Second Upload of Language Resources

Deliverable D4.5

Version 0.5

2012-07-30

METANET4U

www.metanet4u.eu

The central objective of the Metanet4u project is to contribute to the establishment of a pan-European digital platform that makes available language resources and services, encompassing both datasets and software tools, for speech and language processing, and supports a new generation of exchange facilities for them.
This central objective is articulated in terms of the following main goals:
Assessment: to collect, organize and disseminate information that permits an updated insight into the current status and the potential of language related activities, for each of the national and/or language communities represented in the project. This includes organizing and providing a description of: language usage and its economic dimensions; language technologies and resources, products and services; main actors in different areas, including research, industry, government and society in general; public policies and programs; prevailing standards and practices; current level of development, main drivers and roadblocks; etc.
Collection: to assemble and prepare language resources for distribution. This includes collecting languages resources; documenting these language resources; upgrading them to agreed standards and guidelines; linking and cross-lingual aligning them where appropriate.
Distribution: to distribute the assembled language resources through exchange facilities that can be used by language researchers, developers and professionals. This includes collaboration with other projects and, where useful, with other relevant multi-national forums or activities. It also includes helping to build and operate broad inter-connected repositories and exchange facilities.
Dissemination: to mobilize national and regional actors, public bodies and funding agencies by raising awareness with respect to the activities and results of the project, in particular, and of the whole area of language resources and technology, in general.

METANET4U is a project in the META-NET Network of Excellence, a cluster of projects aiming at fostering the mission of META. META is the Multilingual Europe Technology Alliance, dedicated to building the technological foundations of a multilingual European information society.

METANET4U is co-funded by the participating institutions and the ICT Policy Support Programme of the European Commission

and by the participating institutions:

Faculty of Sciences, University of Lisbon

Instituto Superior Técnico

University of Manchester

University *Alexandru Ioan Cuza*

Research Institute for Artificial Intelligence, Romanian Academy

University of Malta

Technical University of Catalonia

Universitat Pompeu Fabra

Revision History

Version	Date	Author	Organisation	Description
v 0.1	July 12, 2012	Georgiana Gilmeanu, Jan Joachimsen, Mike Rosner	UOM	Draft
V 0.2	July 23, 2012	Georgiana Gilmeanu, Jan Joachimsen	UOM	Draft
V 0.3	July 25, 2012	Georgiana Gilmeanu, Jan Joachimsen	UOM	Draft
V 0.4	July 29, 2012	Mike Rosner	UOM	Draft
V 0.5	July 30, 2012	Georgiana Gilmeanu, Jan Joachimsen, Mike Rosner	UOM	Final

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

METANET4U

Second Upload of Language Resources

Document METANET4U-2011-D4.5
EC CIP project #270893

Deliverable D4.5

Completion: Final
Status: Submitted
Dissemination level: Restricted to Project Partners

Responsible: Mike Rosner (WP4 coordinator)

Contributing Partners: FCUL, IST, UNIMAN, UAIC, RACAI, UOM, UPC, UPF

Authors: Georgiana Gilmeanu, Jan Joachimsen, Mike Rosner

Reviewers: Fernando Batista, Thomas Pellegrini

© all rights reserved by FCUL on behalf of METANET4U

Contents

1 Introduction 7

2 Infrastructure among partners 7

2.1 Developments leading to the milestone Batch 2 7

2.1.1 META-SHARE 7

2.1.2 IPR 9

2.1.3 Metadata schema 9

2.2 Work plan 10

2.3 Validation Process 11

2.4 Implementation of the work plan 11

2.5 Achievements 11

Progress Report of Individual Partners on the Installation of the updated META-SHARE software v2.1.1 11

3 Upload of Batch 2 of Resources 12

3.1 Procedures 12

3.2 Resources uploaded for Batch 2 12

3.3 Implementation of the uploading plan 21

3.4 Number of resources in local nodes vs. in deliverable report 26

4 Conclusions 27

5 References 27

6 Appendix 28

6.1 Appendix 1: Complete Narrative Descriptions 28

6.2 Appendix 2: Quick Validation Report 29

1 Introduction

This deliverable deals with the upload of Batch 2 of language resources to the META-SHARE platform. Leading up to this upload, several other steps were involved, i.e., updates of the META-SHARE software, the metadata schema and IPR templates. These will be described in this document as well.

2 Infrastructure among partners

2.1 Developments leading to the milestone Batch 2

In preparation of the upload for Batch 2, several updates were made in three areas: the META-SHARE software, the IPR schema and the metadata schema. The following subsections will summarize the respective changes in these areas.

2.1.1 META-SHARE

For the upload of Batch 1, project partners used a version of the META-SHARE software that was announced as not being the final version. Instead, future upgrades were planned in stages until the upload of Batch 2.

The chronological order of the respective version updates was as follows:

- Version 1.1, released at the end of M10
- Version 2.0 beta, released at the beginning of M14
- Version 2.1, released mid M16
- Versions 2.1.1 and 2.1.2 released at the beginning of M18

Version 1.1 included the following updates over version 1.0:

- support for extended metadata format, covering representations for audio, audio/text, language description, lexicon, text corpus, and tool/service language resources;
- full import/export of XML resources, including a migration path of resources entered via the 1.0 editor, as well as import of schema-compliant descriptions in XML format;
- a unit test-suite allowing the system administrator to verify whether the local installation behaves as expected;

- many minor improvements, including editor robustness and usability, performance improvements, extended filtering in browse mode, and various bug fixes;
- extensions to the installation manual, including a detailed description of XML import/export/migration as well as a new Frequently Asked Questions section, drawing on the many requests for help that the helpdesks received over the month before the release of v 1.0.

Version 2.0 beta was the first open source release of the META-SHARE software. It provided all necessary features to set up a META-SHARE node, to fill it with metadata for language resources (LRs) and to make LRs available for download. The emphasis on the development of v 2.0 beta was on:

- extending the metadata schema;
- rewriting and cleaning up of the program code for the open source release, which included:
  - a complete rewrite of the Editor,
  - an improved search index;
  - a complete redesign of the object model implementing the metadata schema.

Due to the "beta" nature of v 2.0, it was not recommended for production use, therefore not all partners upgraded to this version from v 1.1. It also lacked a migration path from the older v1.1 version of the META-SHARE software. However, a bugfix release v2.1 was announced for the following weeks. For the following release, new features were announced, such as

- synchronization,
- a sophisticated user rights management,
- further improvements of the editor,
- recommendation services,
- support for adding third-party contributions to the code base.

After the release of v2.0, T4ME sent to every PSP a set of metadata records converted to the updated metadata schema in META-SHARE v2.0, for the language resources previously imported by every project into the META-SHARE repository.

Additionally, a XSLT converter was implemented into META-SHARE v2.0, which would convert metadata files in v1.1 format to the new v2.0 format.

Version 2.1 was released in May 2012 as ready for production use. Improvements over version 2.0 beta concentrated on:

- improving the usability of the metadata editor;
- improving the usability of browse, search and filter functionality;
- providing an upgrade path from previous versions of META-SHARE

It was recommended to all users of META-SHARE to upgrade to this v 2.1 at their earliest convenience.

Version 2.1.1 was a bug fix release for v2.1. Due to a bug in v2.1.1, T4ME made a new minor release update quickly after its release, v2.1.2. As the version number suggests, v2.1.2 is almost the same as v2.1.1; it just contains a fix for the aforesaid bug. The download for v2.1.1 has therefore been removed from the download page.

It was recommended that all users of META-SHARE upgrade to this version. Information on the progress of upgrading all existing META-SHARE nodes to the latest version is included in section 2.5 of this report.

2.1.2 IPR

The META-SHARE licences, too, were upgraded around the time that Batch 1 was uploaded. T4ME released v1.0 of the licence package in the end of November 2011 (M10), as a response to comments and suggestions made during the Athens workshop (October 2011) and the interaction cycle which followed after that.

The changes consisted in adding additional licence types for resources not to be redistributed and for commercial distribution. The upgraded licence set therefore consisted of:

- 4 META-SHARE non-commercial "No Redistribution" licences
- 4 META-SHARE commercial "No Redistribution" licences
- 6 META-SHARE Commons licences

Even though T4ME had pointed out that these licences could be regarded as provisional versions subject to discussion, the licences were only agreed to be final by the project partners at the beginning of March 2012 (M14).

2.1.3 Metadata schema

Together with the META-SHARE software v2.0 beta, a version v2.0 of the META-SHARE metadata model was released in March 2012. It included several changes from the previous version v1.1, namely:

- the extension of the metadata schema to cover remaining LR/media types (e.g. multimodal resources, n-gram resources; for details compare Deliverable D4.1 "Metadata Descriptions and other

Interoperability Standards as Agreed with META-NET and Partner Projects", Version 2.1)
- some modifications based on
  - discussions among the metadata core group,
  - suggestions made by PSP collaborating projects,
  - requirements following from the new approach to software implementation which automates the process of code generation from the metadata XML schema files.

By April 2012 (M15), the latest metadata schema version (v2.0) covered all languages resources and media types foreseen by the DoW. It incorporated most of the issues brought up by the collaborating project partners during the previous phase with respect to general information and text and audio corpora.

2.2 Work plan

On the basis of the upgraded metadata schema and META-SHARE software, the following internal work plan was adopted in anticipation of a timely delivery of Batch 2. All dates refer to 2012.

March:

- Installation of new versions of META-SHARE as they were made available by T4ME;
- Distribution of new narrative description templates by RACAI (e.g. for speech, tools, multimodal databases).

June 1:

- introduction of metadata records by partners,
- drafting of narrative descriptions,
- delivering materials to RACAI for quick validation check (including both resources/tools and narrative descriptions)

This plan was later slightly adjusted, due the fact that newer versions of META-SHARE were released after June 1. This meant structural changes in existing META-SHARE nodes so that project partners had to re-enter the metadata for uploaded resources into the nodes.

July 6 (3 weeks before the deadline)

- RACAI produces a first version of report of quick validation check to allow partners to repair problems in their resources/tools

July 13 (2 weeks before deadline)

- partners complete eventual correction of metadata records and upload corrected resources

July 20 (1 week before deadline)

- UOM submits the final draft of Deliverable D4.5 "Second upload of language resources" to the project partner for reviewing.

2.3 Validation Process

The validation process, undertaken by RACAI, followed the same procedure as for Batch 1 (described thoroughly in Deliverable D4.3 "First Upload of Language Resources").

2.4 Implementation of the work plan

There were no significant deviations from the work plan (apart from those commented on in section 2.2). Slight delays were caused by additional conversion work for metadata entries between version 1 and 2 of META-SHARE.

2.5 Achievements

This section lists the projects partners' contributions for the upload of Batch 2 of resources. Additionally to individual preparation of resources and solving of IPR issues, the partners had to deal with the local installation of the META-SHARE software. The following list summarises the results of setting up and working with the META-SHARE software on the respective local servers.

Progress Report of Individual Partners on the Installation of the updated META-SHARE software v2.1.1

1. FCUL - A local version of META-SHARE v 2.1.1 was installed and can be accessed at http://194.117.20.131/

The repository available at http://194.117.20.147/ is still available and is running META-SHARE v1.1 for legacy and review reasons and hosts the batch 1 resources (which were imported to the v2.1 repository).

2. CLUL- The CLUL's language resources for Batch 2 are hosted at FCUL's local node and can be accessed at http://194.117.20.131/.

3. IST – The META-SHARE version 2.1.2 is running and can be accessed at http://metanet4u.l2f.inesc-id.pt

4. UNIMAN – UNIMAN's language resources for Batch 2 are hosted at FCUL's local node and can be accessed at http://194.117.20.131/.

5. UAIC - A local version of META-SHARE version 2.1.1 was installed and can be accessed at http://metashare.infoiasi.ro/.

6. RACAI – A local version of META-SHARE last version 2.1.2 was successfully installed and can be accessed at http://ws.racai.ro:9191.

7. UOM – The UOM's language resources for Batch 2 are hosted at FCUL's local node and can be accessed at http://194.117.20.131/.

8. UPC- UPC installed a local version of META-SHARE v 2.1.1 and can be accessed at http://metashare.talp.cat/. The University of Vigo (UVIGO) and Aholab from Basque Country University (EHU) installed as well the platform and are accessible at http://metashare.gts.uvigo.es and http://aholab.ehu.es/metashare/ respectively.

9. UPF: - A local version of META-SHARE version 2.1.2 was successfully installed and can be accessed at http://metashare.upf.edu/. It still has some of the issues of the previous version that do not allow uploading resource description files (as described in the metashare support website). They have been modified to allow such upload.

3 Upload of Batch 2 of Resources

3.1 Procedures

The changes in the META-SHARE software implemented an updated metadata schema, which required additional work from the project partners, i.e. metadata for resources had to be modified and, in some cases, had to be re-entered via the metadata editor.

Once this was achieved, the project partners sent their resources, metadata and narrative descriptions to RACAI for validation. For the complete narrative descriptions, refer to Annex 1. The results of this validation process are reproduced in Annex 2.

3.2 Resources uploaded for Batch 2

Based on Deliverable 2.5 Table 1 lists the resources that were uploaded for Batch 2. It states the names the project partners, their uploaded resources, the kind of each resource and the languages included in the

respective resource. The adjustments to the original uploading plan (see Deliverable D2.5) are showcased in Table 2 below.

New resources that were not uploaded to any of the project partners' META-SHARE nodes listed above given they were hosted at the ELRA META-SHARE node are marked in Table 1 with "*".

Resources that were not uploaded to any of the project partners' META-SHARE nodes listed above as they were previously hosted by the ELRA META-SHARE node are marked in Table 1 with the call to footnote 1.

Table 1: Resources uploaded for Batch 2

Partner		Name of Resource	Resource Type	Languages covered
1. ULX - University of Lisbon	Endogenous resources	LX-Abbreviations	lexicon, list of abbreviations	Portuguese
		C-ORAL-ROM Portuguese Corpus[1]	corpus, speech	Portuguese
		CINTIL-Internacional Corpus of Portuguese[1]	corpus, speech	Portuguese
		Lexicon of Multiword Expressions	lexicon	Portuguese
		MWN.PT[1]	wordnet	Portuguese
		PAROLE corpus[1]	corpus, annotated corpus	Portuguese
		PAROLE lexicon[1]	lexicon	Portuguese
		CINTIL-PropBank*	corpus, annotated corpus	Portuguese
		SIMPLE lexicon	lexicon	Portuguese
		LX-Stopwords	lexicon, list of stop words	Portuguese
		CINTIL-Treebank*	corpus, annotated corpus	Portuguese
	Exogenous	MBT (Memory-based tagger, generation and tagging)	LT tool, tagger	Language independent
		YamCha: Yet Another Multipurpose Chunk Annotator	LT tool, chunk annotator	Language independent

	resources	TinySVM	LT tool, tagger	Language independent
		Geo-Net-PT01	grammar, ontology	Portuguese
		Ontologia de Nanociência e Nanotecnologia	ontology	Portuguese
		Summ-it	corpus, annotated corpus	Portuguese
		Forma	LT tool, tagger	Portuguese
2. IST - Instituto Superior Técnico	Endogenous resources	TED Talks (3)	Corpus, speech	Portuguese / English
		PTSTAR Golden Collection (New name: CLUE)	Corpus, annotated corpus, multilingual word alignment	6 European languages
	Exogenous resources	UP/ TAP	corpus, text	Portuguese/ English
3. UNIMAN-University of Manchester	Endogenous resources	U-Compare NaCTeM Sentence Detector	LR tools, sentence splitter	English
		U-Compare Cafetiere sentence splitter	LR tools, sentence splitter	English
		U-compare platform	LR tools, platform	language-independent
		U-Compare Workbench	LR tools, NLP development environment	language-independent
	Exogenous resources	U-Compare GENIA Sentence Detector	LR tools, sentence splitter	English
		U-Compare GENIA Tokenizer	LR tools, tokenizer	English
		U-Compare OpenNLP PoSTagger	LR tools, PoS tagger	English
		U-Compare OpenNLP Sentence Detector	LR tools, sentence splitter	English
		U-Compare OpenNLP Tokenizer	LR tools, tokenizer	English, Spanish
		U-Compare Type System	services, specification of linguistic annotations	language-independent
		U-Compare Apertium	LR tools,	English,

		Morphological Analyser	morphological analyser	Spanish, Catalan, Galician, Basque, Portuguese
		U-Compare Apertium POS tagger	LR tools, tagger	English, Spanish, Catalan, Galician, Basque, Portuguese
		UIMA Apertium Translator[4]	LR tool, machine translation	English <-> Spanish, Spanish <-> Catalan, Spanish <-> Galician, Spanish <-> Portuguese, Basque -> Spanish
4. UAIC - University Alexandru Ioan Cuza	Endogenous resources	Categorizer-UAIC	LR tools, document category/ domain identification	English
		DP-UAIC	LR tools, discourse parser system	English/Romanian
		Lemmatizer-UAIC	LR tools, lemmatizer	Romanian
		NP-chunker-UAIC	LR tools, NP-chunker	Romanian
		RARE-RO-UAIC	LR tools, Robust rule-based Anaphora resolution system,	Romanian/ English
		Splitter-UAIC	LR tools, sentence splitter	Romanian/ English
		Summarizer-UAIC	LR tools, summarization system	Romanian/ English
		Tokenizer-UAIC	LR tools, tokenizer	Romanian/ English
5. RACAI – Romanian Academy	Endogenous resources	Mapping list from PWN2.0 to PWN3.0	wordnet synset alignments	English
		NAACL 2003	annotated parallel corpus	Romanian/ English

		ROMORPH	lexicon	Romanian
		RO-WORDNET (part 2)	lexical ontology, semantic dictionary	Romanian
		COLLOC	LR tools, collocation extractor	language independent
		LangId	LR tools, language identification	language independent
		LexChain	LR tools, lexical chain	language independent, wordnet available
		LexPar	LR tools, dependency linker	language independent
		RO-HYPHEN	LR tools, hyphenator	Romanian
		TTL-Lemmatizer	LR tools, lemmatizer	Romanian/ English/ French
		TTL-Tagger	LR tools, morpho-syntactic tagger	Romanian/ English/ French
		TTL-Tokenizer	LR tools, tokenizer	Romanian/ English/ French
		TTL-Chunker	LR tools, chunker	Romanian/ English/ French
6. UOM - University of Malta	Endogenous resources	F_MONA_1	speech corpus, speech data	Maltese
		MalToBI/SPAN Corpus	speech corpus	Maltese
	Exogenous resources	Local Government documentation	raw text corpus	Maltese/ English
		Maltese Wikipedia	text corpus	Maltese
		Basic English-Maltese Dictionary	bilingual wordlist	English/ Maltese
		MFSA Maltese Company Registry	lexicon, company names database	Maltese / English
7. UPC - Technical University of	Endogenous resources	ALBAYZIN[1]	annotated speech database	Catalan, Spanish
		TM2: Technical	corpus, multimodal,	Catalan,

Catalonia		Meetings	annotated	Spanish
		CHIL2007+[1]	multimodal database	English
		Catalan-SpeechDat(I)[1]	annotated speech database	Catalan
		SALA-Mexico[1]	annotated speech database	5 dialects of Spanish Mexican
		SALA-Venezuela[1]	annotated speech database	5 dialects of Spanish Venezuelan
		Spanish SpeechDat (II)[1]	annotated speech database	Spanish from Spain (5 dialects)
		SpeechDat-Car Spain[1]	annotated speech database	Spanish from Spain (5 dialects)
		Speecon Catalan[1]	annotated speech database	5 dialects of Catalan (in Catalonia)
		Interface Emotional Speech database Spanish[1]	annotated speech database	Spanish
		LC-STAR Catalan Phonetic Lexicon[1]	lexicon	Catalan
		LC-STAR Spanish Phonetic Lexicon[1]	lexicon	Spanish
		UPC-ESMA	annotated speech synthesis database	Spanish
		TC-STAR Spanish TTS Baseline Female 10h[1]	annotated speech synthesis database	Spanish
		TC-STAR Spanish TTS Baseline Male 10h[1]	annotated speech synthesis database	Spanish
		TC-STAR Bilingual Expressive Speech[1]	annotated speech database	Spanish/ English
		TC-STAR Bilingual VC[1]	annotated speech synthesis database	Spanish/ English
	Exogenous resources	Ahosyn_male_EU: Large Bilingual Speech Database for Synthesis[2]	annotated speech database	Basque

		Ahosyn_male_ES: Large Bilingual Speech Database for Synthesis[2]	Annotated speech database	Spanish
		Ahosyn_female_EU: Large Bilingual Speech Database for Synthesis[2]	annotated speech database	Basque
		Ahosyn_female_ES: Large Bilingual Speech Database for Synthesis[2]	Annotated speech database	Spanish
		Bizkaifon: speech and video database for the Western dialects of the Basque Language[1]	multimodal database	Basque
		Ahoemo1: Emotional speech and video database in Standard Basque[2]	multimodal database	Basque
		Ahoemo2: Emotional speech database in Standard Basque[2]	speech database	Basque
		Ahoemo3: Emotional speech database in Standard Basque[2]	speech database	Basque
		Galician SpeechDat FDB[3]	annotated speech database	Galician
		LAS CORTES[1]	annotated speech database	Spanish
		SPANISH EPPS[1]	speech, speech database	Spanish
		Speech-Dat like database for Basque[1]	annotated speech database	Basque
		MDB602EU Speech-dat like database for Basque (Mobile)[2].	annotated speech database	Basque
		Transgrigal DB[3]	annotated speech database	Galician
		DOGalicia: Parallel Galician-Spanish Corpus[3]	parallel corpus	Galician/ Spanish
		GCG: GrupoCorreoGalego	annotated speech database	Galician

		Cotovia Transcriber[3]	LR tools, transcriber	Galician
8. UPF - University Pompeu Fabra	Endogenous resources	Converters to LMF 2	LR tool	n/a
	Exogenous resources	TRL V-Subcat Lexicon	Lexicon	Spanish
		Tools for Catalan Corpus Processing	LR tools, corpus processing	Catalan
		Tools for Spanish Corpus Processing	LR tools, corpus processing	Spanish
		Apertium Portuguese dictionary in LMF	lexicon	Portuguese
		Parole/Simple LMF lexicon Catalan	lexicon	Catalan
		SenSem Corpus	corpus, text	Spanish
		SenSem Database (lexicon) Catalan	lexicon	Catalan
		SenSem Database (lexicon) Spanish	lexicon	Spanish
		CESS_EU: The Basque Dependency Treebank	annotated text written corpus	Basque
		Termoteca	lexica, terminological resource	English, French, Galician, Spanish
		Apertium English dictionary	lexicon	English
		Apertium French dictionary	lexicon	French
		Apertium Italian dictionary	lexicon	Italian
		WikiCorpus	Tagged corpus	Spanish/Catalan

Notes:

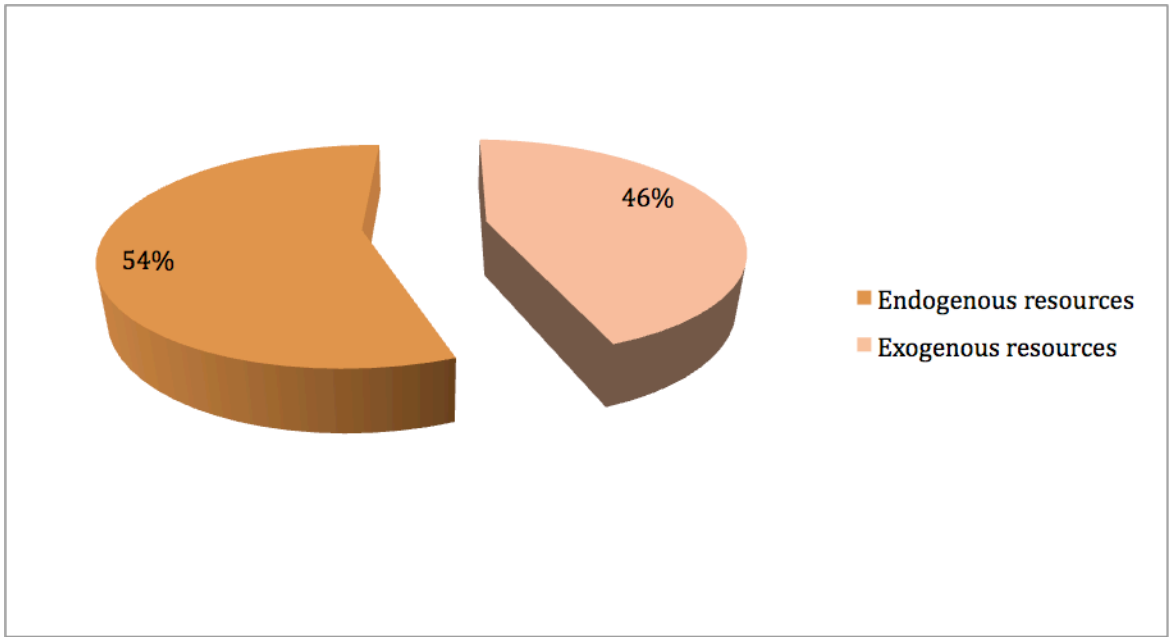
[1, *]: hosted at ELDA META-SHARE node.

[2]: hosted on Aholab META-SHARE node.

[3]: hosted on UVIGO META-SHARE node.

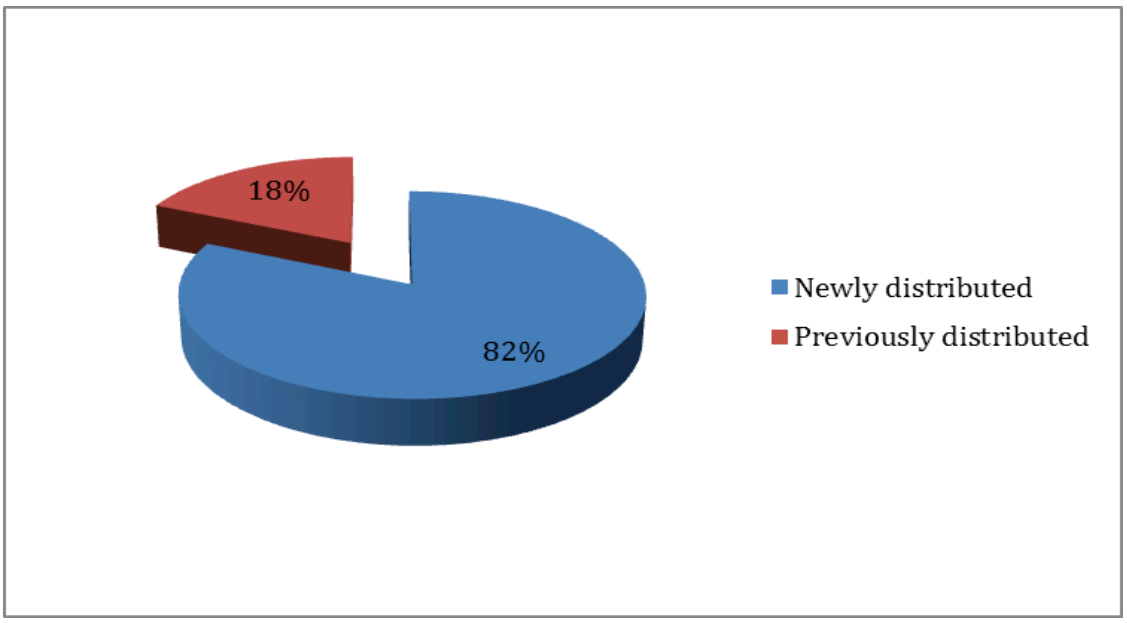
[4]: UNIMAN's UIMA Apertium Translator was provided instead of the originally planned separate "MT transfer" and "Morphological generator" components, as it was realised that it did not make sense to provide these separately.

The pie-chart below visualizes the ratio of endogenous to exogenous resources for Batch 2. The higher amount of endogenous resources is due to the fact that partners put more efforts into updating and correcting their endogenous resources. This was in accordance with the amount of efforts and human resources set out by WP3.



As can be seen in Table 1 (and also in Table 2 below), some resources in Batch2 are hosted in ELDA (Evaluations and Language resources Distribution Agency). This is also the case for some resources in Batch 1.

The figure below shows the percentages of new language resources hosted in META-SHARE nodes compared to those already available through distribution channels like ELDA or LDC before the project's activities.



3.3 Implementation of the uploading plan

Table 2 contains the adjustments to the list of promised language resources. In some cases, new resources originally not promised for Batch 2 were added to it (column 2 in Table 2). In other cases, the upload for some resources was postponed to Batch 3 (column 3), while some resources were pre-drawn from Batch 3 to have an early delivery already at Batch 2 (column 4).

Most of the adjustments concern the exogenous resources. There are several different reasons for these adjustments. While only some of them are delayed by the updating and documentation processes or by the fact that some providers didn't deliver their resources at the time of the second upload, the majority of delayed resources encountered problems related to the incipient stage of the META-SHARE software. Several issues were pointed out by the project partners, connected to the lack, incipient or yet not fully trustful status of e.g. user identification, single entry point/site, billing system for commercial resources or the synchronisation of nodes. These issues made some IPR owners hesitate to agree on licensing their resources for the use in META-SHARE.

In order not to delay the progressive population of META-SHARE, in turn some other resources were advanced from Batch 3 to Batch 2, and new resources were added (see columns New Resources and Pre-drawn from Batch 3 in Table 2).

Table 2: Adjustments to the uploading plan (D2.5)

Note: the resources marked with "END" are endogenous and the resources marked with "EXO" are exogenous.

Partner	New resources uploaded at batch 2 that were not in the uploading plan for this batch	Postponed to Batch 3	Pre-drawn from Batch 3	Cancelled Resources
1. ULX - University of Lisbon	MBT (Memory-based tagger, generation and tagging) (EXO) [1] YamCha: Yet Another Multipurpose Chunk Annotator (EXO) [1] TinySVM (EXO) [1]	Clássicos LP/Porto Editora (EXO) Corpus NILC (EXO) Dicionário de Verbos do Português Medieval (DVPM) (EXO) Glossário (EXO) MorDebe (EXO) Norma Urbana Culta (NURC) (EXO) PORLEX (EXO) Corpus NILC (EXO) CorpusTCC (EXO) PLN-BR Gold (EXO) RHETALHO (EXO) TeMário 2006 (EXO) DiZer 2.0 (EXO) GistSumm (EXO) NILC Taggers (EXO) Ontolp Plugin (EXO)	n.a.	European Parliament Parallel Corpus (Portuguese) (EXO) [4] JRC Acquis (Portuguese) (EXO) [4] COMPARA (EXO) CETEMPúblico (EXO) Panorama do Português Oral de Maputo (PPO) (EXO)

		Stemmer (EXO) Text Aligners(EXO)		
2. IST - Instituto Superior Técnico	n.a.	CALL(END)	n.a.	n.a.
3. UNIMAN-University of Manchester	U-Compare Cafetiere sentence splitter (END) UIMA Apertium Translator (EXO)[2]	n.a.	n.a.	NaCTeM sentence splitte (END)
4. UAIC - University Alexandru Ioan Cuza	n.a.	n.a.	n.a.	n.a.
5. RACAI - Romanian Academy	n.a.	Romanian WEB 1.5T (EXO)	n.a.	n.a.
6. UOM - University of Malta	n.a.	Maltese Speech Engine Corpus (EXO)	MFSA Maltese Company Registry (EXO)	Malta Online Dictionary (EXO)[5]
7. UPC - Technical University of Catalonia	TM2: Technical Meetings (END) CHIL2007+ (END)[3] Ahoemo3 (EXO) Former Ahosyn database has been splited in four databases: Ahosyn_male_EU (EXO), Ahosyn_male_ES (EXO), Ahosyn_female_EU (EXO), Ahosyn_female_ES (EXO) EmodB_EU1 is renamed Ahoemo1 (EXO) EmodB_EU1 is renamed Ahoemo2 (EXO)	EL_PERIODICO_97-07 (EXO)	n.a.	SpeechRate database for Basque (EXO)

8. UPF - University Pompeu Fabra	Apertium Portuguese dictionary in LMF (EXO) Parole/Simple LMF lexicon Catalan (EXO)	Newspaper headlines corpus (END) Corpus CLUVI (END) Corpus Técnico do Galego (END) Dicionario CLUVI inglés-galego (END) Euskal Wordnet 3.0 (END) Tools for automatic UTF-8 conversion (END)	SenSem Corpus (EXO) SenSem Database (lexicon) Catalan (EXO) SenSem Database (lexicon) Spanish (EXO)	n.a.

Notes:

[1]: These are new resources by ULX, which replace COMPARA, CETEMPúblico, Panorama do Português Oral de Maputo (PPO).

2 Two originally planned tools have been combined into one, i.e. "UIMA/U-Compare Apertium MT Transfer" and "UIMA/U-Compare Apertium Morphological Generator" have been combined into "UIMA Apertium Translator"

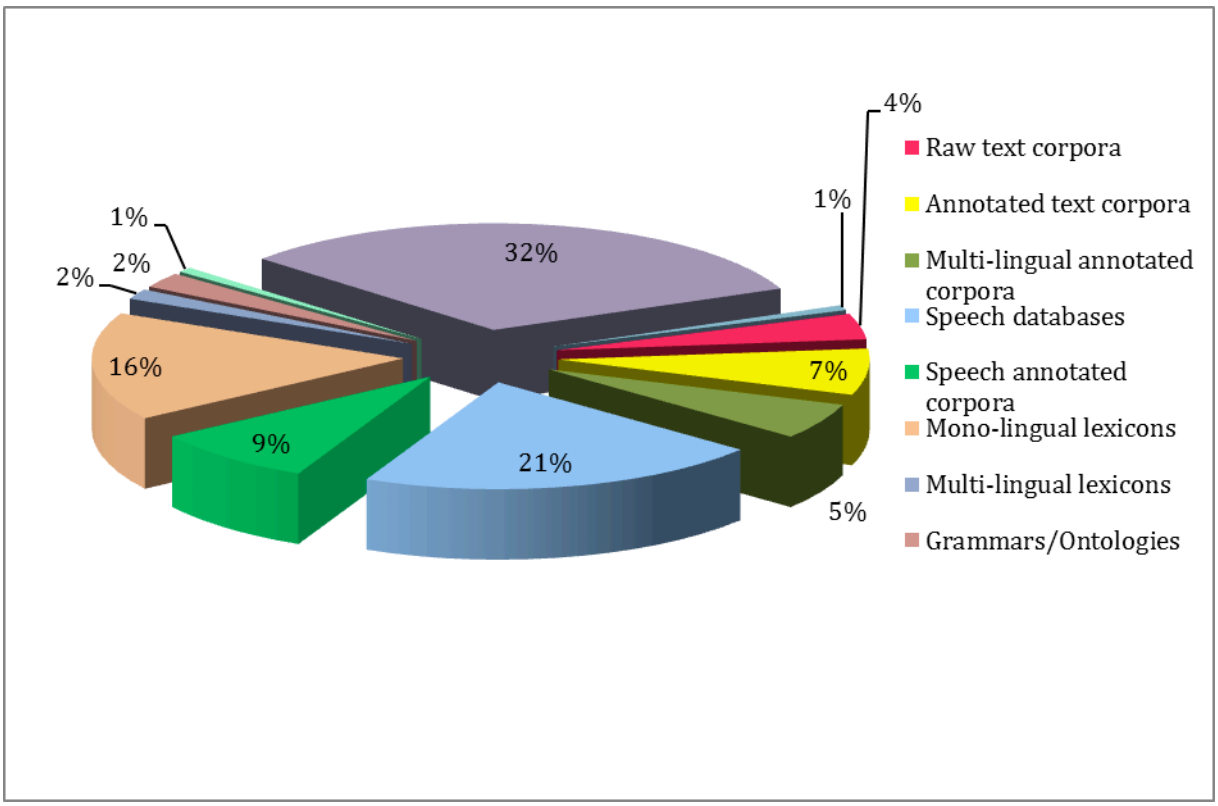
[3]: hosted at ELRA

[4]: These resources were cancelled, since JRC will be hosting these resources on its own META-SHARE node.

[5]: This resource was cancelled, since it turned out to be not suitable as a language resource in METANET4U.

In order to visualise the composition by resource type of the present upload, we present a pie-chart based on the corresponding attribute appearing in deliverable 2.5. The resource types can be summarised into 11 main categories:

1. raw text corpora
2. annotated text corpora
3. multilingual annotated corpora
4. speech databases
5. speech annotated corpora
6. monolingual lexicons
7. multilingual lexicons
8. grammars/ontologies
9. wordnets
10. LR/LT tools
11. services



As the pie-chart clearly shows, the majority of language resources in Batch 2 consists of LR tools, followed by speech databases and monolingual lexicons.

3.4 Number of resources in local nodes vs. in deliverable report

It is worth helping to understand the disparities that happen to occur between the number of resources which appear in D4.5 report and what appears in META-SHARE local nodes.

For instance, in UPF's case there are 48 resources reported as delivered in D4.3 and D4.5 but there are 58 resources uploaded on their local Meta-share node. This is because what is counted as a single resource in Batch 1 or 2 turns out to be a collection of resources with every item having its own metadata description.

For example the "Tools for Catalan Corpus Processing" resource, included in Table 1 for UPF is a zip file containing:

1. IULA lexicon look up Web Service
2. IULA paradigma Web Service
3. IULA processor
4. IULA tagger
5. IULA tagger-graph
6. IULA tokenizer

The same case applies for RACAI partner. The resource which on the META-SHARE node appears under the name "Tokenizing, Tagging and Lemmatizing free running texts" is a collection of 4 tools: TTL tokenizer, TTL tagger, TTL lemmatizer and TTL chunker.

They correspond to RACAI's list in Deliverable 2.5:

TTL---Lemmatizer

TTL---Tagger

TTL---Tokenizer

TTL---Chunker

For other partners (UNIMAN, UPC), a smaller number of resources in their local repositories is due to the fact that they also have resources hosted with ELDA/ELRA.

4 Conclusions

Compared to the upload of Batch 1, more of the anticipated goals were achieved. Moreover, the project partners were able to replaced cancelled resources with new resources.

The updates of the software and IPR licences during the preparation phase did not yet bring the full stable working routine for the META-SHARE software. However, the latest version permitted to upload almost all of the endogenous resources and most of the planned exogenous resources.

The main focus for Batch 3 will be to arrive at a final version of META-SHARE in order to accommodate the exogenous resources which were originally promised.

5 References

Gilmeanu, Georgiana; Joachimsen, Jan; Rosner, Mike (2011): *D4.3 – First Upload of Language Resources*

Moreno, Asunción (2012): *D2.5 – Report on second selection of resources, revising selection in D2.1*

Tufiş, Dan (2012): *Narrative Descriptions for the Resources delivered as BATCH 2.*

Tufiş, Dan (2012): *D3.2 - WP3: Delivery of the BATCH 2 of Resources Validation Report*

6 Appendixes

6.1 Appendix 1: Complete Narrative Descriptions

BATCH 2 Narrative Collection

Table of Narrative Documents

ULX - University of Lisbon 5

Endogenous resources 5

Abbreviations **5**

C-ORAL-ROM Portuguese Corpus **8**

CINTIL-Internacional Corpus of Portuguese **18**

Lexicon of multiword expressions **24**

PropBank **29**

Stopwords **35**

Treebank **39**

IST - Instituto Superior Técnico 45

Endogenous resources 45

TED Talks **45**

PTSTAR Golden Collection – Cross-Language Unit Elicitation alignments (CLUE) **48**

TAP **50**

UNIMAN-University of Manchester 53

Endogenous resources (tools) 53

U-Compare Platform **53**

U-Compare Workbench **57**

Restricted exogenous resources (tools) 62

U-Compare GENIA Sentence Detector **62**

U-Compare GENIA Tokenizer **65**

U-Compare OpenNLP PoSTagger **68**

U-Compare OpenNLP Sentence Detector **71**

U-Compare OpenNLP Tokenizer **74**

U-Compare Type System **77**

Unrestricted exogenous resources (tools) 82

Apertium Morphological Analyser **82**

Apertium POS tagger **88**

Newly added in Batch 2 95

U-Compare Cafetiere sentence splitter **95**

UAIC - University Alexandru Ioan Cuza 99
Endogenous resources (tools) 99
Categorizer-UAIC 99
Discourse Parser 103
Lemmatizer 105
NP Chunker 108
RARE 113
Splitter v1 and v2 117
Summarizer v1 and v2 121
Tokenizer-UAIC 123
RACAI - Romanian Academy 126
Endogenous resources 126
Mapping list from PWN2.0 to PWN3.0 126
NAACL 2003 128
ROMORPH 130
RO-WORDNET (part 2) 133
Endogenous resources (tools) 139
COLLOC 139
LangId 147
LexChain 149
LexPar 152
RO-HYPHEN 155
TTL Package (TTL-Lemmatizer, TTL-Tagger, TTL-Tokenizer, TTL-Chunker) 157
UOM - University of Malta 161
Endogenous resources 161
F-MONA 1 161
MaltoBI Corpus 163
Restricted Exogenous resources 167
Local Government documentation 167
Maltese Wikipedia 169
Unrestricted exogenous resources 171
Basic English-Maltese Dictionary 171
Newly added in Batch 2 173
MFSA Maltese Company Registration.zip 173
UPC – Universitat Politècnica de Catalunya 175

Endogenous resources 175
ALBAYZIN **175**
TM2: technical meetings **179**
CHIL2007+ **182**
Catalan SpeechDat (I) **186**
SALA-Mexico **189**
SALA- Venezuela **192**
Spanish SpeechDat (II) **196**
SpeechDat-Car Spanish **199**
Speecon Catalan **205**
LC-STAR Catalan Phonetic Lexicon **212**
LC-STAR Spanish Phonetic Lexicon **218**
UPC ESMA **224**
TC-STAR Spanish TTS Baseline male 10h **230**
TC-STAR Bilingual (Spanish English) VC **236**
Exogenous resources 239
UVIGO LR Description 239
Galician SpeechDat FDB **239**
Transcrigal DB **243**
DOGalicia (Parallel Galician-Spanish Corpus) **247**
GCG Corpus **250**
Cotovia Text-to-Speech System for Galician and Spanish **252**
The Basque University LR Description 254
Ahosyn male EU: Large Speech Database for Synthesis in Basque **254**
Ahosyn female EU: Large Speech Database for Synthesis in Basque **257**
Ahosyn male ES: Large Speech Database for Synthesis in Spanish **260**
Ahosyn female ES: Large Speech Database for Synthesis in Spanish **263**
Ahoemo1: Emotional speech and video database in Standard Basque **266**
Ahoemo2: Emotional speech database in Standard Basque **269**
Ahoemo3: Emotional speech database in Standard Basque **273**
MDB602EU: Basque SpeechDat like (Mobile Network) **276**
UPF - University Pompeu Fabra **280**
Endogenous resources 280
TRL V-Subcat Lexicon **280**
Restricted exogenous resources 282
CESS_EU: The Basque Dependency Treebank **282**
Termoteca **286**

Unrestricted exogenous resources 289
Apertium English dictionary 289
Apertium French dictionary 293
Apertium Italian dictionary 296
WikiCorpus 300
Endogenous resources (tools) 307
Converters to LMF 2 307
Tools for Catalan Corpus and Spanish Corpus Processing 310
Extra Resources (new or delivered now from Batch 3) 332
Apertium Portuguese dictionary in LMF 332
Parole/Simple LMF lexicon Catalan 336
SenSem Corpus 339
SenSem Database (lexicon) Catalan 343
SenSem Database (lexicon) Spanish 347

ULX - University of Lisbon

Endogenous resources

Abbreviations

1. BASIC INFORMATION

1.1 *Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc.)*

LX-Abbreviations resource is a collection of abbreviations of different types from European Portuguese composed by 208 words. Each abbreviation is annotated with grammatical categories, gender and number. Finally, abbreviations (see Branco & Silva, 2003) are grouped into types, as shown below:

LX-Abbreviations		
Types	Foreign abbreviations	4
	Nouns	5
	Units	3
	Possessives	1
	Personals	40
	Week days	7
	Months	12
	Social titles	125
	Parts for addresses	8
	Noun Phrases	3
Total	**10**	**208**

This resource was collected in the context of TagShare – Tagging and Shallow Tools and Resources project[1] with the following main goals: developing of a set of linguistic resources and software component tools to support the computational processing of Portuguese.

1.2 *Representation of the lexicon (flat files, database, markup)*

The corpus is represented in .txt format.

1.3 *Character encoding*

The characters are in UTF8 code.

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)

Name: António Branco

[1] It can be visited at http://tagshare.di.fc.ul.pt/.

Address: Departamento de Informática NLX − Grupo de Fala e Linguagem Natural, Faculdade de Ciências da Universidade de Lisboa, Edifício C6, Campo Grande 1749-016 Lisboa

Position: Assistant Professor

Affiliation: Faculty of Sciences, University of Lisbon

Telephone: +351 217 500 087

Fax: +351 217 500 084

E-mail: antonio.branco@di.fc.ul.pt

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be available on the META-SHARE platform.

2.3 Copyright statement and information on IPR

This resource is a free license-based for research and for commercial purposes, with attribution and no redistribution allowed. It will be available on the META-SHARE platform.

3. TECHNICAL INFORMATION

3.1 Directories and files

The archive that can be uploaded on the META-SHARE is a .zip file with two files: one .xml and one .xsd, which contains the .xml specification file.

3.2 Data structure of an entry

In the text file, the data is organized by types of abbreviations and each one of them is subdivided into entries with tags: grammatical categories, and grammatical features (gender and number), as exemplified below, when “WD” stands for “Week Days”, and “fs” for “female singular”; and the correspondent list of abbreviations:

```
<entry>
        <tag>_WD#fs</tag>
        <list>
                <abbrev>seg.</abbrev>
                <abbrev>qua.</abbrev>
                <abbrev>qui.</abbrev>
                <abbrev>sex.</abbrev>
                <abbrev>ter.</abbrev>
        </list>
</entry>
<entry>
        <tag>_WD#ms</tag>
        <list>
                <abbrev>sáb.</abbrev>
                <abbrev>dom.</abbrev>
```

</list>
</entry>

3.3 Lexicon size (nmb. of lexical items, KB occupied on disk)

The corpus is composed by 208 words with 3.9 KB compressed (27 KB uncompressed) for disk storage.

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

The language of the LX-Abbreviations is European Portuguese.

4.2 Entry Type

For this information, please see item 3.2.

4.3 Attributes and their values

There are three values for gender − <m> for male, <f> for female, and <g> for male or female − and other three for number − <s> for singular, <p> for plural, and <n> for singular or plural.

Taking as an example the entry exposed at Section 3.2 <_WD#fs>, the first value <WD> is the grammatical category tag (WD: Week Days) followed <#> by the tags for gender <m> and number <f>.

4.4 Coverage of the lexicon

The LX-Abbreviations lexicon covers the general language.

4.5 Intended application of the lexicon

LX-Abbreviations has been used as part of LX-Tokenizer in all NLP applications developed at NLX-Group, as a base list with string types considered hard cases for tokenization of Portuguese texts, involving the ambivalence between the end of a sentence and the end of an abbreviation (see Branco and Silva, 2003).

4.6 POS assignment

Each type of abbreviation was mannually annotated with proper grammatical category tag, according to the POS-Tagger used at NLX-Group (see Silva, 2007).

4.7 Reliability (automatically/manually constructed)

Mannually constructed (open list), under the standard abbreviations considered in grammars and spelling handbooks for Portuguese.

5. RELEVANT REFERENCES AND OTHER INFORMATION

Barreto, Florbela, António Branco, Eduardo Ferreira, Amália Mendes, Maria Fernanda Nascimento, Filipe Nunes e João Silva, 2006, "Open Resources and Tools for the Shallow Processing of Portuguese: The TagShare Project", Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006).

Silva, João, 2007. Shallow Processing of Portuguese: From Sentence Chunking to Nominal Lemmatization. MSc thesis, University of Lisbon. Published as Technical Report DI-FCUL-TR-07-16.

C-ORAL-ROM Portuguese Corpus

1 BASIC INFORMATION

1.1 Corpus composition

The C-ORAL-ROM corpus, created for the project C-ORAL-ROM - Integrated Reference Corpora for Spoken Romance Languages, in 2004, is a multilingual corpus of spoken language of the four main romance languages (Spanish, Portuguese, French and Italian), with approximately 300.000 words each language, covering both formal (152.755 words) and informal (165.838 words) speech. It is constituted by 153 recordings, corresponding to a total of 30 hours of recording. The recordings cover a period that goes from 1970 to 2002, but great part of them fall within the nineties.

INFORMAL			
Familiar/Private	Conversations	24.449	
	Dialogues	62.738	
	Monologues	46.005	**133.192**
Public	Conversations	1.817	
	Dialogues	23.119	
	Monologues	7.710	**32.646**
TOTAL (words)			**165.838**

FORMAL			
Natural Context	Business	10.215	
	Conferences	9.750	
	Law	6.315	
	Political Debate	8.923	
	Political Speech	8.649	
	Professional Explanation	6.473	
	Preaching	6.127	
	Teaching	9.822	**66.274**
Media	Interviews	14.570	
	Meteorology	1.930	
	News	1.859	
	Reportages	10.762	
	Scientific Press	9.923	
	Sport	5.676	

	Talk Shows	17.396	**62.116**
Telephone	Private	24.365	**24.365**
TOTAL (words)			**152.755**

The original corpus (including text files, sound files, text-to-sound aligned files and POS tagged files) is available at ELRA. Here, we present a new version that arises from the need to uniformize this spoken corpus (regarding orthographic transcription, text-to-sound alignment software – aligned with EXMARaLDA software – and POS annotation) with two other resources: Fundamental Portuguese and Spoken Portuguese.

1.2 Representation of the corpora (flat files, database, markup)

The corpus consists of audio files in .wav format, aligned transcriptions in XML EXMARaLDA format and transcriptions in plain text. There is also a version of plain text with POS-tag information, automatically assigned.

1.3 Character encoding

The characters have been encoded in UTF-8.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)

Name: Dr. Amália Mendes
Address: Complexo Interdisciplinar da Universidade de Lisboa
Av. Prof. Gama Pinto, 2
1649-003 Lisboa - Portugal
Affiliation: Centro de Linguística da Universidade de Lisboa
Position: Researcher
Telephone: + 351 21 790 47 00
Fax: + 351 21 796 56 22
e-mail: amalia.mendes@clul.ul.pt

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be available on the MetaShare platform as a set of transcription files in .html (with metadata regarding the recording and the speakers) in XML following the EXMARaLDA basic transcription format (.exb), as sound files (.wav), as plain text files (.txt) and as plain text files tagged with POS information (.pos).

2.3 Copyright statement and information on IPR

The resource will be distributed under an ELRA license.

3 TECHNICAL INFORMATION

3.1 Directories and files

The C-ORAL-ROM_EXM corpus has 153 .wav files, 153 .exb files, 153 .txt files, and 153 .pos files, corresponding to the audio files, the transcriptions with text-to-sound synchronization (EXMARaLDA format), the plain text files and the files automatically tagged with POS information.

3.2 Data structure of an entry

The XML files follow the Exmaralda data structure. Please, consult the Document Type Definitions (DTD) and XML-Schemata of Exmaralda for details.
Available at http://www.exmaralda.org/en_downloads.html.

3.3 Corpora size (nmb. of tokens, MB occupied on disk)

The corpus contains 318,593 tokens, 47.887 utterances and it needs about 4.40 GB for disk storage for the .wav, the .exb, the .txt and the .pos files.

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

This corpus is a monolingual, annotated corpus.

4.2 The natural language(s) of the corpus

This corpus consists of face-to-face informal conversations between acquaintances, friends or relatives, telephonic conversations and formal acts that include media (such as radio and televison programs) and other contexts (such as conferences, teaching, preaching, etc.).

4.3 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

The annotation of the corpus was made at the utterance level and includes discourse mark-up (like pauses, hesitations, reformulations, extralinguistic elements, speaker overlapping, etc.) and prosodic information markers. The annotation also includes a XML mark-up that specifies the alignment between audio and utterance transcription. The XML markup follows the Document Type Definitions (DTD) of the Exmaralda basic transcription format (basic-transcription.dtd) that is included in the files. For more details on the XML mark-up and XML-Schemata of EXMARaLDA we refer to the website (http://www.exmaralda.org/en_downloads.html).

4.4.2 Tags (if POS/MSD/TIME/discourse/etc –tagged or parsed),

The corpus was automatically POS-tagged with an automatic tagger trained on a slightly adapted version of the written part of CINTIL corpus (Barreto et al., 2006) Multi-word units do not receive special POS tags as is the

case in CINTIL, and contracted forms (e.g., “pelo”, “do”) are kept and receive a double tag (e.g., pelo/PREP+DET), while in CINTIL these words are split into two separate tokens.

4.4.3 Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)

The transcriptions were manually aligned with the audio files using the EXMARaLDA software. Every transcription is aligned at utterance level.

4.4.4 Attributes and their values (if annotated)

The following conventions were applied in the transcription of the audio files.

Transcription Conventions	
Symbol	Context of use
//	Indicates the end of an utterance.
?	Indicates an interrogative utterance.
/	.
…	Indicates an utterance which was left incomplete by the speaker because its conclusion is obvious.
+	Indicates an incomplete utterance.
xxx	Indicates an incomprehensible word.
yyyy	Indicates an incomprehensible sequence.
&	Indicates a word fragment or a filled pause.
[/]	Indicates the repetition of a word. In addition, if more than a word is repeated, angle brackets (< >) should be used to delimit the repeated sequence.
[//]	Indicates a reformulation of the discourse. In addition, if more than a word is involved in the reformulation, angle brackets (< >) should be used to delimit the sequence.
[///]	Indicates a total reformulation of the previous discourse.
hhh	Indicates extralinguistic elements, such as laugh, cough, etc.
" "	Indicates direct discourse.
&eh, &ah, &hum	Symbols used to represent filled pauses.

In comparison to the previous version, there are four main differences regarding orthographic transcription:

1. introduction of the symbols [/], [//] and [///] for word repetition and discourse reformulation.
2. overlapping is directly aligned in EXMARaLDA and no longer marked in orthographic transcription;
3. quotation marks (which were used in titles of books, songs, television/radio programs, etc.) indicate now direct discourse;
4. titles of books, songs, television/radio programs are transcribed in capital letter (first letter).

The following tags were applied in the POS-tagging. The column on the right shows the tags used in the first version of the corpus.

POS codification		
Tag	**Category**	**Tag from previous version**
ADJ	Adjectives	ADJ
ADV	Adverbs	ADV
CARD	Cardinals	NUMc
CJ	Conjunctions	CONJc (coordenative) CONJs (subordinative)
CL	Clitics	CL
CN	Common Nouns	Nc
DA	Definite Articles	ARTd
DEM	Demonstratives	DEMv (variable) DEMi (invariable)
DFR	Denominators of Fractions	
DGTR	Roman Numerals	spelled out and marked as NUMc
DGT	Digits	spelled out and marked as NUMc
DM	Discourse Marker	MD
EADR	Electronic Addresses	
EOE	End of Enumeration	
EXC	Exclamatives	RELv (variable) RELi (invaraible)
GER	Gerunds	VG
GERAUX	Gerunds as auxiliary verbs	VAUX
IA	Indefinite Articles	ARTi
IND	Indefinites	INDv (variable) INDi (invariable)
INF	Infinitive	VB
INFAUX	Infinitive auxiliary verb	VAUXB
INT	Interrogatives	RELv (variable) RELi (invaraible)
ITJ	Interjection	INT
LTR	Letters	spelled out and marked as Nc
MGT	Magnitude Classes	Nc
MTH	Months	Nc
NP	Noun Phrases	
ORD	Ordinals	NUMo
PADR	Part of Address	Np
PNM	Part of Name	Np
PNT	Punctuation Marks	
POSS	Possessives	POS
PPA	Past Participles not in compound tenses	PPA

PP	Prepositional Phrases	
PPT	Past Participle in compound tenses	VPP
PREP	Prepositions	PREP
PRS	Personals	PES
QNT	Quantifiers	RELv (variable) RELi (invaraible)
REL	Relatives	RELv (variable) RELi (invaraible)
STT	Social Titles	Nc
SYB	Symbols	
TERMN	Optional Terminations	
UM	"um" or "uma"	ARTi:NUMc
UNIT	Measurement units in abbreviated form	spelled out and marked as Nc
VAUX	Finite "ter" or "haver" in compound tenses	VAUX
V	Verbs (other than PPA, PPT, INF or GER)	Vpi (present indicative) Vppi (past indicative) Vii (imperfect indicative) Vmpi (pluperfect indicative) Vfi (future indicative) Vc (conditional indicative) Vpc (present subjunctive) Vic (imperfect subjunctive) VBf (inflected subjunctive Vimp (imperative)
WD	Week Days	
LADV1… LADVn	Multi-Word Adverbs	LADV… LADV
	Multi-Word Conjunctions	LCONJ… LCONJ
	Multi-Word Prepositions	LPREP… LPREP
	Multi-Word Pronominals	LPRON… LPRON
	Multi-Word Discourse Markers	LD… LD
	Proper Nouns	Np
	Foreign Word	ESTR
	Acronymous	SIGL
Contracted forms	**Combinations of :**	
CL+CL	Two clitics	
PREP+ADV	Preposition and Adverb	PREP+ADV
PREP+DA	Preposition and Definite Articles	PREP+ARTd
PREP+DEM	Preposition and Demonstratives	PREP+DEMv PREP+DEMi
PREP+IND	Preposition and Indefinite	PREP+INDv PREP+INDi
PREP+INT	Preposition and Interrogative	PREP+RELv PREP+RELi
PREP+PRS	Preposition and Personal pronoun	PREP+PES
PREP+QNT	Preposition and Quantifier	PREP+RELv

		PREP+RELi
PREP+REL	Preposition and Relative	PREP+RELv PREP+RELi
PREP+UM	Preposition and "um" or "uma"	PREP+ARTi:NUMc
Tags specific to the spoken corpus		
EMP	Emphasis	ENF
EL	Extra-linguistic	EL
PL	Para-linguistic	PL
FRAG	Fragmente Words or Filled Pauses	FRAG
Pimp	Word Impossible to Transcribe	Pimp
Simp	Sequence Impossible to Transcribe	Simp
	Without Classification	SC

Regarding the metadata, each .exb file has metainformation concerning the following topics.

Metainformation field	
Fixed Attributes	
Project Name	
Transcription Name	
Transcription convention	
Referenced media file(s) (automatically imports sound file)	
Comments (for example: cut in the recording from 04'50'' to 08'27'')	
User defined attributes	
Country	
Date (DD/MM/YYYY)	

Place of the recording	
Length of the recording (m's'')	
Length of the transcribed excerpt	
Location of the transcribed excerpt	
Words	
Acoustic quality	Good Medium Bad
Source	
Code in CRPC	
Recording conditions	
Topic	
Communication interactivity	Unknown Unspecified Interactive (corresponds to dialogues and conversations and may not include the investigator) Non-interactive (corresponds often to monologues) Semi-interactive (corresponds mainly to monologic speech punctuated by repeated interjections from the hearer)
Communication planning	Unknown Unspecified Spontaneous (topic not determined from context or observers: conversation, chatting, joke-telling) Semi-spontaneous (topic directed in some way by an investigator or community member, but actors speak/sing freely within this context) Planned (the speaker prepares in detail the structure and content of his/her performance in advance)
Communication involvement	Unknown Unspecified Elicit (Investigator asks speaker(s) to produce isolated phonemes/words/utterances/grammatical structures) Non-elicit (the researcher does not interfere verbally with the speech event) No-observer (a tape recorder runs continuously in room while people talk (having been for example set there a half hour earlier by

	the investigator, with permission of course)
Communication social context	Unknown Unspecified Family (restrictive to relatives) Private (friends, colleagues, etc.) Public (to the communication event is allowed to whoever, in a free or in a regulated manner) Controlled environment (the communication event undergoes the agreement to elicit a linguistic behaviour)
Communication event structure	Unknown Unspecified Monologue Dialogue Conversation Not natural format
Communication channel	Unknown Undefined Face to Face (spontaneous speech) Experimental setting (takes place within a controlled environment for the purpose of testing hypotheses) Broadcasting (Interview; Meteorology; News; Reportage; Scientific Press; Sport; Talk-show) Formal (Business; Conferences; Law; Political debate; Political speech; preaching; Professional explanation; Teaching) Telephone
Transcriber	
Revisor	
Original physical format	
Physical storage Id. (cassette and CD)	

Speakertable (metadata regarding the participants)	
Fixed Attributes	
Abbreviation	
Sex	
Language(s) used	

First language	
Second language	
Comment (for example: MAR produces “germinada” instead of the correct form “geminada)	
User defined attributes	
Name	
Age	
Geographical origin	
Residence	
Education	Illiterate Primary school Middle school High school University students Graduated
Profession	
Linguistic influence	
Role	Interviewer Informant Participant Announcer Answer

4.5 Intended application of the corpus

The corpus can be used in linguistic research and for improving and developing numerous kinds of Natural Language Processing tools and applications, as well as in developing speech technologies.

4.6 Reliability of the annotations (automatically/manually assigned) – if any

The transcriptions and alignments with the audio files were manually done. The POS-tagging, on the other hand, was done automatically.

5. RELEVANT REFERENCES AND OTHER INFORMATION

Bacelar do Nascimento, M. F., J. Bettencourt Gonçalves, R. Veloso, S. Antunes, N. Martins, F. Barreto, R. Amaro, and M. L. Garcia Marques (2006), C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages, MOSTRA DE LINGUÍSTICA - A Linguística em Portugal: estado da arte, projectos e produtos, CD publication.

Bacelar do Nascimento, M. F., J. Bettencourt Gonçalves, R. Veloso, S. Antunes, F. Barreto and R. Aamaro (2005), "The Portuguese Corpus", in CRESTI, E. and M. MONEGLIA (eds.), *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*, John Benjamins Publishing Company, Studies in Corpus Linguistics nº 15, Amsterdam/Philadelphia, pp. 163-207 (with DVD).

Bacelar do Nascimento, M. F., A. Mendes and R. Amaro (2003) "Reusing Available Resources for Tagging a Spoken Portuguese Corpus", in *TASHA'2003: Workshop on Tagging and Shallow Processing of Portuguese*, University of Lisbon, October 2003.

Bacelar do Nascimento, M. F. (2002), "Quelques considérations sur la constitution et l'exploitation d'un *corpus* de portugais parlé" in SCARANO, A. (a cura di) *Macrosyntaxe et pragmatique: l' analyse de la langue orale*, Bulzoni, Roma, pré-impressão LABLITA, November 2002, pp. 221-228.

Bacelar do Nascimento, M. F., E. Cresti, M. Moneglia, A. Moreno Sandoval, J. Veronis, P. Martin, K. Choucri, V. Mapelli, D. Falavigna, A. Cid and C. Blum (2002), "The C-ORAL-ROM Project. New methods for spoken language archives in a multilingual romance corpus LREC", in RODRIGUES, M. C. and C. SUAREZ ARAUJO (a cura di), Proceedings of the *Third International Conference on Language Resources and Evaluation*, Paris: ELRA, vol. 1, pp. 2-10.

Bacelar do Nascimento, M. F., L. A. S. Pereira e J. Saramago (2000), "Portuguese Corpora at CLUL". In *Second International Conference on Language Resources and Evaluation – Proceedings, Volume II*, Athens: 1603-1607.

F. Barreto, A. Branco, E. Ferreira, A. Mendes, M. F. Nascimento, F. Nunes, and J. Silva (2006), "Open Resources and Tools for the Shallow Processing of Portuguese". In Proceedings of the *Fifth International Conference on Language Resources and Evaluation* (LREC2006), Genoa, Italy.

CINTIL-Internacional Corpus of Portuguese

1 BASIC INFORMATION

1.1 Corpus composition

CINTIL-Corpus Internacional do Português is a linguistically interpreted corpus of Portuguese. At present it is composed of 1 Million annotated tokens, verified by human expert annotators. The annotation comprises information on part-of-speech, open classes lemma and inflection, multi-word expressions pertaining to the class of adverbs and to the closed POS classes, and multi-word proper names (for named entity recognition). The corpus has been developed at the University of Lisbon by the NLX group at the Faculty of Sciences and the Anagrama group at the Cenro de Linguística da Universidade de Lisboa.

1.2 Representation of the corpora (flat files, database, markup)

The corpus consists of 2 files with linguistic annotation (pos, inflection, lemma, named entities). The annotation has been manually revised.

1.3 Character encoding

The characters have been encoded in UTF-8.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)

Name: Prof. António Branco
Address: Departamento de Informática
Faculdade de Ciências da Universidade de Lisboa
Alameda da Universidade
1649-003 Lisboa - Portugal
Affiliation: Faculdade de Ciências da Universidade de Lisboa
Position: Professor
Telephone: +351 21 7500606
Fax: + 351 21
e-mail: Antonio.Branco@di.fc.ul.pt

Name: Dr. Amália Mendes
Address: Complexo Interdisciplinar da Universidade de Lisboa
Av. Prof. Gama Pinto, 2
1649-003 Lisboa - Portugal
Affiliation: Centro de Linguística da Universidade de Lisboa
Position: Researcher
Telephone: +351 21 790 47 00
Fax: + 351 21 796 56 22
e-mail: amalia.mendes@clul.ul.pt

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be available on the MetaShare platform as a set of 2 files annotated with linguistic information. The annotation manual is made available with the corpus.

2.3 Copyright statement and information on IPR

ELRA licence.

3 TECHNICAL INFORMATION

3.1 Directories and files

CINTIL is divided into 2 text files, one for written texts and another for transcriptions of spoken register.

3.2 Data structure of an entry

Each file is structured in one header per text, with sentences, paragraphs and excerpts mark-up, in XML format. The text and its annotation is presented in a 4 column format: token, lemma (for lexical categories), pos tag + inflection tag, named entities tag.

3.3 Corpora size (nmb. of tokens, MB occupied on disk)

The corpus contains 1,191,746 million tokens (689,124 written and 502,622 spoken) and needs about 20MB for disk storage.

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

This corpus is a monolingual, annotated corpus, covering both written and spoken registers.

4.2 The natural language(s) of the corpus

The language of the corpus is European Portuguese.

4. 3 Domain(s)/register(s) of the corpus

The corpus contains a written and a spoken subpart. The written subpart includes fiction, newspapers and technical texts, from 1990 till 2003, plus 3 fiction texts from the 19th century. The spoken subpart are transcriptions of recordings of both formal and informal registers, from 1970 to 2002. These spoken transcriptions cover very diversified situations: conversations, dialogues, phone conversations, radio and television programs, homilies, among others. Here is a brief overview of the different registers and the number of tokens.

- Written = 689,124 tokens:
  - News: 58.7% - 404,690 tokens
  - Fiction: 29% - 200,194 tokens
  - Other: 12.2% - 84,240 tokens
- Spoken = 502,622 tokens:
  - Informal/Private: 43.2% - 217,604 tokens
  - Informal/Public: 9.5% - 48,221 tokens
  - Informal/Phone: 0.4% - 2,287 tokens
  - Formal/Natural: 19.3% - 97,499 tokens
  - Formal/Media: 17.6% - 88,727 tokens
  - Formal/Phone: 9.6% - 48,284 tokens
- Total = 1,191,746 tokens

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

Each written text is described in a header in xml format with information on the type of data, the author's name, title, place and date of edition. The header of each transcription refers to the type of data, title, description of participants, date and place of recording, situation, register, length, number of tokens, accousting quality, transcriber and revisor's name. The corpus contains markup in xml at the following levels: paragraph, sentence (equivalent to speech turn in the spoken subpart) and excerpt.

4.4.2 Tags (if POS/MSD/TIME/discourse/etc –tagged or parsed),

The corpus was automatically tagged with LX-Tagger (Barreto et al.,2006). All levels of linguistic annotation were verified by two human annotators.

4.4.3Alignment information (if the corpus contain aligned documents: level of alignment, how it was achieved)

Does not apply.

4.4.4 Attributes and their values (if annotated)

Part-of-speech tags:

Tag	Category	Examples
ADJ	Adjectives	bom, brilhante, eficaz, …
ADV	Adverbs	hoje, já, sim, felizmente, …
CARD	Cardinals	zero, dez, cem, mil, …
CJ	Conjunctions	e, ou, tal como, …
CL	Clitics	o, lhe, se, …
CN	Common Nouns	computador, cidade, ideia, …
DA	Definite Articles	o, os, …
DEM	Demonstratives	este, esses, aquele, …
DFR	Denominators of Fractions	meio, terço, décimo, %, …
DGTR	Roman Numerals	VI, LX, MMIII, MCMXCIX, …
DGT	Arabic Numerals	0, 1, 42, 12345, 67890, …
DM	Discourse Marker	olá, …
EADR	Electronic Addresses	http://www.di.fc.ul.pt, …
EOE	End of Enumeration	etc
EXC	Exclamation	ah, ei, …
GER	Gerunds	sendo, afirmando, vivendo, …
GERAUX	Gerund "ter"/"haver" in compound tenses	tendo, havendo
IA	Indefinite Articles	uns, umas, …
IND	Indefinites	tudo, alguém, ninguém, …
INF	Infinitive	ser, afirmar, viver, …
INFAUX	Infinitive "ter"/"haver" in compound tenses	ter, haver, …
INT	Interrogatives	quem, como, quando, …
ITJ	Interjection	bolas, caramba, …
LTR	Letters	a, b, c, …
MGT	Magnitude Classes	unidade, dezena, dúzia, resma, …
MTH	Months	Janeiro, Dezembro, …

NP	Noun Phrases	idem, …
ORD	Ordinals	primeiro, centésimo, penúltimo, …
PADR	Part of Address	Rua, av., rot., …
PNM	Part of Name	Lisboa, António, João, …
PNT	Punctuation Marks	., ?, (, …
POSS	Possessives	meu, teu, seu, …
PPA	Past Participles not in compound tenses	sido, afirmados, vivida, …
PP	Prepositional Phrases	algures, …
PPT	Past Participle in compound tenses	sido, afirmado, vivido, …
PREP	Prepositions	de, para, em redor de, …
PRS	Personals	eu, tu, ele, …
QNT	Quantifiers	todos, muitos, nenhum, …
REL	Relatives	que, cujo, tal que, …
STT	Social Titles	Presidente, drª., prof., …
SYB	Symbols	@, #, &, …
TERMN	Optional Terminations	(s), (as), …
UM	"um" or "uma"	um, uma
UNIT	Abbreviated Measurement Unit	kg., km., …
VAUX	Finite "ter" or "haver" in compound tenses	temos, haveriam, …
V	Verbs (other than PPA, PPT, INF or GER)	falou, falaria, …
WD	Week Days	segunda, terça-feira, sábado, …
Tags for multi-word expressions		
LADV1…LADVn	Multi-Word Adverbs	de facto, em suma, um pouco, …
LCJ1…LCJn	Multi-Word Conjunctions	assim como, já que, …
LDEM1…LDEMn	Multi-Word Demonstratives	o mesmo, …
LDFR1…LDFRn	Multi-Word Denominators of Fractions	por cento
LDM1…LDMn	Multi-Word Discourse Markers	pois não, até logo, …
LITJ1…LITJn	Multi-Word Interjections	meu Deus
LPRS1…LPRSn	Multi-Word Personals	a gente, si mesmo, V. Exa., …
LPREP1…LPREPn	Multi-Word Prepositions	através de, a partir de, …
LQD1…LQDn	Multi-Word Quantifiers	uns quantos, …
LREL1…LRELn	Multi-Word Relatives	tal como, …
Tags specific to the spoken corpus		
EMP	Emphasis	
EL	Extra-linguistic	
PL	Para-linguistic	
FRG	Fragment	

Inflection tags:

Tag	Description
Tags for nominal categories	
m	Masculine
f	Feminine
s	Singular
p	Plural
dim	Diminutive
sup	Superlative
comp	Comparative
Tags for verbs	
1	First Person
2	Second Person
3	Third Person
pi	Presente do Indicativo
ppi	Pretérito Perfeito do Indicativo
ii	Pretérito Imperfeito do Indicativo
mpi	Pretérito Mais que Perfeito do Indicativo
fi	Futuro do Indicativo
c	Condicional
pc	Presente do Conjuntivo
ic	Pretérito Imperfeito do Conjuntivo
fc	Futuro do Conjuntivo
imp	Imperativo
Tags for infinitive verbs	
ifl	Inflected
nifl	Not Inflected

Named-entity tags

Position	*description*	Semantic type	*description*	example
B-	beginning	PER	person	...o[O] *João[B-PER] Silva[I-PER]* disse[O]...
I-	inside	ORG	organization	...a[O] *Universidade[B-ORG] de[I-ORG] Lisboa[I-ORG]* comprou[O]...
		LOC	location	...de[O] *Londres[B-LOC]* a[O] *Paris[B-LOC]*...
		WRK	work	...a[O] *Mona[B-WRK] Lisa[I-WRK]* está[O]...
		MSC	other cases	...o[O] *RMS[B-MSC] Titanic[I-MSC]* afundou[O]...
O	outside			

4.5 Intended application of the corpus

The corpus can be used in linguistic research and for improving and developing numerous kinds of Natural Language Processing tools and applications.

4.6 Reliability of the annotations (automatically/manually assigned) – if any

The POS-tagging was done automatically in a first phase and then manually revised by two annotators.

5. RELEVANT REFERENCES AND OTHER INFORMATION

Barreto, F., A. Branco, E. Ferreira, A. Mendes, M. F. Nascimento, F. Nunes, and J. Silva (2006), "Open Resources and Tools for the Shallow Processing of Portuguese". In 5th International Conference on Language Resources and Evaluation (LREC2006), Genoa, Italy.

Barreto, Florbela, António Branco, Eduardo Ferreira, Amália Mendes, Fernanda Bacelar Nascimento, Filipe Nunes and João Silva, 2006, "Linguistic Resources and Software for Shallow Processing", In *Actas do XXI Encontro Anual da Associação Portuguesa de Linguística*, Lisbon, Portugal.

Branco, António and João Silva, 2006, "LX-Suite: Shallow Processing Tools for Portuguese", *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics* (EACL2006), Trento, Italy, pp.179-182.

Lexicon of multiword expressions

1. BASIC INFORMATION

1.1 Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc),

This lexicon includes multiword expressions (MWE) of European Portuguese extracted from a balanced 50,8M word written corpus – a subcorpus of the Reference Corpus of Contemporary Portuguese (CRPC). This corpus covers different genres, being mainly constituted by journalistic texts (59%), but it also includes texts from literature (21%), magazines (15%), miscellaneous, supreme court verdicts, parliament sessions and leaflets (5%). The MWE lexicon covers 1.198 lemmas (composed of single words from different POS categories: nouns, adjectives, verbs and adverbs) and a total of 12.753 MWE lemmas (which include inflectional variants of the MWE lemmas) and 242.233 concordances of those MWE expressions manually verified.

1.2 Representation of the lexicon (flat files, database, markup)

The lexicon is represented in .txt and .html format.

1.3 Character encoding

The characters have been encoded in UTF-8.

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)

Name: Dr. Amália Mendes
Address: Complexo Interdisciplinar da Universidade de Lisboa
Av. Prof. Gama Pinto, 2
1649-003 Lisboa - Portugal
Affiliation: Centro de Linguística da Universidade de Lisboa
Position: Researcher
Telephone: + 351 21 790 47 00
Fax: + 351 21 796 56 22
e-mail: amalia.mendes@clul.ul.pt

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be available on the MetaShare platform.

2.3 Copyright statement and information on IPR

The resource is free licensed-based for research purposes and free license-based for commercial purposes. It is planned to be distributed under a MetaShare Commons BY SA licence.

3. TECHNICAL INFORMATION

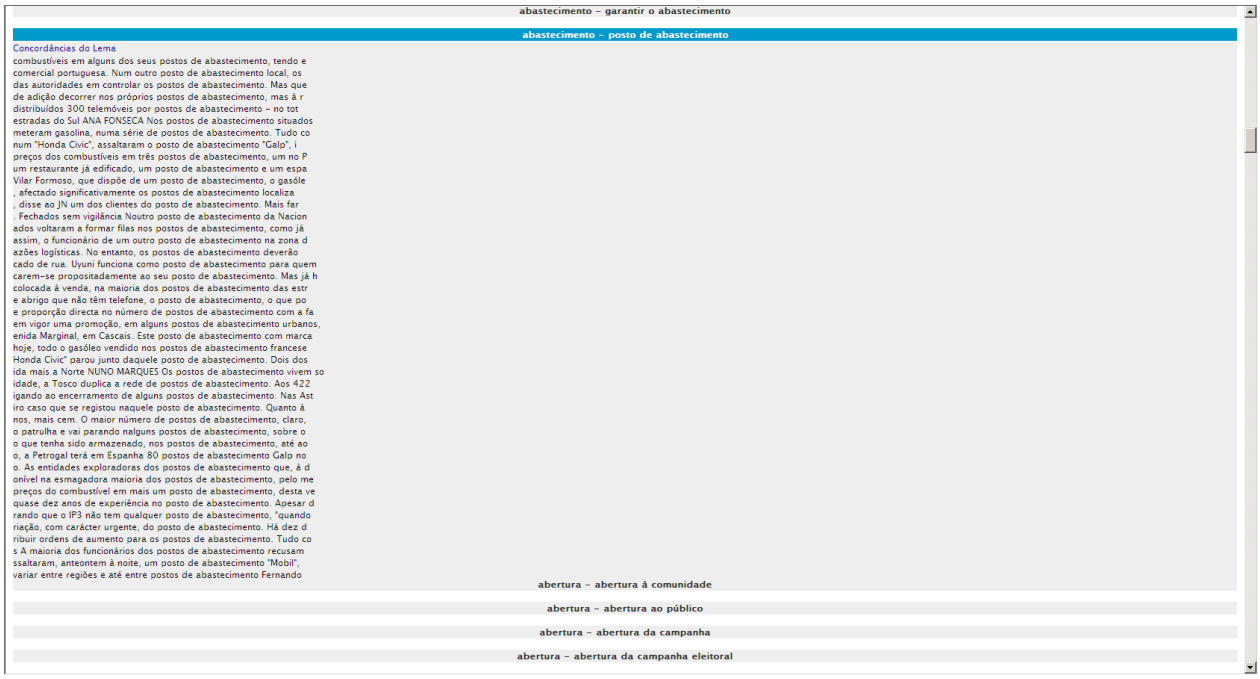
3.1 Directories and files

The lexicon is available in two different formats: .txt and .html (for consultation). The .html version is divided in 17 files, organized in alphabetical order of the main lemma, and the .txt version is a single file.

3.2 Data structure of an entry

Each MWE expression is lemmatized according to the lemma that was under analysis. As an example, the MWE lemma POSTO DE ABASTECIMENTO ('gas station') is associated to the main lemma ABASTECIMENTO (‘supply’), since the selection of MWE was undertaken during the analysis of that lemma. However, it is not associated to the lemma POSTO (‘station’), because that lemma was not analyzed.

HTML version: The files have entries for lemma and each associated MWE lemma. Clicking on the lemma bar will show concordance lines for each context in which the expression – and its selected variants (gender, number, verbal inflection, lexical and/or syntactic variation) – occurred.



TXT version: Each entry is structured with the following information: main lemma (LEM), MWE lemma (GROUP), total frequency of the group in the corpus (FREQ) and list of concordances.

LEM: flor
GROUP: jarra de flores
FREQ: 11
CONCORDANCES:
da democracia não podem ser uma jarra de flores. Neste sentido,
modelo empenhado em pintar, uma jarra de flores, uma paisagem, o
Zé P'reira. Estava adornado com jarras de flores e velas de cera
abros, talha doirada, imagens, e jarras de flores fanadas e fingi
dos ao fundo à esquerda, junto à «Jarra de Flores» de Gauguin, no
eram a entrar para dentro de uma jarra de flores. "Quanto ao luxo
modava os hóspedes. Eu tinha dez jarras de flores sobre uma estan
o de novo no escritório, ponho a jarra de flores na secretária. A
pleto de seixos de bibelots e de jarras de flores que a vazante i
s que se desprendem da mesa, uma jarra de flores, a brancura da c
velmente, por reflexo, a colocar jarras de flores de plástico no

3.3 Lexicon size (nmb. of lexical items, KB occupied on disk)

The lexicon has 1198 main lemmas and 12.753 MWE lemmas. The total number of concordances extracted from the corpus is 242.233. The txt version requires 16 MB and the html version 34 MB for disk storage.

1. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

The language of the lexicon is European Portuguese.

4. 2 Entry Type

HTML version: There is one entry for main lemma, followed by the MWE lemmas associated to it, which include all the variant forms. Clicking on the lemma bar opens the list of concordances of this MWE in the corpus (covering inflection variants and lexical and/or syntactic variation).

TXT version: There is one entry for the main lemma, followed by the MWE lemma, its frequency and concordances.

4.3 Attributes and their values

TXT version: main lemma (LEM), MWE lemma (GROUP), frequency (FREQ) and concordances (CONCORDANCES).

4.4 Coverage of the lexicon

The dimension of the lexicon, along with the variety of types of texts from which it was extracted (journalistic, literary, techno-scientific, etc.), guarantee a wide coverage of the contemporary Portuguese vocabulary.

4.5 Intended application of the lexicon

This lexicon could be used in several areas, ranging from psycholinguistics (development of hypothesis about the representation of the individual mental lexicon, semantic memory and cognitive processes), lexicography (improvement of their coverage in modern dictionaries) or computational linguistics (helping to develop and evaluate language processing tools able of dealing with MW expressions specific issues, like automatic unit recognition, lexical association measures for validation of significant MW units, as well as overgeneration, tagging and parsing problems.

4.6 POS assignment

n.a.

4.7 Reliability (automatically/manually constructed)

All MWE were automatically extracted and sorted using Mutual Information (MI) statistical measure (that calculates the frequency of each group in the corpus and crosses this frequency with the isolated frequency of each word of the group, also in the corpus). Several cut-off options were used in order to reduce noise: (i) groups with internal punctuation were eliminated; (ii) two-word groups with initial or ending grammatical words were also eliminated using a stop-list, since the main goal was to study lexical associations instead of functional compounds or verb valency; (iii) a minimum frequency was selected: 3 occurrences to groups of 3 to 5 tokens and of 10 for 2-token groups. A subset of the results (isolated based on the best MI values) was manually validated taking into consideration the MI value and also the frequency of occurrence. The MWE selected by human experts are the ones included in this lexicon.

5. RELEVANT REFERENCES AND OTHER INFORMATION

Bacelar do Nascimento, M. F., A. Mendes, S. Antunes (2006) "Typologies of MultiWord Expressions Revisited: A Corpus-driven Approach" in Kawaguchi, Y., S. Zaima, T. Takagaki (eds.) *Spoken Language Corpus and Linguistic Informatics*, Jonh Benjamins, Coll. Usage-Based Linguistic Informatics, vol.V, pp. 227-244.

Antunes, S., M. F. Bacelar do Nascimento, J. M. Casteleiro, A. Mendes, L. Pereira, T. Sá (2006) "A Lexical Database of Portuguese Multiword Expressions" in Vieira, R. *et al*. (2006) PROPOR 2006, LNAI 3960, Berlin, Springer-Verlag, pp. 238-243.

Mendes, A., S. Antunes, M. F. Bacelar do Nascimento, J. M. Casteleiro, L. Pereira, T. Sá (2006) "COMBINA-PT: a Large Corpus-extracted and Hand-checked Lexical Database of Portuguese Multiword Expressions", *Proceedings of the V International Conference on Language Resources and Evaluation*, LREC2006, Genoa, May 22-28 2006, pp. 1900-1905.

Antunes, S., M. F. Bacelar do Nascimento, J. M. Casteleiro, A. Mendes, L. Pereira, T. Sá (2006) "Corpus-based extraction and identification of Portuguese Multiword Expressions", Verbum ex machina - *Actes de la 13e conférence sur le traitement automatique des langues naturelles* (TALN 2006), Louvain, Presses Universitaires de Louvain, vol. 1, pp. 389-397.

PropBank

I. Basic Information

1.1. Corpus information

The CINTIL-PropBank (Branco *et al.*, 2012) is a set of sentences annotated with their constituency structure and semantic role tags, composed of 10,039 sentences and 110,166 tokens taken from different sources and domains: news (8,861 sentences; 101,430 tokens), and novels (399 sentences; 3,082 tokens) (cf. 3.2). In addition, there are 779 sentences (5,654 tokens) used for regression testing of the computational grammar that supported the annotation of the corpus (cf. Section 4.6).

For the creation of this PropBank we adopted a semi-automatic analysis with a double-blind annotation followed by adjudication. The resulting dataset contains three information levels: phrase constituency, grammatical functions, and phrase semantic roles.

The main motivation behind the creation of this resource was to build a high quality data set with semantic information that could support the development of automatic semantic role labelers for Portuguese.

The development of this resource started under the project SemanticShare – Resources and Tools for Semantic Processing (see more at: http://nlx.di.fc.ul.pt/projects.html), whose main goal was to generate a deep linguistic annotated corpus of Portuguese, with manually verified grammatical representations.

The following table displays a breakdown of the CINTIL-PropBank corpus:

CINTIL-TreeBank				
Sub-corpus	**id**	**Sentences**	**Tokens**	**Domain**
Sentences for regression testing	aTSTS	779	5,654	Test
CINTIL-International Corpus of Portuguese2	bCINT	1,219	13,516	News
	cCINT	399	3,082	Novels
CETEMPúblico	eCTMP	7,541	86,905	News
Penn TreeBank (translation)	dPENN	101	1,012	News
Total		10,039	110,166	

2 CINTIL-International Corpus of Portuguese was the first corpus, but only partly, to being used to integrate our corpus of syntactic trees of constituencies and, thus, give it the name.

1.2. Representation of the corpora (flat files, database, markup)

The corpus is a single file in a XML format.

1.*3. Character encoding*

The characters are in UTF8 code.

II. Administrative Information

2.1. Contact person

Name: António Branco

Address: Departamento de Informática NLX - Grupo de Fala e Linguagem Natural, Faculdade de Ciências da Universidade de Lisboa, Edifício C6, Campo Grande 1749-016 Lisboa

Position: Assistant professor

Affiliation: Faculty of Sciences, University of Lisbon

Telephone: +351 217 500 087

Fax: +351 217 500 084

E-mail: antonio.branco@di.fc.ul.pt

2.2. Delivery medium (if relevant; description of the content of each piece of medium)

This resource is available through META-SHARE.

2.3. Copyright statement and information on IPR

This resource is available for both research and commercial purposes, with attribution, and no redistribution nor derivatives allowed.

III. Technical Information

3.*1. Directories and files*

The archive that can be uploaded on the META-SHARE is a ZIP file with two files: one XML and one XSD, which contains the XML specification file.

3.2. Data structure of an entry

For the XML file with the set of sentences, the data is organized with one sentence per entry (<sentence>). Each entry contains the sentence identifier (<id>), formed by the sub-corpus id (cf. the Table in Section 1.1) concatenated with a sentence number; the sentence in raw text (<raw>); and the S-expressions or parenthesis format tree (<tree>), as shown in the example below:

<sentence>

<id>aTSTS-001/11</id>

<raw>A criança obedece apenas à mãe.</raw>

<tree>(S (S (NP-SJ-ARG1 (ART-SP A) (N criança)) (VP (V obedece) (PP-IO-ARG2 (ADV-M-ADV apenas) (PP (P a_) (NP-C (ART-SP a) (N mãe)))))) (PNT .))</tree>

</sentence>

3.3. *Corpus size (nmb. of tokens, NB occupied in disk)*

The corpus is composed by 10,143 sentences, which take up 678.1 KB compressed (3.5 MB uncompressed) for disk storage.

IV. Content Information

4.*1. Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)*

This is a monolingual and a semi-automatic annotated corpus.

4.*2. The natural language(s) of the corpus*

The language of the corpus is Portuguese with pre-spelling reform of 19903.

4.*3. Domain(s)/register(s) of the corpus*

Concerning the register of the texts represented in the corpus, it comprises excerpts of news articles from daily and general newspapers (8,861 sentences; 101,430 tokens), literary language from excerpts of novels (339 sentences; 3,082 tokens), and, additionally, test sentences used for regression testing of the grammar used in the annotation process (779 sentences; 5,654 tokens).

4.4. Annotation in the corpus (if an annotated corpus)

4.4.1. Types of annotation (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

The annotation of the corpus consists of three levels of linguistic information: phrase constituency, grammatical functions, and phrase semantic roles.

4.4.2. Tags (if POS/WSD/TIME/discourse/etc – tagged or parsed)

For details on the linguistic options underlying the analyses, see (Branco *et al.*, 2011). In this Section we present the tag sets used at the various levels of annotation:

Phrasal and part-of-speech tags

Tag	Meaning
A	Adjective

3This means that the orthography rules used are those that are described by the Orthography Reform of 1945. The orthographic agreement of 1990 was adopted just in may of 2009 and is being implemented until 2012.

AP	Adjective Phrase
ADV	Adverb
ADVP	Adverb Phrase
C	Complementizer
CP	Complementizer Phrase
CARD	Cardinal
CONJ	Conjuction
CONJP	Conjuction Phrase
D	Determiner
DEM	Demonstrative
N	Noun
NP	Noun Phrase
P	Preposition
PP	Preposition Phrase
POSS	Possessive
QNT	Predeterminer
S	Sentence
V	Verb
VP	Verb Phrase

Grammatical Function Tagset

Tag	Meaning
C	Complement

DO	Direct Object
IO	Indirect Object
M	Modifier
N	Relationship between words and named entities
OBL	Oblique Complement
PRD	Predicate
SJ	Subject
SP	Specifier

Semantic Role Tagset

Tag	Meaning
ARG1	First Argument
ARG2	Second Argument
ARG3	Third Argument
ARG11	Argument 1 of subordinating predicator and Argument 1 in the subordinate clause (semantic function of Subjects of so called Subject Control predicators)
ARG21	Argument 2 of subordinating predicator and Argument 1 in the subordinate clause (semantic function of Subjects of so called Direct Object Control predicators)
ARG1cp	Argument 1 in complex predicate constructions
ARG2cp	Argument 2 in complex predicate constructions
ARG2ac	Argument 2 of anticausative readings
ADV	Adverbial
CAU	Cause
DIR	Direction

EXT	Extension
LOC	Localization
MNR	Mode
PNC	Objective
POV	Viewpoint
TMP	Time
PRED	Secondary predication
NULL	Null

4.4.3. Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)

Not applicable.

4.4.4. Attributes and their values (if annotated)

Not applicable.

4.5. Intended application of the corpus

The corpus can be used in linguistic research and in the development of dependency parsers and semantic role labeling tools.

4.6. Reliability of the annotations (automatically/manually assigned) – if any

In order to achieve a gold-standard corpus with high accuracy, the CINTIL-PropBank is created by a two-phase process, where an automatic annotation is then manually revised by language experts with post-graduate degrees in Linguistics. More specifically, in the first stage, a deep computational grammar (see Branco and Costa, 2008) is used to generate all the possible parses for a given sentence (the parse forest). This is followed by a manual disambiguation stage where the correct parse is chosen from among those in the parse forest. This second stage follows a double-blind annotation method, where two annotators work independently and, for those cases where their decisions differ, a third annotator (the adjudicator) is brought in to make the final decision. For this corpus, the level of inter-annotator agreement (ITA) is 0.83 in terms of the specific inter-annotator metric developed for this kind of corpora and annotation (Castro, 2011).

The automatic annotation assigns only argumental semantic roles, leaving modifiers with a underspecified 'M'. These tags are manually specified, again following the same annotation method as before (double-blind annotation with adjudication). For this task, the ITA is 0.76.

V. Relevant References and Other Information

Branco, A. Carvalheiro, C., Pereira, S., Avelãs, M., Pinto, C., Silveira, S., Costa, F., Silva, J., Castro, S., and Graça, J., 2012, "A PropBank for Portuguese: the CINTIL-PropBank". In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Branco, A., Silva, J., Costa, F., and Castro, S., 2011, "CINTIL-TreeBank Handbook: Design options for the representation of syntactic constituency". Technical Report 2011;02, University of Lisbon, Department of Informatics.

Branco, A. and Costa, F., 2008, "A computational grammar for deep linguistic processing of portuguese: LXGram, version A.4.1". Technical Report DI-FCUL-TR-08-17. University of Lisbon, Department of Informatics, 2008.

Castro, Sérgio, 2011, *Developing Reliability Metrics and Validation Tools for datasets with deep linguistic Information*, MA Dissertation, Universidade de Lisboa, Faculdade de Ciências, Departamento de Informática.

Stopwords

1. BASIC INFORMATION

1.1 *Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc.)*

LX-Stopwords resource is a manual list of words from Portuguese composed by 2631 words of 51 types. The words are grouped in three big classes, arranged according to their morpho-syntactic category and inflectional feature value (closed classes, open classes, and multi-word units). This list was created as a support resource to develop CRIVO/*EtiFac* tool (see Branco & Silva, 2001), a tool for the semiautomatic annotation of corpora. With this in mind, the list seeks to be an as exhaustive as possible repository of all word forms that belong to closed classes, items typically with high frequency and fixity.

Taking into account the ambiguity between words of different categories, which means that some words from closed classes (1866 words) can be part of others categories, two classes were added to the list: open classes (592 words) and multi-word units (173 words), including only the words already contained in closed classes.

This wordlist was collected in the context of NeXing – Natural Negation Modeling and Processing[4] project whose the main goal was to contribute for improving the automated mapping between (orthographic) form and (linguistic) meaning, on the one hand, and between (linguistic) meaning and knowledge (representation), on the other hand, in what concerns natural language negation.

1.2 Representation of the lexicon (flat files, database, markup)

The corpus is represented in .txt format.

1.2 *Character encoding*

The characters are in UTF-8 code.

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)

Name: António Branco

[4] It can be visited at http://www.di.fc.ul.pt/~ahb/nexing.htm.

Address: Departamento de Informática NLX - Grupo de Fala e Linguagem Natural
Faculdade de Ciências da Universidade de Lisboa, Edifício C6
Campo Grande 1749-016 Lisboa
Position: Assistant Professor
Affiliation: Faculty of Sciences, University of Lisbon
Telephone: +351 217 500 087
Fax: +351 217 500 084
E-mail: antonio.branco@di.fc.ul.pt

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be available on the META-SHARE platform.

2.3 Copyright statement and information on IPR

This resource is a free license-based for research purposes and free license-based for commercial purposes, with attribution and no redistribution allowed. It will be available on the META-SHARE platform.

3. TECHNICAL INFORMATION

3.1 Directories and files

The archive that can be uploaded on the META-SHARE is a .zip file with two files: one .xml and one .xsd, which contains the .xml specification file.

3.2 Data structure of an entry

In the .txt file, the data are organized by classes of words according to their morpho-syntactic category which are sub-specified by inflectional feature value. As shown in the examples below, each class of words is divided in sub-classes taking into account the grammatical category introduced by the symbol <_> (cf. example A.), followed, when applicable, by <#> features values (gender <f/m/g(both)>, number <s/p/n(both)> (cf. example B.), and person <1/2/3> (cf. example C.):

A.

```
<entries>
        <sub-class>_PREP</sub-class>
            <list>
                <stopword>juntamente com</stopword>
                <stopword>por causa de</stopword>
                <stopword>até a</stopword>
                <stopword>mediante</stopword>
                <stopword>como</stopword>
                <stopword>enquanto</stopword>
                <stopword>segundo</stopword>
                <stopword>quando de</stopword>
                <stopword>a</stopword>
```

B.

```
<entries>
        <sub-class>_WD#fs</sub-class>
            <list>
                <stopword>segunda</stopword>
```

```
                    <stopword>segunda-feira</stopword>
                    <stopword>terça</stopword>
                    <stopword>terça-feira</stopword>
                    <stopword>quarta</stopword>
                    <stopword>quarta-feira</stopword>
                    <stopword>quinta</stopword>
                    <stopword>quinta-feira</stopword>
                    <stopword>sexta</stopword>
                    <stopword>sexta-feira</stopword>
                </list>
</entries>
```

C.

```
<entries>
        <sub-class>_PRS#gs1</sub-class>
            <list>
                <stopword>eu</stopword>
                <stopword>mim</stopword>
            </list>
</entries>
```

3.3 Lexicon size (nmb. of lexical items, KB occupied on disk)

The corpus is composed by 2631 words with 14.3 KB compressed (137.6 KB uncompressed) for disk storage.

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

The language of the list is European Portuguese.

4.2 Entry Type

For this information, see, please, item 3.2.

4.3 Attributes and their values

For the sake of completeness, the list of tags below contains all the tags that may occur in this lexicon, including open class tags.

Tagset	
Tag	**Description**
_DA	Definite Article
_UM	Occurrences of "um" or "uma"
_IA	Indefinite Articles (except "um" and "uma", see _UM)
_QNT	Quantifiers
_IND	Indefinites

_DEM	Demonstrative
_POSS	Possessive
_PRS	Personals
_CL	Clitics
_INT	Interrogative
_REL	Relatives
_EXC	Exclamatives
_CJ	Conjunctions
_PREP	Prepositions
_CARD	Cardinals (except "um" and "uma", see _UM)
_MGT	Magnitude classes
_ORD	Ordinals
_DFR	Denominators of fractions
_WD	Week Days
_MTH	Months
_ADV	Adverbs
_UNIT	Measurement Units (when in abbreviated form)
_EOE	End of Enumeration
_STT	Social Title
_EMP	Emphasis
_EL	Extra-linguistic
_DM	Discourse marker
_PL	Para-linguistic
_FRG	Fragment
_ITJ	Interjections
_CN	Common noun
_ADJ	Adjective
_VAUX	Auxiliar Verb "ter" and "haver" preceding _PPT in compound tenses
_INFAUX	Auxiliar Verb (Infinitive)
_GERAUX	Auxiliar Verb (Gerund)
_V	Verb (other than PPA, PPT, INF or GER)
_PPT	Past Participle preceded by aux. verb "ter" or "haver" in compound tenses
_PPA	Other Past Participles
_GER	Gerund
_INF	Infinitive
_NP	Noun Phrase
_PP	Prepositional Phrase
_PNM	Part of Name

_PADR	Part of Address
_LTR	Letters
_DGT	Digits
_DGTR	Roman numerals
_PNT	Punctuation
_SYB	Symbol
_EADR	Electronic Address
_TERMN	Terminations (for optional plu./fem./etc")

4.4 Coverage of the lexicon

The LX-Stopwords list works on the general language.

4.5 Intended application of the lexicon

The wordlist can be used in linguistic research and also in some NLP aplications.

4.6 POS assignment

All words were manually grouped and tagged (morpho-syntactic tagging) according to POS-Tagger tagset (see Silva, 2007).

4.7 Reliability (automatically/manually constructed)

The wordlist and the annotation were mannually constructed once the main and final goal was to construct a completely and accurately tagged resource.

5. RELEVANT REFERENCES AND OTHER INFORMATION

Branco, António and João Silva, 2001. EtiFac: A Facilitating Tool for Manual Tagging. In *Actas do XVII Encontro Anual da Associação Portuguesa de Linguística (APL'02)*, pp. 81-90.

Silva, João, 2007. *Shallow Processing of Portuguese: From Sentence Chunking to Nominal Lemmatization*. MSc thesis, University of Lisbon. Published as Technical Report DI-FCUL-TR-07-16.

Treebank

I. Basic Information

1.1. Corpus information

The CINTIL-TreeBank (Branco et al., 2011) is a corpus of syntactic constituency trees of Portuguese texts composed of 10,039 sentences and 110,166 tokens taken from different sources and domains: news (8,861 sentences;

101,430 tokens), novels (399 sentences; 3,082 tokens) (see 3.2.). In addition, there are 779 sentences (5,654 tokens) that are used for regression testing of the computational grammar that supported the annotation of the corpus (cf. section 4.6.).

For the creation of this TreeBank we adopted a semi-automatic analysis with a double-blind annotation followed by adjudication. The resulting dataset contains one information level: phrase constituency.

The main motivation behind the creation of this resource was to build a high quality data set with syntactic information that could support the development of a large set of automatic resources and tools for Portuguese for NLP studies.

The development of this resource started under the project SemanticShare – Resources and Tools for Semantic Processing (at: http://nlx.di.fc.ul.pt/projects.html) whose main goal was to generate a deep linguistic annotated corpus of Portuguese, with manually verified grammatical representations.

The following table displays a breakdown of the CINTIL-TreeBank corpus:

CINTIL-TreeBank				
Sub-corpus	**id**	**Sentences**	**Tokens**	**Domain**
Sentences for regression testing	aTSTS	779	5,654	Test
CINTIL-International Corpus of Portuguese5	bCINT	1,219	13,516	News
	cCINT	399	3,082	Novels
CETEMPúblico	eCTMP	7,541	86,905	News
Penn TreeBank (translation)	dPENN	101	1,012	News
Total		10,039	110,166	

1.2. Representation of the corpora (flat files, database, markup)

The corpus is a single file in a xml format.

1.*3. Character encoding*

5 CINTIL-International Corpus of Portuguese was the first corpus, but only partly, to being used to integrate our corpus of syntactic trees of constituencies and, thus, give it the name.

The characters are in UTF8 code.

II. Administrative Information

2.1. Contact person

Name: António Branco

Address: Departamento de Informática NLX - Grupo de Fala e Linguagem Natural, Faculdade de Ciências da Universidade de Lisboa, Edifício C6, Campo Grande 1749-016 Lisboa

Position: Assistant professor

Affiliation: Faculty of Sciences, University of Lisbon

Telephone: +351 217 500 087

Fax: +351 217 500 084

E-mail: antonio.branco@di.fc.ul.pt

2.2. Delivery medium (if relevant; description of the content of each piece of medium)

This resource is available through META-SHARE.

2.3. Copyright statement and information on IPR

This resource is available for both research and commercial purposes, with attribution, and no redistribution nor derivatives allowed. It will be available on the META-SHARE.

III. Technical Information

3.*1. Directories and files*

The archive that can be uploaded on the Meta-Share is a .zip file with two files: one .xml and one .xsd, which contains the .xml specification file.

3.2. Data structure of an entry

For the .xml file with the set of sentences, the data is organized with one sentence per entry. Each entry contains the sentence id (concatenated with sub-corpus/sentence number), sentence in raw text, and s-expressions or parenthesis format tree, as shown in the example below:

<sentence>

<id>aTSTS-001/11</id>

<raw>A criança obedece apenas à mãe.</raw>

<tree>(S (S (NP (ART A) (N criança)) (VP (V obedece) (PP (ADV apenas) (PP (P a_) (NP (ART a) (N mãe)))))) (PNT .))</tree>

</sentence>

3.3. *Corpus size (nmb. of tokens, NB occupied in disk)*

The corpus is composed by 10.164 sentences with 636 *KB* compressed (3.1 *MB* uncompressed) for disk storage.

IV. Content Information

4.*1. Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)*

This is a monolingual and a semi-automatic annotated corpus.

4.*2. The natural language(s) of the corpus*

The language of the corpus is Portuguese with pre-spelling reform of 19906.

4.*3. Domain(s)/register(s) of the corpus*

Concerning to text registers represented into the corpus, it comprises news from daily and general newspapers (8,861 sentences), literary language from novels (339 sentences), and, additionally, 779 sentences from test set (cf. section 1.1.).

4.4. Annotation in the corpus (if an annotated corpus)

4.4.1. Types of annotation (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

Trees with syntactic constituency.

4.4.2. Tags (if POS/WSD/TIME/discourse/etc – tagged or parsed)

Not applicable.

4.4.3. Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)

It does not apply.

4.4.4. Attributes and their values (if annotated)

This is the tag set used:

Phrasal and part-of-speech tags

Tag	Meaning

6 This means that the orthography rules used are those that are described by the Orthography Reform of 1945. The orthographic agreement of 1990 was adopted just in may of 2009 and is being implemented until 2012.

A	Adjective
AP	Adjective Phrase
ADV	Adverb
ADVP	Adverb Phrase
C	Complementizer
CP	Complementizer Phrase
CARD	Cardinal
CONJ	Conjuction
CONJP	Conjuction Phrase
D	Determiner
DEM	Demonstrative
N	Noun
NP	Noun Phrase
P	Preposition
PP	Preposition Phrase
POSS	Possessive
QNT	Predeterminer
S	Sentence
V	Verb
VP	Verb Phrase

4.5. Intended application of the corpus

The corpus can be used in linguistic research and, on the other hand, to development of constituency parsers.

4.6. Reliability of the annotations (automatically/manually assigned) – if any

In order to achieve a gold-standard corpus with high accuracy, the CINTIL-TreeBank is created by a two-phase process, where an automatic annotation is then manually revised by language experts with post-graduate degrees in Linguistics. More specifically, in the first stage, a deep computational grammar (see Branco and Costa, 2008) is used to generate all the possible parses for a given sentence (the parse forest). This is followed by a manual disambiguation stage where the correct parse is chosen from among those in the parse forest. This second stage follows a double-blind annotation method, where two annotators work independently and, for those cases where their decisions differ, a third annotator (the adjudicator) is brought in to make the final decision. For this corpus, the level of inter-annotator agreement (ITA) is 0.83 in terms of the specific inter-annotator metric developed for this kind of corpora and annotation (Castro, 2011).

The automatic annotation assigns only argumental semantic roles, leaving modifiers with a underspecified 'M'. These tags are manually specified, again following the same annotation method as before (double-blind annotation with adjudication). For this task, the ITA is 0.76.

V. Relevant References and Other Information

Branco, A., Silva, J., Costa, F., and Castro, S., 2011, "CINTIL-TreeBank Handbook: Design options for the representation of syntactic constituency". In *Technical Reports Series*, University of Lisbon, Department of Informatics.

Branco, A. and Costa, F., 2008, "A computational grammar for deep linguistic processing of portuguese: LXGram". In *Technical Reports Series*. University of Lisbon, Department of Informatics, 2008.

Castro, Sérgio, 2011, *Developing Reliability Metrics and Validation Tools for datasets with deep linguistic Information*, MA Dissertation, Universidade de Lisboa, Faculdade de Ciências, Departamento de Informática.

IST - Instituto Superior Técnico

Endogenous resources

TED Talks

I.BASIC INFORMATION

1.Corpus composition

This corpus is composed of the audio, the automatic transcriptions, the manual transcriptions and the translations for Portuguese of TED Talks from Al Gore (On averting climate crisis), Dann Dennett (On Our Consciousness) and Malcolm Gladwell (On Spaghetti Sauce). TED is the acronym for Technology, Entertainment, Design.

2.Representation of the corpora (flat files, database, markup)

Each one of the three Ted Talk is composed of a set of files:

1 a wav file: audio file
2 a xml file: xml file containing the automatic transcription
3 a trs file: containing the manual transcriptions
4 two parallel txt files containing the manual transcription and the respective translation, without any annotation.

3.Character encoding

The characters in the text files are encoded in ISO-8859-1 (Latin1).

II.ADMINISTRATIVE INFORMATION

1.Contact person

Name:Luísa Coheur
Address: Rua Alves Redol, nº 9, 1000-029, Lisboa
Affiliation: IST/INESC-ID
Position: Assistant Professor
Telephone: +351 3100314
Fax: +351-213-145-843
e-mail: luisa.coheur@inesc-id.pt

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive.

2.3 Copyright statement and information on IPR

The resource is free.

III.TECHNICAL INFORMATION

1.Directories and files

The archive that will be uploaded on the MetaShare platform will contain three folders, each one dedicated to one Ted Talk and being composed of the files described in Section 1.2.

2.Data structure of an entry

Both the XML and the TRS files have a data type definition file associated (respectively, alert-pt.dtd and trans-14.dtd) that are provided in the archive.

Considering the xml file, the dtd is alert-pt.dtd an example of a transcript segment is:

<TranscriptSegment>
<TranscriptGUID>7</TranscriptGUID>
<AudioType start='2421' end='2848' conf='0.456900'>Clean</AudioType>
<Time reasons='' start='2421' sns_conf='0.953300' end='2848'/>
<Speaker name='Homem' gender='M' id_conf='0.221200' gender_conf='0.921100' known='F' id='1002'/>
<SpeakerLanguage native='T'>PT</SpeakerLanguage>
<TranscriptWordList>
<Word start='2438' end='2453' conf='0.551049'>a</Word>
<Word start='2454' end='2490' conf='0.736239'>truly</Word>
<Word start='2491' end='2522' conf='0.855250'>great</Word>
<Word start='2523' end='2557' conf='0.696083'>honor</Word>
<Word start='2558' end='2563' conf='0.877075'>to</Word>
<Word start='2566' end='2588' conf='0.809444'>have</Word>
<Word start='2589' end='2602' conf='0.848018'>the</Word>
<Word start='2603' end='2670' conf='0.959818'>opportunity</Word>
<Word start='2671' end='2676' conf='0.894171'>to</Word>
<Word start='2677' end='2701' conf='0.786599'>come</Word>
<Word start='2702' end='2706' conf='0.895611'>to</Word>
<Word start='2707' end='2715' conf='0.828017'>the</Word>
<Word start='2716' end='2763' conf='0.813053'>stage</Word>
<Word start='2764' end='2809' conf='0.873642'>twice</Word>
</TranscriptWordList>
</TranscriptSegment>

Considering the trs file, an example of a turn is:

<Turn speaker="spk2" startTime="134.433" endTime="138.770">
<Sync time="134.433"/>
<this>
<Event desc="rire en fond" type="noise" extent="instantaneous"/>
This was a rented ford taurus.
</Turn>
<Turn startTime="138.770" endTime="141.304">
<Sync time="138.770"/>
<Event desc="rire en fond" type="noise" extent="instantaneous"/>
</Turn>

In the other two files each line contains a single sentence.

3.Corpora size (nmb. of tokens, MB occupied on disk)

The TXT, TRS and XML files have a total of 35,399 word tokens (17,997 word tokens for the English and Portuguese TXT files). The whole resource occupies 114.9 MB.

IV.CONTENT INFORMATION

1.Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

This corpus is in multilingual, parallel and annotated.

2.The natural language(s) of the corpus

The languages of the corpus are English and Portuguese.

4. 3 Domain(s)/register(s) of the corpus

Al Gore's talk is about the climate, Dan Dannett talk is about consciousness and Malcolm Gladwell talk focus on food industry

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

Only the xml and the trs corpora are annotated.

4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),

These tags are described in the DTD files.

4.4.3Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)

Not relevant

4.4.4 Attributes and their values (if annotated)

Described in the DTD files.

1.5 Intended application of the corpus

This corpus can be used to test an English recognizer, as both the audio and the manual transcriptions are provided; it can also be used to test a machine translation system for EN-PT.

1.6 Reliability of the annotations (automatically/manually assigned) – if any

The translation was manually done.

5 RELEVANT REFERENCES AND OTHER INFORMATION

PTSTAR Golden Collection – Cross-Language Unit Elicitation alignments (CLUE)

I.BASIC INFORMATION

1.Corpus composition

This corpus consists of a set of manual alignments of 400 parallel sentences from the Europarl corpora [1] in four languages (pt, en, es, fr), being considered the following pairs: en-es, en-fr, en-pt, es-fr, pt-es. This work deeply extends the corpus detailed in [2].

2.Representation of the corpora (flat files, database, markup)

The corpus is composed of several txt files, namely:

- four files containing the 400 parallel sentences in each language;
- a file for each pair en-es, en-fr, en-pt, es-fr, pt-es and pt-fr containing the word alignments;
- a file for each pair en-es, en-fr, en-pt, es-fr, pt-es and pt-fr containing the multiword units alignments.

3.Character encoding

Characters are encoded in ISO-8859-1 (Latin1).

II.ADMINISTRATIVE INFORMATION

1.Contact person

Name:Luísa Coheur
Address: Rua Alves Redol, nº 9, 1000-029, Lisboa
Affiliation: IST/INESC-ID
Position: Assistant Professor
Telephone: +351 3100314
Fax: +351-213-145-843
e-mail: luisa.coheur@inesc-id.pt

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive.

2.3 Copyright statement and information on IPR

The resource is free.

III.TECHNICAL INFORMATION

1.Directories and files

The archive that will be uploaded on the MetaShare platform will contain one folder with 4 files with extensions pt, es, fr and es (one for each considered language), 6 files with extension wa (word alignments) and 6 files with extension mwu (multiword units' alignments).

2.Data structure of an entry

Each file containing the sentences in the different languages (extensions pt, en, fr and es) has a sentence per line.

Each file with the extension wa contains, in each line, three digits and a S or a P (ex: 1 14 17 P). The first digit represents the sentences/lines in the parallel files that are being aligned; the second an third digits correspond to the position of the words being aligned; S represents a sure alignment and P a possible one. Details about this can be found in [2]. Very detailed guidelines are included in the corpus.
Each file with the extension mwu contains, in each line, five digits and an S or a P. The first digit represents the sentences/lines in the parallel files that are being considered; the second and third digit represent the positions of the word units being aligned in the source language (the fourth and the fifth digit represent the same for the target language); S represents a sure alignment and P a possible one.

3.Corpora size (nmb. of tokens, MB occupied on disk)

Each parallel file contains 400 sentences. The .wa files (word alignments) totalize 48130 alignments and the .wmu files (multi-word units) totalize 22,099 alignments. The whole set of files occupies 1.3 MB.

IV.CONTENT INFORMATION

1.Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

This corpus is multilingual, parallel and lightly annotated.

2.The natural language(s) of the corpus

The languages of the corpus are Portuguese, English, French and Spanish.

4. 3 Domain(s)/register(s) of the corpus

The corpus has sentences from the European parliament sessions [1].

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

Only the files with extensions pt, en, fr and es are annotated (in xml).

4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),

Each sentence of the previously mentioned files is marked with the tag <s snum=N>, where N represents the sentence number.

4.4.3Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)

As previously stated, in order to establish the alignments between the different languages, for each considered languages' pair, two files are provided. The one

with extension wa represents the word alignments; the one with extension mwu represents the multiword units alignments. The used notation is explained in Section 3.2.

4.4.4 Attributes and their values (if annotated)

Not relevant.

1.7 Intended application of the corpus

This corpus can be used as a gold collection for word alignments.

1.8 Reliability of the annotations (automatically/manually assigned) – if any

The alignments were manually built.

5 RELEVANT REFERENCES AND OTHER INFORMATION

[1] P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In MT summit, volume 5.

[2] João Graça and Joana Paulo Pardal and Luísa Coheur and Diamantino António Caseiro, Building a golden collection of parallel Multi-Language Word Alignment, *The 6th International Conference on Language Resources and Evaluation, LREC 2008*, May. 2008

Restricted exogenous resources

TAP

1 BASIC INFORMATION

1.1 Corpus composition

UP/TAP is a parallel/comparable corpus containing articles extracted from the TAP UP-magazine. All the articles are written both in Portuguese and in English, making it a parallel corpus. The UP Magazine focuses on interesting issues and facts from Portugal. Its contents are geared towards travel and travelers, and were created with the collaboration of a wide number of people around the globe.
The current version of the corpus (June 2012) contains data from 51 editions, which corresponds to 33879 aligned sentences. The Portuguese data contains about 771.442 words and the English data contains 775504 words.

1.2 Representation of the corpora (flat files, database, markup)

The corpus is composed of txt files, each one corresponding to the original PDF name from where the text data was extracted:

1.3 Character encoding

All data is encoded in UTF-8.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name:Luísa Coheur
Address: Rua Alves Redol, nº 9, 1000-029, Lisboa
Affiliation: IST/INESC-ID
Position: Assistant Professor
Telephone: +351 3100314
Fax: +351-213-145-843
e-mail: luisa.coheur@inesc-id.pt

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive.

2.3 Copyright statement and information on IPR

The resource is free.

3 TECHNICAL INFORMATION

3.1 Directories and files

The main corpus directory contains 51 subdirectories, each of which containing aligned files extracted from each one of the 51 editions. The extension of each file is either ".pt" for Portuguese content, or ".en" for English content.

3.2 Data structure of an entry

Each file contains one sentence per line, and each line in one language corresponds to the same line in the other language.

3.3 Corpora size (nmb. of tokens, MB occupied on disk)

The size of the distributed archive is 4Mbytes, whereas in its uncompressed format the corpus occupies 12Mbytes.
The current version of the corpus contains 2243 aligned files, corresponding to 33879 aligned sentences, and comprising about 771K words for Portuguese and about 776K words for English.

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

This corpus is comparable multilingual corpus.

4.2 The natural language(s) of the corpus

The languages of the corpus are Portuguese and English.

4. 3 Domain(s)/register(s) of the corpus

The corpus contains text articles extracted from the TAP UP-magazine. Its contents are geared towards travel and travelers, and focuses on interesting issues and facts of Portugal.

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

The content of each topic was split into two files with the same name but a different extension (.pt, .en). The content of each one of the files was split into sentences in the way that the translation of each line in one file corresponds to the same line in the other file.

4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),

None

4.4.3Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)

As previously stated, in order to establish the alignments between the two languages, two files have been created only differing in their extension. The one with extension ".pt" represents the Portuguese data; the one with extension ".en" represents the corresponding English data. The translation of each line in one of the files corresponds to the same line in the other file.

4.4.4 Attributes and their values (if annotated)

Not relevant.

4.5 Intended application of the corpus

This corpus can be used for speech translation.

4.6 Reliability of the annotations (automatically/manually assigned) – if any

Automatically assigned.

5 RELEVANT REFERENCES AND OTHER INFORMATION

UNIMAN-University of Manchester

Endogenous resources (tools)

U-Compare Platform

I. Basic Information

Tool name

U-Compare platform

Overview and purpose of the tool

The purpose of the U-Compare platform (Kano et al., 2011; Kano et al; 2009) is to facilitate easy and rapid development and evaluation of NLP and text mining systems. It includes utilities (including a graphical user interface, the U-Compare workbench, see separate record in META-SHARE) to create workflows from individual, interoperable NLP tools and resources, a customisable system to create and evaluate different workflows, and different utilities to visualise different types of annotations produced by workflows. U-Compare is packaged with the world's largest repository of UIMA components. This repository, which originally consisted largely of tools for processing English biomedical text, in being considerably enlarged as ongoing work, to include tools that can operate on a number of European languages and multilingual tools (Ananiadou et al, 2011; Thompson et al, 2011).

A short description of the algorithm

U-Compare is a platform rather than a tool that performs a specific purpose. Hence, there is no specific algorithm to describe. However, this section provides some further implementation details.

U-Compare is built on top of the Unstructured Management Architecture (UIMA)[7] (Ferruci et al., 2006), which is a generic framework widely adopted by the NLP community to improve the interoperability of tools/components. U-Compare provides several tools that are necessary for NLP application development, but which are not provided by UIMA, due to the more general nature of UIMA.

U-Compare requires that tools and resources used in workflows are UIMA components. UIMA components are interoperable, in the sense that their input/output mechanisms are standardised – they must obtain their input by reading annotations from a data structure, the Common Analysis Structure (CAS), that is accessible to all components in a workflow, and output consists of adding new annotations to the CAS, or updating existing ones. Existing tools can be "wrapped" as UIMA components by writing code to convert their input/output formats to those required by UIMA.

The U-Compare platform allows workflows to be created simply by specification of which components to run, and in which order. This is helped by the Workbench graphical user interface (see separate META-SHARE record), which allows users to create workflows using drag-and-drop actions. Comparison workflows can also be created to compare and evaluate the outputs of several different workflows that perform the same task, possibly against a gold standard annotated corpus if this is available, with results displayed in graphical form.

[7] http://uima.apache.org

Whilst U-Compare workflows can be created using any UIMA components, U-Compare also defines its own type system (described in a separate META-SHARE record), which aims to facilitate semantic interoperability between components. By "types", we mean the categories of annotations that are input/output by the UIMA components. UIMA itself does not define or impose the use of a particular type system. This means that interoperability between components produced by different developers can be difficult to achieve, as they may use different sets of types as input/output. The U-Compare type system aims to resolve this problem, by providing a hierarchical *sharable* system of the most common types of annotations produced by NLP tools. The idea is that compatibility between a large number of UIMA components can be achieved, at least at an intermediate level of the hierarchy, through mapping of types to this type system. All components in the U-Compare repository are compatible with the U-Compare type system.

The U-Compare platform can be used independently of the U-Compare Workbench Graphical User Interface to run workflows directly from the command line. Additionally, U-Compare provides UIMA components for standard I/O streams that communicate in a simple standoff annotation format, which allows easy embedding of workflows into other systems, regardless of programming language.

2. TECHNICAL INFORMATION

Software dependencies and system requirements

The U-Compare platform can be used in any environment in which Java 6 is available. At least the first time the system is run, an Internet connection is required, since the most up-to-date relevant files are downloaded from the internet.

Installation

No specific installation is required. U-Compare can be started directly from the Internet by clicking on the "Start U-Compare" button on this page: http://www.nactem.ac.uk/ucompare/index.html

However, it is preferable to start U-Compare from the command line, by downloading the file UCLoader.classfrom http://u-compare.org/downloads/UCLoader.class.

See also http://www.nactem.ac.uk/ucompare/launch.html for more information

Execution instructions

From the command line, the U-Compare workbench is started by running the UCLoader.class file, e.g.

java -Xms700m -Xmx1000m UCLoader

The –Xms and –Xmx specified the minimum and maximum memory allocated to U-Compare. The more memory is allocated, the quicker U-Compare will run. Note that the first time U-Compare is launched, relevant files will be downloaded from the internet. Therefore, the first time the system is launched, it may take a considerable amount of time to start up.

The default behaviour of UCLoader is to start the U-Compare Workbench interface. However, other options can be specified that allow workflows to be run from the command line, without starting up the interface. The general way to run a workflow from the command line is as follows:

java -cp . -XmsXXXm –XmxXXXm
-Djavaws.workflow.path="path/to/yourworkflow.xml" UCLoader --jnlp http://u-compare.org/lib/u-compare-runworkflow.jnlp

“-cp” is the Java VM option to specify your classpath (in this case the current directory ``.'' is specified). You should include UCLoader.class in your classpath.

“-Xms” and ``-Xmx'' are the Java VM options to specify the amount of heap memory allocation.

An example workflow that can be run on the command line, together with more details, are provided here:

http://www.nactem.ac.uk/ucompare/developerguide/Command_Line_Mode_without_U.html

Input/Output data formats

Input data formats

Workflows that are run in the platform obtain input data via Collection Reader components. Currently, only text may be read in, but in the future, files of other modalities, e.g. speech, are planned. A collection reader reads a text or set of texts, which may be unannotated or may already contain annotations. The U-Compare library includes several types of collection reader, e.g. to read in text from an input window, from a directory of files or, in the case or a workflow being run from the command line, from standard input. Several corpus-specific readers are also provided, that read in annotated data.

Output data format

The result of running a workflow in U-Compare is a set of annotations added to the UIMA CAS. The contents of the CAS may be exported to different formats, e.g., to files in the XMI (XML Metadata Interchange) format (Grose et al, 2002) or inline XML format, or to standard output, in the case that the workflow is run from the command line.

The U-Compare platform also includes annotation visualisation tools, which display annotations produced by the workflow as underlines and arcs superimposed on the document text, and highlight differences discovered during component comparison, as well statistics relating to the comparison, such as F-score, precision, recall, etc.

Integration with external tools

The platform does not require the use of any external tools. As mentioned above, workflows created using the U-Compare platform can by embedded into other applications.

3. Content Information

Figure 1 illustrates two of the visualisation tools available in the U-Compare platform. On the left is the display of workflow comparison/evaluation, and on the right is the display of an annotated document.

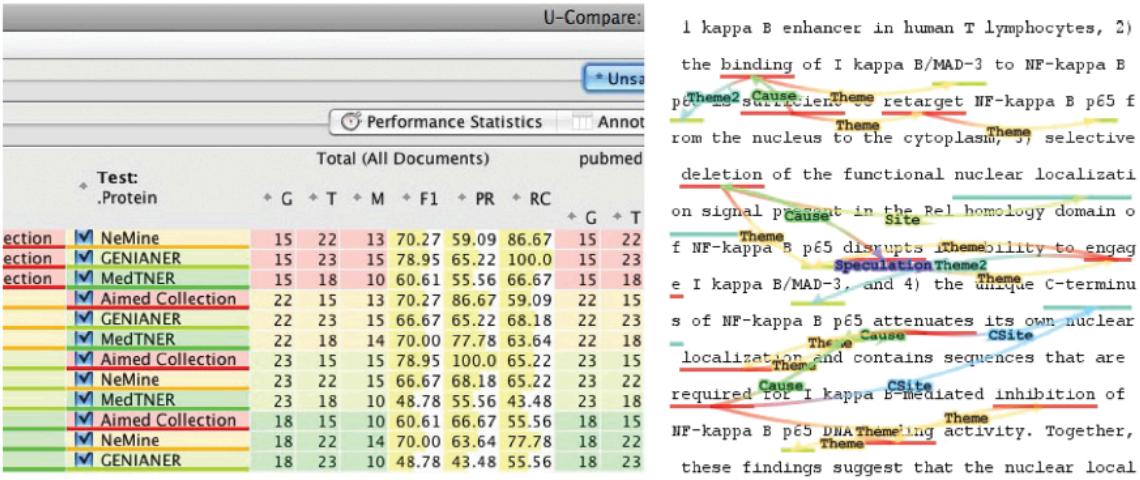


Figure 1: U-Compare visualisation tools

In Figure 2, text based output is shown, as a result of running a workflow from command line. For each annotation, the annotation offsets, annotation type and annotation attributes are displayed.

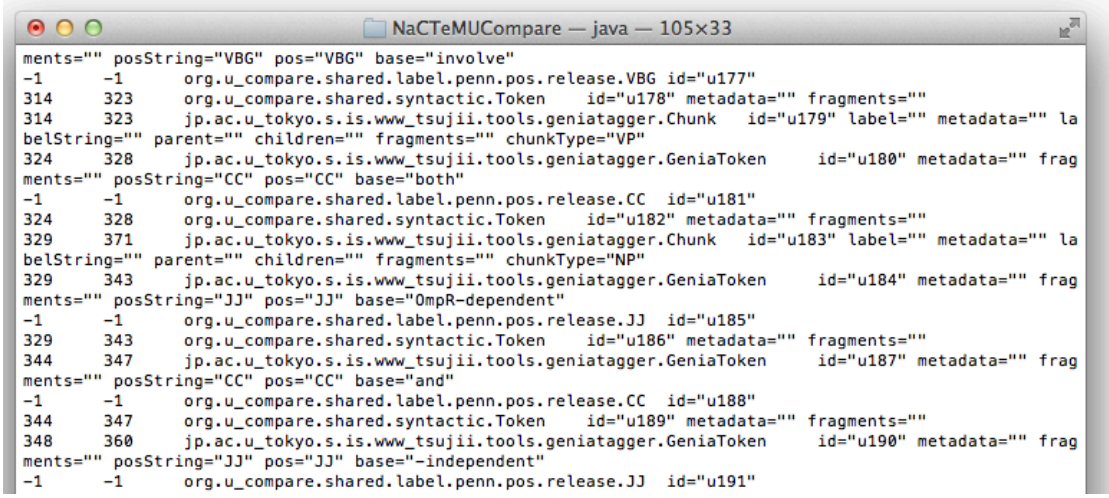


Figure 2: Ouput of a workflow in the command line

4. Administrative Information

Contact

For further information, please contact Sophia Ananiadou:
sophia.ananiadou@manchester.ac.uk

5. References

Ananiadou, S., Thompson, P., Kano, Y., McNaught, J., Attwood, T. K., Day, P. J. R., Keane, J., Jackson, D. and Pettifer, S.. (2011). "Towards Interoperability of European Language Resources". *Ariadne*, 67

Ferrucci, D., Lally, A., Gruhl, D., Epstein, E., Schor, M., Murdock, J. W., Frenkiel, A., Brown, E.W. , Hampp T., Doganata, Y., Welty, C., Amini, L., Kofman, G., Kozakov, L. and Mass, Y. (2006). Towards an Interoperability Standard for Text and Multi-Modal Analytics. IBM Research Report RC24122.

Grose, T.J. Doney, G.C. and Brodsky, S.A. (2002). *Mastering XMI. Java Programming with XMI, XML, and UML.* John Wiley & Sons, Inc.

Kano, Y., Baumgartner Jr., W. A, McCrochon, L., Ananiadou, S., Cohen, K. B., Hunter, L. and Tsujii, J. (2009). U-Compare: share and compare text mining tools with UIMA. *Bioinfomatics*, 25(15), 1997-1998.

Kano, Y., Miwa, M., Cohen, K. B., Hunter, L., Ananiadou, S. and Tsujii, J.. (2011). U-Compare: a modular NLP workflow construction and evaluation system. *IBM Journal of Research and Development*, 55(3), 11:1 - 11:10.

Thompson, P., Kano, Y., McNaught, J., Pettifer, S., Attwood, T. K., Keane, J. and Ananiadou, S. (2011). Promoting Interoperability of Resources in META-SHARE. In *Proceedings of the IJCNLP Workshop on Language Resources, Technology and Services in the Sharing Paradigm (LRTS)*, pp. 50-58

U-Compare Workbench

BASIC INFORMATION

Tool name

U-Compare Workbench

Overview and purpose of the tool

The U-Compare Workbench (Kano et al., 2009; Kano et al., 2011) is a graphical user interface that operates on top of the U-Compare platform. The U-Compare platform allows users to build and evaluate NLP workflows. Workflows consist of one or more components, consisting of corpus readers and tools, such as tokenisers, POS taggers, named entity recognisers, etc. Workflows can be built using any components that are compliant with the UIMA framework[8] (Ferrucci et al., 2006). The Workbench provides several facilities, including the following:

- Rapid construction of workflows by dragging and dropping components from a library onto the workflow canvas
- Graphical display of comparison of the performance of alternative workflows and evaluation against gold standard data
- Import of new UIMA components into the library
- Export of components/workflows

The core library of components provided with U-Compare includes several different types of tools, including sentence splitters, tokenisers, part-of-speech taggers, lemmatisers, named entity recognisers, etc. Currently, the

[8] http://uima.apache.org/

majority of these are for English, with a focus on biomedical text. However, several UIMA components are currently under development for the processing of other European languages, such as Portuguese, Maltese, Romanian, Spanish, Catalan, Basque and Galician and French (Ananiadou et al., 2011; Thompson et al., 2011). These components will be added to the U-Compare library in the near future.

A short description of the algorithm

There is no algorithm to describe as such, as this is tool is a graphical user interface for the U-Compare platform.

TECHNICAL INFORMATION

Software dependencies and system requirements

The U-Compare workbench can be used in any environment in which Java 6 is available. At least the first time the system is run, an internet connection is required, since the most up-to-date relevant files are downloaded from the internet.

Installation

No specific installation is required. U-Compare can be started directly from the Internet by clicking on the "Start U-Compare" button on this page: http://www.nactem.ac.uk/ucompare/index.html

However, it is preferable to start U-Compare from the command line, by downloading the file UCLoader.classfrom http://u-compare.org/downloads/UCLoader.class.

See also http://www.nactem.ac.uk/ucompare/launch.html for more information

Execution instructions

From the command line, the U-Compare workbench is started by running the UCLoader.class file, e.g.

java -Xms700m -Xmx1000m UCLoader

The –Xms and –Xmx specified the minimum and maximum memory allocated to U-Compare. The more memory is allocated, the quicker U-Compare will run. Note that the first time U-Compare is launched, relevant files will be downloaded from the internet. Therefore, the first time the system is launched, it may take a considerable amount of time to start up.

Input/Output data formats

Input data formats

The first component in a workflow must read in some input data. Currently, this can only be text (annotated or not), although other modalities, such as speech, are planned. This first component must be a "collection reader", which reads the data to be processed into the UIMA Common Analysis Structure (CAS). This is the common data structure that can be accessed by all components in a workflow. Components obtain their input by reading annotations from the CAS, while the output of components is written to the CAS by creating new annotations, or updating existing annotations.

The library of U-Compare components includes several generic collection readers to read in plain text, e.g., from an input window or from a directory of files. Several corpus-specific readers (currently mainly are also provided to read annotated texts into the CAS. Collection readers for different annotated corpora can be added as required.

Output data format

As mentioned above, the output of a workflow is a set of annotations (possibly of various different types) that are added to the CAS. Different types of annotation viewers provided in the U-Compare workbench allow annotations to be viewed in different ways e.g., as simple text spans or as tree/HPSG structures (in the case of the display of parser results). It is possible for the annotations in the CAS to be written to an output file using a CAS consumer component. Components are provided in the U-Compare library to produce different types of output files, such as XMI, inline XML annotations, etc.

Integration with external tools

The workbench is intended to be run as a standalone application. However, workflows created using the workbench can be embedded into other applications.

CONTENT INFORMATION

In this section, some screenshots are provided to illustrate the functionality of the U-Compare Workbench. Further information about using U-Compare can be obtained from the documentation pages on the website:

- User manual: http://www.nactem.ac.uk/ucompare/userguide/index.html
- Developer manual: http://www.nactem.ac.uk/ucompare/developerguide/index.html

Figure 1 shows the main window of the U-Compare Workbench. On the right is the library of components, while on the left is the workflow canvas. To create a new workflow, components are simply dragged from the library onto the canvas, in the order in which they are to be executed. Components can also be reordered once they have been placed on the workflow canvas.

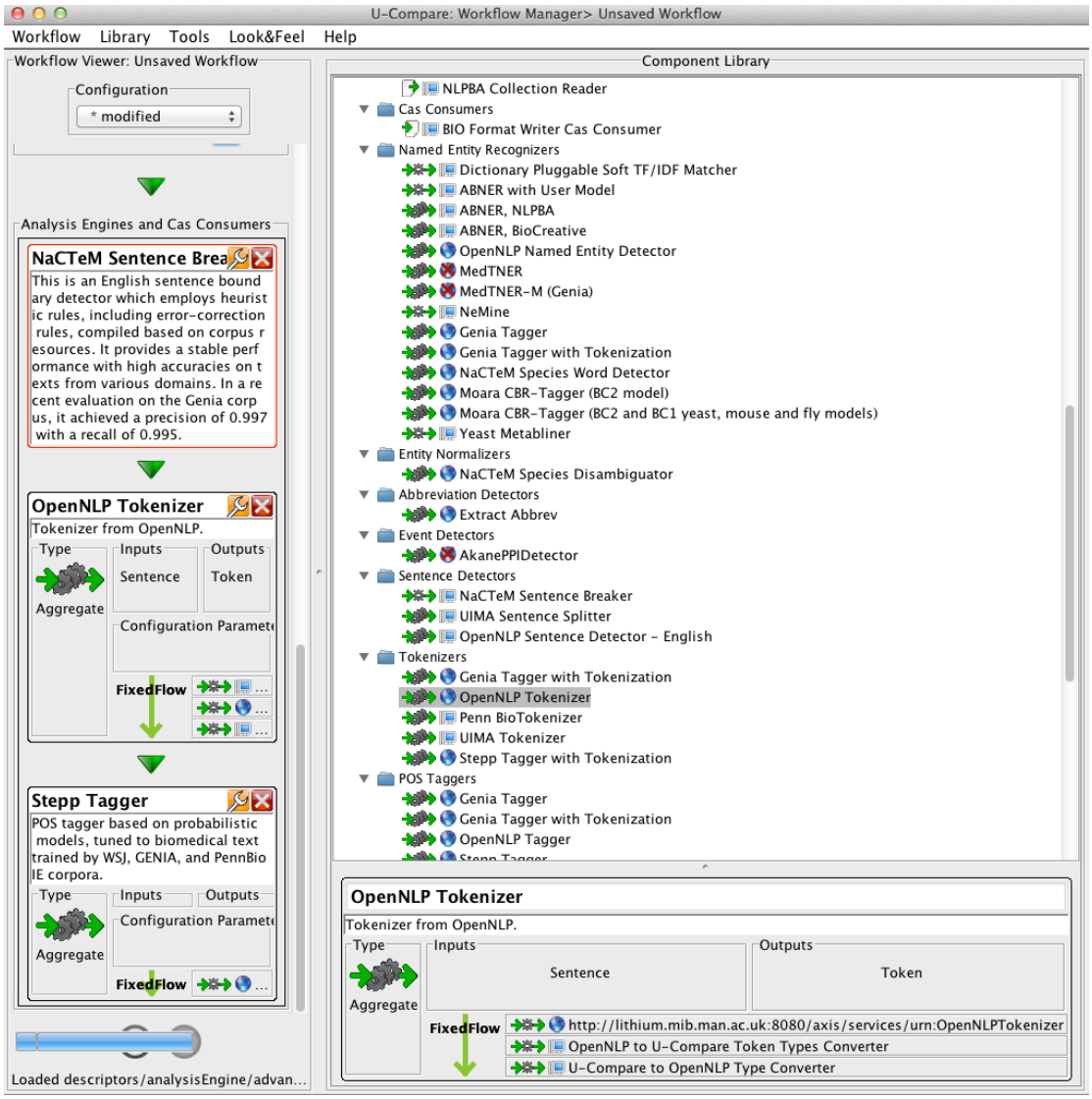


Figure 1: The main window of the U-Compare workbench

Figure 2 shows the display of the annotations in U-Compare's default annotation viewer, produced by the workflow shown in Figure 1. There are three types of annotations, highlighted using different colours, i.e. sentence annotations, tokens and part-of-speech annotations. Attributes associated with the different types of annotations can be viewed in a tabular format. This is illustrated on the right-hand side of Figure 2, which shows the table of attributes associated with the part-of-speech annotations. Each annotation stores the start and end offsets of the annotation, plus the part-of-speech, in the "posString" attribute. Clicking over a row in the table causes the corresponding annotation to be highlighted.

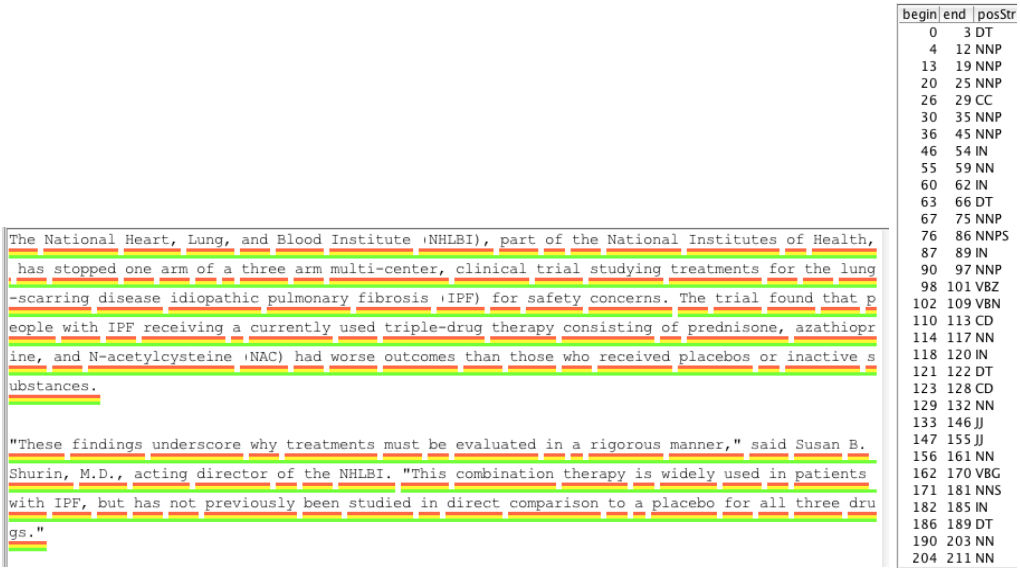


Figure 2: Annotation viewer

Figure 3 shows the output of a comparison workflow. The outputs of 2 named entity recognisers (ABNER-NLPBA and ABNER-BioCreative) are compared against a gold standard corpus (Aimed). The system produces pairwise comparisons of the annotations, with different resources being assumed as the gold standard. Since Aimed is the gold standard corpus, only the rows in which Aimed is in the "Assumed Gold Standard" are meaningful in this comparison. In each row, several pieces of information are shown: the number of relevant annotations in the gold standard corpus (G), the number of annotations produced by the relevant tool (T), the number of matching annotations (M), the F1 score, precision (PR) and recall (RC). These figures are shown both for the collection as a whole, and for the individual documents in the corpus.

Assumed Gold Standard	Comparison Components	Total (All Documents)						pubmed abstract 11780382.xmi show					
		Boundary Match						Boundary Match					
.Protein	.Protein	G	T	M	F1	PR	RC	G	T	M	F1	PR	RC
Aimed	ABNER-NLPBA	15	23	15	78.95	65.22	100.0	15	23	15	78.95	65.22	100.0
ABNER-NLPBA	Aimed	23	15	15	78.95	100.0	65.22	23	15	15	78.95	100.0	65.22
Aimed	ABNER-BioCreative	15	21	15	83.33	71.43	100.0	15	21	15	83.33	71.43	100.0
ABNER-BioCreative	Aimed	21	15	15	83.33	100.0	71.43	21	15	15	83.33	100.0	71.43

Figure 3: Output of comparison workflow

ADMINISTRATIVE INFORMATION

Contact

For further information, please contact Sophia Ananiadou:
sophia.ananiadou@manchester.ac.uk

REFERENCES

Ananiadou, S., Thompson, P., Kano, Y., McNaught, J., Attwood, T. K., Day, P. J. R., Keane, J., Jackson, D. and Pettifer, S.. (2011). “Towards Interoperability of European Language Resources”. *Ariadne*, 67

Ferrucci, D., Lally, A., Gruhl, D., Epstein, E., Schor, M., Murdock, J. W., Frenkiel, A., Brown, E.W. , Hampp T., Doganata, Y., Welty, C., Amini, L., Kofman, G., Kozakov, L. and Mass, Y. (2006). Towards an Interoperability Standard for Text and Multi-Modal Analytics. IBM Research Report RC24122.

Kano, Y., Baumgartner, W. A., Jr., McCrohon, L., Ananiadou, S., Cohen, K. B., Hunter, L. (2009). "U-Compare: share and compare text mining tools with UIMA". *Bioinformatics,* vol. 25, no. 15, 1997-1998.

Kano, Y., Miwa, M., Cohen, K. B., Hunter, L. E., Ananiadou, S., & Tsujii, J. (2011). "U-Compare: A modular NLP workflow construction and evaluation system". *IBM Journal of Research and Development, 55*(3), 11:11-11:10.

Thompson, P., Kano, Y., McNaught, J., Pettifer, S., Attwood, T. K., Keane, J. and Ananiadou, S.. (2011). Promoting Interoperability of Resources in META-SHARE. In *Proceedings of the IJCNLP Workshop on Language Resources, Technology and Services in the Sharing Paradigm (LRTS)*, pp. 50-58

Restricted exogenous resources (tools)

U-Compare GENIA Sentence Detector

1. BASIC INFORMATION

Tool name

U-Compare GENIA Sentence Splitter

Overview and purpose of the tool

The purpose of the tool is to detect sentence boundaries in English text. It is trained on the GENIA corpus of biomedical abstracts (Kim et al., 2003) and so is particularly suitable for splitting sentences in biomedical texts.

The tool is provided as a UIMA[9] (Ferrucci et al., 2006) component, which forms part of the in-built library of components provided with the U-Compare platform (Kano et al., 2009; Kano et al., 2011; see separate META-SHARE record)[10] for building and evaluating text mining workflows. The U-Compare Workbench (see separate META-SHARE record), which provides a graphical drag-and drop interface for the rapid creation of workflows.

A short description of the algorithm

This sentence detector is trained on the GENIA corpus (Kim et al., 2003), using machine learning methods.

[9] http://uima.apache.org/
[10] http://nactem.ac.uk/ucompare/

2. TECHNICAL INFORMATION

Software dependencies and system requirements

In order to run U-Compare, Java 6 must be installed.

The UIMA component calls a web service. Hence, internet access is required.

Installation

There is no specific installation for U-Compare. The file UCLoader.class should be downloaded from http://u-compare.org/downloads/UCLoader.class

Execution instructions

U-Compare is started by running UCLoader.class from the command line. Since U-Compare can consume a large amount of memory, it is suggested to specify minimum and maximum memory usage when running U-Compare, as in the following example:

java –jar -Xms700m –Xmx 1000m UCLoader

The memory usage can be adjusted, but note that a minimum memory usage of 256 MB is recommended. Please also note that when U-compare is first started for the first, a large number of files will be downloaded, and so it will take some time to start. Subsequent launches will be quicker.

Once U-Compare has been started, the sentence detector tool can be executed through inclusion in workflow. This can be done simply by dragging and dropping it onto the workflow canvas using the graphical user interface of the U-Compare workbench. See the META-SHARE record "U-Compare Workbench" for more details.

Input/Output data formats

Input data formats

The tool operates on plain, unannotated text. Thus, the UIMA Common Analysis Structure (CAS) should contain the text to be analysed prior to the tool being executed. In a UIMA workflow, this could be achieved by reading in a single text or corpus of text. For example, U-Compare provides collection readers that can read in text from an input box, or otherwise read a directory of texts.

Output data format

The tool detects the boundaries of sentences and, for each sentence, adds an annotation to the UIMA CAS corresponding to the sentence. Different CAS consumers (such as those provided in U-Compare) can be used to write the contents of the CAS to a file or database format.

Integration with external tools

The tool can be run as part of a UIMA workflow, either using U-Compare or otherwise. For instructions of how to include components in UIMA workflows outside of U-Compare, see:

http://nactem.ac.uk/ucompare/developerguide/Using_U_Compare_Components_.html

3. CONTENT INFORMATION

Figure 1 shows the output of the tool in the U-Compare workbench. One of the sentences is highlighted. The sample text is taken the US National Library of Medicine website (http://www.nlm.nih.gov/databases/alerts/2011_nhlbi_ifp.html)

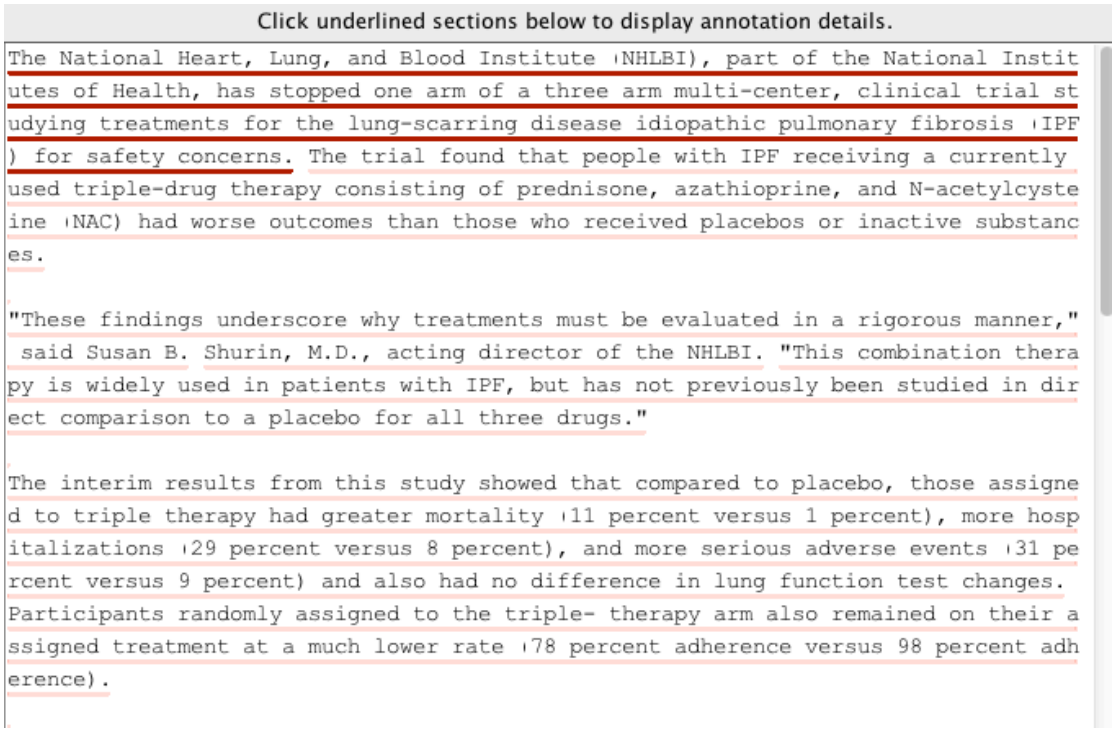


Figure 1: Output of the U-Compare GENIA Sentence Detector in the U-Compare workbench

Running the tool on the 4 KB text on a single core machine with 8 GB RAM takes around 2.43 seconds.

4. ADMINISTRATIVE INFORMATION

Contact

For further information, please contact Sophia Ananiadou:

sophia.ananiadou@manchester.ac.uk

5. REFERENCES

Ferrucci, D., Lally, A., Gruhl, D., Epstein, E., Schor, M., Murdock, J. W., Frenkiel, A., Brown, E.W. , Hampp T., Doganata, Y., Welty, C., Amini, L., Kofman, G., Kozakov, L. and Mass, Y. (2006). Towards an Interoperability Standard for Text and Multi-Modal Analytics. IBM Research Report RC24122.

Kim, Jin-Dong, Tomoko Ohta, Yuka Tateisi and Jun'ichi Tsujii (2003). GENIA corpus - a semantically annotated corpus for bio-textmining. Bioinformatics. 19(suppl. 1). pp. i180-i182, Oxford University Press, 2003.

Kano, Y., Baumgartner Jr., W. A, McCrochon, L., Ananiadou, S., Cohen, K. B., Hunter, L. and Tsujii, J. (2009). U-Compare: share and compare text mining tools with UIMA. *Bioinfomatics*, 25(15), 1997-1998.

Kano, Y., Miwa, M., Cohen, K. B., Hunter, L., Ananiadou, S. and Tsujii, J.. (2011). U-Compare: a modular NLP workflow construction and evaluation system. *IBM Journal of Research and Development*, 55(3), 11:1 - 11:10.

U-Compare GENIA Tokenizer

BASIC INFORMATION

Tool name

U-Compare GENIA Tokeniser (GENIA Tagger)

Overview and purpose of the tool

Tokenisation is one of the functionalities of the GENIA tagger, which additionally outputs the base forms, part-of-speech tags, chunk tags, and named entity tags. The tagger is specifically tuned for biomedical text such as MEDLINE abstracts.

The tool is a UIMA[11] (Ferrucci et al., 2006) component, which forms part of the in-built library of components provided with the U-Compare platform (Kano et al., 2009; Kano et al., 2011; see separate META-SHARE record)[12] for building and evaluating text mining workflows. The U-Compare Workbench (see separate META-SHARE record), which provides a graphical drag-and drop interface for the rapid creation of workflows.

A short description of the algorithm

The tokenisation and POS tagging functionality is based on an algorithm described in Tsuruoka et al. (2005), which uses a cyclic dependency network (Toutanova et al, 2003) with maximum entropy modelling with inequality constraints. The tokenisation and POS tagging functionality was trained on a corpus containing newspaper articles (Wall Street Journal corpus), and the GENIA (Kim et al., 2003) and PennBioIE corpora (Kulick et al., 2003), both containing biomedical text.

TECHNICAL INFORMATION

Software dependencies and system requirements

In order to run U-Compare, Java 6 must be installed.

The UIMA component calls a web service. Hence, internet access is required.

Installation

There is no specific installation for U-Compare. The file UCLoader.class should be downloaded from http://u-compare.org/downloads/UCLoader.class

[11] http://uima.apache.org/
[12] http://nactem.ac.uk/ucompare/

Execution instructions

U-Compare is started by running UCLoader.class from the command line. Since U-Compare can consume a large amount of memory, it is suggested to specify minimum and maximum memory usage when running U-Compare, as in the following example:

java –jar -Xms700m –Xmx 1000m UCLoader

The memory usage can be adjusted, but note that a minimum memory usage of 256 MB is recommended. Please also note that when U-compare is first started for the first, a large number of files will be downloaded, and so it will take some time to start. Subsequent launches will be quicker.

Once U-Compare has been started, the GENIA tool can be executed through inclusion in workflow. This can be done simply by dragging and dropping it onto the workflow canvas using the graphical user interface of the U-Compare workbench. See the META-SHARE record "U-Compare Workbench" for more details.

Input/Output data formats

Input data formats

The tool requires sentence split text as input. Thus, the UIMA Common Analysis Structure (CAS) must contain sentence annotations before this component is run. In a UIMA workflow, this could be achieved either by executing a component that performs sentence splitting prior to this component, or otherwise reading in a corpus of documents that already contains sentence annotations.

Output data format

One of the functionalities of the tool is to detect tokens in the text and assign parts-of-speech and base forms to them. An annotation is thus added to the CAS corresponding to each token in a document. Other annotations are also added by the GENIA tagger (e.g. named entity and chunk annotations), but we only focus on the token annotations here. Different CAS consumers (such as those provided in U-Compare) can be used to write the contents of the CAS to a file or database format.

Integration with external tools

The tool can be run as part of a UIMA workflow, either using U-Compare or otherwise. For instructions of how to include components in UIMA workflows outside of U-Compare, see:

http://nactem.ac.uk/ucompare/developerguide/Using_U_Compare_Components_.html

CONTENT INFORMATION

Figure 1 shows the output of the tool in the U-Compare workbench. Each token recognised is separately underlined. The sample text is taken the US National Library of Medicine website (http://www.nlm.nih.gov/databases/alerts/2011_nhlbi_ifp.html)

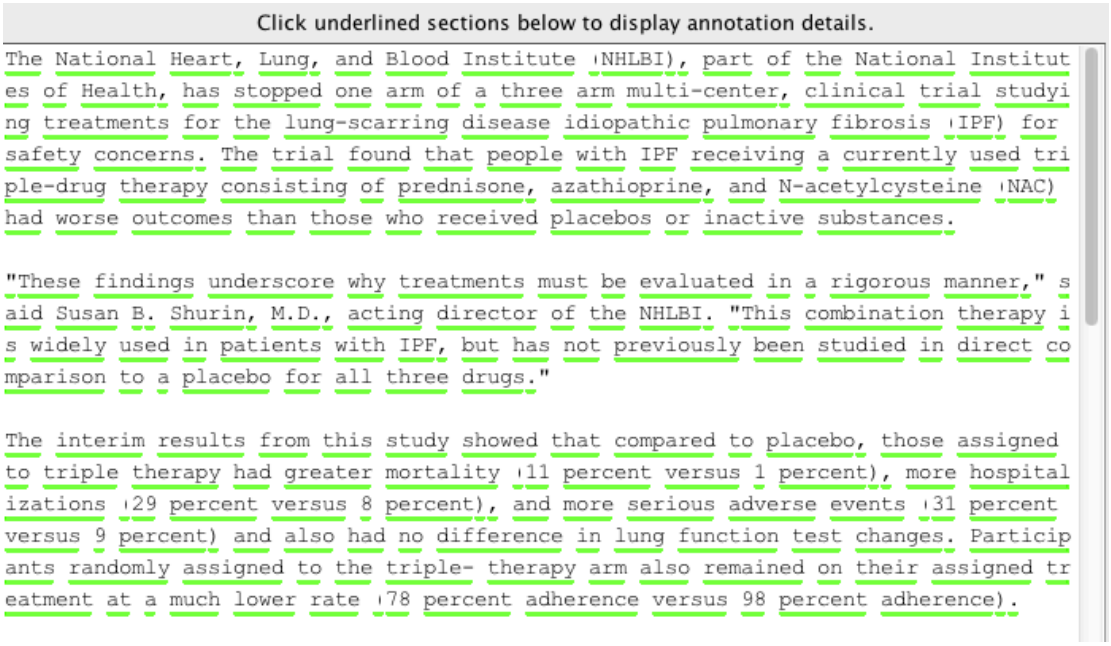


Figure 1: Output of the tokenisation functionality of the GENIA Sentence Detector in the U-Compare workbench

Running the tool on the 4 KB text on a single core machine with 8 GB RAM takes around 2.4 seconds.

ADMINISTRATIVE INFORMATION

Contact

For further information, please contact Sophia Ananiadou:
sophia.ananiadou@manchester.ac.uk

REFERENCES

Ferrucci, D., Lally, A., Gruhl, D., Epstein, E., Schor, M., Murdock, J. W., Frenkiel, A., Brown, E.W. , Hampp T., Doganata, Y., Welty, C., Amini, L., Kofman, G., Kozakov, L. and Mass, Y. (2006). Towards an Interoperability Standard for Text and Multi-Modal Analytics. IBM Research Report RC24122.

Kano, Y., Baumgartner Jr., W. A, McCrochon, L., Ananiadou, S., Cohen, K. B., Hunter, L. and Tsujii, J. (2009). U-Compare: share and compare text mining tools with UIMA. *Bioinfomatics*, 25(15), 1997-1998.

Kano, Y., Miwa, M., Cohen, K. B., Hunter, L., Ananiadou, S. and Tsujii, J.. (2011). U-Compare: a modular NLP workflow construction and evaluation system. *IBM Journal of Research and Development*, 55(3), 11:1 - 11:10.

Kim, J.-D, Ohta, T., Tateisi, Y. and Tsujii, J. (2003). GENIA corpus - a semantically annotated corpus for bio-textmining. Bioinformatics. 19(suppl. 1). pp. i180-i182, Oxford University Press, 2003.

Kulick S, Bies A, Liberman M, Mandel M, McDonald R, Palmer M, Schein A, Ungar L, Winters S, White P (2004). Integrated annotation for biomedical information extraction. Human Language Technology conference/North American Chapter of the Association for Computational Linguistics Annual Meeting, pp. 61-68.

Toutanova, K. and Klein, D. and Manning, C. D. and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of NAACL '03, pp 173- 180.

Tsuruoka, Y., Tateisi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S. and Tsujii, J.. (2005). Developing a Robust Part-of-Speech Tagger for Biomedical Text. *Advances in Informatics - 10th Panhellenic Conference on Informatics*, pp 382--392, Springer-Verlag

U-Compare OpenNLP PoSTagger

BASIC INFORMATION

Tool name

U-Compare OpenNLP POS Tagger

Overview and purpose of the tool

This is a UIMA[13] (Ferrucci et al., 2006) wrapper for the OpenNLP Tokenizer tool. It assigns part-of-speech tags to tokens in English text. The tagset used in from the Penn Treebank (Marcus et al., 1993).

The tool forms part of the in-built library of components provided with the U-Compare platform (Kano et al., 2009; Kano et al., 2011; see separate META-SHARE record)[14] for building and evaluating text mining workflows. The U-Compare Workbench (see separate META-SHARE record), which provides a graphical drag-and drop interface for the rapid creation of workflows.

A short description of the algorithm

OpenNLP tools[15] are trained using machine-learning methods. The tool provided uses the pre-trained tagging model for English, available on the OpenNLP SourceForge website: http://opennlp.sourceforge.net/models-1.5/

TECHNICAL INFORMATION

Software dependencies and system requirements

[13] http://uima.apache.org/
[14] http://nactem.ac.uk/ucompare/
[15] http://opennlp.apache.org/

In order to run U-Compare, Java 6 must be installed.

The UIMA component calls a web service. Hence, internet access is required.

Installation

There is no specific installation for U-Compare. The file UCLoader.class should be downloaded from http://u-compare.org/downloads/UCLoader.class

Execution instructions

U-Compare is started by running UCLoader.class from the command line. Since U-Compare can consume a large amount of memory, it is suggested to specify minimum and maximum memory usage when running U-Compare, as in the following example:

java –jar -Xms700m –Xmx 1000m UCLoader

The memory usage can be adjusted, but note that a minimum memory usage of 256 MB is recommended. Please also note that when U-compare is first started for the first, a large number of files will be downloaded, and so it will take some time to start. Subsequent launches will be quicker.

Once U-Compare has been started, the sentence detector tool can be executed through inclusion in workflow. This can be done simply by dragging and dropping it onto the workflow canvas using the graphical user interface of the U-Compare workbench. See the META-SHARE record "U-Compare Workbench" for more details

Input/Output data formats

Input data formats

The tool requires as input text that has been split into sentences and tokenised. Thus, the UIMA Common Analysis Structure (CAS) must contain both sentence and token annotations before this component is run. In a UIMA workflow, this could be achieved either by executing component(s) that perform sentence splitting and tokenisation prior to this component, or otherwise reading in a corpus of documents that already contains sentence and token annotations.

Output data format

The purpose of the tool is to detect tokens in the text. An annotation of type "POSToken" is thus added to the CAS corresponding to each token in a document. This type of annotation has a "posString" attribute to store the part-of-speech of the token. Different CAS consumers (such as those provided in U-Compare) can be used to write the contents of the CAS to a file or database format.

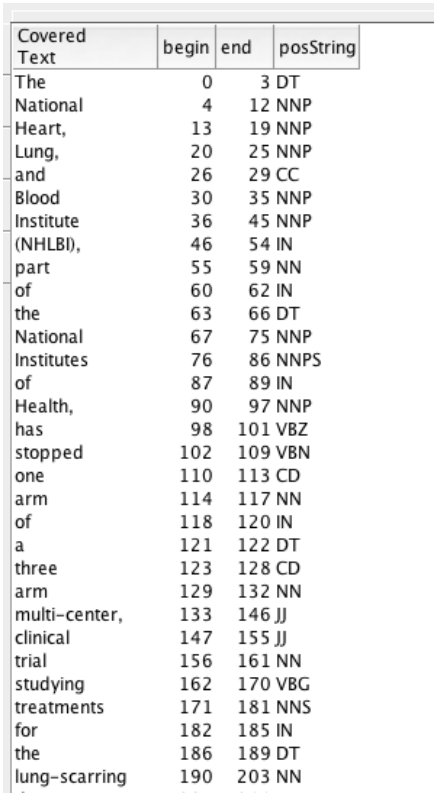
Integration with external tools

The tool can be run as part of a UIMA workflow, either using U-Compare or otherwise. For instructions of how to include components in UIMA workflows outside of U-Compare, see:

http://nactem.ac.uk/ucompare/developerguide/Using_U_Compare_Components_.html

CONTENT INFORMATION

Figure 1 shows the part of the output of the tool that is produced in in the U-Compare workbench. The attributes of POSToken annotations are shown, consisting the of the beginning and end offsets of the token in the text, and the POS tag assigned to it. The sample text is taken the US National Library of Medicine website (http://www.nlm.nih.gov/databases/alerts/2011_nhlbi_ifp.html)



Covered Text	begin	end	posString
The	0	3	DT
National	4	12	NNP
Heart,	13	19	NNP
Lung,	20	25	NNP
and	26	29	CC
Blood	30	35	NNP
Institute	36	45	NNP
(NHLBI),	46	54	IN
part	55	59	NN
of	60	62	IN
the	63	66	DT
National	67	75	NNP
Institutes	76	86	NNPS
of	87	89	IN
Health,	90	97	NNP
has	98	101	VBZ
stopped	102	109	VBN
one	110	113	CD
arm	114	117	NN
of	118	120	IN
a	121	122	DT
three	123	128	CD
arm	129	132	NN
multi-center,	133	146	JJ
clinical	147	155	JJ
trial	156	161	NN
studying	162	170	VBG
treatments	171	181	NNS
for	182	185	IN
the	186	189	DT
lung-scarring	190	203	NN

Figure 1: Output of the U-Compare OpenNLP POS Tagger in the U-Compare workbench

Running the tool on the 4 KB text on a single core machine with 8 GB RAM takes around 3.4 seconds.

ADMINISTRATIVE INFORMATION

Contact

For further information, please contact Sophia Ananiadou:
sophia.ananiadou@manchester.ac.uk

Copyright statement and information on IPR

The OpenNLP Sentence Detector must be used in compliance with the Apache Licence: http://www.apache.org/licenses/

REFERENCES

Marcus, M.P., Marcinkiewicz, M.A. and Santorini, B. (1993), Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics* 19(2), pp 313—330.

Ferrucci, D., Lally, A., Gruhl, D., Epstein, E., Schor, M., Murdock, J. W., Frenkiel, A., Brown, E.W. , Hampp T., Doganata, Y., Welty, C., Amini, L., Kofman, G., Kozakov, L. and Mass, Y. (2006). Towards an Interoperability Standard for Text and Multi-Modal Analytics. IBM Research Report RC24122.

Kano, Y., Baumgartner Jr., W. A, McCrochon, L., Ananiadou, S., Cohen, K. B., Hunter, L. and Tsujii, J. (2009). U-Compare: share and compare text mining tools with UIMA. *Bioinfomatics*, 25(15), 1997-1998.

Kano, Y., Miwa, M., Cohen, K. B., Hunter, L., Ananiadou, S. and Tsujii, J.. (2011). U-Compare: a modular NLP workflow construction and evaluation system. *IBM Journal of Research and Development*, 55(3), 11:1 - 11:10.

U-Compare OpenNLP Sentence Detector

BASIC INFORMATION

Tool name

U-Compare Sentence Detector

Overview and purpose of the tool

This is a UIMA[16] (Ferrucci et al., 2006) wrapper for the OpenNLP Sentence Detector tool. It splits English text into individual sentences.

The tool forms part of the in-built library of components provided with the U-Compare platform (Kano et al., 2009; Kano et al., 2011; see separate META-SHARE record)[17] for building and evaluating text mining workflows. The U-Compare Workbench (see separate META-SHARE record), which provides a graphical drag-and drop interface for the rapid creation of workflows.

A short description of the algorithm

OpenNLP tools[18] are trained using machine-learning methods. The tool provided uses the pre-trained tagging model for English, available on the OpenNLP SourceForge website: http://opennlp.sourceforge.net/models-1.5/

TECHNICAL INFORMATION

Software dependencies and system requirements

In order to run U-Compare, Java 6 must be installed.

[16] http://uima.apache.org/
[17] http://nactem.ac.uk/ucompare/
[18] http://opennlp.apache.org/

The UIMA component calls a web service. Hence, internet access is required.

Installation

There is no specific installation for U-Compare. The file UCLoader.class should be downloaded from http://u-compare.org/downloads/UCLoader.class

Execution instructions

U-Compare is started by running UCLoader.class from the command line. Since U-Compare can consume a large amount of memory, it is suggested to specify minimum and maximum memory usage when running U-Compare, as in the following example:

java –jar -Xms700m –Xmx 1000m UCLoader

The memory usage can be adjusted, but note that a minimum memory usage of 256 MB is recommended. Please also note that when U-compare is first started for the first, a large number of files will be downloaded, and so it will take some time to start. Subsequent launches will be quicker.

Once U-Compare has been started, the sentence detector tool can be executed through inclusion in workflow. This can be done simply by dragging and dropping it onto the workflow canvas using the graphical user interface of the U-Compare workbench. See the META-SHARE record "U-Compare Workbench" for more details.

Input/Output data formats

Input data formats

The tool operates on plain, unannotated text. Thus, the UIMA Common Analysis Structure (CAS) should contain the text to be analysed prior to the tool being executed. In a UIMA workflow, this could be achieved by reading in a single text or corpus of text. For example, U-Compare provides collection readers that can read in text from an input box, or otherwise read a directory of texts.

Output data format

The purpose of the tool is to detect sentences in the text. An annotation is thus added to the CAS corresponding to each sentence in a document. Different CAS consumers (such as those provided in U-Compare) can be used to write the contents of the CAS to a file or database format.

Integration with external tools

The tool can be run as part of a UIMA workflow, either using U-Compare or otherwise. For instructions of how to include components in UIMA workflows outside of U-Compare, see:

http://nactem.ac.uk/ucompare/developerguide/Using_U_Compare_Components_.html

CONTENT INFORMATION

Figure 1 shows the part of the output of the tool that is produced in in the U-Compare workbench. Each sentence detected is separately underlined The sample text is taken the US National Library of Medicine website (http://www.nlm.nih.gov/databases/alerts/2011_nhlbi_ifp.html)

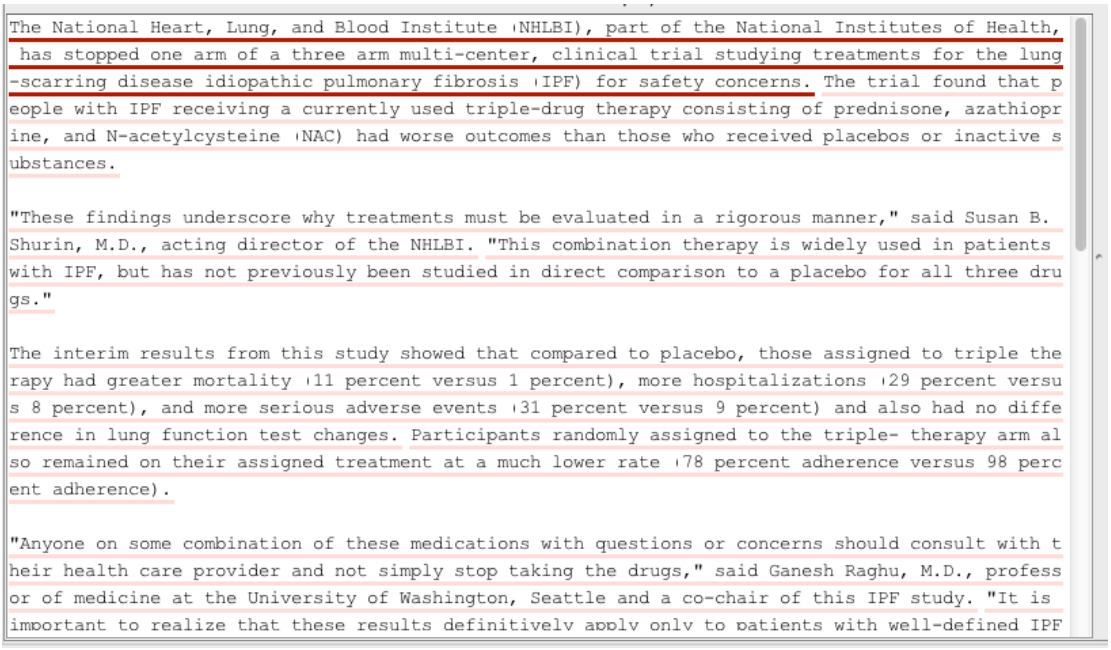


Figure 1: Output of the U-Compare OpenNLP Sentence Detector in the U-Compare workbench

Running the tool on the 4 KB text on a single core machine with 8 GB RAM takes around 0.7 seconds.

ADMINISTRATIVE INFORMATION

Contact

For further information, please contact Sophia Ananiadou:
sophia.ananiadou@manchester.ac.uk

Copyright statement and information on IPR

The OpenNLP Sentence Detector must be used in compliance with the Apache Licence: http://www.apache.org/licenses/

REFERENCES

Ferrucci, D., Lally, A., Gruhl, D., Epstein, E., Schor, M., Murdock, J. W., Frenkiel, A., Brown, E.W. , Hampp T., Doganata, Y., Welty, C., Amini, L., Kofman, G., Kozakov, L. and Mass, Y. (2006). Towards an Interoperability Standard for Text and Multi-Modal Analytics. IBM Research Report RC24122.

Kano, Y., Baumgartner Jr., W. A, McCrochon, L., Ananiadou, S., Cohen, K. B., Hunter, L. and Tsujii, J. (2009). U-Compare: share and compare text mining tools with UIMA. *Bioinfomatics*, 25(15), 1997-1998.

Kano, Y., Miwa, M., Cohen, K. B., Hunter, L., Ananiadou, S. and Tsujii, J.. (2011). U-Compare: a modular NLP workflow construction and evaluation system. *IBM Journal of Research and Development*, 55(3), 11:1 - 11:10.

U-Compare OpenNLP Tokenizer

BASIC INFORMATION

Tool name

U-Compare OpenNLP Tokenizer

Overview and purpose of the tool

This is a UIMA[19] (Ferrucci et al., 2006) wrapper for the OpenNLP Tokenizer tool. It splits English sentences into individual tokens.

The tool forms part of the in-built library of components provided with the U-Compare platform (Kano et al., 2009; Kano et al., 2011; see separate META-SHARE record)[20] for building and evaluating text mining workflows. The U-Compare Workbench (see separate META-SHARE record), which provides a graphical drag-and drop interface for the rapid creation of workflows.

A short description of the algorithm

OpenNLP tools[21] are trained using machine-learning methods. The tool provided uses the pre-trained model for English, available on the OpenNLP SourceForge website: http://opennlp.sourceforge.net/models-1.5/

TECHNICAL INFORMATION

Software dependencies and system requirements

In order to run U-Compare, Java 6 must be installed.

The UIMA component calls a web service. Hence, internet access is required.

Installation

There is no specific installation for U-Compare. The file UCLoader.class should be downloaded from http://u-compare.org/downloads/UCLoader.class

Execution instructions

[19] http://uima.apache.org/
[20] http://nactem.ac.uk/ucompare/
[21] http://opennlp.apache.org/

U-Compare is started by running UCLoader.class from the command line. Since U-Compare can consume a large amount of memory, it is suggested to specify minimum and maximum memory usage when running U-Compare, as in the following example:

java –jar -Xms700m –Xmx 1000m UCLoader

The memory usage can be adjusted, but note that a minimum memory usage of 256 MB is recommended. Please also note that when U-compare is first started for the first, a large number of files will be downloaded, and so it will take some time to start. Subsequent launches will be quicker.

Once U-Compare has been started, the sentence detector tool can be executed through inclusion in workflow. This can be done simply by dragging and dropping it onto the workflow canvas using the graphical user interface of the U-Compare workbench. See the META-SHARE record "U-Compare Workbench" for more details

Input/Output data formats

Input data formats

The tool requires sentence split text as input. Thus, the UIMA Common Analysis Structure (CAS) must contain sentence annotations before this component is run. In a UIMA workflow, this could be achieved either by executing a component that performs sentence splitting prior to this component, or otherwise reading in a corpus of documents that already contains sentence annotations.

Output data format

Since the purpose of the tool is to detect tokens in the text, the result of running the tool is that an annotation corresponding to each token in the text is thus added to the CAS. Different CAS consumers (such as those provided in U-Compare) can be used to write the contents of the CAS to a file or database format.

Integration with external tools

The tool can be run as part of a UIMA workflow, either using U-Compare or otherwise. For instructions of how to include components in UIMA workflows outside of U-Compare, see:

http://nactem.ac.uk/ucompare/developerguide/Using_U_Compare_Components_.html

CONTENT INFORMATION

Figure 1 shows the output of the tool in the U-Compare workbench. Each token recognised is separately underlined. The sample text is taken the US National Library of Medicine website (http://www.nlm.nih.gov/databases/alerts/2011_nhlbi_ifp.html)

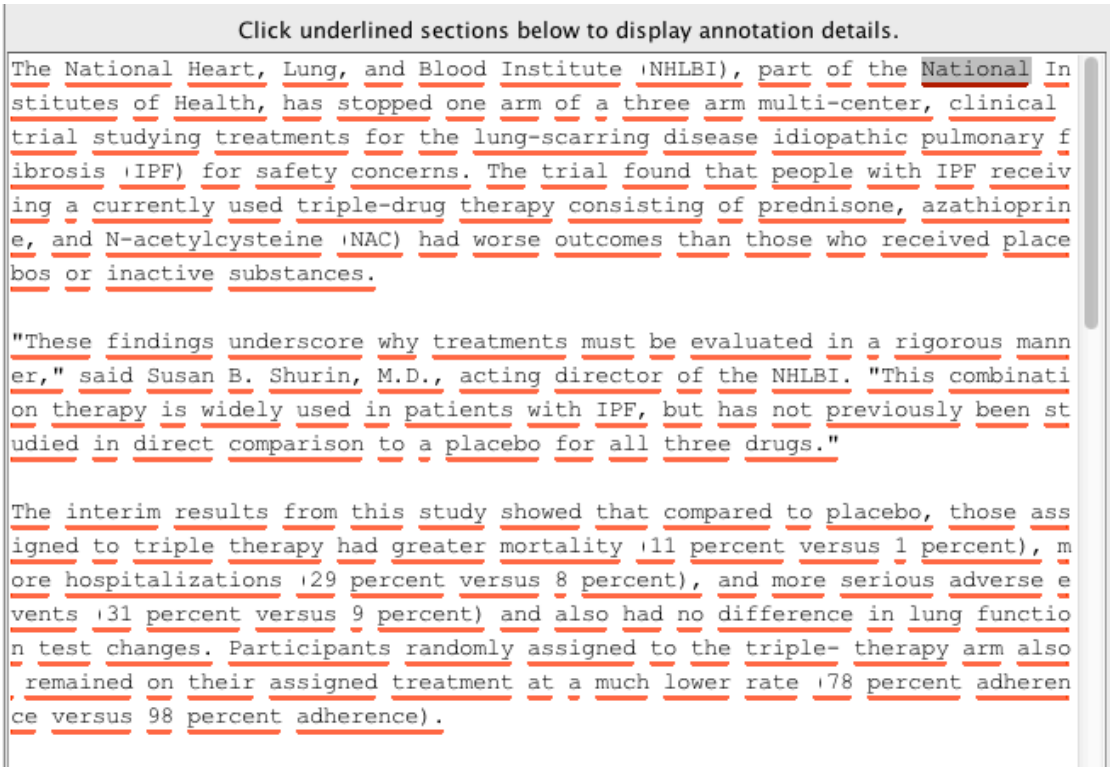


Figure 1: Output of the U-Compare OpenNLP Tokenizer in the U-Compare workbench

Running the tool on the 4 KB text on a single core machine with 8 GB RAM takes around 2.7 seconds.

ADMINISTRATIVE INFORMATION

Contact

For further information, please contact Sophia Ananiadou:
sophia.ananiadou@manchester.ac.uk

Copyright statement and information on IPR

The OpenNLP Sentence Detector must be used in compliance with the Apache Licence: http://www.apache.org/licenses/

REFERENCES

Ferrucci, D., Lally, A., Gruhl, D., Epstein, E., Schor, M., Murdock, J. W., Frenkiel, A., Brown, E.W. , Hampp T., Doganata, Y., Welty, C., Amini, L., Kofman, G., Kozakov, L. and Mass, Y. (2006). Towards an Interoperability Standard for Text and Multi-Modal Analytics. IBM Research Report RC24122.

Kano, Y., Baumgartner Jr., W. A, McCrochon, L., Ananiadou, S., Cohen, K. B., Hunter, L. and Tsujii, J. (2009). U-Compare: share and compare text mining tools with UIMA. *Bioinfomatics*, 25(15), 1997-1998.

Kano, Y., Miwa, M., Cohen, K. B., Hunter, L., Ananiadou, S. and Tsujii, J.. (2011). U-Compare: a modular NLP workflow construction and evaluation system. *IBM Journal of Research and Development*, 55(3), 11:1 - 11:10.

U-Compare Type System

1 BASIC INFORMATION

1.1 Resource composition

The resource constitutes of a hierarchically-structured system of data types, which is intended to be suitable for describing the inputs and output annotation types of a wide range of natural language processing applications which operate within the UIMA Framework[22] (Ferrucci et al, 2006). It is being developed in conjunction with the U-Compare Workbench, but can be used as the base type system for other UIMA components and workflows, to help to ensure greater interoperability.

1.2 Representation of the resource (flat files, database, markup)

The resource is provided as java archive (jar file), UCompareTypeSystem.jar

1.3 Character encoding

The characters are UTF8 encoded.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Sophia Ananiadou
Address: Manchester Interdisciplinary Biocentre,131 Princess Street, Manchester M1 7DN, IK
Affiliation: National Centre for Text Mining, School of Computer Science, University of Manchester
Position: Director
Telephone: +44 161 306 3092
Fax: +44 161 306 5201
e-mail: Sophia.Ananiadou@manchester.ac.uk

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource is available on the MetaShare platform.

2.3 Copyright statement and information on IPR

The resource is available for research purposes under the LGPL licence.

[22] http://uima.apache.org/

3 TECHNICAL INFORMATION

3.1 Directories and files

The jar file contains XML descriptor files corresponding to the core U-Compare type system, as well as various extensions that have been created to accommodate various components in the U-Compare library. Also included in the archive are automatically generated Java source and class files, corresponding to each annotation type in the system. These are required by the UIMA framework, since each annotation in the Common Analysis Structure (CAS; the common data structure that is used to store the results produced by each component in a UIMA workflow) corresponds to a Java object.

3.2 Data structure of an entry

The file follows the required format of UIMA type system descriptor files[23], which are XML files. The following is an example of the description of an individual type.

```
<typeDescription>
 <name>org.u_compare.shared.syntactic.POSToken</name>
 <description/>
    <supertypeName>org.u_compare.shared.syntactic.Token</supertypeName>
  <features>
    <featureDescription>
        <name>pos</name>
                    <description/>                    <rangeTypeName>
                org.u_compare.shared.label.POS
         </rangeTypeName>
    </featureDescription>
    <featureDescription>
          <name>posString</name>
          <description>A special field which internally converts object of pos field into String
    value.
</description>
<rangeTypeName>uima.cas.String
                    </rangeTypeName>
      </featureDescription>
   </features>
  </typeDescription>
```

The tags and attributes are as follows:

- *typeDescription* – contains the description of the type
    - *name* – the name (label) assigned to the annotation
    - *description* – a textual description of the type
    - *supertypeName* – types in the type system are hierarchically structured. This element contains the name assigned to the supertype of the current type.
    - *Features* – contains descriptions of the features (attributes) associated with the current annotation type.

[23] See http://uima.apache.org/downloads/releaseDocs/2.1.0-incubating/docs/html/tutorials_and_users_guides/tutorials_and_users_guides.html#ugr.tug.aae.defining_types

- *featureDescription* – Contains the description of a single feature/attribure
  - *name* – the name of the feaure
  - *description* – a textual description of the feature
  - *rangeType* – the type of the the value of the feature

3.3 Resource size (nmb. of tokens, MB occupied on disk)

The U-Compare type system consists of a hierarchy of 281 different types. The size of the file is 89 KB.

4 CONTENT INFORMATION

4.1 The natural language(s) of the resource

The type system was developed mainy based on the inputs/ouitputs of English tools, although recent work on developing UIMA components for other languages suggests that the current type system is largely suitable for other languages, at least European ones.

4. 2 Entry Type

The data types described by the U-Compare type system can be roughtly split into three different group, i.e., syntactic types, semantic types and document types (i.e, describing the structural aspects of a document).

4.3 Attributes and their values

As mentioned in the section 3.2, each type in the type system may have zero or more attributes associated it, to store additional information about the annotations.

4.4 Coverage of the resource

Currently, the U-Compare type system is only suitable for text-based UIMA components, although extensions are planned for speech-based applications. Figures 1, 2 and 3 provide an overview of the main data types covered by the U-Compare type system, and the hierarchical structure of these types.

4.5 Intended application of the resource

In UIMA, a common data structure called the Common Analysis Structure (CAS) is used to store the outputs of each component in a workflow, in the form of annotations. Each component obtains its input by reading relevant annotations from the CAS, and produces output by creating nre annotations in the CAS, or updating existing annotations. The UIMA framework itself does not attempt to place any restrictions or recommendations regarding the use of a particular system of annotation types. However, some level of commonality of the type systems used by different components is required to try to achieve maximum interoperability and flexibility in the ways in which different components can be combined. For example, if a named entity recogniser requires the input types *Token* and *Chunk*, then it is only possible to use components that produce annotations with these names earlier in the workflow.

Different NLP research groups have produced different repositories of UIMA components, e.g., the BIONLP UIMA Component Repository (Baumgartner et al., 2008) , the CMU UIMA component repository[24] and the UIMA-fr consortium (Hernandez et al., 2010), but generally using their own type systems. This can cause problems for interoperability - components developed by one team cannot be combined easily with components developed by another team, because they use different type systems.

Ideally, to achieve maximum interoperability, a single, common type system would be imposed, to be followed by all developers of NLP UIMA components. However, this is considered not a viable option, as it would be difficult to achieve consensus on exactly which types should be present, given, for example, the various different syntactic and semantic theories on which different tools are based.

The U-Compare type system is a *sharable* type system, which aims to cover the most common types of annotation, both syntactic and semantic, that are produced by NLP applications. The idea is that all components in a UIMA workflow should produce annotations that are compatible with this type system. As the U-Compare type system consists of fairly general types, it is permissible to create new types that correspond to more specialised types of annotations, as long as these new types can form sub-types of one of the existing U-Compare types. This ensures that compatibility between components developed by different groups can at least be achieved at an intermediate level of the hierarchy.

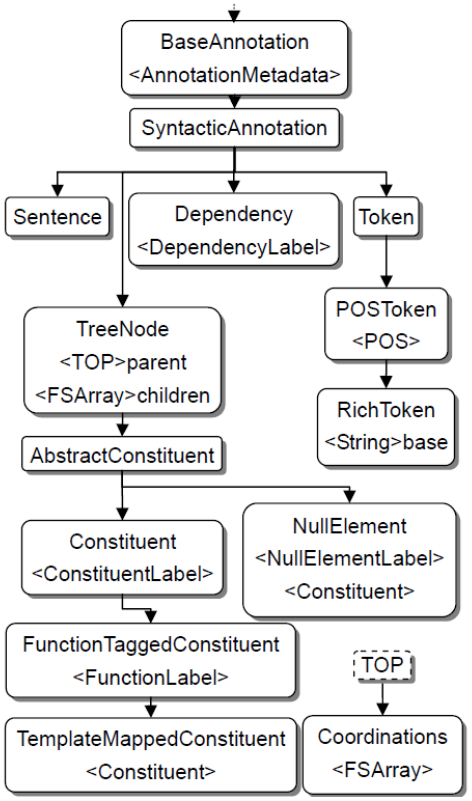


Figure 1: Main syntactic types in the U-Compare type system

[24] http://uima.lti.cs.cmu.edu

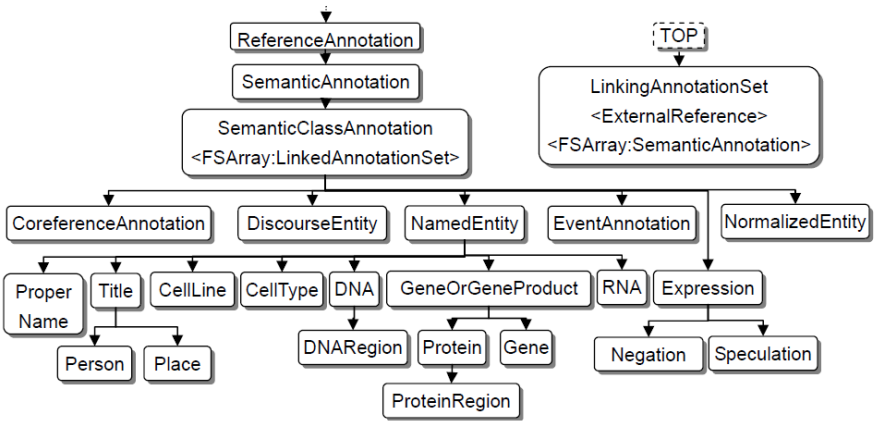


Figure 2: Main semantic types in the U-Compare type system

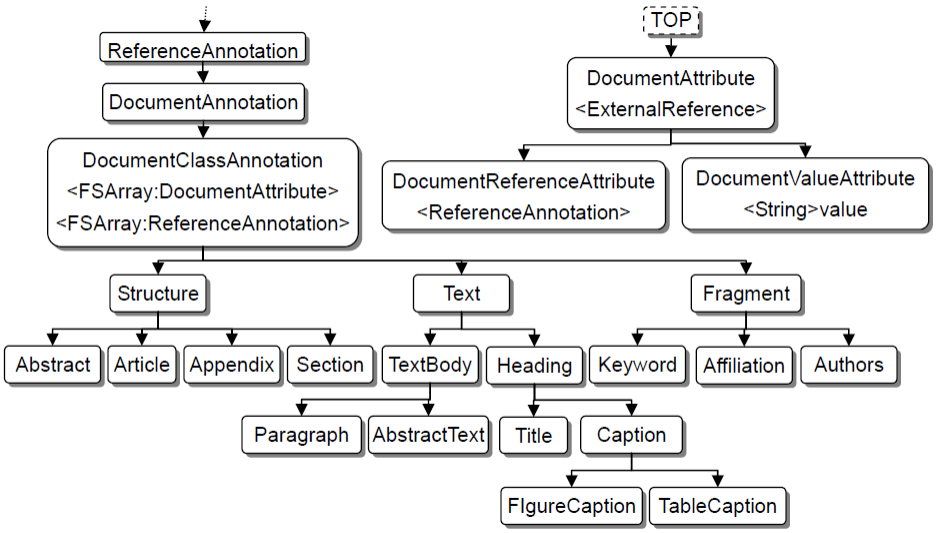


Figure 3: Main document-level types in the U-Compare type system

4.6 Reliability (automatically/manually constructed)

The U-Compare type system has been manually constructed, by considering the different input and output types of a wide range of NLP and text mining tools, and is still evolving. The success of the U-Compare type system in facilitating the construction of interoperable UIMA components can be demonstrated by the fact around 60 components that comply with the U-Compare type system, and which cover a number of different European languages, are now available in the component library of the U-Compare Workbench. This library is being extended as part of the META-NET initiative (Ananiadou et al., 2011; Thompson et al. 2011)

5 RELEVANT REFERENCES AND OTHER INFORMATION

Ananiadou, S., Thompson, P., Kano, Y., McNaught, J., Attwood, T. K., Day, P. J. R., Keane, J., Jackson, D. and Pettifer, S.. (2011). “Towards Interoperability of European Language Resources”. *Ariadne*, 67

Baumgartner, W. A., Cohen, K. B., & Hunter, L. (2008). "An open-source framework for large-scale, flexible evaluation of biomedical text mining systems". *Journal of Biomedical Discovery and Collaboration, 3*, 1.

Ferrucci, D., Lally, A., Gruhl, D., Epstein, E., Schor, M., Murdock, J. W., Frenkiel, A., Brown, E.W. , Hampp T., Doganata, Y., Welty, C., Amini, L., Kofman, G., Kozakov, L. and Mass, Y. (2006). Towards an Interoperability Standard for Text and Multi-Modal Analytics. IBM Research Report RC24122.

Hernandez, N., Poulard, F., Vernier, M., & Rocheteau, J. (2010). "Building a French-speaking community around UIMA, gathering research, education and industrial partners, mainly in Natural Language Processing and Speech Recognizing domains". In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp 41-45

Kano, Y., Baumgartner, W. A., Jr., McCrohon, L., Ananiadou, S., Cohen, K. B., Hunter, L. (2009). "U-Compare: share and compare text mining tools with UIMA". *Bioinformatics,* vol. 25, no. 15, 1997-1998.

Kano, Y., Miwa, M., Cohen, K. B., Hunter, L. E., Ananiadou, S., & Tsujii, J. (2011). "U-Compare: A modular NLP workflow construction and evaluation system". *IBM Journal of Research and Development, 55*(3), 11:11-11:10.

Thompson, P., Kano, Y., McNaught, J., Pettifer, S., Attwood, T. K., Keane, J. and Ananiadou, S.. (2011). Promoting Interoperability of Resources in META-SHARE. In *Proceedings of the IJCNLP Workshop on Language Resources, Technology and Services in the Sharing Paradigm (LRTS)*, pp. 50-58

Unrestricted exogenous resources (tools)

Apertium Morphological Analyser

BASIC INFORMATION

Tool name

UIMA/U-Compare Apertium Morphological Analyser

Overview and purpose of the tool

This tool performs tokenization of text and assigns all possible morphological analyses to each token. These analyses include the base form of the token, part-of-speech, information about number and gender. The morphological analyser is a module of Apertium machine translation system[25] (Armentano-Ollet et al., 2006). The provided tool can currently operate on a subset of the languages that are supported by the Apertium system, namely: English, Spanish, Calatan, Galician, Portuguese, Romanian and Basque.

The tool is provided as a UIMA[26] (Ferrucci et al., 2006) component, specifically as Java archive (jar) file, which can be incorporated within any UIMA workflow. However, it is particularly designed use in the U-Compare text mining plaform (Kano et al., 2009; Kano et al., 2011; see separate META-SHARE record), since the types of annotations it produces are compliant with the U-Compare.

A short description of the algorithm

The morphological analysis is carried using finite-state transducers, in conjunction with morphological dictionaries. See the Apertium documentation (http://wiki.apertium.org/wiki/Documentation) for more information.

TECHNICAL INFORMATION

Software dependencies and system requirements

The tool can most easily be run using the U-Compare platform (see separate META-SHARE record). The only requirement to run U-Compare is for Java 6 to be installed.

To run the tool as a UIMA component independently of U-Compare, Apache UIMA must be installed (see http://uima.apache.org/).

Installation

In order to run the Apertium Morphological analyser in U-Compare, it must be imported into U-Compare. Information about downloading and running U-Compare can be found in the documentation associated with the META-SHARE record for the U-Compare Workbench or the U-Compare website: http://nactem.ac.uk/ucompare/.

Importing the UIMA component (provided as the file ApertiumMorpho.jar) is carried out in U-Compare as follows:

1) From the “Library” menu in the U-Compare Workbench, choose the item “Register External Components (Edit Classpath)”. This will cause a window to appear that allows external components to be managed. It is shown in Figure 1.

[25] http://www.apertium.org/
[26] http://uima.apache.org/

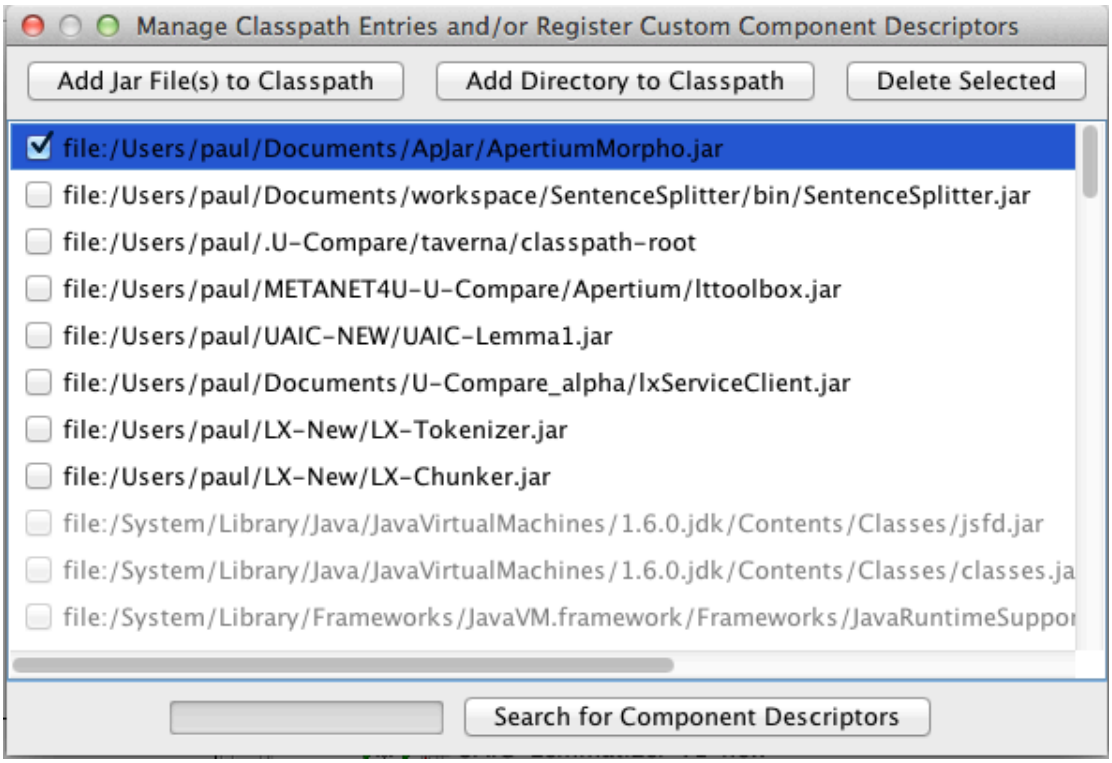


Figure 1: External component management window

2) Click on the button "Add Jar File(s) to Classpath", and browse to the location where the file "ApertiumMorpho.jar" has been saved, and click on "Open". The will cause the file to be displayed in the list of files in the external component manager window.

3) Ensure that box next to the file "ApertiumMorpho.jar" is checked in the external component manager window. Then, click the "Search for Component Descriptors" button. A "Component Descriptors Search Results" window will appear (Figure 2).

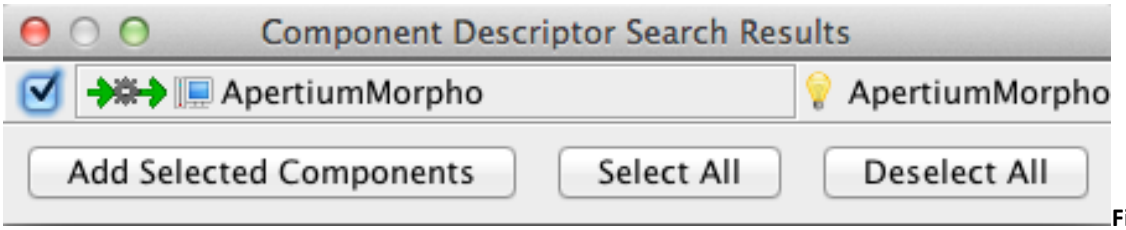


Figure 2: Component Descriptor Search Results window

4) Check the box next to "ApertiumMorpho", and click on the "Add Selected Components" button. The name of the component will then appear in the library on components on the right hand side of the main U-Compare workbench window, under "Custom components", and it can then be used in workflows.

Execution instructions

Within U-Compare, the tool can be executed through inclusion in workflow. This can be done simply by dragging and dropping it onto the workflow canvas using the graphical user interface of the U-Compare workbench. See the META-SHARE record “U-Compare Workbench” for more details. Alternatively, it can be incorporated into other UIMA-based workflows, by following the documentation on the Apache UIMA site: http://uima.apache.org/

The component must be configured before use, to tell the component the languages in which the text(s) to be morphologically analysed are written. This is set by specifying a value for the “languagePair” parameter. Since this tool is a module of a machine translation system, the language data is stored in pairs, i.e., the source language and the target language. The value of the “languagePair” parameter consists of two-letter codes for the languages, joined with a hyphen, where the first language is the source language, and the second is the target language, e.g. “pt-es” is used when Portuguese is the source language and Spanish is the target language.

If this morphological analyser is run in a workflow without the translation module, only the source language is relevant, but the a complete language pair string must still be specified as the value of the “languagePair” parameter, where the language to be analysed appears first in the pair string.

Possible values of the languagePair attribute that can currently be used are as follows: "en-es", "es-en", "gl-es", "es-gl", "es-pt", "pt-es", "es-ca", "ca-es" "ro-es", "es-ro" and "eu-es". If a non-valid value is entered, then the language pair will default to "en-es", i.e,. morphological analysis for English will be carried out

When being run within U-Compare, the value of the “languagePair” parameter can be set by clicking on the  icon. This will cause a parameter configuration window to appear, allowing the user to enter the appropriate language pair string. A part of this window is shown in Figure 3. The value entered means that the component will be configured to carry out morphological analysis for Spanish.

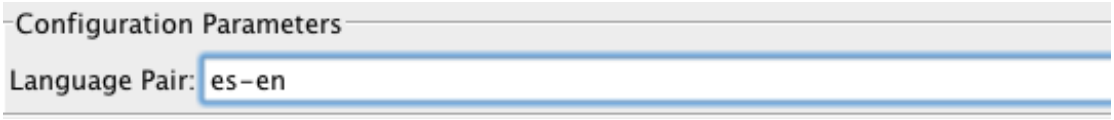


Figure 3: Configuration of LanguagePair parameter

Input/Output data formats

Input data formats

The tool operates on plain, unannotated text. Thus, the UIMA Common Analysis Structure (CAS) should contain the text to be analysed prior to the tool being executed. In a UIMA workflow, this could be achieved by reading in a single text or corpus of text. For example, U-Compare provides collection readers that can read in text from an input box, or otherwise read a directory of texts.

Output data format

An annotation is thus added to the CAS corresponding to each token in a document, with the type “ApertiumToken”. This provides a “morphology” field, in which the possible morphological analyses generated for each token are stored. An example of one of these morphological analyses (for Spanish) is as follows:

mide/medir<vblex><pri><p3><sg>/medir<vblex><imp><p2><sg>

The line begins with the surface form of the token as it appears in the text being analysed. Each possible morphological analysis is separated by a forward slash. The first item on each analysis is the base form, followed by different tags providing morphological information, each enclosed in angled brackets. The first of these is a part-of-speech tag. The other tags will vary according to the part of speech, but in the above example, consist of tense, person and number.

Different CAS consumers (such as those provided in U-Compare) can be used to write the contents of the CAS to a file or database format.

Integration with external tools

As mentioned above, the tool can only be run within U-Compare framework, or more generally, within the UIMA framework.

CONTENT INFORMATION

Figure 4 shows the output of the tool in the U-Compare workbench. Each token recognised is separately underlined. The sample text is taken from the CNN Español site (http://www.cnnespanol.com).

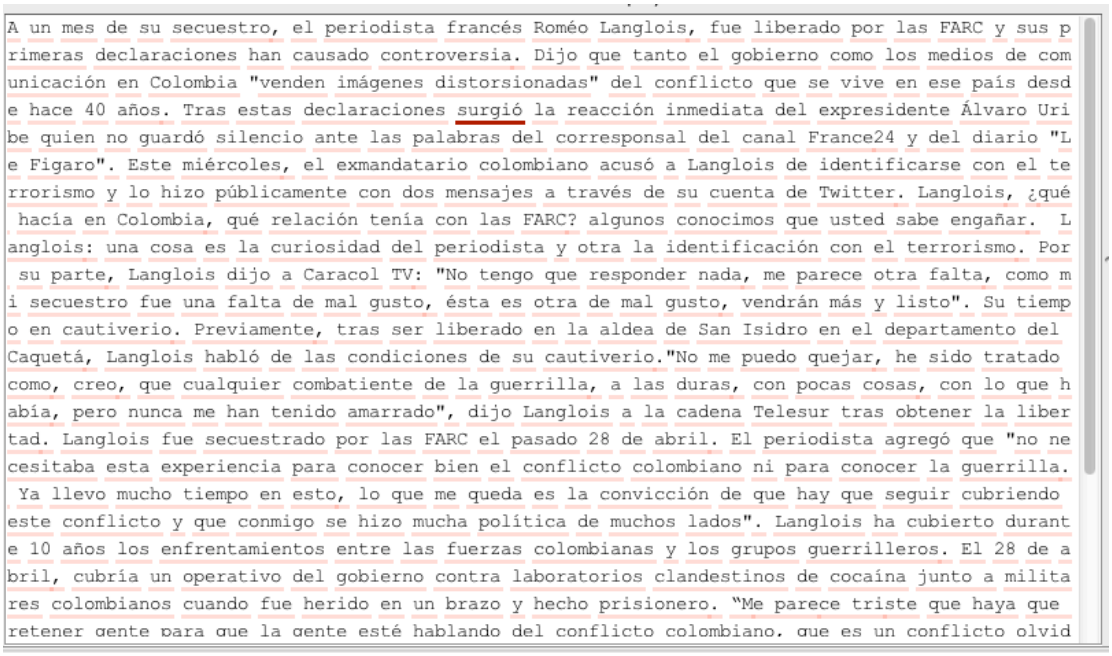


Figure 1: Output of the Apertium morphological analyser in the U-Compare workbench, showing the individual tokens identified

Figure 2 shows another part of analysis displayed in U-Compare, with the attributes of each annotation. These consist of the start and end offsets of each token, together with the different possible morphological analyses for each of them. Words that are unknown by the system are marked with a "*".

begin	end	posString	base	morphology
0	1			A/A<pr>
2	4			un/uno<det><ind><m><sg>
5	8			mes/mes<n><m><sg>
9	11			de/de<pr>
12	14			su/suyo<det><pos><mf><sg>
15	24			secuestro/secuestro<n><m><sg>/secuestrar<vblex><pri><p1><sg>
24	25			,/,<cm>
26	28			el/el<det><def><m><sg>
29	39			periodista/periodista<n><mf><sg>
40	47			francés/francés<n><m><sg>/francés<adj><m><sg>
48	53			Roméo/*Roméo
54	62			Langlois/*Langlois
62	63			,/,<cm>
64	67			fue/ir<vblex><ifi><p3><sg>/ser<vbser><ifi><p3><sg>
68	76			liberado/liberar<vblex><pp><m><sg>
77	80			por/por<pr>
81	84			las/el<det><def><f><pl>/prpers<prn><pro><p3><f><pl>
85	89			FARC/FARC<n><acr><f><pl>
90	91			y/y<cnjcoo>
92	95			sus/suyo<det><pos><mf><pl>
96	104			primeras/primero<adj><f><pl>/primer<det><ord><f><pl>
105	118			declaraciones/declaración<n><f><pl>
119	122			han/haber<vbhaver><pri><p3><pl>
123	130			causado/causar<vblex><pp><m><sg>
131	143			controversia/controversia<n><f><sg>
143	144			./.<sent>
145	149			Dijo/Decir<vblex><ifi><p3><sg>
150	153			que/que<cnjcoo>/que<cnjsub>/que<rel><an><mf><sp>

Figure 2: Attributes for the ApertiumToken annotations, displaying the beginning and end offsets of each token, plus the possible morphological anlyses

Running the tool on the 3 KB Spanish text shown in Figure 1 on a single core machine with 8 GB RAM takes around 0.25 seconds.

ADMINISTRATIVE INFORMATION

Contact

Contacts for the Apertium system can be found here: http://wiki.apertium.org/wiki/Contact

For further information regarding this UIMA wrapper for the Apertium morphological analyser module, please contact Sophia Ananiadou:

sophia.ananiadou@manchester.ac.uk

REFERENCES

Armentano-Oller, C., Carrasco, R., Corbí-Bellot, A.,Forcada, M., Ginestí-Rosell, M., Ortiz-Rojas, S. (2006). Open-source Portuguese–Spanish machine translation. Computational Processing of the Portuguese Language ,50-59

Ferrucci, D., Lally, A., Gruhl, D., Epstein, E., Schor, M., Murdock, J. W., Frenkiel, A., Brown, E.W. , Hampp T., Doganata, Y., Welty, C., Amini, L., Kofman, G., Kozakov, L. and Mass, Y. (2006). Towards an Interoperability Standard for Text and Multi-Modal Analytics. IBM Research Report RC24122.

Kano, Y., Baumgartner Jr., W. A, McCrochon, L., Ananiadou, S., Cohen, K. B., Hunter, L. and Tsujii, J. (2009). U-Compare: share and compare text mining tools with UIMA. *Bioinfomatics*, 25(15), 1997-1998.

Kano, Y., Miwa, M., Cohen, K. B., Hunter, L., Ananiadou, S. and Tsujii, J.. (2011). U-Compare: a modular NLP workflow construction and evaluation system. *IBM Journal of Research and Development*, 55(3), 11:1 - 11:10.

Apertium POS tagger

BASIC INFORMATION

Tool name

UIMA/U-Compare Apertium POS Tagger

Overview and purpose of the tool

This tool assigns a part-of-speech tag and base form to each token in a text. It operates on text that has previously been tokenised and morphologically analysed. The POS tagger is a module of Apertium machine translation system[27] (Armentano-Ollet et al., 2006). The provided tool can currently operate on a subset of the languages that are supported by the Apertium system, namely: English, Spanish, Calatan, Galician, Portuguese, Romanian and Basque.

NOTE: The morphological analysis required prior to running the POS tagger MUST be carried out by running the Apertium morphological analyser (which also performs tokeniaation).

The tool is provided as a UIMA[28] (Ferrucci et al., 2006) component, specifically as Java archive (jar) file, which can be incorporated within any UIMA workflow. However, it is particularly designed use in the U-Compare text mining plaform (Kano et al., 2009; Kano et al., 2011; see separate META-SHARE record), since the types of annotations it produces are compliant with the U-Compare.

The Apertium morphological analyser is also available as a UIMA component (called ApertiumMorpho).

A short description of the algorithm

The part-of-speech tagger determines the appropriate morphological analysis from amongst those produced by the morphological analyser model. The tagger is based on first-order hidden Markov models. The states of the Markov

[27] http://www.apertium.org/
[28] http://uima.apache.org/

model represent parts of speech, and the observable parameters are ambiguity classes formed by groups of parts of speech. For the purposes of part-of-speech tagging, the fine-grained tags produced by the morphological analyser are mapped to more coarse-grained categories. See the Apertium documentation (http://wiki.apertium.org/wiki/Documentation) for more information.

TECHNICAL INFORMATION

Software dependencies and system requirements

The tool can most easily be run using the U-Compare platform (see separate META-SHARE record). The only requirement to run U-Compare is for Java 6 to be installed.

To run the tool as a UIMA component independently of U-Compare, Apache UIMA must be installed (see http://uima.apache.org/).

Installation

In order to run the Apertium POS Tagger in U-Compare, it must be imported into U-Compare. Information about downloading and running U-Compare can be found in the documentation associated with the META-SHARE record for the U-Compare Workbench or the U-Compare website: http://nactem.ac.uk/ucompare/.

Importing the UIMA component (provided as the file ApertiumMorpho.jar) is carried out in U-Compare as follows:

1) From the "Library" menu in the U-Compare Workbench, choose the item "Register External Components (Edit Classpath)". This will cause a window to appear that allows external components to be managed. It is shown in Figure 1.

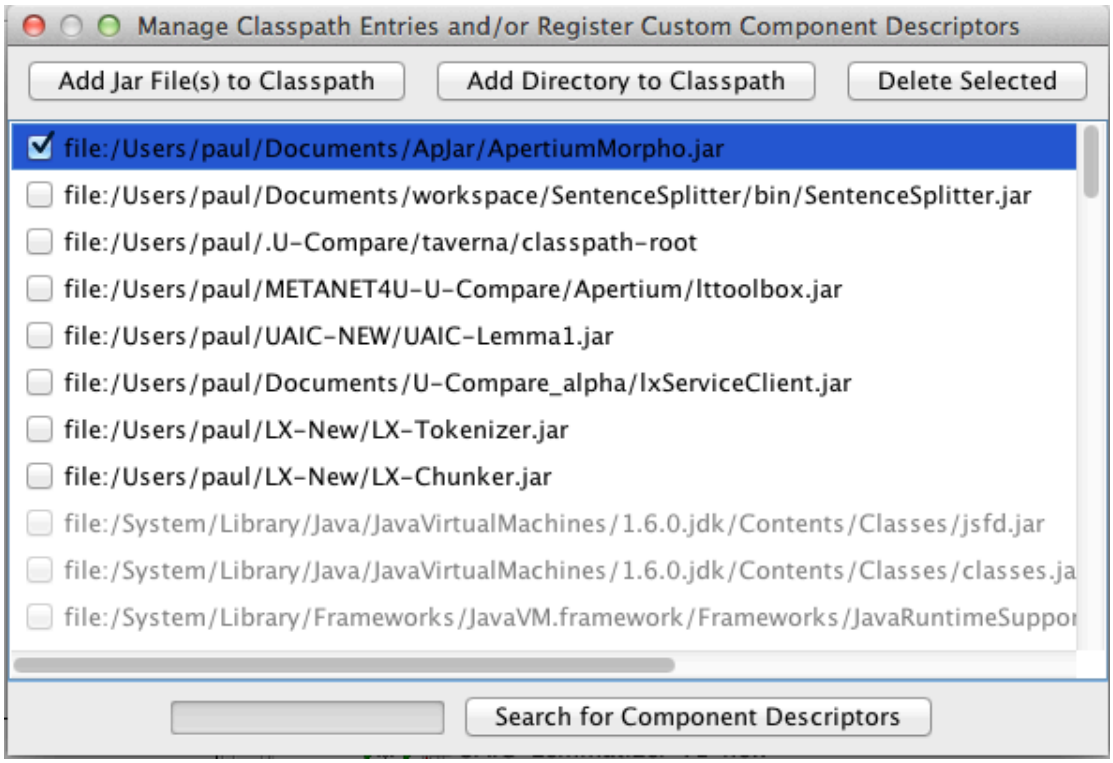


Figure 1: External component management window

2) Click on the button "Add Jar File(s) to Classpath", and browse to the location where the file "ApertiumPOS.jar" has been saved, and click on "Open". The will cause the file to be displayed in the list of files in the external component manager window.

3) Ensure that box next to the file "ApertiumPOS.jar" is checked in the external component manager window. Then, click the "Search for Component Descriptors" button. A "Component Descriptors Search Results" window will appear (Figure 2).

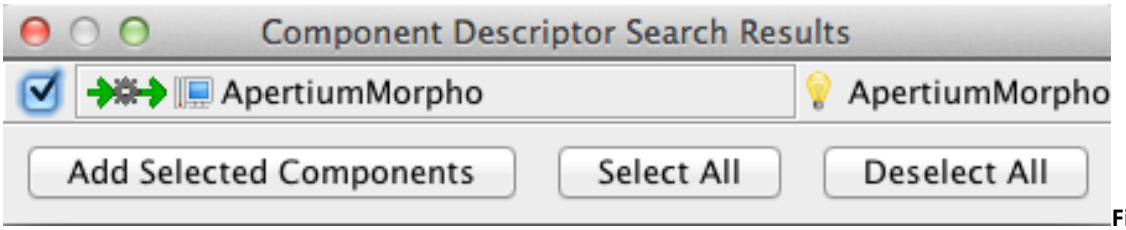


Figure 2: Component Descriptor Search Results window

4) Check the box next to "ApertiumPOS", and click on the "Add Selected Components" button. The name of the component will then appear in the library on components on the right hand side of the main U-Compare workbench window, under "Custom components", and it can then be used in workflows.

Execution instructions

Within U-Compare, the tool can be executed through inclusion in workflow. This can be done simply by dragging and dropping it onto the workflow canvas using the graphical user interface of the U-Compare workbench. See the META-SHARE record "U-Compare Workbench" for more details. Alternatively, it can be incorporated into other UIMA-based workflows, by following the documentation on the Apache UIMA site: http://uima.apache.org/

The component must be configured before use, to tell the component the languages in which the text(s) to be morphologically analysed are written. This is set by specifying a value for the "languagePair" parameter. Since this tool is a module of a machine translation system, the language data is stored in pairs, i.e., the source language and the target language. The value of the "languagePair" parameter consists of two-letter codes for the languages, joined with a hyphen, where the first language is the source language, and the second is the target language, e.g. "pt-es" is used when Portuguese is the source language and Spanish is the target language.

If this POS Tagger is run in a workflow without the translation module, only the source language is relevant, but the a complete language pair string must still be specified as the value of the "languagePair" parameter, where the language to be analysed appears first in the pair string.

Possible values of the languagePair attribute that can currently be used are as follows: "en-es", "es-en", "gl-es", "es-gl", "es-pt", "pt-es", "es-ca", "ca-es" "ro-es", "es-ro" and "eu-es". If a non-valid value is entered, then the language pair will default to "en-es", i.e., POS tagging for English will be carried out

When being run within U-Compare, the value of the "languagePair" parameter can be set by clicking on the  icon. This will cause a parameter configuration window to appear, allowing the user to enter the appropriate language pair string. A part of this window is shown in Figure 3. The value entered means that the component will be configured to carry out morphological analysis for Spanish.

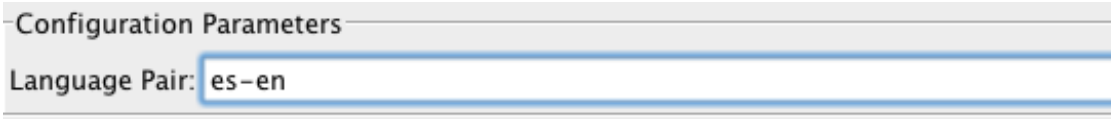


Figure 3: Configuration of LanguagePair parameter

Input/Output data formats

Input data formats

The tool operates on text that has previously been tokenised and morphologically analysed. The type of morphological analysis expected is the one produced by the Apertium morphological analyser module. Hence, the UIMA component corresponding to the Apertium morphological analyser (ApertiumMorpho) MUST be run prior to this POS tagger. This will ensure that annotations of type "ApertiumToken" will be added to the UIMA Common Analysis Structure (CAS), which are required as input to the POS tagger.

Output data format

The result of running the Apertium POS tagger is that the "morphology" attribute of the annotation is updated to contain only the morphological analysis selected by the tagger (thee morphological analyser may have produced several possible morphological analyses). The part-of-speech and base form of the selected morphological analysis are extracted and set as the values of the "posString" and "base" attributes of the annotation, respectively.

Integration with external tools

As mentioned above, the tool can only be run within U-Compare framework, or more generally, within the UIMA framework.

CONTENT INFORMATION

Figure 4 shows the output of the tool in the U-Compare workbench. Each token (corresponding to an "ApertiumToken" annotation) is separately underlined. The sample text is taken from the CNN Español site (http://www.cnnespanol.com).

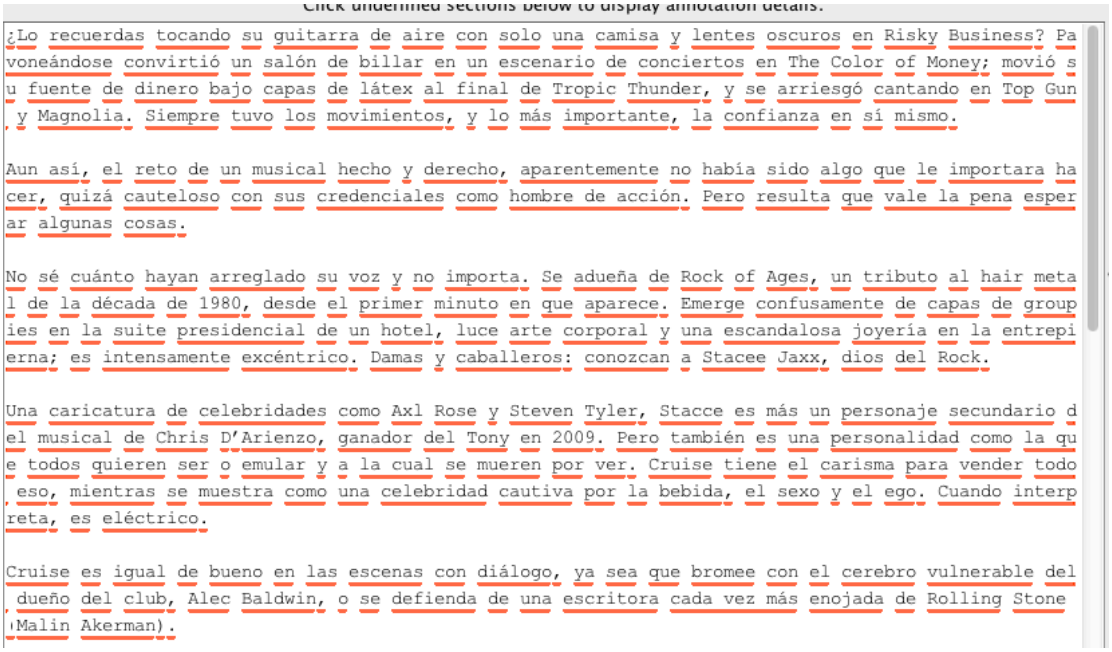


Figure 4: Output of a workflow in which the Apertium morphological analyser is run prior to the Apertium POS tagger, in the U-Compare workbench.

Figure 5 shows another part of analysis displayed in U-Compare, with the attributes of each annotation. These consist of the start and end offsets of each token, together with the part-of-speech tag assigned and the appropriate base form of the word.

Covered Text	begin	end	pos posType	posString	base
¿	0	1	lquest	lquest	¿
Lo	1	3	prn	prn	Lo
recuerdas	4	13	vblex	vblex	recordar
tocando	14	21	vblex	vblex	tocar
su	22	24	det	det	suyo
guitarra	25	33	n	n	guitarra
de	34	36	pr	pr	de
aire	37	41	n	n	aire
con	42	45	pr	pr	con
solo	46	50	adv	adv	solo
una	51	54	det	det	uno
camisa	55	61	n	n	camisa
y	62	63	cnjcoo	cnjcoo	y
lentes	64	70	n	n	lente
oscuros	71	78	adj	adj	oscuro
en	79	81	pr	pr	en
Risky	82	87			
Business	88	96			
?	96	97	sent	sent	?
Pavoneándose	98	110			
convirtió	111	120	vblex	vblex	convertir
un	121	123	det	det	uno
salón	124	129	n	n	salón
de	130	132	pr	pr	de
billar	133	139	n	n	billar
en	140	142	pr	pr	en
un	143	145	det	det	uno
escenario	146	155	n	n	escenario
de	156	158	pr	pr	de
conciertos	159	169	n	n	concierto
en	170	172	pr	pr	en

Figure 5: Attributes for the RichToken annotations, displaying the beginning and end offsets of each token, plus the assigned POS tag and base form

When run as part of a workflow with the Apertium morphological analyser on a single core machine with 8 GB RAM, the Apertium Tagger takes approximately -.87 seconds to run, with the complete workflow taking around 1.13 seconds.

ADMINISTRATIVE INFORMATION

Contact

Contacts for the Apertium system can be found here: http://wiki.apertium.org/wiki/Contact

For further information regarding this UIMA wrapper for the Apertium morphological analyser module, please contact Sophia Ananiadou:
sophia.ananiadou@manchester.ac.uk

REFERENCES

Armentano-Oller, C., Carrasco, R., Corbí-Bellot, A.,Forcada, M., Ginestí-Rosell, M., Ortiz-Rojas, S. (2006). Open-source Portuguese–Spanish machine translation. Computational Processing of the Portuguese Language ,50-59

Ferrucci, D., Lally, A., Gruhl, D., Epstein, E., Schor, M., Murdock, J. W., Frenkiel, A., Brown, E.W. , Hampp T., Doganata, Y., Welty, C., Amini, L., Kofman, G., Kozakov, L. and Mass, Y. (2006). Towards an Interoperability Standard for Text and Multi-Modal Analytics. IBM Research Report RC24122.

Kano, Y., Baumgartner Jr., W. A, McCrochon, L., Ananiadou, S., Cohen, K. B., Hunter, L. and Tsujii, J. (2009). U-Compare: share and compare text mining tools with UIMA. *Bioinfomatics*, 25(15), 1997-1998.

Kano, Y., Miwa, M., Cohen, K. B., Hunter, L., Ananiadou, S. and Tsujii, J.. (2011). U-Compare: a modular NLP workflow construction and evaluation system. *IBM Journal of Research and Development*, 55(3), 11:1 - 11:10.

Apertium MT transfer

Apertium Mophological generator

Newly added in Batch 2

U-Compare Cafetiere sentence splitter

1. Basic Information

Tool name

U-Compare Catetiere English sentence detector

Overview and purpose of the tool

The purpose of the tool is to detect sentence boundaries in English text. The tool is provided as a UIMA[29] (Ferrucci et al., 2006) component, specifically as Java archive (jar) file, which can be incorporated within any UIMA workflow. However, it is particularly designed use in the U-Compare text mining plaform (Kano et al., 2009; Kano et al., 2011; see separate META-SHARE record), since the types of annotations it produces are compliant with the U-Compare.

A short description of the algorithm

The sentence detector uses a set of rules to break texts into sentences.

2. TECHNICAL INFORMATION

Software dependencies and system requirements

The tool can most easily be run using the U-Compare platform (see separate META-SHARE record). The only requirement to run U-Compare is for Java 6 to be installed.

For use in UIMA workflows outside of U-Compare, UIMA will need to be installed; see: http://uima.apache.org/

Installation

In order to run the sentence splitter in U-Compare, it must be imported into the system. Information about downloading and running U-Compare can be found in the documentation associated with the META-SHARE record for the U-Compare Workbench or the U-Compare website: http://nactem.ac.uk/ucompare/.

Importing the UIMA component (provided as the file CafetiereEngilshSentenceSplitter.jar) is carried out in U-Compare as follows:

1) From the “Library” menu in the U-Compare Workbench, choose the item “Register External Components (Edit Classpath)”. This will cause a window to appear that allows external components to be managed. It is shown in Figure 1.

[29] http://uima.apache.org/

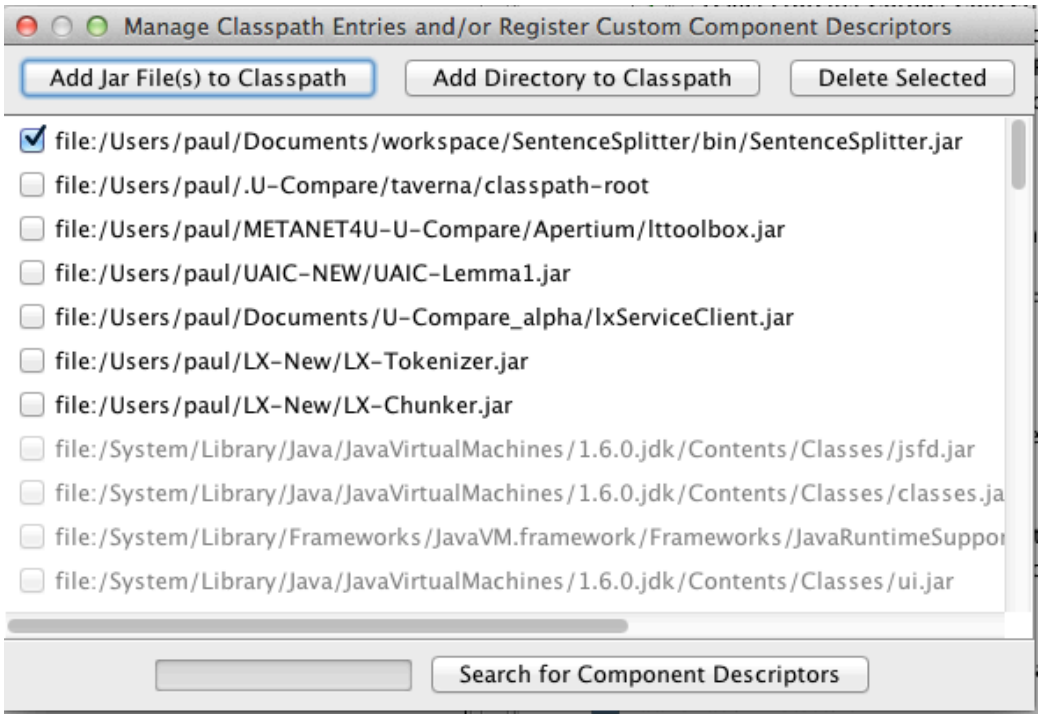


Figure 1: External component management window

2) Click on the button "Add Jar File(s) to Classpath", and browse to the location where the file "CafetiereEngilshSentenceSplitter.jar" has been saved, and click on "Open". The will cause the file to be displayed in the list of files in the external component manager window.

3) Ensure that box next to the file "CafetiereEngilshSentenceSplitter.jar" is checked in the external component manager window. Then, click the "Search for Component Descriptors" button. A "Component Descriptors Search Results" window will appear (Figure 2).

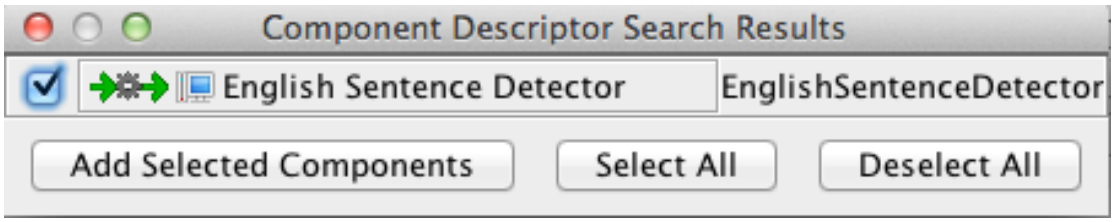


Figure 2: Component Descriptor Search Results window

4) Check the box next to "English Sentence Detector", and click on the "Add Selected Components" button. The name of the component will then appear in the library on components on the right hand side of the main U-Compare workbench window, under "Custom components", and it can then be used in workflows.

Execution instructions

One imported into U-Compare, the sentence splitter can be used simply by dragging and dropping it into a workflow using the graphical user interface of the U-Compare workbench. Alternatively, it can be incorporated into other UIMA-based workflows, by following the documentation on the Apache UIMA site: http://uima.apache.org/

Input/Output data formats

Input data formats

The input is plain text document that has previously been read into the UIMA Common Analysis Stucture (CAS) via a UIMA collection reader component, i.e. this will normally be the first annotation tool that is run in a workflow/

Output data format

The tool detects the boundaries of sentences and adds an annotation of the type org.u_compare.shared.syntactc.Sentence (which is one of the types in the interoperable U-Compare type system) for each sentence in the document to the UIMA CAS.

Integration with external tools

As mentioned above, the tool can only be run within U-Compare framework, or more generally, within the UIMA framework.

3. Content Information

Figure 1 shows the output of the tool in the U-Compare workbench. One of the sentences is highlighted. The sample text is taken the US National Library of Medicine website (http://www.nlm.nih.gov/databases/alerts/2011_nhlbi_ifp.html)

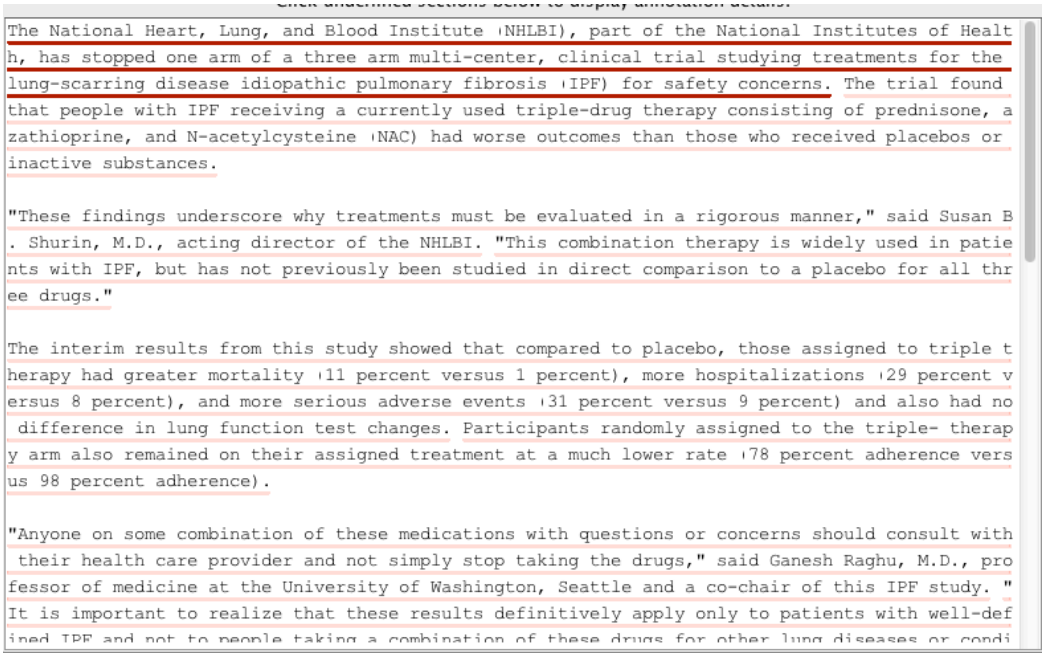


Figure 1: Output of Cafetiere Sentence Detector in the U-Compare workbench

Running the tool on the 4 KB text on a single core machine with 8 GB RAM takes around 30 milliseconds.

4. Administrative Information

Contact

For further information, please contact Sophia Ananiadou:
sophia.ananiadou@manchester.ac.uk

5. References

Ferrucci, D., Lally, A., Gruhl, D., Epstein, E., Schor, M., Murdock, J. W., Frenkiel, A., Brown, E.W. , Hampp T., Doganata, Y., Welty, C., Amini, L., Kofman, G., Kozakov, L. and Mass, Y. (2006). Towards an Interoperability Standard for Text and Multi-Modal Analytics. IBM Research Report RC24122.

Kano, Y., Baumgartner Jr., W. A, McCrochon, L., Ananiadou, S., Cohen, K. B., Hunter, L. and Tsujii, J. (2009). U-Compare: share and compare text mining tools with UIMA. *Bioinfomatics*, 25(15), 1997-1998.

Kano, Y., Miwa, M., Cohen, K. B., Hunter, L., Ananiadou, S. and Tsujii, J.. (2011). U-Compare: a modular NLP workflow construction and evaluation system. *IBM Journal of Research and Development*, 55(3), 11:1 - 11:10.

UAIC - University Alexandru Ioan Cuza

Endogenous resources (tools)

Categorizer-UAIC

1. BASIC INFORMATION

Tool name

CategoriZer provides tools for automatic extraction of language indicators that help researches to monitor language use. Here, language indicators refer to: frequent words, metrics, key words, etc.

Overview and purpose of the tool

The initial purpose of the CategoriZer application was the creation of an interrogation tool for a book database within which the user would find books by using keywords and specifications. Working at the project, we decided to extend this research and now the user can specify a book and its categories are returned due to a very rough text analysis.

The application automatically trains an SVM text classification system to put text in certain categories, by running daily targeted searches on those categories and carefully studying the results using Google's various search technologies.

A short description of the algorithm

Database

The database management system used is Apache Derby. The database is composed of two schemes, APP and CLARIN. Each scheme has its own tables: “Categories”, “Parsed” and “Admin” tables for “APP” scheme. “CLARIN” scheme has a generated table for each category which is found during the indexing process in the library of books.

The “Categories” and “Parsed” tables are automatically generated when the admin requests an indexing process. This process depends on what files are located in the “Books” folder contained in our project.

In the tables (Adventure, Architecture, Art, Biographies, Crafts and Hobbies, Fairy Tales, History, Horror, Humor, Medicine, Music, Mysteries, Nature, Philosophy, Psychology, Racism, Science Fiction) from the “CLARIN” scheme we store the stems and their frequencies found in the books for each category separately. Therefore, we have the following columns: “Id”, “Word”, “Frequency”, “Report” and “Score”.

In order to use our database on another computer we created a backup function that safely allows the backup and restore the database.

The Parsing Algorithm

The parsing algorithm computes the frequencies of the stems found in a text file. This is used intensively on the indexing process but it is also used for categorizing a book.

The algorithm first removes punctuation signs and then tokenizes all the words of the book. Then it applies an aggressive stemming algorithm called "Lovins Stemmer" on each token and returns a list of all stems and their frequencies found in the book given as input. During the indexing process this list is used to update the category to which the parsed book belongs. During the categorization process the book given by the user is parsed and the list returned is compared with the list computed for each category in the database.

The Categorizing Algorithm

Firstly the book given by the user is parsed, after which a prominence score is computed for every category. This is done by comparing relative frequencies of each stem found in the book, and the ones in each category. Every stem in each category has a property named *score*, which is the relevance of that stem for the given category. The score is bigger if the stem is not as frequent in other categories. The final percentage is based on a formula that takes in account both the importance computed and the scores of the stems in the book. In essence, this algorithm is a form of TF/IDF.

Gutenberg

Gutenberg is one of the first producers of free electronic books, having more than 30 000 e-books which were digitized with the help of thousands of volunteers. Since the main idea of the project is to tell the category of a book, we needed a big database which would become our starting point for developing the project. In order to do this, we developed a web crawler that downloaded all the books from 19 main categories in Gutenberg. Once we had the books collection we started processing and obtained statistical results that would help categorizing a book. The books categories in Gutenberg are contributed by a large community of human users.

2. TECHNICAL INFORMATION

Software dependencies and system requirements

The aligner is implemented in the Java programming language. It is deployed as a web application at the address http://nlptools.info.uaic.ro/Categorizer/. It requires only a web browser to run as client. On the server, the following resources are needed:

1. Java Runtime Environment 1.7;
2. Glassfish server;
3. SQL server.

Installation

The application does not require any installation as it is accessible as a web application.

Execution instructions

The program is accessible by opening a web browser and typing the address of the service in the address bar.

The CategoriZer application is composed of two main modules: Book categorization for users and books collection database management for administrators.

Admin interface

If the admin wants to login, he needs to have a valid username and password. Once he is logged in, he has access to the collections of books.

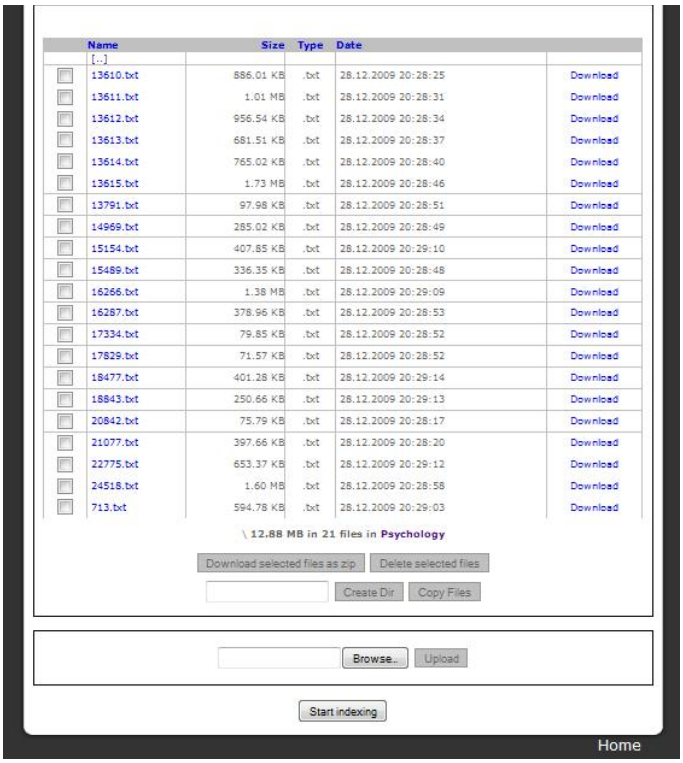


Fig. 1: Administrator interface

Currently we have 19 categories available for indexing, with a total of 2248 books. These categories are: *Adventure, Architecture, Art, Biographies, Crafts and Hobbies, Fairy Tales, History, Horror, Humor, Medicine, Music, Mysteries, Nature, Philosophy, Plays, Poetry, Psychology, Racism, Science Fiction*. In this part, the admin, can delete a category (or more), can copy the files from a category to another, or he can create a brand new category. He is not allowed to add a book in the main root (if he tries this, an error message will appear on the screen). The book must be in a category. When the admin selects a category, the application lists the books it contains.

In this screen the admin can download a book, add a book in the category, delete a book, or even copy books from the current category to another. A book can be indexed as being part of two different categories.

After modifying the structure of the library, the admin can press the “Start indexing” button to make the server parse all the books and modify the database accordingly.

User interface

Normal users can use CategoriZer to determine the best five categories that a book matches according to the database created by the admin. Users can specify the url of a text and request it's categorization. When this happens, the server starts to process and the user is presented with a progress bar.

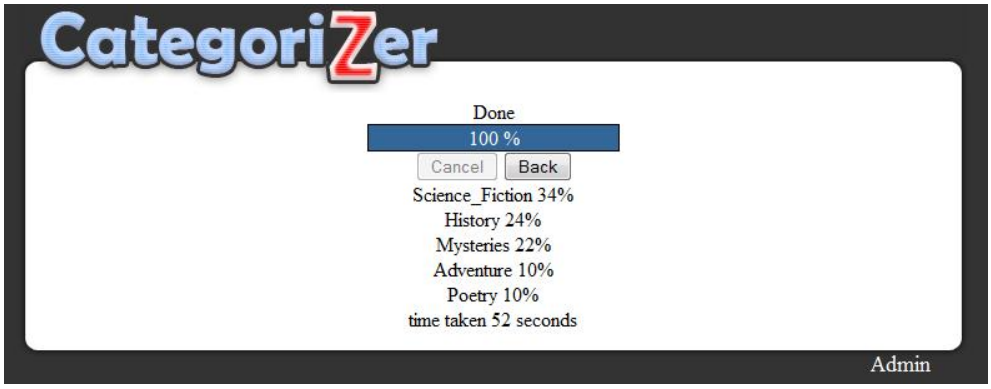


Fig. 2: User interface

Above is an example screen of the results returned. The categorization indicates that the given text has elements from *Science Fiction, History, Mysteries, Adventure* and *Poetry*.

Input/Output data formats

Input data formats

The input of the program is the URL of the text the user wants to categorize.

Output data format

The output of the application is the category information of the input text. An example of the output is given in fig. 2

Integration with external tools

The application does not need any external tool.

3. CONTENT INFORMATION

The input files are generally in text format.

4. ADMINISTRATIVE INFORMATION

Contact

For further information and technical support installing and/or running this tool, please email to Radu Simionescu: radu.simionescu@info.uaic.ro.

Discourse Parser

1. Basic Information

Tool name

DiscourseParser-UAIC: builds an RST-like tree structure of an input document.

Overview and purpose of the tool

The purpose of the tool is to build an a binary RST-like tree structure of an input text. Nodes in the tree are discourse spans, leaves are discourse units (clauses or simple sentences). Nodes of the tree are labelled as nuclear or satellite. Under any node there are either two nuclei or one nucleus and a one satellite.

A short description of the algorithm

First, the text is tokenised, POS-tagged and lemmatised. Further, the process is split into two flows: one that segments the sentences into elementary discourse units (edus) and then constructs elementary discourse trees (edts) of each sentence, and another one that detects NPs and then runs an anaphora resolution (AR) engine to detect coreferential relations. Intermediate files in the processing flow are in the XML format. When two processes join, the resulted files are merged into a single representation. An edt is a discourse tree whose leaf-nodes are the edus of one sentence. Sentence-internal cue-words/phrases trigger the constituency of edts out of each sentence. For each sentence in the original text a set of edts is obtained. At this point a process that simulates the human power of incremental discourse processing is started. At any moment in the developing process, say after *n* steps corresponding to the first *n* sentences, a forest of trees is kept, representing the most promising structures built by combining in all possible ways all edts of all *n* sentences. Each such tree corresponds to one possible interpretation of the text processed so far. Then, at step *n*+1 of the incremental discourse parsing, the following operations are undertaken: first, all edts corresponding to the next sentence are integrated in all possible ways onto all the trees of the existing forest; then the resulted trees are scored according to four independent criteria, sorted and filtered so that only a fraction of them is retained (again the most promising after *n*+1 steps). From the final bunch of trees, only the best ranked tree is taken to be the discourse structure. A general framework to resolve anaphors is used to integrate a model of coreference resolution that deals with most types of anaphors. Centering transitions scores are computed after AR is run, therefore after all references are solved. References and transitions, as well as heuristics for the proper development of a discourse tree, contribute with scores to the overall score of a developing discourse tree. These scores are then used to control the beam-search.

2. TECHNICAL INFORMATION

Software dependencies and system requirements

The tool is developed in Java. It requires the following settings to run:

1. Java VM any newer than 1.7
2. 1+ GB RAM
3. A language model (english version is included in the archive).

Installation

Unpack the archive.

Execution instructions

Edit “config.cfg” to indicate input and output files.

Input/Output data formats

Input data formats

Input is tokenized, part-of-speech tagged, lemmatized and annotated with coreference chains XML document with the following format:

```
<?xml version="1.0" encoding="UTF-8"?>
<DOCUMENT>
<P ID="GB-SB.txt">
        <S ID="S0">
<CLAUSE ID="0" TYPE="SENTENCE">
         <W ID="W0" LEMMA="" POS="N" NUM="SG" PERSONTYPE="OBJECT" oldPOS="NN">"</W>
         <W ID="W1" LEMMA="in" POS="PREP" NUM="" oldPOS="IN">In</W>
…
</CLAUSE>
        </S>
</P>
<DE  ID="0" reList="0,4,12,60,94"/>
…
</DOCUMENT>
```

Output data format

Outputs add to the input data about the nodes in the discourse tree, as follows:

```
<ROOT ID="R0" REL="N_N" LEFT="N1" RIGHT="N4" TYPE="PARAGRAPH" VEIN="[0, 10, 21, 23, 27, 29, 30, 52, 56, 72, 75, 76, 79, 80, 81, 112, 113, 116, 117, 118, 119, 120, 121, 122, 128, 131, 132, 133, 134, 135, 136, 137, 138, 139, 152, 153]"/>
<NODE ID="N1" REL="N_S" LEFT="0" RIGHT="N2" TYPE="SENTENCE" VEIN="[0, 10, 21, 23, 27, 29, 30, 52, 56, 72, 75, 76, 79, 80, 81, 112, 113, 116, 117, 118, 119, 120, 121, 122, 128, 131, 132, 133, 134, 135, 136, 137, 138, 139, 152, 153]"/>
…
</ROOT>
```

The output includes computed veins and types of relations between the two children of each internal node.

Integration with external tools

The application does not need any external tool.

2. CONTENT INFORMATION

An example of an input file and the corresponding output is included in the archive and their names are input.xml and output.xml.

Running this example on a quad core machine with 4 GB RAM would take 0.2 seconds (including writing the output XML).

3. ADMINISTRATIVE INFORMATION

Contact

For further information and technical support installing and/or running this tool, please email to Daniel Anechitei (daniel.anechitei@info.uaic.ro) and Ionut Pistol: ipistol@info.uaic.ro

4. REFERENCES

Pistol I.C. "Romanian Processing Chains in Metanet4U", Proceedings of the 8th International Conference "Linguistic Resources and Tools for Processing of the Romanian Language". 8-9 December 2011, Ed. M. Moruz, D. Cristea, D. Tufis, A. iftene, H Teodorescu, "Al. I. Cuza" University publishing House, ISSN 1843-911X

Cristea, D. (2005): Motivations and Implications of Veins Theory. Bernadette Sharp (Ed.). Natural Language Understanding and Cognitive Science, Proceedings of the 2nd International Workshop on Natural Language Understanding and Cognitive Scienc3, NLUCS 2005, in conjunction with ICEIS 2005, Miami, U.S.A., May 2005, INSTICC Press, Portugal, ISBN 972-8865-23-6X, pp. 32-44

Cristea, D., Postolache, O., Pistol, I. (2005): Summarisation through Discourse Structure. In Alexander Gelbukh (Ed.): Computational Linguistics and Intelligent Text Processing, 6th International Conference CICLing 2005, Mexico City, Mexico, February 2005, Proceedings, Springer LNSC, vol. 3406, ISBN 3-540-24523-5, pp. 632-644

Lemmatizer

1. BASIC INFORMATION

Tool name

Lemmatizer-UAIC: Lemmatizer for the Romanian language

Overview and purpose of the tool

The purpose of the tool is to add the base form for each lexical token in a Romanian language text. The first version of the tools requires just a tokenized UTF-8 input; the second requires part-of-speech data to be present. The second version, benefiting from part-of-speech disambiguation, produces better results for most cases.

A short description of the algorithm

The tool uses an external resource in the form of a dictionary of flexed word forms accompanied by the corresponding base forms. If part-of-speech data is present in the input file, this information is used to select from

possible multiple entries of the same flexed form, otherwise the first corresponding base form encountered is selected.

2. TECHNICAL INFORMATION

Software dependencies and system requirements

The original Python tool was rewritten in Java, updated and integrated in UIMA/U-Compare. It requires the following settings to run:

1. Java VM any newer than 1.5
2. 1+ GB RAM
3. The U-Compare platform, available at http://u-compare.org/
4. The flexed forms dictionary, available as part of the archive.

Installation

The tool is available as U-Compare component and is installed using the U-Compare interface following the instructions available at http://u-compare.org/userguide/Creating_Workflow.html#SECTION00043000000000000000 .

Execution instructions

The first version Lemmatizer_v1

Create a blank workflow in U-Compare, add either "Input Text Reader" or "File System Collection Reader", then a tokenizer component and the "UAIC-Lemmatizer_v1" component. For the "UAIC-Lemmatizer_v1" component select the "bd.txt" file as run parameter. Input or select input files, then use the run button in the U-Compare interface.

The second version Lemmatizer_v2

Create a blank workflow in U-Compare, add either "Input Text Reader" or "File System Collection Reader", then a tokenizer component and a part-of-speech tagging component producing RichToken as output, and the "UAIC-Lemmatizer_v2" component. For the "UAIC-Lemmatizer_v2" component select the "bd.txt" file as run parameter. Input or select input files, then use the run button in the U-Compare interface.

Input/Output data formats

Input data formats

The first version Lemmatizer_v1

As input you can either build the above workflow (working on plain UTF8 text) or produce a Token annotated file according to the U-Compare type system described in http://www.aclweb.org/anthology/W/W09/W09-1504.pdf .

The second version Lemmatizer_v2

As input you can either build the above workflow (working on plain UTF8 text) or produce a RichToken annotated file according to the U-Compare type system described in http://www.aclweb.org/anthology/W/W09/W09-1504.pdf .

Output data format

Both types of lemmatizers produce the same output format. U-Compare can output several formats, the standard U-Compare output being a UTF-8 XML as shown below.

<?xml version="1.0" encoding="UTF-8"?>

<Document>

<org.apache.uima.examples.SourceDocumentInformation sofa="Sofa" begin="0" end="726" uri="C:/U-Compare/temp/interactive..." offsetInSource="0" documentSize="791" lastSegment="true">

<uima.tcas.DocumentAnnotation sofa="Sofa" begin="0" end="726" language="en">

<org.u_compare.shared.syntactic.Token sofa="Sofa" begin="0" end="3" fragments="null" metadata="null">

<org.u_compare.shared.syntactic.RichToken sofa="Sofa" begin="0" end="3" fragments="null" metadata="null" pos="UnknownPOS" posString="V" base="fi">Era</org.u_compare.shared.syntactic.RichToken>

…

</uima.tcas.DocumentAnnotation>

</org.apache.uima.examples.SourceDocumentInformation>

</Document>

Integration with external tools

The application does not need any external tool (other than the U-Compare platform).

3. CONTENT INFORMATION

An example of an input file and the corresponding output is included in the archive and their names are input.txt and output.xml.

Running this example on a quad core machine with 4 GB RAM and U-Compare 1.13 would take 0.3 seconds (including writing the output XML, excluding the tokenization and part-of-speech annotation).

4. ADMINISTRATIVE INFORMATION

Contact

For further information and technical support installing and/or running this tool, please email to Ionut Pistol: ipistol@info.uaic.ro.

5. REFERENCES

Pistol I.C. "Romanian Processing Chains in Metanet4U", Proceedings of the 8th International Conference "Linguistic Resources and Tools for Processing of the Romanian Language". 8-9 December 2011, Ed. M. Moruz, D. Cristea, D. Tufis, A. iftene, H Teodorescu, "Al. I. Cuza" University publishing House, ISSN 1843-911X

NP Chunker

1. BASIC INFORMATION

Tool name

NP-chunker-UAIC: Romanian deep noun phrase chunker

Overview and purpose of the tool

The purpose of the tool is to determine noun phrases boundaries, included nested, in a Romanian language text. An noun phrase (NP) can contain other NPs in a recursive manner. The tool requires tokens and their part-of-speech data to be present.

A short description of the algorithm

The tool is based on the Graphical Grammar Studio (GGS), an in-house tool, developed by the same author. GGS is a grammar applier, similar with Nooj in some aspects, developed in Java. A grammar is represented in a visual fashion. A grammar is composed of nodes which can consume at most one input token at a time. Nodes are organised in networks, and jumps from a node to another network can be made.

The NP chunking grammar for Romanian was developed by the same author. It contains in total 122 graphs and 1288 nodes from which 212 are token matching nodes(the rest are nodes which are used for jumping or which do not consume input) and 466 are jump nodes. It contains in total 1621 arcs.

2. TECHNICAL INFORMATION

Software dependencies and system requirements

The tool is public as a web service and web application. The tool itself, on the server side, requires

1. Java VM any newer than 1.5
2. 1+ GB RAM
3. The GGS engine library
4. The np chunking grammar for romanian

Installation

The tool is available as a web service and a web application. The web application URL is http://nlptools.info.uaic.ro/WebNpChunkerRo/ . The WSDL for the web service can be found at http://nlptools.info.uaic.ro/WebNpChunkerRo/NpChunkerRoWS?wsdl

Execution instructions

The web application

The web application is oriented towards testing and it basically has a demonstrative purpose. The user can give as input a raw Romanian text. The system splits, tokenizes and PoS-taggs the text before feeding it to the GGS engine to obtain NP annotations. The result is displayed in an interactive window. The user can visually see the NP boundaries and can click on tokens to view the PoS information (part of speech and morphologic features)

The web service

The web service is designed as a stand-alone NLP component. It exposes two methods:

- *chunkText* chunks raw unprocessed Romanian text (with diacritics). The processing consists in a chain. The raw text is being split, tokenized and pos tagged using the UAIC pos tagger service from
- *chunkTaggedText* takes as input text which is already split, tokenized and PoS-tagged text (in a format described in the next section) and adds the NP annotations. The returned value is also an xml document, represented also as a string.

Input/Output data formats

The web application

The input is a raw, unformatted, Romanian text. It is highly recommended that this text contains diacritics. The output is offered in the html format, in a visually pleasing manner. There is currently no user friendly method of saving this output to the local drive for further use. For this, one should use the web service instead.

The web service

There are two possibilities to use the web service: through he method *chunkText* and through the method *chunkTaggedText*. For the first method, the input is raw unprocessed Romanian text (with diacritics). The format of input for the second method, the *chunkTaggedText***,** must have sentence tags named S (see example below). The children of these will be considered the tokens. Each token must contain the word form as its inner text and must have a LEMMA attribute and a MSD attribute. The tagset used for the MSD tags is the one presented in the first annex of [Simionescu, 2011a].

Input data format sample for method chunkTaggedText

…

<S id="2" offset="4">

<W LEMMA="între" MSD="Sp" id="1" offset="0">Într-</W>

<W LEMMA="un" MSD="Tifsr" id="2" offset="5">o</W>

<W LEMMA="seară" MSD="Ncfsrn id="3" offset="7">seară</W>

<W LEMMA="de~la" MSD="Sp" id="4" offset="13">de la</W>

<W LEMMA="început" MSD="Ncmsry" id="5" offset="19">începutul</W>

<W LEMMA="lui" MSD="Ds3---s” id="6" offset="29">lui</W>

<W LEMMA="iulie" MSD="Ncmsry" id="7" offset="33">iulie</W>

<W LEMMA="1909" MSD="M" offset="39">1909</W>

<W LEMMA="," MSD="COMMA" id="9" offset="43">,</W>

<W LEMMA="cu" MSD="Sp" id="10" offset="45">cu</W> <S id="2" offset="4">

<W LEMMA="între" MSD="Sp" id="1" offset="0">Într-</W>

<W LEMMA="un" MSD="Tifsr” id="2" offset="5">o</W>

<W LEMMA="seară" MSD="Ncfsrn id="3" offset="7">seară</W>

<W LEMMA="de~la" MSD="Sp" id="4" offset="13">de la</W>

<W LEMMA="început" MSD="Ncmsry” id="5" offset="19">începutul</W>

<W LEMMA="lui" MSD="Ds3---s" id="6" offset="29">lui</W>

<W LEMMA="iulie" MSD="Ncmsry" id="7" offset="33">iulie</W>

<W LEMMA="1909" MSD="M" id="8" offset="39">1909</W>

<W LEMMA="," MSD="COMMA" id="9" offset="43">,</W>

<W LEMMA="cu" MSD="Sp" id="10" offset="45">cu</W>

…

Output data format sample

Any additional attributes (others then LEMMA and MSD which are required for the processing) of the tokens will be preserved. Also, some additional attributes will be added for each token, representing a detailed explanation of the MSD tag provided as input. The NP chunks are annotated as inline NP xml tags. The rest of the input xml structure will be preserved.

…

<S id="2" offset="4">

<W LEMMA="între" MSD="Sp" POS="ADPOSITION" id="1" offset="0">Într-</W>

<NP>

<W Case="direct" Gender="feminine" LEMMA="un" MSD="Tifsr" Number="singular" POS="ARTICLE" Type="indefinite" id="2" offset="5">o</W>

<HEAD>

<W Case="direct" Definiteness="no" Gender="feminine" LEMMA="seară" MSD="Ncfsrn" Number="singular" POS="NOUN" Type="common" id="3" offset="7">seară</W>

</HEAD>

<W LEMMA="de~la" MSD="Sp" POS="ADPOSITION" id="4" offset="13">de la</W>

<NP>

<HEAD>

<W Case="direct" Definiteness="yes" Gender="masculine" LEMMA="început" MSD="Ncmsry" Number="singular" POS="NOUN" Type="common" id="5" offset="19">începutul</W>

</HEAD>

<W LEMMA="lui" MSD="Ds3---s" POS="DETERMINER" Person="third" Possessor_number="singular" Type="possessive" id="6" offset="29">lui</W>

<NP>

<HEAD>

<W Case="direct" Definiteness="yes" Gender="masculine" LEMMA="iulie" MSD="Ncmsry" Number="singular" POS="NOUN" Type="common" id="7" offset="33">iulie</W>

</HEAD>

<W EXTRA="NotInDict" LEMMA="1909" MSD="M" POS="NUMERAL" id="8" offset="39">1909</W>

</NP>

</NP>

</NP>

<W LEMMA="," MSD="COMMA" id="9" offset="43">,</W>

<W LEMMA="cu" MSD="Sp" POS="ADPOSITION" id="10" offset="45">cu</W> <S id="2" offset="4">

<W LEMMA="între" MSD="Sp" POS="ADPOSITION" id="1" offset="0">Într-</W>

<NP>

<W Case="direct" Gender="feminine" LEMMA="un" MSD="Tifsr" Number="singular" POS="ARTICLE" Type="indefinite" id="2" offset="5">o</W>

<HEAD>

<W Case="direct" Definiteness="no" Gender="feminine" LEMMA="seară" MSD="Ncfsrn" Number="singular" POS="NOUN" Type="common" id="3" offset="7">seară</W>

</HEAD>

<W LEMMA="de~la" MSD="Sp" POS="ADPOSITION" id="4" offset="13">de la</W>

<NP>

<HEAD>

<W Case="direct" Definiteness="yes" Gender="masculine" LEMMA="început" MSD="Ncmsry" Number="singular" POS="NOUN" Type="common" id="5" offset="19">începutul</W>

</HEAD>

<W LEMMA="lui" MSD="Ds3---s" POS="DETERMINER" Person="third" Possessor_number="singular" Type="possessive" id="6" offset="29">lui</W>

<NP>

<HEAD>

<W Case="direct" Definiteness="yes" Gender="masculine" LEMMA="iulie" MSD="Ncmsry" Number="singular" POS="NOUN" Type="common" id="7" offset="33">iulie</W>

</HEAD>

<W EXTRA="NotInDict" LEMMA="1909" MSD="M" POS="NUMERAL" id="8" offset="39">1909</W>

</NP>

</NP>

</NP>

<W LEMMA="," MSD="COMMA" id="9" offset="43">,</W>

<W LEMMA="cu" MSD="Sp" POS="ADPOSITION" id="10" offset="45">cu</W>

…

Integration with external tools

The application does not need any external tools.

3. CONTENT INFORMATION

An example of an input xml and the corresponding returned response for the *chunkTaggedText* method are included in the archive and their names are input_samp.xml and output_sampl.xml.

4. **ADMINISTRATIVE INFORMATION**

Contact

For further information and technical support installing and/or running this tool, please email to Radu Simionescu: radu.simionescu@info.uaic.ro or radsimu@gmail.com

5. **REFERENCES**

Simionescu R (2011a) *Hybrid POS-tagger* (in Romanian). Master thesis presented at the Faculty of Computer Science, University Al. I. Cuza of Iași, http://nlptools.infoiasi.ro/WebPosRo/.

Simionescu R. (2011b) *Romanian Deep Noun Phrase Chunking Using Graphical Grammar Studio*, Proceedings of the 8th International Conference "Linguistic Resources and Tools for Processing of the Romanian Language". 8-9 December 2011, Ed. M. Moruz, D. Cristea, D. Tufis, A. Iftene, H Teodorescu, "Al. I. Cuza" University publishing House, ISSN 1843-911X.

Graphical Grammar Studio http://sourceforge.net/projects/ggs/

RARE

1. BASIC INFORMATION

RARE: Robust Anaphora Resolution Engine

Overview and purpose of the tool

The purpose of the tool is to extract coreference chains. Anaphora is a semantic relation that is evidenced during the interpretation of two surface strings, usually called referential expressions (REs), when the interpretation of one of them (anaphor) depends on the other (antecedent), see for instance (Kamp & Reyle, 1993).

In (Cristea and Dima, 2001) a general AR framework capable to handle complex anaphoric phenomena has been introduced. Later, RARE – an incremental system implementing this philosophy – has been built (Postolache and Forascu, 2004). The general assumption behind the framework is that the interpretation of anaphora should also involve a semantic layer that stores information acquired from the textual layer. In between these two, a third layer is placed for bookkeeping operations during the interpretation process.

The type of analysis supported by the framework is incremental. Just like in normal reading, anaphors are mostly resolved at the time of reading, but sometimes decisions are postponed until the acquisition of complementary information that helps the disambiguation process.

A short description of the algorithm

RARE was built having in mind two important principles:

1. References are semantic not textual (Halliday, Hasan, 1976)

2. Incremental parsing

The first principle means that the interpretation of the coreferential relations would need at least two levels: a textual one and a semantic one. A referential expression evokes a semantic representation and is coreferential with an expression introduced later on in the discourse, which evokes a semantic meaning.

The second principle continues the first one, saying that between the textual level and the semantic level there is another level, the projection one. The projection level is useful as long as it helps with the identification of the referent, because when the reference is resolved this level is no longer necessary.

A referential expression projects a feature structure on the projection level, which evokes a centre at the semantic level. The feature structure from the projection level is destroyed after this centre is built.

2. TECHNICAL INFORMATION

Software dependencies and system requirements

RARE is implemented in the programming language Java, under the Java 1.6 JDK. It requires the following settings to run:

1. Java 1.6 JDK

2. 1+ GB RAM (2 GB preferred)

Installation

The application does not require any installation aside that of Java 1.6 JDK, which is publicly available.

Execution instructions

Given that the user's machine has Java 1.6 JDK installed, the application can be run as an executable file both under Windows and Linux platforms.

The ".jar" file must be placed in a working folder, containing two subfolders: "lib" and "language folder". The input should be also found in the working folder.

The "lib" folder has jdom.jar library.

The "language folder" folders will contain the resource file used by RARE. These resource files are as follows: stopwords.txt – a file containing the stopwords; window.xml – a xml file that describes where to look for the anaphora, a number of sentences; tagset.xml – a file that does the mapping of a tagset to the internal tagset used by RARE; constraints.xml – is the file that describes the rules used by RARE

The "constraints.xml" file, reproduced below, is self-explanatory:

<rule ID="numberAgreement" type="promoting" score="1.1">

; Promoting rule for anaphors of type pronoun.

; If the constraints are met, the current score for the DE to act as an antecedent of the current RE is augmented by 10%

<constraint on="RE" attr="POS" val="pron">

; If equal(RE.POS, pron) is TRUE

<relation pred="equal" attrList="NUM"/>

; then a DE such that equal(RE.NUM, DE.NUM)=TRUE will be selected

</rule>

<rule ID="pronounAgreementOnNumGen" type="promoting" score="1.2">

; A pronoun should agree on number and gender with a person

; antecedent

<constraint on="RE" attr="POS" val="pron">

<constraint on="RE" attr="PERS" val="3">

<constraint on="RE" attr="NUM" val="SG">

<constraint on="DE" attr="ENTITYTYPE" val="PERSON"/>

<relation pred="equal" attrList="NUM GEN"/>

</rule>

<rule ID="heIsMale" type="promoting" score="1.2">

; "he/him/his" mainly refers an entity of type PERSON, MALE

<constraint on="RE" attr="POS" val="pron">

<constraint on="RE" attr="PERS" val="3">

<constraint on="RE" attr="GEN" val="M">

<constraint on="RE" attr="NUM" val="SG">

<constraint on="DE" attr="ENTITYTYPE" val="PERSON"/>

<constraint on="DE" attr="PERSONTYPE" val="MALE"/>

</rule>

Input/Output data formats

Input data formats

The input text of the documents must have the format, and must be placed in the same folder with the “rare.jar” file:

<W ID="9" LEMMA="live" POS="VBD">lived</W>

<NP ID="4" HEADID="11">

<W ID="10" DETTYPE="UNDEF" LEMMA="a" POS="DET">a</W>

<W ID="11" LEMMA="king" NUM="SG" PERSONTYPE="PERSON" POS="N">king</W>

</NP>…

Output data format

The program outputs a XML UTF-8 file in the format:

<W ID="9" LEMMA="live" POS="VBD">lived</W>

<NP ID="4" HEADID="11" ref=”4”>

<W ID="10" DETTYPE="UNDEF" LEMMA="a" POS="DET">a</W>

<W ID="11" LEMMA="king" NUM="SG" PERSONTYPE="PERSON" POS="N">king</W>

</NP>…

<DE ID="0" reList="0,3" />

<DE ID="1" reList="4,23,18,19,25,73,84,127" />

Integration with external tools

The application does not need any external tool.

3. CONTENT INFORMATION

Examples of input files are included into the archive and their names are input.xml and output.xml. The result of processing the two input file is shown in the file output.xml. Running this example on a dual core machine with 4 GB RAM would take 11 seconds.

4. ADMINISTRATIVE INFORMATION

Contact

For further information and technical support installing and/or running this tool, please email to Eugen Ignat: eugen.ignat@info.uaic.ro.

5. REFERENCES

Cristea,D., Dima,G.E. *An integrating framework for anaphora resolution*. In Information Science and Technology, Romanian Academy Publishing House, Bucharest, vol. 4, no. 3-4, p 273-291, 2001.

Gamallo, P. 2008 *Evaluating two different methods for the task of extracting bilingual lexicons from comparable corpora*. In Proceedings of LREC 2008 Workshop on Comparable Corpora, Marrakech, Morocco, pp. 19-26. ISBN: 2-9517408-4-0.

Halliday, M. A. K., and Ruqaiya Hasan. 1976.Cohesion in English. London: Longman.

Hans Kamp and Uwe Reyle. *From Discourse to the Lexicon: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer Academic Publishers, 1993.

Postolache, O., Forăscu, C. *A Coreference Resolution Model on Excerpts from a Novel*, in Proceedings of the ESSLLI Student Session, August 2004, Nancy, France, pp. 202-213, 2004.

Rapp, R. 1999. *Automatic Identification of Word Translations from Unrelated English and German Corpora*. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), pages 519-526, college Park, Maryland, USA.

Splitter v1 and v2

1. BASIC INFORMATION

Tool name

Splitter-UAIC: a tool that segments a text in discourse elementary units (clauses).

Overview and purpose of the tool

The purpose of the tool is to add delimitations around the clauses present in the input text. Two versions are available, a basic one working on plain text input, and a complex one requiring previous annotations to be available, as well as a trained model, which is made available for English.

A short description of the algorithm

Splitter-UAIC_v1

A set of heuristics are used to determine if markers found in text separate adjacent clauses.

Splitter-UAIC_v2

A clause is headed by a main verb or verbal compound. Verbs and verb compounds are considered pivots and clause boundaries are looked for in-between them. Verb compounds are either complex verbal constructions (main verb plus auxiliaries) or made of a main active verb and other verbs filling up compulsory role slots. An example of a compound verb is “like to swim” (1) where “like” and “swim” are verbs in-between which there should not appear any clause boundary, or otherwise the two resultant clauses would lose coherence. Identically, “managed to annotate” (2) is considered as being a compound verb.

<When I go to the river,>< I like to swim with friends.> *(1)*

<John told me>< he managed to annotate the whole file.> *(2)*

The exact place of a clause boundary between verbal constructs are best indicated by markers (key words or expressions) like in examples (3) and (4).When they are missing, boundaries are found statistically, based on explicit annotations in training files.

<Markers are good><because they can give information about the discourse structure.>(3)

<Verbs and verb compounds are considered pivots><and clause boundaries are looked for in-between them.> *(4)*

In order to recognize the markers, the training is made by using a window of *n*POS tags to the left of the marker and *m*POS tags to the right. For the cases in which we don't have any marker at the boundary between clauses, a symmetrical window of *l*POS tags is used. The values of the three parameters *m*, *n* and *l*is determined after running optimisation tests.

2. TECHNICAL INFORMATION

Software dependencies and system requirements

Both versions are written in Java and require the following settings to run:

1. Java VM any newer than 1.7
2. 1+ GB RAM
3. For the second version, a trained language model is required (English version included as part of the archive).

Installation

No installation required, just unpack the archives.

Execution instructions

Splitter-UAIC_v1

Add UTF8 text files in the same directory as the "segmenter-v1.jar" executable. They need to have the "txt" file extension. When running "segmenter-v1.jar", files with the same name as the originals (but with added "done" extension) will be created in the same directory.

Splitter-UAIC_v2

In the “segmenter-v2” directory edit “preferencesSegmenter.pref” to indicate the input and output files, then run “runSegmenterModule.jar”.

Input/Output data formats

Input data formats

Splitter-UAIC_v1

Input is UTF8 text found in files with the “txt” extension in the same directory as the executable.

Splitter-UAIC_v2

Input is tokenized and part-of-speech tagged XML, according to the format:

```
<?xml version="1.0" encoding="utf-8"?>
<DOCUMENT>
<P ID="Grimm Brothers - Sleeping Beauty.txt">
        <S ID="S00">
                <W ID="W0" POS="VBG" LEMMA="sleeping">Sleeping</W>
…
        </S>
</P>
</DOCUMENT>
```

Output data format

Splitter-UAIC_v1

Output is produced as UTF8 text with clauses separated on individual lines created in the same directory as the executable.

Splitter-UAIC_v2

Output is written in the file specified by the user, according to the format:

```
<?xml version="1.0" encoding="UTF-8"?>
<DOCUMENT>
<P ID="Grimm Brothers - Sleeping Beauty.txt">
        <S ID="S00">
<CLAUSE ID="0" TYPE="SENTENCE">
```

<W ID="W0" LEMMA="sleeping" POS="VBG">Sleeping</W>

…

</CLAUSE>

</S>

</P>

</DOCUMENT>

Integration with external tools

The application does not need any external tool.

3. CONTENT INFORMATION

Examples of input files and the corresponding output are included for both versions of the splitter. For segmenter-v1 they are “inputRO.txt” and “outputRO.txt”, for segmenter-v2 they are “input.xml” and “output.xml” – found in the Segmenter-v2 directory.

Running this example on a quad core machine with 4 GB RAM and U-Compare 1.13 would take 0.1 seconds for segmenter-v1 and 0.2 for segmneter-v2.

4. ADMINISTRATIVE INFORMATION

Contact

For further information and technical support installing and/or running this tool, please email to Daniel Anechitei (daniel.anechitei@info.uaic.ro) and Ionut Pistol: ipistol@info.uaic.ro.

5. REFERENCES

Pistol I.C. "Romanian Processing Chains in Metanet4U", Proceedings of the 8th International Conference "Linguistic Resources and Tools for Processing of the Romanian Language". 8-9 December 2011, Ed. M. Moruz, D. Cristea, D. Tufis, A. Iftene, H Teodorescu, "Al. I. Cuza" University publishing House, ISSN 1843-911X

Cristea D., Anechitei D., Ignat E. "Clause Level Multilingual Segmentation", Proceedings of the 8th International Conference "Linguistic Resources and Tools for Processing of the Romanian Language". 8-9 December 2011, Ed. M. Moruz, D. Cristea, D. Tufis, A. Iftene, H Teodorescu, "Al. I. Cuza" University publishing House, ISSN 1843-911X

Summarizer v1 and v2

1. BASIC INFORMATION

Tool name

Summarizer-UAIC: a tool that produces a summary of an input text

Overview and purpose of the tool

The purpose of the tool is to produce summaries of input texts. Two versions are available, a very simple one working on plain text input, and a complex one requiring previous annotations to be available.

A short description of the algorithm

The simple version summarizer_v1

Clauses are identified and scored according to a set of heuristics and the best ranked ones are used to compose a summary, which can be configured in terms of a percentage of the size of the original text. This version works both on English and Romanian languages.

The complex version summarizer_v2

Two types of summaries are generated: general and focused.

For the general summary, from the discourse tree presented as output of the previous module, the Discourse Parser (see DiscourseParser-UAIC documentation), the vein expression of the root node of the discourse tree represents a summary of the corresponding text. The vein expression of a discourse unit *u* is a list of discourse units, including *u*, which is meant to express the sequence of units that are significant to understand *u* in the context of the whole discourse. The vein expression of the root node is a sequence of units relevant for the whole text, thus for the general summary.

The combination of veins of all units which contain a particular noun phrase (NP) give a summary focused on that particular NP. A trained model is made available for English.

TECHNICAL INFORMATION

Software dependencies and system requirements

The first version requires a working Internet connection and a compatible browser (Internet Explorer, Chrome, Opera, Safari).

The second version is written in Java and requires the following settings to run:

1. Java VM any newer than 1.7

2. 1+ GB RAM

3. A trained language model (English version included as part of the archive).

Installation

No installation required, just unpack the archives.

Execution instructions

The simple version summarizer_v1

Access http://profs.info.uaic.ro/~ipistol/ALPE/res/rez/info.php and complete the form.

The complex version summarizer_v2

In the "Summarizer-v2" directory edit "config.cfg" to indicate the input and output files, then run "runExtractSummary.jar".

Input/Output data formats

Input data formats

The simple version summarizer_v1

Input is UTF8 text.

The complex version summarizer_v2

Input is a bin file produced by the DiscourseParser-UAIC component. Also, in the "config.cfg" file the user should specify for which NP the focused summary should be produced.

Output data format

The simple version summarizer_v1

Output is produced as UTF8 text.

The complex version summarizer_v2

Output is composed of two files: one general summary (summary.txt) and one focused (summary_NP.txt). They are both text files with one discourse unit per line.

Integration with external tools

The application does not need any external tool. It integrates the Clause Segmenter and the Discourse Parser.

3. CONTENT INFORMATION

Examples of input files and the corresponding output are included for both versions of the summarizer. For summarizer-v1 (with 30% selected as percentage form original text) they are “inputRO.txt” and “outputRO.txt”, for summarizer-v2 they are “tree.bin” and “summary.txt”, “summary_NP.txt” – found in the Summarizer-v2 directory.

Running this example on a quad core machine with 4 GB RAM and U-Compare 1.13 would take 0.1 seconds for summarizer-v1 and 0.2 for summarizer-v2.

4. ADMINISTRATIVE INFORMATION

Contact

For further information and technical support installing and/or running this tool, please email to Daniel Anechitei (daniel.anechitei@info.uaic.ro) and Ionut Pistol: ipistol@info.uaic.ro.

5. REFERENCES

Pistol I.C. "Romanian Processing Chains in Metanet4U", Proceedings of the 8th International Conference "Linguistic Resources and Tools for Processing of the Romanian Language". 8-9 December 2011, Ed. M. Moruz, D. Cristea, D. Tufis, A. Iftene, H Teodorescu, "Al. I. Cuza" University publishing House, ISSN 1843-911X

Cristea, D. (2005): Motivations and Implications of Veins Theory. Bernadette Sharp (Ed.). Natural Language Understanding and Cognitive Science, Proceedings of the 2nd International Workshop on Natural Language Understanding and Cognitive Scienc3, NLUCS 2005, in conjunction with ICEIS 2005, Miami, U.S.A., May 2005, INSTICC Press, Portugal, ISBN 972-8865-23-6X, pp. 32-44

Cristea, D., Postolache, O., Pistol, I. (2005): Summarisation through Discourse Structure. In Alexander Gelbukh (Ed.): Computational Linguistics and Intelligent Text Processing, 6th International Conference CICLing 2005, Mexico City, Mexico, February 2005, Proceedings, Springer LNSC, vol. 3406, ISBN 3-540-24523-5, pp. 632-644

Tokenizer-UAIC

1. BASIC INFORMATION

Tool name

Tokenizer-UAIC: Basic text tokenizer

Overview and purpose of the tool

The purpose of the tool is to mark individual lexical tokens from input text. The only corpus restriction is that it has to be UTF-8 encoded. The tools work for any Indo-European languages and any size of input document.

A short description of the algorithm

The algorithm identifies isolated punctuation marks (single space or other with at least one empty space before or after it) and marks it as token delimitation.

2. TECHNICAL INFORMATION

Software dependencies and system requirements

The original Java tool was updated and integrated in UIMA/U-Compare. It requires the following settings to run:

1. Java VM any newer than 1.5
2. 1+ GB RAM
3. The U-Compare platform, available at http://u-compare.org/

Installation

The tool is available as U-Compare component and is installed using the U-Compare interface following the instructions available at http://u-compare.org/userguide/Creating_Workflow.html#SECTION00043000000000000000 .

Execution instructions

Create a blank workflow in U-Compare, add either “Input Text Reader” or “File System Collection Reader”, then the “UAICTokenizerDescriptor” component. Input or select input files, then use the run button in the U-Compare interface.

Input/Output data formats

Input data formats

The input text or directory will contain only UTF8 encoded data, otherwise the output may change non-UTF8 characters.

Output data format

U-Compare can output several formats, the standard U-Compare output being a UTF-8 XML as shown below.

<?xml version="1.0" encoding="UTF-8"?>

<Document>

<org.apache.uima.examples.SourceDocumentInformation sofa="Sofa" begin="0" end="261" uri="file:/C: U-Compare/temp/interactive..." offsetInSource="0" documentSize="261" lastSegment="true">

<uima.tcas.DocumentAnnotation sofa="Sofa" begin="0" end="261" language="en">

<org.u_compare.shared.syntactic.Token sofa="Sofa" begin="0" end="9" fragments="null" metadata="null">President</org.u_compare.shared.syntactic.Token>

…

</uima.tcas.DocumentAnnotation>

</org.apache.uima.examples.SourceDocumentInformation>

</Document>

Integration with external tools

The application does not need any external tool (other than the U-Compare platform).

3. CONTENT INFORMATION

An example of an input file and the corresponding output is included in the archive and their names are input.txt and output.xml.

Running this example on a quad core machine with 4 GB RAM and U-Compare 1.13 would take 0.2 seconds (including writing the output XML).

4. ADMINISTRATIVE INFORMATION

Contact

For further information and technical support installing and/or running this tool, please email to Ionut Pistol: ipistol@info.uaic.ro.

5. REFERENCES

Pistol I.C. "Romanian Processing Chains in Metanet4U", Proceedings of the 8th International Conference "Linguistic Resources and Tools for Processing of the Romanian Language". 8-9 December 2011, Ed. M. Moruz, D. Cristea, D. Tufis, A. Iftene, H Teodorescu, "Al. I. Cuza" University publishing House, ISSN 1843-911X

RACAI - Romanian Academy

Endogenous resources

Mapping list from PWN2.0 to PWN3.0

1 BASIC INFORMATION

1.1 Resource composition

PWN 3.0-2.0 Concept Mapping is a resource containing the correpondences between Princeton WordNet (Fellbaum, 1998) concepts / synsets for versions 3.0 and 2.0. These concepts / synsets are encoded as Inter Lingual Indexes (ILIs), which are 8-digit numbers (e.g.: 00001740).

1.2 Representation of the resource (flat files, database, markup)

This resource contains 4 files, each corresponding to the existing Part-of-Speech (POS) categories (nouns, verbs, adjectives, adverbs). Each file contains raw text.

1.3 Character encoding

Character Encoding is ANSI.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)

NAME: Dan TUFIŞ / Dan ŞTEFĂNESCU
ADDRESS: Calea 13 Septembrie, No. 13, CASA ACADEMIEI, Bucharest 050711, ROMANIA
AFFILIATION: Research Institute for Artificial Intelligence, Romanian Academy
POSITION: Director / Researcher
TELEPHONE: +40213 188 103
FAX: +40213 188 142
E-MAIL: tufis@racai.ro / danstef@racai.ro

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

This resource can be downloaded from any of the links below:
http://nlptools.racai.ro/nlptools/index.php?page=pwn3to2
http://www.racai.ro/ResearchActivity/WebServicesandResources/WordNetMappings/tabid/138/Default.aspx

2.3 Copyright statement and information on IPR

This is a Free resource.

3 TECHNICAL INFORMATION

3.1 Directories and files

Folder *PWN_3.0-2.0_Concept_Mapping* has 4 files, each containing PWN 3.0 to PWN 2.0 concept mapping for the existing POS categories:

- adj.txt
- adv.txt
- noun.txt
- verb.txt

3.2 Data structure of an entry

An entry has the following structure:

ILI_PWN_3.0 <tab> ILI_PWN_2.0

An ILI is the Inter Lingual Index encoding a unique PWN concept / synset.

3.3 Resource size (no. of rules, MB occupied on disk)

Resource size is detailed by the following table:

POS Category	File Name	Disk Space	Concepts mapped (lines)
Adjectives	adj.txt	330.7 KB	17,828
Adverbs	adv.txt	66.6 KB	3,594
Nouns	noun.txt	1.4 MB	79,647
Verbs	verb.txt	250.3 KB	13,492
Total		2MB	114,561

Table 1: Resource size for all POS categories

4 CONTENT INFORMATION

4.1 Type of the resource

This is a language-dependent resource. It works for ENGLISH PWN.

4.2 The natural language(s) for the resource is applicable (if language dependent)

This resource refers to ENGLISH PWN concepts / synsets.

4.3 Domain(s)/register(s) of the corpus

This resource falls into General Domain.

4.4 Annotations in the corpus (if an annotated corpus)

No annotation is present.

4.5 Intended application of the resource

Update tools that work for PWN 2.0

4.6 Reliability of the annotations (automatically/manually assigned) – if any

This resource has been created using both automatic and manual operations. After the automatic step, which involved finding identical definitions, similar hyper/hypo-nims, similar sense numbers for the literals, etc., this resource was compared with a similar one, automatically developed by the UPC NLP Research Group[30]. All common mappings were considered correct and the differences were manually analyzed by linguists. During this analysis, the incorrect mappings were removed (Tufiş et al., 2011).

5 REFERENCES

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Tufiş, D., Ion, R., Mititelu, V., Irimia, E., Ştefănescu, D., Mihăilă, C. (2011). Extending and completing the Ro-

WordNet lexical ontology by eliminating the existing semantic conflicts and by validating the differential semantics model based on Ro-WordNet. Academic report (in Romanian). Bucharest, Romania.

NAACL 2003

1 BASIC INFORMATION

1.1 Corpus composition

The English-Romanian parallel corpus NAACL 2003 supplied in Batch 2 of the MetaNet4U project consists of the News part of the word alignment traning corpus from the HLT-NAACL 2003 workshop "Building and Using Parallel Texts: Data Driven Machine Translation and Beyond" (Mihalcea and Pedersen, 2003). It was collected in 2003, from Romanian newspapers archives with English versions of the published articles. It was automatically sentence aligned and then manual validation was performed on the resulting sentence alignments.

1.2 Representation of the corpus

The corpus is encoded in the XCES XML format (http://www.xces.org/), with one XML file per language.

1.3 Character encoding

The characters are UTF-8 encoded.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Dan Tufiş
Address: Calea 13 Septembrie, no. 13, 050711
Affiliation: Research Institute for Artificial Intelligence, Romanian Academy
Position: Director
Telephone: +40 21 3188103
Fax: +40 21 3188142
e-mail: tufis@racai.ro

2.2 Delivery medium

The resource will be uploaded on the MetaShare platform as an archive.

[30] http://nlp.lsi.upc.edu/web/index.php?option=com_content&task=view&id=21&Itemid=59

2.3 Copyright statement and information on IPR

The resource is free, license-based, for research purposes.

3 TECHNICAL INFORMATION

3.1 Directories and files

The corpus consists of two XML files: 'NAACL_news-en.xml' and 'NAACL_news-ro.xml'.

3.2 Data structure of an entry

The corpus is structured in paragraphs, divided into sentences. Each sentence is segmented into tokens, including punctuation. Each token has a descriptor attribute ('msd') containing syntactic information about its grammatical category and its morpho-lexical attributes, and a base form attribute ('base') containing the lemma.

3.3 Corpus size

The corpus contains 39956 sentence pairs. There are 843,832 words (no punctuation) in English and 757,550 words in Romanian. The English file (XML encoded) has 58.4MB and the Romanian file has 54.6MB.

4 CONTENT INFORMATION

4.1 Type of the corpus

Parallel corpus, written, annotated, XCES-encoded.

4.2 The natural language(s) of the corpus

English and Romanian.

4.3 Domain(s)/register(s) of the corpus

The corpus is automatically collected from the web, from Romanian newspapers archives having English translations. The domain of the corpus is "News" with subdomain "Political News".

4.4 Annotations in the corpus

4.4.1 Types of annotations

The corpus is sentence split, tokenized, POS-tagged (with the MSD tagset in both English and Romanian, see http://nl.ijs.si/ME/V3/msd/msd.pdf) and lemmatized. **The POS tagging for both English and Romanian was checked by hand at the Research Institute for Artificial Intelligence of the Romanian Academy** and thus this is a manually validated POS-tagged corpus. It is suitable for training POS tagging models.

4.4.2 Tags

We used the Multext-East Morpho-Syntactic Descriptors to POS tag this corpus. For further information about MSDs, one can refer to http://nl.ijs.si/ME/V3/msd/msd.pdf.

4.4.3 Alignment information

The corpus contains 39956 sentence pairs. The alignment of the sentences is implicit, i.e. there is no external alignment file.
Every sentence (e.g. '<xces:s id="NAACL_2003_en_1">…</xces:s>') in the English file has a unique integer identifier (1) which corresponds to the parallel sentence in the Romanian file ('<xces:s id="NAACL_2003_ro_1">…</xces:s>').

4.4.4 Attributes and their values

Each token has three attributes: 'type' which can be 'word' or 'punctuation', 'base' which contains the lemma of the token (if the token is of type 'punctuation', then the value of the base attribute is the punctuation string) and 'msd' which contains the correct MSD of that token in context.

4.5 Intended application of the corpus

Being manually validated, the corpus may be used to train POS tagging/lemmatization statistical models. It can also be used to extract English-Romanian translation equivalents and to study word alignment algorithms. It is too small for being useful in deriving n-gram langauge models but parts of it can constitute test/development sets for SMT.

4.6 Reliability of the annotations

The POS annotations are reliable in that the POS tagging has been checked and corrected by hand. Initially, we ran the POS tagging training and testing on this corpus for both English and Romanian doing POS label corrections where necessary, until the tagger's (a variant of the TnT POS tagger (Brants, 2000) called TTL (Tufiș et al., 2008)) accuracy was above 99%. After that, we manually checked every pair of sentences in order to see if the POS tagging is correct.
Lemmatization is also reliable in that every lemma is assigned from a lexicon containing for every word form, its correct lemma and MSD label. The ambiguity cases where for the same MSD label we had different lemmas, were resolved by hand.

5 REFERENCES

Brants, T. (2000). TnT – A Statistical Part-Of-Speech Tagger. In *Proceedings of the 6th Applied NLP Conference ANLP-2000*. Seattle, WA, pp 224--231.

Mihalcea, R. and Pedersen, T. An Evaluation Exercise for Word Alignment. In *Proceedings of the HLT/NAACL Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, Canada, May 2003.

Tufiș, D., Ion, R., Ceauşu, A., and Ştefănescu, D. RACAI's Linguistic Web Services. In *Proceedings of the 6th Language Resources and Evaluation Conference* - LREC 2008, Marrakech, Morocco, May 2008. ELRA - European Language Ressources Association. ISBN 2-9517408-4-0.

ROMORPH

1 BASIC INFORMATION

1.1 Resource composition

The paradigmatic morphology of Romanian has been developed for several years in different formats and variants (Tufiș, 1989), the most complete being implemented in the LISP-based ELU linguistic programming environment (Estival et al., 1994).

This unification-based implementation of the paradigmatic morphology, together with lexical repositories, associating paradigms and lexical roots to almost 35.000 of Romanian lemmas, was documented in a flat (theory neutral) attribute-value representation (FAVR, see Tufiș, Barbu, 1997).

In the context of the paradigmatic morphology theory, a word is treated as an entity made of two fundamental units: a *root* and an *ending* (built of one or more desinences and/or suffixes). The root usually carries context-free information, while the ending is a bearer of contextual information.

Some contextual information - consisting of restrictions on its use in conjunction with the specified endings - can be associated with the root if there is root alternation (for the same lemma and the same part-of-speech, the different inflected forms can share two or more roots).

The information associated with the root is stored in a dictionary (lexical repository) entry corresponding to the lemma of the corresponding root. Such an entry has the following structure:

pos
@lemma
root_1 root_2 … root_k associated_paradigm1
root_k+1 … associated_paradigm2
…

The information associated with the ending is stored in ROPMORPH, the file containing a complete inventory of the Romanian paradigms for verbs, nouns, pronouns, articles and adjectives.

Any lemma can be associated to one or more inflectional paradigms. An inflectional paradigm is a tree structure that identifies all the legal endings (and the associated restrictions) which can be associated to a root (or more roots) of a given lemma.

For a detailed description see 5.b and 5.c.

1.2 Representation of the resource

XML markup

1.3 Character encoding

The characters are UTF8 encoded

2 ADMINISTRATIVE INFORMATION

2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)

Name: Dan Tufis,
Address: Calea 13 Septembrie, no. 13, 050711
Affiliation: Research Institute for Artificial Intelligence, Romanian Academy
Position: Director
Telephone: +4021 3188103
Fax: +40 21 3188142
e-mail: tufis@racai.ro

2.2 Delivery medium

The resource will be uploaded on the MetaShare platform as an archive.

2.3 Copyright statement and information on IPR

The resource is free, license-based, for research purposes and fee license-based for commercial purposes.

3 TECHNICAL INFORMATION

3.1 Directories and files

A single xml file, named *morphaltUTF8.xml*

3.2 Data structure of an entry

```
<PARADIGM PARADIGM=value of type string >
  <TYPE TYPE= value of type string >
   <NUM NUM = value of type string >
    <ENCL ENCL= value of type string >
     <CASE CASE= value of type string >
      <TERM TERM= value of type string  ALT = value of type string />
     </CASE>
    </ENCL>
   </NUM>
  </TYPE>
  </PARADIGM
```

3.3 Resource size (nmb. of rules, MB occupied on disk)

286 entries: 286 paradigms for nouns, verbs, adjective, articles, pronouns.
Compressed:18k; uncompressed:603k.

4 CONTENT INFORMATION

4.1 Type of the resource (language (in(dependent)

Language dependent

4.2 The natural language for the resource

Romanian

4.3 Domain(s)/register(s) of the resource

Not applicable

4.4 Annotations in the resource

4.4.1 Types of annotations

Morphological Annotation

4.4.2 Tags (if POS/MSD/TIME/discourse/etc –tagged or parsed)

The representation is in the form of the attribute-value pairs.

4.4.3 *Attributes and their values (if annotated)*

- *Attributes for the PARADIGM tag*
  PARADIGM = the name of the paradigm
  CAT=”n/adv”) GEN="masculine/feminine”
  TYPE="manner/common"(“manner” for adverbs and “common” for nouns)
  INTENSIFY="none/diminutive/augmentative"
- *Attributes for the TYPE tag*
  TYPE= "{proper common}/common/proper"
- *Attributes for the NUM tag*
  NUM="singular/plural"
- *Attributes for the CASE tag*
  CASE="{nominative/genitive/ dative/accusative/vocative}">
  Attributes for the HUM (human) tag
  HUM=”imperson/person”
  CASE="{nominative/genitive/ dative/accusative/vocative}">
- *Attributes for the ENCL tag*
  ENCL="no/yes"
- *Attributes for the TERM tag*
  TERM=string; this string is an ending for a morphological form in Romanian
  ALT =”1/2” for nouns and adjectives and ALT=”1/2/3/4/5/6/7/8/9” this shows to a morphological generator on which alternative root to apply the ending in TERM
- *Attributes for the VOICE tag*
  VOICE= “{active reflexive}/active/passive”
  GEN="masculine" – for the active VOICE

NUM="singular" –for the active VOICE

- ü *Attributes for the TENSED tag*
  TENSED="yes/no" PRD="yes/no"
  For TENSED="yes"→ PRD="no" MOOD="infinitive" TENSE="present"
- ü *Attributes for the MOOD tag*
  MOOD="indicative/conjunctive/imperative/participle/gerund/supine/infinitive"
- ü *Attributes for the TENSE tag*
  TENSE="present/imperfect/simpleperfect/pastperfect"
- ü *Attributes for the PERS tag*
  PERS="1/2/3"

4.5 Intended application of the resource

This is a very reliable resource to be used in implementing a Romanian morphological generator. We also used this resource in developing a procedure of lexical acquisition for Part-of-Speech tagging. (see 5.d, 5.e)

4.6 Reliability of the annotations

Manually assigned.

5 RELEVANT REFERENCES AND OTHER INFORMATION

a. D.Tufiş. "It Would Be Much Easier If WENT Were GOED", in Proceedings of the 4th European Conference of the Association for Computational Linguistics, Manchester, 1989

b. Dominique Estival, Dan Tufiş, Octavian Popescu. 1994. Devélopment d'outils et des données linguistiques pour le traîtement du langage naturel. Rapport Final – Projet EST (7RUPJO38421) ISSCO, Geneve, September 1994

c. Dan Tufiş Barbu A.M. "A Reversible and Reusable Morpho-Lexical Description of Romanian". In Dan Tufiş, Andersen P. (eds.) "Recent Advances in Romanian Language Technology", Editura Academiei, 1997.

d. Elena Irimia. 2007. RomGen - A Romanian morphological generator, essential component of an English-to-Romanian MT system. In proceedings of the Doctoral Consortium, EUROLAN 2007 Summer School, Iaşi, România, July 23 - August 3, 2007 pp. 54-58.

e. Elena Irimia. 2007. ROG - A Paradigmatic Morphological Generator for Romanian. In proceedings of "The 3rd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics", October 5-7, 2007, Poznań, Poland, pp. 408-412, ISBN 978-83-7177-407-2.

f. Dan Tufiş, Radu Ion, Elena Irimia şi Alexandru Ceauşu. 2007. Achiziţie lexicală nesupervizată pentru adnotare morfo-lexicală. Atelierul de Lucru "Resurse Lingvistice Româneşti şi Instrumente pentru Prelucrarea Limbii Române", 14-15 decembrie 2007, Iaşi, România.

g. Dan Tufiş, Elena Irimia, Radu Ion, Alexandru Ceauşu. 2008. Unsupervised Lexical Acquisition for Part of Speech Tagging. In Proceedings of LREC 2008 (Language Resources and Evaluation Conference), May 26 - June 1, Marrakech, Morocco. ELRA - European Language Resources Association. ISBN: 2-9517408-4-0.

RO-WORDNET (part 2)

1 BASIC INFORMATION

1.1 Resource composition

Ro-WordNet (RoWN) (Tufiș et al., 2008; Tufiș et al., 2004b) is a lexical ontology following the Princeton WordNet (PWN) (Miller, 1990; Fellbaum, 1998) organizational principles. The synsets in RWN are aligned with PWN3.0 (Tufiș et al., 2012) and, additionally, they are associated with SUMO/MILO (Neal & Pease, 2001) concepts and labeled with DOMAINS3.0 (Bentivogli et al., 2004) categories.

1.2 Representation of the resource (flat files, database, markup)

RoWN is distributed as an XML file, observing the structure of EuroWordNet (Vossen, 1998) and BalkaNet wordnets (Tufiș, 2004; Tufiș et al., 2004a). The file can be loaded and browsed in VisDic (Horak & Smrz, 2004) (as well as in its descendant versions), the official editor and browser of the BalkaNet project.

1.3 Character encoding

The characters have been encoded in UTF8.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)

NAME: Dan TUFIŞ

ADDRESS: Calea 13 Septembrie, No. 13, CASA ACADEMIEI, Bucharest 050711, ROMANIA

AFFILIATION: Research Institute for Artificial Intelligence, Romanian Academy

POSITION: Director

TELEPHONE: +40213 188 103

FAX: +40213 188 142

E-MAIL: tufis@racai.ro / danstef@racai.ro

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive. It can be downloaded from: http://nlptools.racai.ro/nlptools/index.php?page=rown

2.3 Copyright statement and information on IPR

The resource is free license-based for research purposes and fee license-based for commercial purposes.

3 TECHNICAL INFORMATION

3.1 Directories and files

RoWordNet_3.0 – the directory containing the following files:

- wnrom_3_v2.0.xml – the proper Romanian WordNet file

- .\VisDic_cfg_file\wnrom_3_v2.0.cfg – the VisDic configuration file for RoWN

The VisDic editor and browser (if needed) can be freely downloaded from http://nlp.fi.muni.cz/projects/visdic/.

3.2 Data structure of an entry

The structure of an entry in RWN is exemplified below:

<SYNSET>

<ID>ENG30-xxxxxxxx-C </ID>

<POS>cat</POS>

<SYNONYM>

[<LITERAL>literal

<SENSE>k</SENSE>

</LITERAL>]$^{+}$

</SYNONYM>

<DEF> a definition </DEF>

[<BCS>n</BCS>]

[<ILR>synset-ID<TYPE>name-of-relation</TYPE></ILR>]$^{+}$

[<DOMAIN>a domain</DOMAIN>]$^{+}$

[<SUMO>a sumo-concept<TYPE> a type of mapping</TYPE></SUMO>]

<\SYNSET>

The structure of an entry for a non-lexicalized synset is the following:

<SYNSET>

<ID>ENG30-xxxxxxxx-C </ID>

<POS>cat</POS>

<NL>yes</NL>

<SYNONYM></SYNONYM>

<DEF> a definition </DEF>

[<BCS>n</BCS>]

[<ILR>synset-ID<TYPE>name-of-relation</TYPE></ILR>]$^{+}$

[<DOMAIN>a domain</DOMAIN>]$^{+}$

[<SUMO>a sumo-concept<TYPE> a type of mapping</TYPE></SUMO>]

<\SYNSET>

3.3 Resource size (no. of rules, MB occupied on disk)

The current (validated) version contains 59,015 synsets, with the following distribution:

Noun synsets	Verb synsets	Adj. synsets	Adv. synsets	Total
41,063	10,064	4,822	3,066	59,015

The needed disk space is about 29.3 Mb.

4 CONTENT INFORMATION

4.1 Type of the resource

This is a language-dependent resource (Romanian).

4.2 The natural language(s) for the resource is applicable (if language dependent)

The language of the lexical ontology is Romanian. Via alignment with PWN, it is virtually a bilingual English-Romanian dictionary.

4.3 Domain(s)/register(s) of the corpus

This resource falls into General Domain.

4.4 Annotations in the corpus (if an annotated corpus)

There are four types of entries, all of them having the same structure: entries for nouns, for verbs, for adjectives and for adverbs.

See section 3.2:

The value of the <ID> tag is a unique identifier for the aligned synset in PWN3.0 (the numerical value is the offset of the respective synset in the PWN database). The trailing character C in the ID value is one of N, V, A, R.

The value of the <POS> is one of the N, V, A, R (identical to the character C) identifying the part of speech of the literals in the current synset. One should notice that in the Romanian wordnet the adjectival satellites (marked with the category S in PWN) are included into the A category.

Under the tag <SYNONYM> there are one or more <LITERAL> immediately followed by a sense number. Unlike in PWN, here the numbering is not related to the frequency of the respective sense of the literal, but it follows the numbering conventions from the Romanian Explanatory Dictionary (DEX), the reference dictionary by the Romanian Academy. In the case of non-lexicalized concepts, the tag <SYNONYM> is empty.

The tag <DEF> marks up the definition from DEX. In some cases (namely when the respective sense was not documented in DEX, the definition is a professional translation of the corresponding PWD definition).

The <BCS> tag is optional and it marks up the so called base concept synsets. The value of the tag is 1, 2 or 3, according to what was called in BalkaNet BCS1, BCS2 and BCS3 synsets (see Tufiș et al, 2004).

The current synset entry contains one or more relations towards other synsets. This information is encoded as: [<ILR>synset-ID<TYPE>name-of-relation</TYPE></ILR>]+ where the <ILR> tag (Internal Language Relation) uniquely identifies the target synset of the relation specified by the tag <TYPE>. The relations are transferred from PWN3.0.

The tag <DOMAIN> is one of the labels specified by the DOMAINS-3 taxonomy and it was imported from the PWN3.0 via synset alignment, as well.

The tag <SUMO> marks up the SUMO/MILO concept transferred from the SUM/MILO - PWN3.0 alignment via PWN-RWN synset alignment. The tag <TYPE> embedded into content of the <SUMO> tag describes the type of mapping: "=" defining exact mapping and "+" defining an approximate mapping (the SUMO concept is more general than the meaning of the current entry.

The <NL> tag signals the non-lexicalized concepts in Romanian. For them there is no literal in the <SYNONYM> tag, but there is a gloss.

The design procedure of the RWN followed the *conceptual density principle* (Tufiș et al., 2004) in a top-down strategy and the literals chosen for implementation were selected on the basis of frequency and definitional productivity (the number of entries in DEX definitions containing the specific literals). The lexical stock covers the basic general language vocabulary of Romanian.

4.5 Intended application of the resource

The lexical ontology has been used in practically all NLP-enhanced applications developed at RACAI: tagging, lemmatization, word-sense disambiguation (Tufiş et al., 2004c), word alignment, collocation extraction, document classification, question-answering, machine translation and others (Tufiș & Barbu, 2004).

4.6 Reliability of the annotations (automatically/manually assigned) – if any

The lexical ontology has been based on several reference published dictionaries: Explanatory Dictionary of Romanian, Dictionary of Synonyms, Dictionary of Antonyms. The mapping to the translation equivalent synsets from Princeton WordNet has been manually done by experienced lexicographers and NLP researchers. Based on the manual synset alignment, the semantic relations have been automatically transferred from PWN onto RWN, while the lexical relations were transferred (when it was possible) under the validation of a lexicographer.

5 REFERENCES

References on the Romanian WordNet:

Dan Tufiş, Verginica Barbu Mititelu, Dan Ştefănescu, Radu Ion. (2012). *The Lexical Ontology for Romanian*. In Nicoletta Calzolari and Nancy Ide (eds.): Language Resources and Evaluation, Special Issue on Wordnets, Springer 2012, ISSN 1574-020X.

Dan Tufiş, Radu Ion, Luigi Bozianu, Alexandru Ceauşu, Dan Ştefănescu. (2008). *RO-Wordnet*. In Proceedings of the 4th Global WordNet Association Conference, Szeged, Hungary.

Dan Tufiş (ed.). (2004). *Special Issue on BalkaNet*, Romanian Academy, vol7, no. 2-3, ISSN 1453-8245 .

Dan Tufiş, D. Cristea, S. Stamou. (2004a). In *BalkaNet: Aims, Methods, Results and Perspectives. A General Overview*. In Romanian Journal on Information Science and Technology, Dan Tufiş (ed.) Special Issue on BalkaNet, Romanian Academy, vol7, no. 2-3, 2004, pp. 9-34, ISSN 1453-8245

Dan Tufiş, Eduard Barbu, Verginica Barbu Mititelu, Radu Ion, Luigi Bozianu. (2004b). *The Romanian Wordnet*. Romanian Journal on Information Science and Technology, vol. 7, no. 2-3 (pp. 107-124).

Dan Tufiş, Radu Ion, Nancy Ide. (2004c). *Word sense disambiguation as a wordnets validation method in Balkanet*. In Proceedings of the 4th LREC Conference, Lisbon, 741-744; 1071-1074

Dan Tufiş, Eduard Barbu. (2004). *A Methodology and Associated Tools for Building Interlingual Wordnets*. In Proceedings of LREC2004, Lisbon, Portugal (pp. 1067-1070).

References to Princeton WordNet, DOMAINS taxonomy and Sumo/MILO ontology:

Horak, A., Smrz, P. (2004). *VisDic - Wordnet Browsing and Editing Tool*. In *Proceedings of the Second International WordNet Conference - GWC 2004*. Brno, Czech Republic : Masaryk University, 2003. pp. 136-141. ISBN 80-210-3302-9.

Bentivogli, L., Forner, P., Magnini, B., Pianta, E. (2004). *Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing*. Proceedings of COLING 2004 Workshop on "Multilingual Linguistic Resources" (pp. 101-108).

Niles, I., Pease, A. (2001). Towards a Standard Upper Ontology. Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (pp. 2-9).

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. (1990). Introduction to WordNet: An On-Line Lexical Database. International Journal of Lexicography, Vol. 3, No. 4 (pp. 235-244).

Fellbaum, Ch. (Ed.) (1998). WordNet: An electronic lexical database. Cambridge, MA: MIT Press.

Vossen, P. (Ed.) (1998). EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Dordrecht: Kluwer Academic Publishers.

Endogenous resources (tools)

COLLOC

1 BASIC INFORMATION

1.1 Tool name

The tool is called **Collocation Extractor**.

1.2 Overview and purpose of the tool

Collocation Extractor identifies and extracts collocations along with their contexts of occurence from a given preprocessed text. The tool is a stand-alone application developed in C#.

1.3 A short description of the algorithm

The algorithm and different studies on its performance have been described in several papers: Ştefănescu et al. (2006), Todiraşcu et al. (2007), Ştefănescu et al. (2008), Todiraşcu et al. (2009), Ştefănescu (2010).
In this approach, we considered a collocation to be an expression formed by 2 principal content words which satisfy the following constraints:
– the distance between them is relatively constant;
– they appear together more often than expected by chance: Log-Likelihood.
Looking at this definition, one can notice, that from a strict linguistic point of view, such a construction can be seen as a strong co-occurrence, rather than a collocation.
The first component of our solution is based on a method developed by Smadja (1993). This uses the average and the standard deviation computed on distances between words to identify pairs of words that regularly appear together at the same distance, a fact which is considered to be the manifestation of a certain relation between those words. Collocations can be found by looking for such pairs for which standard deviation is small.
In order to find certain types of collocations, the application allows for POS (Part Of Speech) filtering, computing the standard deviation only for pairs having certain POS-es within a user-defined window of non-functional words. It stores all the pairs for which standard deviation is smaller than a user-defined threshold. According to Manning and Schütze (1999), a good value for this threshold is 1.5. This method can identify good candidates for multi-word expressions but not good enough. *Collocation Extractor* further filters out some of the pairs in order to keep only those composed by words which appear together more often than expected by chance. This is done by computing the Log-Likelihood scores for all the above obtained pairs and keeping only those above a user-defined threshold.
This tool is language-independent and can be also used for finding multi-word terminological expressions (Ştefănescu, 2012).

2 TECHNICAL INFORMATION

2.1 Software dependencies and system requirements

This tool requires Windows machines with Microsoft .Net Framework 3.5 installed.

2.2 Installation

This tool requires Microsoft .Net Framework 3.5. No other installation is required.

2.3 Execution instructions

Run the executable *Collocation Extractor.exe* and select the menu entry '*Collocations*'.

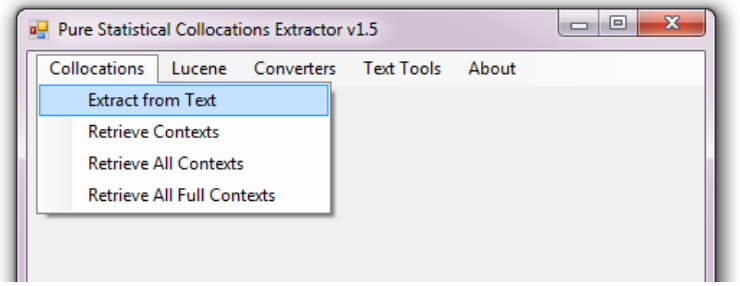


Figure 1: *Collocation Extractor* menu options

In order to extract the collocations from a text, the user needs to select the menu item '*Extract from Text*'. A configuration window will appear, allowing the user to set the parameters of collocation extraction.

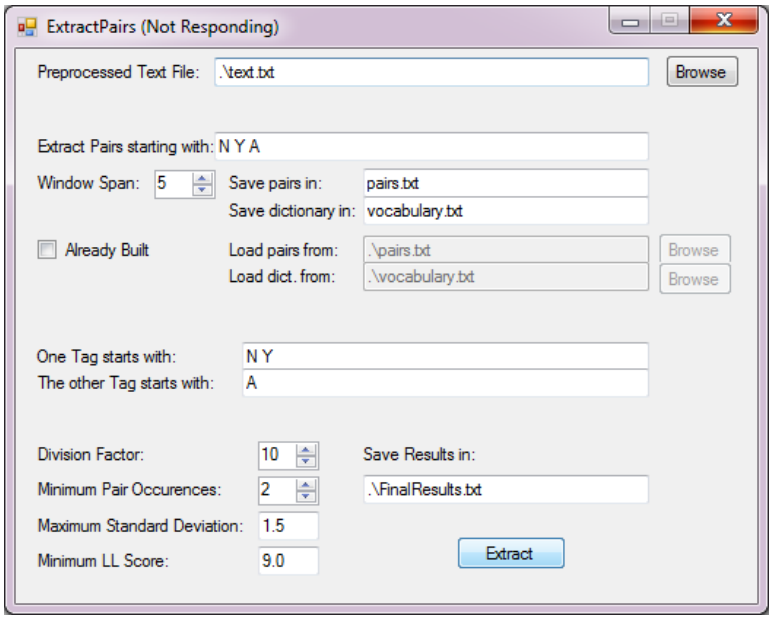


Figure 2: Configuration window for collocations extraction

The user is required to input the path to the preprocessed text file and to define the parameters of the extraction process:

- the POS-es of the words the application should take into account (*Extract Pairs starting with*). In the example given in Fig. 2, the user intends to look only at nouns (all words having the POS-tag starting with 'N' and 'Y' (proper nouns)) and adjectives (all words having the POS-tag starting with 'A'). Obviously, this field should be set according to the tagset used for pre-processing the text.

- the *Window Span* of content words which is going to be considered (e.g.: a window span of 5 (see Fig. 2) means a window of 11 content words having the current considered content word in the middle);

- the intermediary file names which will contain all word pairs and vocabulary satisfying the user constraints. If these files already exist, the user has the possibility of checking the box '*Already Build*' and give the path to these files;

- the POS tags a pair should be formed of (*One / The other Tag starts with*). In the example given in Fig. 2, the user searches for noun(N, Y)-adjective(A) pairs. It practice, it does not matter which one is first, since negative distances are also considered.

- division factor (default value is 10) refers to the number of parts in which the application divides the problem in order to consume less memory and still be efficient. This is similar to MapReduce algorithm, yet this is not parallelized. It allows us to correctly determine the frequencies of the existing pairs without consuming too much memory. The division factor should be set depending on the size of the input file. The larger the file, the higher the division factor;

- the minimum number of occurrences for a pair in order to be taken into account (default value is 2);

- the maximum standard deviation allowed for a pair (default value is 1.5);

- the minimum Log-Likelihood pair allowed for a pair (default value is 9).

This step finds the word pairs which define the collocations we are looking for. In order to find the real collocations, one needs to extract the occurring expressions formed by these principal words. In order to do this fast, the application can be used to index the sentences of the text as documents by selecting the menu item '*Index text*' (see Fig. 3).

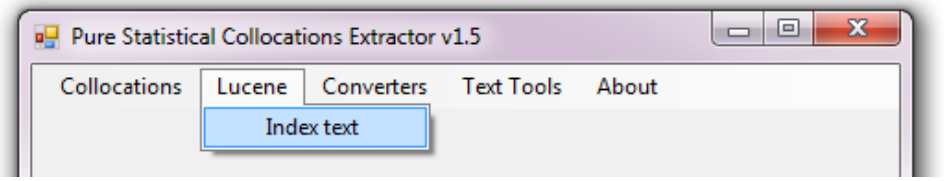


Figure 3: Indexing the input text with Lucene

After the index is created, the user can fully retrieve the collocational expressions and the general context in which they occur, by accessing the other options available in the menu entry '*Collocations*' (see Fig. 1).

Selecting the menu item '*Retrieve Contexts*' allows the user to get the context dependent data within the user interface. In this new window, the user is required to '*Load Data From*' the file containing the final results (*FinalResults.txt* in our example) and then double click a collocation from the left-side panel (see Fig. 4).

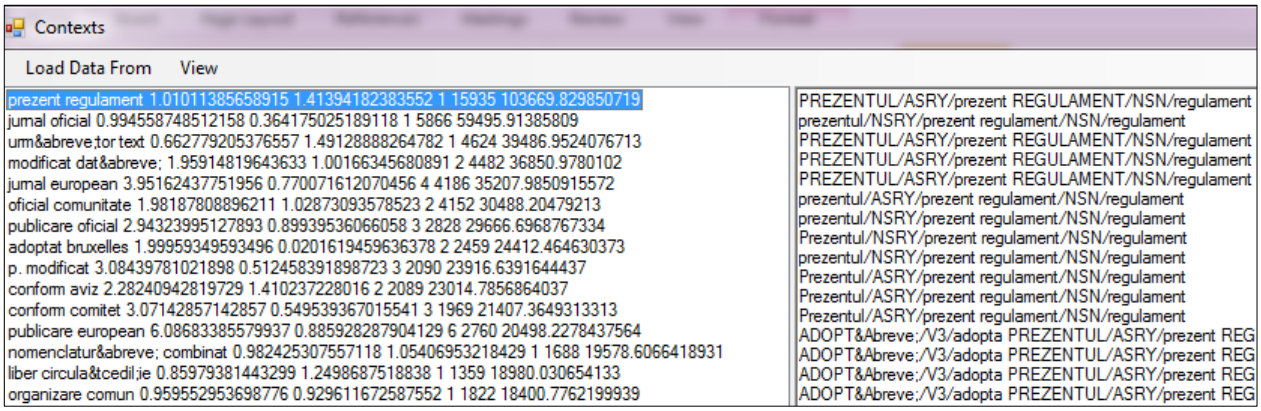


Figure 4: User interface allowing the user to see the occurences of the collocations in text

Selecting the menu item '*Expressions only*' from the menu entry '*View*' will compact all the data in the right panel into unique expressions and their frequencies (see Fig. 5).

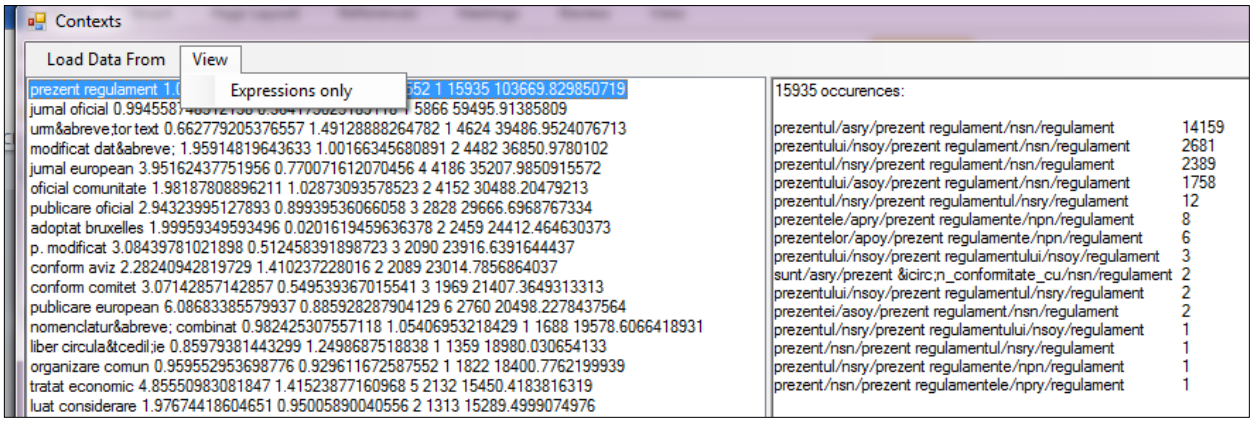


Figure 5: Compacted collocations

The user has also the option to save this data on disk by selecting the menu item '*Retrieve All Contexts*' from '*Collocations*' (see Fig. 1) (the user is required to '*Load Data From*' the file containing the final results (*FinalResults.txt* in our example)). The output would be similar to that in Fig 5. The application allows the user to also save the entire sentences containing the collocations through the option '*Retrieve All Full Contexts*' (the user is required to '*Load Data From*' the file containing the final results (*FinalResults.txt* in our example)).

In case the input file does not have the correct encoding, the application offers several possibilities of conversion (see Fig. 6) which were implemented due to practical reasons. Some other options are still under construction (see the menu entry '*Text Tools*').

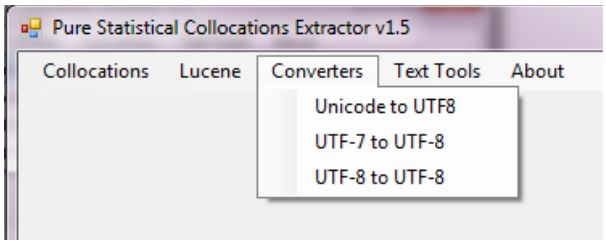


Figure 6: Options for encoding conversions of text

2.4 Input / Output data formats

The input should be a pre-processed text file of the following format (see Fig. 7):

word_form <tab> *POS-tag* <tab> *lemma*

```
investit&abreve;        ASN     investit
de      S       de
Tratat  NSN     tratat
.       PERIOD  .
Agen&tcedil;ia  NSRY    agen&tcedil;ie
este    V3      fi
reglementat&abreve;     ASN     reglementat
de      S       de
dispozi&tcedil;ile      NPRY    dispozi&tcedil;ie
Tratatului      NSOY    tratat
&scedil;i       CR      &scedil;i
ale     TP      al
prezentului     NSOY    prezent
statut  NSN     statut
.       PERIOD  .
```

Figure 7: Format of the input text

The output depends on the selected menu item:

Extract from Text:

The output file is a list of collocations ordered according to the LL score (see Fig. 8):

<*word_1*> <*word_2*> <*avg*> <*st_dev*> <*round*(*avg*)> <*freq*> <*LL*>

where:

<*word_1*> <*word_2*> are the principal words forming the collocation;
<*avg*> is the average distance at which the two words occur in text;
< *st_dev*> is the standard deviation from the average for the two words;
<*round*(*avg*)> is the actual distance used for this collocation. It is computed using *round* math function, since we need an integer for distance and *avg* is usually not an integer;
<*freq*> is the frequency of the pair;
<*LL*> is the Log Likelihood score for that pair.

```
jurnal oficial 0.994558748512158 0.364175025189118 1 5866 59495.91385809
urm&abreve;tor text 0.662779205376557 1.49128888264782 1 4624 39486.9524076713
modificat dat&abreve; 1.95914819643633 1.00166345680891 2 4482 36850.9780102
jurnal european 3.95162437751956 0.770071612070456 4 4186 35207.9850915572
oficial comunitate 1.98187808896211 1.02873093578523 2 4152 30488.20479213
publicare oficial 2.94323995127893 0.899395360660058 3 2828 29666.6968767334
adoptat bruxelles 1.99959349593496 0.0201619459636378 2 2459 24412.464630373
p. modificat 3.08439781021898 0.512458391898723 3 2090 23916.6391644437
conform aviz 2.28240942819729 1.410237228016 2 2089 23014.7856864037
conform comitet 3.07142857142857 0.549539367015541 3 1969 21407.3649313313
publicare european 6.08683385579937 0.885928287904129 6 2760 20498.2278437564
nomenclatur&abreve; combinat 0.982425307557118 1.05406953218429 1 1688 19578.6066418931
liber circula&tcedil;ie 0.85979381443299 1.2498687518838 1 1359 18980.030654133
organizare comun 0.959552953698776 0.929611672587552 1 1822 18400.7762199939
tratat economic 4.85550983081847 1.41523877160968 5 2132 15450.4183816319
luat considerare 1.97674418604651 0.95005890040556 2 1313 15289.4999074976
reglementare administrativ 2.07810499359795 0.998548334796929 2 1281 14589.118712215
```

Figure 8: Collocations list given as output

Retrieve All Contexts

The output is the same list as above, but each collocation is followed by the list of its real unique occurrences in the given text, along with their corresponding frequencies (see Fig. 9).

```
1. prezent regulament 1.01011385658915 1.41394182383552 1 15935 103669.829850719
------------------------------------------------------
prezentul/asry/prezent regulament/nsn/regulament        14159
prezentului/nsoy/prezent regulament/nsn/regulament      2681
prezentul/nsry/prezent regulament/nsn/regulament        2389
prezentului/asoy/prezent regulament/nsn/regulament      1758
prezentul/nsry/prezent regulamentul/nsry/regulament     12
prezentele/apry/prezent regulamente/npn/regulament      8
prezentelor/apoy/prezent regulamente/npn/regulament     6
prezentului/nsoy/prezent regulamentului/nsoy/regulament 3
sunt/asry/prezent &icirc;n_conformitate_cu/nsn/regulament       2
prezentului/nsoy/prezent regulamentul/nsry/regulament   2
prezentei/asoy/prezent regulament/nsn/regulament        2
prezentul/nsry/prezent regulamentului/nsoy/regulament   1
prezent/nsn/prezent regulamentul/nsry/regulament        1
prezentul/nsry/prezent regulamente/npn/regulament       1
prezent/nsn/prezent regulamentele/npry/regulament       1
######################################################
2. jurnal oficial 0.994558748512158 0.364175025189118 1 5866 59495.91385809
------------------------------------------------------
jurnalul/nsry/jurnal oficial/asn/oficial        5257
jurnal/nsn/jurnal oficial/asn/oficial   520
jurnalului/nsoy/jurnal oficial/asn/oficial      78
jurnale/npn/jurnal oficiale/apn/oficial 9
jurnalele/npry/jurnal lor/ps/lui oficiale/apn/oficial   2
jurnalele/npry/jurnal oficiale/apn/oficial      2
######################################################
3. urm&abreve;tor text 0.662779205376557 1.49128888264782 1 4624 39486.9524076713
------------------------------------------------------
urm&abreve;torul/asry/urm&abreve;tor text/nsn/text      4597
urm&abreve;toarele/apry/urm&abreve;tor texte/npn/text   26
urm&abreve;torului/asoy/urm&abreve;tor text/nsn/text    1
######################################################
4. modificat dat&abreve; 1.95914819643633 1.00166345680891 2 4482 36850.9780102
------------------------------------------------------
modificat&abreve;/asn/modificat ultima/m/ultima dat&abreve;/nsrn/dat&abreve;    3447
modificat/asn/modificat ultima/m/ultima dat&abreve;/nsrn/dat&abreve;    944
modificate/apn/modificat ultima/m/ultima dat&abreve;/nsrn/dat&abreve;   57
modificat&abreve;/asn/modificat ultima/m/ultima data/nsry/dat&abreve;   7
modificat&abreve;/asn/modificat la/s/la data/nsry/dat&abreve;   5
modificate/apn/modificat la/s/la data/nsry/dat&abreve;  3
modificat&abreve;/asn/modificat ultim&abreve;/m/ultim&abreve;
dat&abreve;/nsrn/dat&abreve;    3
modificat/asn/modificat la/s/la data/nsry/dat&abreve;   2
modificat/asn/modificat ultima/m/ultima data/nsry/dat&abreve;   2
modificat&abreve;/asn/modificat &icirc;nainte_de/s/&icirc;nainte_de
data/nsry/dat&abreve;   2
```

Figure 9: *'Retrieve All Contexts'* output

Retrieve All Full Contexts

The output resembles that of the '*Retrieve All Contexts*' but in this case each collocation is followed by the list of the unique sentences in the given text which contain that collocation, along with their corresponding frequencies (see Fig. 10).

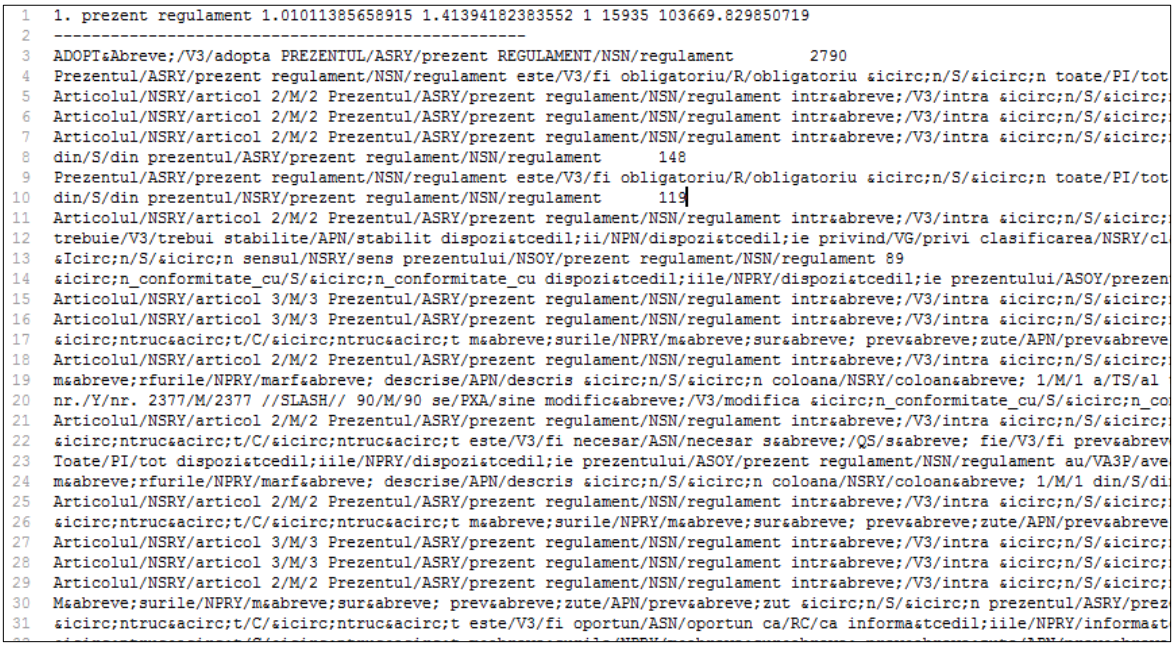


Figure 101: '*Retrieve All Full Contexts'* output

Another output is the Lucene index which is constructed when selected the command '*Lucene*'-> '*Index text*' (see Fig. 3). This will create a folder named '*LuceneIndex*' in the current folder, containing the Lucene index for the input text.

2.5 Integration with external tools

Collocation Extractor is fully self-contained.

3 CONTENT INFORMATION

This application can be tested by using the example provided in the 'test/' folder. The user should run the executable *Collocation Extractor.exe* and then go through the above explained procedures (see Section 2.3) using as input the file 'input.txt' in the same folder.

3.1 Test input files

See the testing kit in the 'test/' folder. The input file is 'input.txt' which contains a preprocessed Romanian 346.9 Mb text from the JRC-Acquis corpus (Steinberger et al., 2006).

3.2 Output files

One may obtain the following output files:

- FinalResults.txt – file containing the collocation list extracted from the given pre-processed text (see Section 2.4 and Fig. 8);

- vocabulary.txt – file containing all lemmas (with their corresponding POS tags) in the given text and their frequencies. This file is used for generating the FinalResults.txt file;
- pairs.txt – file containing all pairs extracting from the given text according to user preferences / constraints defined as in Fig. 2. On each line it contains a word pair, the distance between the words and the POS tags of the two. This file is used for generating the FinalResults.txt file;
- log.txt – contains the running times for different stages of the extraction process;
- LuceneIndex – folder which contains the Lucene index constructed as in Fig. 3;
- Contexts.txt – file containing the data described in Section 2.4 and Fig. 9;
- FullContexts.txt – file containing the data described in Section 2.4 and Fig. 10;

3.3 Running times

In order to report the running times for this tool, we run it on a 64bit 12-core Intel(R) Core(TM) i7 CPU 980 @ 3.33GHz and 16 GB of RAM.

For the example given in 'test/' folder we obtained the following timings (see log.txt file in 'reference_results/' folder):

- '*Extract from Text*' completed in 4:44 minutes;
- Lucene index completed in 1:01 minutes;
- '*Retrieve All Contexts*' completed in 5:11 minutes;
- '*Retrieve All Contexts*' completed in 5:15 minutes;

4 ADMINISTRATIVE INFORMATION

4.1 Contact

For further information, please contact Dan ŞTEFĂNESCU (http://www.racai.ro/~danstef/; danstef@racai.ro).

5 REFERENCES

Manning C., Schütze H. (1999). Foundations of Statistical Natural Language Processing, MIT Press, Cambridge.

Ştefănescu, D. (2010). Intelligent Information Mining from Multilingual Corpora. PhD thesis (in Romanian). Romanian Academy, Bucharest.

Ştefănescu, D. (2012). Mining for Term Translations in Comparable Corpora. In Proceedings of the 5th Workshop on Building and Using Comparable Corpora (BUCC 2012), Istanbul, Turkey.

Ştefănescu, D., Tufiş, D., Irimia, E. (2006). Automatic Identification and Extraction of Collocations from Texts. In Proceedings of the 2nd Romanian Workshop for Linguistic Tools and Resources Volume, 3 Nov. 2006, Bucharest, Romania (in Romanian).

Ştefănescu, D., Ceauşu, A., Ion, R., Todiraşcu, A., Heid, U., Gledhill, C., Rousselot, F. (2008). Extraction de collocations monolingues et bilingues: application a la traduction. In Proceedings of the Latin Union Conference, 28-29 Feb. 2008, Bucharest, Romania (in French). ISBN 978-9-291220-37-3.

Smadja F. (1993). Retrieving Collocations from Text: Xtract.Computational Linguistics 19, pp. 143-175.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), pp. 2142--2147. Genoa, Italy, 24-26 May 2006.

Todiraşcu, A., Gledhill, C., Ştefănescu, D. (2007). Extracting Collocations in Context: the case of Romanian VN constructions. In Proceedings of RANLP 2007, 27-29 Sep. 2007, Borovets, Bulgaria.

Todiraşcu, A., Gledhill, C., Ştefănescu, D. (2009). Extracting Collocations in Contexts. In Human Language Technology. Challenges of the Information Society, Lecture Notes in Computer Science Series, Springer Berlin / Heidelberg. ISSN: 0302-9743 (Print) 1611-3349 (Online), Volume 5603/2009, pp. 336-349, 2009. ISBN 978-3-642-04234-8.

LangId

1 BASIC INFORMATION

1.1 Tool name

The tool is called **LangId** which is the short for “Language Identifier”.

1.2 Overview and purpose of the tool

LangId identifies the language of a written raw text. It is developed in C# and it is encapsulated as a SOAP web service which is WSDL described at:

http://www.racai.ro/webservices/LangId.asmx?WSDL

LangId is able to identify 54 commonly known languages (*Afrikaans*, *Alemannic German*, *Arabic*, *Azerbaijani*, *Bavarian*, *Belarusian*, *Bosnian*, *Breton*, *Bulgarian*, *Catalan*, *Chinese* (*Standard*), *Croatian*, *Czech*, *Danish*, *Dutch*, *English*, *Esperanto*, *Estonian*, *Filipino*, *Finnish*, *French*, *Galician*, *German*, *Greek*, *Hebrew*, *Hungarian*, *Indonesian*, *Irish Gaelic*, *Italian*, *Japanese*, *Korean*, *Latin*, *Latvian*, *Lithuanian*, *Maltese*, *Norwegian*, *Occitan*, *Polish*, *Portuguese*, *Romanian*, *Russian*, *Serbian*, *Serbo-Croatian*, *Sicilian*, *Slovak*, *Slovene*, *Spanish*, *Swedish*, *Thai*, *Turkish*, *Ukrainian*, *Volapük*, *Welsh*, *Yiddish*) and 3 rare languages (*Aweti*, *Beaver*, *Teop*).

1.3 A short description of the algorithm

The algorithm and the web service are thoroughly described in Tufiş et al. (2008) and Ştefănescu (2010). The algorithm implements a statistical solution which requires for training raw balanced reference texts for each of the languages one intends to identify. In the training phase, the application builds a model for each language of interest, which contains the distribution of affixes and small words in the reference texts corresponding to that language. In the prediction phase, when a new, unseen text is presented, the tool generates a similar model for it

and then compares this model with the reference models learned in the training phase. The application computes a similarity score for each language and the language code corresponding to the highest score is returned as output.
In the training phase, for each of the 57 languages, we used Wikipedia documents containing 3 to 10 Mb of raw text.
For the moment, the web service is a quite slow since it is necessary to compare the models for 57 language pairs in order to identify the language. In the near future we plan to optimize it, in order to run faster.

2 TECHNICAL INFORMATION

2.1 Software dependencies and system requirements

This tool is available as a web service and therefore all computations are performed on the machine hosting the service. Consequenly, there are no system requirements for the end-user.

2.2 Installation

There is no installation required for this tool.

2.3 Execution instructions

After adding the Web Service as a web reference, identifying the language of a text can be done using the following lines of code (this is a C# example; see *Program.cs*):

```
LangIdWebService langIdService = new LangIdWebService();
string text = DataStructReader.readWholeTextFile("input.txt", Encoding.UTF8);
LangIDResult result = langIdService.IdentifyLanguage(text, true, false);
```

2.4 Input / Output data formats

The input should be a UTF-8 raw text string. The output is a *LangIDResult* object containg:

- the native name of the identified language (.*LanguageNative*)
- the English name of the identified language (.*LanguageEn*)
- ISO 639-1 code of the identified language (.*Language*)
- the confidence score for the prediction, which is a *double* value between 0 and 100 (.*Confidence*)

2.5 Integration with external tools

LangId is fully self-contained.

3 CONTENT INFORMATION

This Web Service can be tested by using the example provided in the ‘test/’ directory, or by using the web application available at:

http://www.racai.ro/webservices/LangId.aspx

For testing using the local provided executable (in ‘test/’ directory), it is necessary to have .Net Framework 4 installed.

Program.cs contains a C# example for using the web service.

3.1 Test input files

See the testing kit in the 'test/' directory. The input file is 'input.txt' which contains a raw Latin UTF-8 text.

In case of using the web application, the input text should be UTF-8 encoded and it should be inserted into the '*Input text:*' textbox.

3.2 Output files

There is no output file. The test application will display results at Standard Output.

3.3 Running times

LangId Web Service is hosted on a 4-core Intel(R) Xeon(R) x86 CPU @ 3.20GHz and 2 GB of RAM. The usual running time for identifiying the language of a text (as in the provided example) is about 13.5 seconds.

4 ADMINISTRATIVE INFORMATION

4.1 Contact

For further information, please contact Dan ŞTEFĂNESCU (http://www.racai.ro/~danstef/; danstef@racai.ro).

5 REFERENCES

Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D. (2008). RACAI's Linguistic Web Services. In Proceedings of the 6th Language Resources and Evaluation Conference - LREC 2008, 28-30 May 2008, Marrakech, Morocco

Ştefănescu, D. (2010). Intelligent Information Mining from Multilingual Corpora. PhD thesis (in Romanian). Romanian Academy, Bucharest.

LexChain

1 BASIC INFORMATION

1.1 Web Service name

The WebService is called **LexChains** which is the short for "Lexical Chains".

1.2 Overview and purpose of the Web Service

LexChains generates lexical chains between words in English or English Princeton WordNet 3.0 (Fellbaum, 1998) (PWN) concepts, using the PWN structure. It is developed in C# as a REST web service accessible at:

http://khufu.racai.ro:8001/lexchains.ashx

1.3 A short description of the service

Lexical Chains as returned by this service are described in Ion and Ştefănescu (2011) and Ştefănescu (2010).

The term “lexical chain” refers to a set of words which, in a given context (sentence, paragraph, section and so on), are semantically related to each other and are all bound to a specific topic. For instance, words like “tennis”, “ball”, “net”, “racket”, “court” all may form a lexical chain if it happens that a paragraph in a text contains all of them.

Moldovan and Novischi (2002) used an extended version of the PWN (XWN[31]) to derive lexical chains between the meanings of two words by finding relation paths in the PWN hierarchy.

In a similar fashion, the LexChains Web Service exploits the PWN structure returning such lexical chains for a given pair of words or concepts. In this paradigm, a lexical chain is not simply a set of topically related words, but becomes a path of synsets in the PWN hierarchy. This is the derived definition that we have used to implement our version of lexical chains as meaning paths through PWN.

2 TECHNICAL INFORMATION

2.1 WSDL code for invoking the web-service

This is a *REpresentational State Transfer* (REST) Web Service, for which there is no WSDL code.

2.2 Software dependencies and system requirements (if any)

There is no need for local installations and there are no software dependencies or system requirements for this tool, since it is hosted by a remote machine (khufu.racai.ro).

2.3 Execution instructions

Since this is a REST Web Service, it can be directly accessed by users or applications. This can be done in multiple ways:

i. given 2 words as input

e.g.: http://khufu.racai.ro:8001/lexchains.ashx?w1=tree&w2=leaf

The user must use GET parameters *w1* and *w2*

ii. given 2 words and their part of speech as input

e.g.: http://khufu.racai.ro:8001/lexchains.ashx?w1=farmer&pos1=n&w2=grow&pos2=v

The user must use GET parameters *w1*, *w2*, *pos1* and *pos2*

iii. given 2 concepts as input, by using the Inter Lingual Indexes (ILIs) as codification

e.g.: http://khufu.racai.ro:8001/lexchains.ashx?ili1=09450163-n&ili2=09394007-n

The user must use GET parameters *ili1* and *ili2*

The user has also the possibility to limit the number of chains returned by the service by using GET parameter *max*; e.g.:

http://khufu.racai.ro:8001/lexchains.ashx?ili1=09450163-n&ili2=09394007-n&max=10

[31] http://xwn.hlt.utdallas.edu/

2.4 Input / Output data formats

The Input requires the GET parameters to be correctly set.

The Output has the following format (see Fig. 1):

<*word_1/concept_1*> (<*PWN_relation*_1>) <*intermediary_concept*_1> (<*PWN_relation*_2>) <*intermediary_concept*_2> … (<*PWN_relation*_k>) <*intermediary_concept*_k> **<*word_2/concept_2*>**

<*concept*_i> and <*intermediary_concept*_i> have the following format (see Fig. 1):

ILI[literal_*1*#sense_no(pos), … , literal_*n*#sense_no(pos)]

O

```
09450163-n[sun#1(n),Sun#1(n)] (member_holonym) 09439433-n[solar_system#1(n)] (member_holonym) 09394007-n[planet#1(n),major_planet#1(n)]
09450163-n[sun#1(n),Sun#1(n)] (member_holonym) 09439433-n[solar_system#1(n)] (member_meronym) 09394007-n[planet#1(n),major_planet#1(n)]
09450163-n[sun#1(n),Sun#1(n)] (member_holonym) 09439433-n[solar_system#1(n)] (member_meronym) 09381480-n[outer_planet#1(n)] (hyponym) 09
09450163-n[sun#1(n),Sun#1(n)] (member_meronym) 09439433-n[solar_system#1(n)] (member_holonym) 09394007-n[planet#1(n),major_planet#1(n)]
09450163-n[sun#1(n),Sun#1(n)] (member_holonym) 09439433-n[solar_system#1(n)] (member_holonym) 09381480-n[outer_planet#1(n)] (hyponym) 09
09450163-n[sun#1(n),Sun#1(n)] (member_holonym) 09439433-n[solar_system#1(n)] (member_meronym) 09355623-n[minor_planet#1(n),planetoid#1(n
n[planet#1(n),major_planet#1(n)]
09450163-n[sun#1(n),Sun#1(n)] (member_holonym) 09439433-n[solar_system#1(n)] (member_meronym) 09355623-n[minor_planet#1(n),planetoid#1(n
n[planet#1(n),major_planet#1(n)]
09450163-n[sun#1(n),Sun#1(n)] (member_holonym) 09439433-n[solar_system#1(n)] (member_meronym) 09394007-n[planet#1(n),major_planet#1(n)]
n[planet#1(n),major_planet#1(n)]
```

Figure 2: LexChains Output format

2.5 Integration with external tools

LexChains is fully self-contained.

3 CONTENT INFORMATION

3.1 An usage example with associated data

Examples of how to use this Web Service are given in Section 2.3

3.2 An example of the output data

An Output example for query (see also Fig. 1):

http://khufu.racai.ro:8001/lexchains.ashx?ili1=09450163-n&ili2=09394007-n&max=5

09450163-n[sun#1(n),Sun#1(n)] (member_holonym) 09439433-n[solar_system#1(n)] (member_holonym) 09394007-n[planet#1(n),major_planet#1(n)]

09450163-n[sun#1(n),Sun#1(n)] (member_holonym) 09439433-n[solar_system#1(n)] (member_meronym) 09394007-n[planet#1(n),major_planet#1(n)]

09450163-n[sun#1(n),Sun#1(n)] (member_holonym) 09439433-n[solar_system#1(n)] (member_meronym) 09381480-n[outer_planet#1(n)] (hyponym) 09394007-n[planet#1(n),major_planet#1(n)]

09450163-n[sun#1(n),Sun#1(n)] (member_meronym) 09439433-n[solar_system#1(n)] (member_holonym) 09394007-n[planet#1(n),major_planet#1(n)]

09450163-n[sun#1(n),Sun#1(n)] (member_holonym) 09439433-n[solar_system#1(n)] (member_holonym) 09381480-n[outer_planet#1(n)] (hyponym) 09394007-n[planet#1(n),major_planet#1(n)]

3.3 Approximation of the time necessary to process the test input file

LexChains Web Service is hosted on a 8-core Intel(R) Xeon(R) x64 CPU E5504 @ 2.00GHz and 8 GB of RAM. Running time for a query is limited to maximum 100 miliseconds.

4 ADMINISTRATIVE INFORMATION

4.1 Contact

For further information, please contact Dan ŞTEFĂNESCU (http://www.racai.ro/~danstef/; danstef@racai.ro).

5 REFERENCES

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Ion, R., Ştefănescu, D. (2011). Unsupervised Word Sense Disambiguation with Lexical Chains and Graph-Based Context Formalization. In Zygmunt Vetulani (ed.): LTC 2009, Lecture Notes in Artificial Intelligence, Volume 6562/2011, pp. 435—443, Springer, Heidelberg

Moldovan, D., Novischi, A. (2002). Lexical chains for question answering. In: Proceedings of the 19th International Conference on Computational Linguistics, August 24 – September 01, 2002, Taipei, Taiwan, pp. 1-7.

Ştefănescu, D. (2010). Intelligent Information Mining from Multilingual Corpora. PhD thesis (in Romanian). Romanian Academy, Bucharest.

LexPar

1 BASIC INFORMATION

1.1 Tool name

The tool is called **LexPar** which is the short for “Lexicalized Parsing” in Romanian and English.

1.2 Overview and purpose of the tool

LexPar performs a pseudo-syntatic analysis of a sentence using an improved version of the Lexical Attraction Models of Deniz Yuret (1998). It is a dependency-like analysis but without relation orientation and labeling. It is written in Perl and it is encapsulated as a SOAP web service which is WSDL described at http://ws.racai.ro/lxpws.wsdl. The version number of LexPar is currently 6.0.

1.3 A short description of the algorithm

Lexical Attraction Models (LAMs) were first introduced by Deniz Yuret (1998) to exemplify how an algorithm can learn word dependencies from raw text. The “syntactic structure” of a sentence is given by a dependency structure which is an undirected, connected and planar graph with no cycles. We will refer to this structure using the term “linkage” (not a proper syntactic structure as defined by Meľčuk (1988) because the links are not oriented and they

have no names; what has been retained from Meľčuk's definition is the planarity condition which states that two links are not allowed to intersect except at a word position in a sentence). Yuret's general thesis is that lexical attraction is the likelihood of a syntactic relation. Therefore, the highest probability assigned by the model to a sentence is to be obtained by the syntactically correct linkage. He proves the result by formalizing the dependency structure of a sentence as a Markov network (or Markov random field (Kindermann and Snell 1950)). A node of the network is a word of the sentence and the potential function is the pointwise mutual information (MI) of a link.

LexPar (Ion and Barbu Mititelu, 2006) produces a linkage with the same properties as that of (Yuret 1998). It will be generated using the same basic idea: intermixing learning with parsing. The novelty of our approach resides in the use of with two additional devices:

1. linking rules that allow the formation of certain links while rejecting others;
2. use of POS information to increase the score of a weak (or unseen) link.

2 TECHNICAL INFORMATION

2.1 Software dependencies and system requirements

This kit contains a Perl wrapper to the LexPar linkage generator web service. The text to be analyzed is sentence split and then POS tagged and lemmatized. We (transparently) call the TTL web service (also a deliverable of RACAI in Batch 2 of the MetaNet4U project) to achieve these annotations. To run these scripts one must install the following:

- Perl, a recent version from http://www.perl.org/get.html;
- Perl package Unicode::String from http://www.cpan.org/;
- Perl package SOAP::Lite from http://www.cpan.org/.

2.2 Installation

Other than what is mentioned in the previous section, there is no installation required for this tool.

2.3 Execution instructions

LexPar (script name 'lexparws.pl') is called from a command line interface giving it two arguments: the language of the text to be linked, either English, ('en') or Romanain ('ro'), and the UTF-8 encoded text file to be processed. The distribution kit contains a test directory with two UTF-8 encoded text files: 'en-test-file.txt' and 'ro-test-file.txt'. If we want to run LexPar on the Romanian file, we should type:

./lexparws.pl ro test/ro-test-file.txt >result.txt

The output is written to the STDOUT in UTF-8 so that 'lexparws.pl' can be integrated into command line pipelines. In the command above, the result is redicted into 'result.txt' file in the same working directory as 'lexparws.pl'.

2.4 Input/Output data formats

The input data to LexPar is a text file, UTF-8 encoded with no byte order markings. This file is pre-processed (internally) using the TTL web service.

The output data produced by LexPar is the text in the input file, sentence split, POS tagged and lemmatized, with linking information. Each token along with its annotations is printed on a line (separated by '\r\n'). Sentences are separated by a blank line (a single '\r\n').

Each token in a sentence is counted beginning with '0' up to the number of tokens in that sentence. The annotations of the current token are tab-separated ('\t') from the token and are (in the order of appearance): word form, POS tag, lemma, chunk and link index. The link index is the index of the word with which the current word links taking into acound the numbering of the tokens in the current sentence. **Please note that the link index may be missing** due to the fact that, in the process of linkage formation, that word could not be linked to any other word (e.g. some linking rule denied the formation of the link or the resulting graph would have not been planar). Also, the linkage annotation is not symmetric: if token A is linked to token B, then this information is present on either A or B, not both! See the next example for reference.

0	It		Pp3ns	it		Vp#1		
1	helps	Vmip3s	help		Vp#1		0	
2	the		Dd	the				7
3	poor		Afp	poor			Ap#1	7
4	and		Cc-n	and				5
5	many		Pi3-p	many		Np#1	7	
6	middle-class	Afp		middle-class	Np#1,Ap#2		7	
7	people	Ncn		people	Np#1		1	
8	afford	Vmn		afford	Vp#2		1	
9	the		Dd	the			Np#2	10
10	cost		Ncns	cost			Np#2	8
11	.		PERIOD	.				

2.5 Integration with external tools

LexPar dependes on the availability of the TTL web service also hosted by RACAI to obtain sentence splitting, POS tagging and lemmatization. The WSDL of the TTL web service is located at http://ws.racai.ro/ttlws.wsdl.

3 CONTENT INFORMATION

3.1 Test input files

See the distribution kit in the 'test/' directory. There is one test file for each of the languages supported by LexPar, namely English and Romanian.

3.2 Output files

For each input file in the 'test/' directory, there is the corresponding output file in the 'sample-output/' directory in the distribution kit.

3.3 Running times

The speeds that we are going to report have been obtained by calling the LexPar web service in the local network on a wireless connection of 54Mbps. One should consider the fact that the time the text is passed around the network is going to affect the overall performance of LexPar. Also, LexPar is running on a Ubuntu Server machine with a 8-core Intel(R) Xeon(R) CPU E5405 @ 2.00GHz CPU and 8 GB of RAM.

We have considered two types of measures: 'bytes/second' and 'tokens/second'. Bytes/second is going to help us predict the number of seconds after which an input UTF-8 text file of N bytes is going to be finished. Tokens/second gives the average processed tokens per second that the LexPar web service is capable. Please note that the LexPar web service also calls the TTL web sevice and its speed is going to be dependent on the TTL speed.

	English	Romanian
tokens/second	19	17
bytes/second	111	98

Table 1: LexPar speeds on each supported language

4 ADMINISTRATIVE INFORMATION

4.1 Contact

For further information, please contact Radu ION (radu@racai.ro).

5 REFERENCES

Ion, R., Barbu Mititelu, V. (2006). Constrained Lexical Attraction Models. In Proceedings of the Nineteenth International Florida Artificial Intelligence Research Society Conference. Menlo Park, Calif.: AAAI Press, pp. 297--302.

Kindermann, R. and Snell, J. L. 1950. Markov Random Fields and their Applications. American Mathematical Society, Providence Rhode Island.

Meľčuk, I. 1988. Dependency Syntax: Theory and Practice, New York , SUNY Press.

Yuret, D. 1998. Discovery of linguistic relations using lexical attraction. Ph.D. diss., Department of Computer Science and Electrical Engineering, MIT.

RO-HYPHEN

1 BASIC INFORMATION

1.1 Web Service name

The Web Service name is Bermuda Hyphenator (part of the VoiceForge delivrable in Batch 3). It is a substitute for the ro-Hyphen tool in Batch 2.

1.2 Overview and purpose of the Web Service

Bermuda Hyphenator is used to split words into syllables and add stress information at syllable level.

1.3 A short description of the service

Bermuda Hyphenator uses a MaxEnt classifier for splitting the word into syllables based on lexical features and then applies a custom algorithm (also based on MaxEnt) in order to assign stress to a syllable. The modification of the original MaxEnt classifier for stress assignment consists in the fact that the MaxEnt classifier is ***forced*** to assign exactly one lexical stress. The tool uses different stress patterns based on the part-of-speech of the word.

2 TECHNICAL INFORMATION

2.1 WSDL code for invoking the web-service

The WSDL code is located at http://khufu.racai.ro:8081/Hyphenator.asmx?WSDL.

2.2 Software dependencies and system requirements (if any)

There is no need for local installations and there are no software dependencies or system requirements for this tool, since it is hosted by a remote machine (khufu.racai.ro).

2.3 Execution instructions

The GetHyphenation method is used to hyphenate a given word (parameter 1) and assign stress to a syllable based on the part-of-speech (POS) (parameter 2) of the given word. Use “N” for nouns, “V” for verbs, “A” for adjectives and “R” for adverbs. For other or unknown use “OTHER” as POS. Also use “OTHER” if you don’t have enough information about the word’s POS.

2.4 Input / Output data formats

It returns the word’s syllables separated with hyphens and it marks the stressed syllable with the “’” character. An example output for the word “testare” (en. testing) is: “**tes-'ta-re”**

2.5 Integration with external tools

Bermuda Hyphenator is fully self-contained.

3 CONTENT INFORMATION

3.1 An usage example with associated data

For the request: word:’testare’ PartOfSpeech: ‘N’ the Web Service returns the following response:
<string>tes-'ta-re</string>

3.2 An example of the output data

An example of the output data is given in section 3.1.

3.3 Approximation of the time necessary to process the test input file

Bermuda Hyphenator Web Service is hosted on a 8-core Intel(R) Xeon(R) x64 CPU E5504 @ 2.00GHz and 8 GB of RAM. It takes about 1-2ms to process a request (if the resources are pre-loaded).

4 ADMINISTRATIVE INFORMATION

4.1 Contact

For further information, please contact Tiberiu BOROȘ (tibi@racai.ro).

TTL Package (TTL-Lemmatizer, TTL-Tagger, TTL-Tokenizer, TTL-Chunker)

1 BASIC INFORMATION

1.1 Tool name

The tool is called **TTL** which is the short for "Tokenizing, Tagging and Lemmatizing free running texts" in Romanian, English and French.

1.2 Overview and purpose of the tool

TTL performs sentence splitting, tokenization, POS tagging, lemmatization and shallow parsing (chunking) on Romanian, English and French texts. It can be extended to support other languages provided the necessary resources exist: POS tagged corpora and lexicons. It is written in Perl and it is encapsulated as a SOAP web service which is WSDL described at http://ws.racai.ro/ttlws.wsdl. The version number of TTL is currently 8.5.

1.3 A short description of the algorithm

TTL (Ion, 2007) is a text preprocessing module developed in Perl. Its functions are: Named Entity Recognition (by means of regular expressions defined over sequences of characters), sentence splitting, tokenization, POS tagging, lemmatization and chunking.

The NER function is included as a preprocessing stage to sentence splitting because end of sentence markers may constitute parts of an NE string (i.e. a period may be a part of an abbreviation). POS tagging is achieved through the HMM tagging technology.

The POS tagger of TTL follows the description of HMM tagger given in (Brants, 2000) but it extends it in several ways allowing for tiered tagging, for a more accurate processing of unknown words and also for tagging of named entities (which are practically labeled by the NER module before actual POS tagging). The TTL's tagset is the MSD[32] with its smaller superset CTAG. TTL tagging methodology follows the tiered tagging approach (Tufiş, 1999) where MSDs are recovered from an initial CTAG annotation.

Lemmatization is achieved after POS tagging by lexicon lookup (in general, a word form and its POS tag uniquely identify the lemma). In the case of out-of-lexicon word forms the lemmatization is performed by a statistical module which automatically learns normalization rules from the existing lexical stock (for details see (Ion, 2007)).

Chunking is implemented with regular expressions over sequences of POS tags. It is not recursive and it does not perform attachments (PPs to NPs for instance).

[32] http://nl.ijs.si/ME/V3/msd/html/

2 TECHNICAL INFORMATION

2.1 Software dependencies and system requirements

This kit contains Perl wrappers to the TTL web service basic operations: tokenization (which includes sentence splitting), POS tagging, lemmatization and chunking. To run these scripts one must install the following:

- Perl, a recent version from http://www.perl.org/get.html;
- Perl package Unicode::String from http://www.cpan.org/;
- Perl package SOAP::Lite from http://www.cpan.org/.

2.2 Installation

Other than what is mentioned in the previous section, there is no installation required for this tool.

2.3 Execution instructions

There are four operations that are available through TTL: tokenization (which includes sentence splitting), POS tagging, lemmatization and chunking. These operations can be chained through the standard UNIX pipe operator '|' in this **strict order**: tokenization (script name 'ttlws-tokenizer.pl'), POS tagging ('ttlws-postagger.pl'), lemmatization ('ttlws-lemmatizer.pl') and chunking ('ttlws-chunker.pl').

Each script will receive the language code ('en', 'ro' or 'fr') and an input file (which may be the result of a previous processing step) and will write the result of the processing to STDOUT. If called with no arguments, the script will output the command line arguments it expects. The possible workflows which may be constructed with these operations are:

- **Tokenization**: call ttlws-tokenizer.pl and obtaind a list of tokenized sentences separated with '\r\n';
- **POS tagging**: call ttlws-postagger.pl on the result of ttlws-tokenizer.pl and obtain POS tags for each token;
- **Lemmatization**: call ttlws-lemmatizer.pl on the result of ttlws-postagger.pl to obtain lemmas for each word;
- **Chunking**: call ttlws-chunker.pl on the result of ttlws-lemmatizer.pl to obtain the chunk name and id to which the current word belongs;

For instance, if we want to POS tag a Romanian UTF-8 encoded file (all input files must be text and UTF-8 encoded) which resides in the 'test/' directory under the current working directory in which the Perl scripts ttlws-tokenizer.pl and ttlws-postagger.pl are, then by calling

```
cat test/ro-test-file.txt | \

        ttlws-tokenizer.pl ro - | \

        ttlws-postagger.pl ro - >result.txt
```

the results of the tokenization and POS tagging of the file 'ro-test-file.txt' will be written in 'result.txt'. The dash '-' stands for the input file read from STDIN. Alternatively, one can run the following to achieve the same result:

```
ttlws-tokenizer.pl ro test/ro-test-file.txt >result-1.txt

ttlws-postagger.pl ro result-1.txt >result.txt
```

The full workflow can be run as follows:

```
cat test/ro-test-file.txt | \
        ttlws-tokenizer.pl ro - | \
        ttlws-postagger.pl ro - | \
        ttlws-lemmatizer.pl ro - | \
        ttlws-chunker.pl ro - >result.txt
```

2.4 Input/Output data formats

The input data for the first step of the workflow, i.e. tokenization, is UTF-8 text file with no Byte Order Markings (BOM). Each processing step will output the data in text format, column style. Thus, for the input text "This is a test sentence in English.", the tokenizer ttlws-tokenizer.pl will output the text:

```
This
is
a
test
sentence
in
English
.       PERIOD
```

in which each token is on a separate line (end of line follows Windows conventions: '\r\n'). If the token is a punctuation sequence or a named entity, after a tab character ('\t') follows a label that will be used by the POS tagger ttlws-postagger.pl. Sentences are separated by an empty line '\r\n'.

Each processing step produces an output that is used by the next step in line. The output are tokens per line with tab characters separating the annotations produced up the current processing step. The full pipeline for our example above produces the output (with extra tabs inserted to ease the reading):

```
This            Pd3-s   this
is              Vmip3s  be              Vp#1
a               Ti-s            a               Np#1
test            Ncns            test            Np#1
sentence        Ncns            sentence        Np#1
in              Sp              in              Ap#1
English Afp             English Ap#1
.               PERIOD  .
```

2.5 Integration with external tools

TTL is fully self-contained. There are no external dependencies other than the Perl packages required for the web service wrappers to run.

3 CONTENT INFORMATION

3.1 Test input files

See the distribution kit in the 'test/' directory. There is one test file for each of the languages supported by TTL.

3.2 Output files

For each input file in the 'test/' directory, there is the corresponding output file in the 'sample-output/' directory in the distribution kit.

3.3 Running times

The speeds that we are going to report have been obtained by calling the TTL web service in the local network on a wireless connection of 54Mbps. One should consider the fact that the time the text is passed around the network is going to affect the overall performance of TTL. Also, TTL is running on a Ubuntu Server machine with a 8-core Intel(R) Xeon(R) CPU E5405 @ 2.00GHz CPU and 8 GB of RAM.

We have considered two types of measures: 'bytes/second' and 'tokens/second'. Bytes/second is going to help us predict the number of seconds after which an input UTF-8 text file of N bytes is going to be finished. Tokens/second gives the average processed tokens per second that the TTL web service is capable. We measure the speeds of all possible workflows presented in section 2.3. The server was not idle through the tests and it had 2GB of RAM and 1 core assigned to other process. This means that the reported speeds could be indicative of a low load of the server and could be higher if the server is assigned only to TTL. They can also be lower if the server is fully loaded.

	English	Romanian	French
1. Tokenization	2603	2255	2348
2. POS Tagging	882	1101	872
3. Lemmatization	782	914	755
4. Chunking	681	652	659

Table 1: TTL speeds in bytes/second on each language. Please note that the processing step no. N includes all processing steps from 1 to N - 1

	English	Romanian	French
1. Tokenization	456	391	439
2. POS Tagging	155	191	163
3. Lemmatization	137	158	141
4. Chunking	119	113	123

Table 2: TTL speeds in tokens/second on each language. Please note that the processing step no. N includes all processing steps from 1 to N - 1

4 ADMINISTRATIVE INFORMATION

4.1 Contact

For further information, please contact Radu ION (radu@racai.ro).

5 REFERENCES

Brants, T. (2000). TnT – A Statistical Part-Of-Speech Tagger. In *Proceedings of the 6th Applied NLP Conference ANLP-2000*. Seattle, WA, pp 224--231.

Ion, R. (2007). Word Sense Disambiguation Methods Applied to English and Romanian. PhD thesis (in Romanian). Romanian Academy, Bucharest.

Tufiş, D. (1999). Tiered Tagging and Combined Classifiers. In F. Jelinek, E. Nth (Eds.), *Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence*. Springer, pp. 28--33.

UOM - University of Malta

Endogenous resources

F-MONA 1

1 BASIC INFORMATION

1.1 Corpus composition

108 WAV files subdivided into 12 directories with a variable number of sentences (sometimes: clauses) each. They come together with transcriptions and tables of phoneme durations (see 1.2 below).

1.2 Representation of the corpora (flat files, database, markup)

12 directories with several sentences (or clauses) each. Each sentence/clause is represented by four files:

i. A file with the speech sentence/clause - .wav

ii. A file with the sentence - .txt (This is a UTF-8 based file that uses the Maltese keyboard to include the 4 non-ASCII Maltese characters)

iii. An Excel file (xlsx, UTF-8) with 5 columns for

1. name of WAV file and beginning of phoneme (times given are in seconds)
2. end of phoneme
3. duration of phoneme
4. phoneme
5. word

1.3 Character encoding

UTF-8

2 ADMINISTRATIVE INFORMATION

2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)

Name: Paul Micallef
Address:
Affiliation: University of Malta, Department of Communications & Computer Engineering
Position: Professor Engineer
Telephone: +356 2340 2520
Fax: (+356) 21343577
e-mail: paul.micallef@um.edu.mt

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform. (also available under: http://staff.um.edu.mt/paul.micallef/speech_annotation)

2.3 Copyright statement and information on IPR

The resource is licensed under META-SHARE Commons BY-NC

3 TECHNICAL INFORMATION

3.1 Directories and files

12 directories, 108 WAV files, 12 txt files, 12 xlsx files

3.2 Data structure of an entry

5 columns (see above under 1.2)

3.3 Corpora size (nmb. of tokens, MB occupied on disk)

14.3 MB zipped, 108 WAV files for 128 sentences

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

monolingual, audio, raw text transcriptions, phoneme alignment information

4.2 The natural language(s) of the corpus

Maltese

4. 3 Domain(s)/register(s) of the corpus

Newspaper texts

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

n.a.

4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),

n.a.

4.4.3 Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)

alignment on: phoneme start, phoneme end, duration, phoneme, word

values retrieved automatically, partially manually checked

4.4.4 Attributes and their values (if annotated)

phoneme start: time
phoneme end: time
duration: time
phoneme: phoneme, sil(ence), breath, pause

4.5 Intended application of the corpus
Speech synthesis

4.6 Reliability of the annotations (automatically/manually assigned) – if any
automatically, partially checked

5 RELEVANT REFERENCES AND OTHER INFORMATION

MaltoBI Corpus

MalToBI / SPAN Corpus

Author(s): Dr. Alexandra Vella

Institute: Institute of Linguistics, University of Malta

Address:
University of Malta
Msida MSD 2080
MALTA

Email: alexandra.vella@um.edu.mt

Date: 06.12.2004 and 21.03.2005

Version: n.a.

CONTENTS
1. INTRODUCTION
1.1 SPEECH FILE FORMATS
WAV (sample rate: 44.1 kHz, resolution: 16-bit, Channels: stereo)

1.2 DIRECTORY STRUCTURE
Currently, the corpus contains 8 subfolders with a total of 913 .wav files (1.37 GB of disk storage, total time: 2h51m59s), 8 .html files (252 KB of disk storage), 1 text file (4 KB of disk storage) and 2 Text.Grid files (8 KB of disk storage).

MalToBI/SPAN has 8 subfolders, accompanied by a .html file which lists its contents. It also contains a file "Format.txt", which explains the data and file name structure of the subfolders and contains the sentences and phrases read out by the speakers (except the story "The North Wind and the Sun"). A folder normally contains 118 .wav files:

- 2 sound files containing a read story ("The North Wind and the Sun", each by speaker A and speaker B)
- 2 sound files containing each 30 read sentences (each by speaker A and speaker B)
- 2 x each of the 30 sentences as a single sound file (each by speaker A and speaker B)
- 2 x 26 phrases in individual files (each by speaker A and speaker B)
- 1 sound file with a map task (speaker A and B interacting)
- 1 sound file with a simulated conversation (speaker A and B interacting)

Exceptions from this structure are folders “0501221” (only 92 .wav files; lacking the single files for phrases (F) spoken by speaker B) and “050241” (only 113 .wav files; lacking 5 single files: 0502041A3_F20.wav, 0502041A3_S12.wav, 0502041A3_S16.wav, 0502041A3_S25.wav, 0502041A3_S28.wav). It is not clear whether the files were not uploaded or lost – however, they could be reconstructed from the corresponding long .wav files.

Folder “0412061” contains also two Text.Grid files (“0412061A3_S01.TextGrid.txt” and “0412061A3_S02.TextGrid.txt”) which are transcriptions of the respective .wav files.

1.3 FILE NAMING CONVENTIONS

FileName: YYMMDDXST.wav

where

YY: Two digit year of recording

MM: Two digit month of recording

DD: Two digit day of recording

X: One digit session number, starting from 1.

S: Single letter speaker A, B or C (where C means both). Speaker A is by convention the Map Task leader is always recorded on the left audio channel for joined tasks.

T: Single digit identifying task 1 (map task), 2 (story), 3 (phrases), 4 (dialogue.)

1.4 LABEL FILES

Annotation files in PRAAT Text.Grid format (currently only 2 in folder “0412061”, missing files will be added after revision)

The annotation of the corpus was made on different linguistic levels (tones, tone breaks, words, (miscellaneous)). In the present Text.Grid files, intervals (for words and “misc”) and pointers (for tones and tone breaks) are encoded by time data, based on the time-line which is given by the respective sound recording.

2. DATABASE DESIGN AND COLLECTION

2.1 RECORDING PLATFORM

unknown

2.2 SPEAKER RECRUITMENT

recruitment details unknown; 2 speakers per session (= per folder = per recording date)

2.3 PROMPTING DESIGN

see 3.

3. DATABASE CONTENTS DEFINITION

The categories 3.1 to 3.15 are not applicable, since the tasks were

- to read out a story (“The North Wind and the Sun”)
- to read out a collection of sentences (“Sentenzi”, see below)
- to read out a collection of phrases (“Frazijiet”, see below)
- a free conversation in a simulated situation (job interview)
- a free conversation during a map task

3.1 ISOLATED DIGITS
- 3.1.1 Single digit
- 3.1.2 Isolated digits string

3.2 CONNECTED DIGITS
- 3.2.1 Sheet number)
- 3.2.2 Telephone number
- 3.2.3 Credit card number
- 3.2.4 PIN code
- 3.2.5 Connected digit strings

3.3 NATURAL NUMBERS

3.4 MONEY AMOUNTS

3.5 YES/NO

3.6 DATES
- 3.6.1 Spontaneous date
- 3.6.2 Prompted date
- 3.6.3 Relative and general date expressions

3.7 TIMES
- 3.7.1 Spontaneous time
- 3.7.2 Prompted time

3.8 APPLICATION WORDS

3.9 APPLICATION WORD PHRASE

3.10 DIRECTORY ASSISTANCE NAMES
- 3.10.1 Spontaneous first name
- 3.10.2 Spontaneous city name
- 3.10.3 Prompted city name
- 3.10.4 Prompted company name
- 3.10.5 Prompted personal name

3.11 SPELT ITEMS
- 3.11.1 Strings of letters in MSA
- 3.11.2 Spelt word
- 3.11.3 Spelt city name
- 3.11.4 Spelt personal name

3.12 PHONETICALLY RICH WORDS

3.13 PHONETICALLY RICH SENTENCES

3.14 FREE SPONTANEOUS SPEECH

3.15 SPONTANEOUS CONTROL ITEMS
- 3.15.1 City name
- 3.15.2 Age
- 3.15.3 Place of call
- 3.15.4 Phone type

4. TRANSCRIPTION

Where available, the transcription files of the corpus are PRAAT Text.Grid files, using PRAAT's specific entry structure for each tier, point and interval. Each transcription file has four tiers: "tones", "words", "breaks", "misc(ellaneous)". The tiers "words" and "misc" are interval tiers, while "tones" and "breaks" are point tiers.

The following are not transcriptions but the sentences and phrases being read out by the speakers.

4.1 Sentenzi

1. Malta issa qieghda fl-Ewropa, hux hekk?
2. Inti gejja ghada, hux?
3. Ix-xoghol kollu spiccawh, hux tassew?
4. Iridu jmorru jghumu t-tfal, hux veru?
5. Ghamluh dak li kellu jsir, mhux tassew?

6. Vera ghamluh bil-galbu x-xoghol, mhux veru?
7. Tistaí tigi tghinni, Romina?
8. Lil Daniel sibtu, Angela?
9. Ix-xoghol jidher li se jkompli ghaddej.
10. F'dawn l-ahhar jiem gew iffirmati zewg kuntratti.

11. Kull m'ghandkom taghmlu huwa li timlew l-applikazzjoni mahruga mill-Ufficju tal-Kunsill.
12. Ghidulu jghaddi malajr.
13. Morru ixtru l-hobz minghand tal-kantuniera.
14. Oqghodu attenti intom w ghaddejjin minn hdejn tal-laham.
15. Ghadhom hemm id-diffikultajiet.

16. Ghandu l-qargha?
17. Kaxxa gallettini xtrawlha lil ommhom?
18. Gejja maghna allura?
19. Irregistrajtu ghalih is-seminar?
20. Fhimt x'kont qed nghidlek?

21. Fejn hu l-ktieb li kont qed naqra?
22. Meta ha tmur l-Ingilterra?
23. Ghalfejn qieghda taghggibha?!
24. X'jigifieri? X'inti thawwad?
25. Anna taf ismi u tieghek ukoll. Anna taf isimna?

26. X'ser tiehu minn dawn?
27. Fejn kont il-bierah meta gejna nfittxuk?
28. Forsi baghat ghalih is-Sultan biex jixtri xi wahda mill-invenzjonijiet stupendi tieghu?
29. Min gej? Francesca jew Daniela?
30. Inti x'se taghmel? Sejra maghhom jew gejja maghna?

4.2 Frazijiet
1. L-Gholja tar-Re
2. Iz-Zona ta' l-Ghajnejn
3. Triq Rainier
4. Hotel Pergola fi Triq Ermola
5. Triq l-Ewwel ta' Mejju

6. Zona Laramill
7. Triq Amery
8. Torri Mulé
9. Misrah il-Lejl

5. LEXICON
n.a.
6. SPEAKER DEMOGRAPHIC INFORMATION
There is no speaker information contained in the corpus
- 6.1 ACCENT/REGIONS
- 6.2 SPEAKER AGES
- 6.3 SPEAKER OVERLAP

7. RECORDING CONDITIONS
Studio recording
- 7.1 ENVIRONMENTS FOR FIXED NETWORK
- 7.2 ENVIRONMENTS FOR MOBILE NETWORK
- 7.3 HANDSETS

8. TEST MATERIAL
n.a.

The resource will be uploaded to META-SHARE as a set of .wav files with accompanying transcription files in PRAAT Text.Grid format. The current original files are available at: http://staff.um.edu.mt/cbor7/mlexweb/public_html2/maltobi/

Restricted Exogenous resources

Local Government documentation

1 BASIC INFORMATION

1.1 Corpus composition

This corpus is a collection of different governmental resources, containing two types of documents: minutes, which were taken during local council meetings (covering the years from 2007 till 2010) and memorandums (covering from 2008 till 2011).
This corpus, consisting of raw text files and comma separated values (CSV) files, is the percentage that could be extracted from the original corpus.
Some issues arise due to Maltese characters. It is important to note that not all documents contain the right Maltese characters. Some documents may replace:

ġ -> g ; ż -> z ; ħ -> h ; ċ -> c

With those being on the right hand side also Maltese characters, except *c*. Furthermore, in some of the documents, the keyboard equivalence of the character is printed, rather than the character itself (and this is also dependent on whether the user made use of the 47 or 48-key keyboard layout).

1.2 Representation of the corpora (flat files, database, markup)

The corpus is organized in two folders:

a. TXT – this folder contains a collection of text files; minutes and memos which were extracted from the original corpus.
b. CSV – this folder contains a collection of CSV files, most of which contain financial information, extracted from excel files in the original corpus

1.3 Character encoding

UTF-8

2 ADMINISTRATIVE INFORMATION

2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)

Name: Dr. Alexiei Dingli
Affiliation: University of Malta, Department of Communications & Computer Engineering
Position: Senior Lecturer, Department of Intelligent Computer Systems
Telephone: +356 2340 2486
e-mail: alexiei.dingli@um.edu.mt

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform

2.3 Copyright statement and information on IPR

META-SHARE Commons BY NC SA

3 TECHNICAL INFORMATION

3.1 Directories and files

2 Directories:

a. TXT directory – 4402 text files
b. CSV directory – 1275 CSV files

3.2 Data structure of an entry

n.a.

3.3 Corpora size (nmb. of tokens, MB occupied on disk)

TXT directory size on disc: 53.5MB
CSV directory size on disc: 184 MB
Corpus size on disc: 238MB

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

The data is mostly monolingual but is in two languages - both English and Maltese. The corpus is raw.

4.2 The natural language(s) of the corpus

Maltese and English

4. 3 Domain(s)/register(s) of the corpus

Local councils' meeting minutes, government memorandums and financial information (with respect to decisions undertaken by the local councils).

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

n.a.

4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),

n.a.

4.4.3Alignment information (if the corpus contain aligned documents: level of alignment, how it was achieved)

n.a.

4.4.4 Attributes and their values (if annotated)

n.a.

4.5 Intended application of the corpus

Information extraction techniques involving for instance named entity recognition and topic analysis to identify key elements of well known document types and to build gazetteers that include the names of people, organisations, places and quantities.

4.6 Reliability of the annotations (automatically/manually assigned) – if any
n.a.

5 RELEVANT REFERENCES AND OTHER INFORMATION

Maltese Wikipedia

1 BASIC INFORMATION

1.1 Corpus composition

This corpus is part of the collection of the Wikipedia Dumps which was retrieved from wikipedia.org on April 8, 2010. This resource was downloaded from: http://archive.org/details/mtwiki_20100610. It comes with two individual XML files, one containing the Wikipedia articles and another containing the metadata about it.

1.2 Representation of the corpora (flat files, database, markup)

2 XML files:

1. The main one contains:
    1.1 Site information, from where the data was retrieved
    1.2 The Wikipedia articles. Each article is enclosed between an XML tag, with the name *page*
2. A small file containing metadata about the retrieval of the articles

1.3 Character encoding

UTF-8

2 ADMINISTRATIVE INFORMATION

2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)
n.a.

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as a ZIP file, containing 2 files:

1. XML of Maltese Wikipedia
2. XML of its metadata (downloaded from the same site)

2.3 Copyright statement and information on IPR

Creative Commons license: Attribution-Noncommercial-Share Alike 3.0 United States

3 TECHNICAL INFORMATION

3.1 Directories and files

2 XML files:

1. An XML file containing the Maltese Wikipedia articles
2. An XML containing the metadata of (1)

3.2 Data structure of an entry

```
<page>
  <title> ... </title>
  <id> ... </id>
  <revision>
    <id> ... </id>
    <timestamp> ... </timestamp>
    <contributor>
      <username> ... </username>
      <id> ... </id>
    </contributor>
    <text xml:space="preserve"> ... </text>
  </revision>
 </page>
```

3.3 Corpora size (nmb. of tokens, MB occupied on disk)

22.1 MB, 9089 pages

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

The Wikipedia articles are monolingual - Maltese.
The file containing the metadata is in English.

4.2 The natural language(s) of the corpus

Maltese

4. 3 Domain(s)/register(s) of the corpus

Wikipedia articles

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

XML mark-up

4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),

Media Wiki web-page tags

4.4.3Alignment information (if the corpus contain aligned documents: level of alignment, how it was achieved)

n.a.

4.4.4 Attributes and their values (if annotated)

n.a.

4.5 Intended application of the corpus
Information extraction, data mining, ...
4.6 Reliability of the annotations (automatically/manually assigned) – if any
n.a.

5 RELEVANT REFERENCES AND OTHER INFORMATION
This resource was retrieved from http://archive.org/details/mtwiki_20100610, part of the Wikipedia dumps.

Unrestricted exogenous resources

Basic English-Maltese Dictionary

1. BASIC INFORMATION

1.1 Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc),
Bilingual wordlist, consisting of alphabetically ordered English lemmas with their Maltese translation and Maltese pronunciation (transcribed in ad-hoc system by the original author).
1.2 Representation of the lexicon (flat files, database, markup)
Originally a HTML file, the upload is a TEI-compliant XML dictionary file.

1.3 Character encoding
UTF-8

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)

Name: Dr. Grazio Falzon
Email : grazfalf@gmail.com
Address: 25405 Dana Drive, South Bend, IN 46619, U.S.A
Position: retired
2.2 Delivery medium (if relevant; description of the content of each piece of medium)
The resource will be uploaded on the MetaShare platform as the derived XML file in Version 1.

2.3 Copyright statement and information on IPR
META-SHARE Commons BY NC SA

3. TECHNICAL INFORMATION

3.1 Directories and files
one XML file
3.2 Data structure of an entry

Dictionary entries are marked using the XML schema for dictionaries after TEI P5:

```
<entry>
<form>
  <orth>ABBEY</orth>
</form>
<sense>
  <cit xml:lang="mt">
    <quote>abbazija</quote>
    <gramGrp>
      <pos>n</pos>
      <gen>F</gen>
    </gramGrp>
    <pron>abbatsi'ya</pron>
  </cit>
</sense>
</entry>
```

Multiple Maltese translations for one English entry are encoded with several (counted) <sense>-tags (see example for ABANDON in the XML file).

3.3 Lexicon size (nmb. of lexical items, KB occupied on disk)
5458 entries, ca. 2 MB on disk

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon
English, Maltese (direction: English to Maltese)

4. 2 Entry Type
XML markup
4.3 Attributes and their values
<orth>: string
<pos>: v(erb), n(oun), adj(ective), …
<gen> (gender of a noun): M, F
<number> sg, pl(ural)
<quote> string (translation)
<pron> string (pronunciation of the Maltese translation)
… (see TEI specs for dictionaries)

4.4 Coverage of the lexicon
Everyday life, no special domain

4.5 Intended application of the lexicon
Get by in Malta in everyday situations in Maltese.
4.6 POS assignment
nouns, verbs, adjectives

4.7 Reliability (automatically/manually constructed)
some mistakes/inconsistencies, no special characters, manually constructed;
conversion was done automatically (manual inconsistencies were taken over and will have to be cleaned from the XML file manually)

5. RELEVANT REFERENCES AND OTHER INFORMATION
Original HTML source is here: http://aboutmalta.com/language/engmal.htm

Newly added in Batch 2

MFSA Maltese Company Registration.zip

1. BASIC INFORMATION

1.1 Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc),
List of companies with further information

1.2 Representation of the lexicon (flat files, database, markup)
Database (Excel file)

1.3 Character encoding
unknown

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)

Name: Mr. Joseph Caruana
Address: Notabile Road, Attard BKR14
Telephone: (+356) 25485170
Position: Director Registry of Companies, Malta Financial Services Authority
email: JCaruana@mfsa.com.mt,

2.2 Delivery medium (if relevant; description of the content of each piece of medium)
The resource will be uploaded on the MetaShare platform.

2.3 Copyright statement and information on IPR
The resource is licensed under META-SHARE Commons BY NC SA.

3. TECHNICAL INFORMATION

3.1 Directories and files
1 file (xls)

3.2 Data structure of an entry
6 columns: Company ID, Company name, Registration date, Company address, Company locality, Company postcode

3.3 Lexicon size (nmb. of lexical items, KB occupied on disk)
45,993 entries

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon
English, Maltese

4. 2 Entry Type
see 3.2

4.3 Attributes and their values
n.a.
4.4 Coverage of the lexicon
n.a.
4.5 Intended application of the lexicon
named entity recognition

4.6 POS assignment
n.a.
4.7 Reliability (automatically/manually constructed)
unknown

5. RELEVANT REFERENCES AND OTHER INFORMATION

UPC – Universitat Politècnica de Catalunya

Endogenous resources

ALBAYZIN

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

This corpus consists of 3 sub-corpora of 16 kHz 16 bits signals, recorded by 304 Castilian Spanish speakers. The 3 sub-corpora are: 1) Phonetic corpus: 6,800 utterances of phonetically balanced sentences, including 1000 with phonetic segmentation. 2) Geographic corpus: 6,800 utterances of sentences extracted from a Spanish geographic database. and 3) "Lombard" corpus: 2,000 utterances corresponding to sentences from the other two corpora.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Climent Nadeu
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 401 6438
Fax: +34 93 401 6447
e-mail: climent.nadeu@upc.edu

2.2 . Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be downloadable as an archive.

2.3 . Copyright statement and information on IPR

The resource has copyright. The Copyright belongs to Universitat Politècnica de Catalunya, Universidad Politécnica de Madrid, Universidad Politécnica de Valencia, Universidad de Granada, and Universitat Autònoma de Barcelona. The resource is fee, license-based, for research and commercial purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The database includes documentation (copyright, readme, specification documents, …), speech files (one per utterance) and label files (each speech file has an accompanying label file) in a nested structured directory.

3.2. Encoding

Documentation is encoded in txt, word and pdf text.

Speech files (extension .ses) are stored as sequences of 16 kHz 16 bits uncompressed speech samples.

Each speech file has an accompanying ASCII SAM label file (extension .seo) which includes the prompted sentence, number of samples, etc.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The corpus contains a set of 15600 speech utterance files, which corresponds to about 2 GB.

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database for training ASR systems in Spanish. Standards proposed in the EU-funded project SAM (ESPRIT project 2589) have been followed closely for the design, recording and annotation of the database.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

The database is divided in three sections: a) a phonetic database, b) an application database, c) a Lombard speech database.

a) The corpus of the phonetic database is divided in two parts:

Subcorpus 1: Composed by utterances from a set of 200 phonetically balanced sentences. The meaning of "phonetically balanced" is described below. 4 speakers utter the overall set, and each of 160 speakers utter a phonetically balanced subset of 25 sentences. Consequently, 24 utterances result for each sentence. Subcorpus 1 is intended for training purposes.

Subcorpus 2: Composed by utterances from a set of 500 phonetically balanced sentences. 40 speakers utter a group of 50 sentences, so each sentence is uttered by 4 speakers. Subcorpus 2 is designed to test the phonetic decoding performance of a recognition system.

b) In the application database, which is based on the task of information retrieval from a geographic database, the starting point is a textual corpus of 3900 sentences. The corpus is divided in a training set, composed by 2700 sentences, and a test set composed by 1200 sentences. The corpus

has been divided into 78 subsets of 50 sentences each. Each speaker from a set of 136 speakers (48 are shared with the phonetic corpus) utters a subset of 50 sentences.

c) The Lombard database is produced from a subset of sentences from the above mentioned databases. Each of 40 speakers (shared with the other corpora; 20 from each one) utters 25 sentences. During recordings, a noise source is applied to the speakers via headphones to produce the Lombard effect.

The total number of utterances is 15600 (6800+6800+2000). Their average duration is 4 seconds. So there are around 17 hours of speech.

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

The annotation files (which use SAM extension .seo), one for each recorded utterance, include the prompted sentence as well as other information about the recording session (time, place,…) and the signal (sampling frequency, beginning and ending times, min and max levels, etc).

In the documentation, the following annotations can be found:

- Strong departs from the reference phonetic transcription of the sentence.
- Phonetic segmentation and labeling of the speech signals. It was done for 1000 utterances from the phonetic corpus, that correspond to 40 speakers out of the 160 speakers from subcorpus 2.
- Semantic transcription of all the sentences of the geographic application corpus.

4.4. Lexicon. Description of the lexicon (if applicable)

The database doesn't include a lexicon. The process of designing the set of sentences for the phonetic corpus is included in the documentation.

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

The total number of speakers is 304, half women and half men, with a balanced coverage of ages from 18 to 55.

All the selected speakers are from the central dialect of Castilian Spanish, spoken in Castilla-La Mancha, Castilla-León, Cantabria o Madrid. Those speakers from this dialect that showed specific features from a given geographic area or from a restricted social group were removed. Half of the selected speakers are less than 31 years old.

4.6. Recording platform

4.6.1.Domain(s), environments,

Recordings were performed in an acoustically conditioned room.

4.6.2 Recording platform

- Microphone: headphone Shure SM 10
- Acquisition board: OROS AU-22 (SAM compatible)
- Device compatible with SONY multifunction optical disk.
- DAT for acoustic recording of the evolution of each of the recording sessions.
- Software EUROPEC from SAM project

All signal files have 200 ms of silence (environment sound) at their beginning and end. Endpoints have been detected and manually validated for all signals.

Signals have been low-pass filtered to remove very low frequency noise included in the recording process. The filter was designed to cancel (40dB) a frequency component around 16 Hz, and to pass with low attenuation (0.1dB) frequency components higher than 150 Hz.

TM2: technical meetings

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

This resource is intended to upgrade the CHIL2007+ corpus with two newly recorded and annotated audiovisual technical meetings. The new recordings are in Spanish and Catalan. A relevant objective is to include situations in which the semantic content can not be extracted from only a single modality.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person:

Name: Climent Nadeu
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 4016438
Fax: +34 93 401 6447
e-mail: climent.nadeu@upc.edu

2.2. Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive.

2.3. Copyright statement and information on IPR

The resource belongs to UPC. The resource is free, license-based, for research and commercial purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The archive contains signal files, annotation files, and general documentation files, in a nested directory structure.

3.2. Encoding

Audio signals are in WAV format, at 44.1 KHz and 24 bits/sample.

Video signals are sequences of JPEG-compressed images, at a resolution of 752x582 pixels at 25 files per second.

Three different label file formats are used to annotate the corpus:

- XML format.
- ELAN XML [12] format.
- Text ASCII format.

Documentation is encoded in word/pdf/plain text.

3.3. Resource size (size of recorded data/MB occupied on disk)

The corpus contains about 175 GB of data.

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

The corpus is intended for doing research on video, audio and speech technologies, and in particular on the integration of their outputs for a multi-level based scene analysis. It was designed to upgrade and extend the already existing CHIL2007 corpus. While that existing corpus was in English, the new recorded data has been produced in two other languages, Catalan and Spanish. A relevant objective in the two new recorded sessions is to include situations in which the semantic content can not be extracted from only a single modality. Also, annotations have been largely extended with respect to the original CHIL2007 corpus.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

The corpus consists of two technical meetings of about 30' each, one in Spanish and the other in Catalan. Four speakers talk rather spontaneously, and long speeches are usual. There is also activity in terms of people entering/leaving the room, opening and closing the door, standing up and going to the screen, interaction among the attendees, coffee breaks, etc.

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

The CHIL2007 corpus includes manual annotations from both audio and video modalities.

It contains a detailed multichannel verbatim orthographic transcription of the audio modality, which, besides speech transcription, includes speaker turns and identities, speech endpoints, vocalizations (e.g. <uh>, <uhm>, <Smack>, <B>), and acoustic events (from a set of 12 predefined events, like cough, laugh, door slam, chair moving,…). Also named entities and topics have been labeled from the ortographic transcriptions of speech.

Video annotations provide 3D multiperson head locations, movement, focus of attention, hand gestures, head gestures, and spatial role labeling (spatial relations).

From both modalities, annotations are included regarding activity classification, emotions, and links between different tiers.

The annotations were validated internally using inter-annotator agreement, by taking into consideration 6.6% of the total amount of annotated data.

4.4. Lexicon. Description of the lexicon (if applicable)

Not applicable

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

4 speakers per session, 6 speakers in total. The meeting participants were staff members and students. One speaker, who participated in both sessions was non-native in both languages.

4.6. Recording platform

4.6.1.Domain(s), environments,

Smart-room equipped with several cameras and microphones on the walls.

4.6.2. Recording platform

A set of audio sensors:

- 6 clusters of four-channel T-shaped omnidirectional microphone (24 microphones);
- 4 tabletop directional cardioid microphones;
- 4 close-talking directional wireless microphones.

A set of video sensors:

- four fixed cameras located at the room corners;
- one fixed, wide-angle panoramic camera located under the room ceiling.

For audio data capture, all microphones were connected to a number of RME Octamic eight-channel pre-amplifiers/digitizers. The pre-amplifier outputs were sampled at 44.1 kHz and 24 bits per sample, and were recorded to a computer in WAV format via an RME Hammerfall HDSP9652 I/O card.

The cameras provide images of 752x582 pixels, and frame rate 25 fps. All video streams were saved as sequences of JPEG-compressed images.

CHIL2007+

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

This resource includes the already existing CHIL2007 Evaluation Package distributed by ELRA and extends largely its annotations. The CHIL2007 database contains audiovisual recordings of English seminars produced by several partners of the EU-funded CHIL project in their smart-rooms. It was intended to evaluate audio and video technologies developed during the project lifetime. Data, annotations, evaluation tools and documentation are provided. The upgraded CHIL2007+ version includes up to 9 new annotation tiers.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person for the new annotations:

Name: Climent Nadeu
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 4016438
Fax: +34 93 401 6447
e-mail: climent.nadeu@upc.edu

2.2. Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be downloadable as an archive.

2.3. Copyright statement and information on IPR

The resource belongs to the CHIL Consortium. The resource is fee, license-based, for research and commercial purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The archive contains signal files, annotation files, and general documentation files, in a nested directory structure.

3.2. Encoding

Audio signals are in either WAV or SPHERE format, at 44.1 KHz and 24 bits/sample.

Video signals are sequences of JPEG-compressed images.

Documentation is encoded in word/pdf/plain text.

3.3. Resource size (size of recorded data/MB occupied on disk)

The corpus contains about 900 GB of data.

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

The CHIL2007 corpus was produced within the CHIL project and includes audiovisual recordings of scientific presentations given by students, faculty members or invited speakers in the field of multimodal interfaces and speech processing. It was intended to evaluate audio and video technologies developed during the project lifetime. The objective of the project was to create environments in which computers serve humans who focus on interacting with other humans as opposed to having to attend to and being preoccupied with the machines themselves.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

The CHIL2007 Evaluation Package includes a set of audiovisual recordings of interactive seminars produced by AIT (Athens Information Technology), IBM, ITC (Instituto Trentino di Cultura), UKA (Universität Karlsruhe), and UPC (Universitat Politècnica de Catalunya) in their smart-rooms. Each seminar usually consists of a presentation of 20 to 30 minutes to a group of three to five attendees in a meeting room. During and after the presentation, there are questions from the attendees with answers from the presenter. There is also activity in terms of people entering/leaving the room, opening and closing the door, standing up and going to the screen, discussion among the attendees, coffee breaks, etc. There are 2 types of audio-visual documents presented in CHIL2007: full sessions of about 20 minutes and excerpts of 5 min. The full sessions contain the the whole meeting session. The excerpts include short and most prominent pieces of meeting where interactions, discussions and movements happen frequently. The database includes 1 full session and 8 excerpts from each participant site.

As CHIL2007 is an evaluation package, it also contains complete documentation about evaluations that were carried out with the provided database, including the definition and description of the evaluation methodologies, protocols, and metrics, as well as the software scoring tools necessary to evaluate developed systems in a given technology.

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

The CHIL2007 corpus includes manual annotations from both audio and video modalities.

It contains a detailed multichannel verbatim orthographic transcription of the audio modality, which, besides speech transcription, includes speaker turns and identities, speech endpoints, vocalizations (e.g. <uh>, <uhm>, <Smack>, <B>), and acoustic events (from a set of 12 predefined events, like cough, laugh, door slam, chair moving,…).

Video labels provide 3D multiperson head locations.It also includes information on the seminars setup plus background pictures and calibration data.

All the annotations were validated in various ways, internally in the CHIL project, using cross-validations between three different CHIL partners.

In the extended version CHIL2007+, the following new annotation tiers have been included: movement, focus of attention, hand gestures, head gestures, spatial role labeling (spatial relations), activity classification, emotions, named entities, and topics.

4.4. Lexicon. Description of the lexicon (if applicable)

Not applicable

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

The number of people performing in the room during recordings was chosen to be between 3 and 7. The seminar participants were mainly staff and students from the respective recording sites, but national and international visitors participated also. This led to a broad variety in participants' nationalities and to many different native and nonnative accents in English in the collected data. CHIL2007 has 71 individual speakers, including only five female voices, and comprises speakers originating from 17 different countries, with the biggest groups being Spaniards (23%), Italians (15%), and Greeks and Germans (each 14%).

4.6. Recording platform

4.6.1.Domain(s), environments,

Smart-rooms equipped with several cameras and microphones on the walls.

4.6.2. Recording platform

The smart-rooms and recording platforms differ among the five sites. Though all five sites complied with a set of minimum requirements, but often added additional sensors. The minimal setup consisted of:

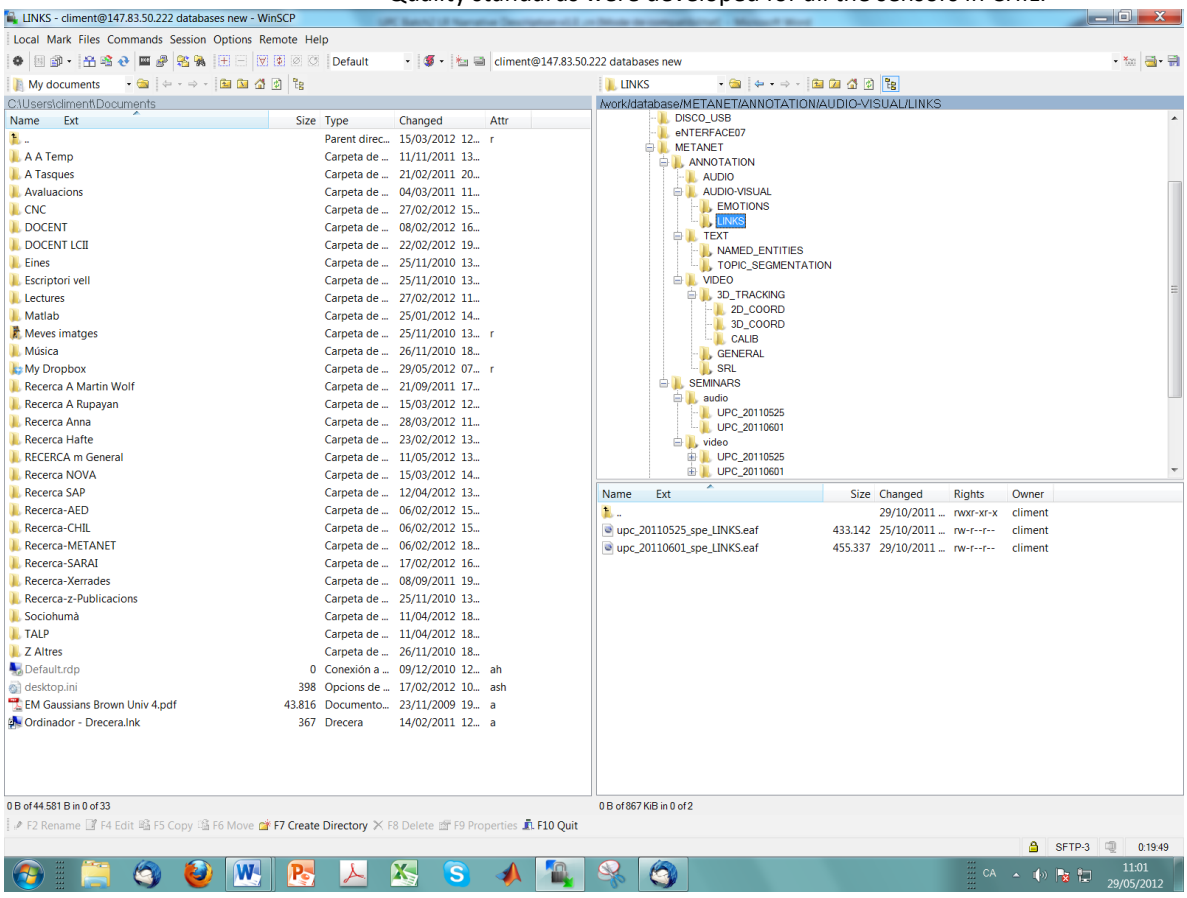
- A set of audio sensors:
  - a 64-channel linear NIST Mark III microphone array;
  - three four-channel T-shaped microphone clusters;
  - three tabletop microphones;
  - close-talking microphones worn by the lecturer and each of the meeting participants.
- A set of video sensors:

- four fixed cameras located at the room corners;
- one fixed, wide-angle panoramic camera located under the room ceiling;
- one active pan-tilt-zoom camera.

For audio data capture, all microphones beside the NIST Mark III microphones were connected to a number of RME Octamic eight-channel pre-amplifiers/digitizers. The pre-amplifier outputs were sampled at 44.1 kHz and 24 bits per sample, and were recorded to a computer in WAV format via an RME Hammerfall HDSP9652 I/O card. The 64-channel NIST Mark III data were similarly sampled and recorded in SPHERE format, but were fed into a recording computer via an Ethernet connection in the form of multiplexed IP packets.

The type of cameras installed varied among the sites, being either firewire or analog, providing images in resolutions ranging from 640x480 to 1024x768 pixels, and frame rates from 15 to 30 fps. All video streams were saved as sequences of JPEG-compressed images.

Quality standards were developed for all the sensors in CHIL.



Catalan SpeechDat (I)

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

The SpeechDat Catalan FDB database contains the recordings of 1,005 Catalan speakers (474 males, 531 females) recorded over the Spanish fixed telephone network. The database is partitioned into 4 CD-ROMs, in ISO 9660 format. Speech samples are stored as sequences of 8-bit 8 kHz A-law, uncompressed. Each prompted utterance is stored in a separate file, and each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information. Each speaker uttered about 41 items. A pronunciation lexicon with the phonetic transcription in SAMPA is also included.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Climent Nadeu
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 401 6438
Fax: +34 93 401 6447
e-mail: climent.nadeu@upc.edu

2.2 . Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive.

2.3 . Copyright statement and information on IPR

The resource has copyright. The Copyright belongs to Universitat Politècnica de Catalunya. The resource is fee, license-based, for research and commercial purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The database includes documentation (copyright, readme, documents, …), speech files (one per utterance) and label files (each speech file has an accompanying label file) in a nested structured directory.

3.2. Encoding

Documentation is encoded in word and pdf text.

Speech files are stored as sequences of 8-bit 8 kHz A-law uncompressed speech sample. Each prompted utterance is stored within a separate file.

Each speech file has an accompanying ASCII SAM label file.
Label files contain information about the database, speech signal coding, speakers, labeling session, transcriptions and annotations.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The corpus contains about 36.000 speech files

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database for training ASR systems in Catalan. The database fulfills the SpeechDat (www.speechDat.org) specifications.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

Each speaker utters 41 items:

- 3 application words
- 1 sequence of 10 isolated digits
- 4 connected digits (prompt sheet number -6 digits, telephone number –9/11 digits, credit card number –14/16 digits, PIN code –6 digits)
- 3 dates (spontaneous date e.g. birthday, prompted date, relative and general date expression)
- 1 word spotting phrase using embedded application words
- 1 Internet address (either URL or e-mail address)
- 1 isolated digit
- 3 spelled words (1 surname, 1 directory assistance city name, 1 real/artificial name for coverage)
- 1 currency money amount
- 1 natural number
- 5 directory assistance names (1 spontaneous, e.g. own surname, 1 city of birth/growing up, 1 most frequent city out of a set of 500, 1 most frequent company/agency out of a set of 500, 1 “forename surname” out of a set of 150)
- 2 yes/no questions (1 predominantly “yes” question, 1 predominantly “no” question, including fuzzy questions)
- 9 phonetically rich sentences
- 2 time phrases (1 spontaneous time of day, 1 word style time phrase)
- 4 phonetically rich words.

Utterances are both, read and spontaneous

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

The transcription included in this database is an orthographic, lexical transcription with a few details that represent audible acoustic events (speech and non speech) present in the corresponding waveform files. The extra marks contained in the transcription aid in interpreting the text form of the utterance. Transcriptions were made in two passes: one pass in which words are transcribed, and a second pass in which the additional details are added.

Transcriptions were checked periodically for quality control. Two persons transcribed samples of the same transcription task and results were compared

4.4. Lexicon. Description of the lexicon (if applicable)

The lexicon file is an alphabetically ordered list of distinct lexical items (i.e. splitting tokens at every space, essentially words in our case) which occur in the corpus with the corresponding pronunciation information. Each distinct word has a separate entry. As the lexicon is derived from the corpus it uses the same alphabetic encoding for special and accented characters as used in the transcriptions

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

The database contains recordings from 1005 adult speakers.

All the speakers are Catalan native. Speakers were selected from the two main Catalan dialects spoken in Catalonia: central and western.

Age distribution: 13 speakers are under 16, 473 speakers are between 15 and 29 years old, 286 speakers are between 30 and 45, 192 speakers are between 46 and 60, and 41 speakers are over 60.

4.6. Recording platform

4.6.1.Domain(s), environments,

Speakers were calling from the fixed telephone network

4.6.2 Recording platform

The recording platform is based on a PC with an ISDN-BRI interface.

SALA-Mexico

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

The SALA Spanish Mexican Database comprises 1260 Mexican speakers (554 males, 706 females) recorded over the Mexican fixed telephone network. The speech databases made within the SALA project were validated by SPEX, the Netherlands, to assess their compliance with the SALA format and content specifications.The speech files are stored as sequences of 8-bit, 8kHz A-law speech files and are not compressed, according to the specifications of SALA. Each prompt utterance is stored within a separate file and has an accompanying ASCII SAM label file.Each speaker uttered 52 items. The following age distribution has been obtained: 20 speakers are under 16 years old, 801 speakers are between 16 and 30, 291 speakers are between 31 and 45, 124 speakers are between 46 and 60, and 24 speakers are over 60. A phonetic lexicon with canonical transcriptions in SAMPA is also provided.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Asuncion Moreno
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 401 6437
Fax: +34 93 401 6447
e-mail: asuncion.moreno@upc.edu

2.2 . Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive.

2.3 . Copyright statement and information on IPR

The resource has copyright. The Copyright belongs to Universitat Politècnica de Catalunya. The resource is fee, license-based, for research and commercial purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The archive that will be uploaded on the MetaShare platform will contain several files containing documentation (copyright, readme, documents, …), speech files (one per utterance) and label files (each speech file has an accompanying label file) in a nested structured directory.

3.2. Encoding

Documentation is encoded in word and pdf text.

Speech files are stored as sequences of 8-bit 8 kHz A-law uncompressed speech samples (CCITT G.711 recommendation). Each prompted utterance is stored within a separate file.

Each speech file has an accompanying ASCII SAM label file
Label files contain information about the database, speech signal coding, speakers, segmentation, labeling session, transcriptions and annotations.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The corpus contains about 75.000 speech files

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database for training ASR systems in Mexican Spanish. The database fulfills the SALA specifications.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

Each speaker uttered the following items: * 6 application words; * 1 sequence of 10 isolated digits; * 4 connected digits: 1 sheet number (6 digits), 1 telephone number (9-11 digits), 1 credit card number (14-16 digits), 1 PIN code (6 digits); * 3 dates: 1 spontaneous date (e.g. birthday), 1 prompted date (word style), 1 relative and general date expression; * 1 spotting phrase using an application word (embedded); * 1 isolated digit; * 3 spelled-out words (letter sequences): 1 spelling of surname; 1 spelling of directory assistance city name; 1 real/artificial name for coverage; * 1 currency money amount; * 1 natural number; * 5 directory assistance names: 1 surname (out of 500); 1 city of birth / growing up (spontaneous); 1 most frequent city (out of 500); 1 most frequent company/agency (out of 500); 1 "forename surname" (set of 150) * 2 questions, including "fuzzy" yes/no: 1 predominantly "yes" question, 1 predominantly "no" question; * 9 phonetically rich sentences; * 9 additional spontaneous items * 2 time phrases: 1 time of day (spontaneous), 1 time phrase (word style); * 4 phonetically rich words. Utterances are both, read and spontaneous

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

The transcription included in this database is an orthographic, lexical transcription with a few details that represent audible acoustic events (speech and non speech) present in the corresponding waveform files. The extra marks contained in the transcription aid in interpreting the text form of the utterance. Transcriptions were made in two passes: one pass in which words are transcribed, and a second pass in which the additional details are added.

Transcriptions were checked periodically for quality control. Two persons transcribed samples of the same transcription task and results were compared

4.4. Lexicon. Description of the lexicon (if applicable)

The lexicon file is an alphabetically ordered list of distinct lexical items (i.e. splitting tokens at every space, essentially words in our case) which occur in the corpus with the corresponding pronunciation information. Each distinct word has a separate entry. As the lexicon is derived from the corpus it uses the same alphabetic encoding for special and accented characters as used in the transcriptions

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

The database contains recordings from 1260 adult speakers. The following age distribution has been obtained: 20 speakers are under 16 years old, 801 speakers are between 16 and 30, 291 speakers are between 31 and 45, 124 speakers are between 46 and 60, and 24 speakers are over 60.

4.6. Recording platform

4.6.1.Domain(s), environments,

Speakers were calling from the fixed telephone network

4.6.2 Recording platform
The recording platform is based on a PC with an ISDN-BRI interface. Recording software is ADA.

SALA- Venezuela

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

The SALA Spanish Venezuelan database contains the recordings of 1,000 Venezuelan speakers (504 males, 496 females) recorded over the Venezuelan fixed telephone network. This database is partitioned into 5 CD-ROMs The speech files are stored as sequences of 8-bit, 8kHz mu-law speech files and are not compressed, according to the specifications of SALA. Each prompt utterance is stored within a separate file and has an accompanying ASCII SAM label file.This speech database was validated by SPEX (the Netherlands) to assess its compliance with the SALA format and content specifications. Each speaker uttered 44 items: * 6 application words * 1 sequence of 10 isolated digits * 4 connected digits (1 sheet number -6 digits, 1 telephone number –9/11 digits, 1 credit card number –14/16 digits, 1 PIN code -6 digits) * 3 dates (1 spontaneous date e.g. birthday, 1 word style prompted date, 1 relative and general date expression) * 1 spotting phrase using an embedded application word * 1 isolated digit * 3 spelled words (1surname, 1 directory assistance city name, 1 real/artificial name for coverage) * 1 currency money amount * 1 natural number * 5 directory assistance names (1 surname out of a set of 500, 1 city of birth/growing up, 1 most frequent city out of a set of 500, 1 most frequent company/agency out of a set of 500, 1 "forename surname" out of a set of 150) * 2 yes/no questions (1 predominantly "yes" question, 1 predominantly "no" question) * 9 phonetically rich sentences * 1 additional sentence * 2 time phrases (1 spontaneous time of day, 1word style time phrase) * 4 phonetically rich words. The following age distribution has been obtained: 7 speakers are under 16, 476 speakers are between 16 and 30, 330 speakers are between 31 and 45, 177 speakers are between 46 and 60, and 10 speakers are over 60.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Asuncion Moreno
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 401 6437
Fax: +34 93 401 6447
e-mail: asuncion.moreno@upc.edu

2.2 . Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive.

2.3 . Copyright statement and information on IPR

The resource has copyright. The Copyright belongs to Universitat Politècnica de Catalunya. The resource is fee, license-based, for research and commercial purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The archive that will be uploaded on the MetaShare platform will contain several files containing documentation (copyright, readme, documents, …), speech files (one per utterance) and label files (each speech file has an accompanying label file) in a nested structured directory.

3.2. Encoding

Documentation is encoded in word and pdf text.

Speech files are stored as sequences of 8-bit 8 kHz mu-law uncompressed speech sample. Each prompted utterance is stored within a separate file.

Each speech file has an accompanying ASCII SAM label file
Label files contain information about the database, speech signal coding, speakers, labeling session, transcriptions and annotations.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The corpus contains about 44.000 speech files

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database for training ASR systems in Venezuelan Spanish. The database fulfills the SALA specifications.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

* 6 application words * 1 sequence of 10 isolated digits * 4 connected digits (1 sheet number -6 digits, 1 telephone number –9/11 digits, 1 credit card number – 14/16 digits, 1 PIN code -6 digits) * 3 dates (1 spontaneous date e.g. birthday, 1 word style prompted date, 1 relative and general date expression) * 1 spotting phrase using an embedded application word * 1 isolated digit * 3 spelled words (1surname, 1 directory assistance city name, 1 real/artificial name for coverage) * 1 currency money amount * 1 natural number * 5 directory assistance names (1 surname out of a set of 500, 1 city of birth/growing up, 1 most frequent city out of a set of 500, 1 most frequent company/agency out of a set of 500, 1 "forename surname" out of a set of 150) * 2 yes/no questions (1 predominantly "yes" question, 1 predominantly "no" question) * 9 phonetically rich sentences * 1 additional sentence * 2 time phrases (1 spontaneous time of day, 1word style time phrase) * 4 phonetically rich words. Utterances are both, read and spontaneous

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

The transcription included in this database is an orthographic, lexical transcription with a few details that represent audible acoustic events (speech and non speech) present in the corresponding waveform files. The extra marks contained in the transcription aid in interpreting the text form of the utterance. Transcriptions were made in two passes: one pass in which words are transcribed, and a second pass in which the additional details are added.
Transcriptions were checked periodically for quality control. Two persons transcribed samples of the same transcription task and results were compared

4.4. Lexicon. Description of the lexicon (if applicable)

The lexicon file is an alphabetically ordered list of distinct lexical items (i.e. splitting tokens at every space, essentially words in our case) which occur in the corpus with the corresponding pronunciation information. Each distinct word has a separate entry. As the lexicon is derived from the corpus it uses the same alphabetic encoding for special and accented characters as used in the transcriptions

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

The database contains recordings from 1000 adult speakers. All the speakers are Venezuela native. Speakers were selected from several different accents from Venezuela. The following age distribution has been obtained: 7 speakers are under 16, 476 speakers are between 16 and 30, 330 speakers are between 31 and 45, 177 speakers are between 46 and 60, and 10 speakers are over 60.

4.6. Recording platform

4.6.1.Domain(s), environments,

Speakers were calling from the fixed telephone network

4.6.2 Recording platform
The recording platform is based on a PC with an ISDN-BRI interface. Recording software is ADA.

Spanish SpeechDat (II)

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

The Castillian Spanish SpeechDat(II) FDB-4000 contains the recordings of 4,000 Castillian Spanish speakers (2,061 males, 1,939 females) recorded over the Spanish fixed networ.Speech samples are stored as sequences of 8-bit 8 kHz A-law. Each prompted utterance is stored in a separate file. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.This speech database was validated by SPEX (the Netherlands) to assess its compliance with the SpeechDat format and content specifications.Each speaker uttered the following 40 items: * 3 application words * 1 sequence of 10 isolated digits * 4 connected digits (1 sheet number -6 digits, 1 telephone number –9/11 digits, 1 credit card number –14/16 digits, 1 PIN code -6 digits out of a set of 150) * 3 dates (1 spontaneous date e.g. birthday, 1 word style prompted date, 1 relative and general date expression) * 1 word spotting phrase using an embedded application word * 1 isolated digit * 3 spelled word (1 surname, 1 directory assistance city name, 1 real/artificial for coverage) * 1 currency money amount * 1 natural number * 5 directory assistance names (1 spontaneous e.g. own forename, 1 city of birth/growing up, 1most frequent city out of a set of 500, 1 most frequent company/agency out of a set of 500, 1 "forename surname" out of a set of 150) * 2 yes/no questions (1 predominantly "yes" question, 1 predominantly "no" question) * 9 phonetically rich sentences * 2 time phrases (1 spontaneous time of day, 1 word style time phrase) * 4 phonetically rich words. The following age distribution has been obtained: 42 speakers are under 16, 2,234 are between 16 and 30, 844 are between 31 and 45, 764 are between 46 and 60, and 116 are over 60.A pronunciation lexicon with a phonemic transcription in SAMPA is also included.This database includes the Spanish SpeechDat(II) FDB-1000 (ref. ELRA-S0101).

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Asuncion Moreno
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 401 6437
Fax: +34 93 401 6447
e-mail: asuncion.moreno@upc.edu

2.2 . Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive.

2.3 . Copyright statement and information on IPR

The resource has copyright. The Copyright belongs to Universitat Politècnica de Catalunya. The resource is fee, license-based, for research and commercial purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The archive that will be uploaded on the MetaShare platform will contain several files containing documentation (copyright, readme, documents, …), speech files (one per utterance) and label files (each speech file has an accompanying label file) in a nested structured directory.

3.2. Encoding

Documentation is encoded in word and pdf text.

Speech files are stored as sequences of 8-bit 8 kHz mu-law uncompressed speech sample. Each prompted utterance is stored within a separate file.

Each speech file has an accompanying ASCII SAM label file
Label files contain information about the database, speech signal coding, speakers, labeling session, transcriptions and annotations.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The resource contains about 160.000 speech files

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database for training ASR systems in Castillian Spanish. The database fulfills the SpeechDat (www.SpeechDat.org) specifications.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

* 3 application words * 1 sequence of 10 isolated digits * 4 connected digits (1 sheet number -6 digits, 1 telephone number –9/11 digits, 1 credit card number – 14/16 digits, 1 PIN code -6 digits out of a set of 150) * 3 dates (1 spontaneous date e.g. birthday, 1 word style prompted date, 1 relative and general date expression) * 1 word spotting phrase using an embedded application word * 1 isolated digit * 3 spelled word (1 surname, 1 directory assistance city name, 1 real/artificial for coverage) * 1 currency money amount * 1 natural number * 5 directory assistance names (1 spontaneous e.g. own forename, 1 city of birth/growing up, 1most frequent city out of a set of 500, 1 most frequent company/agency out of a set of 500, 1 "forename surname" out of a set of 150) * 2 yes/no questions (1 predominantly "yes" question, 1 predominantly "no" question) * 9 phonetically rich sentences * 2 time phrases (1 spontaneous time of day, 1 word style time phrase) * 4 phonetically rich words. Utterances are both, read and spontaneous

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

The transcription included in this database is an orthographic, lexical transcription with a few details that represent audible acoustic events (speech and non speech) present in the corresponding waveform files. The extra marks contained in the transcription aid in interpreting the text form of the utterance. Transcriptions were made in two passes: one pass in which words are transcribed, and a second pass in which the additional details are added.
Transcriptions were checked periodically for quality control. Two persons transcribed samples of the same transcription task and results were compared

4.4. Lexicon. Description of the lexicon (if applicable)

The lexicon file is an alphabetically ordered list of distinct lexical items (i.e. splitting tokens at every space, essentially words in our case) which occur in the corpus with the corresponding pronunciation information. Each distinct word has a separate entry. As the lexicon is derived from the corpus it uses the same alphabetic encoding for special and accented characters as used in the transcriptions

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

The database contains recordings from 4000 adult speakers. All the speakers are Spanish native. Speakers were selected from several different accents from Spain. The following age distribution has been obtained: 42 speakers are under 16, 2,234 are between 16 and 30, 844 are between 31 and 45, 764 are between 46 and 60, and 116 are over 60.

4.6. Recording platform

4.6.1.Domain(s), environments,

Speakers were calling from the fixed telephone network

4.6.2 Recording platform

The recording platform is based on a PC with an ISDN-BRI interface. Recording software is ADA.

SpeechDat-Car Spanish

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

The Spanish SpeechDat-Car database contains the recordings of 306 Spanish speakers from 4 different regions (156 males, 150 females), recorded over the Spanish GSM telephone network, and in a car. This database is partitioned into 89 CDs (DVDs are also available).The speech data files are in two formats. Four of the 5 microphones were recorded on the computer in the boot of the car. The speech data are stored as sequences of 16 kHz, 16 bit and uncompressed. The fifth microphone was connected to the cell phone, and was recorded on a remote machine. The data are stored as sequences of 8 kHz 8 bit A-law. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.This speech database was validated by SPEX (the Netherlands) to assess its compliance with the SpeechDat-Car format and content specifications.Each speaker recorded two sessions. In each session they uttered 129 items:- 2 voice activation keywords- 1 sequence of 10 isolated digits- 7 connected digits (1 sheet number -5 digits, 1 spontaneous telephone number, 3 read telephone numbers, 1 credit card number ?14/16 digits, 1 PIN code -6 digits)- 3 dates (1 spontaneous date e.g. birthday, 1 prompted date, 1 relative or general date expression)- 2 word spotting phrases using an embedded application word- 4 isolated digits- 7 spelled words (1 spontaneous e.g. own forename or surname, 1 directory city name, 4 real word/name, 1 artificial name for coverage)- 1 money amount- 1 natural number- 7 directory assistance names (1 spontaneous e.g. own forename or surname, 1 city of birth/growing up, 2 most frequent cities, 2 most frequent company/agency, 1 ?forename surname?)- 9 phonetically rich sentences- 2 time phrases (1 spontaneous time of day, 1word style time phrase)- 4 phonetically rich words- 67 application words (13 mobile phone application words, 22 IVR function keywords, 32 car products keywords)- 2 additional language dependent keywords- Prompts for spontaneous speechThe

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Asuncion Moreno
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 401 6437
Fax: +34 93 401 6447
e-mail: asuncion.moreno@upc.edu

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive.

2.3 Copyright statement and information on IPR

The resource has copyright. The Copyright belongs to Universitat Politècnica de Catalunya. The resource is fee, license-based, for research and for commercial purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The archive that will be uploaded on the MetaShare platform will contain several files containing documentation (copyright, readme, documents, …), speech files (five per utterance) and label files (each speech file has an accompanying label file) in a nested structured directory.

3.2. Encoding

Documentation is encoded in word and plain text.

Four high quality audio channels are recorded in a car in a mobile platform Plt_M and are stored as sequences of 16bit, 16 kHz uncompressed and multiplexed. Channels are sequentially multiplexed in short unsigned.

One telephone channel is recorded via GSM mobile phone on a stationary ISDN speech server Plt_F. Speech files are stored as sequences of 8-bit 8 kHz A-law uncompressed speech samples (CCITT G.711 recommendation).

Each speech file has an accompanying ASCII SAM label file
Label files contain information about the database, speech signal coding, speakers, labeling session, transcriptions and annotations.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The corpus contains about 77.400 x 5 speech files of recorded speech.

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database for training ASR systems in Spanish for in-car applications and GSM applications. The database fulfills the Speechdat Car (www.speechdat.org) specifications.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

In each session, each speaker uttered the following items:
- 2 voice activation keywords
- 1 sequence of 10 isolated digits

- 7 connected digits (1 sheet number -5 digits, 1 spontaneous telephone number, 3 read telephone numbers, 1 credit card number ?14/16 digits, 1 PIN code -6 digits)
- 3 dates (1 spontaneous date e.g. birthday, 1 prompted date, 1 relative or general date expression)
- 2 word spotting phrases using an embedded application word
- 4 isolated digits
- 7 spelled words (1 spontaneous e.g. own forename or surname, 1 directory city name, 4 real word/name, 1 artificial name for coverage)
- 1 money amount
- 1 natural number
- 7 directory assistance names (1 spontaneous e.g. own forename or surname, 1 city of birth/growing up, 2 most frequent cities, 2 most frequent company/agency, 1 ?forename surname?)
- 9 phonetically rich sentences
- 2 time phrases (1 spontaneous time of day, 1word style time phrase)
- 4 phonetically rich words
- 67 application words (13 mobile phone application words, 22 IVR function keywords, 32 car products keywords)
- 2 additional language dependent keywords
- 10 Prompts for spontaneous speech

Utterances are both, read and spontaneous

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

The transcription included in this database is an orthographic, lexical transcription with a few details that represent audible acoustic events (speech and non speech) present in the corresponding waveform files. The extra marks contained in the transcription aid in interpreting the text form of the utterance. Transcriptions were made in two passes: one pass in which words are transcribed, and a second pass in which the additional details are added.

Transcriptions were checked periodically for quality control. Two persons transcribed samples of the same transcription task and results were compared

4.4 Lexicon. Description of the lexicon (if applicable)

The lexicon file is an alphabetically ordered list of distinct lexical items (i.e. splitting tokens at every space, essentially words in our case) which occur in the corpus with the corresponding pronunciation information. Each distinct word has a separate entry. As the lexicon is derived from the corpus it uses the same alphabetic encoding for special and accented characters as used in the transcriptions

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

database contains the recordings of 306 Spanish speakers from 4 different regions (156 males, 150 females). The following age distribution has been obtained: 160 speakers are

between 18 and 30, 80 speakers are between 31 and 45, 65 speakers are between 46 and 60, and 1speaker is over 60.

4.6. Recording platform

4.6.1.Domain(s), environments,

All the recordings were performed in cars for 4 or 5 passengers. There are defined 6 environment conditions:

1. car stopped by motor running,
2. car in town traffic,
3. car in town traffic,
4. car moving at a low speed with rough road conditions,
5. car moving at a low speed with rough road conditions
6. car moving at a high speed with good road conditions

In addition, some information was collected during the recordings:

- Weather conditions : rain, sun chine, wind …
- Accessories used during recordings: windscreen wipers, ventilation, fan, radio …
- Level of fan: on/off

4.6.2 Recording platform

Two types of recordings compose the database. First, wideband recordings (60-7000 Hz) for systems which are installed and operate in the car itself; second, narrow band recordings (300-3400 Hz) for systems that operate centrally outside the car and obtain their spoken input from the driver over the cellular telephone network. Two recording platforms were used

A 'mobile' recording platform (PltM) installed inside the car, recording multi-channel speech utterances in a high bandwidth mode (16 kHz sample frequency).

A 'fixed' recording platform (PltF) located at the far-end fixed side of the GSM communications simultaneously recording the speech utterances coming from the car (8 kHz sample frequency, A law encoding).

Multi-channel recordings are performed simultaneously in the car and through the GSM network. The recordings are made through an Acoustic front-end (AFE) installed inside the car and connected to the recording platform PltM.. Three kinds of AFEs are used simultaneously during the recordings: a close-talking microphone, a remote noise cancelling microphone with 3 Handsfree microphones placed at different locations in the car and a commercial Handsfree car-kit equipment for GSM radiotelephones in cars.

The synchronisation mode between the PltM and PltF is based on use of DTMF tones emitted from the GSM terminal placed in car.

The mobile recording platform in the car (PltM) is the Master platform. It uses a PC to drive the recording process and to control the remote PltF platform. Data

Acquisition is performed by a dedicated hardware in the PC and the storage is made directly on hard disk. The recordings are always made on four channels (1 close-talk signal as reference and 3 far-talk signals). The positions for the far-talk microphones are:

- A_Column: at the ceiling of the car near the A-pillar
- Sunvisor: at the ceiling of the car in front of the driver behind the sunvisor
- Center: at the ceiling of the car over the mid-console (near the rear mirror)

The GSM phone with hands-free car-kit and special hands-free microphone is installed in the car. The hands-free microphone is mounted at the ceiling of the car over the mid-console, i.e. beside the far-talk microphone position Center.

The remote control of the GSM phone comprises the following basic functions:

- Make a call
- Generate DTMF tones
- Hang up
- Detect dropped call

The remote control of these functions is performed by a simple Windows NT program using the standard AT command set without using any special API. The GSM mobile mounted in the car-kit is connected to the serial port of the PC. The computer for in-car use powered by the car battery (12V DC), is mounted in the boot of the car. The PC features a Pentium II 266MHz processor, 64MB RAM, 4.5 GB SCSI hard disk drive, CPU board with sound chip. (Operating system is Windows NT4). It hosts the data acquisition board and anti-aliasing filter board. As backup medium an internal 2GB SCSI JAZ drive or a removable SCSI Hard Disk Drive is used. A flat panel TFT colour-display for in-vehicle use is attached to the windscreen or the dashboard of the car.

The data acquisition board installed in the Car-PC is a combination of two plug-in boards:

- Multifunction data acquisition board
- Anti-aliasing filter board

The software has been developed by MATRA and can be decomposed into following independent tasks:

- ***Telephone control API***
- ***Multi-channel board recording API***
- ***User Interface (MMI)***

Prompt file management

The fixed recording platform, located at the far-end fixed side of the GSM communication, record simultaneously the speech utterances coming from the car. The software was developed by UPC. The main characteristics of PltF are:

Direct connection to an ISDN line
Recording of speech

DTMF detection (simultaneously with recording of speech)
Full duplex operation (record while playing).

A synchronization and communication protocol between the two platforms were defined. The objectives of this protocol were:

- Detecting if PltF is still alive during the recordings (and to repair a hang up);
- Allowing synchronisation of the recordings on the two platforms;
- Allowing the separation of the items in individual files.

The protocols comprise a series of beeps and DTMF-codes transmitted by PltM to PltF to ensure that each recorded item is preceded by a simultaneous beep on all recording channels to allow rapid off-line synchronization of the recordings on both platforms.

DTMF tones used for the communication and synchronization of platforms were recorded in the fixed platform only. Delays and differences in length of the transmitted DTMF tones made them impractical for a further off-line synchronization. The prompt beep used to indicate the starting point of speech recordings is recorded in both platforms and can be use for synchronization of the speech signals if necessary.

Speecon Catalan

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

The Catalan Speecon database comprises the recordings of 550 adult Catalan speakers who uttered over 290 items (read and spontaneous). The data were recorded over 4 microphone channels in 4 recording environments (office, entertainment, car, public place). The speech database follows the specifications made within the UE funded Speecon project. The database was validated by UVIGO. The Catalan-Speecon Database was funded by the Catalan Government

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Asuncion Moreno
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 401 6437
Fax: +34 93 401 6447
e-mail: asuncion.moreno@upc.edu

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive.

2.3 Copyright statement and information on IPR

The resource has copyright. The Copyright belongs to Universitat Politècnica de Catalunya. The resource is free, license-based, for research and fee, license-based, for commercial purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The archive that will be uploaded on the MetaShare platform will contain several files containing documentation (copyright, readme, documents, …), speech files (four channels per utterance) and label files (each speech file has an accompanying label file) in a nested structured directory.

3.2. Encoding

Documentation is encoded in word and plain text.

The signals are stored in a raw file format, i.e. without headers in the signal file. Each of the four speech channels is recorded at 16 kHz with 16 bit quantization with the least significant byte first ("lohi" or Intel format) as (signed) integers.

A description of the sample rate, the quantization, and byte order used is held in the SAM label file that corresponds to each speech file. This label file also contains information about the noise level at recording time, and the signal quality value of the speech file.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The corpus contains about 176.000 speech utterances recorded simultaneously by 4 channels.

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database for training ASR systems in Spanish for in-car applications and GSM applications. The database fulfills the Speechdat Car (www.speechdat.org) specifications.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

In each session, the following items were recorded:
- 6 noise recordings
- 1 silence
-30 free spontaneous, rich context items
- 17 elicited spontaneous items
- 30 phonetically rich sentences, read
- 5 phonetically rich words, read
- 31 general words and phrases, read
- 216 ***application specific words and phrases***, read

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

The transcription included in this database is an orthographic, lexical transcription with a few details that represent audible acoustic events (speech and non speech) present in the corresponding waveform files. The extra marks contained in the transcription aid in interpreting the text form of the utterance. Transcriptions were made in two passes: one pass in which words are transcribed, and a second pass in which the additional details are added.
Transcriptions were checked periodically for quality control. Two persons transcribed samples of the same transcription task and results were compared

4.4 Lexicon. Description of the lexicon (if applicable)

The lexicon file is an alphabetically ordered list of distinct lexical items (i.e. splitting tokens at every space, essentially words in our case) which occur in the corpus with the corresponding pronunciation information. Each distinct word has a separate entry. As the lexicon is derived from the corpus it uses the same alphabetic encoding for special and accented characters as used in the transcriptions

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

Database contains the recordings of 550 Catalan speakers from 4 different regions (261 males, 289 females). The following age distribution has been obtained: 311 speakers are between 18 and 30, 169 speakers are between 31 and 45, 54 speakers are between 46 and 60, and 16 speakers over 60.

4.6. Recording platform

4.6.1.Domain(s), environments,

5 recording environments were defined:

Office: An office, i.e. a room where people are working at desks, usually or possibly with a computer. No discussions or meetings should be held in the office during the recordings.

Entertainment (household): Living room i.e. a room with some furniture, places where people may sit down. A table, a TV or some audio equipment may be present. Instead of a living room also a hotel room may be possible.

Car: Vehicle for 4 or 5 passengers

Public places: A very large room (hall) or open-air. A hall should have at least 3 walls and a ceiling; more or less busy people, but not too quiet. An open area has no walls and no closed ceiling. Of course, it can be marked off by the walls of the surrounding building. In such a case, at most 2 walls may be closer than 10 meters. This allows recordings at the corner of two buildings. In all cases trees, small shops, an open cafe area, traffic as well as a pedestrian way are possible.

Each of these recording environments have its specific noise characteristics, number and kind of microphones to be recorded simultaneously and position of the recording platform and microphones.

From the view of recording set-ups, the office and the entertainment (household) environment are treated the same. Therefore, four scenarios with different hardware set-ups apply.

4.6.2 Recording platform

The recording platform is based on a laptop using its two Type II PCMCIA slots as interface to the audio equipment. As operating system Windows ’98 applies; the

interfaces cards rely on it. The recording software used was developed by UPC. Up to four microphones are recorded simultaneously.

Interface Emotional Speech database (Spanish)

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

This database contains the recordings of one male and one female Spanish professional speakers recorded in a noise-reduced room. It consists of the recordings and annotations of read text material in neutral style plus six MPEG expressions and also in fast, slow, softly and loudly speech styles. The text material is composed by 184 items including phonetically balanced sentences, digits and isolated words. The text material was the same for all the modes and styles, giving a total of 3h 59min recorded speech for the male speaker and 3h 53min for the female speaker. The Emotional Speech Synthesis Database was created within the scope of the Interface EU funded project. Databases with same specificactions were created for English, French and Slovenian

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Asuncion Moreno
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 401 6437
Fax: +34 93 401 6447
e-mail: asuncion.moreno@upc.edu

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive.

2.3 Copyright statement and information on IPR

The resource is copyright. The Copyright belongs to Universitat Politècnica de Catalunya. The resource is free, license-based, for research and for commercial purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The archive that will be uploaded on the MetaShare platform will contain several files containing documentation (copyright, readme, prompts, contents, vocabulary and phonetic transcription), and speech files.

3.2. Encoding

Documentation is encoded in plain text.

The speech files are stored as a sequence of 16-bit, 16kHz speech files without header nor compression (Linear PCM, Intel byte format). Each file corresponds to one item (one isolated word or one sentence).

The files are stored as signals (file extension .l16).

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The corpus contains about 8 hours of recorded speech.

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database for training emotion detection and modeling.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

To study only the influence of emotions on speech, the same speakers are asked to utter the same speech material in different emotional styles. Basic specifications concern neutral style and the 6 MPEG4 emotions proposed for video analysis: anger, sadness, joy, surprise, disgust and fear. Neutral style was recorded in five variations: normal, soft, loud, slow and fast

The corpus contains 184 different sentences:
- 100 Affirmative sentences including short and longer ones
- 34 Interrogative and (5) stressed sentences.
- 16 Paragraphs
- 10 Digits
- 24 Isolated Words

The corpus was selected to have occurrences of all the Spanish phonemes in different parts of the sentences. An additional criteria to select the sentences is to have two realizations of all the Spanish diphones that appears in inner word position, that means excluding diphones that only appears from coarticulation effects between words.

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

The contents of the speech file was supervised at recording time. If mispronunciation or other deviation from script were detected, the recordings were redone. No mispronunciations are expected.

4.4 Lexicon. Description of the lexicon (if applicable)

The contents and transcriptions of the complete database is included

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

For each language, speakers are required to be professional actors (one male one female speaker). No further specifications were given (age/accent).

4.6. Recording platform

4.6.1.Domain(s), environments,

All recordings are performed in quiet conditions.

The Spanish database was recorded in a silent room. A wall with a glass window divides the room in two parts. The speaker reads the sentences displayed directly from the PC. To avoid extra noises, the display, PC and recording system were placed in one side of the silent room and the speaker in the other side.

Two operators supervised the recordings at recording time. One of them checked the utterances correspond exactly to the text to be read. The other operator checked the recording system.

4.6.2 Recording platform

Recordings have been made using an electrodynamics microphone AKG 320. Speech signals were first recorded at 32 kHz and down sampled to the required 16kHz samples. Recording levels are adapted to each emotional style to avoid saturations and to keep a high dynamic range. The relative recording gain levels are noted down each time that the platform allows it.

LC-STAR Catalan Phonetic Lexicon

1. BASIC INFORMATION

1.1 Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc),

The LC-STAR Catalan phonetic lexicon comprises more than 100,000 words, distributed over three categories: a set of 53,225 common word entries; a list of closed set (function) word classes containing numbers, letters, abbreviations and specific vocabulary for applications controlled by voice; and a set of 45,306 proper names (including person names, family names, cities, streets, companies and brand names)
The LC-STAR Catalan phonetic lexicon was created within the scope of the LC-STAR project (IST 2001-32216) which was sponsored by the European Commission and the Spanish Government.

1.2 Representation of the lexicon (flat files, database, markup)

The lexicon is provided in XML format and includes phonetic transcriptions in SAMPA.

1.3 Character encoding

The lexica is delivered in UTF16 character encoding [http://www.unicode.org/].

A number of documentation files are provided on the CD-ROM containing the overall database description. They can be in ASCII (.TXT), MS Word (.DOC), in Postscript (.PS), in Portable Document Format (.PDF) or Hypertext Markup Language (.HTM) and stored in the root.

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)

Name: Asuncion Moreno
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 401 6437
Fax: +34 93 401 6447
e-mail: asuncion.moreno@upc.edu

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive.

2.3 Copyright statement and information on IPR

The resource has copyright. The Copyright belongs to Universitat Politècnica de Catalunya. The resource is fee, license-based, for research and for commercial purposes.

3. TECHNICAL INFORMATION

3.1 Directories and files

The lexicon has the following directory structure:

\COPYRIGH.TXT	Copyright notice
\README.TXT	Readme file
\<database>\DOC\	
DESIGN.{PDF|DOC}	Documentation file
SAMPALEX.PDF	List of SAMPA symbols
VALREP.{DOC|PDF}	Validation report by SPEX
\<database>\LEXICON	
LEXIC<nn>.XML	Lexicon files nn
LEXICON.DTD	DTD File

3.2 Data structure of an entry

The structure of an entry is exemplified below:

```
<?xml version="1.0" encoding="UTF-16"?>
<!DOCTYPE LEXICA SYSTEM "LEXICON.DTD">
<LEXICA xml:lang="ES">
  <ENTRYGROUP orthography="història">
    <ENTRY>
      <NOM class="common" gender="feminine" number="singular" appreciative="not_specified"/>
      <LEMMA>història</LEMMA>
      <PHONETIC>i s - " t O - r i - @</PHONETIC>
      <APP>
        <SBD entries="29" type="6.2.1."/>
      </APP>
    </ENTRY>
  </ENTRYGROUP>
  <ENTRYGROUP orthography="hi">
    <ENTRY>
      <PRO person="not_specified" case="not_specified" type="personal" gender="invariant"
number="invariant" politeness="no"/>
      <LEMMA>hi</LEMMA>
      <PHONETIC>" i</PHONETIC>
    </ENTRY>
  </ENTRYGROUP>
  <ENTRYGROUP orthography="adaptar-s'hi">
    <ENTRY_COMP>
      <PHONETIC>@ - D @ p - " t a rr - s i</PHONETIC>
      <ENTRY_EL orthography="adaptar">
        <VER number="not_specified" person="not_specified" mood="infinitive"/>
      </ENTRY_EL>
      <ENTRY_EL orthography="s'">
        <PRO person="3" type="reflexive" gender="invariant" number="invariant"
case="not_specified"/>
      </ENTRY_EL>
      <ENTRY_EL orthography="hi">
```

```
            <PRO person="3" type="personal" gender="invariant" number="invariant"
case="not_specified"/>
          </ENTRY_EL>
        </ENTRY_COMP>
    </ENTRYGROUP>
    <ENTRYGROUP orthography="marques">
        <ENTRY>
            <NOM class="common" gender="feminine" number="plural" appreciative="not_specified"/>
            <LEMMA>marca</LEMMA>
            <PHONETIC>" m a rr - k @ s</PHONETIC>
            <APP>
                <SBD entries="114" type="5.1.4."/>
            </APP>
        </ENTRY>
        <ENTRY>
            <VER tense="present" number="singular" person="2" mood="indicative"
gender="not_specified"/>
            <LEMMA>marcar</LEMMA>
            <PHONETIC>" m a rr - k @ s</PHONETIC>
        </ENTRY>
    </ENTRYGROUP>
    <ENTRYGROUP orthography="Puigmal">
        <ENTRY>
            <NOM class="STR" gender="invariant" number="invariant" appreciative="not_specified"/>
            <LEMMA>Puigmal</LEMMA>
            <PHONETIC>p u dZ - " m a l</PHONETIC>
        </ENTRY>
        <ENTRY>
            <NOM class="GEO" gender="invariant" number="invariant" appreciative="not_specified"/>
            <LEMMA>Puigmal</LEMMA>
            <PHONETIC>p u dZ - " m a l</PHONETIC>
        </ENTRY>
    </ENTRYGROUP>
          </LEXICA>
```

3.3 Lexicon size (nmb. of lexical items, KB occupied on disk)

The lexicon comprises more than 100,000 words, distributed over three categories:

- a set of 53,225 common word entries. This set is extracted from a corpus of more than 20 million words distributed over 6 different domains (sports/games, news, finance, culture/entertainment, consumer information, personal communications). This was done with the aim of reaching a target for each domain of at least 95% self coverage. In addition to extracting word lists from the corpus, a list of closed set (function) word classes are included in the final word list.
- a set of 45,306 proper names (including person names, family names, cities, streets, companies and brand names) divided into 3 domains. Multiple word names such as New_York are kept together in all three domains, and they count as one entry. The 3 domains consist of first and last names (21,868 different entries), place names (8,279 different entries), and organisations (16,004 different entries).

- a list of 7,498 special application words translated from English terms defined by the LC-STAR consortium. This list contains: numbers, letters, abbreviations and specific vocabulary for applications controlled by voice (information retrieval, controlling of consumer devices, etc.).

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

The natural language of the LC-STAR Catalan Phonetic Lexicon is Catalan

4.2 Entry Type

The lexicon is composed by entry groups. An entry group refers to a generic entry in a vocabulary. For each entry group, it is mandatory to specify, among others, one or more entry or compound entry or abbreviation elements. An entry refers to one specific grammatical/ morphological meaning of a vocabulary entry. So, and entry group (i.e. a word form) can be multiple tagged.

As is defined in [2], lexica consist of a set of entry group elements.

- An entry group refers to a generic entry in a vocabulary. For each entry group, it is mandatory to specify orthography; zero or more alternative spelling elements; one or more entry or compound entry or abbreviation elements.
- An entry refers to one specific grammatical/morphological meaning of a vocabulary entry. For each entry, it is mandatory to specify: one POS; one lemma; and one phonetic transcription.
- For application words, one APP tag has to be specified.
- Compound entries have the following structure: phonetic transcription; two or more entry elements (a subset of an entry).
- Abbreviations from application wordlists are tagged using the ABB tag.
- Each attribute has the default value NS (=Not Specified), which is always implied.
- In each entry the possibility of inserting a comment is also provided by the XML formalism <!—insert here your comment --> that can be used in any part of the Lexica.

4.3 Attributes and their values

As defined in [2] the following attrbutes for POS have been defined:

NOM (Common and proper nouns): Class, Number, Gender, Person, Case, Type, Appreciative, Possessive_agreement

ADJ (Descriptive/qualificative adjective): Number, Gender, Case, Degree, Form, Type, Appreciative, Possessive_agreement

DET (Determinative adjective or determiner): Number, Gender, Person, Case, Type, Degree

NUM (Numeral adjective or numerals): Number, Gender, Case, Type

VER (Verb): Number, Gender, Person, Case, Mood, Tense, Voice, Polarity, Aspect, Form, Degree, Copula, Type

AUX (Auxiliary verb): Number, Gender, Person, Case, Mood, Tense, Voice, Aspect, Degree, Type, Form
PRO (Pronoun): Number, Possessive_agreement, Gender, Person, Case, Type, Politeness
ART (Article): Number, Gender, Case, Type
ADV (Adverbs and adverbial phrases): Degree, Type
CON (Conjunctions and conjunctive phrases): Degree
ADP (Adpositions and prepositional phrases) , Number, Gender, Person, Type
INT (Interjection)
PAR (Particles and clitics): Number, Person, Tense, Mood,
PRE (Predicative)
ONO (Onomatopoeia words)
MEW (Measure words)
AUW (Auxiliary words)
IDI (Idiom)
PUN (Punctuation marks)
ABB (Abbreviations)
COMPOUND TAGS

4.4 Coverage of the lexicon

The lexicon comprises a set of more than 100,000 entries. Entries were generated from more than 45,000 common words entries selected from a corpora of more than 20 million words distributed in appropriated domains, a set of more than 45,000 names including person names, family names, cities, streets, companies and brand names and a list of 5,000 Special Application Words translated from English terms defined by the LC-STAR consortium.

4.5 Intended application of the lexicon

The lexicon is intended for speech recognition and speech synthesis systems support

4.6 POS assignment

The following references were used:

- Fabra, Pompeu: Gramàtica catalana, 4th ed. Editorial Teide, Barcelona, 1968
- Fabra, Pompeu: Gramàtica catalana, 4th ed, Institut d'estudis catalans, Palau de la Diputació, Barcelona ,1971 (MCMXXVI)
- Jané, A: Gramàtica Catalana, Salvat Editores, S.A. Barcelona, 1968
- Xuriguera, Joan Baptista: Els verbs catalans conjugats, Editorial Claret, Col-leció Pompeu Fabra #6, Barcelona, 1997.

4.7 Reliability (automatically/manually constructed)

Electronic Dictionaries were used to uniform spellings

Phonetic transcription was based in an in-house Catalan grapheme to phoneme transcriber and transcription rules and exceptions were discussed with that expert linguistics that did the transcription of proper names and foreign words

Tagging was done automatically and fully manually supervised. Clitics were manually done.

A 10% of all the manual work was double checked every week. If some discrepancies between linguists were detected, the work was redone

5. RELEVANT REFERENCES AND OTHER INFORMATION

[1] Ziegenhain, U. et al. “Specification of corpora and word lists in 12 languages”. LC-STAR Project Deliverable D1.1.

[2] Maltese, G. Montecchio, C. et al. “General and language specific specification of contents of lexica”. LC-STAR Project Deliverables D2. 2003.

LC-STAR Spanish Phonetic Lexicon

1. BASIC INFORMATION

1.1 Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc),

The LC-STAR Spanish phonetic lexicon comprises more than 100,000 words, distributed over three categories: a set of 55,854 common word entries; a list of closed set (function) word classes containing numbers, letters, abbreviations and specific vocabulary for applications controlled by voice; and a set of 45,403 proper names (including person names, family names, cities, streets, companies and brand names)
The LC-STAR Spanish phonetic lexicon was created within the scope of the LC-STAR project (IST 2001-32216) which was sponsored by the European Commission and the Spanish Government.

1.2 Representation of the lexicon (flat files, database, markup)

The lexicon is provided in XML format and includes phonetic transcriptions in SAMPA.

1.3 Character encoding

The lexica is delivered in UTF16 character encoding [http://www.unicode.org/].

A number of documentation files are provided on the CD-ROM containing the overall database description. They can be in ASCII (.TXT), MS Word (.DOC), in Postscript (.PS), in Portable Document Format (.PDF) or Hypertext Markup Language (.HTM) and stored in the root.

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)

Name: Asuncion Moreno
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 401 6437
Fax: +34 93 401 6447
e-mail: asuncion.moreno@upc.edu

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive.

2.3 Copyright statement and information on IPR

The resource has copyright. The Copyright belongs to Universitat Politècnica de Catalunya. The resource is fee, license-based, for research and for commercial purposes.

3. TECHNICAL INFORMATION

3.1 Directories and files

The lexicon has the following directory structure:

\COPYRIGH.TXT	Copyright notice
\README.TXT	Readme file
\<database>\DOC\	
DESIGN.{PDF|DOC}	Documentation file
SAMPALEX.PDF	List of SAMPA symbols
VALREP.{DOC|PDF}	Validation report by SPEX
\<database>\LEXICON	
LEXIC<nn>.XML	Lexicon files nn
LEXICON.DTD	DTD File

3.2 Data structure of an entry

The structure of an entry is exemplified below:

```
<?xml version="1.0" encoding="UTF-16"?>
<!DOCTYPE LEXICA SYSTEM "LEXICON.DTD">
<LEXICA xml:lang=”ES”>
   <ENTRYGROUP orthography="minotauro">
      <ENTRY>
         <NOM class="common" gender="masculine" number="singular"
appreciative="not_specified"/>
         <LEMMA>minotauro</LEMMA>
         <PHONETIC>m i - n o - " t a w - r o</PHONETIC>
      </ENTRY>
   </ENTRYGROUP>
   <ENTRYGROUP orthography="nosotras">
      <ENTRY>
         <PRO person="1" case="not_specified" type="personal" gender="feminine"
number="plural" politeness="no"/>
         <LEMMA>yo</LEMMA>
         <PHONETIC>n o - " s o - t r a s</PHONETIC>
      </ENTRY>
                                    </ENTRYGROUP>
  <ENTRYGROUP orthography="comprobarlo">
          <ENTRY_COMP>
                  <PHONETIC>k o m - p r o - " B a r - l o</PHONETIC>
                  <ENTRY_EL orthography="comprobar">
                        <VER number="not_specified" person="not_specified"
mood="infinitive" />
                        </ENTRY_EL>
                        <ENTRY_EL orthography="lo">
                              <PRO person="3" type="personal" gender="masculine"
number="singular" />
```

```
                    </ENTRY_EL>
                </ENTRY_COMP>
  </ENTRYGROUP>
  <ENTRYGROUP orthography="compras">
        <ENTRY>
                <NOM class="common" gender="feminine" number="plural"
appreciative="not_specified" />
                <LEMMA>compra</LEMMA>
                <PHONETIC>" k o m - p r a s</PHONETIC>
                <APP>
                        <SBD entries="53" type="2.1.2." />
                        <SBD entries="105" type="3.1.2." />
                </APP>
        </ENTRY>
        <ENTRY>
                <VER tense="present" number="singular" person="2" mood="indicative"
gender="not_specified"/>
                <LEMMA>comprar</LEMMA>
                <PHONETIC>" k o m - p r a s</PHONETIC>
                <APP>
                        <SBD entries="49" type="2.1.4." />
                        <SBD entries="168" type="2.2.2." />
                </APP>
        </ENTRY>
   <ENTRYGROUP orthography="Mónica">
      <ENTRY>
         <NOM class="PER" gender="feminine" number="invariant"
appreciative="not_specified"/>
        <LEMMA>Mónica</LEMMA>
        <PHONETIC>" m o - n i - k a</PHONETIC>
      </ENTRY>
  </ENTRYGROUP>
       </LEXICA>
```

3.3 Lexicon size (nmb. of lexical items, KB occupied on disk)

The lexicon comprises more than 100,000 words, distributed over three categories:

- a set of 55,854 common word entries. This set is extracted from a corpus of more than 37 million words distributed over 6 different domains (sports/games, news, finance, culture/entertainment, consumer information, personal communications). This was done with the aim of reaching a target for each domain of at least 95% self coverage. In addition to extracting word lists from the corpus, a list of closed set (function) word classes are included in the final word list.
- a set of 45,403 proper names (including person names, family names, cities, streets, companies and brand names) divided into 3 domains. Multiple word names

such as New_York are kept together in all three domains, and they count as one entry. The 3 domains consist of first and last names (23,114 different entries), place names (15,423 different entries), and organisations (7,777 different entries).

- a list of 7,498 special application words translated from English terms defined by the LC-STAR consortium. This list contains: numbers, letters, abbreviations and specific vocabulary for applications controlled by voice (information retrieval, controlling of consumer devices, etc.).

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

The natural language of the LC-STAR Spanish Phonetic Lexicon is Castillian Spanish

4.2 Entry Type

The lexicon is composed by entry groups. An entry group refers to a generic entry in a vocabulary. For each entry group, it is mandatory to specify, among others, one or more entry or compound entry or abbreviation elements. An entry refers to one specific grammatical/ morphological meaning of a vocabulary entry. So, and entry group (i.e. a word form) can be multiple tagged.

As is defined in [2], lexica consist of a set of entry group elements.

- An entry group refers to a generic entry in a vocabulary. For each entry group, it is mandatory to specify orthography; zero or more alternative spelling elements; one or more entry or compound entry or abbreviation elements.
- An entry refers to one specific grammatical/morphological meaning of a vocabulary entry. For each entry, it is mandatory to specify: one POS; one lemma; and one phonetic transcription.
- For application words, one APP tag has to be specified.
- Compound entries have the following structure: phonetic transcription; two or more entry elements (a subset of an entry).
- Abbreviations from application wordlists are tagged using the ABB tag.
- Each attribute has the default value NS (=Not Specified), which is always implied.
- In each entry the possibility of inserting a comment is also provided by the XML formalism <!—insert here your comment --> that can be used in any part of the Lexica.

4.3 Attributes and their values

As defined in [2] the following attrbutes for POS have been defined:

NOM (Common and proper nouns): Class, Number, Gender, Person, Case, Type, Appreciative, Possessive_agreement

ADJ (Descriptive/qualificative adjective): Number, Gender, Case, Degree, Form, Type, Appreciative, Possessive_agreement

DET (Determinative adjective or determiner): Number, Gender, Person, Case, Type, Degree

NUM (Numeral adjective or numerals): Number, Gender, Case, Type

VER (Verb): Number, Gender, Person, Case, Mood, Tense, Voice, Polarity, Aspect, Form, Degree, Copula, Type
AUX (Auxiliary verb): Number, Gender, Person, Case, Mood, Tense, Voice, Aspect, Degree, Type, Form
PRO (Pronoun): Number, Possessive_agreement, Gender, Person, Case, Type, Politeness
ART (Article): Number, Gender, Case, Type
ADV (Adverbs and adverbial phrases): Degree, Type
CON (Conjunctions and conjunctive phrases): Degree
ADP (Adpositions and prepositional phrases) , Number, Gender, Person, Type
INT (Interjection)
PAR (Particles and clitics): Number, Person, Tense, Mood,
PRE (Predicative)
ONO (Onomatopoeia words)
MEW (Measure words)
AUW (Auxiliary words)
IDI (Idiom)
PUN (Punctuation marks)
ABB (Abbreviations)
COMPOUND TAGS

4.4 Coverage of the lexicon

The lexicon comprises a set of more than 100,000 entries. Entries were generated from more than 45,000 common words entries selected from a corpora of more than 37 million words distributed in appropriated domains, a set of more than 45,000 names including person names, family names, cities, streets, companies and brand names and a list of 5,000 Special Application Words translated from English terms defined by the LC-STAR consortium.

4.5 Intended application of the lexicon

The lexicon is intended for speech recognition and speech synthesis systems support

4.6 POS assignment

The following references were used:

- María Moliner: Diccionario de uso del espanol.
- Emilio Alarcos Llorach: Gramática de la lengua espanola (Real Academia Espanola, Colección Nebrija y Bello).

4.7 Reliability (automatically/manually constructed)

Electronic Dictionaries were used to uniform spellings

Phonetic transcription was based in an in-house Spanish grapheme to phoneme transcriber and transcription rules and exceptions were discussed with that expert linguistics that did the transcription of proper names and foreign words

Tagging was done automatically and fully manually supervised.

A 10% of all the manual work was double checked every week. If some discrepancies between linguists were detected, the work was redone

5. RELEVANT REFERENCES AND OTHER INFORMATION

[1] Ziegenhain, U. et al. “Specification of corpora and word lists in 12 languages”. LC-STAR Project Deliverable D1.1.

[2] Maltese, G. Montecchio, C. et al. “General and language specific specification of contents of lexica”. LC-STAR Project Deliverables D2. 2003.

UPC ESMA

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

This database contains the recordings of

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Antonio Bonafonte
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 401 0764
Fax: +34 93 401 6447
e-mail: antonio.bonafonte@upc.edu

2.2 . Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive.

2.3 . Copyright statement and information on IPR

The resource is copyright. The Copyright belongs to Universitat Politècnica de Catalunya. The resource is fee, license-based, for research and commercial purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The archive that will be uploaded on the MetaShare platform will contain several files containing documentation (copyright, readme, phoneme statistics), speech files (one per utterance) and manual and automatic label files including the phonetic transcription and segmentation. (each speech file has an accompanying label file).

3.2. Encoding

Documentation is encoded in plain text.

The database is made of three subcorpus: sentences (SE subcorpus, 30 minutes), paragraphs (PA subcorpus, 30 minutes) and literary paragraphs (PL, 45 minutes).

For each utterance, two WAV signals are provided, one for the speech waveform (.wav extension) and one for the laryngograph signal (.lwv extension). These signals are stored as sequences of 16-bit 16 kHz with the least significant byte first ("lohi" or Intel format) as (signed) integers. Each prompted utterance is stored in a

separate file. Furthermore, for each utterance, several label files provide the phonetic transcription, segme

The labelling include the phonetic transcription, phonetic segmentation and CGI (closure glottal instant) labeling. Two sets of transcriptions are provided: manual (supervised) and automatic.

Automatic transcription:
For each utterance the phonetic transcription is derived using Saga (UPC open source tool). HMM based forced-alignment is used to detect pauses and segment the speech files into phonemes.

For each utterance, the following files are provided:

- .txt: prompt text. It includes the tag "<S>" to indicate pauses.
- .pho: phonetic transcription in SAM-PA, including syllable and word bondaries.
- .seg: phonetic segmentation: start and ending time of each phone.
- .prb: probability assigned to each segmentation (useful to detect potencial segmentation problems).
- .cgi: instants of the glottal closure (and implicitely, voiced/unvoiced and pitch labeling).

Manual transcription:
The manual transcription includes, for each utterance, the same files than the automatic transcription, except the .prb. Therefore, the prompt text (.txt), phonetic transcription (.pho) and segmentation (.seg) and the cgi instants (.cgi). However, manual labels are not available for the whole database. In particular, there is no manual information for the PL subcorpus and for some utterances in the PA subcorpus.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The corpus contains about 776 files / 1h 45min of recorded speech and needs about 450MB for disk storage.

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database for training TTS systems in Spanish. The recordings were produced at the same time and by the same female speaker than the Interface Emotional Speech Database.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

The corpus consist of three subcorpora.

- Subcorpus SE: 506 phonetically ballanced sentences designed to maximize the phonetic variability, taking into account the stress and position in the sentence.

- Subcorpus PA: 206 short paragraphs from news.
- Subcorpus PL: 62 long paragraphs. The speaker portrayed a bit the text of this subcorpus.

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

As already mentioned, phonetic transcription, phonetic segmentation, pauses and closure glottal instants (CGI) are labelled automatically for all the files; and manualy for the sentences and some of the short paragraphs.

The automatic phonetic segmentation was created using UPC tools. Phonetic segmentation uses HMM-based forced alignments. In the first step, the HMM toolkit finds the pauses and the pronunciation variants. In the second step, the phonetic segmentation is derived. The automatic CGI labelling used the Praat program.

Note that the manual transcription was generated several years before the automatic transcription. There maybe small inconsistencies in the transcription.

4.4. Lexicon. Description of the lexicon (if applicable)

The lexicon file is an alphabetically ordered list of distinct lexical items (i.e. splitting tokens at every space, essentially words in our case) which occur in the corpus with the corresponding pronunciation information. Each distinct word has a separate entry. As the lexicon is derived from the corpus it uses the same alphabetic encoding for special and accented characters as used in the transcriptions.

4.5. Speaker.

The database was recorded by a female actress, from Barcelona.

4.6. Recording platform

The database was recorded in a isolated room at Universitat Politècnica de Catalunya. The studio includes two isolated rooms, one for the speaker and other for the operators. The speaker recorded simulatenously in two channels: membrane microphone, and laringograph. The recording platform was developed at UPC. It consists of a computer program that allows to navigate through the selected prompts and controls the synchronous recordings (asio interface). The platform shows the recording levels, clipping, etc. The signals were recorded at 32kHz and with 16 bits per sample..

TC-STAR Spanish TTS Baseline female 10h

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

This database contains the recordings of one female Spanish professional speaker recorded in a noise-reduced room simultaneously through a close talk microphone, a mid distance microphone and a laryngograph signal. It consists of the recordings and annotations of read text material of approximately 10 hours of speech for baseline applications (Text-to-Speech systems). The TC-STAR Spanish TTS Baseline Female Speech Database was created within the scope of the TC-STAR project funded by the European Government. The database complies with the common specifications created in the TC-STAR project. The annotation of the database includes manual orthographic transcriptions, the automatic segmentation into phonemes and automatic generation of pitch marks. A certain percentage (20%) of phonetic segments and pitch marks has been manually checked. A pronunciation lexicon in SAMPA with POS, lemma and phonetic transcription of all the words prompted and spoken is also provided.Speech samples are stored as sequences of 24-bit 96 kHz with the least significant byte first ("lohi" or Intel format) as (signed) integers. Each prompted utterance is stored in a separate file. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Antonio Bonafonte
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 401 0764
Fax: +34 93 401 6447
e-mail: antonio.bonafonte@upc.edu

2.2 . Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive.

2.3 . Copyright statement and information on IPR

The resource has copyright. The Copyright belongs to Universitat Politècnica de Catalunya. The resource is fee, license-based, for research and commercial purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The archive that will be uploaded on the MetaShare platform will contain several files containing documentation (copyright, readme, documents, …), speech files (one per utterance) and label files (each speech file has an accompanying label file).

3.2. Encoding

Documentation is encoded in plain text.

Speech samples are stored as sequences of 24-bit 96 kHz with the least significant byte first ("lohi" or Intel format) as (signed) integers. Each prompted utterance is stored in a separate file. Each prompted utterance is stored in three separate file (one file for each channel).

Each speech file has three accompanying label files, identified by the file extesion:

- .esL: ascii file with time of the closure glottal instants (pitch marks). The file was derived automatically from the signal file using praat.
- .esP: ascii file with phonetic transcription and phonetic segmentation. The segmentation has been derived automatically by our in-house HMM-based segmentation tooklkit.
- .esS: SAM label file including the orthograhic information, the phonetic transcription and a rough prosodic labeling.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The corpus contains about 3,700 speech files/ 10 hours of recorded speech and needs about 34 GB for disk storage.

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database for training TTS systems in Spansih. The database is designed based on the TCSTAR specifications for baseline voices. (www.tcstar.org , deliverable document D6).

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

The corpus consist of several subcorpora. Each subcorpora is split in sentences or short paragraphs. Most of corpus was designed to provide high phonetic and prosodic coverage, including spoken style. These subcorpora are built from transcriptons of parliamentary speeches, news, novels. Furthermore some sentences are designed to increase the triphone coverage and to include special words in several domains.

4.3. Transcriptions, annotations:

All the speech data is labeled with orthographic, broad prosodic, phonetic transcription and pitch. The orthographic, prosodic and phonetic transcriptions were supervised manually. The pitch and segmentation labels were produced automatically and twenty per cent of the labels were manually supervised.
The orthographic annotations is a transliteration of what was actually read by the speaker, without ambiguity at the word level. The phonetic transcription annotate what the sepaker really said, including elision, reduction or assimilations. The prosodic annotation consist on minor breaks (intermediate intonational phrases), major breaks (intonational phrases), pitch accent (intonational prominence) normal and emphatic. The phonetic segmentation matches the phonetic transcription. The pitch labelling consist of one temporal mark for each pitch period.

4.4. Lexicon. Description of the lexicon (if applicable)

The lexicon file is an alphabetically ordered list of distinct lexical items (i.e. splitting tokens at every space, essentially words in our case) which occur in the corpus with the corresponding pronunciation information. Each distinct word has a separate entry. As the lexicon is derived from the corpus it uses the same alphabetic encoding for special and accented characters as used in the transcriptions. Each lexicon entry contains the possible phonetic transcription and POS. The lexicon follows the convention defined in the LC-STAR lexica.

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

The database was recorded by a female professional speaker, adult, from the central Castillian variant. Five speakers with the same profile were considered. The selection of this voice took into account different aspects: phonetics, pronunciation/articulatory, appropriateness of the voices for signal processing manipulation and preference tests.

4.6. Recording platform

The database was recorded in a isolated room at Universitat Politècnica de Catalunya. The studio includes two isolated rooms, one for the speaker and other for the operators. The speaker recorded simultaneously in three channels: membrane microphone, close-talk microphone and laringograph. The recording platform was developed at UPC. It consists of a computer program that allows to navigate through the selected prompts and controls the synchronous recordings (asio interface). The platform shows the recording levels, clipping, etc. The signals were recorded at 96kHz and with 24 bits per sample

TC-STAR Spanish TTS Baseline male 10h

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

This database contains the recordings of one male Spanish professional speaker recorded in a noise-reduced room simultaneously through a close talk microphone, a mid distance microphone and a laryngograph signal. It consists of the recordings and annotations of read text material of approximately 10 hours of speech for baseline applications (Text-to-Speech systems). The TC-STAR Spanish TTS Baseline Male Speech Database was created within the scope of the TC-STAR project funded by the European Government. The database complies with the common specifications created in the TC-STAR project. The annotation of the database includes manual orthographic transcriptions, the automatic segmentation into phonemes and automatic generation of pitch marks. A certain percentage (20%) of phonetic segments and pitch marks has been manually checked. A pronunciation lexicon in SAMPA with POS, lemma and phonetic transcription of all the words prompted and spoken is also provided.Speech samples are stored as sequences of 24-bit 96 kHz with the least significant byte first ("lohi" or Intel format) as (signed) integers. Each prompted utterance is stored in a separate file. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Antonio Bonafonte
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 401 0764
Fax: +34 93 401 6447
e-mail: antonio.bonafonte@upc.edu

2.2 . Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive.

2.3 . Copyright statement and information on IPR

The resource has copyright. The Copyright belongs to Universitat Politècnica de Catalunya. The resource is fee, license-based, for research and commercial purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The archive that will be uploaded on the MetaShare platform will contain several files containing documentation (copyright, readme, documents, …), speech files (one per utterance) and label files (each speech file has an accompanying label file).

3.2. Encoding

Documentation is encoded in plain text.

Speech samples are stored as sequences of 24-bit 96 kHz with the least significant byte first ("lohi" or Intel format) as (signed) integers. Each prompted utterance is stored in three separate file (one file for each channel).

Each speech file has three accompanying label files, identified by the file extesion:

- .esL: ascii file with time of the closure glottal instants (pitch marks). The file was derived automatically from the signal file using praat.
- .esP: ascii file with phonetic transcription and phonetic segmentation. The segmentation has been derived automatically by our in-house HMM-based segmentation tooklkit.
- .esS: SAM label file including the orthograhic information, the phonetic transcription and a rough prosodic labeling.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The corpus contains about 3,700 speech files/ 10 hours of recorded speech and needs about 32 GB for disk storage.

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database for training TTS systems in Spansih. The database is designed based on the TCSTAR specifications for baseline voices. (www.tcstar.org , deliverable document D6).

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

The corpus consist of several subcorpora. Each subcorpora is split in sentences or short paragraphs. Most of corpus was designed to provide high phonetic and prosodic coverage, including spoken style. These subcorpora are built from transcriptons of parliamentary speeches, news, novels. Furthermore some sentences are designed to increase the triphone coverage and to include special words in several domains.

4.3. Transcriptions, annotations:

All the speech data is labeled with orthographic, broad prosodic, phonetic transcription and pitch. The orthographic, prosodic and phonetic transcriptions were supervised manually. The pitch and segmentation labels were produced automatically and twenty per cent of the labels were manually supervised.
The orthographic annotations is a transliteration of what was actually read by the speaker, without ambiguity at the word level. The phonetic transcription annotate what the sepaker really said, including elision, reduction or assimilations. The prosodic annotation consist on minor breaks (intermediate intonational phrases), major breaks (intonational phrases), pitch accent (intonational prominence) normal and emphatic. The phonetic segmentation matches the phonetic transcription. The pitch labelling consist of one temporal mark for each pitch period.

4.4. Lexicon. Description of the lexicon (if applicable)

The lexicon file is an alphabetically ordered list of distinct lexical items (i.e. splitting tokens at every space, essentially words in our case) which occur in the corpus with the corresponding pronunciation information. Each distinct word has a separate entry. As the lexicon is derived from the corpus it uses the same alphabetic encoding for special and accented characters as used in the transcriptions. Each lexicon entry contains the possible phonetic transcription and POS. The lexicon follows the convention defined in the LC-STAR lexica.

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

The database was recorded by a male professional speaker, adult, from the central Castillian variant. Five speakers with the same profile were considered. The selection of this voice took into account different aspects: phonetics, pronunciation/articulatory, appropriateness of the voices for signal processing manipulation and preference tests.

4.6. Recording platform

The database was recorded in a isolated room at Universitat Politècnica de Catalunya. The studio includes two isolated rooms, one for the speaker and other for the operators. The speaker recorded simultaneously in three channels: membrane microphone, close-talk microphone and laringograph. The recording platform was developed at UPC. It consists of a computer program that allows to navigate through the selected prompts and controls the synchronous recordings (asio interface). The platform shows the recording levels, clipping, etc. The signals were recorded at 96kHz and with 24 bits per sample

TC-STAR Bilingual (Spanish English) Expressive speech

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

This database contains the recordings of two female and two male bilingual (Spanish/English) professionalspeakers recorded in a noise-reduced room simultaneously through a close talk microphone, a mid distance microphone and a laryngograph signal. It consists of the recordings and annotations of read text material of approximately 1 hours of speech per speaker and language. 220 English paragraphs were selected from the EPPS (European Parliamentary Plenary Sessions) to represent different situations or styles. The Spanish translation was also collected. The speakers read some excerpt of the parliamentary session and the read, in parliamentary style, the English text and the Spanish translation. The same speakers recorded the TC-STAR Bilingual Voice-Conversion speech.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Antonio Bonafonte
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 401 0764
Fax: +34 93 401 6447
e-mail: antonio.bonafonte@upc.edu

2.2 . Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive.

2.3 . Copyright statement and information on IPR

The resource has copyright. The Copyright belongs to Universitat Politècnica de Catalunya. The resource is fee, license-based, for research and for commercial purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The archive that will be uploaded on the MetaShare platform will contain several files containing documentation (copyright, readme, documents, …), speech files (one per utterance) and label files (each speech file has an accompanying label file).

3.2. Encoding

Documentation is MS-WORD formant and encoded in plain text.

Speech samples are stored as sequences of 24-bit 96 kHz with the least significant byte first ("lohi" or Intel format) as (signed) integers. Each prompted utterance is stored in three separate files (one file for each channel).

Each speech file has four accompanying label files, identified by the file extesion:

- .xxL: ascii file with time of the closure glottal instants (pitch marks).
- .xxP: text file with phonetic transcription and phonetic segmentation.
- .xxW: text file with word segmentation.
- .xxS: SAM label file including the orthograhic information, the phonetic transcription and a rough prosodic labeling.

where 'xx' is either 'en' (for English utterances) or 'es' (for Spanish ones).

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

For each speaker and language, the corpus contains about 221 speech files/ 1 hours of recorded speech and needs about 3 GB for disk storage. In total, the disk storage is 23.5GB.

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database designed for research on expressive speech in the domain of the European Parliament and to compare the prosody in different languages. Selecting bilingual speakers and reading the English and the Spanish utterantes one after the other tries to minimize external factors. The database is designed based on the TCSTAR specifications for expressive voices. (www.tcstar.org , deliverable document D6).

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

The corpus consist 220 paragraphs selected from the EPPS (European Parliament Plenary Sessions). The paragraphs were selected based on the speech delivered by the politicians, to cover different styles. The Spanish translation is also read (paralel corpus). Before reading the text, the speakers listen to the orginal recording of the Parliamentary. Then they are asked to read in the same style (not reading style). Furthermore, the speakers read each paragraph first in English, then in Spanish, so that the same style is set in both languages.

4.3. Transcriptions, annotations:

All the speech data is labeled with orthographic, broad prosodic, phonetic transcription and pitch. The orthographic, prosodic and phonetic transcriptions were supervised manually. The pitch and segmentation labels were produced automatically.

The orthographic annotations is a transliteration of what was actually read by the speaker, without ambiguity at the word level. The phonetic transcription annotate what the sepaker really said, including elision, reduction or assimilations. The prosodic annotation consist on minor breaks (intermediate intonational phrases), major breaks (intonational phrases), pitch accent (intonational prominence) normal and emphatic. The phonetic segmentation matches the phonetic transcription. The pitch labelling consist of one temporal mark for each pitch period.

4.4. Lexicon. Description of the lexicon (if applicable)

The lexicon file is an alphabetically ordered list of distinct lexical items (i.e. splitting tokens at every space, essentially words in our case) which occur in the corpus with the corresponding pronunciation information. Each distinct word has a separate entry. As the lexicon is derived from the corpus it uses the same alphabetic encoding for special and accented characters as used in the transcriptions. Each lexicon entry contains the possible phonetic transcription and POS. The lexicon follows the convention defined in the LC-STAR lexica.

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

The database was recorded by 2 male and 2 females bilingual professional speakers. The speakers were selected from more than 10 speakers taking into account that they are native both in UK English and Catillian Spanish, and also phonetics, pronunciation/articulatory and preference tests.

4.6. Recording platform

The database was recorded in a isolated room at Universitat Politècnica de Catalunya. The studio includes two isolated rooms, one for the speaker and other for the operators. The speaker recorded simultaneously in three channels: membrane microphone, close-talk microphone and laringograph. The recording platform was developed at UPC. It consists of a computer program that allows to navigate through the selected prompts and controls the synchronous recordings (asio interface). The platform shows the recording levels, clipping, etc. The signals were recorded at 96kHz and with 24 bits per sample

TC-STAR Bilingual (Spanish English) VC

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

This database contains the recordings of two female and two male bilingual (Spanish/English) professional speakers recorded in a noise-reduced room simultaneously through a close talk microphone, a mid distance microphone and a laryngograph signal. It consists of the recordings and annotations of read text material of approximately 75 minutes of speech per speaker and language. For each language and speaker, first 15 minutes of short sentences are read in mimicking style: the speakers listen to a sentence an repeat with same rhythm and intonation. Additionally, a corpus of 1 hour is read. The corpus was produced to support research on intra-lingual and cross-lingual voice conversion. The corpus can also be used to produce bilingual synthetic voices (TTS). The same speakers recorded the TC-STAR Bilingual Voice-Conversion speech.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Antonio Bonafonte
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 401 0764
Fax: +34 93 401 6447
e-mail: antonio.bonafonte@upc.edu

2.2 . Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive.

2.3 . Copyright statement and information on IPR

The resource has copyright. The Copyright belongs to Universitat Politècnica de Catalunya. The resource is fee, license-based, for research and for commercial purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The archive that will be uploaded on the MetaShare platform will contain several files containing documentation (copyright, readme, documents, …), speech files (one per utterance) and label files (each speech file has an accompanying label file).

3.2. Encoding

Documentation is MS-WORD formant and encoded in plain text.

Speech samples are stored as sequences of 24-bit 96 kHz with the least significant byte first ("lohi" or Intel format) as (signed) integers. Each prompted utterance is stored in three separate file (one file for each channel).

Each speech file has four accompanying label files, identified by the file extesion:

- .xxL: ascii file with time of the closure glottal instants (pitch marks).
- .xxP: text file with phonetic transcription and phonetic segmentation.
- .xxS: SAM label file including the orthograhic information, the phonetic transcription and a rough prosodic labeling.

where 'xx' is either 'en' (for English utterances) or 'es' (for Spanish ones).

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

For each speaker and language, the corpus contains about 200 speech files of short sentences in the mimick style (15 minutes) and 200 short paragrpahs in reading style (1 hour). The recorded speech needs about 4 GB for disk storage per speaker and language. In total, the disk storage is 32.5GB.

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database designed for research on intra-lingual voice conversion and cross-lingual voice conversion. The corpus '33' is read in mimick style: the speaker listen to a short sentence, then listen and repeat the same sentence and finally read the sentence which is recorded. In this way all the sentences are uttered by the different speaker with same intonation and rhythm. This is useful to investigate on voice conversion of the segmental aspects of the speech. The corpus '11' contains read speech. It can be used to investigate intralingual voice conversion considering also the prosody. As the speakers are bilingual and both, Spanish and English data is produce, this can be used for research cross-lingual voice conversion.
The database is designed based on the TCSTAR specifications for voice converssion (www.tcstar.org , deliverable document D6).

This resource can also be used to create synthetic voices, specially statistical based speech synthesis, either monolingual (Spanish and English) or bilingual voices.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

The corpus consist of:

- Subcorpus 11: approximately 214 short sentences in mimicking style, in Spanish and in English.

- Subcorpus 33: approximately 1 hour (200 Spanish paragraphs, 170 English paragraphs) selected from the EPPS (European Parliament Plenary Sessions).

4.3. Transcriptions, annotations:

All the speech data is labeled with orthographic, broad prosodic, phonetic transcription and pitch. The orthographic, prosodic and phonetic transcriptions were supervised manually. The pitch and segmentation labels were produced automatically and 5% of them were manually checked and corrected.
The orthographic annotations is a transliteration of what was actually read by the speaker, without ambiguity at the word level. The phonetic transcription annotate what the sepaker really said, including elision, reduction or assimilations. The prosodic annotation consist on minor breaks (intermediate intonational phrases), major breaks (intonational phrases), pitch accent (intonational prominence) normal and emphatic. The phonetic segmentation matches the phonetic transcription. The pitch labelling consist of one temporal mark for each pitch period.

4.4. Lexicon. Description of the lexicon (if applicable)

The lexicon file is an alphabetically ordered list of distinct lexical items (i.e. splitting tokens at every space, essentially words in our case) which occur in the corpus with the corresponding pronunciation information. Each distinct word has a separate entry. As the lexicon is derived from the corpus it uses the same alphabetic encoding for special and accented characters as used in the transcriptions. Each lexicon entry contains the possible phonetic transcription and POS. The lexicon follows the convention defined in the LC-STAR lexica.

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

The database was recorded by 2 male and 2 females bilingual professional speakers. The speakers were selected from more than 10 speakers taking into account that they are native both in UK English and Catillian Spanish, and also phonetics, pronunciation/articulatory and preference tests.

4.6. Recording platform

The database was recorded in a isolated room at Universitat Politècnica de Catalunya. The studio includes two isolated rooms, one for the speaker and other for the operators. The speaker recorded simultaneously in three channels: membrane microphone, close-talk microphone and laringograph. The recording platform was developed at UPC. It consists of a computer program that allows to navigate through the selected prompts and controls the synchronous recordings (asio interface). The platform shows the recording levels, clipping, etc. The signals were recorded at 96kHz and with 24 bits per sample

Exogenous resources

UVIGO LR Description

Galician SpeechDat FDB

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

The SpeechDat Galician Database for the Fixed Telephone Network FDB contains the recordings of 653 Galician speakers (217 males, 436 females) recorded over the Spanish fixed telephone network. The database is partitioned into 3 CD-ROMs, in ISO 9660 format. Speech samples are stored as sequences of 8-bit 8 kHz A-law, uncompressed. Each prompted utterance is stored in a separate file, and each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information. Each speaker uttered about 44 items. A pronunciation lexicon with the phonetic transcription in SAMPA is also included.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Carmen Garcia-Mateo
Address: E.E. Telecomunicacion, Campus Universitario, 36310 Vigo, Spain
Affiliation: Multimedia Technologies Group (GTM). Universidade de Vigo
Position: Professor
Telephone: +34 986 812133
Fax: +34 986 812116
e-mail: carmen.garcia@uvigo.es

2.2. Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive.

2.3. Copyright statement and information on IPR

The resource has copyright. The Copyright belongs to Universidade de Vigo and University of Santiago de Compostela. The resource is fee, license-based, for research and commercial purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The database includes documentation (copyright, readme, documents, …), speech files (one per utterance) and label files (each speech file has an accompanying label file) in a nested structured directory.

3.2. Encoding

Documentation is encoded in word and pdf text.

Speech files are stored as sequences of 8-bit 8 kHz A-law uncompressed speech sample. Each prompted utterance is stored within a separate file.

Each speech file has an accompanying ASCII SAM label file.
Label files contain information about the database, speech signal coding, speakers, labeling session, transcriptions and annotations.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The corpus contains about 38.333 speech files, which corresponds to about 1,2 GB.

4. CONTENT INFORMATION

4.1. Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database for training ASR systems in Galician. The database fulfills the SpeechDat (http://www.speechdat.org) specifications.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

Each speaker utters 44 items, 40 of these items are mandatory within SpeechDat:

- 3 application words
- 1 sequence of 10 isolated digits
- 4 connected digits (prompt sheet number -6 digits, telephone number –9/11 digits, credit card number –14/16 digits, PIN code –6 digits)
- 1 telephone number
- 1 personal identification number
- 3 dates (spontaneous date e.g. birthday, prompted date, relative and general date expression)
- 1 word spotting phrase using embedded application words
- 1 isolated digit

- 3 spelled words (1 surname, 1 directory assistance city name, 1 real/artificial name for coverage)
- 1 currency money amount
- 2 natural number
- 5 directory assistance names (1 spontaneous, e.g. own forename, 1 city of birth/growing up, 1 most frequent city out of a set of 500, 1 most frequent company/agency out of a set of 500, 1 "forename surname" out of a set of 150)
- 2 yes/no questions (1 predominantly "yes" question, 1 predominantly "no" question, including fuzzy questions)
- 10 phonetically rich sentences
- 2 time phrases (1 spontaneous time of day, 1 word style time phrase)
- 4 phonetically rich words.

Utterances are both, read and spontaneous.

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

The transcription included in this database is an orthographic, lexical transcription with a few details that represent audible acoustic events (speech and non speech) present in the corresponding waveform files. The extra marks contained in the transcription aid in interpreting the text form of the utterance. Transcriptions were made in two passes: one pass in which words are transcribed, and a second pass in which the additional details are added.

Transcriptions were checked periodically for quality control. Two persons transcribed samples of the same transcription task and results were compared.

4.4. Lexicon. Description of the lexicon (if applicable)

The lexicon file is an alphabetically ordered list of distinct lexical items (i.e. splitting tokens at every space, essentially words in our case) which occur in the corpus with the corresponding pronunciation information. Each distinct word has a separate entry. As the lexicon is derived from the corpus it uses the same alphabetic encoding for special and accented characters as used in the transcriptions.

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

The database contains recordings from 653 adult speakers.

All the speakers are Galician native. Speakers were selected from the three main Galician dialectal areas: western, central and eastern.

Age distribution: 12 speakers are under 16, 375 speakers are between 16 and 30 years old, 164 speakers are between 31 and 45, 88 speakers are between 46 and 60, 9 speakers are over 60, and 5 speakers are unknown.

4.6. Recording platform

4.6.1. Domain(s), environments,

Speakers were calling from the fixed telephone network.

4.6.2. Recording platform

The recording platform is based on a PC with an E-1 interface.

Transcrigal DB

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

The Transcrigal DB database contains 31 hours of recordings of broadcast news programs recorded from the Television de Galicia (TVG). TVG is the public Galician television. The speech recordings present variations of topic, speaker, acoustic channel, speaking mode, etc. The whole corpus has been segmented, labelled and transcribed manually using the tool developed by DGA (DélégationGénérale pour l'Armement, France) and LDC (Linguistic Data Consortium, USA), called "Transcriber", with conventions similar to those adopted by LDC for the DARPA HUB-4 corpora. Transcriptions include speaker turns, topics, channel information.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Carmen Garcia-Mateo
Address: E.E. Telecomunicacion, Campus Universitario, 36310 Vigo, Spain
Affiliation: Multimedia Technologies Group (GTM). Universidade de Vigo
Position: Professor
Telephone: +34 986 812133
Fax: +34 986 812116
e-mail: carmen.garcia@uvigo.es

2.2. Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive.

2.3. Copyright statement and information on IPR

The resource has copyright. The Copyright belongs to Universidade de Vigo, University of Santiago de Compostela and TVG. The resource is fee, license-based, for research purposes and fee license-based for commercial purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The archive that will be uploaded on the MetaShare platform will contain several files containing documentation (copyright, readme, documents,

…), 63 audio files in the WAV directory, each one having an accompanying transcription file in the TRS directory.
File names follow the following file name conventions:

E YYMMDD[T] . EXT

where:
E - 'M' for midday broadcast news or 'S' for evening broadcast news
YYMMDD - date code of recording: YY (two digits year), MM (month), DD (day)
EXT - file extenssion
[T] – This code is optional and indicates the "Topic" in the audio file. Its meaning is as follows: 'N' news, 'D' sports, and 'T' weather.

3.2. Encoding

Documentation is encoded in word/pdf/plain text.

The recorded audio files are stored in linear PCM (.wav) format, 16 kHz, 16 bit, single channel uncompressed audio samples. Files have an average duration of 1 hour.

Each speech file has an accompanying XML transcription file.
The XML transcription files contain information about the database, speech signal coding, speakers (public figures identified by names, others by their role within the broadcast, e.g. reporter, speaker), turns, segmentation, speaking style, channel, background sounds, acoustic events and literal transcriptions of the speech.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The corpus contains about 63 audio files /31 hours of recorded speech and needs about 3,9 GB for disk storage.

4. CONTENT INFORMATION

4.1. Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database for training ASR systems in Galician, but also contains minor proportions of Spanish.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

The database consists of 31 news programs (14 broadcast from the 2:30 pm news and 17 broadcasts from the 8:30 pm news), each one lasting approximately 1 hour. The recordings were carried out at six different months between October 2002 and March 2004. Each news program include interviews, reports from different recording environments,

segmented in speaker turns, phonetically rich, read and spontaneous speaking style.

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

The transcription included in this database is an orthographic, lexical transcription with a few details that represent audible acoustic events (speech and non speech) present in the corresponding waveform files. The extra marks contained in the transcription aid in interpreting the text form of the utterance. Transcriptions were made in three passes: one pass in which speaker segments and environmental conditions are added, a second pass adding acoustic events and their time stamps and, a third pass transcribing those segments not featuring music and speech overlap. Transcriptions follow the TRS format produced by the Transcriber transcribing tool.

4.4. Lexicon. Description of the lexicon (if applicable)

Not applicable.

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

The database recordings contain segments from 998 adult Galician speakers of unknown accent (489 male, 509 female), and 525 adult Spanish speakers (420 male, 105 female). Speakers are unbalanced in gender favouring male speakers in total duration.

4.6. Recording platform

4.6.1. Domain(s), environments,

Four recording environments are present:

Studio: segments originate from speakers located in the studio.

Telephone: segments originate from speakers over a telephone.

Outside: segments originate from speakers outside of buildings, e.g. on streets, public space

None: segments that have non of the above classification.

4.6.2. Recording platform

The recordings were made through aHauppauge WinTV PCI card connected to the TV antennaand the AVI_IO software. All files were recorded in linear PCM (.wav) format, 16 kHz, 16 bit, single channel.

DOGalicia (Parallel Galician-Spanish Corpus)

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

DOGalicia is a parallel Galician-Spanish corpus designed for statistical translation purposes. This corpus was obtained from some sections of the "Diario Oficial de Galicia (DOG)" (Official Regional Gazette of Galicia), from 1996 to 2010, and contains more than 42 million words for each language. The data are aligned at sentence level and stored in text files, in a one sentence per line basis. Data are provided in UTF-8 plain text.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Eduardo Rodriguez-Banga
Address: E.E. Telecomunicacion, Campus Universitario, 36310 Vigo, Spain
Affiliation: Multimedia Technologies Group (GTM). Universidade de Vigo
Position: Associate Professor
Telephone: +34 986 812676
Fax: +34 93 986 812116
e-mail: erbanga@gts.uvigo.es

2.2. Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive.

2.3. Copyright statement and information on IPR

This resource is free for academic or research purposes, giving due credit to the "Diario Oficial de Galicia (DOG)" and to the "Universidade de Vigo".

3. TECHNICAL INFORMATION

3.1. Directories and files

The corpus that will be uploaded on the MetaShare platform will contain several files under a common directory.

3.2. Encoding

Documentation is encoded in word, pdf or plain text.

The corpus files are encoded as UTF-8 plain text.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The corpus consists of two files for each language corresponding to the time periods of 1996-2003 and 2004-2010. The file names are as follows, where es-gl denotes the language pair and the final extension (es or gl) refers to the language of the text:

DOG1996-2003.es-gl.es size: 140 MB
DOG1996-2003.es-gl.gl size: 136 MB
DOG2004-2010.es-gl.es size: 152 MB
DOG2004-2010.es-gl.gl size: 145 MB

4. CONTENT INFORMATION

4.1. Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a parallel Galician-Spanish corpus designed for statistical translation purposes

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

This corpus consists of two files for each language corresponding to the time periods of 1996-2003 and 2004-2010.

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

The sentences of the two languages were automatically aligned based on the structure of the source documents. Due to the similarity between Galician and Spanish languages, a Levenshtein distance measure was used to check the proper alignment of the sentences. Therefore, alignments errors are improbable.

4.4. Lexicon. Description of the lexicon (if applicable)

Not applicable.

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

Not applicable.

4.6. Recording platform

Not applicable.

GCG Corpus

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

The GCG ("Grupo Correo Galego") corpus is a morphosyntactically annotated corpus for the Galician language.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Eduardo Rodriguez-Banga
Address: E.E. Telecomunicacion, Campus Universitario, 36310 Vigo, Spain
Affiliation: Multimedia Technologies Group (GTM). Universidade de Vigo
Position: Associate Professor
Telephone: +34 986 812676
Fax: +34 93 986 812116
e-mail: erbanga@gts.uvigo.es

2.2. Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive.

2.3. Copyright statement and information on IPR

Copyright belongs to "Grupo Correo Galego", "Universidade de Vigo" and "Centro Ramón Piñeiro para a investigación en humanidades". This resource is free for academic and research purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The annotated corpus consists of a single file distributed in csv and xlsx formats. The original text file is also distributed together with some additional documentation. All the files are under a common directory.

3.2. Encoding

Text in the csv and text files is encoded as ISO-8859. Different fields in the csv file are delimited by tabs.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The size of any file is very small (< 40KB)

4. CONTENT INFORMATION

4.1. Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a Galician morphosyntactic corpus, semi-automatically annotated and manually reviewed,

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

This corpus contains almost 400,000 words, their corresponding part-of-speech (POS) tags and some additional linguistic information.

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

POS tags were manually reviewed after automatic tagging. Therefore these tags are quite reliable. Some other linguistic information, such as lemma and detailed verbal analysis, was generated automatically.

4.4. Lexicon. Description of the lexicon (if applicable)

Not applicable.

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

Not applicable.

4.6. Recording platform

4.6.1. Domain(s), environments,

Not applicable.

4.6.2. Recording platform
Not applicable.

Cotovia Text-to-Speech System for Galician and Spanish

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

Cotovía is a unit-selection text-to-speech system for Galician and Spanish. Cotovia has been developed by the University de Vigo and the center 'Ramón Piñeiro' for Research in Humanities, both in Galicia, Spain. Its development has involved a research group of linguists and engineers.

Cotovía has been developed as a research project, therefore most of the work has been focused on the most interesting aspects from a scientific point of view. Although the performance of the whole TTS system is quite good, there are some parts that could be clearly improved.

Cotovía can be also used for other related purposes such as automatic phonetic transcription or POS tagging.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Eduardo Rodriguez-Banga
Address: E.E. Telecomunicacion, Campus Universitario, 36310 Vigo, Spain
Affiliation: Multimedia Technologies Group (GTM). Universidade de Vigo
Position: Associate Professor
Telephone: +34 986 812676
Fax: +34 93 986 812116
e-mail: erbanga@gts.uvigo.es

2.2. Delivery medium (if relevant; description of the content of each piece of medium)

Cotovia is available at http://sourceforge.net/projects/cotovia/ .

2.3. Copyright statement and information on IPR

The resource has copyright. The Copyright belongs to the 'Universidade de Vigo' and the 'Centro Ramón Piñeiro'. Cotovía C++ code is distributed under the GPL3.0+ license, while each of the avaliable speaker voices has its own license. The voices available at SourceForge are free for commercial and non-commercial uses. Some other voices, free for non-commercial uses, may be available in the next future through external links.

3. TECHNICAL INFORMATION

3.1. Directories and files

Cotovia files are available at http://sourceforge.net/projects/cotovia/ at the Files section or at the Git repository.

3.2. Encoding

Cotovia code is written in C++.
Documentation files are in different formants (pdf, raw text, …).

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The disk space needs of Cotovia mainly depend on the voices installed (ranging from about 180 MB to 2.1 GB).

4. CONTENT INFORMATION

4.1. Type of the resource (language, ASR, BN, dialogues, TTS….)

Cotovia is a TTS system for Galician and Spanish.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

At this time the original recordings of the speakers are not available in an organized manner at SourceForge.

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

At this time the transcriptions and annotations of the original recordings of the speakers are not available in an organized manner at SourceForge.

4.4. Lexicon. Description of the lexicon (if applicable)

Not applicable.

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

At this time two speakers (one male and one female) are available at SourceForge.

4.6. Recording platform

4.6.1. Domain(s), environments,

Read corpus.

4.6.2. Recording platform

Recording studio.

The Basque University LR Description

Ahosyn male EU: Large Speech Database for Synthesis in Basque

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

3798 phonetically balanced sentences in Basque recorded by a male voice talent (KJ) in neutral style. It was registered at 48kHz, 16bits, semi-professional room, 1 microphones (Diaphragm Neumann TLM103) and laryngograph included.
The voice talent KJ is the same as in 'Ahosyn male ES'. The read sentences are the same as in 'Ahosyn female EU'.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Inma Hernaez
Address: ETSI Alda. Urkijo s/n 48013 Bilbao SPAIN
Affiliation: Aholab Signal Processing Laboratory. University of the Basque Country (UPV/EHU)
Position: Professor
Telephone: +34 94 601 3969
Fax: +34 94 601 4259
e-mail: inma.hernaez@ehu.es

2.2 . Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be downloadable as an archive through the metashare platform.

2.3 . Copyright statement and information on IPR

The resource has copyright. The Copyright belongs to University of the Basque Country . The resource is fee, license-based, for research and commercial purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The database includes documentation (readme file), speech and glottal speech files (one file per utterance) and label files (each speech file has 3 accompanying label files) each in different directories according to the following structure:
wav: Recorded speech, wav format, 48Khz, mono, 16 bits
egg: Recorded laryngograph signals , wav format, 48Khz, mono, 16 bits
txt: Text files
mrk: Phone labels with the following format:
Beginning(ms)/Lenght(ms) : Phone(sampa)
(Basque sampa is available at http://aholab.ehu.es/sampa_basque.htm)

Ling: Linguistic information.
Phonemes separated by white space
Stress marked with „'" symbol before de stressed vowel
Syllables separated by symbol „-„.
Words separated by symbol „|" .

3.2. Encoding

Documentation is encoded in a text file.
Speech and glottal speech files (extension .wav) are stored as wav files, 48kHz sampling frequency, mono, 16 bits.
Annotation files are encoded in raw text files.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The corpus contains a set of 3798 speech utterance files, about 6 hours of speech. The whole database occupies 4,6GB (2,3GB of speech files).

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database for developing general domain TTS systems for Basque. The written texts are in Standard Basque. The pronunciation corresponds to Basque spoken in the Southern region of the Basque speaking areas (the Spanish regions).

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

The database contains the recording of 3798 read sentences in standard Basque (Batua) . The sentences were selected from a big amount of texts collected from several domains (News, Literature, Arts, Science and others). After transcription of the texts, the selection of the final set of sentences was performed considering phonetic balance and restricting the final length of the sentence.

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

The speech files were automatically segmented. Phonetic transcription and linguistic annotation was automatically done. No manual revision was done.

4.4. Lexicon. Description of the lexicon (if applicable)

The database doesn't include a lexicon.

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

The speaker is a native bilingual Basque/Spanish speaker. He is a professional radio broadcaster.

4.6. Recording platform

4.6.1.Domain(s), environments,

Recordings were performed in an acoustically conditioned semi-professional room at the dependencies of Aholab.

4.6.2 Recording platform

- Microphones: Neumann TLM103 (diaphragm), Shure Beta54 (close-talk)headphone Shure SM 10 (not provided)
- Audio interface: RME Fireface 400
- Laryngograph PCLX (LTD)
- Software NannyRecord (UPC) , Fireface Mixer

All signal files have some ms of silence (environment sound) at their beginning and end. Endpoints were manually supervised at the recording site.

Ahosyn female EU: Large Speech Database for Synthesis in Basque

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

3798 phonetically balanced sentences in Basque, recorded by a male voice talent (AG) in neutral style. It was registered at 48kHz, 16bits, semi-professional room, 1 microphones (Diaphragm Neumann TLM103) and laryngograph included.
The voice talent AG is the same as in 'Ahosyn female ES'. The read sentences are the same as in 'Ahosyn male EU'.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Inma Hernaez
Address: ETSI Alda. Urkijo s/n 48013 Bilbao SPAIN
Affiliation: Aholab Signal Processing Laboratory. University of the Basque Country (UPV/EHU)
Position: Professor
Telephone: +34 94 601 3969
Fax: +34 94 601 4259
e-mail: inma.hernaez@ehu.es

2.2 . Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be downloadable as an archive through the metashare platform.

2.3 . Copyright statement and information on IPR

The resource has copyright. The Copyright belongs to University of the Basque Country . The resource is fee, license-based, for research purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The database includes documentation (readme file), speech and glottal speech files (one file per utterance) and label files (each speech file has 3 accompanying label files) each in different directories according to the following structure:
wav: Recorded speech, wav format, 48Khz, mono, 16 bits
egg: Recorded laryngograph signals , wav format, 48Khz, mono, 16 bits
txt: Text files
mrk: Phone labels with the following format:
Beginning(ms)/Lenght(ms) : Phone(sampa)
(Basque sampa is available at http://aholab.ehu.es/sampa_basque.htm)

Ling: Linguistic information.
Phonemes separated by white space
Stress marked with „'" symbol before de stressed vowel

Syllables separated by symbol „-„.
Words separated by symbol „|" .

3.2. Encoding

Documentation is encoded in a text file.
Speech and glottal speech files (extension .wav) are stored as wav files, 48kHz sampling frequency, mono, 16 bits.
Annotation files are encoded in raw text files.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The corpus contains a set of 3798 speech utterance files, about 6 hours of speech. The whole database occupies 4,1GB (2GB of speech files).

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database for developing general domain TTS systems for Basque. The written texts are in Standard Basque. The pronunciation corresponds to Basque spoken in the Southern region of the Basque speaking areas (the Spanish regions).

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

The database contains the recording of 3798 read sentences in standard Basque (Batua) . The sentences were selected from a big amount of texts collected from several domains (News, Literature, Arts, Science and others). After transcription of the texts, the selection of the final set of sentences was performed considering phonetic balance and restricting the final length of the sentence.

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

The speech files were automatically segmented. Phonetic transcription and linguistic annotation is fully automatic. No manual revision has been done.

4.4. Lexicon. Description of the lexicon (if applicable)

The database doesn't include a lexicon.

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

The speaker is a native bilingual Basque/Spanish speaker. She is a professional radio broadcaster.

4.6. Recording platform

4.6.1.Domain(s), environments,

Recordings were performed in an acoustically conditioned semi-professional room at the dependencies of Aholab.

4.6.2 Recording platform

- Microphones: Neumann TLM103 (diaphragm), Shure Beta54 (close-talk)headphone Shure SM 10 (not provided)
- Audio interface: RME Fireface 400
- Laryngograph PCLX (LTD)
- Software NannyRecord (UPC) , Fireface Mixer

All signal files have some ms of silence (environment sound) at their beginning and end. Endpoints were manually supervised at the recording site.

Ahosyn male ES: Large Speech Database for Synthesis in Spanish

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

3330 phonetically balanced sentences in Spanish, recorded by a female voice talent (KJ) in neutral style. It was registered at 48kHz, 16bits, semi-professional room, 1 microphones (Diaphragm Neumann TLM103) and laryngograph included.
The voice talent KJ is the same as in 'Ahosyn male EU'. The first 3329 read sentences are the same as in 'Ahosyn female ES'.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Inma Hernaez
Address: ETSI Alda. Urkijo s/n 48013 Bilbao SPAIN
Affiliation: Aholab Signal Processing Laboratory. University of the Basque Country (UPV/EHU)
Position: Professor
Telephone: +34 94 601 3969
Fax: +34 94 601 4259
e-mail: inma.hernaez@ehu.es

2.2 . Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be downloadable as an archive through the metashare platform.

2.3 . Copyright statement and information on IPR

The resource has copyright. The Copyright belongs to University of the Basque Country . The resource is fee, license-based, for research and commercial purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The database includes documentation (readme file), speech and glottal speech files (one file per utterance) and label files (each speech file has 3 accompanying label files) each in different directories according to the following structure:
wav: Recorded speech, wav format, 48Khz, mono, 16 bits
egg: Recorded laryngograph signals , wav format, 48Khz, mono, 16 bits
txt: Text files
mrk: Phone labels with the following format:
Beginning(ms)/Lenght(ms) : Phone(sampa)

Ling: Linguistic information.
Phonemes separated by white space

Stress marked with „'" symbol before de stressed vowel
Syllables separated by symbol „-„.
Words separated by symbol „|" .

3.2. Encoding

Documentation is encoded in a text file.
Speech and glottal speech files (extension .wav) are stored as wav files, 48kHz sampling frequency, mono, 16 bits.
Annotation files are encoded in raw text files.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The corpus contains a set of 3330 speech utterance files, about 6 hours of speech. The whole database occupies 4,5GB (2,2GB of speech files).

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database for developing general domain TTS systems for Spanish. The pronunciation corresponds to Castilian Spanish.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

The database contains the recording of 3330 read sentences in Spanish. The sentences were selected from a big amount of texts collected from several domains (News, Literature, Arts, Science and others). After transcription of the texts, the selection of the final set of sentences was performed considering phonetic balance and restricting the final length of the sentence.

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

The speech files were automatically segmented. Phonetic transcription and linguistic annotation was automatically done. No manual revision was done.

4.4. Lexicon. Description of the lexicon (if applicable)

The database doesn't include a lexicon.

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

The speaker is a native bilingual Basque/Spanish speaker. She is a professional radio broadcaster.

4.6. Recording platform

4.6.1.Domain(s), environments,

Recordings were performed in an acoustically conditioned semi-professional room at the dependencies of Aholab.

4.6.2 Recording platform

- Microphones: Neumann TLM103 (diaphragm), Shure Beta54 (close-talk)headphone Shure SM 10 (not provided)
- Audio interface: RME Fireface 400
- Laryngograph PCLX (LTD)
- Software NannyRecord (UPC) , Fireface Mixer

All signal files have some ms of silence (environment sound) at their beginning and end. Endpoints were manually supervised at the recording site.

Ahosyn female ES: Large Speech Database for Synthesis in Spanish

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

3329 phonetically balanced sentences in Spanish, recorded by a female voice talent (AG) in neutral style. It was registered at 48kHz, 16bits, semi-professional room, 1 microphones (Diaphragm Neumann TLM103) and laryngograph included.
The voice talent AG is the same as in 'Ahosyn female EU'. The read sentences are the same as the first 3329 sentences in 'Ahosyn male ES'.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Inma Hernaez
Address: ETSI Alda. Urkijo s/n 48013 Bilbao SPAIN
Affiliation: Aholab Signal Processing Laboratory. University of the Basque Country (UPV/EHU)
Position: Professor
Telephone: +34 94 601 3969
Fax: +34 94 601 4259
e-mail: inma.hernaez@ehu.es

2.2 . Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be downloadable as an archive through the metashare platform.

2.3 . Copyright statement and information on IPR

The resource has copyright. The Copyright belongs to University of the Basque Country . The resource is fee, license-based, for research purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The database includes documentation (readme file), speech and glottal speech files (one file per utterance) and label files (each speech file has 3 accompanying label files) each in different directories according to the following structure:
wav: Recorded speech, wav format, 48Khz, mono, 16 bits
egg: Recorded laryngograph signals , wav format, 48Khz, mono, 16 bits
txt: Text files
mrk: Phone labels with the following format:
Beginning(ms)/Lenght(ms) : Phone(sampa)

Ling: Linguistic information.
Phonemes separated by white space
Stress marked with „'" symbol before de stressed vowel

Syllables separated by symbol „-„.
Words separated by symbol „|" .

3.2. Encoding

Documentation is encoded in a text file.
Speech and glottal speech files (extension .wav) are stored as wav files, 48kHz sampling frequency, mono, 16 bits.
Annotation files are encoded in raw text files.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The corpus contains a set of 3329 speech utterance files, about 6 hours of speech. The whole database occupies 4,1GB (2,1GB of speech files).

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database for developing general domain TTS systems for Spanish. The pronunciation corresponds to Castilian Spanish.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

The database contains the recording of 3329 read sentences in Spanish. The sentences were selected from a big amount of texts collected from several domains (News, Literature, Arts, Science and others). After transcription of the texts, the selection of the final set of sentences was performed considering phonetic balance and restricting the final length of the sentence.

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

The speech files were automatically segmented. Phonetic transcription and linguistic annotation was automatically done. No manual revision was done.

4.4. Lexicon. Description of the lexicon (if applicable)

The database doesn't include a lexicon.

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

The speaker is a native bilingual Basque/Spanish speaker. She is a professional radio broadcaster.

4.6. Recording platform

4.6.1.Domain(s), environments,

Recordings were performed in an acoustically conditioned semi-professional room at the dependencies of Aholab.

4.6.2 Recording platform

- Microphones: Neumann TLM103 (diaphragm), Shure Beta54 (close-talk)headphone Shure SM 10 (not provided)
- Audio interface: RME Fireface 400
- Laryngograph PCLX (LTD)
- Software NannyRecord (UPC) , Fireface Mixer

All signal files have some ms of silence (environment sound) at their beginning and end. Endpoints were manually supervised at the recording site.

Ahoemo1: Emotional speech and video database in Standard Basque

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

This audio and video resource contains 170 items (sentences and words) repeated for the 6 MPEG emotions and 95 items for neutral style, for a female voice in Basque. The audio was registered at 32kHz, 16bits, professional studio, 1 microphone and laryngograph included. Two cameras were used for video recordings (one frontal, one lateral). Face markers were used.

The voice has been segmented at phone level and manually revised for the neutral style. The video is unprocessed.

It is available for research use with a fee.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person:

Name: Inma Hernaez
Address: ETSI Alda. Urkijo s/n 48013 Bilbao SPAIN
Affiliation: Aholab Signal Processing Laboratory. University of the Basque Country (UPV/EHU)
Position: Professor
Telephone: +34 94 601 3969
Fax: +34 94 601 4259
e-mail: inma.hernaez@ehu.es

2.2. Delivery medium (if relevant; description of the content of each piece of medium)

The audio files will be downloadable as an archive through the metashare platform. The video files will be provided in 2 DVDs.

2.3. Copyright statement and information on IPR

The resource belongs to UPV/EHU. The resource is fee, license-based, for research purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The audio part of the database includes documentation (readme file), speech and glottal speech files (one file per utterance) and label files (each speech file has 3 accompanying label files) each in different directories according to the following structure:
wav: Recorded speech, wav format, 32Khz, mono, 16 bits.
Egg: recorded laringograph glotal signal
txt: Text files

mrk: Phone labels with the following format:
Beginning(ms)/Lenght(ms) : Phone(sampa)
Basque sampa: http://aholab.ehu.es/sampa_basque.htm
ling: Linguistic information.
Phonemes separated by white space
Stress marked with „'" symbol before de stressed vowel
Syllables separated by symbol „-„.
Words separated by symbol „|" .

Each directory contains the data files organized by emotions:
wav|egg|txt|mrk|ling/anger|disgust|fear|happiness|neutral|sadness|surprise

Each emotion directory except neutral contains the data files separeted according to the content (see section 4.2):

wav|egg|txt|mrk|ling/anger|disgust|fear|happiness|sadness|surprise/common
|dependent

The neutral directory contains only one folder ‚common':
wav|egg|txt|mrk|ling/neutral/common

3.2. Encoding

Documentation is encoded in a text file.
Speech and glottal speech files (extension .wav) are stored as wav files, 32kHz sampling frequency, mono, 16 bits.
Annotation files are encoded in raw text files.
Video is stored in DVD. It has not been processed.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The recorded database has 1 hour and 25 minutes length. 50 minutes come from the common texts, 35 minutes from the texts semantically related with emotion. The audio part occupies 590MB. The video files use 2 DVD.

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

The intended purpose of the database was the study of emotions in speech for the development of TTS systems in Basque, as well as facial emotions for avatar development.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

The selected texts for the database into two different groups.

– One group consists of emotion independent texts, which are common for all emotions, as well as for the neutral style. This common group of texts was phonetically balanced in order to achieve a phoneme distribution similar to the one that occurs in natural oral language. These texts have neutral semantic content and are called common texts.

– The other group includes texts semantically related to each emotion, and therefore, this group is different for each of the emotions considered in the database. Neutral style was not considered in this part of the corpus. These texts are called specific texts.

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

The following annotation is provided:

txt: Text files
mrk: Phone labels with the following format:
Beginning(ms)/Lenght(ms) : Phone(sampa)
Basque sampa: http://aholab.ehu.es/sampa_basque.htm
ling: Linguistic information.
Phonemes separated by white space
Stress marked with „'" symbol before de stressed vowel
Syllables separated by symbol „-„.
Words separated by symbol „|" .

Segmentation and linguistic annotation is fully automatic.

4.4. Lexicon. Description of the lexicon (if applicable)

Not applicable

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

The speaker is a native bilingual Basque/Spanish speaker. She is a professional dubbing actress.

4.6. Recording platform

4.6.1.Domain(s), environments,

Recordings were performed in professional studio.

4.6.2. Recording platform

Ahoemo2: Emotional speech database in Standard Basque

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

702 phonetically balanced sentences in Standard Basque repeated for the 6 MPEG emotions and neutral style, for one male and one female voices. Recorded laryngograph signals and automatic segmentation are provided. Part of the emotions have been manually checked for the female voice.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Inma Hernaez
Address: ETSI Alda. Urkijo s/n 48013 Bilbao SPAIN
Affiliation: Aholab Signal Processing Laboratory. University of the Basque Country (UPV/EHU)
Position: Professor
Telephone: +34 94 601 3969
Fax: +34 94 601 4259
e-mail: inma.hernaez@ehu.es

2.2 . Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be downloadable as an archive through the metashare platform.

2.3 . Copyright statement and information on IPR

The resource has copyright. The Copyright belongs to University of the Basque Country . The resource is fee, license-based, for research purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The database includes documentation (readme file), speech and glottal speech files (one file per utterance) and label files (each speech file has 3 accompanying label files) each in different directories according to the following structure:
/male: data for male voice
/male/wav: Recorded speech, wav format, 48Khz, mono, 16 bits.
/male/egg: Recorded laryngograph signals , wav format, 48Khz, mono, 16 bits
/male/txt: Text files
/male/mrk: Phone labels with the following format:
Beginning(ms)/Lenght(ms) : Phone(sampa)
Basque sampa: http://aholab.ehu.es/sampa_basque.htm
/male/ling: Linguistic information.
Phonemes separated by white space
Stress marked with „'" symbol before de stressed vowel
Syllables separated by symbol „-„.

Words separated by symbol „|” .

/female: data for female voice
/female: data for male voice
/female/wav: Recorded speech, wav format, 48Khz, mono, 16 bits
/female/egg: Recorded laryngograph signals , wav format, 48Khz, mono, 16 bits
/female/txt: Text files
/female/mrk: Phone labels with the following format:
Beginning(ms)/Lenght(ms) : Phone(sampa)
Basque sampa: http://aholab.ehu.es/sampa_basque.htm
/female/ling: Linguistic information.
Phonemes separated by white space
Stress marked with „'” symbol before de stressed vowel
Syllables separated by symbol „-„.
Words separated by symbol „|” .
The files corresponding to each emotion are organized in different directories (anger, disgust, fear, happiness, neutral, sadness, surprise) in every category (wav, egg, txt, mrk, ling).

3.2. Encoding

Documentation is encoded in a text file.
Speech and glottal speech files (extension .wav) are stored as wav files, 48kHz sampling frequency, mono, 16 bits.
Annotation files are encoded in raw text files.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The recorded database consists on approximately 1.5 hours per emotion which makes up 10.5 hours of recordings per speaker, more than 20 hours in total. The corpus contains a set of 702 speech utterance files. The whole database occupies 12GB (5,7GB of speech files).

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database developed to model emotional intonation for TTS in Basque. The pronunciation and intonation corresponds to Standard Basque of the southern region of the Basque speaking areas.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

The database contains the recording of 702 read sentences in Basque. The sentences were selected from a big amount of texts collected from several domains (News, Literature, Arts, Science and others). After transcription of the texts, the selection of the final set of sentences was

performed considering phonetic balance and restricting the final length of the sentence.

The neutral texts were read with acted emotional (anger, disgust, fear, happiness, sadness, surprise) and neutral styles .

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

The speech files were automatically segmented. Phonetic transcription and linguistic annotation was automatically done.
For the female voice, a manual revision of neutral, hpappiness and sadness styles was done. For the male voice no manual revision was done.

4.4. Lexicon. Description of the lexicon (if applicable)

The database doesn't include a lexicon.

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

The speakers are a native bilingual Basque/Spanish speaker. They are both professional radio broadcasters.

4.6. Recording platform

4.6.1.Domain(s), environments,

Recordings were performed in an acoustically conditioned semi-professional room at the dependencies of Aholab.

4.6.2 Recording platform

- Microphones: BeyerDynamic MC740 (Membrane) Emkay VR-3576 (Close-talk) (not provided)
- Mixing desk: Soundcraft Spirit F1
- Laryngograph PCLX (Laryngograph LTD)
- Audio Card: VXPOcket 440 (Digigram)
- Software NannyRecord (UPC) , Digitram Wave Mixer

All signal files have some ms of silence (environment sound) at their beginning and end. Endpoints were manually supervised at the recording site.

Ahoemo3: Emotional speech database in Standard Basque

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

500 phonetically balanced sentences in Standard Basque repeated for 3 emotions (anger, hapiness and sadness) and neutral style, for one male and one female voices. Recorded laryngograph signals and automatic segmentation are provided.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Inma Hernaez
Address: ETSI Alda. Urkijo s/n 48013 Bilbao SPAIN
Affiliation: Aholab Signal Processing Laboratory. University of the Basque Country (UPV/EHU)
Position: Professor
Telephone: +34 94 601 3969
Fax: +34 94 601 4259
e-mail: inma.hernaez@ehu.es

2.2 . Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be downloadable as an archive through the metashare platform.

2.3 . Copyright statement and information on IPR

The resource has copyright. The Copyright belongs to University of the Basque Country . The resource is fee, license-based, for research purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The database includes documentation (readme file), speech and glottal speech files (one file per utterance) and label files (each speech file has 3 accompanying label files) each in different directories according to the following structure:
/JI: data for male voice
/JI/wav: Recorded speech, wav format, 48Khz, mono, 16 bits.
/JI/egg: Recorded laryngograph signals , wav format, 48Khz, mono, 16 bits
/JI/txt: Text files
/JI/mrk: Phone labels with the following format:
Beginning(ms)/Lenght(ms) : Phone(sampa)
Basque sampa: http://aholab.ehu.es/sampa_basque.htm
/JI/ling: Linguistic information.
Phonemes separated by white space
Stress marked with "'" symbol before de stressed vowel

Syllables separated by symbol ”- ”.
Words separated by symbol ”|” .

/KC: data for female voice
/KC/wav: Recorded speech, wav format, 48Khz, mono, 16 bits
/KC/egg: Recorded laryngograph signals , wav format, 48Khz, mono, 16 bits
/KC/txt: Text files
/KC/mrk: Phone labels with the following format:
Beginning(ms)/Lenght(ms) : Phone(sampa)
Basque sampa: http://aholab.ehu.es/sampa_basque.htm
/KC/ling: Linguistic information.
Phonemes separated by white space
Stress marked with ”'” symbol before de stressed vowel
Syllables separated by symbol ”- ”.
Words separated by symbol ”|” .
The files corresponding to each emotion are organized in different directories (anger, happiness, neutral, sadness) in every category (wav, egg, txt, mrk, ling).

3.2. Encoding

Documentation is encoded in a text file.
Speech and glottal speech files (extension .wav) are stored as wav files, 48kHz sampling frequency, mono, 16 bits.
Annotation files are encoded in raw text files.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The recorded database consists on approximately 50 minutes per emotion which makes up 3.5 hours of recordings per speaker, about 7 hours in total. The corpus contains a set of 500 speech utterance files. The whole database occupies 4.9GB (2,5GB of speech files).

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database for developing emotional TTS systems for Basque. The pronunciation and intonation corresponds to Standard Basque of the southern region of the Basque speaking areas.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

The database contains the recording of 500 read sentences in Standard Basque. The sentences were selected from a big amount of texts collected from several domains (News, Literature, Arts, Science and others). After transcription of the texts, the selection of the final set of sentences was

performed considering phonetic balance and restricting the final length of the sentence.

The same texts with neutral content were read with acted emotions (anger, happiness and sadness) and in neutral style.

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

The speech files were automatically segmented. Phonetic transcription and linguistic annotation was automatically done. No manual revision was done.

4.4. Lexicon. Description of the lexicon (if applicable)

The database doesn't include a lexicon.

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

The speakers are native bilingual Basque/Spanish speakers. They are both professional radio broadcasters.

4.6. Recording platform

4.6.1. Domain(s), environments,

Recordings were performed in an acoustically conditioned semi-professional room at the dependencies of Aholab.

4.6.2 Recording platform

- Microphones: Neumann TLM103 (diaphragm), Shure Beta54 (close-talk)
- Audio interface: RME Fireface 400
- Laryngograph PCLX (LTD)
- Software NannyRecord (UPC) , Fireface Mixer

All signal files have some ms of silence (environment sound) at their beginning and end. Endpoints were manually supervised at the recording site.

MDB602EU: Basque SpeechDat like (Mobile Network)

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

The SpeechDat Basque MDB database contains the recordings of 600 Basque speakers (287 males, 315 females) collected over the mobile telephone network. The database is partitioned into 4 CD-ROMs, in ISO 9660 format. Speech samples are stored as sequences of 8-bit 8 kHz A-law, uncompressed. Each prompted utterance is stored in a separate file, and each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information. Each speaker uttered 54 items. A pronunciation lexicon with the phonetic transcription in SAMPA is also included.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Inma Hernaez
Address: ETSI Alda. Urkijo s/n 48013 Bilbao SPAIN
Affiliation: Aholab Signal Processing Laboratory. University of the Basque Country (UPV/EHU)
Position: Professor
Telephone: +34 94 601 3969
Fax: +34 94 601 4259
e-mail: inma.hernaez@ehu.es

2.2 . Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be provided as 4 CDROM. 3 CDs contain each data from 200 speakers, 1 CD contains data from 2 speakers.

2.3 . Copyright statement and information on IPR

The resource has copyright. The Copyright belongs to the Basque Government. The resource can be available after negotiation with the Basque Government.

3. TECHNICAL INFORMATION

3.1. Directories and files

The database includes documentation (copyright, readme, documents, …), speech files (one per utterance) and label files (each speech file has an accompanying label file) in a nested structured directory.

3.2. Encoding

Documentation is encoded in raw text, word and ps text.

Speech files are stored as sequences of 8-bit 8 kHz A-law uncompressed speech sample. Each prompted utterance is stored within a separate file.

Each speech file has an accompanying ASCII SAM label file.
Label files contain information about the database, speech signal coding, speakers, labeling session, transcriptions and annotations.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The database occupies 2Gb. There are 32508 speech and labels files.

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database for training ASR systems in Basque. The database fulfills the SpeechDat (www.speechDat.org) specifications.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

Each speaker utters 54 items:

- 3 Isolated digit items:
  - 2 single isolated digit
  - 1 sequence of 10 isolated digits in one utterance (for SDB)
- 4 Digit/number strings:
  - 1 prompt sheet number (5+ digits, including any check digit)
  - 1 telephone number (9 digits)
  - 1 credit card number (14-16 digits, including any check digit) (corresponding to a set of 150 SDB numbers)
  - 1 6-digit PIN code (corresponding to a set of 150 SDB codes)
- 1 Natural number
- 1 Money amounts:
  - 1 currency amount, mixed size and units
- 3 Yes/no questions:
  - 1 predominantly yes including 'fuzzy' yes/no (spontaneous)
  - 1 predominantly no including 'fuzzy' yes/no (spontaneous)
  - 1 Background knowledge of the language (spontaneous)
- 3 Dates:
  - 1 birthdate (spontaneous)
  - 1 prompted date phrase, in words not using digital format
  - 1 relative and general date expression
- 2 Times:
  - 1 time of day (spontaneous)

        - 1 prompted time phrase, in analogue not digital form
    - 9 Application keywords/keyphrases
    - 1 Word spotting phrase using embedded application words
    - 1 Spontaneous sentences
    - 7 Directory assistance names:
        - 1 city of birth/growing up (spontaneous)
        - 2 most frequent cities (set of 500)
        - 2 most frequent companies/agencies (set of 500)
        - 1 proper name (forename and surname) (150 names)
        - 1 proper name
    - 5 Spellings:
        - 2 real/artificial words to maximise letter coverage
        - 2 spelling e.g. of directory assistance city name
        - 1 spelling of proper name e.g. own forename (spontaneous)
    - 4 Phonetically rich words
    - 9 Phonetically rich sentences
    - 1 Acoustic conditions (spontaneous)

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

The transcription included in this database is an orthographic, lexical transcription with a few details that represent audible acoustic events (speech and non speech) present in the corresponding waveform files. The extra marks contained in the transcription aid in interpreting the text form of the utterance.

4.4. Lexicon. Description of the lexicon (if applicable)

The lexicon file is an alphabetically ordered list of distinct lexical items (i.e. splitting tokens at every space, essentially words in our case) which occur in the corpus with the corresponding pronunciation information. Each distinct word has a separate entry. As the lexicon is derived from the corpus it uses the same alphabetic encoding for special and accented characters as used in the transcriptions.

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

The database contains recordings from 602 adult speakers.

Speakers accent have been categorized to 5 geographical regions. The geographical regions have been defined in order to cover the major regional variants of Basque. These 5 regions have been grouped into two sets, the standard Basque “Batua” and the dialectal form “Bizkaiera”. For the dialect “Bizkaiera” prompt sheets using specific material of the dialect were generated. For the rest of the regions (group “Batua”) standard Basque was used.

Age distribution: 4 speakers are under 16, 281 speakers are between 15 and 30 years old, 185 speakers are between 31 and 45, 118 speakers are between 46 and 60, and 14 speakers are over 60.

4.6. Recording platform

4.6.1.Domain(s), environments,

Speakers were calling from the mobile telephone network

4.6.2 Recording platform

The recording platform is based on a PC with an ISDN-BRI interface.

UPF - University Pompeu Fabra

Endogenous resources

TRL V-Subcat Lexicon

Gold-standard for Spanish verbal subcategorization frames. The gold-standard was built merging two manually developed dictionaries: the Spanish working lexicon of the Incyta Machine Translation system (Alonso and Bocsák, 2005) and the Spanish Resource Grammar (Marimon 2010).

BASIC INFORMATION

Identification Information

- resourceName: TRL Spanish V-SUBCAT lexicon: LMF Format
- resourceShortName: TRL Spanish V-SUBCAT lexicon
- url: http://panacea-lr.eu/en/info-for-researchers/test-sets-gold-standards-and-other-material/subcategorization-frames/spanish-scf-gold-standard/trl-spanish-v-subcat-lexicon-general-domain
- metaShareId: NOT_DEFINED_FOR_V2

- description:

Gold-standard for Spanish verbal subcategorization frames. The gold-standard was built merging two manually developed dictionaries: the Spanish working lexicon of the Incyta Machine Translation system (Alonso and Bocsák, 2005) and the Spanish Resource Grammar (Marimon 2010).

See Necsulescu et al (2011) for details about the process of merging both dictionaries and about the information encoded in the feature structures.

ADMINISTRATIVE INFORMATION

Contact Person(s)

- surname: Padró
- givenName: Muntsa
- communicationInfo:
    - email: muntsa.padro@upf.edu
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
- affiliation:
    - organizationName: Universitat Pompeu Fabra - Institut Universitari Lingüística Aplicada
    - organizationShortName: IULA - UPF
    - departmentName: Institut Universitari Lingüística Aplicada (UPF)
    - communicationInfo:
        - email: muntsa.padro@upf.edu
        - url: http://www.iula.upf.edu/

- address: Roc Boronat, 138
- zipCode: 08018
- city: Barcelona
- country: Spain
- telephoneNumber: +34 93 5421207
- faxNumber: +34 93 5422321

Distribution Information

- Availability type: available-unrestrictedUse
- Licence: CC_BY-SA

TECHNICAL INFORMATION

- Resource type: lexicalConceptualResource
- Format: text/xml
- Character Encoding: UTF-8
- Validation Information:
  - validated: true
  - validationType: formal
  - validationMode: automatic
  - validationModeDetails: The lexicon validates against the LMF DTD v.16

CONTENT INFORMATION

- Encoding information:
  - encodingLevel: syntax
  - linguisticInformation: lemma
  - linguisticInformation: syntax-SubcatFrame
  - conformanceToStandardsBestPractices: LMF
- Linguality: monolingual
- Language(s): Spanish

RELEVANT REFERENCES AND OTHER INFORMATION

Resource Documentation

- · documentUnstructured: Necsulescu, Silvia; Bel, Núria; Padró, Muntsa; Marimon, Montserrat; Revilla, Eva. 2011. Towards the Automatic Merging of Language Resources, Woler 2011. Ljubljana, Slovenia. Available at http://alpage.inria.fr/~sagot/woler2011/WoLeR2011/Program_files/WoLeR%202011%20-%20Necsulescu%20Bel%20Padro%20Marimon%20Revilla.pdf

Resource Creation Information

Resource Creator(s):

- organizationInfo:
  - organizationName: Universitat Pompeu Fabra - Institut Universitari Lingüística Aplicada
  - organizationShortName: IULA - UPF
  - departmentName: Institut Universitari Lingüística Aplicada (UPF)
  - communicationInfo:

- email: muntsa.padro@upf.edu
- url: http://www.iula.upf.edu/
- address: Roc Boronat, 138
- zipCode: 08018
- city: Barcelona
- country: Spain
- telephoneNumber: +34 93 5421207
- faxNumber: +34 93 5422321

Resource Funding:

- projectName: PANACEA
- projectShortName: PANACEA
- url: http://panacea-lr.eu/
- fundingType: ownFunds
- fundingType: euFunds
- funder: UPF
- funder: EU
- fundingCountry: Spain

- projectName: METANET4U
- projectShortName: METANET4U
- url: http://metanet4u.eu/
- fundingType: ownFunds
- fundingType: euFunds
- funder: UPF
- funder: EU
- fundingCountry: Spain

Restricted exogenous resources

CESS_EU: The Basque Dependency Treebank

This is the stand-off GrAF version of the Basque Dependency Treebank (BDT). It is the Reference Corpus for the Processing of Basque (EPEC) annotated at syntactic level. EPEC is a 300,000 word corpus of standard written journal texts which aims to be a training corpus for the development and inprovement of several Natural Language Procesing tools. It has been manually tagged at different levels: morphology, partial syntax and semantic This is the stand-off GrAF version of the Constituent Basque Treebank.

BASIC INFORMATION

Identification Information

- resourceName: GrAF version of the Basque Dependency Treebank
- resourceShortName: GrAF version of the Basque Dependency Treebank
- url: http://ixa.si.ehu.es/Ixa/Produktuak/1306407157

- metaShareId: NOT_DEFINED_FOR_V2

- description:

This is the stand-off GrAF version of the Basque Dependency Treebank (BDT). It is the Reference Corpus for the Processing of Basque (EPEC) annotated at syntactic level. EPEC is a 300,000 word corpus of standard written journal texts which aims to be a training corpus for the development and inprovement of several Natural Language Procesing tools. It has been manually tagged at different levels: morphology, partial syntax and semantic This is the stand-off GrAF version of the Constituent Basque Treebank.

ADMINISTRATIVE INFORMATION

Contact Person(s)

- surname: Aranzabe
- givenName: Maxux
- communicationInfo:
    - email: maxux.aranzabe@ehu.es
    - country: Spain
- affiliation:
    - organizationName: University of the Basque Country - IXA Group
    - organizationShortName: UPV
    - communicationInfo:
        - email: ixa@ehu.es
        - url: https://ixa.si.ehu.es/Ixa/index_html
        - country: Spain

- surname: Vivaldi
- givenName: Jorge
- communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
- affiliation:
    - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
    - organizationShortName: IULA UPF
    - departmentName: Institut Universitari Lingüística Aplicada
    - communicationInfo:
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/
        - address: Roc Boronat, 138
        - zipCode: 08018
        - city: Barcelona
        - country: Spain
        - telephoneNumber: +34 93 5422332
        - faxNumber: +34 93 5422321

Distribution Information

- Availability type: available-unrestrictedUse
- Licence: CC_BY-NC-SA

TECHNICAL INFORMATION

- Resource type: corpus
- Character Encoding: UTF-8
- Validation Information:
    - validated: true
    - validationType: formal
    - validationMode: automatic
    - validationModeDetails: The documents validates against the GrAF DTD v.1.0.4

CONTENT INFORMATION

- Linguality: monolingual
- Language(s): Euskera
- Size: 30.672.000 bytes
- Annotation information:
    - annotationType: syntacticAnnotation-treebanks
    - annotationStandoff: true
    - segmentationLevel: phrase
    - annotationFormat: GrAF
    - tagset: other

RELEVANT REFERENCES AND OTHER INFORMATION

Resource Documentation

- · documentUnstructured: Arantza Diaz de Ilarraza Sánchez, Enrique Fernández Terrones, Izaskun Aldezabal Roteta, Maria Jesús Aranzabe Urruzola (2008). "From Dependencies to Constituents in the Reference Corpus for the Processing of Basque (EPEC)". In Procesamiento del lenguaje Natural, nº 41 (2008), pp. 147-154.

- · documentUnstructured: Web site of the CESS-ECE project: http://clic.ub.edu/cessece/

Resource Creation Information

Resource Creator(s):

- organizationInfo:
    - organizationName: IXA
    - organizationShortName: IXA
    - communicationInfo:
        - email: ixa@ehu.es
        - url: https://ixa.si.ehu.es/Ixa/index_html
        - country: Spain

- personInfo:
    - surname: Aldezabal

    - givenName: Izaskun
    - communicationInfo:
        - email: izaskun.aldezabal@ehu.es

- personInfo:
    - surname: Aranzabe
    - givenName: Maxux
    - communicationInfo:
        - email: maxux.aranzabe@ehu.es

- personInfo:
    - surname: Arriola
    - givenName: Jose Mari
    - communicationInfo:

- personInfo:
    - surname: Atutxa
    - givenName: Aitziber
    - communicationInfo:

- personInfo:
    - surname: Díaz de Ilarraza
    - givenName: Arantza
    - communicationInfo:
        - email: a.diazdeilarraza@ehu.es

- personInfo:
    - surname: Estarrona
    - givenName: Ainara
    - communicationInfo:
        - email: ainara.estarrona@ehu.es

- personInfo:
    - surname: Fernandez
    - givenName: Kike
    - communicationInfo:
        - email: sisfetek@ehu.es

- personInfo:
    - surname: Iruskieta
    - givenName: Mikel
    - communicationInfo:

- personInfo:
    - surname: Uria
    - givenName: Larraitz
    - communicationInfo:

- organizationInfo:
    - organizationName: Universitat Pompeu Fabra - Institut Universitari Lingüística Aplicada
    - organizationShortName: IULA - UPF

- departmentName: Institut Universitari Lingüística Aplicada (UPF)
- communicationInfo:
  - email: jorge.vivaldi@upf.edu
  - url: http://www.iula.upf.edu/
  - address: Roc Boronat, 138
  - zipCode: 08018
  - city: Barcelona
  - country: Spain
  - telephoneNumber: +34 93 5421207
  - faxNumber: +34 93 5422321

Resource Funding:

- projectName: CESS-ECE Syntactically and Semantically Annotated Corpora (Spanish, Catalan, Basque)
- projectShortName: CESS-ECE Project
- url: http://clic.ub.edu/cessece
- fundingType: nationalFunds
- funder: Ministerio de Educación y Ciencia. Gobierno de España ((HUM2004-21127)
- fundingCountry: Spain

- projectName: METANET4U
- projectShortName: METANET4U
- url: http://metanet4u.eu/
- fundingType: ownFunds
- fundingType: euFunds
- funder: UPF
- funder: EU
- fundingCountry: Spain

Termoteca

This lexical resource is the LMF version of the Termoteca, a multilingual terminological database based on the monolingual and parallel speciality texts collected in the corpora of the University of Vigo, namely in the CLUVI Corpus and in the Galician Technical Corpus

BASIC INFORMATION

Identification Information

- resourceName: Termoteca
- resourceShortName: Termoteca
- url: http://sli.uvigo.es/termoteca/
- metaShareId: NOT_DEFINED_FOR_V2

- description:

This lexical resource is the LMF version of the Termoteca, a multilingual terminological database based on the monolingual and parallel speciality texts collected in the corpora of the University of Vigo, namely in the CLUVI Corpus and in the Galician Technical Corpus

ADMINISTRATIVE INFORMATION

Contact Person(s)

- surname: Gómez Guinovart
- givenName: Xavier
- communicationInfo:
    - email: xgg@uvigo.es
    - address: Campus da Universidade de Vigo
    - zipCode: 36310
    - city: Vigo
    - country: Spain
- affiliation:
    - organizationName: Universidade de Vigo
    - organizationShortName: UVIGO
    - departmentName: Tecnoloxías e Aplicacións da Lingua Galega (Grupo TALG)
    - communicationInfo:
        - email: xgg@uvigo.es
        - url: http://sli.uvigo.es
        - address: Campus da Universidade de Vigo
        - zipCode: 36310
        - city: Vigo
        - country: Spain
        - telephoneNumber: +34 986 813858
        - faxNumber: +34 986 812380

Distribution Information

- Availability type: available-restrictedUse
- Licence: AGPL

TECHNICAL INFORMATION

- Resource type: lexicalConceptualResource
- Format: text/xml
- Character Encoding: UTF-8
- Creation information:
    - originalSource:
        - targetResourceNameURI: Termoteca located at http://sli.uvigo.es/termoteca/
    - creationMode: automatic
- Validation Information:
    - validated: true
    - validationType: formal
    - validationMode: automatic
    - validationModeDetails: The lexicon validates against the LMF DTD v.16

CONTENT INFORMATION

- Encoding information:
    - encodingLevel: morphology
    - linguisticInformation: lemma
    - linguisticInformation: definition/gloss
    - linguisticInformation: usage-Examples
    - linguisticInformation: usage-Frequency
    - linguisticInformation: semantics-Domain
    - linguisticInformation: semantics-SemanticRoles
    - linguisticInformation: semantics-CrossReferences
    - conformanceToStandardsBestPractices: LMF
- Linguality: multilingual
- Language(s): Galician Spanish English French Portuguese

RELEVANT REFERENCES AND OTHER INFORMATION

Resource Documentation

- · documentUnstructured: Gómez Guinovart, Xavier (2012): A Hybrid Corpus-Based Approach to Bilingual Terminology Extraction. In I. Moskowich-Spiegel Fandiño, B. Crespo (eds.). Encoding the Past, Decoding The Future: Corpora in the 21st Century. Cambridge Scholar Publishing: Newcastle upon Tyne, pp. 147-175 (ISBN 1-4438-3581-1). Available at http://webs.uvigo.es/sli/arquivos/bi_term_extraction.pdf

Resource Creation Information

Resource Creator(s):

- organizationInfo:
    - organizationName: Universidade de Vigo
    - organizationShortName: UVIGO
    - departmentName: Tecnoloxías e Aplicacións da Lingua Galega (Grupo TALG)
    - communicationInfo:
        - email: xgg@uvigo.es
        - url: http://sli.uvigo.es
        - address: Campus da Universidade de Vigo
        - zipCode: 36310
        - city: Vigo
        - country: Spain
        - telephoneNumber: +34 986 813858
        - faxNumber: +34 986 812380

- organizationInfo:
    - organizationName: Universitat Pompeu Fabra - Institut Universitari Lingüística Aplicada
    - organizationShortName: IULA - UPF
    - departmentName: Institut Universitari Lingüística Aplicada (UPF)
    - communicationInfo:
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/
        - address: Roc Boronat, 138
        - zipCode: 08018
        - city: Barcelona
        - country: Spain

- telephoneNumber: +34 93 5421207
- faxNumber: +34 93 5422321

Resource Funding:

- projectName: METANET4U
- projectShortName: METANET4U
- url: http://metanet4u.eu/
- fundingType: ownFunds
- fundingType: euFunds
- funder: UPF
- funder: EU
- fundingCountry: Spain

Unrestricted exogenous resources

Apertium English dictionary

This is the LMF version of the Apertium English dictionary. Monolingual dictionary for English was generated from the Apertium expanded lexicon of the en-es pair system (English/Spanish).

BASIC INFORMATION

Identification Information

- resourceName: English LMF Apertium Dictionary
- resourceShortName: LMF Apertium En
- url: http://www.apertium.org
- metaShareId: NOT_DEFINED_FOR_V2

- description:

This is the LMF version of the Apertium English dictionary. Monolingual dictionary for English was generated from the Apertium expanded lexicon of the en-es pair system (English/Spanish).

Apertium is a free/open-source machine translation platform, initially aimed at related-language pairs but recently expanded to deal with more divergent language pairs (such as English-Catalan).

The platform provides: a language-independent machine translation engine; tools to manage the linguistic data necessary to build a machine translation system for a given language pair and linguistic data for a growing number of language pairs.

ADMINISTRATIVE INFORMATION

Contact Person(s)

- surname: Forcada
- givenName: Mikel
- communicationInfo:
    - email: mlf@dlsi.ua.es
    - zipCode: 03080
    - city: Alicante
    - country: Spain
- affiliation:
    - organizationName: Universitat d'Alacant, (Grup Transducens)
    - organizationShortName: Universitat d'Alacant, Grup Transducens
    - communicationInfo:
        - email: info@prompsit.com
        - url: http://transducens.dlsi.ua.es/
        - address: Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante. 03080 Alicante
        - zipCode: 03080
        - city: Alicante
        - country: Spain
        - telephoneNumber: +34 96 5903772
        - faxNumber: +34 96 5909326

- surname: Vivaldi
- givenName: Jorge
- communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
- affiliation:
    - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
    - organizationShortName: IULA UPF
    - departmentName: Institut Universitari Lingüística Aplicada
    - communicationInfo:
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/
        - address: Roc Boronat, 138
        - zipCode: 08018
        - city: Barcelona
        - country: Spain
        - telephoneNumber: +34 93 5421207
        - faxNumber: +34 93 5422321

Distribution Information

- Availability type: available-unrestrictedUse
- Licence: GPL

TECHNICAL INFORMATION

- Resource type: lexicalConceptualResource
- Format: text/xml
- Character Encoding: UTF-8
- Creation information:
    - originalSource:
        - targetResourceNameURI: Apertium Monolingual dictionary as described in http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
    - creationMode: automatic
    - creationModeDetails: This lexicon was created with the ApertiumMonolingual2LMF.pl script
    - creationTool:
        - targetResourceNameURI: ApertiumMonolingual2LMF.pl
- Validation Information:
    - validated: true
    - validationType: formal
    - validationMode: automatic
    - validationModeDetails: The lexicon validates against the LMF DTD v.16

CONTENT INFORMATION

- Encoding information:
    - encodingLevel: morphology
    - linguisticInformation: lemma
    - linguisticInformation: partOfSpeech
    - conformanceToStandardsBestPractices: LMF
- Linguality: monolingual
- Language(s): English

RELEVANT REFERENCES AND OTHER INFORMATION

Resource Documentation

- · documentUnstructured: Forcada,M.L., Ivanov, B., Ortiz, S., Pérez, J.A., Ramírez, G., Sánchez, F., Armentano, C., Montava, M.A., Tyers, F.M., (2010). 'Documentation of the Open Source Shallow-Transfer Machine Translation Platform Apertium'. Departament de Llenguatges i Sistemes Informàtics Universitat d'Alacant, March 2010. http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf

Resource Creation Information

Resource Creator(s):

- organizationInfo:
    - organizationName: Universitat d'Alacant, (Grup Transducens)
    - organizationShortName: Universitat d'Alacant, Grup Transducens
    - communicationInfo:
        - email: info@prompsit.com
        - url: http://transducens.dlsi.ua.es/
        - address: Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante. 03080 Alicante
        - zipCode: 03080
        - city: Alicante
        - country: Spain

        - telephoneNumber: +34 96 5903772
        - faxNumber: +34 96 5909326

- organizationInfo:
    - organizationName: Universitat Pompeu Fabra - Institut Universitari Lingüística Aplicada
    - organizationShortName: IULA - UPF
    - departmentName: Institut Universitari Lingüística Aplicada (UPF)
    - communicationInfo:
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/
        - address: Roc Boronat, 138
        - zipCode: 08018
        - city: Barcelona
        - country: Spain
        - telephoneNumber: +34 93 5421207
        - faxNumber: +34 93 5422321

- organizationInfo:
    - organizationName: Universitat Politècnica de Catalunya
    - organizationShortName: UPC
    - communicationInfo:
        - url: http://upc.edu
        - city: Barcelona
        - country: Spain

- personInfo:
    - surname: Jimmy O'Reagan
    - communicationInfo:

- personInfo:
    - surname: Paul "greenbreen" Breen
    - communicationInfo:

Resource Funding:

- projectName: Opentrad
- projectShortName: Opentrad
- url: http://www.opentrad.com/
- fundingType: nationalFunds
- funder: Ministerio de Industria, Turismo y Comercio. Gobierno de España
- fundingCountry: Spain

- projectName: Apertium 2.0
- projectShortName: Apertium
- url: http://www.apertium.org
- fundingType: nationalFunds
- funder: Ministerio de Industria, Turismo y Comercio. Gobierno de España
- funder: Secretariat de Telecomunicacions i Societat de la Informació de Generalitat de Catalunya
- fundingCountry: Spain

- projectName: METANET4U

- projectShortName: METANET4U
- url: http://metanet4u.eu/
- fundingType: ownFunds
- fundingType: euFunds
- funder: UPF
- funder: EU
- fundingCountry: Spain

Apertium French dictionary

This is the LMF version of the Apertium French dictionary. Monolingual dictionariy for French was generated from the Apertium expanded lexicon of the fr-es pair system (French/Spanish).

BASIC INFORMATION

Identification Information

- resourceName: French LMF Apertium Dictionary
- resourceShortName: LMF Apertium Es
- url: http://www.apertium.org
- metaShareId: NOT_DEFINED_FOR_V2

- description:

This is the LMF version of the Apertium French dictionary. Monolingual dictionariy for French was generated from the Apertium expanded lexicon of the fr-es pair system (French/Spanish).

Apertium is a free/open-source machine translation platform, initially aimed at related-language pairs but recently expanded to deal with more divergent language pairs (such as English-Catalan).

The platform provides: a language-independent machine translation engine; tools to manage the linguistic data necessary to build a machine translation system for a given language pair and linguistic data for a growing number of language pairs.

ADMINISTRATIVE INFORMATION

Contact Person(s)

- surname: Forcada
- givenName: Mikel
- communicationInfo:
    - email: mlf@dlsi.ua.es
    - zipCode: 03080
    - city: Alicante
    - country: Spain
- affiliation:
    - organizationName: Universitat d'Alacant, (Grup Transducens)
    - organizationShortName: Universitat d'Alacant, Grup Transducens
    - communicationInfo:
        - email: info@prompsit.com

            - url: http://transducens.dlsi.ua.es/
            - address: Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante. 03080 Alicante
            - zipCode: 03080
            - city: Alicante
            - country: Spain
            - telephoneNumber: +34 96 5903772
            - faxNumber: +34 96 5909326

- surname: Vivaldi
- givenName: Jorge
- communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
- affiliation:
    - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
    - organizationShortName: IULA UPF
    - departmentName: Institut Universitari Lingüística Aplicada
    - communicationInfo:
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/
        - address: Roc Boronat, 138
        - zipCode: 08018
        - city: Barcelona
        - country: Spain
        - telephoneNumber: +34 93 5421207
        - faxNumber: +34 93 5422321

Distribution Information

- Availability type: available-unrestrictedUse
- Licence: GPL

TECHNICAL INFORMATION

- Resource type: lexicalConceptualResource
- Format: text/xml
- Character Encoding: UTF-8
- Creation information:
    - originalSource:
        - targetResourceNameURI: Apertium Monolingual dictionary as described in http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
    - creationMode: automatic
    - creationModeDetails: This lexicon was created with the ApertiumMonolingual2LMF.pl script
    - creationTool:
        - targetResourceNameURI: ApertiumMonolingual2LMF.pl
- Validation Information:
    - validated: true
    - validationType: formal

- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

CONTENT INFORMATION

- Encoding information:
  - encodingLevel: morphology
  - linguisticInformation: lemma
  - linguisticInformation: partOfSpeech
  - conformanceToStandardsBestPractices: LMF
- Linguality: monolingual
- Language(s): French

RELEVANT REFERENCES AND OTHER INFORMATION

Resource Documentation

- · documentUnstructured: Forcada,M.L., Ivanov, B., Ortiz, S., Pérez, J.A., Ramírez, G., Sánchez, F., Armentano, C., Montava, M.A., Tyers, F.M., (2010). 'Documentation of the Open Source Shallow-Transfer Machine Translation Platform Apertium'. Departament de Llenguatges i Sistemes Informàtics Universitat d'Alacant, March 2010. http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf

Resource Creation Information

Resource Creator(s):

- organizationInfo:
  - organizationName: Prompsit Language Engineering, S.L
  - organizationShortName: Prompsit
  - communicationInfo:
    - email: info@prompsit.com
    - url: http://www.prompsit.com/
    - address: Avenida Universidad, s/n. Edificio Quorum III. 03202 Elche (Alicante). España.
    - zipCode: 03202
    - city: Elche
    - country: Spain
    - telephoneNumber: (+34) 965457549

- organizationInfo:
  - organizationName: Eleka Ingenieritza Linguistikoa S.L
  - organizationShortName: ELEKA SL
  - communicationInfo:
    - email: info@eleka.net
    - url: http://www.eleka.net
    - address: Zelai Haundi kalea 3 - Polígono Industrial Osinalde - 20170 Usurbil - Gipuzkoa
    - zipCode: 20170
    - city: Usurbil
    - country: Spain
    - telephoneNumber: (+34) 943 377 225

- organizationInfo:

- organizationName: Universitat Pompeu Fabra - Institut Universitari Lingüística Aplicada
- organizationShortName: IULA - UPF
- departmentName: Institut Universitari Lingüística Aplicada (UPF)
- communicationInfo:
  - email: jorge.vivaldi@upf.edu
  - url: http://www.iula.upf.edu/
  - address: Roc Boronat, 138
  - zipCode: 08018
  - city: Barcelona
  - country: Spain
  - telephoneNumber: +34 93 5421207
  - faxNumber: +34 93 5422321

Resource Funding:

- projectName: Opentrad
- projectShortName: Opentrad
- url: http://www.opentrad.com/
- fundingType: nationalFunds
- funder: Ministerio de Industria, Turismo y Comercio. Gobierno de España
- fundingCountry: Spain

- projectName: Apertium 2.0
- projectShortName: Apertium
- url: http://www.apertium.org
- fundingType: nationalFunds
- funder: Ministerio de Industria, Turismo y Comercio. Gobierno de España
- funder: Secretariat de Telecomunicacions i Societat de la Informació de Generalitat de Catalunya
- fundingCountry: Spain

- projectName: METANET4U
- projectShortName: METANET4U
- url: http://metanet4u.eu/
- fundingType: ownFunds
- fundingType: euFunds
- funder: UPF
- funder: EU
- fundingCountry: Spain

Apertium Italian dictionary

This is the LMF version of the Apertium Italian dictionary. Monolingual dictionary for Italian was generated from the Apertium expanded lexicon of the es-it.

BASIC INFORMATION

Identification Information

- resourceName: Italian LMF Apertium Dictionary
- resourceShortName: LMF Apertium It

- url: http://www.apertium.org
- metaShareId: NOT_DEFINED_FOR_V2

- description:

This is the LMF version of the Apertium Italian dictionary. Monolingual dictionary for Italian was generated from the Apertium expanded lexicon of the es-it.

Apertium is a free/open-source machine translation platform, initially aimed at related-language pairs but recently expanded to deal with more divergent language pairs (such as English-Catalan).

The platform provides: a language-independent machine translation engine; tools to manage the linguistic data necessary to build a machine translation system for a given language pair and linguistic data for a growing number of language pairs.

ADMINISTRATIVE INFORMATION

Contact Person(s)

- surname: Forcada
- givenName: Mikel
- communicationInfo:
    - email: mlf@dlsi.ua.es
    - zipCode: 03080
    - city: Alicante
    - country: Spain
- affiliation:
    - organizationName: Universitat d'Alacant, (Grup Transducens)
    - organizationShortName: Universitat d'Alacant, Grup Transducens
    - communicationInfo:
        - email: info@prompsit.com
        - url: http://transducens.dlsi.ua.es/
        - address: Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante. 03080 Alicante
        - zipCode: 03080
        - city: Alicante
        - country: Spain
        - telephoneNumber: +34 96 5903772
        - faxNumber: +34 96 5909326

- surname: Vivaldi
- givenName: Jorge
- communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
- affiliation:
    - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
    - organizationShortName: IULA UPF

- departmentName: Institut Universitari Lingüística Aplicada
- communicationInfo:
  - email: jorge.vivaldi@upf.edu
  - url: http://www.iula.upf.edu/
  - address: Roc Boronat, 138
  - zipCode: 08018
  - city: Barcelona
  - country: Spain
  - telephoneNumber: +34 93 5421207
  - faxNumber: +34 93 5422321

Distribution Information

- Availability type: available-unrestrictedUse
- Licence: GPL

TECHNICAL INFORMATION

- Resource type: lexicalConceptualResource
- Format: text/xml
- Character Encoding: UTF-8
- Creation information:
  - originalSource:
    - targetResourceNameURI: Apertium Monolingual dictionary as described in http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
  - creationMode: automatic
  - creationModeDetails: This lexicon was created with the ApertiumMonolingual2LMF.pl script
  - creationTool:
    - targetResourceNameURI: ApertiumMonolingual2LMF.pl
- Validation Information:
  - validated: true
  - validationType: formal
  - validationMode: automatic
  - validationModeDetails: The lexicon validates against the LMF DTD v.16

CONTENT INFORMATION

- Encoding information:
  - encodingLevel: morphology
  - linguisticInformation: lemma
  - linguisticInformation: semantics-CrossReferences
  - linguisticInformation: partOfSpeech
  - conformanceToStandardsBestPractices: LMF
- Linguality: monolingual
- Language(s): Italian

RELEVANT REFERENCES AND OTHER INFORMATION

Resource Documentation

- · documentUnstructured: http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf

- · documentUnstructured: Forcada,M.L., Ivanov, B., Ortiz, S., Pérez, J.A., Ramírez, G., Sánchez, F., Armentano, C., Montava, M.A., Tyers, F.M., (2010). 'Documentation of the Open Source Shallow-Transfer Machine Translation Platform Apertium'. Departament de Llenguatges i Sistemes Informàtics Universitat d'Alacant, March 2010.

Resource Creation Information

Resource Creator(s):

- organizationInfo:
  - organizationName: Prompsit Language Engineering, S.L
  - organizationShortName: Prompsit
  - communicationInfo:
    - email: info@prompsit.com
    - url: http://www.prompsit.com/
    - address: Avenida Universidad, s/n. Edificio Quorum III. 03202 Elche (Alicante). España.
    - zipCode: 03202
    - city: Elche
    - country: Spain
    - telephoneNumber: (+34) 965457549

- organizationInfo:
  - organizationName: Universitat d'Alacant, (Grup Transducens)
  - organizationShortName: Universitat d'Alacant, Grup Transducens
  - communicationInfo:
    - email: info@prompsit.com
    - url: http://transducens.dlsi.ua.es/
    - address: Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante. 03080 Alicante
    - zipCode: 03080
    - city: Alicante
    - country: Spain
    - telephoneNumber: +34 96 5903772
    - faxNumber: +34 96 5909326

- organizationInfo:
  - organizationName: Universitat Pompeu Fabra - Institut Universitari Lingüística Aplicada
  - organizationShortName: IULA - UPF
  - departmentName: Institut Universitari Lingüística Aplicada (UPF)
  - communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - url: http://www.iula.upf.edu/
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - telephoneNumber: +34 93 5421207
    - faxNumber: +34 93 5422321

Resource Funding:

- projectName: Opentrad

- projectShortName: Opentrad
- url: http://www.opentrad.com/
- fundingType: nationalFunds
- funder: Ministerio de Industria, Turismo y Comercio. Gobierno de España
- fundingCountry: Spain

- projectName: Apertium 2.0
- projectShortName: Apertium
- url: http://www.apertium.org
- fundingType: nationalFunds
- funder: Ministerio de Industria, Turismo y Comercio. Gobierno de España
- funder: Secretariat de Telecomunicacions i Societat de la Informació de Generalitat de Catalunya
- fundingCountry: Spain

- projectName: METANET4U
- projectShortName: METANET4U
- url: http://metanet4u.eu/
- fundingType: ownFunds
- fundingType: euFunds
- funder: UPF
- funder: EU
- fundingCountry: Spain

WikiCorpus

Documentation for the 'GrAF version of Spanish portions of Wikipedia Corpus'

This is the stand-off GrAF version of Spanish portions of the Wikipedia (based on a 2006 dump). The original Wikipedia Spanish Corpus contains about 120 million words in raw text format. It has been cleaned by erase disambiguation pages, removing some XML tags and homengenizing lists ending tag. Then, the corpus has been processed for adding structural tagging (head, paragraph, sentence, list, etc.) and morphosyntactic information.

BASIC INFORMATION

Identification Information

- resourceName: GrAF version of Spanish portions of Wikipedia Corpus
- resourceShortName: GrAF version of the Spanish Wikipedia
- url: http://www.lsi.upc.edu/~nlp/wikicorpus/
- metaShareId: NOT_DEFINED_FOR_V2

- description:

This is the stand-off GrAF version of Spanish portions of the Wikipedia (based on a 2006 dump). The original Wikipedia Spanish Corpus contains about 120 million words in raw text format. It has been cleaned by erase disambiguation pages, removing some XML tags and homengenizing lists ending tag. Then, the corpus has been processed for adding structural tagging (head, paragraph, sentence, list, etc.) and morphosyntactic information.

ADMINISTRATIVE INFORMATION

Contact Person(s)

- surname: Boleda
- givenName: Gemma
- communicationInfo:
    - email: gemma.boleda@upf.edu
    - country: Spain
- affiliation:
    - organizationName: Universitat Politécnica de Catalunya
    - organizationShortName: UPC
    - communicationInfo:
        - email: gemma.boleda@upf.edu
        - url: http://www.upc.edu/
        - country: Spain

- surname: Vivaldi
- givenName: Jorge
- communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
- affiliation:
    - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
    - organizationShortName: IULA UPF
    - departmentName: Institut Universitari Lingüística Aplicada
    - communicationInfo:
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/
        - address: Roc Boronat, 138
        - zipCode: 08018
        - city: Barcelona
        - country: Spain
        - telephoneNumber: +34 93 5422332
        - faxNumber: +34 93 5422321

Distribution Information

- Availability type: available-unrestrictedUse
- Licence: GFDL

TECHNICAL INFORMATION

- Resource type: corpus
- Character Encoding: UTF-8
- Validation Information:
    - validated: true
    - validationType: formal

- validationMode: automatic
- validationModeDetails: The documents validates against the GrAF DTD v.1.0.4

CONTENT INFORMATION

- Linguality: monolingual
- Language(s): Spanish
- Size: 113.800.200 tokens
- Annotation information:
  - annotationType: morphosyntacticAnnotation-posTagging
  - annotationStandoff: true
  - segmentationLevel: phrase
  - annotationFormat: GrAF
  - tagset: PAROLE

RELEVANT REFERENCES AND OTHER INFORMATION

Resource Documentation

- · documentUnstructured: Samuel Reese, Gemma Boleda, Montse Cuadros, Lluís Padró, German Rigau (2010). "Wikicorpus: A Word-Sense Disambiguated Multilingual Wikipedia Corpus". In Proceedings of 7th Language Resources and Evaluation Conference (LREC'10), La Valleta, Malta. May, 2010.

- · documentUnstructured: Samuel Reese. 2009. WikiNet: Construction d'une ressource lexico-sémantique multilingue à partir de Wikipedia. Master's thesis. ISAE (Institut Supérieur de l'Aéronautique et de l'Espace), Toulouse, France.

- · documentUnstructured: Web site of the Wikicorpus project: http://www.lsi.upc.edu/~nlp/wikicorpus/

Resource Creation Information

Resource Creator(s):

- organizationInfo:
  - organizationName: Research Group on Natural Language Processing - Grup de recerca consolidat de la Generalitat de Catalunya (2009 SGR-01723)
  - organizationShortName: NLP Research Group
  - communicationInfo:
    - email: castell@lsi.upc.edu
    - url: http://www.lsi.upc.edu/~nlp/
    - country: Spain

- personInfo:
  - surname: Boleda
  - givenName: Gemma
  - communicationInfo:
    - email: gemma.boleda@upf.edu

- organizationInfo:
  - organizationName: Universitat Pompeu Fabra - Institut Universitari Lingüística Aplicada
  - organizationShortName: IULA - UPF

- departmentName: Institut Universitari Lingüística Aplicada (UPF)
- communicationInfo:
  - email: jorge.vivaldi@upf.edu
  - url: http://www.iula.upf.edu/
  - address: Roc Boronat, 138
  - zipCode: 08018
  - city: Barcelona
  - country: Spain
  - telephoneNumber: +34 93 5421207
  - faxNumber: +34 93 5422321

Resource Funding:

- projectName: KNOW: Developing large-scale multilingual technologies for language understanding
- projectShortName: KNOW
- url: http://ixa.si.ehu.es/know/
- fundingType: nationalFunds
- funder: Ministerio de educación y Ciencia. Gobierno de España (TIN2006-15049-C03)
- fundingCountry: Spain

- projectName: Language understanding technologies for multilingual domain-oriented information access
- projectShortName: KNOW2
- url: http://ixa.si.ehu.es/know2
- fundingType: nationalFunds
- funder: Ministerio de educación y Ciencia. Gobierno de España (TIN2009-14715-C04)
- fundingCountry: Spain

- projectName: PASCAL 2 Pattern Analysis, Statistical Modelling and Computational Learning
- projectShortName: Pascal Network of Excellence
- url: http://www.pascal-network.org/
- fundingType: euFunds
- funder: European Union (FP7-ICT-216886)

- projectName: Programa Juan de la Cierva
- projectShortName: Programa Juan de la Cierva
- url: http://www.idi.mineco.gob.es/portal/site/MICINN/menuitem.dbc68b34d11ccbd5d52ffeb801432ea0/?vgnextoid=f900759903236210VgnVCM1000001d04140aRCRD
- fundingType: nationalFunds
- funder: Ministerio de educación y Ciencia. Gobierno de España (JCI-2007-57-1479)
- fundingCountry: Spain

- projectName: METANET4U
- projectShortName: METANET4U
- url: http://metanet4u.eu/
- fundingType: ownFunds
- fundingType: euFunds
- funder: UPF
- funder: EU
- fundingCountry: Spain

Documentation for the 'GrAF version of Catalan portions of Wikipedia Corpus'

This is the stand-off GrAF version of Catalan portions of the Wikipedia (based on a 2006 dump). The original Wikipedia Catalan Corpus contains about 120 million words in raw text format. It has been cleaned by erase disambiguation pages, removing some XML tags and homengenizing lists ending tag. Then, the corpus has been processed for adding structural tagging (head, paragraph, sentence, list, etc.) and morphosyntactic information.

BASIC INFORMATION

Identification Information

- resourceName: GrAF version of Catalan portions of Wikipedia Corpus
- resourceShortName: GrAF version of the Catalan Wikipedia
- url: http://www.lsi.upc.edu/~nlp/wikicorpus/
- metaShareId: NOT_DEFINED_FOR_V2

- description:

This is the stand-off GrAF version of Catalan portions of the Wikipedia (based on a 2006 dump). The original Wikipedia Catalan Corpus contains about 120 million words in raw text format. It has been cleaned by erase disambiguation pages, removing some XML tags and homengenizing lists ending tag. Then, the corpus has been processed for adding structural tagging (head, paragraph, sentence, list, etc.) and morphosyntactic information.

ADMINISTRATIVE INFORMATION

Contact Person(s)

- surname: Boleda
- givenName: Gemma
- communicationInfo:
    - email: gemma.boleda@upf.edu
    - country: Spain
- affiliation:
    - organizationName: Universitat Politécnica de Catalunya
    - organizationShortName: UPC
    - communicationInfo:
        - email: gemma.boleda@upf.edu
        - url: http://www.upc.edu/
        - country: Spain

- surname: Vivaldi
- givenName: Jorge
- communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - address: Roc Boronat, 138
    - zipCode: 08018

    - city: Barcelona
    - country: Spain
- affiliation:
    - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
    - organizationShortName: IULA UPF
    - departmentName: Institut Universitari Lingüística Aplicada
    - communicationInfo:
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/
        - address: Roc Boronat, 138
        - zipCode: 08018
        - city: Barcelona
        - country: Spain
        - telephoneNumber: +34 93 5422332
        - faxNumber: +34 93 5422321

Distribution Information

- Availability type: available-unrestrictedUse
- Licence: GFDL

TECHNICAL INFORMATION

- Resource type: corpus
- Character Encoding: UTF-8
- Validation Information:
    - validated: true
    - validationType: formal
    - validationMode: automatic
    - validationModeDetails: The documents validates against the GrAF DTD v.1.0.4

CONTENT INFORMATION

- Linguality: monolingual
- Language(s): Catalan
- Size: XXXXXXXX bytes
- Annotation information:
    - annotationType: morphosyntacticAnnotation-posTagging
    - annotationStandoff: true
    - segmentationLevel: phrase
    - annotationFormat: GrAF
    - tagset: PAROLE

RELEVANT REFERENCES AND OTHER INFORMATION

Resource Documentation

- · documentUnstructured: Samuel Reese, Gemma Boleda, Montse Cuadros, Lluís Padró, German Rigau (2010). "Wikicorpus: A Word-Sense Disambiguated Multilingual Wikipedia Corpus". In Proceedings of 7th Language Resources and Evaluation Conference (LREC'10), La Valleta, Malta. May, 2010.

- · documentUnstructured: Samuel Reese. 2009. WikiNet: Construction d'une ressource lexico-sémantique multilingue à partir de Wikipedia. Master's thesis. ISAE (Institut Supérieur de l'Aéronautique et de l'Espace), Toulouse, France.

- · documentUnstructured: Web site of the Wikicorpus project: http://www.lsi.upc.edu/~nlp/wikicorpus/

Resource Creation Information

Resource Creator(s):

- organizationInfo:
  - organizationName: Research Group on Natural Language Processing - Grup de recerca consolidat de la Generalitat de Catalunya (2009 SGR-01723)
  - organizationShortName: NLP Research Group
  - communicationInfo:
    - email: castell@lsi.upc.edu
    - url: http://www.lsi.upc.edu/~nlp/
    - country: Spain

- personInfo:
  - surname: Boleda
  - givenName: Gemma
  - communicationInfo:
    - email: gemma.boleda@upf.edu

- organizationInfo:
  - organizationName: Universitat Pompeu Fabra - Institut Universitari Lingüística Aplicada
  - organizationShortName: IULA - UPF
  - departmentName: Institut Universitari Lingüística Aplicada (UPF)
  - communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - url: http://www.iula.upf.edu/
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - telephoneNumber: +34 93 5421207
    - faxNumber: +34 93 5422321

Resource Funding:

- projectName: KNOW: Developing large-scale multilingual technologies for language understanding
- projectShortName: KNOW
- url: http://ixa.si.ehu.es/know/
- fundingType: nationalFunds
- funder: Ministerio de educación y Ciencia. Gobierno de España (TIN2006-15049-C03)
- fundingCountry: Spain

- projectName: Language understanding technologies for multilingual domain-oriented information access
- projectShortName: KNOW2
- url: http://ixa.si.ehu.es/know2

- fundingType: nationalFunds
- funder: Ministerio de educación y Ciencia. Gobierno de España (TIN2009-14715-C04)
- fundingCountry: Spain

- projectName: PASCAL 2 Pattern Analysis, Statistical Modelling and Computational Learning
- projectShortName: Pascal Network of Excellence
- url: http://www.pascal-network.org/
- fundingType: euFunds
- funder: European Union (FP7-ICT-216886)

- projectName: Programa Juan de la Cierva
- projectShortName: Programa Juan de la Cierva
- url: http://www.idi.mineco.gob.es/portal/site/MICINN/menuitem.dbc68b34d11ccbd5d52ffeb801432ea0/?vgnextoid=f900759903236210VgnVCM1000001d04140aRCRD
- fundingType: nationalFunds
- funder: Ministerio de educación y Ciencia. Gobierno de España (JCI-2007-57-1479)
- fundingCountry: Spain

- projectName: METANET4U
- projectShortName: METANET4U
- url: http://metanet4u.eu/
- fundingType: ownFunds
- fundingType: euFunds
- funder: UPF
- funder: EU
- fundingCountry: Spain

Endogenous resources (tools)

Converters to LMF 2

This tool generates the LMF version of Apertium monolingual lexicons. The script takes as input an expanded monolingual Apertium lexicon (generated using: lt-expand apertium.dix > apertium.expanded) and generates the corresponding LMF version.

BASIC INFORMATION

Identification Information

- resourceName: IULA tagger Web Service
- resourceShortName: FreeLing2LMF converter
- url: http://www.uila.upf.edu
- metaShareId: NOT_DEFINED_FOR_V2

- description:

This tool generates the LMF version of Apertium monolingual lexicons. The script takes as input an expanded monolingual Apertium lexicon (generated using: lt-expand apertium.dix > apertium.expanded) and generates the corresponding LMF version.

In the Apertium expended lexicons, the first tag corresponds to the part of speech. The rest of tags (all enclosed in angle brackets) encode additional information depending on the lemma and PoS tag.

Run "perl ApertiumMonolingual2LMF.pl --help" to get more information

ADMINISTRATIVE INFORMATION

Contact Person(s)

- surname: Vivaldi
- givenName: Jorge
- communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
- affiliation:
    - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
    - organizationShortName: IULA UPF
    - departmentName: Institut Universitari Lingüística Aplicada
    - communicationInfo:
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/
        - address: Roc Boronat, 138
        - zipCode: 08018
        - city: Barcelona
        - country: Spain
        - telephoneNumber: +34 93 5421207
        - faxNumber: +34 93 5422321

Distribution Information

- Availability type: available-unrestrictedUse
- Licence: CC_BY-SA

TECHNICAL INFORMATION

- Tool Type: format conversion tool
- Language dependant: false
- WSDL code and information:
- Input Information:
    - mediaType: text
    - resourceType: lexicalConceptualResource
    - modalityType: writtenLanguage
    - mimeType: text / plain
    - characterEncoding: UTF-8

    - annotationType: morphosyntacticAnnotation-posTagging
    - annotationFormat: expanded Apertium monolingual lexicon format
    - conformanceToStandardsBestPractices: other
- Output Information:
    - mediaType: text
    - resourceType: lexicalConceptualResource
    - modalityType: writtenLanguage
    - mimeType: text / xml
    - characterEncoding: UTF-8
    - annotationType: morphosyntacticAnnotation-posTagging
    - annotationFormat: LMF
    - conformanceToStandardsBestPractices: LMF
- Operating System: os-independent
- Required Software: http://www.perl.org/
- Execution Information:

In the Apertium expended lexicons, the first tag corresponds to the part of speech. The rest of tags (all enclosed in angle brackets) encode additional information depending on the lemma and PoS tag.

Run "perl ApertiumMonolingual2LMF.pl --help" to get more information

- Validation Information:
    - validated: true
    - validationType: formal
    - validationMode: automatic
    - validationModeDetails: The output lexicon validates against the LMF DTD v.16

RELEVANT REFERENCES AND OTHER INFORMATION

Resource Documentation

· Resource Creation Information

Resource Creator(s):

- organizationInfo:
    - organizationName: Institut Universitari Lingüística Aplicada - UPF
    - organizationShortName: IULA - UPF
    - departmentName: Institut Universitari Lingüística Aplicada (UPF)
    - communicationInfo:
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/
        - address: Roc Boronat, 138
        - zipCode: 08018
        - city: Barcelona
        - country: Spain
        - telephoneNumber: +34 93 5421207
        - faxNumber: +34 93 5422321

Resource Funding:

- projectName: METANET4U
- projectShortName: METANET4U
- url: http://metanet4u.eu/
- fundingType: ownFunds
- fundingType: euFunds
- funder: UPF
- funder: EU
- fundingCountry: Spain

Tools for Catalan Corpus and Spanish Corpus Processing

Documentation for the 'IULA tokenizer Web Wervice'

This is a text tokenizer service developed and deployed as a SOAP Web Service by the IULA at the Universitat Pompeu Fabra. The text tokenizer requires that the input text be in plain text format (file.txt) and UTF-8 encoded.)POS tagger deployed as web service by the IULA at Universitat Pompeu Fabra. The input file must be in plain text format (file.txt) and UTF-8 encoded.

BASIC INFORMATION

Identification Information

- resourceName: IULA tokenizer Web Wervice
- resourceShortName: IULA tokenizer
- url: http://kurwenal.upf.edu/soaplab2-axis/#tokenization.iula_tokenizer_row
- metaShareId: NOT_DEFINED_FOR_V2

- description:

This is a text tokenizer service developed and deployed as a SOAP Web Service by the IULA at the Universitat Pompeu Fabra. The text tokenizer requires that the input text be in plain text format (file.txt) and UTF-8 encoded.)POS tagger deployed as web service by the IULA at Universitat Pompeu Fabra. The input file must be in plain text format (file.txt) and UTF-8 encoded.

Most IULA SOAP web services are deployed using Soaplab2. Soaplab is a tool that can automatically generate and deploy Web Services on top of existing command-line analysis programs.

Soaplab services allow for asynchronous services. However, for testing purposes, the 'runAndWaitFor' operation can be used. This operation starts a job, waits until it is completed and returns the job results.

As a general rule, IULA SOAP web services accept input data either as 'direct string' data or as URL. Output results are given both as 'direct string' data and as URL. For large outputs, the 'direct string' option is dissabled.

ADMINISTRATIVE INFORMATION

Contact Person(s)

- surname: Vivaldi
- givenName: Jorge
- communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
- affiliation:
    - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
    - organizationShortName: IULA UPF
    - departmentName: Institut Universitari Lingüística Aplicada
    - communicationInfo:
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/
        - address: Roc Boronat, 138
        - zipCode: 08018
        - city: Barcelona
        - country: Spain
        - telephoneNumber: +34 93 5421207
        - faxNumber: +34 93 5422321

Distribution Information

- Availability type: available-unrestrictedUse
- Licence: GPL

TECHNICAL INFORMATION

- Tool Type: Tokenizer service
- Language dependant: true
- WSDL code and information:

WSDL file: http://kurwenal.upf.edu:8080/soaplab2-axis/typed/services/tokenization.iula_tokenizer?wsdl

WSDL description file: http://kurwenal.upf.edu:8080/ws_docs/tokenization.iula_tokenizerwsdl.html

- Input Information:
    - mediaType: text
    - modalityType: writtenLanguage
    - languageName: Spanish
    - languageName: Catalan
    - languageName: English
    - mimeType: text / plain
    - characterEncoding: UTF-8
- Output Information:
    - mediaType: text
    - resourceType: corpus
    - modalityType: writtenLanguage
    - mimeType: text / xml
    - characterEncoding: UTF-8

    - annotationType: segmentation
- Operating System: os-independent
- Required Software:
- Execution Information:

Most IULA SOAP web services are deployed using Soaplab2. Soaplab is a tool that can automatically generate and deploy Web Services on top of existing command-line analysis programs.

Soaplab services allow for asynchronous services. However, for testing purposes, the 'runAndWaitFor' operation can be used. This operation starts a job, waits until it is completed and returns the job results.

As a general rule, IULA SOAP web services accept input data either as 'direct string' data or as URL. Output results are given both as 'direct string' data and as URL. For large outputs, the 'direct string' option is dissabled.

- Request message:
- <SOAP-ENV:Envelope xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
- xmlns:iu="http://soaplab.org/iula_tokenizer">
- <SOAP-ENV:Header/>
- <SOAP-ENV:Body>
- <iu:runAndWaitFor>
- <input_direct_data>Hola! esto es una prueba.</input_direct_data>
- <language>es</language>
- </iu:runAndWaitFor>
- </SOAP-ENV:Body>
- </SOAP-ENV:Envelope>

- Response message:
- <S:Envelope xmlns:S="http://schemas.xmlsoap.org/soap/envelope/">
- <S:Body>
- <ns3:runAndWaitForResponse xmlns:ns2="http://soaplab.org/typedws"
- xmlns:ns3="http://soaplab.org/iula_tokenizer">
- <report>Summary: Completed: Maybe Termination status: 0 Started: 2012-may-31 11:11:55
- (CEST) Ended: 2012-may-31 11:12:01 (CEST) Duration: 0:00:05.890 Report: Some error
- messages were reported. Name: tokenization.iula_tokenizer Job ID:
- [tokenization.iula_tokenizer]72a3c172.1372d38a3a0._7fbf Program and parameters:
- /usr/local/apache-tomcat-6.0.29_PRODUC/webapps/soaplab2-axis/WEB-INF/run/hector.sh
- -inputtext i_input -language es -annotationformat verticalized -outputtext o_output
- --- end of parameters Exit: 0 Standard error stream: Segmentado en frases
- BD-Diccionario: TreeTaggerDB:iula05v.upf.edu:3306 Reconocimiento de expresiones no
- analizables BD-Diccionario: TreeTaggerDB:iula05v.upf.edu:3306 Reconocimiento de
- expresiones no analizables math BD-Diccionario: TreeTaggerDB:iula05v.upf.edu:3306
- Reconocimiento de expresiones foráneas Reconocimiento de locuciones Reconocimiento
- de fechas Reconocimiento de números Reconocimiento de nombres propios
- BD-Diccionario: TreeTaggerDB:iula05v.upf.edu:3306 Reconocimiento de expresiones no
- analizables despues del preproceso Preproceso para 4 tokens realizado en 5 segundos </report>
- <detailed_status>0</detailed_status>
- <output><div1> <s> Hola ! DLD esto es una prueba . DLD </s>
- </div1></output>

- <output_url>http://kurwenal.upf.edu/soaplab2-axis/results/[tokenization.iula_tokenizer]72a3c172.1372d38a3a0._7fbf_output</output_url>
- </ns3:runAndWaitForResponse>
- </S:Body>
- </S:Envelope>

RELEVANT REFERENCES AND OTHER INFORMATION

Resource Documentation

- · documentUnstructured: WSDL file: http://kurwenal.upf.edu:8080/soaplab2-axis/typed/services/tokenization.iula_tokenizer?wsdl

- · documentUnstructured: WSDL description file: http://kurwenal.upf.edu:8080/ws_docs/tokenization.iula_tokenizerwsdl.html

Resource Creation Information

Resource Creator(s):

- organizationInfo:
  - organizationName: Institut Universitari Lingüística Aplicada - UPF
  - organizationShortName: IULA - UPF
  - departmentName: Institut Universitari Lingüística Aplicada (UPF)
  - communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - url: http://www.iula.upf.edu/
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - telephoneNumber: +34 93 5421207
    - faxNumber: +34 93 5422321

Resource Funding:

- projectName: METANET4U
- projectShortName: METANET4U
- url: http://metanet4u.eu/
- fundingType: ownFunds
- fundingType: euFunds
- funder: UPF
- funder: EU
- fundingCountry: Spain

Documentation for the 'GrAF IULA tagger Web Wervice'

POS tagger service deployed as a SOAP web service by the IULA at Universitat Pompeu Fabra. The input file must be in plain text format (file.txt) and UTF-8 encoded. The disambiguation process is done by a TreeTagger instance trained by the IULA. The output of the service is in stand-off style and follows the GrAF format.

BASIC INFORMATION

Identification Information

- resourceName: GrAF IULA tagger Web Wervice
- resourceShortName: IULA GraF tagger
- url: http://kurwenal.upf.edu/soaplab2-axis/#morphosintactic_tagging.iula_tagger_graf_row
- metaShareId: NOT_DEFINED_FOR_V2

- description:

POS tagger service deployed as a SOAP web service by the IULA at Universitat Pompeu Fabra. The input file must be in plain text format (file.txt) and UTF-8 encoded. The disambiguation process is done by a TreeTagger instance trained by the IULA. The output of the service is in stand-off style and follows the GrAF format.

Most IULA SOAP web services are deployed using Soaplab2. Soaplab is a tool that can automatically generate and deploy Web Services on top of existing command-line analysis programs.

Soaplab services allow for asynchronous services. However, for testing purposes, the 'runAndWaitFor' operation can be used. This operation starts a job, waits until it is completed and returns the job results.

As a general rule, IULA SOAP web services accept input data either as 'direct string' data or as URL. Output results are given both as 'direct string' data and as URL. For large outputs, the 'direct string' option is dissabled.

ADMINISTRATIVE INFORMATION

Contact Person(s)

- surname: Vivaldi
- givenName: Jorge
- communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
- affiliation:
    - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
    - organizationShortName: IULA UPF
    - departmentName: Institut Universitari Lingüística Aplicada
    - communicationInfo:
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/
        - address: Roc Boronat, 138
        - zipCode: 08018
        - city: Barcelona
        - country: Spain

        - telephoneNumber: +34 93 5421207
        - faxNumber: +34 93 5422321

Distribution Information

- Availability type: available-unrestrictedUse
- Licence: GPL

TECHNICAL INFORMATION

- Tool Type: Part of speech annotation tool
- Language dependant: true
- WSDL code and information:

WSDL file: http://kurwenal.upf.edu:8080/soaplab2-axis/typed/services/morphosintactic_tagging.iula_tagger_graf?wsdl

WSDL description file: http://kurwenal.upf.edu:8080/ws_docs/morphosintactic_tagging.iula_tagger_grafwsdl.html

- Input Information:
    - mediaType: text
    - modalityType: writtenLanguage
    - languageName: Spanish
    - languageName: Catalan
    - mimeType: text / plain
    - characterEncoding: UTF-8
- Output Information:
    - mediaType: text
    - resourceType: corpus
    - modalityType: writtenLanguage
    - mimeType: text / xml
    - characterEncoding: UTF-8
    - annotationType: morphosyntacticAnnotation-posTagging
    - annotationFormat: stand-off
    - conformanceToStandardsBestPractices: GrAF
- Operating System: os-independent
- Required Software:
- Execution Information:

Most IULA SOAP web services are deployed using Soaplab2. Soaplab is a tool that can automatically generate and deploy Web Services on top of existing command-line analysis programs.

Soaplab services allow for asynchronous services. However, for testing purposes, the 'runAndWaitFor' operation can be used. This operation starts a job, waits until it is completed and returns the job results.

As a general rule, IULA SOAP web services accept input data either as 'direct string' data or as URL. Output results are given both as 'direct string' data and as URL. For large outputs, the 'direct string' option is dissabled.

- Request message:
- <SOAP-ENV:Envelope xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
- xmlns:iu="http://soaplab.org/iula_tagger_graf">

- <SOAP-ENV:Header/>
- <SOAP-ENV:Body>
- <iu:runAndWaitFor>
- <input_direct_data>Hola esto es una prueba!</input_direct_data>
- <language>es</language>
-
- </iu:runAndWaitFor>
- </SOAP-ENV:Body>
- </SOAP-ENV:Envelope>

- Response message:
- <S:Envelope xmlns:S="http://schemas.xmlsoap.org/soap/envelope/">
- <S:Body>
- <ns3:runAndWaitForResponse xmlns:ns2="http://soaplab.org/typedws"
- xmlns:ns3="http://soaplab.org/iula_tagger_graf">
- <report>Summary: Completed: Maybe Termination status: 0 Started: 2012-may-17 17:40:39
- (CEST) Ended: 2012-may-17 17:40:43 (CEST) Duration: 0:00:03.212 Report: Some error
- messages were reported. Name: morphosintactic_tagging.iula_tagger_graf Job ID:
- [morphosintactic_tagging.iula_tagger_graf]72a3c172.1372d38a3a0._7fd3 Program and
- parameters:
- /usr/local/apache-tomcat-6.0.29_PRODUC/webapps/soaplab2-axis/WEB-INF/run/hector.sh
- -inputtext i_input -language es -annotationformat graph -opsent sent -oanc
- o_headerxml -opos o_posxml -osent o_sentxml -oseg o_segxml --- end of parameters
- Exit: 0 Standard error stream: Segmentado en frases BD-Diccionario:
- TreeTaggerDB:iula05v.upf.edu:3306 Reconocimiento de expresiones no analizables
- BD-Diccionario: TreeTaggerDB:iula05v.upf.edu:3306 Reconocimiento de expresiones no
- analizables math BD-Diccionario: TreeTaggerDB:iula05v.upf.edu:3306 Reconocimiento de
- expresiones foráneas Reconocimiento de locuciones Reconocimiento de fechas
- Reconocimiento de números Reconocimiento de nombres propios BD-Diccionario:
- TreeTaggerDB:iula05v.upf.edu:3306 Reconocimiento de expresiones no analizables
- despues del preproceso Preproceso para 4 tokens realizado en 2 segundos </report>
- <detailed_status>0</detailed_status>
- <header>empty_output2. Size limit! use URL.</header>
- <header_url>http://kurwenal.upf.edu/soaplab2-
axis/results/[morphosintactic_tagging.iula_tagger_graf]72a3c172.1372d38a3a0._7fd3_header.xml</head
er_url>
- <pos>empty_output2. Size limit! use URL.</pos>
- <pos_url>http://kurwenal.upf.edu/soaplab2-
axis/results/[morphosintactic_tagging.iula_tagger_graf]72a3c172.1372d38a3a0._7fd3_pos.xml</pos_url>
- <sent><?xml version="1.0" encoding="UTF-8"?> <graph
- xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
- xsi:schemaLocation="http://www.xces.org/ns/GrAF/0.99/ graf-0.99.xsd"
- xmlns="http://www.xces.org/ns/GrAF/0.99/"> <header> <tagsDecl> <tagUsage
- gi="s" occurs="1"/> </tagsDecl> <annotationSets><annotationSet name="xces"
- type="http://www.xces.org/schema/2003"/> </annotationSets> </header>
- <region xml:id="head-r1" anchors="0 24"/> <node xml:id="head-n1"> <link
- targets="head-r1"/> </node> <fs>
- <f name="id" value="div10-head1"/> </fs> </graph></sent>

- <sent_url>http://kurwenal.upf.edu/soaplab2-axis/results/[morphosintactic_tagging.iula_tagger_graf]72a3c172.1372d38a3a0._7fd3_sent.xml</sent_url>
- <seg><?xml version="1.0" encoding="UTF-8"?> <graph
- xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
- xsi:schemaLocation="http://www.xces.org/ns/GrAF/0.99/ graf-0.99.xsd"
- xmlns="http://www.xces.org/ns/GrAF/0.99/"> <header> <tagsDecl> </tagsDecl>
- </header> <region xml:id="seg-r0" anchors="0 4"/> <region xml:id="seg-r1"
- anchors="5 9"/> <region xml:id="seg-r2" anchors="10 12"/> <region
- xml:id="seg-r3" anchors="13 16"/> <region xml:id="seg-r4" anchors="17 23"/>
- <region xml:id="seg-r5" anchors="23 24"/> </graph></seg>
- <seg_url>http://kurwenal.upf.edu/soaplab2-axis/results/[morphosintactic_tagging.iula_tagger_graf]72a3c172.1372d38a3a0._7fd3_seg.xml</seg_url>
- </ns3:runAndWaitForResponse>
- </S:Body>
- </S:Envelope>

RELEVANT REFERENCES AND OTHER INFORMATION

Resource Documentation

- · documentUnstructured: WSDL file: http://kurwenal.upf.edu:8080/soaplab2-axis/typed/services/morphosintactic_tagging.iula_tagger_graf?wsdl

- · documentUnstructured: WSDL description file: http://kurwenal.upf.edu:8080/ws_docs/morphosintactic_tagging.iula_tagger_grafwsdl.html

Resource Creation Information

Resource Creator(s):

- organizationInfo:
  - organizationName: Institut Universitari Lingüística Aplicada - UPF
  - organizationShortName: IULA - UPF
  - departmentName: Institut Universitari Lingüística Aplicada (UPF)
  - communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - url: http://www.iula.upf.edu/
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - telephoneNumber: +34 93 5421207
    - faxNumber: +34 93 5422321

Resource Funding:

- projectName: METANET4U
- projectShortName: METANET4U

- url: http://metanet4u.eu/
- fundingType: ownFunds
- fundingType: euFunds
- funder: UPF
- funder: EU
- fundingCountry: Spain

Documentation for the 'IULA tagger Web Wervice'

POS tagger service deployed as a SOAP web service by the IULA at Universitat Pompeu Fabra. The input file must be in plain text format (file.txt) and UTF-8 encoded. The disambiguation process is done by a TreeTagger instance trained by the IULA.

BASIC INFORMATION

Identification Information

- resourceName: IULA tagger Web Wervice
- resourceShortName: IULA tagger
- url: http://kurwenal.upf.edu/soaplab2-axis/#morphosintactic_tagging.iula_tagger_row
- metaShareId: NOT_DEFINED_FOR_V2

- description:

POS tagger service deployed as a SOAP web service by the IULA at Universitat Pompeu Fabra. The input file must be in plain text format (file.txt) and UTF-8 encoded. The disambiguation process is done by a TreeTagger instance trained by the IULA.

Most IULA SOAP web services are deployed using Soaplab2. Soaplab is a tool that can automatically generate and deploy Web Services on top of existing command-line analysis programs.

Soaplab services allow for asynchronous services. However, for testing purposes, the 'runAndWaitFor' operation can be used. This operation starts a job, waits until it is completed and returns the job results.

As a general rule, IULA SOAP web services accept input data either as 'direct string' data or as URL. Output results are given both as 'direct string' data and as URL. For large outputs, the 'direct string' option is dissabled.

ADMINISTRATIVE INFORMATION

Contact Person(s)

- surname: Vivaldi
- givenName: Jorge
- communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona

    - country: Spain
- affiliation:
    - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
    - organizationShortName: IULA UPF
    - departmentName: Institut Universitari Lingüística Aplicada
    - communicationInfo:
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/
        - address: Roc Boronat, 138
        - zipCode: 08018
        - city: Barcelona
        - country: Spain
        - telephoneNumber: +34 93 5421207
        - faxNumber: +34 93 5422321

Distribution Information

- Availability type: available-unrestrictedUse
- Licence: GPL

TECHNICAL INFORMATION

- Tool Type: annotation tool
- Language dependant: true
- WSDL code and information:

WSDL file: http://kurwenal.upf.edu:8080/soaplab2-axis/typed/services/morphosintactic_tagging.iula_tagger?wsdl

WSDL description file: http://kurwenal.upf.edu:8080/ws_docs/morphosintactic_tagging.iula_taggerwsdl.html

- Input Information:
    - mediaType: text
    - modalityType: writtenLanguage
    - languageName: Spanish
    - languageName: Catalan
    - mimeType: text / plain
    - characterEncoding: UTF-8
- Output Information:
    - mediaType: text
    - resourceType: lexicalConceptualResource
    - modalityType: writtenLanguage
    - mimeType: text / xml
    - characterEncoding: UTF-8
    - annotationType: morphosyntacticAnnotation-posTagging
- Operating System: os-independent
- Required Software:
- Execution Information:

Most IULA SOAP web services are deployed using Soaplab2. Soaplab is a tool that can automatically generate and deploy Web Services on top of existing command-line analysis programs.

Soaplab services allow for asynchronous services. However, for testing purposes, the 'runAndWaitFor' operation can be used. This operation starts a job, waits until it is completed and returns the job results.

As a general rule, IULA SOAP web services accept input data either as 'direct string' data or as URL. Output results are given both as 'direct string' data and as URL. For large outputs, the 'direct string' option is dissabled.

- Request message:
- <SOAP-ENV:Envelope xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
- xmlns:iu="http://soaplab.org/iula_tagger">
- <SOAP-ENV:Header/>
- <SOAP-ENV:Body>
- <iu:runAndWaitFor>
- <input_direct_data>Hola esto es una prueba!</input_direct_data>
- <language>es</language>
- <output_format/>
- </iu:runAndWaitFor>
- </SOAP-ENV:Body>
- </SOAP-ENV:Envelope>

- Response message:
- <S:Envelope xmlns:S="http://schemas.xmlsoap.org/soap/envelope/">
- <S:Body>
- <ns3:runAndWaitForResponse xmlns:ns2="http://soaplab.org/typedws"
- xmlns:ns3="http://soaplab.org/iula_tagger">
- <report>Summary: Completed: Maybe Termination status: 0 Started: 2012-may-31 11:15:48
- (CEST) Ended: 2012-may-31 11:15:56 (CEST) Duration: 0:00:07.399 Report: Some error
- messages were reported. Name: morphosintactic_tagging.iula_tagger Job ID:
- [morphosintactic_tagging.iula_tagger]72a3c172.1372d38a3a0._7fbe Program and
- parameters:
- /usr/local/apache-tomcat-6.0.29_PRODUC/webapps/soaplab2-axis/WEB-INF/run/hector.sh
- -inputtext i_input -language es -annotationformat treetagger -outputtext o_output
- --- end of parameters Exit: 0 Standard error stream: reading parameters ...
- Segmentado en frases BD-Diccionario: TreeTaggerDB:iula05v.upf.edu:3306
- Reconocimiento de expresiones no analizables BD-Diccionario:
- TreeTaggerDB:iula05v.upf.edu:3306 Reconocimiento de expresiones no analizables math
- BD-Diccionario: TreeTaggerDB:iula05v.upf.edu:3306 Reconocimiento de expresiones
- foráneas Reconocimiento de locuciones Reconocimiento de fechas Reconocimiento de
- números tagging ... Reconocimiento de nombres propios BD-Diccionario:
- TreeTaggerDB:iula05v.upf.edu:3306 Reconocimiento de expresiones no analizables
- despues del preproceso Preproceso para 4 tokens realizado en 5 segundos finished. </report>
- <detailed_status>0</detailed_status>
- <output><div1> <head> Hola I hola esto RD---NS éste es VDR3S- ser una E6--FS uno
- prueba N5-FS prueba ! DLD ! </head> </div1> </output>
- <output_url>http://kurwenal.upf.edu/soaplab2-
axis/results/[morphosintactic_tagging.iula_tagger]72a3c172.1372d38a3a0._7fbe_output</output_url>
- </ns3:runAndWaitForResponse>
- </S:Body>
- </S:Envelope>

RELEVANT REFERENCES AND OTHER INFORMATION

Resource Documentation

- · documentUnstructured: WSDL file: http://kurwenal.upf.edu:8080/soaplab2-axis/typed/services/morphosintactic_tagging.iula_tagger?wsdl

- · documentUnstructured: WSDL description file: http://kurwenal.upf.edu:8080/ws_docs/morphosintactic_tagging.iula_taggerwsdl.html

Resource Creation Information

Resource Creator(s):

- organizationInfo:
    - organizationName: Institut Universitari Lingüística Aplicada - UPF
    - organizationShortName: IULA - UPF
    - departmentName: Institut Universitari Lingüística Aplicada (UPF)
    - communicationInfo:
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/
        - address: Roc Boronat, 138
        - zipCode: 08018
        - city: Barcelona
        - country: Spain
        - telephoneNumber: +34 93 5421207
        - faxNumber: +34 93 5422321

Resource Funding:

- projectName: METANET4U
- projectShortName: METANET4U
- url: http://metanet4u.eu/
- fundingType: ownFunds
- fundingType: euFunds
- funder: UPF
- funder: EU
- fundingCountry: Spain

Documentation for the 'IULA pre-process web service'

Text preprocess SOAP web service developed and deployed by the IULA at the Universitat Pompeu Fabra. This preprocess service requires that the input text be in plain text format (file .txt) and UTF-8. Basically, it carries out: (i) text segmentation into minor structural units (titles, paragraphs, sentences, etc.); (ii) detection of entities not found in dictionaries (numbers, abbreviations, URLs, emails, proper nouns, etc.); and (iii) the keeping of sequences of two or more words in a single block (dates, phrases, proper nouns, etc.).

BASIC INFORMATION

Identification Information

- resourceName: IULA pre-process web service
- resourceShortName: IULA pre-process
- url: http://kurwenal.upf.edu/soaplab2-axis/#chunking_segmentation.iula_preprocess_row
- metaShareId: NOT_DEFINED_FOR_V2

- description:

Text preprocess SOAP web service developed and deployed by the IULA at the Universitat Pompeu Fabra. This preprocess service requires that the input text be in plain text format (file .txt) and UTF-8. Basically, it carries out: (i) text segmentation into minor structural units (titles, paragraphs, sentences, etc.); (ii) detection of entities not found in dictionaries (numbers, abbreviations, URLs, emails, proper nouns, etc.); and (iii) the keeping of sequences of two or more words in a single block (dates, phrases, proper nouns, etc.).

Most IULA SOAP web services are deployed using Soaplab2. Soaplab is a tool that can automatically generate and deploy Web Services on top of existing command-line analysis programs.

Soaplab services allow for asynchronous services. However, for testing purposes, the 'runAndWaitFor' operation can be used. This operation starts a job, waits until it is completed and returns the job results.

As a general rule, IULA SOAP web services accept input data either as 'direct string' data or as URL. Output results are given both as 'direct string' data and as URL. For large outputs, the 'direct string' option is dissabled.

ADMINISTRATIVE INFORMATION

Contact Person(s)

- surname: Vivaldi
- givenName: Jorge
- communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
- affiliation:
    - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
    - organizationShortName: IULA UPF
    - departmentName: Institut Universitari Lingüística Aplicada
    - communicationInfo:
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/
        - address: Roc Boronat, 138
        - zipCode: 08018
        - city: Barcelona
        - country: Spain
        - telephoneNumber: +34 93 5421207
        - faxNumber: +34 93 5422321

Distribution Information

- Availability type: available-unrestrictedUse
- Licence: GPL

TECHNICAL INFORMATION

- Tool Type: pre-processing tool
- Language dependant: true
- WSDL code and information:

WSDL file: http://kurwenal.upf.edu:8080/soaplab2-axis/typed/services/chunking_segmentation.iula_preprocess?wsdl

WSDL description file: http://kurwenal.upf.edu:8080/ws_docs/chunking_segmentation.iula_preprocesswsdl.html

- Input Information:
    - mediaType: text
    - modalityType: writtenLanguage
    - languageName: Spanish
    - languageName: Catalan
    - mimeType: text / plain
    - characterEncoding: UTF-8
- Output Information:
    - mediaType: text
    - resourceType: lexicalConceptualResource
    - modalityType: writtenLanguage
    - mimeType: text / xml
    - characterEncoding: UTF-8
    - annotationType: segmentation
- Operating System: os-independent
- Required Software:
- Execution Information:

Most IULA SOAP web services are deployed using Soaplab2. Soaplab is a tool that can automatically generate and deploy Web Services on top of existing command-line analysis programs.

Soaplab services allow for asynchronous services. However, for testing purposes, the 'runAndWaitFor' operation can be used. This operation starts a job, waits until it is completed and returns the job results.

As a general rule, IULA SOAP web services accept input data either as 'direct string' data or as URL. Output results are given both as 'direct string' data and as URL. For large outputs, the 'direct string' option is dissabled.

- Request message:
- <SOAP-ENV:Envelope xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
- xmlns:iu="http://soaplab.org/iula_preprocess">
- <SOAP-ENV:Header/>
- <SOAP-ENV:Body>
- <iu:runAndWaitFor>
- <input_direct_data>Hola esto es una prueba! Hoy es jueves 17 de mayo del 2012</input_direct_data>

- <language>es</language>
- <encoding>utf-8</encoding>
- </iu:runAndWaitFor>
- </SOAP-ENV:Body>
- </SOAP-ENV:Envelope>

- Response message:
- <S:Envelope xmlns:S="http://schemas.xmlsoap.org/soap/envelope/">
- <S:Body>
- <ns3:runAndWaitForResponse xmlns:ns2="http://soaplab.org/typedws"
- xmlns:ns3="http://soaplab.org/iula_preprocess">
- <report>Summary: Completed: Maybe Termination status: 0 Started: 2012-may-17 10:44:36
- (CEST) Ended: 2012-may-17 10:44:41 (CEST) Duration: 0:00:04.453 Report: Some error
- messages were reported. Name: chunking_segmentation.iula_preprocess Job ID:
- [chunking_segmentation.iula_preprocess]72a3c172.1372d38a3a0._7fe9 Program and
- parameters:
- /usr/local/apache-tomcat-6.0.29_PRODUC/webapps/soaplab2-axis/WEB-INF/run/hector.sh
- -inputtext i_input -language es -annotationformat xmltag -inputmimetype utf-8
- -outputtext o_output --- end of parameters Exit: 0 Standard error stream: Segmentado
- en frases BD-Diccionario: TreeTaggerDB:iula05v.upf.edu:3306 Reconocimiento de
- expresiones no analizables BD-Diccionario: TreeTaggerDB:iula05v.upf.edu:3306
- Reconocimiento de expresiones no analizables math BD-Diccionario:
- TreeTaggerDB:iula05v.upf.edu:3306 Reconocimiento de expresiones foráneas
- Reconocimiento de locuciones Reconocimiento de fechas Reconocimiento de números
- Reconocimiento de nombres propios BD-Diccionario: TreeTaggerDB:iula05v.upf.edu:3306
- Reconocimiento de expresiones no analizables despues del preproceso Preproceso para
- 12 tokens realizado en 4 segundos </report>
- <detailed_status>0</detailed_status>
- <output><div1> <s>Hola esto es una prueba!</s><s>Hoy es jueves
- <date ISO8601="2012-05-17">17 de mayo del 2012</date></s>
- </div1></output>
- <output_url>http://kurwenal.upf.edu/soaplab2-
axis/results/[chunking_segmentation.iula_preprocess]72a3c172.1372d38a3a0._7fe9_output</output_url
>
- </ns3:runAndWaitForResponse>
- </S:Body>

RELEVANT REFERENCES AND OTHER INFORMATION

Resource Documentation

- · documentUnstructured: WSDL file: http://kurwenal.upf.edu:8080/soaplab2-
axis/typed/services/chunking_segmentation.iula_preprocess?wsdl

- · documentUnstructured: WSDL description file:
http://kurwenal.upf.edu:8080/ws_docs/chunking_segmentation.iula_preprocesswsdl.html

Resource Creation Information

Resource Creator(s):

- organizationInfo:
  - organizationName: Institut Universitari Lingüística Aplicada - UPF
  - organizationShortName: IULA - UPF
  - departmentName: Institut Universitari Lingüística Aplicada (UPF)
  - communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - url: http://www.iula.upf.edu/
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - telephoneNumber: +34 93 5421207
    - faxNumber: +34 93 5422321

Resource Funding:

- projectName: METANET4U
- projectShortName: METANET4U
- url: http://metanet4u.eu/
- fundingType: ownFunds
- fundingType: euFunds
- funder: UPF
- funder: EU
- fundingCountry: Spain

Documentation for the 'IULA paradigma Web Wervice'

The paradigma service was developed and deployed as a SOAP Web Service by the IULA at the Universitat Pompeu Fabra. Given a verb (infinitive or a verbal form) the service outputs its verbal paradigm grouped according tense and mode.

BASIC INFORMATION

Identification Information

- resourceName: IULA paradigma Web Wervice
- resourceShortName: IULA paradigma
- url: http://kurwenal.upf.edu/soaplab2-axis/#verbal_conjugation.iula_paradigma_row
- metaShareId: NOT_DEFINED_FOR_V2

- description:

The paradigma service was developed and deployed as a SOAP Web Service by the IULA at the Universitat Pompeu Fabra. Given a verb (infinitive or a verbal form) the service outputs its verbal paradigm grouped according tense and mode.

Most IULA SOAP web services are deployed using Soaplab2. Soaplab is a tool that can automatically generate and deploy Web Services on top of existing command-line analysis programs.

Soaplab services allow for asynchronous services. However, for testing purposes, the 'runAndWaitFor' operation can be used. This operation starts a job, waits until it is completed and returns the job results.

As a general rule, IULA SOAP web services accept input data either as 'direct string' data or as URL. Output results are given both as 'direct string' data and as URL. For large outputs, the 'direct string' option is dissabled.

ADMINISTRATIVE INFORMATION

Contact Person(s)

- surname: Vivaldi
- givenName: Jorge
- communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
- affiliation:
    - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
    - organizationShortName: IULA UPF
    - departmentName: Institut Universitari Lingüística Aplicada
    - communicationInfo:
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/
        - address: Roc Boronat, 138
        - zipCode: 08018
        - city: Barcelona
        - country: Spain
        - telephoneNumber: +34 93 5421207
        - faxNumber: +34 93 5422321

Distribution Information

- Availability type: available-unrestrictedUse
- Licence: GPL

TECHNICAL INFORMATION

- Tool Type: Lexicon look up service
- Language dependant: true
- WSDL code and information:

WSDL file: http://kurwenal.upf.edu/soaplab2-axis/services/verbal_conjugation.iula_paradigma?wsdl

WSDL description file: http://kurwenal.upf.edu:8080/ws_docs/verbal_conjugation.iula_paradigmawsdl.html

- Input Information:
  - mediaType: text
  - modalityType: writtenLanguage
  - languageName: Spanish
  - languageName: Catalan
  - languageName: English
  - mimeType: text / plain
  - characterEncoding: UTF-8
- Output Information:
  - mediaType: text
  - modalityType: writtenLanguage
  - mimeType: text / xml
  - characterEncoding: UTF-8
  - annotationType: morphosyntacticAnnotation-posTagging
- Operating System: os-independent
- Required Software:
- Execution Information:

Most IULA SOAP web services are deployed using Soaplab2. Soaplab is a tool that can automatically generate and deploy Web Services on top of existing command-line analysis programs.

Soaplab services allow for asynchronous services. However, for testing purposes, the 'runAndWaitFor' operation can be used. This operation starts a job, waits until it is completed and returns the job results.

As a general rule, IULA SOAP web services accept input data either as 'direct string' data or as URL. Output results are given both as 'direct string' data and as URL. For large outputs, the 'direct string' option is dissabled.

- Request message:
- <SOAP-ENV:Envelope xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
- xmlns:iu="http://soaplab.org/iula_paradigma">
- <SOAP-ENV:Header/>
- <SOAP-ENV:Body>
- <iu:runAndWaitFor>
- <form>comeremos</form>
- <language>es</language>
- </iu:runAndWaitFor>
- </SOAP-ENV:Body>
- </SOAP-ENV:Envelope>

- Response message:
- <S:Envelope xmlns:S="http://schemas.xmlsoap.org/soap/envelope/">
- <S:Body>
- <ns3:runAndWaitForResponse xmlns:ns2="http://soaplab.org/typedws"
- xmlns:ns3="http://soaplab.org/iula_paradigma">
- <report>Summary: Completed: Successfully Termination status: 0 Started: 2012-may-17
- 17:48:02 (CEST) Ended: 2012-may-17 17:48:03 (CEST) Duration: 0:00:00.718 Report:
- Name: verbal_conjugation.iula_paradigma Job ID:

- [verbal_conjugation.iula_paradigma]72a3c172.1372d38a3a0._7fcd Program and
- parameters:
- /usr/local/apache-tomcat-6.0.29_PRODUC/webapps/soaplab2-axis/WEB-INF/run/paradigma.sh
- -f comeremos -l es --- end of parameters Exit: 0 </report>
- <detailed_status>0</detailed_status>
- <output>empty_output2. Size limit! use URL.</output>
- <output_url>http://kurwenal.upf.edu/soaplab2-axis/results/[verbal_conjugation.iula_paradigma]72a3c172.1372d38a3a0._7fcd_output</output_url>
- </ns3:runAndWaitForResponse>
- </S:Body>
- </S:Envelope>

RELEVANT REFERENCES AND OTHER INFORMATION

Resource Documentation

- · documentUnstructured: WSDL file: http://kurwenal.upf.edu/soaplab2-axis/services/verbal_conjugation.iula_paradigma?wsdl

- · documentUnstructured: WSDL description file: http://kurwenal.upf.edu:8080/ws_docs/verbal_conjugation.iula_paradigmawsdl.html

Resource Creation Information

Resource Creator(s):

- organizationInfo:
  - organizationName: Institut Universitari Lingüística Aplicada - UPF
  - organizationShortName: IULA - UPF
  - departmentName: Institut Universitari Lingüística Aplicada (UPF)
  - communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - url: http://www.iula.upf.edu/
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - telephoneNumber: +34 93 5421207
    - faxNumber: +34 93 5422321

Resource Funding:

- projectName: METANET4U
- projectShortName: METANET4U
- url: http://metanet4u.eu/
- fundingType: ownFunds
- fundingType: euFunds
- funder: UPF

- funder: EU
- fundingCountry: Spain

Documentation for the 'IULA lexicon look up Web Wervice'

The lexicon look up service was developed and deployed as a SOAP Web Service by the IULA at the Universitat Pompeu Fabra. Given a word form, the webservice returns the information in the IULA's lexicon.

BASIC INFORMATION

Identification Information

- resourceName: IULA lexicon look up Web Wervice
- resourceShortName: IULA lexicon lookup
- url: http://kurwenal.upf.edu/soaplab2-axis/#stemming_lemmatization.iula_lexicon_lookup_row
- metaShareId: NOT_DEFINED_FOR_V2

- description:

The lexicon look up service was developed and deployed as a SOAP Web Service by the IULA at the Universitat Pompeu Fabra. Given a word form, the webservice returns the information in the IULA's lexicon.

Most IULA SOAP web services are deployed using Soaplab2. Soaplab is a tool that can automatically generate and deploy Web Services on top of existing command-line analysis programs.

Soaplab services allow for asynchronous services. However, for testing purposes, the 'runAndWaitFor' operation can be used. This operation starts a job, waits until it is completed and returns the job results.

As a general rule, IULA SOAP web services accept input data either as 'direct string' data or as URL. Output results are given both as 'direct string' data and as URL. For large outputs, the 'direct string' option is dissabled.

ADMINISTRATIVE INFORMATION

Contact Person(s)

- surname: Vivaldi
- givenName: Jorge
- communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
- affiliation:
    - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
    - organizationShortName: IULA UPF
    - departmentName: Institut Universitari Lingüística Aplicada
    - communicationInfo:

- email: jorge.vivaldi@upf.edu
- url: http://www.iula.upf.edu/
- address: Roc Boronat, 138
- zipCode: 08018
- city: Barcelona
- country: Spain
- telephoneNumber: +34 93 5421207
- faxNumber: +34 93 5422321

Distribution Information

- Availability type: available-unrestrictedUse
- Licence: GPL

TECHNICAL INFORMATION

- Tool Type: Lexicon look up service
- Language dependant: true
- WSDL code and information:

WSDL file: http://kurwenal.upf.edu:8080/soaplab2-axis/typed/services/stemming_lemmatization.iula_lexicon_lookup?wsdl

WSDL description file in html for human reading: http://kurwenal.upf.edu:8080/ws_docs/stemming_lemmatization.iula_lexicon_lookupwsdl.html

- Input Information:
  - mediaType: text
  - modalityType: writtenLanguage
  - languageName: Spanish
  - languageName: Catalan
  - languageName: English
  - mimeType: text / plain
  - characterEncoding: UTF-8
- Output Information:
  - mediaType: text
  - modalityType: writtenLanguage
  - mimeType: text / xml
  - characterEncoding: UTF-8
  - annotationType: morphosyntacticAnnotation-posTagging
- Operating System: os-independent
- Required Software:
- Execution Information:

Most IULA SOAP web services are deployed using Soaplab2. Soaplab is a tool that can automatically generate and deploy Web Services on top of existing command-line analysis programs.

Soaplab services allow for asynchronous services. However, for testing purposes, the 'runAndWaitFor' operation can be used. This operation starts a job, waits until it is completed and returns the job results.

As a general rule, IULA SOAP web services accept input data either as 'direct string' data or as URL. Output results are given both as 'direct string' data and as URL. For large outputs, the 'direct string' option is dissabled.

- Request message:
- <SOAP-ENV:Envelope xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
- xmlns:iu="http://soaplab.org/iula_lexicon_lookup">
- <SOAP-ENV:Header/>
- <SOAP-ENV:Body>
- <iu:runAndWaitFor>
- <language>ES</language>
- <wordForm>comeremos</wordForm>
- </iu:runAndWaitFor>
- </SOAP-ENV:Body>
- </SOAP-ENV:Envelope>

- Response message:
- <S:Envelope xmlns:S="http://schemas.xmlsoap.org/soap/envelope/">
- <S:Body>
- <ns3:runAndWaitForResponse xmlns:ns2="http://soaplab.org/typedws"
- xmlns:ns3="http://soaplab.org/iula_lexicon_lookup">
- <report>Summary: Completed: Successfully Termination status: 0 Started: 2012-may-17
- 17:44:43 (CEST) Ended: 2012-may-17 17:44:44 (CEST) Duration: 0:00:00.363 Report:
- Name: stemming_lemmatization.iula_lexicon_lookup Job ID:
- [stemming_lemmatization.iula_lexicon_lookup]72a3c172.1372d38a3a0._7fcf Program and
- parameters:
- /usr/local/apache-tomcat-6.0.29_PRODUC/webapps/soaplab2-axis/WEB-INF/run/lexiconlookup.sh
- -language ES -wordForm comeremos --- end of parameters Exit: 0 </report>
- <detailed_status>0</detailed_status>
- <output><LexicalEntriesList><lexicalEntry><lemma>comer</lemma><PoSTag>VDU1P-</PoSTag></lexicalEntry></LexicalEntriesList></output>
- <output_url>http://kurwenal.upf.edu/soaplab2-axis/results/[stemming_lemmatization.iula_lexicon_lookup]72a3c172.1372d38a3a0._7fcf_output</output_url>
- </ns3:runAndWaitForResponse>
- </S:Body>
- </S:Envelope>

RELEVANT REFERENCES AND OTHER INFORMATION

Resource Documentation

- · documentUnstructured: WSDL file: http://kurwenal.upf.edu:8080/soaplab2-axis/typed/services/stemming_lemmatization.iula_lexicon_lookup?wsdl

- · documentUnstructured: WSDL description file in html for human reading: http://kurwenal.upf.edu:8080/ws_docs/stemming_lemmatization.iula_lexicon_lookupwsdl.html

Resource Creation Information

Resource Creator(s):

- organizationInfo:
  - organizationName: Institut Universitari Lingüística Aplicada - UPF
  - organizationShortName: IULA - UPF
  - departmentName: Institut Universitari Lingüística Aplicada (UPF)
  - communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - url: http://www.iula.upf.edu/
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - telephoneNumber: +34 93 5421207
    - faxNumber: +34 93 5422321

Resource Funding:

- projectName: METANET4U
- projectShortName: METANET4U
- url: http://metanet4u.eu/
- fundingType: ownFunds
- fundingType: euFunds
- funder: UPF
- funder: EU
- fundingCountry: Spain

Extra Resources (new or delivered now from Batch 3)

Apertium Portuguese dictionary in LMF

This is the LMF version of the Apertium Portuguese dictionary. Monolingual dictionary for Portuguese was generated from the Apertium expanded lexicon of the pt-ca pair system.

BASIC INFORMATION

Identification Information

- resourceName: Portuguese LMF Apertium Dictionary
- resourceShortName: LMF Apertium Pt
- url: http://www.apertium.org
- metaShareId: NOT_DEFINED_FOR_V2

- description:

This is the LMF version of the Apertium Portuguese dictionary. Monolingual dictionary for Portuguese was generated from the Apertium expanded lexicon of the pt-ca pair system.

Apertium is a free/open-source machine translation platform, initially aimed at related-language pairs but recently expanded to deal with more divergent language pairs (such as English-Catalan).

The platform provides: a language-independent machine translation engine; tools to manage the linguistic data necessary to build a machine translation system for a given language pair and linguistic data for a growing number of language pairs.

ADMINISTRATIVE INFORMATION

Contact Person(s)

- surname: Forcada
- givenName: Mikel
- communicationInfo:
    - email: mlf@dlsi.ua.es
    - zipCode: 03080
    - city: Alicante
    - country: Spain
- affiliation:
    - organizationName: Universitat d'Alacant, (Grup Transducens)
    - organizationShortName: Universitat d'Alacant, Grup Transducens
    - communicationInfo:
        - email: info@prompsit.com
        - url: http://transducens.dlsi.ua.es/
        - address: Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante. 03080 Alicante
        - zipCode: 03080
        - city: Alicante
        - country: Spain
        - telephoneNumber: +34 96 5903772
        - faxNumber: +34 96 5909326

- surname: Vivaldi
- givenName: Jorge
- communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
- affiliation:
    - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
    - organizationShortName: IULA UPF
    - departmentName: Institut Universitari Lingüística Aplicada
    - communicationInfo:
        - email: jorge.vivaldi@upf.edu

        - url: http://www.iula.upf.edu/
        - address: Roc Boronat, 138
        - zipCode: 08018
        - city: Barcelona
        - country: Spain
        - telephoneNumber: +34 93 5421207
        - faxNumber: +34 93 5422321

Distribution Information

- Availability type: available-unrestrictedUse
- Licence: GPL

TECHNICAL INFORMATION

- Resource type: lexicalConceptualResource
- Format: text/xml
- Character Encoding: UTF-8
- Creation information:
    - originalSource:
        - targetResourceNameURI: Apertium Monolingual dictionary as described in http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
    - creationMode: automatic
    - creationModeDetails: This lexicon was created with the ApertiumMonolingual2LMF.pl script
    - creationTool:
        - targetResourceNameURI: ApertiumMonolingual2LMF.pl
- Validation Information:
    - validated: true
    - validationType: formal
    - validationMode: automatic
    - validationModeDetails: The lexicon validates against the LMF DTD v.16

CONTENT INFORMATION

- Encoding information:
    - encodingLevel: morphology
    - linguisticInformation: lemma
    - linguisticInformation: partOfSpeech
    - conformanceToStandardsBestPractices: LMF
- Linguality: monolingual
- Language(s): Portuguese

RELEVANT REFERENCES AND OTHER INFORMATION

Resource Documentation

- · documentUnstructured: http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf

- · documentUnstructured: Forcada,M.L., Ivanov, B., Ortiz, S., Pérez, J.A., Ramírez, G., Sánchez, F., Armentano, C., Montava, M.A., Tyers, F.M., (2010). 'Documentation of the Open Source Shallow-Transfer

Machine Translation Platform Apertium'. Departament de Llenguatges i Sistemes Informàtics Universitat d'Alacant, March 2010.

Resource Creation Information

Resource Creator(s):

- organizationInfo:
  - organizationName: Universitat Politècnica de Catalunya
  - organizationShortName: UPC
  - communicationInfo:
    - url: http://upc.edu
    - city: Barcelona
    - country: Spain

- organizationInfo:
  - organizationName: Universitat d'Alacant, (Grup Transducens)
  - organizationShortName: Universitat d'Alacant, Grup Transducens
  - communicationInfo:
    - email: info@prompsit.com
    - url: http://transducens.dlsi.ua.es/
    - address: Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante. 03080 Alicante
    - zipCode: 03080
    - city: Alicante
    - country: Spain
    - telephoneNumber: +34 96 5903772
    - faxNumber: +34 96 5909326

- personInfo:
  - surname: Carmen Armentano Oller
  - communicationInfo:

- organizationInfo:
  - organizationName: Universitat Pompeu Fabra - Institut Universitari Lingüística Aplicada
  - organizationShortName: IULA - UPF
  - departmentName: Institut Universitari Lingüística Aplicada (UPF)
  - communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - url: http://www.iula.upf.edu/
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - telephoneNumber: +34 93 5421207
    - faxNumber: +34 93 5422321

Resource Funding:

- projectName: Opentrad
- projectShortName: Opentrad

- url: http://www.opentrad.com/
- fundingType: nationalFunds
- funder: Ministerio de Industria, Turismo y Comercio. Gobierno de España
- fundingCountry: Spain

- projectName: Apertium 2.0
- projectShortName: Apertium
- url: http://www.apertium.org
- fundingType: nationalFunds
- funder: Ministerio de Industria, Turismo y Comercio. Gobierno de España
- funder: Secretariat de Telecomunicacions i Societat de la Informació de Generalitat de Catalunya
- fundingCountry: Spain

- projectName: METANET4U
- projectShortName: METANET4U
- url: http://metanet4u.eu/
- fundingType: ownFunds
- fundingType: euFunds
- funder: UPF
- funder: EU
- fundingCountry: Spain

Parole/Simple LMF lexicon Catalan

Documentation for the 'Catalan LMF ParoleSimple Lexicon'

This is the LMF version of the Catalan Parole-Simple lexicon.

BASIC INFORMATION

Identification Information

- resourceName: Catalan LMF ParoleSimple Lexicon
- resourceShortName: Catalan LMF ParoleSimple Lexicon
- url: http://www.iec.es
- metaShareId: NOT_DEFINED_FOR_V2

- description:

This is the LMF version of the Catalan Parole-Simple lexicon.

The original PAROLE lexica (20,000 entries per language) were built conform to a model based on EAGLES guidelines and GENELEX results, underlying a common lexical tool adapted from the EUREKA-GENELEX project. This software tool was extended to support the PAROLE model and conversion and management processes of the resulting resources. The languages involved in PAROLE lexica are: Catalan, Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portugese, Spanish and Swedish.

The goal of SIMPLE project was to add semantic information, selected for its relevance for LE applications, to the set of harmonised multifunctional lexica built for 12 European languages by the PAROLE consortium. PAROLE

+SIMPLE lexicons contain morphological, syntactic and semantic information, organised according to a common model and to common linguistic specifications.

ADMINISTRATIVE INFORMATION

Contact Person(s)

- surname: Soler
- givenName: Joan
- communicationInfo:
    - email: jsoler@iec.es
    - address: Carme, 47
    - zipCode: 08001
    - city: Barcelona
    - country: Spain
- affiliation:
    - organizationName: Institut d'Estudis Catalans
    - organizationShortName: IEC
    - communicationInfo:
        - email: jsoler@iec.es
        - address: Carme, 47
        - zipCode: 08001
        - city: Barcelona
        - country: Spain

- surname: Vivaldi
- givenName: Jorge
- communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
- affiliation:
    - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
    - organizationShortName: IULA UPF
    - departmentName: Institut Universitari Lingüística Aplicada
    - communicationInfo:
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/
        - address: Roc Boronat, 138
        - zipCode: 08018
        - city: Barcelona
        - country: Spain
        - telephoneNumber: +34 93 5421207
        - faxNumber: +34 93 5422321

Distribution Information

- Availability type: available-restrictedUse
- Licence: MSCommons_BY-NC-SA

TECHNICAL INFORMATION

- Resource type: lexicalConceptualResource
- Format: text/xml
- Character Encoding: UTF-8
- Creation information:
    - originalSource:
        - targetResourceNameURI: Catalan PAROLE lexicon as described in http://www.ub.edu/gilcub/lascosas/pubYreps/index_par.html.
    - originalSource:
        - targetResourceNameURI: The original source was updated and corrected so that it validates against the parole DTD and eventually against the LMF DTD.
    - creationMode: automatic
    - creationModeDetails: This lexicon was created with the ParoleSimple2LMF.xsl stylesheet.
    - creationTool:
        - targetResourceNameURI: ParoleSimple2LMF.xsl
- Validation Information:
    - validated: true
    - validationType: formal
    - validationMode: automatic
    - validationModeDetails: The lexicon validates against the LMF DTD v.16

CONTENT INFORMATION

- Encoding information:
    - encodingLevel: morphology
    - encodingLevel: syntax
    - encodingLevel: semantics
    - linguisticInformation: lemma
    - linguisticInformation: morpho-Inflection
    - linguisticInformation: partOfSpeech
    - linguisticInformation: syntax-SubcatFrame
    - linguisticInformation: syntacticoSemanticLinks
    - conformanceToStandardsBestPractices: LMF
    - theoreticModel: Parole
    - theoreticModel: Genelex
- Linguality: monolingual
- Language(s): Catalan

RELEVANT REFERENCES AND OTHER INFORMATION

Resource Documentation

- · documentUnstructured: http://institucional.iec.cat/gc/ViewPage.action?siteNodeId=929&languageId=5&contentId=6554

- · documentUnstructured: http://www.ub.edu/gilcub/SIMPLE/reports/simple/D31_BARfin.rtf

- · documentUnstructured: http://www.ilc.cnr.it/AZ_bibliography/Z179.PDF

Resource Creation Information

Resource Creator(s):

- organizationInfo:
  - organizationName: Institut d'Estudis Catalans
  - organizationShortName: IEC
  - communicationInfo:
    - email: jsoler@iec.es
    - address: Carme, 47
    - zipCode: 08001
    - city: Barcelona
    - country: Spain

- organizationInfo:
  - organizationName: Institut Universitari Lingüística Aplicada - UPF
  - organizationShortName: IULA - UPF
  - departmentName: Institut Universitari Lingüística Aplicada (UPF)
  - communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - url: http://www.iula.upf.edu/
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - telephoneNumber: +34 93 5421207
    - faxNumber: +34 93 5422321

Resource Funding:

- projectName: SIMPLE (LE4-8346)
- projectShortName: SIMPLE
- url: http://www.ub.edu/gilcub/SIMPLE/simple.html
- fundingType: euFunds
- funder: EU

- projectName: METANET4U
- projectShortName: METANET4U
- url: http://metanet4u.eu/
- fundingType: ownFunds
- fundingType: euFunds
- funder: UPF
- funder: EU
- fundingCountry: Spain

SenSem Corpus

This is the stand-off GrAF version of the SenSem Spanish Corpus.

BASIC INFORMATION

Identification Information

- resourceName: GrAF version of the SenSem Spanish Corpus
- resourceShortName: GrAF version of the SenSem Spanish Corpus
- url: http://grial.uab.es/
- metaShareId: NOT_DEFINED_FOR_V2

- description:

This is the stand-off GrAF version of the SenSem Spanish Corpus.

The original SenSem Spanish Corpus includes syntactic and semantic annotations for a number of Spanish texts from the press domain developed by the GRIAL group (Grup de recerca consolidat de la Generalitat de Catalunya). The corpus contains one million words 300,000 of which were manually annotated at the syntactic and semantic level with syntagmatic categories, syntactic functions and semantic roles.

ADMINISTRATIVE INFORMATION

Contact Person(s)

- surname: Fernandez Montraveta
- givenName: Anna
- communicationInfo:
    - email: Ana.Fernandez@uab.es
    - country: Spain
- affiliation:
    - organizationName: Universitat Autònoma de Barcelona
    - organizationShortName: UAB
    - communicationInfo:
        - email: Ana.Fernandez@uab.es
        - url: http://www.uab.es/
        - country: Spain

- surname: Vivaldi
- givenName: Jorge
- communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
- affiliation:
    - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
    - organizationShortName: IULA UPF
    - departmentName: Institut Universitari Lingüística Aplicada
    - communicationInfo:
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/
        - address: Roc Boronat, 138
        - zipCode: 08018
        - city: Barcelona

        - country: Spain
        - telephoneNumber: +34 93 5422332
        - faxNumber: +34 93 5422321

Distribution Information

- Availability type: available-unrestrictedUse
- Licence: GPL

TECHNICAL INFORMATION

- Resource type: corpus
- Character Encoding: UTF-8
- Validation Information:
    - validated: true
    - validationType: formal
    - validationMode: automatic
    - validationModeDetails: The documents validates against the GrAF DTD v.1.0.4

CONTENT INFORMATION

- Linguality: monolingual
- Language(s): Spanish
- Size: 893379 tokens
- Annotation information:
    - annotationType: semanticAnnotation-semanticRoles
    - annotationStandoff: true
    - segmentationLevel: phrase
    - annotationFormat: GrAF
    - tagset: other
- Annotation information:
    - annotationType: syntacticAnnotation-shallowParsing
    - annotationStandoff: true
    - segmentationLevel: phrase
    - annotationFormat: GrAF
    - tagset: other

RELEVANT REFERENCES AND OTHER INFORMATION

Resource Documentation

- · documentUnstructured: Alonso, L., J.A. Capilla, I. Castellón, A. Fernández, G. Vázquez (2007). "The Sensem Project: Syntactico-Semantic Annotation of Sentences in Spanish". N.Nikolov, K. Bontcheva, G.Angelova and R. Mitkov. (ed.), Recent Advances in Natural Language Processing IV. Selected papers from RANLP 2005. Current Issues in Linguistic Theory 292John Benjamins Publishing Co, p. 89-98. ISBN: 978 90 272 4807 7

- · documentUnstructured: Castellón, I., A. Fernández, G. Vázquez, L. Alonso, J.A. Capilla (2006). "The Sensem Corpus: a Corpus Annotated at the Syntactic and Semantic Level", Fifth International Conference on Language Resources and Evaluation (LREC), p. 355-359

- · documentUnstructured: Web site of the SenSem project: http://grial.uab.es/fproj.php?id=1

Resource Creation Information

Resource Creator(s):

- organizationInfo:
  - organizationName: GRIAL - Grup de recerca consolidat de la Generalitat de Catalunya (SGR 2009-00049)
  - organizationShortName: GRIAL
  - communicationInfo:
    - email: Ana.Fernandez@uab.es
    - url: http://grial.uab.es/index.php
    - country: Spain

- personInfo:
  - surname: Fernandez Montraveta
  - givenName: Ana
  - communicationInfo:
    - email: ana.fernandez@uab.es

- personInfo:
  - surname: FVázquez
  - givenName: AGlòriana
  - communicationInfo:
    - email: gvazquez@dal.udl.cat

- personInfo:
  - surname: Castellón
  - givenName: Irene
  - communicationInfo:
    - email: icastellon@ub.edu

- organizationInfo:
  - organizationName: Universitat Pompeu Fabra - Institut Universitari Lingüística Aplicada
  - organizationShortName: IULA - UPF
  - departmentName: Institut Universitari Lingüística Aplicada (UPF)
  - communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - url: http://www.iula.upf.edu/
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - telephoneNumber: +34 93 5421207
    - faxNumber: +34 93 5422321

Resource Funding:

- projectName: SenSem
- projectShortName: SenSem

- url: http://grial.uab.es/fproj.php?id=10
- fundingType: nationalFunds
- funder: Ministerio de educación y Ciencia. Gobierno de España (HUM2007-65267)
- fundingCountry: Spain

- projectName: METANET4U
- projectShortName: METANET4U
- url: http://metanet4u.eu/
- fundingType: ownFunds
- fundingType: euFunds
- funder: UPF
- funder: EU
- fundingCountry: Spain

SenSem Database (lexicon) Catalan

This is the LMF version of the SenSem database created by the Spanish Inter-University Research Group GRIAL.

BASIC INFORMATION

Identification Information

- resourceName: LMF version of the SenSem Catalan Data Base
- resourceShortName: LMF version of the SenSem Catalan Data Base
- url: http://grial.uab.es/
- metaShareId: NOT_DEFINED_FOR_V2

- description:

This is the LMF version of the SenSem database created by the Spanish Inter-University Research Group GRIAL.

As part of SenSem project, a corpus of sentences annotated at the semantic and syntactic levels was created.

The source corpus is made up of around 13 million words extracted from the online versions of a Spanish newspaper. From this corpus, 25.000 sentences have been randomly selected, 100 for each of the 250 more frequent verbs in current Spanish. Each sentence has been labeled according to the verb sense it exemplifies, the type of complements it takes (arguments or adjunts), their syntactic category and function, and finally each argument has been labelled with a semantic role. The sentence has also been annotated as to its semantics both in relation with aspectual information and the type of construction being expressed.

From this annotated corpus a lexical data base of verbs was created in which all the previous information will be recollected. The unit of description of the verbs is the sense. In the description of the verbs, argument structure is included, incorporating subcategorization patterns, with the information of frequency of them, semantic roles and information regarding sentence semantics.

The lexicon and the corpus are associated at sense level and together shape up what we call the data bank of the sentential semantic of the Spanish verbs. Both resources are available via web and will form a very important

source of linguistic information which we hope will be of utility in different areas of the natural language processing and linguistic research in general.

The LMF conversion has been done by the Universitat Pompeu Fabra.

ADMINISTRATIVE INFORMATION

Contact Person(s)

- surname: Fernandez Montraveta
- givenName: Anna
- communicationInfo:
    - email: Ana.Fernandez@uab.es
    - country: Spain
- affiliation:
    - organizationName: Universitat Autònoma de Barcelona
    - organizationShortName: UAB
    - communicationInfo:
        - email: Ana.Fernandez@uab.es
        - url: http://www.uab.es/
        - country: Spain

- surname: Vivaldi
- givenName: Jorge
- communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
- affiliation:
    - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
    - organizationShortName: IULA UPF
    - departmentName: Institut Universitari Lingüística Aplicada
    - communicationInfo:
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/
        - address: Roc Boronat, 138
        - zipCode: 08018
        - city: Barcelona
        - country: Spain
        - telephoneNumber: +34 93 5422332
        - faxNumber: +34 93 5422321

Distribution Information

- Availability type: available-unrestrictedUse
- Licence: GPL

TECHNICAL INFORMATION

- Resource type: lexicalConceptualResource
- Format: text/xml
- Character Encoding: UTF-8
- Creation information:
    - originalSource:
        - targetResourceNameURI: SenSem data base located at http://grial.uab.es/index.php
    - creationMode: automatic
- Validation Information:
    - validated: true
    - validationType: formal
    - validationMode: automatic
    - validationModeDetails: The documents validates against the LMF DTD v 16

CONTENT INFORMATION

- Encoding information:
    - encodingLevel: morphology
    - linguisticInformation: lemma
    - linguisticInformation: definition/gloss
    - linguisticInformation: usage-Examples
    - linguisticInformation: usage-Frequency
    - linguisticInformation: semantics-SemanticRoles
    - linguisticInformation: semantics-CrossReferences
    - conformanceToStandardsBestPractices: LMF
- Linguality: monolingual
- Language(s): Catalan

RELEVANT REFERENCES AND OTHER INFORMATION

Resource Documentation

- · documentUnstructured: Alonso, L., J.A. Capilla, I. Castellón, A. Fernández, G. Vázquez (2007). "The Sensem Project: Syntactico-Semantic Annotation of Sentences in Spanish". N.Nikolov, K. Bontcheva, G.Angelova and R. Mitkov. (ed.), Recent Advances in Natural Language Processing IV. Selected papers from RANLP 2005. Current Issues in Linguistic Theory 292John Benjamins Publishing Co, p. 89-98. ISBN: 978 90 272 4807 7

- · documentUnstructured: Castellón, I., A. Fernández, G. Vázquez, L. Alonso, J.A. Capilla (2006). "The Sensem Corpus: a Corpus Annotated at the Syntactic and Semantic Level", Fifth International Conference on Language Resources and Evaluation (LREC), p. 355-359

- · documentUnstructured: Web site of the SenSem project: http://grial.uab.es/fproj.php?id=1

Resource Creation Information

Resource Creator(s):

- organizationInfo:
    - organizationName: GRIAL - Grup de recerca consolidat de la Generalitat de Catalunya (SGR 2009-00049)
    - organizationShortName: GRIAL

    - communicationInfo:
        - email: Ana.Fernandez@uab.es
        - url: http://grial.uab.es/index.php
        - country: Spain

- personInfo:
    - surname: Fernandez Montraveta
    - givenName: Ana
    - communicationInfo:
        - email: ana.fernandez@uab.es

- personInfo:
    - surname: Vázquez
    - givenName: Glòria
    - communicationInfo:
        - email: gvazquez@dal.udl.cat

- personInfo:
    - surname: Castellón
    - givenName: Irene
    - communicationInfo:
        - email: icastellon@ub.edu

- organizationInfo:
    - organizationName: Universitat Pompeu Fabra - Institut Universitari Lingüística Aplicada
    - organizationShortName: IULA - UPF
    - departmentName: Institut Universitari Lingüística Aplicada (UPF)
    - communicationInfo:
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/
        - address: Roc Boronat, 138
        - zipCode: 08018
        - city: Barcelona
        - country: Spain
        - telephoneNumber: +34 93 5421207
        - faxNumber: +34 93 5422321

Resource Funding:

- projectName: SenSem
- projectShortName: SenSem
- url: http://grial.uab.es/fproj.php?id=10
- fundingType: nationalFunds
- funder: Ministerio de educación y Ciencia. Gobierno de España (HUM2007-65267)
- fundingCountry: Spain

- projectName: METANET4U
- projectShortName: METANET4U
- url: http://metanet4u.eu/
- fundingType: ownFunds
- fundingType: euFunds

- funder: UPF
- funder: EU
- fundingCountry: Spain

SenSem Database (lexicon) Spanish

This is the LMF version of the SenSem database created by the Spanish Inter-University Research Group GRIAL.

BASIC INFORMATION

Identification Information

- resourceName: LMF version of the SenSem Spanish Data Base
- resourceShortName: LMF version of the SenSem Spanish Data Base
- url: http://grial.uab.es/
- metaShareId: NOT_DEFINED_FOR_V2

- description:

This is the LMF version of the SenSem database created by the Spanish Inter-University Research Group GRIAL.

As part of SenSem project, a corpus of sentences annotated at the semantic and syntactic levels was created.

The source corpus is made up of around 13 million words extracted from the online versions of a Spanish newspaper. From this corpus, 25.000 sentences have been randomly selected, 100 for each of the 250 more frequent verbs in current Spanish. Each sentence has been labeled according to the verb sense it exemplifies, the type of complements it takes (arguments or adjunts), their syntactic category and function, and finally each argument has been labelled with a semantic role. The sentence has also been annotated as to its semantics both in relation with aspectual information and the type of construction being expressed.

From this annotated corpus a lexical data base of verbs was created in which all the previous information will be recollected. The unit of description of the verbs is the sense. In the description of the verbs, argument structure is included, incorporating subcategorization patterns, with the information of frequency of them, semantic roles and information regarding sentence semantics.

The lexicon and the corpus are associated at sense level and together shape up what we call the data bank of the sentential semantic of the Spanish verbs. Both resources are available via web and will form a very important source of linguistic information which we hope will be of utility in different areas of the natural language processing and linguistic research in general.

The LMF conversion has been done by the Universitat Pompeu Fabra.

ADMINISTRATIVE INFORMATION

Contact Person(s)

- surname: Fernandez Montraveta

- givenName: Anna
- communicationInfo:
    - email: Ana.Fernandez@uab.es
    - country: Spain
- affiliation:
    - organizationName: Universitat Autònoma de Barcelona
    - organizationShortName: UAB
    - communicationInfo:
        - email: Ana.Fernandez@uab.es
        - url: http://www.uab.es/
        - country: Spain

- surname: Vivaldi
- givenName: Jorge
- communicationInfo:
    - email: jorge.vivaldi@upf.edu
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
- affiliation:
    - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
    - organizationShortName: IULA UPF
    - departmentName: Institut Universitari Lingüística Aplicada
    - communicationInfo:
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/
        - address: Roc Boronat, 138
        - zipCode: 08018
        - city: Barcelona
        - country: Spain
        - telephoneNumber: +34 93 5422332
        - faxNumber: +34 93 5422321

Distribution Information

- Availability type: available-unrestrictedUse
- Licence: GPL

TECHNICAL INFORMATION

- Resource type: lexicalConceptualResource
- Format: text/xml
- Character Encoding: UTF-8
- Creation information:
    - originalSource:
        - targetResourceNameURI: SenSem data base located at http://grial.uab.es/index.php
    - creationMode: automatic
- Validation Information:
    - validated: true
    - validationType: formal

- validationMode: automatic
- validationModeDetails: The documents validates against the LMF DTD v 16

CONTENT INFORMATION

- Encoding information:
  - encodingLevel: morphology
  - linguisticInformation: lemma
  - linguisticInformation: definition/gloss
  - linguisticInformation: usage-Examples
  - linguisticInformation: usage-Frequency
  - linguisticInformation: semantics-SemanticRoles
  - linguisticInformation: semantics-CrossReferences
  - conformanceToStandardsBestPractices: LMF
- Linguality: monolingual
- Language(s): Spanish

RELEVANT REFERENCES AND OTHER INFORMATION

Resource Documentation

- · documentUnstructured: Alonso, L., J.A. Capilla, I. Castellón, A. Fernández, G. Vázquez (2007). "The Sensem Project: Syntactico-Semantic Annotation of Sentences in Spanish". N.Nikolov, K. Bontcheva, G.Angelova and R. Mitkov. (ed.), Recent Advances in Natural Language Processing IV. Selected papers from RANLP 2005. Current Issues in Linguistic Theory 292John Benjamins Publishing Co, p. 89-98. ISBN: 978 90 272 4807 7

- · documentUnstructured: Castellón, I., A. Fernández, G. Vázquez, L. Alonso, J.A. Capilla (2006). "The Sensem Corpus: a Corpus Annotated at the Syntactic and Semantic Level", Fifth International Conference on Language Resources and Evaluation (LREC), p. 355-359

- · documentUnstructured: Web site of the SenSem project: http://grial.uab.es/fproj.php?id=1

Resource Creation Information

Resource Creator(s):

- organizationInfo:
  - organizationName: GRIAL - Grup de recerca consolidat de la Generalitat de Catalunya (SGR 2009-00049)
  - organizationShortName: GRIAL
  - communicationInfo:
    - email: Ana.Fernandez@uab.es
    - url: http://grial.uab.es/index.php
    - country: Spain

- personInfo:
  - surname: Fernandez Montraveta
  - givenName: Ana
  - communicationInfo:
    - email: ana.fernandez@uab.es

- personInfo:
    - surname: Vázquez
    - givenName: Glòria
    - communicationInfo:
        - email: gvazquez@dal.udl.cat

- personInfo:
    - surname: Castellón
    - givenName: Irene
    - communicationInfo:
        - email: icastellon@ub.edu

- organizationInfo:
    - organizationName: Universitat Pompeu Fabra - Institut Universitari Lingüística Aplicada
    - organizationShortName: IULA - UPF
    - departmentName: Institut Universitari Lingüística Aplicada (UPF)
    - communicationInfo:
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/
        - address: Roc Boronat, 138
        - zipCode: 08018
        - city: Barcelona
        - country: Spain
        - telephoneNumber: +34 93 5421207
        - faxNumber: +34 93 5422321

Resource Funding:

- projectName: SenSem
- projectShortName: SenSem
- url: http://grial.uab.es/fproj.php?id=10
- fundingType: nationalFunds
- funder: Ministerio de educación y Ciencia. Gobierno de España (HUM2007-65267)
- fundingCountry: Spain

- projectName: METANET4U
- projectShortName: METANET4U
- url: http://metanet4u.eu/
- fundingType: ownFunds
- fundingType: euFunds
- funder: UPF
- funder: EU
- fundingCountry: Spain

6.2 Appendix 2: Quick Validation Report

BATCH 2 Validation Results

Document Updated: 23.07.2012

Color code:

Green – Resource is validated

Orange – Resource received but fails complete validation

Red – Resource NOT received

Blue – Resource does not need validation, reason in parantheses

Contents

ULX - University of Lisbon 2

IST - Instituto Superior Técnico 5

UNIMAN-University of Manchester 6

UAIC – University Alexandru Ioan Cuza 9

RACAI – Romanian Academy 11

UOM - University of Malta 14

UPC – Universitat Politècnica de Catalunya 16

UPF - University Pompeu Fabra 19

ULX - University of Lisbon

Endogenous resources

Abbreviations - [LX-Abbreviations]

- All files are present and folder structure is correct.
- XML Valid.
- Contains 208 abbreviations in 20 groups.

C-ORAL-ROM Portuguese Corpus – [C-ORAL]

- All files are present and folder structure is correct.
- Corpus counts correspond to the Narrative file.

CINTIL-Internacional Corpus of Portuguese - [CINTIL Corpus]

- All files are present and folder structure is correct.
- XML Invalid:
  1. Files require missing dtd files : „<!DOCTYPE cintil SYSTEM "cintil-spoken.dtd">” and „<!DOCTYPE cintil SYSTEM "cintil-written.dtd">”
  2. Errors like: CINTIL-WRITTEN_withxml.col:57: parser error : Opening and ending tag mismatch: cintil line 3 and p : „</s></p>”
- CINTIL Spoken contains 10635 sentences and 211396 tokens including punctuation (as opposed to 502622 as stated in the narrative). CINTIL written contains 30344 sentences with 689602 tokens (as opposed to 689124 as stated in the narrative). The original text is present under the tokenized form, one per line. Combined file size is 18.6MB: 3.8MB for Spoken and 14.8 for Written.

Lexicon of multiword expressions – [LEX-MWE-PT]

- All files are present and folder structure is correct.
- We have found 13442 LEM items in the text file as opposed to 1198 main lemmas and 12753 MWE lemmas. Do the numbers combine in some for or another to give out 13442 items in total (is it possible that lemmas overlap?)
- The html files have a problem displaying on standard english browsers. That is because even though the files are UTF8 encoded, the meta tag in the <head> section does not force UTF8 display. The correction that needs to be done is to change the tag from :
  - <meta http-equiv="Content-Type" content="text/html">
  - <meta http-equiv="Content-Type" content="text/html**; charset=UTF-8**">

MWN.PT - [already in ELRA]

PAROLE corpus - [already in ELRA]

PAROLE lexicon - [already in ELRA]

PropBank – [CINTIL Propbank]

- All files are present and folder structure is correct.

- XML Valid.

SIMPLE lexicon – [SIMPLElexicon]

- All files are present and folder structure is correct.
- The file contains 10438 entries.
- The file is not UTF8 encoded, but rather ASCII (127-256) encoded. Normally, it displays ok (as usually ANSI is the default encoding, but when converted to be viewed as UTF8, all ASCII >127 characters display strangely :
  - In ANSI (ASCII encoding): example="A composição das listas eleitorais gerou polémica."
  - In UTF8: example="A composi裨 das listas eleitorais gerou pol⍰ca."

Stopwords – [LX-Stopwords]

- All files are present and folder structure is correct.
- XML Valid.
- Contains 2631 words in different 3 classes.

Treebank – [CINTIL Treebank]

- All files are present and folder structure is correct.
- XML Valid.

Restricted exogenous resources

Clássicos LP/Porto Editora - [exogenous resource]
COMPARA - [exogenous resource]
Corpus NILC - [exogenous resource]
Dicionário de Verbos do Português Medieval (DVPM) - [exogenous resource]
European Parliament Parallel Corpus - [exogenous resource]
Geo-Net-PT01 - [exogenous resource]
Glossário 0,15- [exogenous resource]
MorDebe 0,15- [exogenous resource]
Norma Urbana Culta (NURC) - [exogenous resource]
Ontologia de Nanociência e Nanotecnologia - [exogenous resource]
Panorama do Português Oral de Maputo (PPOM) - [exogenous resource]
PORLEX - [exogenous resource]
The JRC-Acquis Multilingual Parallel Corpus - [exogenous resource]

Unrestricted Exogenous resources

CETEMPúblico - [exogenous resource]
Corpus NILC - [exogenous resource]
CorpusTCC - [exogenous resource]
PLN-BR Gold - [exogenous resource]
RHETALHO - [exogenous resource]

Summ-it - [exogenous resource]
TeMário 2006 - [exogenous resource]

Unrestricted exogenous resources (tools)

DiZer 2.0 - [exogenous resource]
Forma - [exogenous resource]
GistSumm - [exogenous resource]
NILC Taggers - [exogenous resource]
Ontolp Plugin - [exogenous resource]
Stemmer - [exogenous resource]
Text Aligners - [exogenous resource]

IST - Instituto Superior Técnico

Endogenous resources

TED Talks (3) – [TED_txt/wav.zip]

- The files are present and the folder structure is accuratly described in the Narrative file.
- The AlGore TED resource contains 386 transcript segments enclosing 2071 tokens in the xml file. The DanDennett resource contains 467 transcript segments enclosing 2889 tokens in the xml file. The final MalcolmGladwell resource contains 675 transcript segments enclosing 3318 tokens in the xml file. In total, there are 8278 tokens in all three xml files.

CALL - NOT RECEIVED

PTSTAR Golden Collection - Cross-Language Unit Elicitation alignments (CLUE) – [europarl-alignments.zip]

- The files are present and the folder structure is accuratly described in the Narrative file.
- The token and alignment count are correct:
  For the .wa files:
  8035 en-es-Alinhamentos.wa
  7594 en-fr-Alinhamentos.wa
  7100 en-pt-Alinhamentos.wa
  9075 es-fr-Alinhamentos.wa
  8515 pt-es-Alinhamentos.wa
  7811 pt-fr-Alinhamentos.wa
  48130 alignments in total
  - For the .mwu files:

  3902 en-es-Alinhamentos.mwu
  4278 en-fr-Alinhamentos.mwu
  3195 en-pt-Alinhamentos.mwu
  3229 es-fr-Alinhamentos.mwu
  3544 pt-es-Alinhamentos.mwu
  3951 pt-fr-Alinhamentos.mwu
  22099 alignments in total

Restricted exogenous resources

TAP – [TAP]

- The files are present and the folder structure is accuratly described in the Narrative file.
- The corpus has 51 folders that contain 2243 text files for each of the pt and en languages.
- The files (in total) have 33879 lines. The english side has 775504 tokens (4.6 MB) while the portughese side has 771442 tokens (4.7MB)

UNIMAN-University of Manchester

Endogenous resources (tools)

U-Compare NaCTeM Sentence Detector – [no longer included in Batch 2]

The tool works correctly but there is no narrative document. This resource works but will no longer be included in Batch 2

U-Compare Platform – [online at http://www.nactem.ac.uk/ucompare/]

The documentation is, for the most part, reflective of reality. The "Technical Information" section has been validated in that the "UCLoader.class" has been downloaded and successfully run following the indications present in the documentation. We have tried (and confirmed that they work) a couple of workflows: Genia Tagger and TTL.
We have also tried to run the sample workflow from the command line by following the indications at

http://www.nactem.ac.uk/ucompare/developerguide/Command_Line_Mode_without_U.html

but the program stalled after approx. half an hour of running in which it output the information given below:

```
[]

[]

StdinProteinTagging-CPE.xml => file:/E:/temp/metanet/batch-2-eval/3%20UNIMAN/StdinProteinTagging-CPE.xml

Jun 13, 2012 10:58:36 AM org.apache.axis.utils.JavaUtils isAttachmentSupported

WARNING: Unable to find required classes (javax.activation.DataHandler and javax.mail.internet.MimeMultipart). Attachment support is disabled.

System starting.

workflow initialization completed.

0 0 uima.tcas.DocumentAnnotation id="u0" language="x-unspecified"

-1 -1 uima.cas.Sofa id="u1" sofaID="_InitialView" mimeType="text" sofaURI="" sofaNum="1" sofaString="" sofaArray=""
```

We would suggest that Section 3 "Content Information" is enlarged and the user guided in more detail when constructing the comparison given in Figure 1. As it is, is difficult (impossible to us) to reconstruct the example given in that figure.

U-Compare Workbench – [online at http://www.nactem.ac.uk/ucompare/]

The tool works correctly / validates.

U-Compare GENIA Sentence Detector

The tool works correctly / validates.

U-Compare GENIA Tokenizer

The tool works correctly / validates.

U-Compare OpenNLP PoSTagger

The annotation has been created (I have manually inspected the output file) but when displaying the results in U-Compare, I obtained the error:

```
Jun 13, 2012 2:34:47 PM org.apache.uima.collection.impl.cpm.engine.CPMThreadGroup process

SEVERE: The CPM thread group caught the following unhandled error: java.lang.StackOverflowError
(Thread Name: [Procesing Pipeline#1 Thread]::)
```

Even with increased heap and/or other test sentences, the error is the same which is not surprising since this is a stack overflow. Maybe there is a bug in `org.apache.uima.collection.impl.cpm.engine.CPMThreadGroup` or the XML, even if it's valid, is not compatible with U-Compare type system.

Command line for starting U-Compare: `java -Xms800m -Xmx1024m UCLoader`

Java version:

```
java version "1.7.0_02"

Java(TM) SE Runtime Environment (build 1.7.0_02-b13)

Java HotSpot(TM) Client VM (build 22.0-b10, mixed mode, sharing)
```

Text to process: "This is a test English sentence. It's supposed to show the functionality of the RACAI text preprocessing tools."

The workflow: OpenNLP Sentence Detector, OpenNLP Tokenizer and OpenNLP Tagger

U-Compare OpenNLP Sentence Detector

The annotation has been created (I have manually inspected the output file) but when displaying the results in U-Compare, I obtained the same error as in the case of OpenNLP POS Tagger. The test environment was the same.

U-Compare OpenNLP Tokenizer

The annotation has been created (I have manually inspected the output file) but when displaying the results in U-Compare, I obtained the same error as in the case of OpenNLP POS Tagger. The test environment was the same.

U-Compare Type System

The type system validates.

Unrestricted exogenous resources (tools)

Apertium Morphological Analyser – [ApertiumMorpho.jar]

I have followed the steps described in the documentation to add ApertiumMorpho to U-Compare.

I have obtained the error:

```
java.lang.NoClassDefFoundError: org/apertium/lttoolbox/LTProc
```

Test settings:

Input text in Romanian: “Aceasta este o propoziție de test pentru Apertium. Să vedem cum funcționează.” from the U-Compare component “Input Text Reader”.

Language pair: ‘ro-es’ (but after I have saved the changes, the default value was still present: ‘en-es’). After changing back to ‘en-es’, the displayed language pair was ‘ro-es’ so this is a bug in U-Compare?

I obtain the same error with ‘en-es’ on an English sentence. Maybe the required class was not packaged?

Apertium POS tagger – [ApertiumPOS.jar]

Testing was not performed since it depends on ApertiumMorpho which does not run.

Apertium MT transfer

Apertium Mophological generator

U-Compare Cafetiere sentence splitter

The tool works correctly / validates.

UAIC – University Alexandru Ioan Cuza

Endogenous resources (tools)

UAIC-Tokenizer

The tool works as an U-Compare component, according to specifications.

Segmenter

The tool works as specified but only with java 1.7 (it doesn't work with java 1.6) and the specifications mention that it should work with any version above 1.5. The documentation needs to be modified accordingly.

Lemmatizer

Both versions of the lemmatizer work as U-Compare components, according to specifications.

NP chunker

The web application works as specified. The Web Service is working, but not according to specifications. As input, it needs raw text files and not xml files with an intermediate level of annotation. The documentation needs to be modified accordingly.

Summarizer

summarizer_v1 is a web application which works according to specifications. However, the UTF-8 characters are correctly displayed if the user manually sets the right encoding in the browser. Still, this should be done automatically on the server side: try to set the MIME type in the output HTTP headers to UTF-8.

ï»¿Mircea Radulian, directorul Institutului de FizicÄƒ a PÄƒmÃ¢ntului,Radulian a explicat cÄƒ zona Vrancea este una imprevizibilÄƒ ÅŸi cÄƒ nu se poate preciza cu exactitate cÃ¢nd ÅŸi mai ales dacÄƒ va avea loc un cutremur puternic.

summarizer_v2 works according to specifications.

Categorizer

This is a web application that works according to the description. We have test it with several raw text files. No malfunction has been identified.

Discourse Parser

The tool works according to specifications.

RARE

After review/recheck: we have run the app. according to the instructions provided in the readme.txt file, which state to use the command:
java -jar rare.jar ENG input.xml out.xml
provided that ENG folder contains all the required resources. However, the app. terminated with an error. Here is the app.'s output:

```
equal
lemmaEquality
equal
equal
Rule numberAgreement was created
Rule LemmaRule was created
Rule PRPounAgreementOnNumGen was created
Rule heIsMale was created
Rule sequencesOfPRPouns was created
Rule SheIsFemale was created
Rule ItIsNotPerson was created
Rule SheIsShip was created
Rule ItIsBaby was created
Rule ItIsChild was created
Rule antecedentCollectiveNoun was created
Rule antecedentExquisitePerson was created
Exception in thread "main" java.lang.NullPointerExce
        at rare.model.rules.Rule.run(Rule.java:345)
        at rare.Engine.processPS(Engine.java:149)
        at rare.Engine.run(Engine.java:71)
    at infoiasi.Infoiasi.main(Infoiasi.java:44)
```

RACAI – Romanian Academy

Endogenous resources

Mapping list from PWN3.0 to PWN2.0

This resource contains all the data described by the documentation. The size of the resource matches the specifications. The provided links for downloading this resource are also functioning.

NAACL 2003

The corpus is XML XCES encoded and validates against the XCES schema found at http://www.xces.org. It is POS tagged and lemmatized and from our quick check of the annotations, no obvious errors were present.

It would have been nice to have an explicit alignment file that would indicate the sentence alignments but the explanations found in the narrative are sufficient for anyone trying to reconstruct the sentence alignments.

ROMORPH

- number of entries correct;
- XML Valid
- UTF8 encoding;
- the structure of tags, value and attributes is well detailed

RO-WORDNET (part 2)

This resource contains a part of the Romanian WordNet as an xml file. The format and the size of the resource match the specifications. Also, it is important to notice that, exactly as described in the documentation, the structure of this lexical ontology has no gaps.

Endogenous resources (tools)

COLLOC

We have run the executable provided in the package, using the input provided as example. The tool worked according to the specifications in a similar amount of time. We also modified different parameters (LL threshold, stdev threshold, POS filters, etc.) but no error occurred. The output format was as specified in the documentation.

LangId

First, we have tested this SOAP Web Service using the Web application available at:

http://www.racai.ro/webservices/LangId.aspx
We fed the application with German, Latvian and Greek texts mainly taken from Wikipedia. The application correctly identified the language of these texts according to specifications. The output was also conformant to specifications: for a given text, the user has access to 4 types of information:

(i) the name of the identified language of the text in English
(ii) the name of the identified language of the text in the identified language
(iii) the ISO code for that language
(iv) the confidence score

Second, we have created a C# application for testing this Web Service, using the example provided by the documentation. It worked as described and the running was verified.

LexChain

We have tested LexChain REST Web Service with multiple queries corresponding to the specifications described in the documentation. Here are a few examples we have tested:

http://khufu.racai.ro:8001/lexchains.ashx?w1=mouse&w2=computer
http://khufu.racai.ro:8001/lexchains.ashx?w1=mouse&w2=computer&max=3
http://khufu.racai.ro:8001/lexchains.ashx?w1=mouse&w2=computer&max=10
http://khufu.racai.ro:8001/lexchains.ashx?w1=mouse&w2=computer&max=50
http://khufu.racai.ro:8001/lexchains.ashx?w1=boat&w2=sea
http://khufu.racai.ro:8001/lexchains.ashx?w1=boat&pos1=n&w2=boat&pos2=v
http://khufu.racai.ro:8001/lexchains.ashx?ili1=09450163-n&ili2=09394007-n

The output for this queries is conformant to the specifications.

LexPar

We tested the web service both in the local network and form the Internet. The documentation is clear about the way the LexPar Perl wrapper works. We noticed processing speed variations which can be explained by the fact that the data takes longer to get to the web service server when running from the Internet. Other than that, the web service produced the expected output (the one found in the 'sample-output/' directory in the distribution kit) for each of the tested languages: English, French and Romanian.

RO-HYPHEN

The tool works according to specifications. It can be tested online at :
http://nlptools.racai.ro/nlptools/index.php?page=hyp_stress (contains html wrapper over the webservice)

The web service is available at :
http://khufu.racai.ro:8081/Hyphenator.asmx?op=GetHyphenation

TTL Package (TTL-Lemmatizer, TTL-Tagger, TTL-Tokenizer , TTL-Chunker)

The documentation clearly explains how the workflow is constructed. We tested all the TTL tools in the package (for all supported languages) both in the workflow and using separate intermediate files and they produce the advertised results. As in the case of LexPar, we also noticed speed degradations when running from the Internet. Also, speed can vary quite a bit most likely due to web service server load.

UOM - University of Malta

Endogenous resources

F_MONA_1 – [F-Mona]

- All files are present and folder structure is correct. Resource validates correctly.

MalToBI Corpus – [MalToBI Speech Corpus.zip]

- All files are present and folder structure is correct. Resource validates correctly.
- Narrative is in SR format.

Restricted Exogenous resources

Local Government documentation – [Governmental Resources.zip]

- All files are present and folder structure is correct. Resource validates correctly.
- <<The sentence *„This corpus, consisting of raw text files and comma separated values (CSV) files, is the percentage that could be extracted from the original corpus"* its not explanatory for the way the percentage was obtained (how was the extraction process validated). Randomly choosing a couple of files shows that the text is in bad shape and contains a lot of invalid characters. (RACAI)>>, UOM will address these conversion issues.

Malta Online Dictionary – [t.b.d. in Batch 3]

Maltese Speech Engine Corpus

Maltese Wikipedia – [Maltese Wikipedia.zip]

The size and filecount is ok.

Observations:

- The documentation file lacks clarity. No mention about the „namespace" tag and its content. What does „<namespace key="-2" case="first-letter">Medja</namespace>" mean?
- If another (clearer) documentation exists for the corpus it should be mentioned

SPAN – [same as MalToBI]

Unrestricted exogenous resources

Basic English-Maltese Dictionary – [Bilingual Dictionary.zip]

The XML file format is valid. Resource validates.

MFSA Maltese Company Registration

The counts correspond to the number of tokens. Resource validates.

UPC – Universitat Politècnica de Catalunya

Endogenous resources

ALBAYZIN 0,1 – [already in ELRA]
Catalan-SpeechDat(I) – [already in ELRA]
SALA-Mexico – [already in ELRA]
SALA-Venezuela – [already in ELRA]
Spanish SpeechDat (II) – [already in ELRA]
SpeechDat-Car Spain – [already in ELRA]
Speecon Catalan – [already in ELRA]
Interface expressive database – [already in ELRA]
LC-STAR Catalan Phonetic Lexicon – [already in ELRA]
LC-STAR Spanish Phonetic Lexicon – [already in ELRA]
UPC-ESMA – [upc_esma.06_2008.tgz]

- Folder structure is accuratly described in the Narrative file.
- The Narrative correctly and sufficiently describes all file types.
- There are 776 wav files (and corresponding lwv files) : 208 in /pa, 62 in /pl and 506 in /se.

TC-STAR Spanish Baseline Female – [already in ELRA]
TC-STAR Spanish Baseline Male – [already in ELRA]
TC-STAR Bilingual Expressive Speech – [already in ELRA]
TC-STAR Bilingual VC – [already in ELRA]
TM2 – Technical Meetings – [validation by UPC]

- Folder structure is accuratly described in the Narrative file.
- The Narrative correctly and sufficiently describes all file types.
- The numer of found files in each directory are:

DATA/
ANNOTATION/
AUDIO - 2 files
AUDIO-VISUAL/
EMOTIONS - 2 files
LINKS - 2 files
TEXT/
NAMED_ENTITIES - 2 files
TOPIC_SEGMENTATION - 2 files
VIDEO/
3D_TRACKING
/2D_COORD - 10 files
/3D_COORD - 2 files
/CALIB - 5 files
GENERAL - 2 files

SRL - 2 files SEMINARS/

AUDIO/
UPC_20110525 - 30 files
UPC_20110601 - 32 files
VIDEO/
UPC_20110525
cam2 - 56051 files
cam3 - 56051 files
cam4 - 56051 files
cam5 - 56051files
cam7 - 56051 files
UPC_20110601
cam2 - 50746 files
cam3 - 50746 files
cam4 - 50746 files
cam5 - 50746 files
cam7 - 50746 files

- During the final validation of the recorded material it was found that 2 close-talk channels (ctm_1.wav and ctm_4.wav) of the Spanish session are strongly distorted, so they were excluded from the database release. These channels correspond to the speakers UPC_030 and UPC_035 respectively.

CHIL2007+ – [already in ELRA]

Restricted exogenous resources

Ahosyn: Large Bilingual Speech Database for Synthesis – [external resource]

Restricted exogenous resources

Bizkaifon: speech and video database for the Western dialects of the Basque Language – [external resource]

EL_PERIODICO_97-07 – [t.b.d. in Batch 3]

EmodB_EU1: Emotional speechand video database in Standard Basque – [external resource]

EmodB_EU2: Emotional speech database in Standard Basque – [external resource]

Galician SpeechDat FDB – [external resource]

LAS CORTES – [already in ELRA]

SPANISH EPPS – [already in ELRA]

Speech Rate database for Basque – [external resource]

Speech-Dat like database for Basque – [external resource]

Speech-dat like database for Basque (Mobile) – [external resource]
Transgrigal DB 0,2 – [external resource]
DOGalicia: Parallel Galician-Spanish Corpus – [external resource]
GCG: GrupoCorreoGalego – [external resource]

Restricted Exogenous resources (tools)

Cotovia Transcriber – [external resource]

UPF - University Pompeu Fabra

Endogenous resources

News paper headlines corpus – [t.b.d. in Batch3]

TRL V-Subcat Lexicon – [PKG-TRL-V-SUbcat-Lexicon.tar.gz in PKG-IULA.tar.gz]

- Xml validation
- the documentation file does not contain important informations like: the number of lemmas in the lexicon, structure of the annotation, descriptions/examples of the entries, etc.
- The number of entries in the file corresponds to the number specified in the METADATA document (sizeInfo)

Restricted exogenous resources

CESS_EU: The Basque Dependency Treebank – [corpusCESS_EU.tar.gz]

- annotation in the specified format
- encoded utf8
- size of the corpus: corresponding to the documentation.
- segmentationLevel: phrase: the corpus seems to be segmented at the word level and at the sentence level
- XML not valid
    - Example of errors:

      eebs.450520392-dep.xml:2: validity error : Validation failed: no DTD found !
      w.xces.org/ns/GrAF/0.99/graf-0.99.xsd" xmlns="http://www.xces.org/ns/GrAF/0.99/"
      ! there is a space in the schema url!

      eebs.450520392-dep.xml:32: parser error : Opening and ending tag mismatch: header line 3 and annotationSet
      </annotationSet>

      eebs.450520392-dep.xml:33: parser error : Opening and ending tag mismatch: graph line 2 and header
      </header>

      eebs.450520392-dep.xml:34: parser error : Extra content at the end of the document
      <node xml:id="dep-nt1"/>

    - If the second error is corrected (Opening and ending tag mismatch: header line 3 and annotationSet):
      eebs.450430612/eebs.450430612-dep.xml:1756: element node: validity error : ID dep-nt4 already defined
      <node xml:id="dep-nt4"/>

- validationModeDetails: 1. The documents validates against the GrAF DTD v.1.0.4: the DTD mentioned is not in the resource folder
  2.the documentation file does not contain important informations like: the numer of words in the corpus, the file structure of the corpus, descriptions/examples of the entries, etc.

Corpus CLUVI – [t.b.d. in Batch3]

Corpus Técnico do Galego – [t.b.d. in Batch3]

Dicionario CLUVI inglés-galego – [t.b.d. in Batch3]

Euskal Wordnet 3.0 – [t.b.d. in Batch3]

Termoteca – [PKG-termoteca_LMF.tar.gz in file PKG-IULA.tar.gz]

- Xml valid
- sizeInfo – correct
- size per languages : correct
- UTF8 encoded

Unrestricted exogenous resources

Apertium English dictionary – [PKG-Apertium-EN.tar.gz in file PKG-IULA.tar.gz]

- Xml valid
- UTF8 encoded
- sizeInfo: 33384 instead of 33385
- linguisticInformation: checked
- more information than specified (morphological properties for the wordforms)

Apertium French dictionary – [PKG-Apertium-FR.tar.gz in file PKG-IULA.tar.gz]

- Xml validation
- UTF8 encoded
- sizeInfo: 21362 instead of 21363
- linguisticInformation:checked
- more information than specified (morphological properties for the wordforms)

Apertium Italian dictionary – [PKG-Apertium-IT.tar.gz in file PKG-IULA.tar.gz]

- Xml validation
- UTF8 encoded
- sizeInfo: 12083 instead of 12084
- linguisticInformation:checked
- more information than specified (morphological properties for the wordforms)

WikiCorpus - [WikiCorpus_CA.tar.gz *and* WikiCorpus_ES.tar.gz]

WikiCorpus_CA.tar.gz :

1. the size of the corpus is unclear:
   - in the narrative file, the size in bytes is unspecified
   - the size specified in „tokens" it is ambigous: what does token means?
   - We counted 10,100,591 words
   - No number of folders or documents is specified and the structure of the corpus is not described.
2. XML validation: there is no schema in the archive to validate against.

WikiCorpus_ES.tar.gz :

1. the size of the corpus is unclear:
   - in the narrative file, the size in bytes is unspecified
   - the size specified in „tokens" it is ambigous: what does token means?
   - We counted 10,100,591 words
   - No number of folders or documents is specified and the structure of the corpus is not described.
2. XML validation: there is no schema in the archive to validate against.

Endogenous resources (tools)

Converters to LMF 2 – [PKG-Apertium2LMF.tar.gz in file PKG-IULA.tar.gz]

- The tool was succesfully tested to generate a LMF form of a sample dictionary.
- The LMF dictionary validated against DTD_LMF_REV_16.dtd

Tools for automatic UTF-8 conversion – [t.b.d. in Batch3]

Tools for Catalan Corpus Processing - [IULA-WebServices.zip]

Mentions: we tested the tools for other examples than that in the documentation files.
Some of the documentation files do not offer test data.
None of the documentation files offers test data for catalan.
Some of the tools mentioned in the description that they can be applied to english, but the web service is not including the english option: iula tokenizer, iula paradigma

IULA lexicon look up Web Service

The tool was

- successfully tested for the following words: „textos", "totes", "associada"
- unsuccesfully tested for the following words: „expressions", "informació", "morfosintàctic"

  ✓ *<LexicalEntriesList><lexicalEntry><lemma>expressi□</lemma><PoSTag>N5-FP</PoSTag></lexicalEntry></LexicalEntriesList>*

- ✓ *<LexicalEntriesList><lexicalEntry><lemma>informaci□</lemma><PoSTag>N5-FS</PoSTag></lexicalEntry></LexicalEntriesList>*
- ✓ *<LexicalEntriesList><lexicalEntry><lemma>morfosint□ctic</lemma><PoSTag>JQ--MS</PoSTag></lexicalEntry></LexicalEntriesList>*

This led us to the conclusion that the output is in the extended ASCII format, and not in UTF-8.

IULA paradigma Web Service

- the output is in the extended ASCII format, and not in UTF-8

IULA processor

- The tool is not able to recognize a known word if it is in its capitalized form
  <div1>
  * <s>Va rebre el **<name>Premi</name>** Pr?ncep d'Ast?ries en <num>2005</num>en la categoria de Comunicaci? i <name>Humanitats</name>.</s> *
  </div1>
- the output is in the extended ASCII format, and not in UTF-8

IULA tagger

- the output is in the extended ASCII format, and not in UTF-8

IULA tagger-graph

- the output is in the extended ASCII format, and not in UTF-8

```
<fs>
<f name="word" value="proc?s"/>
<f name="lemma" value="proc?s"/>
<f name="postag" value="N5-FP"/>
</fs>
</a>
```

- Errors in the standard error stream:

```
utf8 "\xE0" does not map to Unicode at /usr3/rails-apps/hector/Preproceso//tokenizerGenerator.pm line 177.
utf8 "\xE9" does not map to Unicode at /usr3/rails-apps/hector/Preproceso//tokenizerGenerator.pm line 177.
utf8 "\xF3" does not map to Unicode at /usr3/rails-apps/hector/Preproceso//tokenizerGenerator.pm line 177.
utf8 "\xE7" does not map to Unicode at /usr3/rails-apps/hector/Preproceso//tokenizerGenerator.pm line 177.
utf8 "\xE0" does not map to Unicode at /usr3/rails-apps/hector/Preproceso//tokenizerGenerator.pm line 177.
```

IULA tokenizer

- the output is in the extended ASCII format, and not in UTF-8

```
<div1>
<p>
<s>
Consulta
diccionari
```

. DLD
</s>
<s>
Donada
una
forma
, DLD
el
servei
retorna
la
informaci?
associada
al
l?xic
. DLD
</s>
</p>
</div1>

Tools for Spanish Corpus Processing - [IULA-WebServices.zip]

Mention: we tested the tools for other examples than that in the documentation file.

IULA lexicon look up Web Wervice

The tool was :

- successfully tested for the following words: „textos", "totes", "asociada"
- unsuccesfully tested for the following words: "categoría" "información", "morfosintáctico"

  - ✓ *<LexicalEntriesList><lexicalEntry><lemma>categor□a</lemma><PoSTag>N5-FS</PoSTag></lexicalEntry></LexicalEntriesList>*
  - ✓ *<LexicalEntriesList><lexicalEntry><lemma>informaci□n</lemma><PoSTag>N5-FS</PoSTag></lexicalEntry></LexicalEntriesList>*
  - ✓ *<LexicalEntriesList><lexicalEntry><lemma>morfosint□ctic</lemma><PoSTag>JQ--MS</PoSTag></lexicalEntry></LexicalEntriesList>*

This led us to the conclusion that the output is in the extended ASCII format, and not in UTF-8.

Trying to test the web service for English too, we found the following error:

Standard error stream:
Use of uninitialized value in concatenation (.) or string at /usr3/rails-apps/hector/Preproceso/lexiconlookup.pl line 98.

IULA paradigma Web Service

- the output is in the extended ASCII format, and not in UTF-8
- in the documentation, the tool is said to be available for english too, but this option is not available

IULA processor

- the output is in the extended ASCII format, and not in UTF-8
- The tool is not able to recognize a known word if it is in its cappitalized form

Ex:
<div1>
* <s>Con m?s de <num>80.000</num>estudiantes al a?o, el <name type="organization">Instituto Cervantes</name> es la mayor instituci?n mundial dedicada a la ense?anza del espa?ol.</s><s>Recibi? el **<name>Premio</name>** Pr?ncipe de <name>Asturias</name> en <num>2005</num>en la categor?a de Comunicaci?n y <name>Humanidades</name>.</s> *
</div1>

IULA tagger

- the output is in the extended ASCII format, and not in UTF-8

IULA tagger-graph

- the output is in the extended ASCII format, and not in UTF-8

```
<a label="TOK" ref="iula-n1" as="xces">
<fs>
<f name="word" value="morfosint?ctico"/>
<f name="lemma" value="morfosint?ctico"/>
<f name="postag" value="JQ--MS"/>
</fs>
</a>
```

- Errors in the standard error stream:
  utf8 "\xE1" does not map to Unicode at /usr3/rails-apps/hector/Preproceso//tokenizerGenerator.pm line 177.
  utf8 "\xE9" does not map to Unicode at /usr3/rails-apps/hector/Preproceso//tokenizerGenerator.pm line 177.
  utf8 "\xF3" does not map to Unicode at /usr3/rails-apps/hector/Preproceso//tokenizerGenerator.pm line 177.

IULA tokenizer

- the output is in the extended ASCII format, and not in UTF-8

<div1>

<s>
Dado
un
lema
y
una
categor?a
, DLD
devuelve
las
frases

del
corpus
<name>
IULA N4666
</name>
donde
este
lema
aparece
. DLD
</s>
<s>
Se
puede
restringir
la
b?squeda
por
dominio
. DLD
</s>
</p>
</div1>

Extra Resources (new or delivered now from Batch 3)

Apertium Portuguese dictionary in LMF – [PKG-Apertium-PT.tar.gz – new resource]

- 6258 entries instead of 6259
- Xml Valid
- UTF8 encoded

Parole/Simple LMF lexicon Catalan – [PKG-SimpleParoleCatalanLMF.tar.gz – new resource]

- Xml Valid
- The number of lexical entries is corect
- What does it mean “semantic units”?
- UTF8 ecoded

SenSem Corpus – [PKG-SenSemCorpus.tar.gz – from Batch3]

- The number of words in the corpus is correct
- UTF8 econded
- The documents should validate against the GrAF DTD v.1.0.4, but the DTD mentioned is not in the resource folder

SenSem Database (lexicon) Catalan – [PKG-SenSemDataBase-CAT.tar.gz – from Batch3]

- UTF8 encoded
- sizeINFO: lexical entries number correct
- sizeINFO: semantic units number correct
- XML valid

SenSem Database (lexicon) Spanish – [PKG-SenSemDataBase-ES.tar.gz – from Batch3]

- UTF8 encoded
- sizeINFO: lexical entries number correct
- sizeINFO: semantic units number correct
- XML valid