

Documentation for

JULIE Lab ACE Reader

Ekaterina Buyko

Oleg Lichtenwald

Jena University Language & Information Engineering (JULIE) Lab

Fürstengraben 30

D-07743 Jena, Germany

{buyko, lichtenwald}@coling-uni-jena.de

1 Objective

The JULIE LAB ACE READER is an UIMA Collection Reader (CR). It reads the English section of the ACE 2005 Multilingual Training Corpus data, which is given as XML files, and converts it to types defined in the UIMA type system that we provide as well. The JULIE LAB ACE READER is part of the JULIE NLP tool suite¹ which contains several NLP components (all UIMA compliant) from sentence splitting to named entity recognition and normalization as well as a comprehensive UIMA type system.

The JULIE LAB ACE READER is currently available in version 2.0. For more detailed information about the ACE data, please read [DMP⁺04].

2 Requirements and Dependencies

The JULIE LAB ACE READER is written in Java 5.0 using Apache UIMA version 2.2.1-incubation². It was not tested with other UIMA versions.

The input of the JULIE LAB ACE READER can be purchased at the Linguistic Data Consortium (LDC)³. The Output of the CR takes place by annotation objects. The classes corresponding to these objects are part of a *JULIE UIMA Type System*⁴.

¹<http://www.julielab.de/>

²<http://incubator.apache.org/uima/>

³<http://www ldc.upenn.edu/>

⁴The *JULIE UIMA type system* can be obtained from <http://www.julielab.de/>

The CR comes as a UIMA pear file. Run the Pear-Installer (e.g., `./runPearInstaller.sh` for Linux) from your UIMA-bin directory. After installation, you will find a subfolder `desc` in your installation folder. This directory contains a descriptor `ACEReadersDescriptor.xml`. You may now e.g. run UIMA's Collection Proceessing Engine Configurator (`cpeGUI.sh`) and add the wrapper as a component into your NLP pipeline.

3 Using the CR – Descriptor Configuration

In UIMA, each component is configured by a descriptor in XML. In the following we describe how the descriptor required by this CR can be created with *Component Descriptor Editor*, an Eclipse plugin which is part of the UIMA SDK.

A descriptor contains information on different aspects. The following subsection refers to each sub aspect of the descriptor which is, in the Component Descriptor Editor, a separate *tabbed page*. For an indepth description of the respective configuration aspects or tabs, please refer to the *UIMA SDK User's Guide*⁵, especially chapter 12 on “Component Descriptor Editor User's Guide”.

To define your descriptor go through each tabbed pages mentioned here, make your respective entries (especially in page *Parameter Settings* you will be able to configure JULIE LAB ACE READER to your needs) and save the descriptor as `ACEReadersDescriptor.xml`.

Overview This tab provides general informtion about the component. For the JULIE LAB ACE READER you need to provide the information as specified in Table 1.

Aggregate Not needed here, as this CR is a primitive.

Parameters See Table 2 for a specification of the configuration parameters of this CR. Do not check “Use Parameter Groups” in this tab.

Parameter Settings The specific parameter settings are filled in here. For each of the parameters defined in 3, add the respective values here (has to be done at least for each parameter that is defined as mandatory). See Table 3 for the respective parameter settings of this CR.

Type System On this page, go to *Imported Type* and add the *JULIE UIMA Type System*. (Use “Import by Location”).

Capabilities Nothing needs to be done here.

⁵<http://incubator.apache.org/uima/>

Subsection	Key	Value
Implementation Details	Implementation Language	JAVA
	Engine Type	Primitive
Runtime Information	updates the CAS	yes
	multiple deployment allowed	yes
	outputs new CASes	no
	Name of the Java class file	de.julielab.jules.reader.AceReader
Overall Identification Information	Name	JULES-ACE-READER
	Version	2.0
	Vendor	julielab
	Description	see above

Table 1: Overview/General Settings for CR.

Parameter Name	Parameter Type	Mandatory	Multivalued	Description
inputDirectory	String	yes	no	Path to the ACE files
generateJulesTypes	Boolean	no	no	Determines if JULIE Lab Types (julie-semantics-ace-types.xml) should be generated in addition to types from julie-ace-types.xml

Table 2: Parameters of this CR.

Parameter Name	Parameter Syntax	Example
inputDirectory	valid path to the ACE files	resources/AceData
generateJulesTypes	boolean variable	true

Table 3: Parameter settings of this CR.

Index Nothing needs to be done here.

Resources Nothing needs to be done here.

4 Copyright and License

This software is Copyright (C) 2008 Jena University Language & Information Engineering Lab (Friedrich-Schiller University Jena, Germany), and is licensed under the terms of the Common Public License, Version 1.0 or (at your option) any subsequent version.

The license is approved by the Open Source Initiative, and is available from their website at <http://www.opensource.org>.

References

- [DMP⁺04] George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. The Automatic Content Extraction (ACE) Program: Tasks, data, & evaluation. In *LREC 2004 – Proceedings of the 4th International Conference on Language Resources and Evaluation. In Memory of Antonio Zampolli. Vol. 3*, pages 837–840. Lisbon, Portugal, 26-28 May 2004. Paris: European Language Resources Association (ELRA), 2004.