently for different tasks (an expert on child care and an expert on nuclear physics may have the same high reliability associated when speaking on their respective domains of expertise, yet low reliability when speaking on the other's). Reported speech is thus a form of valence shifter (Bergler, 2005), which marks the embedded information as not simply factual. Reported speech is of increasing importance outside newspaper articles in new media such as blogs and discussion groups, where citations and quotations form a major backbone for the structure of discussion.

NewsExplorer only identifies material in direct quotes, which is mostly of importance for possibly contentious material or claims. But most reported speech is in form of indirect reported speech. We have developed a set of resources that identifies and tags the *source*, *reporting verb* and *content* of reported speech sentences.

These resources have been implemented as components for the GATE framework (Cunningham et al., 2002) and are distributed under an open source license.[3] This is intended as a first module for more sophisticated representation and reasoning with attributed information, such as belief reasoning based on nested belief structures as suggested by (Ballim and Wilks, 1991) and recently illustrated for the reported speech context by the *Fuzzy Believer System* (Krestel et al., 2007a; Krestel et al., 2007b).

2. Reported Speech

The function of reported speech is to convey information in two steps: from a *source* to a *reporter*, and from the *reporter* to a *reader*. The utterance and its context will be interpreted by the reporter, encoded by the reporter for the reader, and decoded by the reader. The reporter can use the mechanism of reported speech to not only reproduce the content of the utterance, but to reproduce and clarify the whole speech act (Austin, 1962; Searle, 1969).

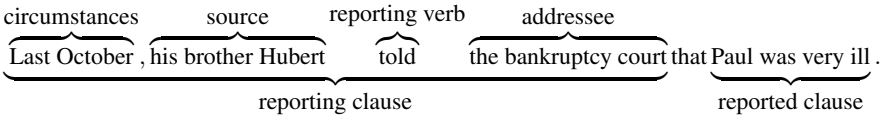
From the reader's point of view, reading a newspaper article is a multilevel process, as illustrated in Figure 2. The reader has to:

1. Understand the content of what is expressed in the article (the reported clause);
2. Evaluate the additional information given by the reporter to reconstruct not only the original utterance but the original speech act (the reporting clause);

[1]EU EMM NewsExplorer, http://press.jrc.it/NewsExplorer/

[2]In contrast to Quirk (Quirk, 1985, page 1021), who considers only indirect speech under the term "reported speech."

[3]See http://semanticsoftware.info



(1)

Figure 1: Example for a reported speech construct and its constituents

3. Interpret the article as presented by the reporter;
4. Reconstruct the original situation; and
5. Interpret the assumed original situation.

This encoding typically takes the form of *reported speech*. The function of *direct* and *indirect* speech is the same, with the distinction that in direct speech the reporter commits to a literal transcription of the original utterance, given in quotes, whereas he gives a summary interpretation when using indirect speech.

Reported speech usually consists of the *reporting clause* and the *reported clause* (Quirk, 1985). The reporting clause contains information about the source of the utterance, the circumstances in which it was made, and possibly a characterization of the manner or force, with which it was made. Figure 1 shows an example from the *Wall Street Journal* 03.03.1988.

The reported clause can consist of direct speech or indirect speech. The same example as in (1) now with direct speech as reported clause:

Last October, his brother Hubert said: "Paul is very ill." (2)

There is also a form of *free direct and indirect speech*. It is used for example to express a stream of consciousness in fictional writing. The reporting clause is omitted in that kind of reported speech. Free form is more widely used in German newspapers, for instance, but is virtually absent in the North American newspaper tradition (Bergler, 1995) and here we will concentrate on direct and indirect speech only.

Direct Speech. Quotation marks usually indicate direct speech. In the domain of newspaper articles, quotation marks are obligatory. An example (from the *Wall Street Journal* 09.14.1987) is:

In a statement yesterday, Towle President Paul Dunphy said, "We look to a closing in the very near future." (3)

The position of the reporting clause can vary: at the beginning of a sentence, as in (3), in the middle, as in (4), or at the end of the sentence, as in (5). When the reporting clause is located within or after the reported clause, subject and verb may reverse positions, as in (4) (from *Wall Street Journal* 12.09.1986).

"Other national advertising will be up 8.7%," { Mr. Coen said, / said Mr. Coen, } "led largely by the strength of such sectors as direct mail." (4)

In this *Wall Street Journal* (12.17.1986) example, the reporting verb is at the end:

"We think this is the bottom year," a Nissan official said. (5)

Direct speech can span over more then one sentence. In that case, the reporting clause is usually found within the first sentence. As a grammatical relation, the direct speech can function as a subordinate clause, for example in (3). But it can also be a subject complement, an apposition to a direct object, or an adverbial construct. A special case is the mixture of direct and indirect speech, where the direct speech forms only part of the reported clause. An example (from the *Wall Street Journal* 03.05.1987) is:

In a televised address, the president concluded that the initiative "was a mistake." (6)

A number of verbs that are frequently used within direct speech can be found in (Quirk, 1985, page 1024). Additionally, verbs indicating the manner of speaking like *mumble*, *mutter*, or *sob* can also be indicators of direct speech.

Indirect Speech. In newspaper articles, indirect speech is ubiquitous. Like direct speech, it relates information from a source, but in a summary form, designed to convey the essence of a larger discourse. In addition, *circumstantial information* that indicates additional features of the context of the original utterance is usually conveyed. Consider this example from the *Wall Street Journal* (02.05.1987), where *as a result* puts the information of the reported clause in the necessary context:

As a result, the company said that it will restate its 1986 earnings. (7)

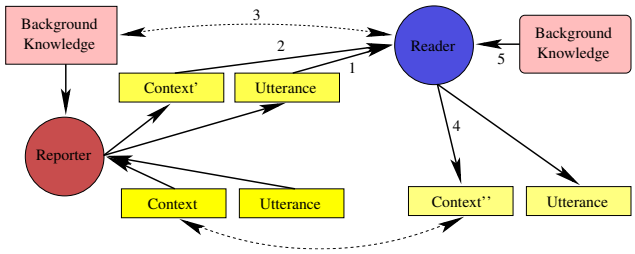


Figure 2: Steps in reported speech analysis

3. Resource Description

Our reported speech analysis resource consists of two interdependent NLP components that have been implemented for the GATE framework (Cunningham et al., 2002):

Reporting Verb Marker: Detects and tags verbs that trigger a reported speech interpretation.

Reported Speech Finder: Finds reported speech constructs and tags its constituents (source, reporting verb, circumstantial information).

according	accuse	acknowledge	add	admit
agree	allege	announce	argue	assert
believe	blame	charge	cite	claim
complain	concede	conclude	confirm	contend
criticize	declare	decline	deny	describe
disagree	disclose	estimate	explain	fear
hope	insist	maintain	mention	note
order	predict	promise	recall	recommend
reply	report	say	state	stress
suggest	tell	testify	think	urge
warn	worry	write	observe	

Table 1: Reporting verbs identified by the reporting verb marker

These components can easily be embedded in more complex NLP pipelines, depending on the concrete application scenario. In turn, they rely on lower-level analysis components, such as a noun phrase chunker and verb grouper, which are distributed with the GATE system (see Section 6.).

3.1. Reporting Verb Marker

The Reporting Verb Marker tags verbs used to express reported speech using a finite state transducer. This component was first developed and implemented by (Doandes, 2003) to extract information related to evidential analysis. Currently, we recognize only the most frequent reported speech verbs as listed in Table 1.

The reporting verb marker is implemented using GATE's *Java Annotation Patterns Engine* (JAPE) (Cunningham et al., 2000). It works with the chunker notion of *verb groups*, contiguous sequences of auxiliaries and verbs. When one of the listed verbs is detected as the head of a verb group, the reporting verb finder marks it as a reported speech verb by adding a corresponding annotation containing the lemma of the reported speech verb.

3.2. Reported Speech Finder

To identify reported speech in newspaper articles, we extract six general patterns. They differ in the position of the reporting verb, the source, and the reporting clause. An overview of those six patterns is shown in Table 2.

Source	Verb	Content
Verb	Source	Content
Content	Source	Verb
Content	Verb	Source
Content	Source Verb	Content
Content	Verb Source	Content

Table 2: Six patterns for finding reported speech in newspaper articles

Identifying reported speech sentences enables us to label the different elements for further analysis. Our components have been designed to allow extracting statements in form of declarative sentences. Currently, we exclude structures where the reported clause is not a grammatical sentence, since infinitival and other omitted constructs no longer report the speech of others, but interpret their actions or utterances, which requires a different treatment; e.g.:

The President denied to sign the bill. (8)

The six patterns are not exhaustive, for example, they will ignore the second source and reporting verb in (*Wall Street Journal* 12.09.1986):

> **Mr. Coen predicted** that a weak sector in 1987 will be national print – newspapers and magazines – which **he said** will see only a 4.8% increase in advertising expenditures. (9)

Constructs that do not fit into our six basic patterns are rare,[4] and additional patterns can be easily added in the future.

4. Implementation

This section contains more details concerning our implementation, which is realized in form of processing resources in the GATE (Cunningham et al., 2002) framework.

4.1. Reporting Verb Marker

A number of verbs used within the reporting verb marker can be found in Table 1. We use a gazetteer to detect these verbs within the original article by comparing the root forms of the verbs with our verb list.

The reporting verb marker is implemented as a JAPE (Cunningham et al., 2000) grammar. Figure 3 shows a code snippet of the JAPE-rule for the verb "concede."[5] After detecting these verbs within a document, the reporting verb marker marks them as reported speech verbs by adding an annotation containing the root of the main verb of the reported speech verb phrase.

4.2. Reported Speech Finder

The reported speech finder is implemented as a regular grammar using the Montréal Transducer, which supports an enhanced version of the JAPE language.[6] Our grammar consists of a set of rules to identify reported speech structures and to label the source, reporting verb, and the reported clause. An example annotation for two sentences from the *Wall Street Journal* (03.19.1987) is given as *reported clause*, source, and **reporting verb**:

[4] Numbers depending on the literary style of the newspaper, around 3%

[5] This rule also specifies the semantic dimensions of the verb, which are not of importance for the remainder of this paper, but are used for further interpretation, see (Doandes, 2003).

[6] Developed by Luc Plamondon, Université de Montréal, see http://www.iro.umontreal.ca/~plamondl/mtltransducer/

WSJ Article	Reported Clause			Source/Verb		
	Precision	Recall	F-measure	Precision	Recall	F-measure
861203-0054	1.00	0.50	0.67	1.00	0.63	0.77
861209-0078	1.00	0.77	0.87	1.00	0.79	0.88
861211-0015	0.97	0.88	0.96	1.00	0.89	0.94
870129-0051	1.00	0.71	0.83	1.00	0.71	0.83
870220-0006	0.96	0.74	0.84	1.00	0.93	0.96
870226-0033	1.00	0.58	0.74	1.00	0.58	0.74
870409-0026	1.00	1.00	1.00	1.00	1.00	1.00

Table 3: Reported speech extraction results for our resource

```
Rule: concede
(
  {VG.voice == "active", VG.MAINVERB == "concede"}
):rvg -->
{
  gate.AnnotationSet rvgAnn = (gate.AnnotationSet)bindings.get("rvg");
  gate.AnnotationSet vgclac = rvgAnn.get("VG");
  gate.Annotation vgclacAnn = (gate.Annotation)vgclac.iterator (). next ();
  gate.FeatureMap features = Factory.newFeatureMap() ;
  features.put("mainverb", vgclacAnn.getFeatures().get("MAINVERB"));
  features.put("semanticDimensions",
    "VOICE:unmarked
     EXPLICITNESS:explicit
     FORMALITY:unmarked
     AUDIENCE:unmarked
     POLARITY:positive
     PRESUPPOSITION:presupposed
     SPEECH_ACT:inform
     AFFECTEDNESS:negative
     STRENGTH:unmarked");
  outputAS.add(rvgAnn.firstNode(),rvgAnn.lastNode(),"RVG",features);
}
```

Figure 3: A sample for the JAPE grammar to find and mark reported speech verbs

> *"The diversion of funds from the Iran arms sales is only a part of the puzzle, and maybe a very small part,"* a congressional source **said**. *"We first want to focus on how the private network which supplied the Contras got set up in 1984, and whether (President) Reagan authorized it."*

In case of contention of more than one applicable rule, the first match is chosen. The grammar rule for one of the six rules to tag reported speech constructs are shown in Figure 4.

```
Rule: Profile1
(START)
((
        (       (DIRQUOT) |
                (ANY))          :cont
        (COMMA)?
        ((SOURCE))              :source
        ((RS_VG))               :verb
        ((CIRC)?)               : circ
        (END)
        ((START)?
                ((DIRQUOT))     :cont2
                (END)?
        )?
)                               : profile )
--> MAKE_ANNOS
```

Figure 4: A JAPE grammar rule to find and mark one type of reported speech sentences

5. Evaluation

For evaluation, we randomly picked 7 newspaper articles (∼6100 words) from the WSJ corpus and created a gold standard containing the reported speech elements: *Source, Reporting Verb,* and *Reported Clause* (that is, we did not evaluate the detection of circumstantial information). The articles contain about 400 sentences and among them 133 reported speech constructs.

Apart from correct and incorrect identification of reported speech, we also measure partial correctness: If the system annotates a reported speech sentence nearly correct, but, for example, mixes up one or two terms of circumstantial information and reported clause, we speak of partially correct detection, if the meaning of the reported speech in general is maintained.

5.1. Results

For the detection of reporting verb and source, i.e., partial correctness, our system achieved a recall value of 0.79 and a precision value of 1.00, thus an F-measure of 0.88. The results for the reported clause, together with a detailed overview of the results obtained for the different test documents, can be seen in Table 3. The results for the extraction of the reported clause (content) suffers mostly from the misinterpretation of parts of reported clauses as circumstantial information.

5.2. Error Analysis

We also performed a detailed analysis of error cases introduced by our system and their root causes. In Table 4, a listing of sample errors that reduce our system's performance is shown. The examples are taken from the WSJ corpus:

1. In the first example, the "max-NP transducer",[7] which the reported speech component uses, failed to identify the NP printed in bold. This leads to a missing match because the patterns of the reported speech finder expects a noun phrase.
2. Complex circumstantial information like in Example 2 can not be detected by the current patterns and the component fails to discover the reported speech.
3. Likewise, in the next example, the boundary of the circumstantial information can not be syntactically determined, because three occurrences of *"that"* are possible starting points of the reported clause after the reporting verb.
4. Example 4 contains misguiding quotation marks, excluding the subject of the reported clause, "1987".

[7] A custom component to combine some of the base NPs into complex NP structures, e.g., for appositions.

No	Example Sentence	Source of Fault
1	"I was doing those things before, but we felt we needed to create a position where I can spend full time accelerating the creation of these (cooperative) arrangements," **Mr. Sick, 52 years old,** said.	max-NP Transducer
2	Mr. Furmark**, asked whether he had a role in the arms sale,** said, "I'm not in that business, I'm an oil man."	Reported Speech Finder
3	However, Charles Redman, the department's spokesman, said **during a briefing *that* followed the meeting** *that* the administration hadn't altered its view *that* sanctions aren't the way to resolve South Africa's problems.	Reported Speech Finder
4	In his annual forecast, Robert Coen, senior vice president and director of forecasting at the ad agency McCann-Erickson Inc., said **1987 "looks good for the advertising industry."**	Reported Speech Finder
5	**Praising the economic penalties imposed by Congress last year,** he said **it was "necessary to pursue the question of sanctions further."**	Reported Speech Finder
6	He **did acknowledge** that he knows Mr. Casey from their past association with Mr. Shaheen.	Verb Phrase Marker
7	He **declined to be more specific, but stressed** that Texas Instruments wasn't concentrating on alliances with Japanese electronics concerns, but also would explore cooperative agreements with other domestic and European companies.	Reported Speech Finder

Table 4: Samples showing different sources of errors for the reported speech finding component

It is not clear if this is a circumstantial information referring to the date the utterance was made or, more likely, a non-contiguous part of the quoted reported clause.

5. Example 5 also contains complex circumstantial information and the phenomenon of a partly quoted reported clause.
6. Example 6 was not recognized as reported speech by the system because the verb grouper component, whose output is used by the reported speech finder, failed to mark the verb construct correctly.
7. Example 7 illustrates a coordinated reported speech structure, juxtaposing two reported clauses, where the second one is not a full clause. Since this is a problem not particular to reported speech but chunking in general, we leave this to future work. Our system only tags the first reported clause as reported speech.

6. Deployment and Application

In this section, we describe how to deploy our reported speech components in practice.

6.1. Pipeline Configuration

Our reported speech components are designed to be embedded within a complete GATE analysis pipeline. They rely on annotations added by a number of existing GATE processing resources, in particular tokenization, sentence splitting, part-of-speech tagging, noun phrase chunking, verb grouping, and rudimentary morphological analysis. A complete example for a possible pipeline configuration, with the components needed for complete reported speech tagging, is shown in Figure 5.

6.2. Annotations

Our components then add annotations for the detected reporting verbs and reported speech constructs (see Figure 6).

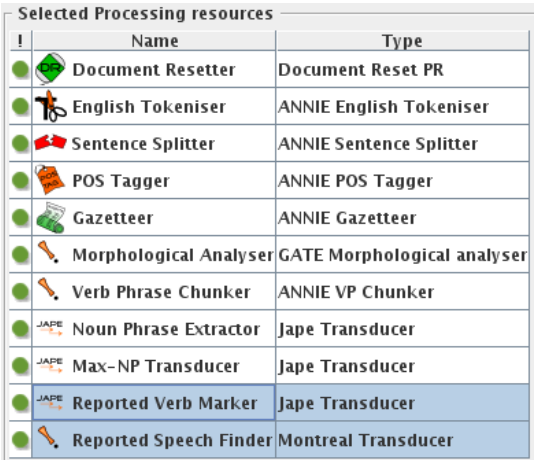


Figure 5: GATE Pipeline

All potential reporting verbs are annotated with their semantic dimensions and their main verb. This annotation is labeled *reporting verb group* ("RVG") and is generated by the reporting verb marker. All detected reported speech occurrences receive a "Reported Speech" annotation generated by the reported speech finder. Different features of the generated annotation contain detailed information about the reported speech construct, see Table 5.

Name	Value
source	source of the reported speech sentence
sourceStart	start offset of the source
sourceEnd	end offset of the source
verb	reporting verb
cont1-3	content, possibly fragmented (up to 3 parts)
cont1-3Start	start offset of the content
cont1-3End	end offset of the content

Table 5: Features of the reported speech annotation

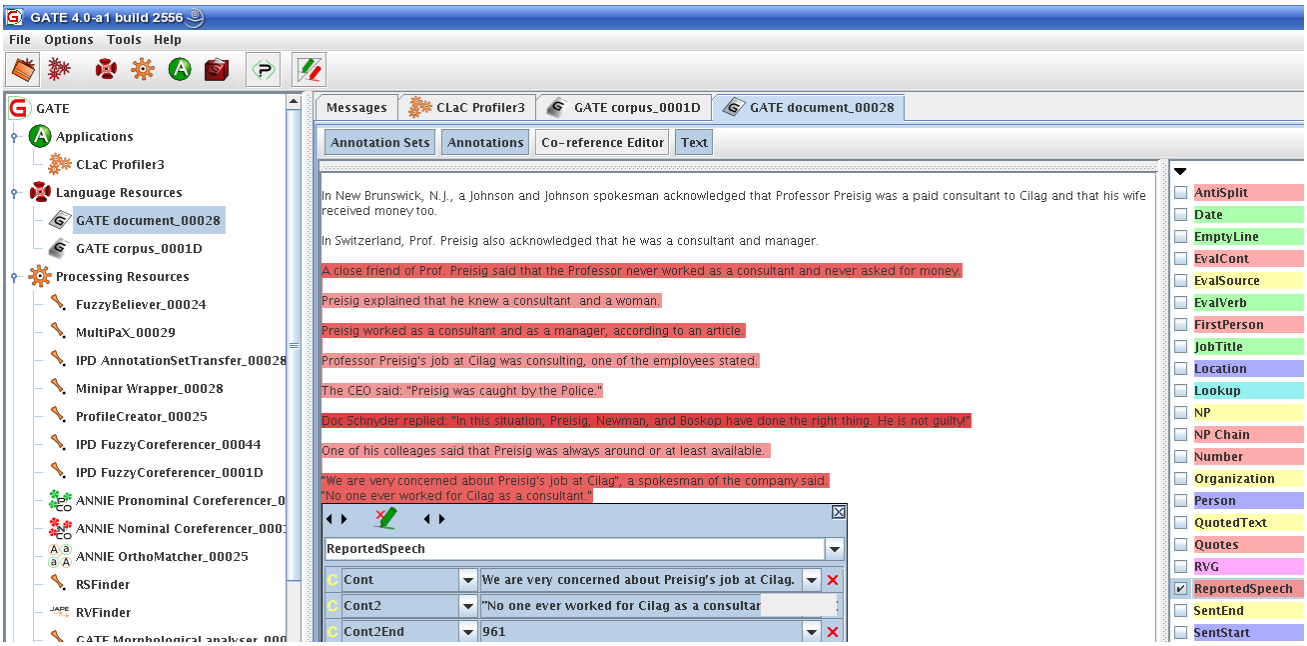


Figure 6: Example reported speech annotation in GATE

6.3. Application

Further processing of the generated reported speech results can then be performed in application-specific pipelines utilizing further components, e.g., for quote attribution as performed by *NewsExplorer* or belief analysis of reporting clauses as performed by the *Fuzzy Believer* (Krestel et al., 2007a; Krestel et al., 2007b).

7. Conclusion

Reported Speech is an important linguistic phenomenon in newspaper articles, where it serves to provide evidential scope for second hand information. It serves this function also in many of the new, more argumentative media, such as blogs, where proper attribution is part of good style and carries with it subtle information about both content and social structure. Our system takes the first step in making this information explicit and accessible. We provide an open source GATE component to identify and functionally annotate reported speech sentences for both direct and indirect speech, extracting the reported clause, the source, the reporting verb, and circumstantial information. For our test corpus we achieve 83% recall and 98% precision.

8. References

J. A. Austin. 1962. *How to Do Things with Words*. Harvard University Press.

Afzal Ballim and Yorick Wilks. 1991. *Artificial Believers: The Ascription of Belief*. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, USA.

Sabine Bergler. 1992. *The Evidential Analysis of Reported Speech*. Ph.D. thesis, Brandeis University, Massachusetts, USA.

Sabine Bergler. 1995. Generative lexicon principles for machine translation: A case for meta-lexical structure. *Journal of Machine Translation*, 9(3).

Sabine Bergler. 2005. Conveying Attitude with Reported Speech. In James C. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Applications*. Springer Verlag.

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

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proc. of the 40th Anniversary Meeting of the ACL*.

Monia Doandes. 2003. Profiling For Belief Acquisition From Reported Speech. Master's thesis, Concordia University, Montréal, Québec, Canada.

Ralf Krestel, René Witte, and Sabine Bergler. 2007a. Creating a Fuzzy Believer to Model Human Newspaper Readers. In Z. Kobti and D. Wu, editors, *Proc. of the 20th Canadian Conference on Artificial Intelligence (Canadian A.I. 2007)*, LNAI 4509, pages 489–501, Montréal, Québec, Canada, May 28–30. Springer.

Ralf Krestel, René Witte, and Sabine Bergler. 2007b. Processing of Beliefs extracted from Reported Speech in Newspaper Articles. In *Proc. of Recent Advances in Natural Language Processing (RANLP-2007)*, Borovets, Bulgaria, September 27–29.

Bruno Pouliquen, Ralf Steinberger, and Clive Best. 2007. Automatic Detection of Quotations in Multilingual News. In *Proc. of Recent Advances in Natural Language Processing (RANLP-2007)*, Borovets, Bulgaria, September 27–29.

Randolph Quirk. 1985. *A comprehensive grammar of the English language*. Longman Group Limited.

John R. Searle. 1969. *Speech Acts*. Cambridge University Press, New York.