EU-IST Project IST-2003-506826 SEKT
SEKT: Semantically Enabled Knowledge Technologies


# D2.5.1 Quantitative Evaluation Tools and Corpora V1

Wim Peters (University of Sheffield)
Niraj Aswani (University of Sheffield)
Kalina Bontcheva (University of Sheffield)
Hamish Cunningham (University of Sheffield)

**Abstract**

This deliverable covers the description and production of a semantically annotated corpus. This is available within the Sekt consortium as training and test data for the machine learning algorithms for semantic annotation and as a gold standard for the evaluation of techniques.

The document briefly describes the ontology used for the annotation (for a full description of the ontology see D1.8.1) and the various annotation principles that have been defined for this task. The report ends with a description of the annotation tool that has been created for enabling semantic annotation in GATE.

Keyword list: knowledge engineering, knowledge modelling, metadata

WP2 Metadata Generation
Report/Data
Contractual date of delivery PU (Report)/ RE (Data)
31/12/2004
Actual Date of delivery
26/01/2004

## SEKT Consortium

This document is part of a research project partially funded by the IST Programme of the Commission of the European Communities as project number IST-2003-506826.

**British Telecommunications plc.**
Orion 5/12, Adastral Park
Ipswich IP5 3RE
UK
Tel: +44 1473 609583, Fax: +44 1473 609832
Contact person: John Davies
E-mail: john.nj.davies@bt.com

**Empolis GmbH**
Europaallee 10
67657 Kaiserslautern
Germany
Tel: +49 631 303 5540
Fax: +49 631 303 5507
Contact person: Ralph Traphöner
E-mail: ralph.traphoener@empolis.com

**Jozef Stefan Institute**
Jamova 39
1000 Ljubljana
Slovenia
Tel: +386 1 4773 778, Fax: +386 1 4251 038
Contact person: Marko Grobelnik
E-mail: marko.grobelnik@ijs.si

**University of Karlsruh**e, Institute AIFB
Englerstr. 28
D-76128 Karlsruhe
Germany
Tel: +49 721 608 6592
Fax: +49 721 608 6580
Contact person: York Sure
E-mail: sure@aifb.uni-karlsruhe.de

**University of Sheffield**
Department of Computer Science
Regent Court, 211 Portobello St.
Sheffield S1 4DP
UK
Tel: +44 114 222 1891
Fax: +44 114 222 1810
Contact person: Hamish Cunningham
E-mail: hamish@dcs.shef.ac.uk

**University of Innsbruck**
Institute of Computer Science
Techikerstraße 13
6020 Innsbruck
Austria
Tel: +43 512 507 6475
Fax: +43 512 507 9872
Contact person: Jos de Bruijn
E-mail: jos.de-bruijn@deri.ie

**Intelligent Software Components S.A.**
Pedro de Valdivia, 10
28006
Madrid
Spain
Tel: +34 913 349 797
Fax: +49 34 913 349 799
Contact person: Richard Benjamins
E-mail: rbenjamins@isoco.com

**Kea-pro GmbH**
Tal
6464 Springen
Switzerland
Tel: +41 41 879 00
Fax: 41 41 879 00 13
Contact person: Tom Bösser
E-mail: tb@keapro.net

**Ontoprise GmbH**
Amalienbadstr. 36
76227 Karlsruhe
Germany
Tel: +49 721 50980912
Fax: +49 721 50980911
Contact person: Hans-Peter Schnurr
E-mail: schnurr@ontoprise.de

**Sirma AI EAD, Ontotext Lab**
135 Tsarigradsko Shose
Sofia 1784
Bulgaria
Tel: +359 2 9768 303, Fax: +359 2 9768 311
Contact person: Atanas Kiryakov
E-mail: naso@sirma.bg

**Vrije Universiteit Amsterdam (VUA)**
Department of Computer Sciences
De Boelelaan 1081a
1081 HV Amsterdam
The Netherlands
Tel: +31 20 444 7731, Fax: +31 84 221 4294
Contact person: Frank van Harmelen
E-mail: frank.van.harmelen@cs.vu.nl

**Universitat Autonoma de Barcelona**
Edifici B, Campus de la UAB
08193 Bellaterra (Cerdanyola del Vall` es)
Barcelona
Spain
Tel: +34 93 581 22 35, Fax: +34 93 581 29 88
Contact person: Pompeu Casanovas Romeu
E-mail: pompeu.casanovas@uab.es

## Executive Summary

This deliverable covers the description of the production of a semantically annotated corpus. This can be used within the Sekt consortium as training and test data for the machine learning algorithms for semantic annotation and as a gold standard for the evaluation of techniques.

The document briefly describes the ontology used for the annotation (for a full description of the ontology see D1.8.1) and the various annotation principles that have been defined for this task. The report ends with a description of the annotation tool that has been created for enabling semantic annotation in GATE.

# Contents

# 1 Overview

The semantically annotated corpus consists of 292 news articles from three news agencies: The Guardian, The Independent and The Financial Times. The news articles cover the period of August to October, 2001.
The articles belong to three general topics or domains of news gathering: International politics, UK politics and Business. Information about approximate numbers of (unique) wordforms can be found in table 1 below. The number of unique wordforms is computed over the whole corpus.

| | Number of documents | Number of wordforms | Number of unique wordforms |
|---|---|---|---|
| UK politics | 101 | 75500 | 10500 |
| International politics | 99 | 80000 | 11000 |
| Business | 92 | 64000 | 8700 |
| | | | |
| Total | 292 | 220000 | 21500 |

Table 1: Corpus Statistics words

# 2 The Ontology

The ontology used in the generation of the ontological metadata is an earlier version of the PROTON ontology, called BULO. The architecture and ontological coverage is described in detail in deliverable D1.8.1. BULO is a development of the KIMO[1] ontology, which was created and used in the scope of the KIM platform for semantic annotation, indexing, and retrieval [4]. The home page of the KIM platform is http://www.ontotext.com/kim.

The BULO ontology forms part of an automatic annotation tool for automatic ontology population and open-domain dynamic semantic annotation of unstructured and semi-structured content for Semantic Web knowledge management applications. Its main features are domain-independence and the inclusion of light-weight logical definitions.

The ontology consists of around 250 classes and 100 relations in a hierarchy with three unique beginners (top level concepts):

- Abstract (with direct hyponyms such as BusinessAbstraction and SocialAbstraction);
- Object (with direct hyponyms such as InformationResource, Agent and Organization);
- Happening (with direct hyponyms such as Situation, Event and TimeInterval).

[1] http://www.ontotext.com/kim/kimo.rdfs

The Gate platform [5,6] enables the annotation of document texts with the semantic metadata from BULO by means of a newly developed annotation tool (see section 5). Table 2 below presents an overview of the most frequent semantic classes in the annotation of the corpus, which therefore appear most significantly in the business and politics newspaper domain.

| Semantic Class | Frequency |
|---|---|
| Country | 2553 |
| Man | 2279 |
| TimeInterval | 1543 |
| Number | 1501 |
| SocialAbstraction | 1464 |
| Company | 1075 |
| Person | 1013 |
| EconomicAbstraction | 894 |
| Organization | 890 |
| Money | 878 |
| Event | 869 |
| ofCountry | 572 |
| Date | 520 |
| Newspaper | 507 |
| MilitaryConflict | 502 |

**Table 2: Corpus Statistics Semantic Classes**

# 3 Levels of Annotation

The overall objective of the created annotation was to create manually a gold standard with a high level of annotated knowledge. The result is an annotation set that should be able to cover a variety of levels and types of semantic annotation, and is decomposable into lower level task oriented ontologies or sets of classes. The sections below describe the various aspects of the annotation structure and the principles that have regulated the annotation effort.

## 3.1 Named Entities

Named entities (NEs) are considered to be entities such as *people, organizations, locations,* and others referred by name. Within a wider interpretation, NEs can be considered also to represent some scalar values (*numbers, percentages, amounts of money, dates*) and addresses.

Our named entity annotation follows in broad lines the criteria of initiatives in the direction of named entity annotation, such as MUC and ACE (http://www.ldc.upenn.edu/Projects/ACE/).

### *3.1.1 MUC*

The categories of named entities defined by the Message Understanding Conference (MUC) are the following [2,3]:

MUC distinguishes the following classes:

- **Organization**: named corporate, governmental, or other organizational entity
- **Person**: named person or family
- **Location**: name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains, etc.)
- **Date**: complete or partial date expression
- **Time**: complete or partial expression of time of day
- **Money**: monetary expression
- **Percent**: percentage

Some of these classes have one or more subclasses.

An example piece of text annotated according to the MUC categories follows:
The <ENAMEX TYPE="LOCATION">U.K.</ENAMEX> satellite television broadcaster said its subscriber base grew <NUMEX TYPE="PERCENT">17.5 percent</NUMEX> during <TIMEX TYPE="DATE">the past year</TIMEX> to 5.35 million

### *3.1.2 ACE*

The Automatic Content Extraction program (ACE[2]), uses seven types of named entities [1]:

- **Person** - Person entities are limited to humans. A person may be a single individual or a group.
- **Organization** - Organization entities are limited to corporations, agencies, and other groups of people defined by an established organizational structure.
- **Facility** - Facility entities are limited to buildings and other permanent man-made structures and real estate improvements.
- **Location** - Location entities are limited to geographical entities such as geographical areas and landmasses, bodies of water, and geological formations.
- **GPE** (Geo-political Entity) - GPE entities are geographical regions defined by political and/or social groups. A GPE entity subsumes and does not distinguish between a nation, its region, its government, or its people.
- **Vehicle** – A vehicle entity is a physical device primarily designed to move an object from one location to another, by (for example) carrying, pulling, or pushing the transported object. Vehicle entities may or may not have their own power source.

[2] http://www.ldc.upenn.edu/Projects/ACE/

- **Weapon** – Weapon entities are limited to physical devices primarily used as instruments for physically harming or destroying animals (often humans), buildings, or other constructions.

*3.1.3 Our corpus*

Our annotation of named entities is, from a taxonomic classification point of view, more comprehensive than the other initiatives described above. The reasons for this are that we are able to use a much more complete ontology than just a list of named entities. We use, in principle, all BULO classes for NE annotation. Further, we apply less restrictions on the selection of mentions in the texts for annotation (see below). For example, in MUC, expressions such as "Dow Jones Industrial Average" are not annotated, because these are not considered to be named entities. In our annotation we would have used the class 'economic abstraction' for this mention. Further, a mention such as 'Ford Focus' would have been annotated in MUC such that the make but not model is annotated (as shown below):

<ENAMEX TYPE="ORGANIZATION">Ford</ENAMEX> Focus.

Our annotation also covers the car model:

<CAR_MODEL><COMMERCIAL_ORGANIZATION>Ford
</COMMERCIAL_ORGANIZATION>Focus</CAR_MODEL>

These semantic labels from the BULO ontology are the most specific concepts that are applicable to the mention 'Ford Focus'. The BULO concepts are organized into a number of hierarchies that go up to the unique beginners mentioned in Section 2 above. The BULO hierarchical chains associated with these semantic labels are the following:

Commercial Organization > Organization > Agent > Object

Car Model > Product > Business Object > Object

Similarly, MUC's <ENAMEX TYPE="LOCATION">Plymouth
Airport</ENAMEX> would look like this:
<AIRPORT><CITY>Plymouth</CITY>airport</AIRPORT>

Associated hierarchical chains:

City > Populated Place > Location > Object

Airport > Transport Facility > Facility > Location > Object

Lastly, MUC did not take into account names of groups of people (e.g. 'Republicans') or adjectival forms of location names (e.g. 'American', 'Japanese'). Our annotation covers these instances.

Overall, the use of BULO as a more or less fully-fledged ontology in our annotation significantly extends the semantic coverage of our annotation compared with previous and ongoing initiatives. On the other hand, because annotation is more complex than initiatives such as MUC and ACE, it does have the drawbacks that it is more time-consuming to create, and that decisions may be more subjective, leading to a slightly lower quality of annotation (because the level of inter-annotator agreement may be lower). The annotation is also only valid with respect to the particular ontology used, so is not as versatile in some ways as a non-ontology based annotation scheme. These disadvantages are unavoidable though, if a true semantic-based annotation is required.

### 3.2 Common Nouns

This type of annotation takes semantic coverage beyond that of proper names, and concentrates on the annotation of common nouns as they appear in the text. Its scope is more extended than MUC, and is in line with the ACE guidelines, which we partly follow (see section 4). As with MUC, the main difference with ACE is the number of semantic classes we use in our annotation.

The advantage of extending the scope of semantic annotation to common nouns is that the result is a much more detailed and varied semantic characterisation of the domain involved and the entities that play a significant role in it. This extra information is necessary for more fine-grained semantic processing tasks. The drawbacks of this approach havev already been mentioned above.

In the annotation process we have followed the following general strategy. We annotate the following text occurrences:

a) Common nouns with the same orthography as any BULO class (e.g. "bank", "government", "airline" and "president" occur as BULO classes)
b) Common nouns different from any BULO class but important for the topic (see section 4.2); e.g. "attack" in a report on the dangers of anthrax as biological warfare; "deal" in a report on business.

## 4. Annotation Principles

This section describes a number of principles that we have defined and adhered to in the annotation process. In the definition phase particular attention has been paid to previous and ongoing initiatives in semantic annotation. The most important annotation system in this respect is that of ACE, which has created a set of guidelines for its annotation procedure. These try to be as comprehensive as possible in their attempt to select relevant linguistic phenomena that determine semantic annotation.

The choices we have made for our annotation have mostly been determined by the available time to produce the annotation, and form in some respects a compromise.
It is our belief that in the majority of cases where we have left out information that is explicitly annotated in ACE, a number of small post processing algorithms can make

implicitly available information explicit. Note that the choices discussed here are not imposed by the limitations of the OCAT tool, and therefore future possible annotation initiatives using OCAT are in no way restricted to the same set of guidelines.

## 4.1 Orthography

Possessive endings ('s) and plural endings on "-s" are, as opposed to ACE, not treated as separate tokens, and annotated as an integral part of the entity.
Lemmatization will allow in a later stage the addition of explicit annotations regarding possessives and plurals.

## 4.2 Topicality

The decision which entities, mentioned in the text by means of a common noun, should be annotated depends on the rather subjective criterion of relevance, which we define as follows:

Only when the mention of an entity is considered relevant to the overall discourse topicality of a text, the entity is annotated.

For example, in a text on the American led assault on targets in Afghanistan, entities such as military persons, weapons, locations and dates will be relevant. In general, news texts, because of their very nature, contain predominantly mentions of relevant entities. However, no annotation is added to entities in sentences such as:

"On a clear moonlit *night* in *Afghanistan* and an autumnal *Sunday* *lunchtime* in *Washington*, the moment everyone expected finally arrived."

## 4.3 Phrasal Annotation

The annotation covers phrasal structures, with the exception of determiners and quantifiers at the beginning of the phrase. This principle maintains the syntactically expressed semantic dependencies between entities, and implies the embedding of annotations. For instance, apposition is expressed in the following way:

"The Canadian premier Mike Harris"
The <Man><Premier><OfCountry>Canadian</OfCountry> premier</Premier> Mike Harris</Man>

In cases of conjunction and disjunction con- or disjuncted entities have been annotated as one whole:

"..., unlike in the Gulf or Kosovo wars,..."
"…, unlike in the <MilitaryConflict><Gulf>Gulf</Gulf> or <PoliticalRegion>Kosovo</PoliticalRegion>wars</MilitaryConflict>,…"

"The US carriers Enterprise and Carl Vinson..."

"The <Ship><Country>US</Country> carriers Enterprise and Karl Vinson</Ship>…"

A small set of post processing algorithms involving detection of capitalization, deletion of embedded determination, quantification and prepositional phrases will automatically identify the name of the entity.

Further, ACE uses nested phrasal annotation with explicit attributive links between appositions and nouns, and also marks the head of the annotated phrase. We have decided not to follow this, but to use, if necessary, automatic techniques in a post processing stage to detect the phrasal head and create apposition links between the phrasal elements.

### 4.4 Generics

An entity is generic when the entity being referred to is not a particular, unique referent. Instead, generic entities refer to a kind or type of entity.

Generic entities have, as opposed to the ACE guidelines, not been annotated in this corpus. An example of generic use is marked by asterisks in the following example sentences:

"The Pakistani authorities have declared parts of the border off limits for *journalists*."

"He noted that in recent weeks there had been a significant increase in *Middle Eastern and eastern European nationals* trying to cross illegally into the US from Mexico."

Overall, generic use is generally characterized by either:

- a definite singular ("The American diet is lethal".),
- an indefinite singular ("A president should know where Afghanistan is"),
- a singular with the zero article ("War is crime"),
- an indefinite plural ("It's an action against terrorists, terrorism, and their sanctuaries and their supporters")

The adopted annotation strategy only covers referential expressions. In all cases, a decision needed to be made by the annotator whether definite and indefinite cases such as the ones above are used in a generic sense, or whether they either refer to an already introduced entity into the discourse or introduce potentially relevant entities. If they are judged generic, they receive no annotation. If they are regarded as indefinites, they are mostly only annotated when they serve the greater purpose of adding relevant semantic content to the semantic metadata.

The following examples are all deemed to be non-generic references, and have therefore been anotated with a BULO class:

Where a particular weapon is annotated (25 Grad missiles), it is logical to annotate the other weapons that are mentioned, even if these are indefinites (markes by asterisks):
"25 Grad missiles swooped towards the Taliban lines from the valley. Volleys of *mortars and 130mm shells* were traded, booming off the hills..."

"…Pakistan's airspace was used by *US and British forces*…"

"Glaxo introduces a new scheme under which *Aids treatments* were made available to African governments..."

Sometimes the use of indefinites forms an integral part of the language use in press releases. This does not indicate generic use and a decision should be made on the relevance of each concept introduced in this way: "*Security* has been tightened at *airports, ports, railway stations*....

"The Pentagon issued a *statement*…"

### 4.5 Negated entities

Non-existing and negated entities are in general not annotated:

"American and British forces are not planning a sustained *war* against the Taliban."

This type of war is not actualized, not even planned.

"No *country* lightly commits forces to military action"

This sentence lacks any annotation, because , "country" is negated, and "forces" and "military action" are used in a generic sense.

### 4.6 Coreference

In the corpus annotation there is no explicit coreference of all mentions of a particular entity. This type of information does exist in ACE, and it is even possible for a noun phrase to contain an embedded mention of the same entity. For instance, the phrase "The historian who taught herself COBOL" evokes a Person entity with three mentions: the entire phrase, and the words "*herself*" and "*who*" (example taken from the ACE guidelines).

In version 1 of the corpus, the annotation would be limited to "historian", without explicit links to the pronominal expressions. Further work in years 2 and 3 will address this.

## 5 The Ontology-based Corpus Annotation Tool (OCAT)

The Ontology-based Corpus Annotation Tool (OCAT) is a GATE plugin [5,6], which uses one or more ontologies for annotation. The required ontology can be selected from a pull-down list of available ontologies, and can be changed at any time during the annotation process. The ontology provided can be in any format that can be read by the GATE ontology support (currently OWL, DAML and RDF). This means that new formats can be added in future, if supported by GATE. Version 1 of OCAT

supports only annotation with information about the ontology class. Future versions will support annotation with instance information and properties.

## 5.1 Viewing Annotated Texts

Ontology-based annotations in the text can be viewed by selecting in the ontology tree the desired classes (see Figure 1). By default, when a class is selected, all of its sub-classes are also automatically selected and their mentions are highlighted in the text. There is an option to disable this default behaviour (see Section 5.4).

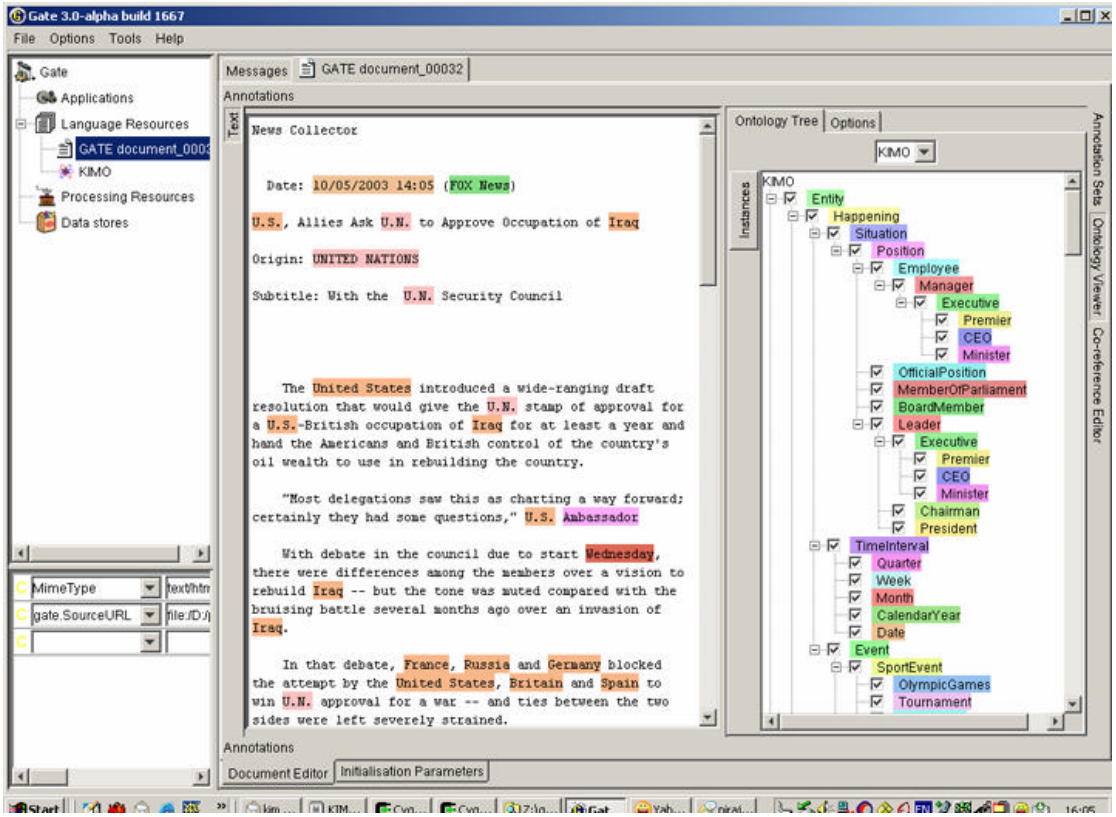

Fig.1 Viewing Ontology-Based Annotations

Figure 1 shows the mentions of each class in a different colour. These colours can be customised by the user by clicking on the class names in the ontology tree. It is also possible to expand and collapse branches of the ontology.

## 5.2. Editing Existing Annotations

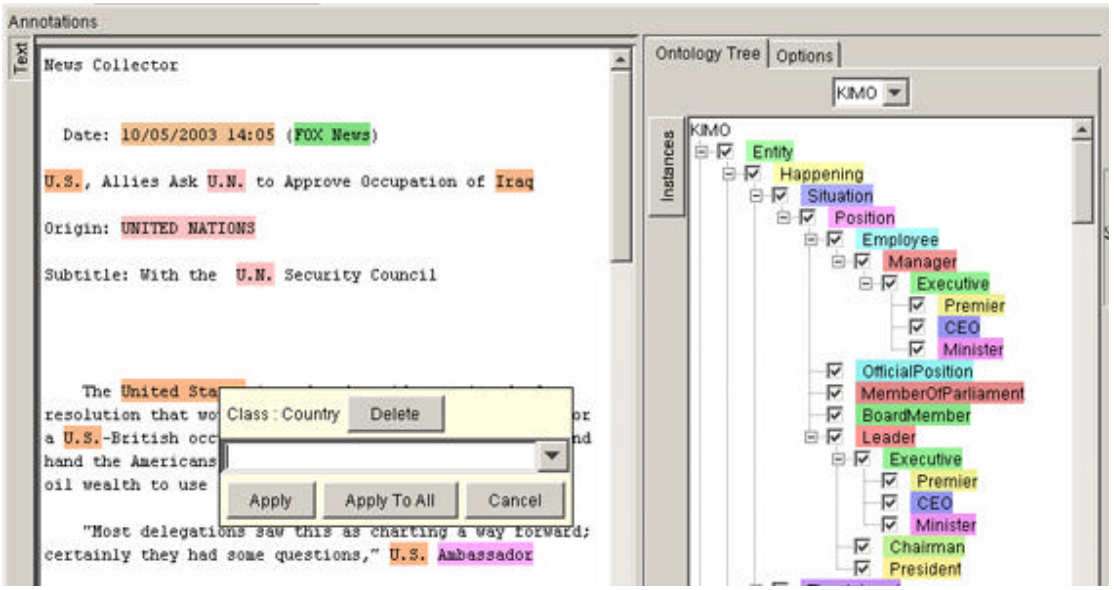

Fig. 2 Editing Existing Annotations

In order to view the class of a highlighted annotation in the text (e.g., United States - see Figure 2), hover the mouse over it and an edit dialog will appear. It shows the current class (Country in our example) and allows the user to delete it or change the class. To delete an existing annotation, press the Delete button.

A class can be changed by starting to type the name of the new class in the combo-box. Then it displays a list of class names, which start with the typed string. For example, if we want to change the type from Country to Location, we can type ``Lo'' and all classes which names start with Lo will be displayed. The more characters are typed, the fewer matching classes remain in the list. As soon as one sees the desired class in the list, it is chosen by clicking on it.

It is possible to apply the changes to all occurrences of the same string and the same previous class, not just to the current one. This is useful when annotating long texts. It is known as the "one sense per discourse" assumption, which is not always true. So the user needs to make sure that they still check the classes of annotations further down in the text, in case the same string has a different meaning (e.g., bank as a building vs. bank as a river bank).

## 5.3. Adding New Annotations

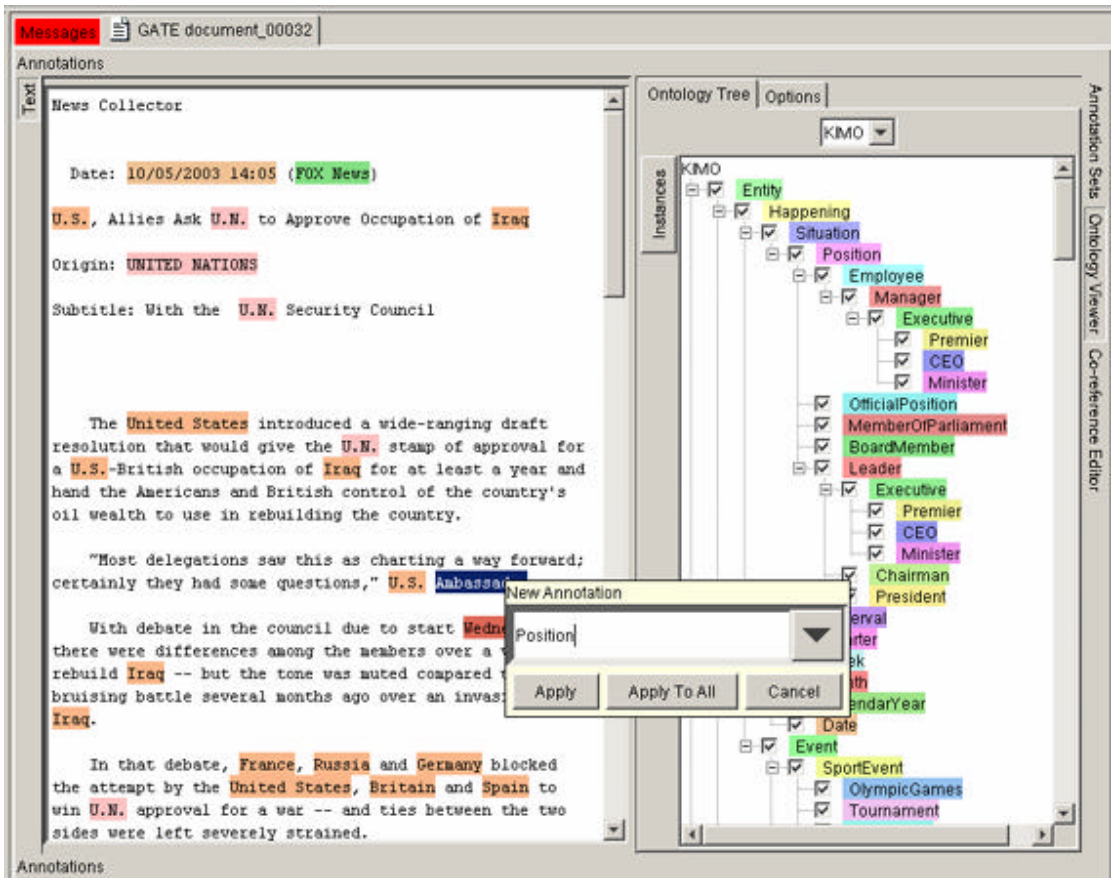


Fig.3 Add New Annotation Dialog

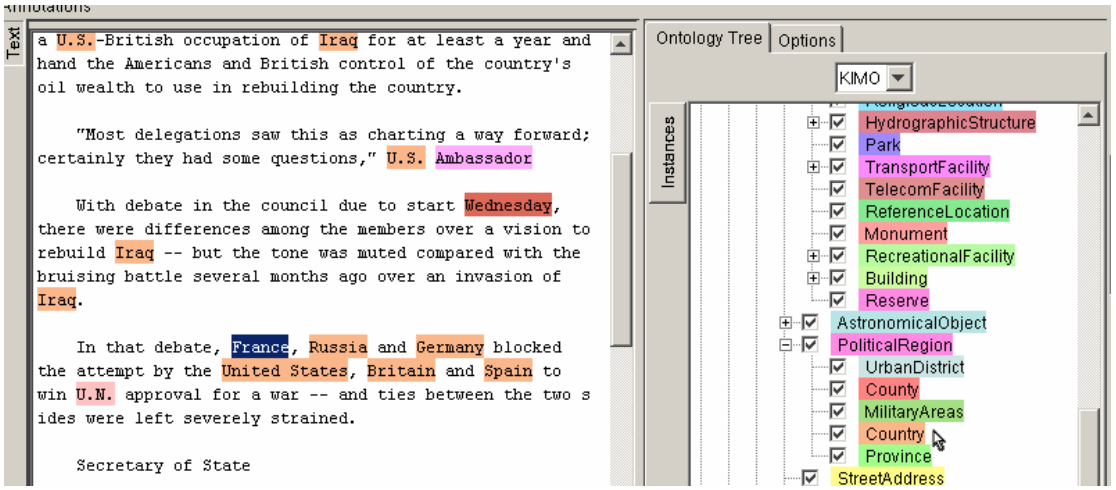


Fig. 4 Adding New Annotation by Clicking

New annotations can be added in two ways: using a dialogue (see Figure 3 or by selecting the text and clicking on the desired class in the ontology tree (see Figure 4).

When adding a new annotation using the dialogue, select a text and after a very short while, if the mouse is not moved, a dialogue will appear (see Figure 3). Start typing the name of the desired class, until you see it listed in the combo-box, then select it with the mouse. This operation is the same, as in changing the class of an existing

annotation. One has the option of applying this choice to the current selection only or to all mentions of the selected string in the current document (Apply to All button).

### 5.4. Options

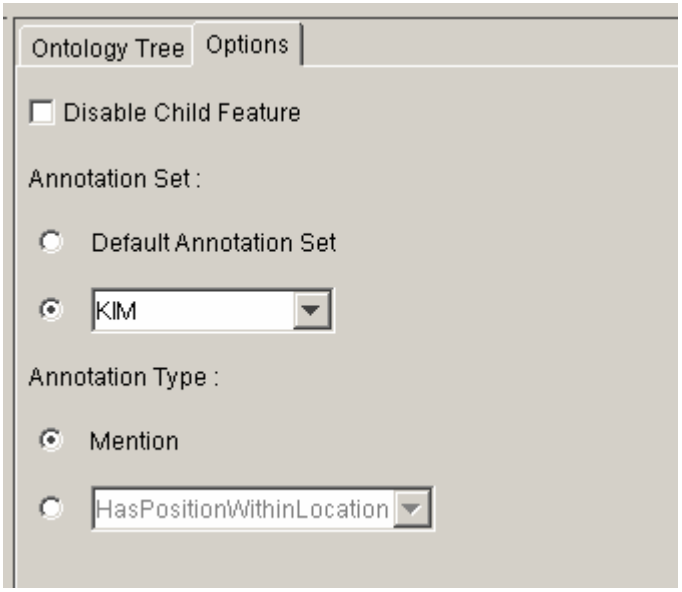


Fig.5 OCAT Tool Options

There are several options that control the OCAT behaviour (see Figure 5):

- **Disable child feature**: By default, when a class is selected, all of its sub-classes are also automatically selected and their mentions are highlighted in the text. This option disables that behaviour, so only mentions of the selected class are highlighted.
- **Annotation Set**: GATE stores information in annotation sets and OCAT allows you to select which set to use as input and output.
- **Annotation Type**: By default, this is annotation of type Mention, but that can be changed to any other name. This option is required because OCAT uses Gate annotations to store and read the ontological data. However, to do that, it needs a type (i.e., name) so ontology-based annotations can be distinguished easily from other annotations (e.g., tokens, gazetteer lookups).
- **Delete confirmation**: By default, OCAT deletes ontological information without asking for confirmation, when the delete button is pressed. However, if this leads to too many mistakes, it is possible to enable delete confirmations from this option.

## 6. Technical details

The OCAT tool is freely available as part of the open-source GATE platform, distributed under the LGPL licence from http://gate.ac.uk.

The corpus itself is currently only available within the SEKT consortium. After version 2, we plan to negotiate its public release via the Linguistic Data Consortium (LDC) http://www.ldc.upenn.edu/ , which is a well-established organisation for sharing linguistic resources: data, tools and standards.

The corpus is in XML format and comprises 3 sub-directories, one for each type of news: business, international political and UK political. The information about which source this file has come from is available in the file name: files starting with ft are from Financial Times, gu are from the Guardian, and ind – from the Independent.

## Bibliography and references

[1] EDT Guidelines for English V4.2.6
Available from http://www.ldc.upenn.edu/Projects/ACE/Annotation/

[2] MUC& Named Entity Task Definition
http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html

[3] R. Grishman, B. Sundheim, "Message Understanding Conference - 6: A Brief History", Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, June 1996.

[4] A. Kiryakov, B. Popov, D. Ognyanoff, D. Manov, A. Kirilov, M. Goranov, *Semantic Annotation, Indexing, and Retrieval.* To appear in Elsevier's Journal of Web Semantics, Vol. 1, ISWC2003 special issue (2), 2004. http://www.websemanticsjournal.org/

[5] K. Bontcheva, V. Tablan, D. Maynard, H. Cunningham. Evolving GATE to Meet New Challenges in Language Engineering. *Natural Language Engineering*. **10**(3/4): 349-373. 2004.

[6] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*. Philadelphia, July 2002.