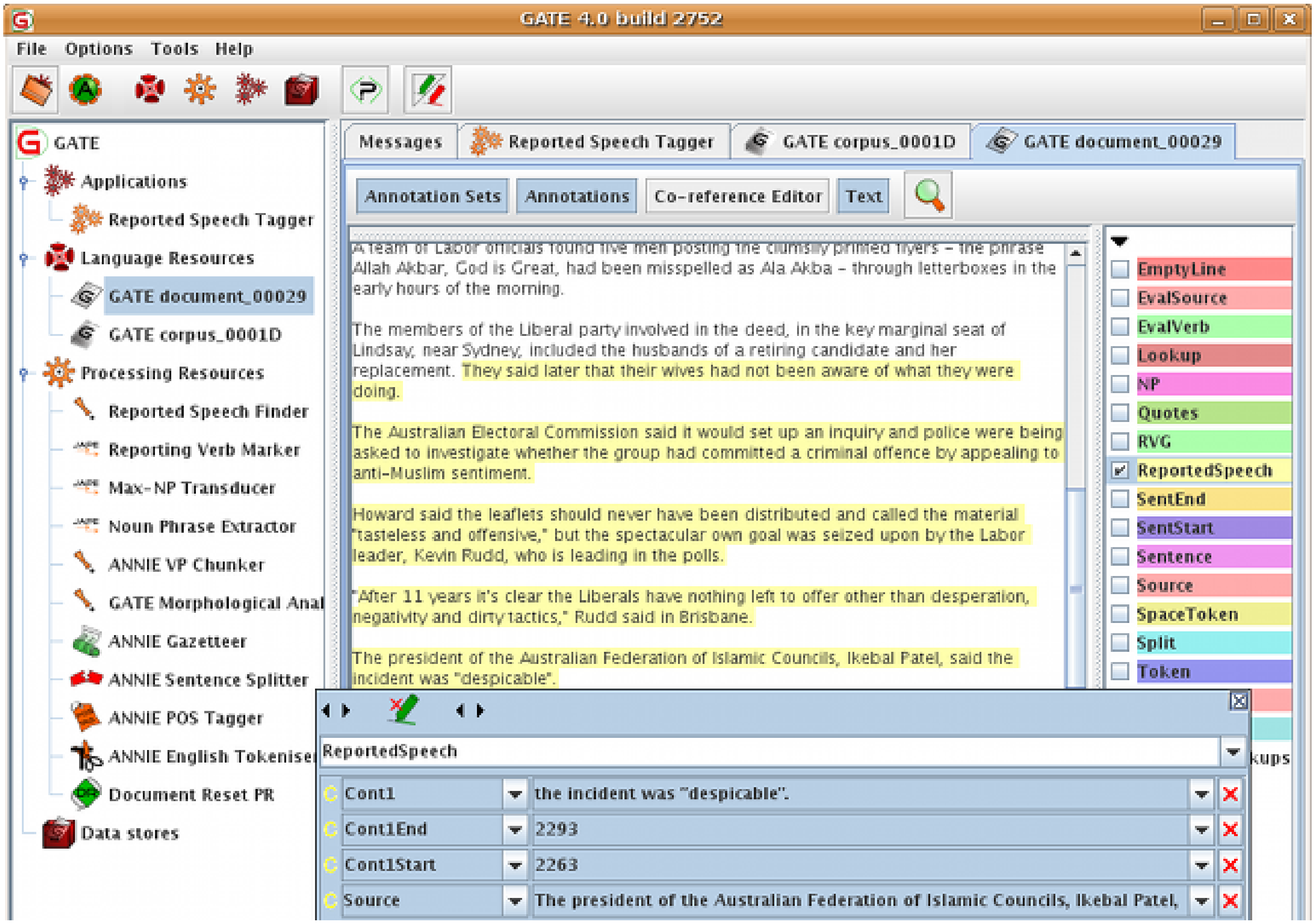
Reported Speech Tagger

- Fuzzy Believer
- GATE Components
- Reported Speech Tagger
- Reported Speech

toc_collapse=0; Table of Contents

1. Overview

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. We developed an NLP component in form of a GATE resource that can automatically detect and tag reported speech constructs, in particular the *source*, *reporting verb* and *content*. 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. For one example application, have a look at our Fuzzy Believer system.



A team of Labor officials found five men posting the clumsily printed flyers – the phrase Allah Akbar, God is Great, had been misspelled as Ala Akba – through letterboxes in the early hours of the morning.

The members of the Liberal party involved in the deed, in the key marginal seat of Lindsay, near Sydney, included the husbands of a retiring candidate and her replacement. They said later that their wives had not been aware of what they were doing.

The Australian Electoral Commission said it would set up an inquiry and police were being asked to investigate whether the group had committed a criminal offence by appealing to anti-Muslim sentiment.

Howard said the leaflets should never have been distributed and called the material "tasteless and offensive," but the spectacular own goal was seized upon by the Labor leader, Kevin Rudd, who is leading in the polls.

"After 11 years it's clear the Liberals have nothing left to offer other than desperation, negativity and dirty tactics," Rudd said in Brisbane.

The president of the Australian Federation of Islamic Councils, Ikebal Patel, said the incident was "despicable".

ReportedSpeech

Cont1	the incident was "despicable".
Cont1End	2293
Cont1Start	2263
Source	The president of the Australian Federation of Islamic Councils, Ikebal Patel,

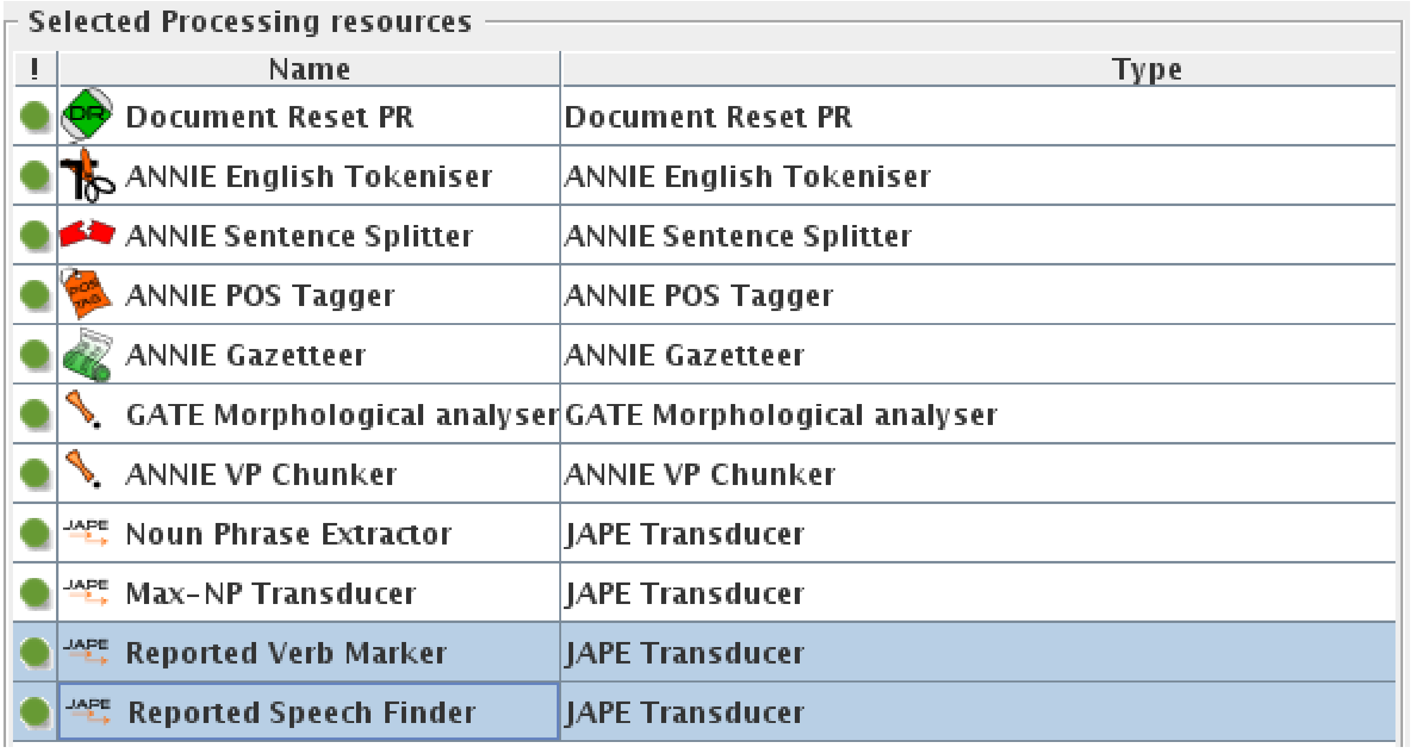
For the theoretical background and motivation please have a look at our paper *Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles*, which appeared at LREC 2008, Marrakech, Morocco. If you use our components or resources, a reference to this paper would be welcome.

2. Prerequisites

As our reported speech tagger comes in form of a number of GATE components, you will obviously need GATE itself. Most of the required pre-processing components are included in the GATE distribution. In particular, you will need (see the pipeline configuration details below): 1. Tokenizer, 2. Sentence Splitter, 3. POS-Tagger, 4. Gazetteer, 5. Morphological Analyzer, 6. Verb Phrase Chunker and 7. MuNPEx Noun Phrase Chunker. Of these, you only need to download the MuNPEx NP Chunker seperately.

3. Documentation

The reported speech tagger is designed to be embedded in more complex pipelines. Here, we describe the minimum requirements for obtaining reported speech annotations. **Note** that in the following discussion we assume you know how to work with GATE, for tasks like adding a new CREOLE repository or loading new components into a processing pipeline. If you haven't done this before, please read the GATE user's guide first!



Selected Processing resources

!	Name	Type
●	Document Reset PR	Document Reset PR
●	ANNIE English Tokeniser	ANNIE English Tokeniser
●	ANNIE Sentence Splitter	ANNIE Sentence Splitter
●	ANNIE POS Tagger	ANNIE POS Tagger
●	ANNIE Gazetteer	ANNIE Gazetteer
●	GATE Morphological analyser	GATE Morphological analyser
●	ANNIE VP Chunker	ANNIE VP Chunker
●	Noun Phrase Extractor	JAPE Transducer
●	Max-NP Transducer	JAPE Transducer
●	Reported Verb Marker	JAPE Transducer
●	Reported Speech Finder	JAPE Transducer

3.1. 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 the image to the right.

Note that our pipeline contains an additional component, the *Max NP Transducer*, which combines some of the base NPs to complex NPs (PP Attachements). This component is not yet distributed with the MuNPEx chunker. We recommend you use one of the available parsers to improve detection of source NPs.

3.2. Result Annotations

Our components then add annotations for the detected reporting verbs and reported speech constructs. 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, as shown in the following table:

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

4. Evaluation

Unfortuntely, the manually annotated texts described in our paper are currently not available due to copyright restrictions.

However, the archive below contains a small example text `WSJ-Example-Sentences.xml` with a number of reported speech sentences from our corpus.

5. Download

Latest version is v1.0 from 28.05.2008. Download:

- the complete gzipped tar archive with all JAPE grammars: ReportedSpeechTagger-1.0.tgz
- our research paper *Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles*
- the GNU GPL license, under which you can use all the components

If you use our components or resources, please cite our LREC 2008 paper: Krestel, R., S. Bergler, and R. Witte, "Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles", *Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC 2008)*, Marrakech, Morocco : European Language Resources Association (ELRA), May 28–30, 2008.

6. Feedback

For questions, comments, etc., please use the Forum.

7. Version History

- v1.0 (28.05.2008): initial public release

Source URL (retrieved on *2026-01-25 14:50*): https://www.semanticsoftware.info/reported-speech-tagger