Extending an interoperable platform to facilitate the creation of multilingual and multimodal NLP applications

Georgios Kontonatsios*, Paul Thompson*, Riza Theresa Batista-Navarro*, Claudiu Mihăilă*, Ioannis Korkontzelos and Sophia Ananiadou
The National Centre for Text Mining,
School of Computer Science, The University of Manchester
131 Princess Street, Manchester M1 7DN, UK
{kontonag,batistar,thompsop,mihailac,
korkonti,ananiads}@cs.man.ac.uk

Abstract

U-Compare is a UIMA-based workflow construction platform for building natural language processing (NLP) applications from heterogeneous language resources (LRs), without the need for programming skills. U-Compare has been adopted within the context of the META-NET Network of Excellence, and over 40 LRs that process 15 European languages have been added to the U-Compare component library. In line with META-NET's aims of increasing communication between citizens of different European countries, U-Compare has been extended to facilitate the development of a wider range of applications, including both multilingual and multimodal workflows. The enhancements exploit the UIMA *Subject of Analysis* (*Sofa*) mechanism, that allows different facets of the input data to be represented. We demonstrate how our customised extensions to U-Compare allow the construction and testing of NLP applications that transform the input data in different ways, e.g., machine translation, automatic summarisation and text-to-speech.

1 Introduction

Currently, there are many repositories that contain a range of NLP components, e.g., OpenNLP[1], Stanford CoreNLP[2], JULIE NLP Toolsuite[3] and NaCTeM software tools[4]. The ability to chain components from these repositories into pipelines is a prerequisite to facilitate the development of complex NLP applications. Combining together heterogeneous components is not, however, always straightforward. The various components used in a pipeline may be implemented using different programming languages, may have incompatible input/output formats, e.g., stand-off or inline annotations, or may require or produce incompatible data types, e.g., a particular named entity recogniser (NER) may require specific types of syntactic constituents as input, making it important to choose the right type of syntactic parser to run prior to the NER. Thus, the tools required to build a new application may not be *interoperable* with each other, and considerable extra work may be required to make the tools talk to each other.

The Unstructured Information Management Architecture (UIMA) (Ferrucci and Lally, 2004) was created as a means to alleviate such problems. It is a framework that facilitates the straightforward combination of LRs, i.e., tools and corpora, into workflow applications. UIMA is an OASIS standard that enables interoperability of LRs by defining a standard workflow metadata format and standard input/output representations.

U-Compare (Kano et al., 2011) is a graphical NLP workflow construction platform built on top of UIMA. It facilitates the rapid construction, testing and evaluation of NLP workflows using drag-and-drop actions within its graphical user interface (GUI). U-Compare enhances interoperability among UIMA-compliant LRs, by defining a common and sharable *Type System*, i.e., a hierarchy of annotation types, which models a wide range of NLP data types, e.g., sentence, token, part-of-speech tag, named entity and discourse annotations. The aim is for all components in U-Compare's library to be compliant with this type system. In the context of META-NET, U-Compare's library has been extended with 46 new LRs supporting 15 European languages, all of which are compliant with the same type system.

*The authors have contributed equally to the development of this work and production of the manuscript.

[1] http://opennlp.sourceforge.net/projects.html

[2] http://nlp.stanford.edu/software/corenlp.shtml

[3] http://www.julielab.de/Resources/Software/NLP_Tools.html

[4] http://nactem.ac.uk/software.php

This makes U-Compare the world's largest repository of type system-compatible LRs, allowing users to seamlessly combine together resources to create a range of NLP applications.

Previously, U-Compare was able to support the development of a wide range of monolingual lexical, syntactic and semantic processing tasks applications that enriched textual input documents by adding annotations of various types. However, not all NLP applications operate in this way; some workflows *transform* the input data to create new "views" of the input data. The META-NET project aims to ensure equal access to information by all European citizens. This aim implies the development of both multilingual applications, which transform input data from one language into another, or multimodal applications, in which text may be transformed into speech, or vice versa.

U-Compare has been extended in several ways to support the construction of these more complex workflow types. Specifically, information about both the original and transformed data, together with annotations associated with each view, can now be visualised in a straightforward manner. The changes support two new categories of workflow. Firstly, workflows that produce two or more textual views of an input text are useful not only for multilingual applications, such as those that carry out machine translation, but also applications that transform the input text in other ways, such as those that produce a summary of an input text. Secondly, workflows that output audio as well as textual views, e.g., text-to-speech applications, are also supported.

2 Related work

Over the past few years, an increasing numbers of researchers have begun to create and distribute their own workflow construction architectures (Ferrucci and Lally, 2004; Cunningham et al., 2002; Grishman et al., 1997; Schäfer, 2006) or platforms (Kano et al., 2011; Rak et al., 2012; Ogrodniczuk and Karagiozov, 2011; Savova et al., 2010) that allow the rapid development of NLP applications.

GATE (Cunningham et al., 2002) is a workflow construction framework that has been used to develop several types of NLP applications, including summarisation systems. It facilitates the development of a wide range of NLP applications by providing a collection of components that can process various languages, together with Java libraries that handle character encoding for approximately 100 languages. However, GATE does not formally define any standards to model multilingual or multimodal applications, but rather aims to boost the development process of NLP applications.

TIPSTER (Grishman et al., 1997) is a generic framework for the development of NLP applications. TIPSTER provides multilingual functionalities by associating text segments of a parallel document with one or more languages. This allows language-dependent NLP components to process only the appropriate mono-lingual subdocuments. However, TIPSTER does not provide explicit guidelines regarding the annotation types and attributes that are produced by components. This lack of a common and sharable system of annotation types discourages interoperability between LRs. However, TIPSTER does not provide a mechanism that facilitates the development of multilingual or multimodal NLP applications.

Heart of Gold (Schäfer, 2006) is an XML-based workflow construction architecture that enables interoperability of tools developed in different programming languages to be combined into pipelines. Heart of Gold contains a rich library of shallow and deep parsing components supporting several languages, e.g., English, German, Japanese and Greek. Nonetheless, Heart of Gold does not specifically support the construction of multilingual or multimodal workflows.

In contrast to the other frameworks introduced above, UIMA (Ferrucci and Lally, 2004) provides an abstract-level mechanism that can be used to support the development of workflows that carry out transformations of the input data. This mechanism is called the *Subject of Analysis* or *Sofa*. Multiple Sofas can be linked with an input file, each of which stores different data and associated annotations. This mechanism can thus be exploited to represent alternative "views" of the input data, such as a source text and its translation. The data stored in different Sofas is not restricted to textual information; it can also correspond to other modalities, such as audio data. This makes the Sofa mechanism equally suitable for storing the output of text-to-speech workflows. Our extensions to U-Compare are thus implemented by reading and displaying the contents of different types of Sofas.

The Sofa mechanism has previously been

under-exploited by UIMA developers, despite its power in allowing more complex NLP workflows to be constructed. Indeed, no other existing UIMA-based platform (Kano et al., 2011; Rak et al., 2012; Savova et al., 2010; Hahn et al., 2008) has demonstrated the use of Sofas to construct multilingual or multimodal applications. Thus, to our knowledge, our enhancements to U-Compare constitute the first attempt to make the construction of workflows that carry out transformations of input data more readily available to UIMA users, without the need for programming skills.

3 METANET4U Components in U-Compare

The two dozen national and many regional languages of Europe present linguistic barriers that can severely limit the free flow of goods, information and services. The META-NET Network of Excellence was created to respond to this issue. Consisting of 60 research centres from 34 countries, META-NET has aimed to stimulate a concerted, substantial and continent-wide effort to push forward language technology research and engineering, in order to ensure equal access to information and knowledge for all European citizens.

META-NET's aims are dependent on the ready availability of LRs that can carry out NLP and text mining (TM) on a range of European languages. Such resources constitute the building blocks for constructing language technology applications that can help European citizens to gain easy access to the information they require. One of the major outcomes of META-NET has been the development of META-SHARE, an open, distributed facility for sharing and exchange of LRs in a large number of European languages.

Within the context of META-NET, interoperability of LRs is clearly of utmost importance, to expedite the process of developing new NLP applications. In order to provide a concrete demonstration of the utility and power of promoting interoperability within META-SHARE, one of the sub-projects of META-NET, i.e., METANET4U, has carried out a pilot study on interoperability, making use of the UIMA framework and the U-Compare platform. It is in this context that a set of 46 new LRs, available in META-SHARE, were wrapped as UIMA components and made available in U-Compare. Of these components, 37 operate on one or more specific languages other than English and 4 are language-independent. Table 1 shows the full set of categories of UIMA components created during the METANET4U project, together with the languages supported.

Several of these new components output multiple Sofas, i.e., two machine translation components, two automatic summarisation components and a text-to-speech component. It is hoped that our U-Compare extensions will help to stimulate the development of a greater number of related UIMA components, and thus promote a new level of complexity for future UIMA workflows.

Table 1: METANET4U UIMA components

Component Function	Supported Languages
Language Identifier	54 modern languages
Paragraph breaker	pt, mt
Sentence splitter	en, pt ,mt, es, ca, ast, cy, gl, it
Tokeniser	en, pt, mt, es, ca, ast, cy, gl, it, fr
Morphological Analyser	en, pt, es, ca, ast, cy, gl, it, ro, eu, fr
POS Tagger	en, es, ca, cy, gl, it, pt, ro, eu, fr, mt
Syntactic chunker	en, es, ca, gl, ast, ro, fr
NP chunker	ro
Segmenter	ro, en
FDG Parser	ro
Dependency Parser	en, es, ca, gl, ast
Discourse Parser	ro
NER	Language independent, trainable
Summariser	ro, en
Machine translation	en↔es, es↔gl, es↔pt, es↔ca, eu→es

4 Enhancements to U-Compare

In UIMA, an artefact, i.e., raw text, audio, image, video, and its annotations, e.g., part-of-speech tags, are represented in a standard format, namely the *Common Analysis Structure* (CAS). A CAS can contain any number of smaller sub-CASes, i.e., Sofas, that carry different artefacts with their linked annotations. Figure 1 illustrates the different types of Sofas that are created by the three types of workflows that we will demonstrate. Firstly, for a machine translation workflow, at least two CAS views, i.e., Sofas, are created, the first

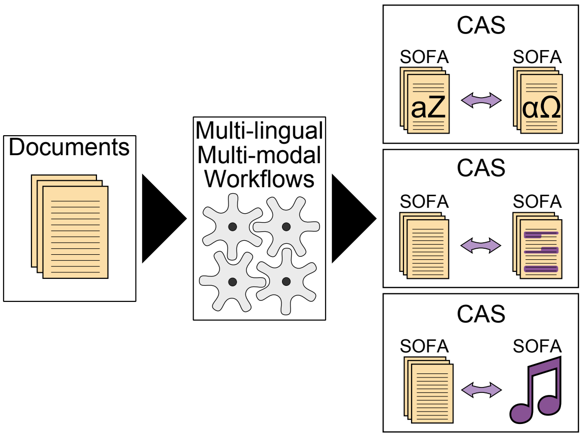


Figure 1: UIMA based multilingual and multimodal workflow architecture

corresponding to the text in the source language, and the other Sofas corresponding to the translation(s) of the source text into target language(s). The second type of workflow, i.e., automatic summarisation, is related to the former workflow, in that the two Sofas produced by the workflow are both textual, one containing the input text and one containing a summary of the original text. The third type of workflow is different, in that a Sofa containing audio data is used to represent the output of a multimodal workflow.

Two specific extensions have been made to U-Compare to handle both textual and audio Sofas. When the output of a workflow consists of multiple textual views (Sofas), the default annotation viewer is automatically split to allow multiple views of the text to be displayed and side-by-side. This can be useful, e.g., to allow careful comparison of a source text and target translation in a machine translation workflow. To handle audio Sofas, we have developed a new, customised viewer that can visualise and play audio data. The visualisation consists of a graphical display of the waveform, power information and spectrogram, as well as segmentation of the audio data into regions (such as individual tokens) and transcriptions, if such information is present in the audio Sofa. The viewer makes use the open-source library Java Speech Toolkit (JSTK)[5].

5 Workflow applications

In order to provide a practical demonstration of the enhanced capabilities of U-Compare, we show three different workflows that transform the input data in different ways, namely translation, automatic summarisation and speech synthesis. In this section, we provide brief details of these workflows.

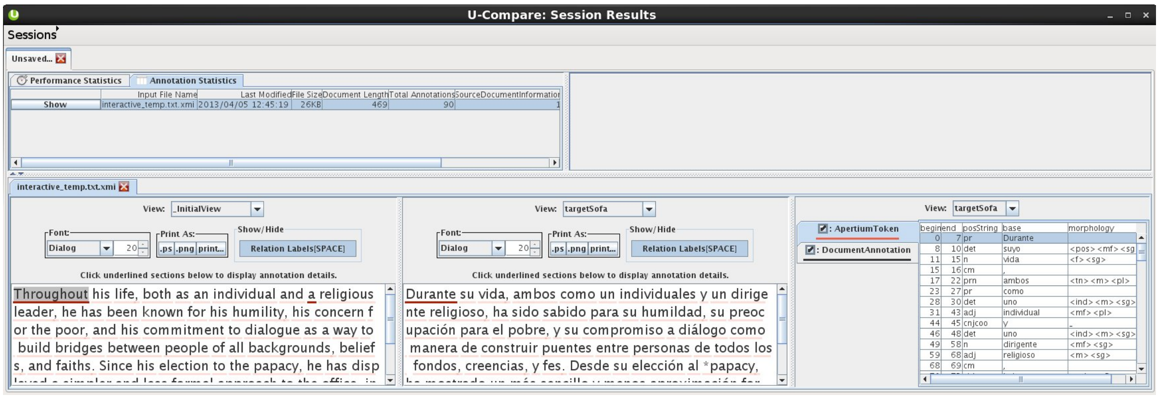
5.1 Machine translation

The University of Manchester has created UIMA wrapper components corresponding to different modules of Apertium (Corbí-Bellot et al., 2005), a free rule-based machine translation engine. These components consist of a morphological analyser, POS tagger and translator. The three components must be run in sequence to carry out translation, although the first two components can be used in other workflows to carry out monolingual analyses. The UIMA components currently handle a subset of the 27 languages dealt with by the complete Apertium system, corresponding to the languages of the METANET4U partners, i.e., English↔Spanish, Galician↔Spanish, Portuguese↔Spanish, Catalan↔Spanish and Basque→Spanish. However, additional language pairs can be added straightforwardly. Our sample workflow includes as its initial component the Language Identifier from the Romanian Academy Research Institute for Artificial Intelligence (RACAI), to automatically detect the language of the text in the input Sofa. The subsequent components in the workflow are the Apertium modules. The workflow demonstrates how heterogeneous components from different research groups can be combined into workflows to create new NLP applications. A sample output from running the workflow is shown in Figure 2. The input text was detected as English by the RACAI Language Identifier. The English text was subsequently analysed by the morphological analyser and POS Tagger, and translated to Spanish by the translator. Figure 2 illustrates the side-by-side display of the contents of the two Sofas.

5.2 Automatic summarisation

Automatic summarisation for Romanian text can be carried out by creating a workflow consisting of two components developed by the Universitatea "Alexandru Ioan Cuza" din Iaşi (UAIC). Firstly, a segmenter (UAICSeg) splits the input text into fragments, which are in turn used as input to the summariser component (UAICSum). The length of the output summary (percentage of the whole document) is parameterised. As can be seen in Figure 3, the output of this workflow is displayed

[5] http://code.google.com/p/jstk



	Input File Name	Last Modified	File Size	Document Length	Total Annotations	SourceDocumentInformation
Show	interactive_temp.txt.xmi	2013/04/05 12:45:19	26KB	469	90	1

View: _InitialView

Font: Dialog 20

Click underlined sections below to display annotation details.

Throughout his life, both as an individual and a religious leader, he has been known for his humility, his concern f or the poor, and his commitment to dialogue as a way to build bridges between people of all backgrounds, belief s, and faiths. Since his election to the papacy, he has disp

View: targetSofa

Font: Dialog 20

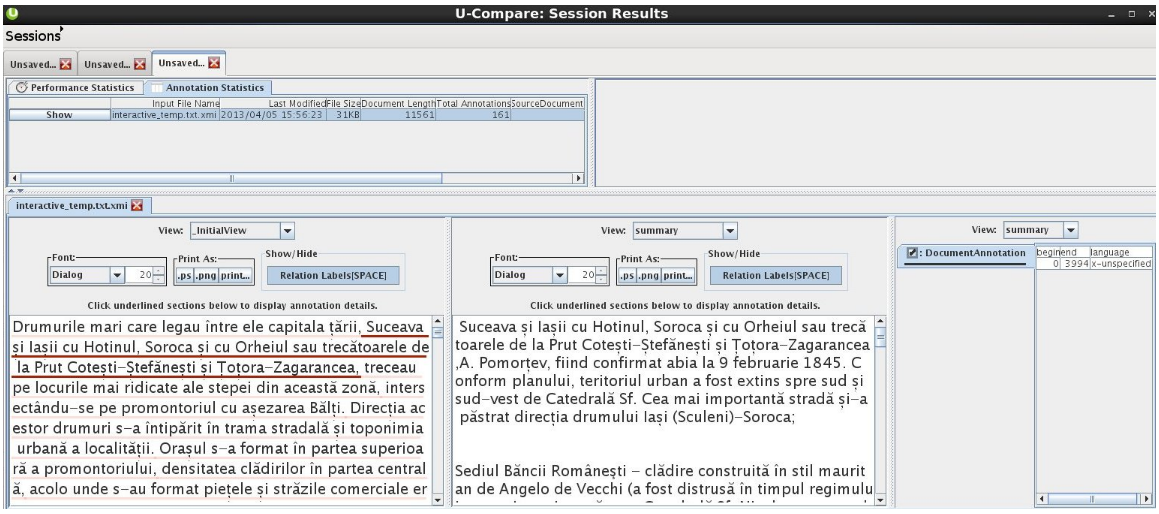
Click underlined sections below to display annotation details.

Durante su vida, ambos como un individuales y un dirige nte religioso, ha sido sabido para su humildad, su preoc upación para el pobre, y su compromiso a diálogo como manera de construir puentes entre personas de todos los fondos, creencias, y fes. Desde su elección al *papacy,

View: targetSofa

begin	end	posString	base	morphology
0	7	pr	Durante	
8	10	det	suyo	<pos><mf><sg
11	15	n	vida	<f><sg>
15	16	cm	,	
17	22	prn	ambos	<tn><m><pl>
23	27	pr	como	
28	30	det	uno	<ind><m><sg>
31	43	adj	individual	<mf><pl>
44	45	cnjcoo	y	_
46	48	det	uno	<ind><m><sg>
49	58	n	dirigente	<mf><sg>
59	68	adj	religioso	<m><sg>
68	69	cm	,	

Figure 2: Translation of English text to Spanish



	Input File Name	Last Modified	File Size	Document Length	Total Annotations	SourceDocument
Show	interactive_temp.txt.xmi	2013/04/05 15:56:23	31KB	11561	161	

View: _InitialView

Font: Dialog 20

Click underlined sections below to display annotation details.

Drumurile mari care legau între ele capitala țării, Suceava și Iașii cu Hotinul, Soroca și cu Orheiul sau trecătoarele de la Prut Cotești–Ștefănești și Țoțora–Zagarancea, treceau pe locurile mai ridicate ale stepei din această zonă, inters ectându–se pe promontoriul cu așezarea Bălți. Direcția ac estor drumuri s–a întipărit în trama stradală și toponimia urbană a localității. Orașul s–a format în partea superioa ră a promontoriului, densitatea clădirilor în partea central ă, acolo unde s–au format piețele și străzile comerciale er

View: summary

Font: Dialog 20

Click underlined sections below to display annotation details.

Suceava și Iașii cu Hotinul, Soroca și cu Orheiul sau trecă toarele de la Prut Cotești–Ștefănești și Țoțora–Zagarancea ,A. Pomorțev, fiind confirmat abia la 9 februarie 1845. C onform planului, teritoriul urban a fost extins spre sud și sud–vest de Catedrală Sf. Cea mai importantă stradă și–a păstrat direcția drumului Iași (Sculeni)–Soroca;

Sediul Băncii Româneşti – clădire construită în stil maurit an de Angelo de Vecchi (a fost distrusă în timpul regimulu

View: summary

begin	end	language
0	3994	x–unspecified

Figure 3: Summarisation of Romanian text

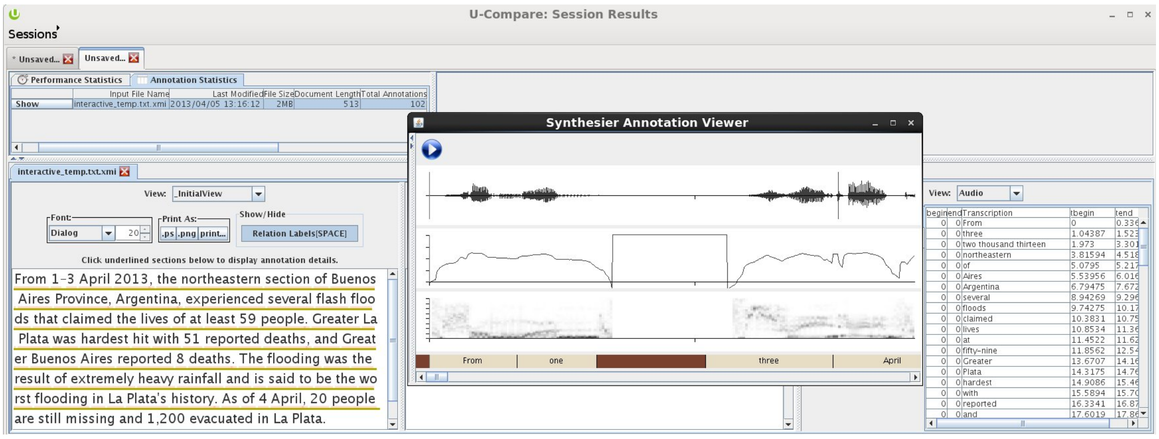
using the same parallel Sofa viewer. In this case, the full text is displayed in the left-hand pane and the summary is shown in the right-hand pane.

5.3 Speech synthesis

The Universitat Politècnica de Catalunya (UPC) developed a speech synthesiser component that is based around their Ogmios text-to-speech system (Bonafonte et al., 2006). The UIMA component version of this tool generates separate text and audio Sofas; the former stores the textual tokens and textual representations of their pronunciations, whilst the latter stores the start and end time offsets of each of the tokens in the audio file, together with their transcriptions. Fig. 4 shows how the textual Sofa information is displayed in U-Compare's default annotation viewer, whilst the audio Sofa information is shown in the new audio visualiser mentioned above. The three different types of visual information are displayed below each other, and the segments (tokens) of the audio file, together with their transcriptions, are displayed at the bottom of the window. A "Play" button allows either the complete file or a selected segment to be played.

6 Conclusions

The requirements of META-NET have motivated several new enhancements to the U-Compare platform, which, to our knowledge, make it the first UIMA-based workflow construction platform that is fully geared towards the development of NLP applications that support a wide range of European languages. The 46 new UIMA-wrapped LRs that have been made available through U-Compare, supporting 15 different European languages and all compliant with the same type system, mean that the improved U-Compare is essentially a hub of multilingual resources, which can be freely and flexibly combined to create new workflows. In addition, our enhancements to U-Compare mean



	Input File Name	Last Modified	File Size	Document Length	Total Annotations
Show	interactive_temp.txt.xmi	2013/04/05 13:16:12	2MB	513	102

View: _InitialView

Font: Dialog 20

Print As: .ps .png print...

Show/Hide: Relation Labels[SPACE]

Click underlined sections below to display annotation details.

From 1-3 April 2013, the northeastern section of Buenos
Aires Province, Argentina, experienced several flash floo
ds that claimed the lives of at least 59 people. Greater La
Plata was hardest hit with 51 reported deaths, and Great
er Buenos Aires reported 8 deaths. The flooding was the
result of extremely heavy rainfall and is said to be the wo
rst flooding in La Plata's history. As of 4 April, 20 people
are still missing and 1,200 evacuated in La Plata.

View: Audio

begin	end	Transcription	tbegin	tend
0	0	From	0	0.336
0	0	three	1.04387	1.523
0	0	two thousand thirteen	1.973	3.301
0	0	northeastern	3.81594	4.518
0	0	of	5.0795	5.217
0	0	Aires	5.53956	6.016
0	0	Argentina	6.79475	7.672
0	0	several	8.94269	9.296
0	0	floods	9.74275	10.17
0	0	claimed	10.3831	10.75
0	0	lives	10.8534	11.36
0	0	at	11.4522	11.62
0	0	fifty-nine	11.8562	12.54
0	0	Greater	13.6707	14.16
0	0	Plata	14.3175	14.76
0	0	hardest	14.9086	15.46
0	0	with	15.5894	15.70
0	0	reported	16.3341	16.87
0	0	and	17.6019	17.86

Figure 4: Speech Synthesis

that various types of multilingual and multimodal workflows can now be created with the minimum effort. These enhancements are intended to make U-Compare more attractive to users, and to help stimulate the development of a new generation of more complex UIMA-based NLP applications. As future work, we intend to extend the library of components that output multiple Sofas, and further extend the functionalities of U-Compare to handle other data modalities, e.g., video.

Acknowledgements

This work was partially funded by the European Community's Seventh Framework Program (FP7/2007-2013) [grant number 318736 (OSS-METER)]; MetaNet4U project (ICT PSP Programme) [grant number 270893]; and Engineering and Physical Sciences Research Council [grant numbers EP/P505631/1, EP/J50032X/1].

References

A. Bonafonte, P. Agüero, J. Adell, J. Pérez, and A. Moreno. 2006. Ogmios: The upc text-to-speech synthesis system for spoken translation. In *TC-STAR Workshop on Speech-to-Speech Translation*, pages 199–204.

A. Corbí-Bellot, M. Forcada, S. Ortiz-Rojas, J. Pérez-Ortiz, G. Ramírez-Sánchez, F. Sánchez-Martínez, I. Alegria, A. Mayor, and K. Sarasola. 2005. An open-source shallow-transfer machine translation engine for the romance languages of Spain. In *Proceedings of the 10th Conference of the EAMT*, pages 79–86.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: an architecture for development of robust HLT applications.

D. Ferrucci and A. Lally. 2004. Building an example application with the unstructured information management architecture. *IBM Systems Journal*, 43(3):455–475.

R. Grishman, B. Caid, J. Callan, J. Conley, H. Corbin, J. Cowie, K. DiBella, P. Jacobs, M. Mettler, B. Ogden, et al. 1997. TIPSTER text phase ii architecture design version 2.1 p 19 june 1996.

U. Hahn, E. Buyko, R. Landefeld, M. Mühlhausen, M. Poprat, K. Tomanek, and J. Wermter. 2008. An overview of JCoRe, the JULIE lab UIMA component repository. In *LREC'08 Workshop 'Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP'*, pages 1–7, Marrakech, Morocco, May.

Y. Kano, M. Miwa, K. Cohen, L. Hunter, S. Ananiadou, and J. Tsujii. 2011. U-compare: A modular nlp workflow construction and evaluation system. *IBM Journal of Research and Development*, 55(3):11.

M. Ogrodniczuk and D. Karagiozov. 2011. Atlas - the multilingual language processing platform. *Procesamiento de Lenguaje Natural*, 47(0):241–248.

R. Rak, A. Rowley, W. Black, and S. Ananiadou. 2012. Argo: an integrative, interactive, text mining-based workbench supporting curation. *Database: The Journal of Biological Databases and Curation*, 2012.

G. Savova, J. Masanz, P. Ogren, J. Zheng, S. Sohn, K. Kipper-Schuler, and C. Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

U. Schäfer. 2006. Middleware for creating and combining multi-dimensional nlp markup. In *Proceedings of the 5th Workshop on NLP and XML: Multi-Dimensional Markup in Natural Language Processing*, pages 81–84. ACL.