Tools for Natural Language Processing

GATE

General Architecture for Text Engineering

Marián Šimko

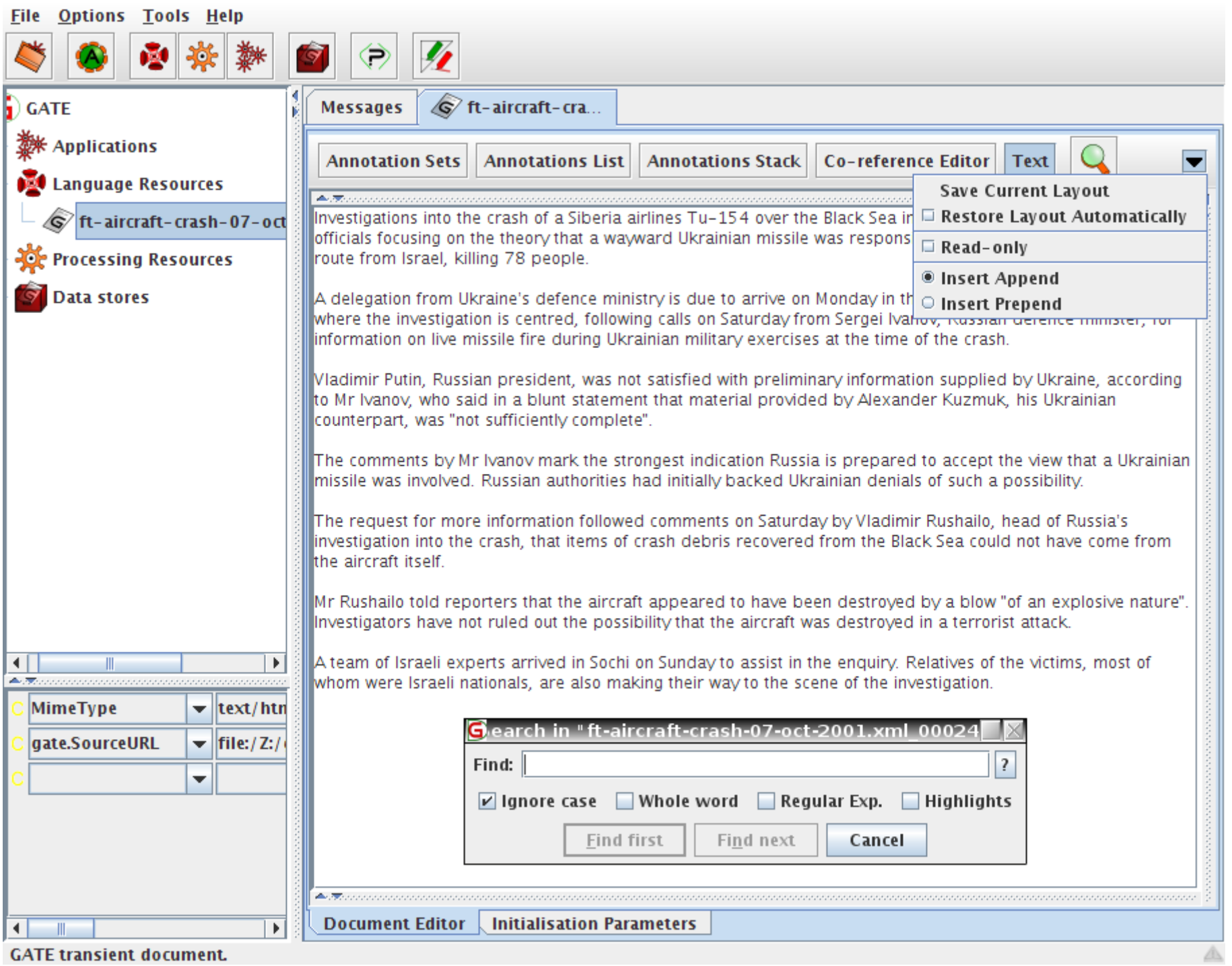
PeWe 13.11.2012

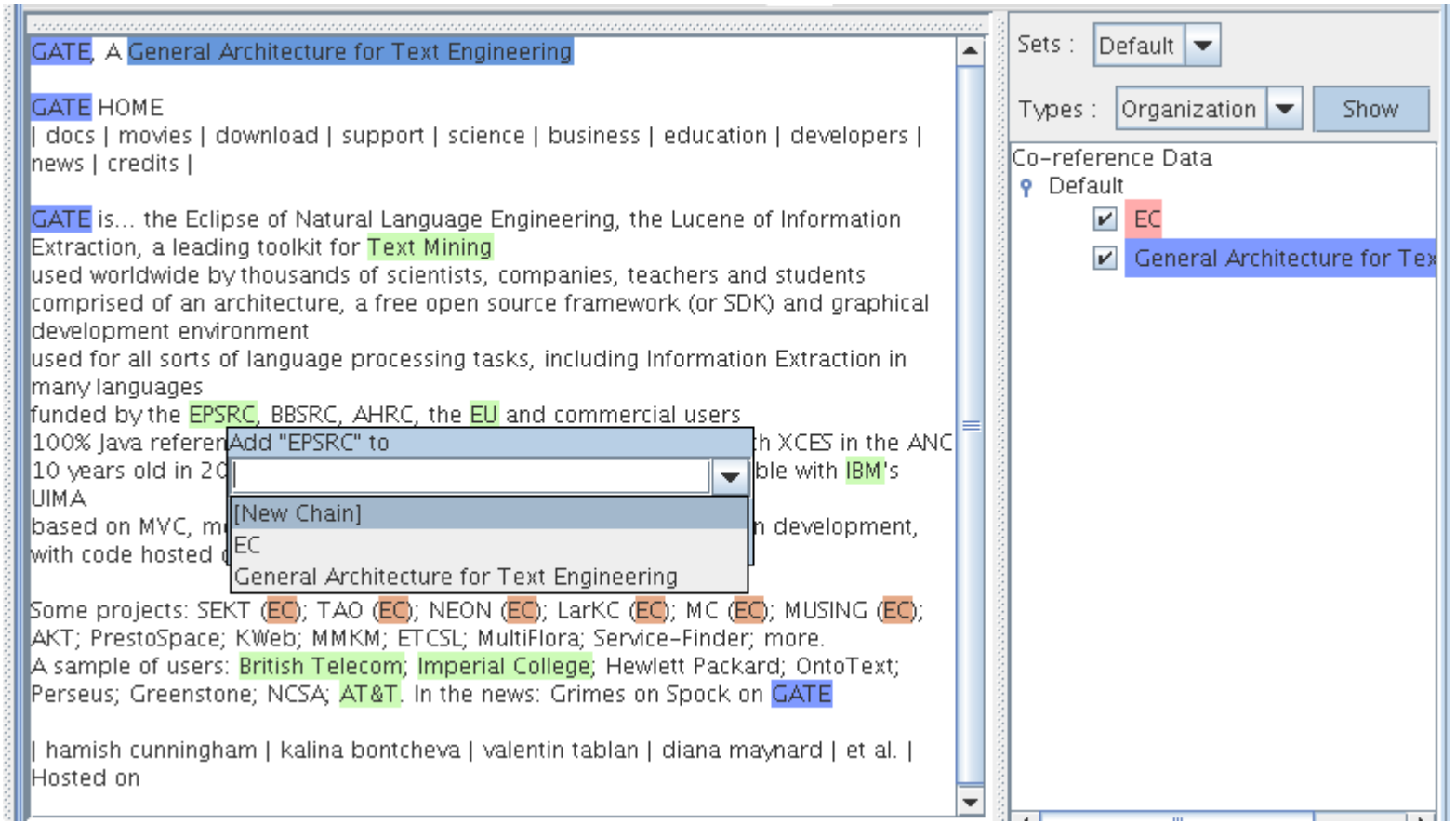
Čo je GATE

- „... an infrastructure for developing and deploying software components that process human language“
- 1995
- open source
- University of Sheffield
- JAVA
- GATE, NTLK, R, RapidMiner

Čo je GATE

- GATE Developer – IDE
- GATE Cloud – cloud
- GATE Teamware – web app
- GATE Mimir – search repo
- GATE Embedded – lib

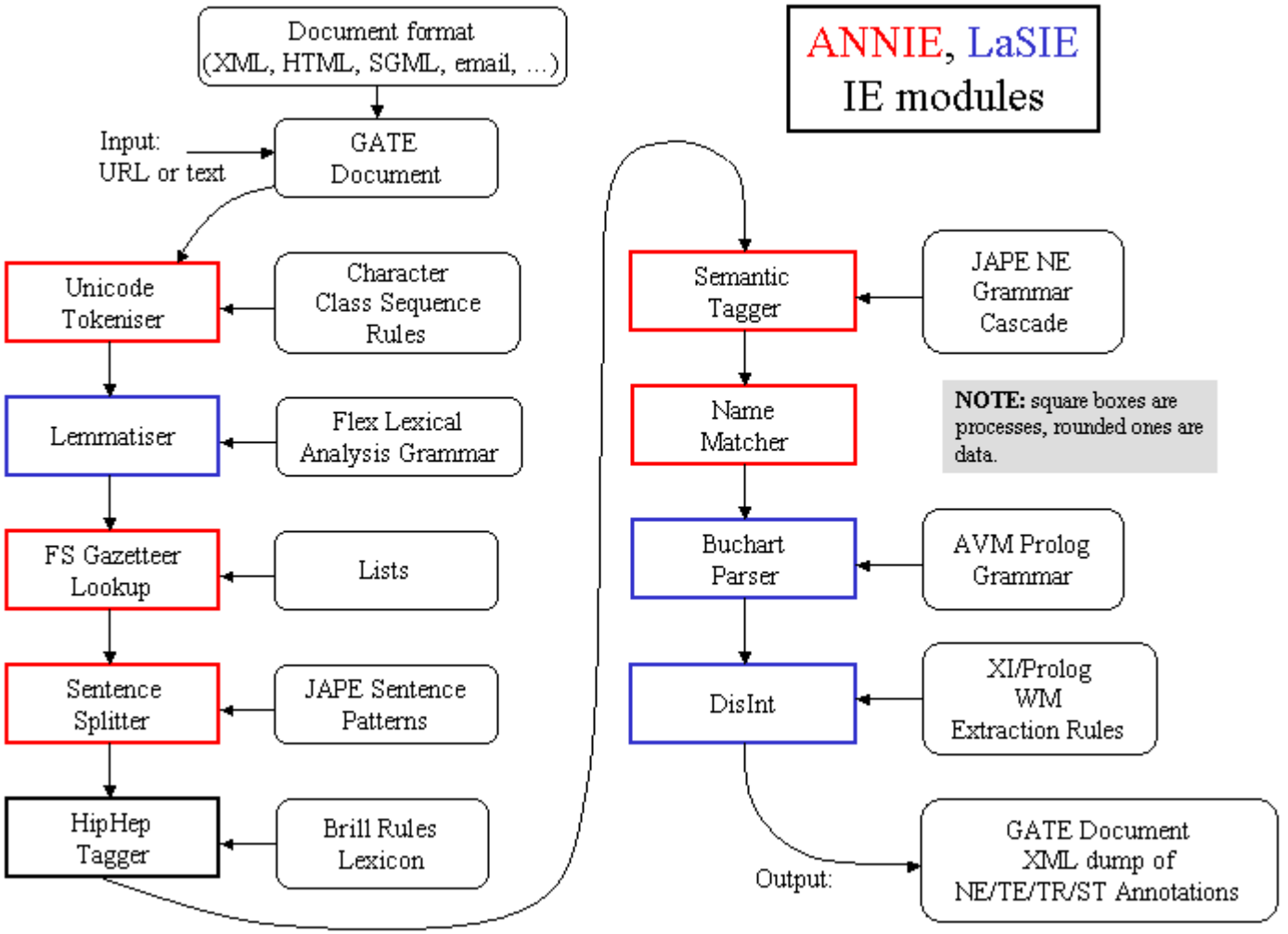

File Options Tools Help
GATE
Applications
Language Resources
ft-aircraft-crash-07-oct
Processing Resources
Data stores
MimeType
text/htn
gate.SourceURL
file:/Z:/
Messages
ft-aircraft-cra...
Annotation Sets
Annotations List
Annotations Stack
Co-reference Editor
Text
Save Current Layout
Restore Layout Automatically
Read-only
Insert Append
Insert Prepend
Investigations into the crash of a Siberia airlines Tu-154 over the Black Sea in
officials focusing on the theory that a wayward Ukrainian missile was respons
route from Israel, killing 78 people.
A delegation from Ukraine's defence ministry is due to arrive on Monday in th
where the investigation is centred, following calls on Saturday from Sergei Ivanov,
information on live missile fire during Ukrainian military exercises at the time of the crash.
Vladimir Putin, Russian president, was not satisfied with preliminary information supplied by Ukraine, according to Mr Ivanov, who said in a blunt statement that material provided by Alexander Kuzmuk, his Ukrainian counterpart, was "not sufficiently complete".
The comments by Mr Ivanov mark the strongest indication Russia is prepared to accept the view that a Ukrainian missile was involved. Russian authorities had initially backed Ukrainian denials of such a possibility.
The request for more information followed comments on Saturday by Vladimir Rushailo, head of Russia's investigation into the crash, that items of crash debris recovered from the Black Sea could not have come from the aircraft itself.
Mr Rushailo told reporters that the aircraft appeared to have been destroyed by a blow "of an explosive nature". Investigators have not ruled out the possibility that the aircraft was destroyed in a terrorist attack.
A team of Israeli experts arrived in Sochi on Sunday to assist in the enquiry. Relatives of the victims, most of whom were Israeli nationals, are also making their way to the scene of the investigation.
earch in "ft-aircraft-crash-07-oct-2001.xml_00024
Find:
?
Ignore case
Whole word
Regular Exp.
Highlights
Find first
Find next
Cancel
Document Editor
Initialisation Parameters
GATE transient document.


GATE, A General Architecture for Text Engineering
GATE HOME
| docs | movies | download | support | science | business | education | developers |
news | credits |
GATE is... the Eclipse of Natural Language Engineering, the Lucene of Information
Extraction, a leading toolkit for Text Mining
used worldwide by thousands of scientists, companies, teachers and students
comprised of an architecture, a free open source framework (or SDK) and graphical
development environment
used for all sorts of language processing tasks, including Information Extraction in
many languages
funded by the EPSRC, BBSRC, AHRC, the EU and commercial users
Add "EPSRC" to
[New Chain]
EC
General Architecture for Text Engineering
Some projects: SEKT (EC); TAO (EC); NEON (EC); LarKC (EC); MC (EC); MUSING (EC);
AKT; PrestoSpace; KWeb; MMKM; ETCSL; MultiFlora; Service-Finder; more.
A sample of users: British Telecom; Imperial College; Hewlett Packard; OntoText;
Perseus; Greenstone; NCSA; AT&T. In the news: Grimes on Spock on GATE
| hamish cunningham | kalina bontcheva | valentin tablan | diana maynard | et al. |
Hosted on
Sets :
Default
Types :
Organization
Show
Co-reference Data
Default
EC
General Architecture for Tex

GATE pod lupou

- CREOLE – komponentový model
  - Language Resources
  - Processing Resources
  - Visual Resources
- ANNIE – basic text extraction pipeline
  - implementácia vybraných resource
- JAPE – jazyk pre reg. výrazy nad anotáciami

ANNIE


Document format
(XML, HTML, SGML, email, ...)
Input:
URL or text
GATE
Document
Unicode
Tokeniser
Character
Class Sequence
Rules
Lemmatiser
Flex Lexical
Analysis Grammar
FS Gazetteer
Lookup
Lists
Sentence
Splitter
JAPE Sentence
Patterns
HipHep
Tagger
Brill Rules
Lexicon
ANNIE, LaSIE
IE modules
Semantic
Tagger
JAPE NE
Grammar
Cascade
Name
Matcher
NOTE: square boxes are processes, rounded ones are data.
Buchart
Parser
AVM Prolog
Grammar
DisInt
XI/Prolog
WM
Extraction Rules
Output:
GATE Document
XML dump of
NE/TE/TR/ST Annotations

JAPE

```
Phase: UrlPre
Input:  Token SpaceToken
Options: control = appelt

Rule: Urlpre

( (({Token.string == "http"} |
  {Token.string == "ftp"})
 {Token.string == ":"}
 {Token.string == "/"}
         {Token.string == "/"}
        ) |
({Token.string == "www"}
         {Token.string == "."}
        )
):urlpre
-->
:urlpre.UrlPre = {rule = "UrlPre"}
```

References

- http://gate.ac.uk