GATE and UIMA in Language Technology Teaching

Graham Wilcock (University of Helsinki)

Structured Documents Processing

The students have experience of Java and XML, and of Apache open source software. They use Xerces, Xalan and FOP in a course *Structured Documents Processing* [1] to validate and transform Shakespeare's sonnets from XML to XHTML, XSL-FO, PDF and SVG, as shown in Doug Tidwell's IBM tutorials [2]. Sonnets are good examples for learning to validate with XML Schemas: the 14 lines are structured differently in Italian sonnets (1 octet and 1 sestet) and English sonnets (3 quartets and 1 couplet).

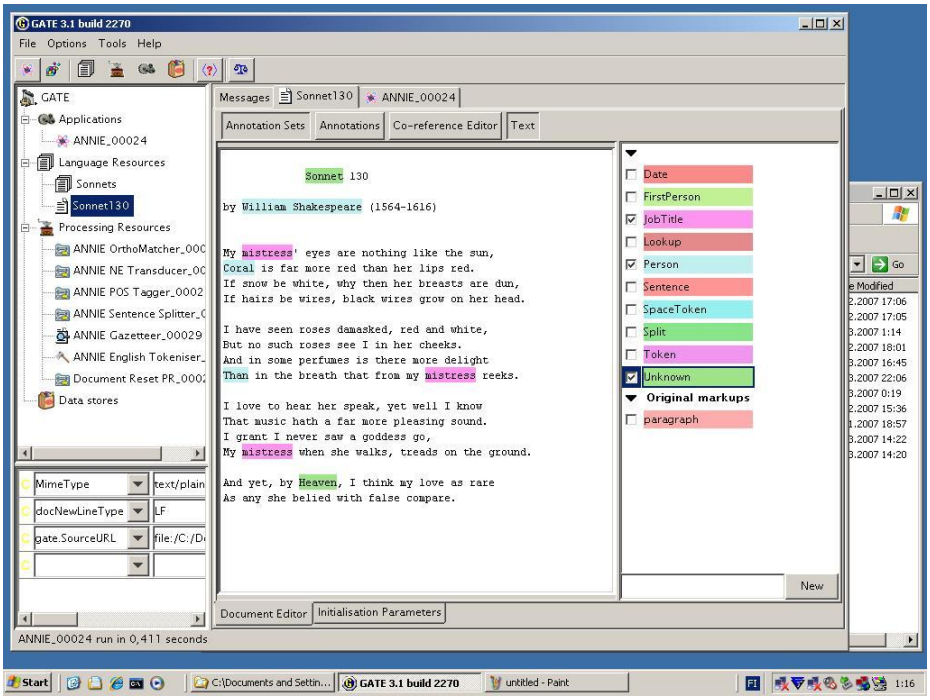

Fig. 1. Shakespeare's Sonnet 130 with ANNIE annotations in GATE.

Unstructured Information Extraction

In a course *Open Source Language Technology* [3], students use GATE with the same sonnet examples. One advantage of GATE is the ready-made information extraction system ANNIE. Sonnets with ANNIE annotations arouse students' interest: in Fig. 1, *William Shakespeare*, *Coral* and *Than* are all annotated as Person, *mistress* is amusingly classified as JobTitle, and *Sonnet* and *Heaven* are marked as Unknown. Students find that the Unknowns are known in WordNet. After seeing the ready-made ANNIE results, students add their own annotations by writing JAPE rules for NPs, PPs, etc.

Moving to Eclipse and UIMA

Currently students use jEdit with Ant as Java IDE and as XML editor [4]. The Eclipse learning curve is steeper, but Eclipse experience will be good for future employment.

Students will set up the OpenNLP tools in UIMA as a practical assignment, using the very clear instructions provided. They will also install the Stanford Named Entity Recognizer [5] using the wrapper by F. Laws [6]. Compared with ANNIE (Fig. 1), the Stanford NER in Fig. 2 has classified *William Shakespeare* and *Coral* (but not *Than*) as Person, while *Sonnet* is an Organization and *Heaven* is a Location.

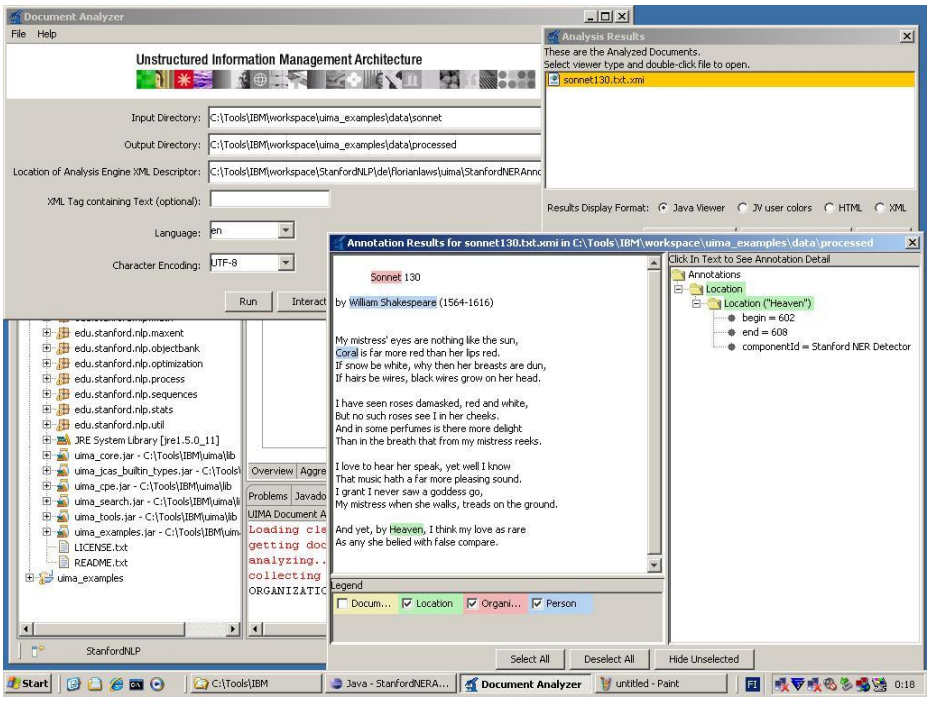

Fig. 2. Sonnet 130 with Stanford NER annotations in UIMA.

References

[1] G. Wilcock (2002-6) *Structured Documents Processing* course materials
http://www.ling.helsinki.fi/kit/2006s/clt232/materiaali.shtml
[2] D. Tidwell (1999-2000) *Transforming XML Documents*
http://www.ling.helsinki.fi/kit/2006k/clt232/tutorials/XMLtoHTML.html
[3] G. Wilcock (2003-5) *Open Source Language Technology* course materials
http://www.ling.helsinki.fi/kit/2005s/clt262/materiaali.shtml
[4] W. Le Page, P. Wellens (2002-3) *jEdit as an Advanced XML Editor*
http://www.adrem.ua.ac.be/~wellenslepage/jedit_as_axe/
[5] Stanford NLP Group (2006) *Stanford Named Entity Recognizer*
http://nlp.stanford.edu/software/CRF-NER.shtml
[6] F.Laws (2006) *UIMA Integration for the Stanford Named Entity Recognizer*
http://www.florianlaws.de/