First Upload of Language Resources

Deliverable D4.3

Version 1.1

2011-11-23

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 1.1	November 23, 2011	Georgiana Gilmeanu, Jan Joachimsen, Mike Rosner	UOM	First version
v 1.0	November 20, 2011	Georgiana Gilmeanu, Jan Joachimsen, Mike Rosner	UOM	Inclusion of new cover pages (p. 2 and 3)

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

First Upload of Language Resources

Document METANET4U-2011-D4.3
EC CIP project #270893

Deliverable D4.3
Completion: Final
Status: Submitted
Dissemination level: Public

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

Contents

1 Introduction

This deliverable deals with the steps taken by the project partners of METANET4U for the upload of Batch 1 of resources to the META-SHARE platform. The process of preparing the upload involved several phases each of which will be described in this document.

2 Infrastructure among partners

2.1 Preparation Phase

UOM as WP4 leaders asked the partners whether they would prefer to install their own nodes or whether they would rather use another option. The following possibilities were presented for discussion and feedback.

1. one central repository for all partners of METANET4U (physically implemented at the project coordinator's site, i.e., FCUL)
2. one local repository per partner.
3. some intermediate position.

In cases 2 and 3, each partner needed to decide (i) whether they wished to implement a local repository at their own site, or else (ii) if they did not wish to implement a local repository, with which local repository they would propose to partner in fulfilling the terms of D4.3.

2.2 Work plan

The first version of the META-SHARE software was released by T4ME on 5th October 2011. At the project board meeting in Berlin (October 20th, 2011), it was decided that willing partners should try to install an associated META-SHARE node (which was needed to run the metadata editor). The rationale for this was to provide information for a subsequent decision, before mid-November, concerning which partners could become associated META-SHARE nodes for Batch 1, on the understanding that this decision might be changed for the subsequent Batches 2 and 3.

UOM issued a six-week work plan to be followed until the end of M10. This plan consisted of the following steps.

1. In the first and second week (October 24th - November 4th), each partner tries to install a META-SHARE node. Technical difficulties and questions were to be directed to the email address helpdesk-technical@meta-share.eu. Also, the metadata for each resource in Batch 1 is filled out using the META-SHARE metadata editor. Technical difficulties and questions regarding metadata are directed to the email address helpdesk-metadata@meta-share.eu. Additionally, a template for a narrative description of resources supplied by RACAI is filled out and sent to RACAI for validation.
2. In the third week (November 7th-11th), partners deliver their resources to RACAI for a quick validation check. This included, for each resource, the metadata records, the narrative descriptions, and an active link to enable access the data. The data pertaining to most resources were accessed via an ftp server.
3. In the fourth week (November 14th-18th), RACAI sends warnings to partners for metadata records possibly not consistent with respective resources. Partners report progress with installation of version 1 of the META-SHARE software. They would then correct their metadata and/or resources accordingly and deliver them back to RACAI for a final check. The end of the fourth week was also the deadline for indicating which repository the partners intended to use for uploading their resources.
4. In the fifth week (November 21th-25th), partners install version 2 of the metadata editor and start to upload METADATA records for the remaining resources and also upload these resources to their nodes. RACAI releases the final quick evaluation reports for written resources and lexica and sends out warnings to partners for possible remaining resources not consistent with respective metadata records. The partners then correct their metadata records and/or remaining resources and deliver them back to RACAI for final check. By Wednesday (November 23th), UOM sends a first draft of Deliverable D4.3 for reviewing to the respective validators.
5. In the sixth week (November 28th-30th), RACAI releases the second and last quick evaluation reports for remaining resources with the general evaluation report to be finalized, released and integrated in D4.3. Partners with local repositories send final versions of the zip files containing updated metadata to UOM. UOM then submits Batch 1 by submitting the final Deliverable D4.3 and connects the associated META-SHARE nodes to the managing nodes by sending all zip files to DFKI.

2.3 Validation Process

Before the actual upload of the language resources were to be validated. The partner responsible for validating the language resources was RACAI. For the validation process, RACAI circulated a narrative description template (in Word format) originating from ELDA among the project

partners, which was discussed and finally agreed upon. All narrative descriptions are included as Appendix 2 of this document.

For each resource, this narrative description consisted of various information (e.g. size, owner details, licence agreement, content description etc.), which was filled out by the project partners and sent to RACAI accompanied by the respective resources. RACAI's task was to perform a validity check of the resources by comparing the information provided in the descriptions with the data in the resources. Due to the different nature and file formats of the language resources (e.g. corpora consisting of text files, XML files or audio in WAV format), RACAI used a variety of methods, both manual and automated, for the validation process. Overall, the validation process consisted of five laborious steps:

1. checking whether the files listed in the descriptions were actually delivered, checking their structure and space occupied,
2. checking whether the files were in the format stated in the documentation file.
3. checking XML files for well-formedness and validating them against XSD/DTD schemes either provided or internally referenced in the XML files themselves.
4. checking the values provided in the narrative descriptions against the delivered files to ensure they corresponded (e.g. counting the number of tokens, sentences, xml units, etc).
5. checking UTF8 compatibility of the files

The results of the validation for each resource received can be found in RACAI's documentation, which is included as Appendix 3 of this document.

Before transferring resources for validation, a confidentiality agreement was signed between RACAI and each project partner, stating that RACAI would use the received copies of the resources only for the purpose of validation and would afterwards destroy them.

2.4 Issues/Deviations from the work plan

1. Installing the META-SHARE software presented several problems to many partners. Contacting the provided helpdesk email addresses resulted in immediate answers, which solved the installation problems for most partners. Solving these problems, however, resulted in some delays in the work plan.
2. The first version of the metadata editor contained some errors, causing problems uploading / typing in / editing metadata; problems logging in, problems with metadata export.
3. Some features were not yet included in version 1 of the META-SHARE software, e.g. the automatic synchronisation between the metadata of the partners' META-SHARE nodes/repositories with the central META-SHARE repository. Thus the local metadata had to be

manually exported to XML from the respective repository and sent to the central repository at DFKI.

4. Several concepts were not clear among project partners, e.g.:
    a. there was confusion about the difference between uploading/sending zipped metadata xml files and uploading the actual resources
    b. there was confusion about the places where to upload the resources
    c. there was confusion as how to retrieve metadata descriptions after submission to a repository in order to send an exported XML file to the central repository
    d. it was unclear how to export the metadata of a repository to an XML file
    e. it was unclear how the physical resources relate to the storage objects inside the exported XML files
    f. it was unclear how resources that already existing in external repositories (e.g. ELRA) should be included in META-SHARE
    g. it was unclear where to see all metadata put into META-SHARE
    h. it was unclear how and when the export of metadata of ELRA-hosted resources into META-SHARE metadata should take place

Help to these issues was provided by the T4ME team, who created several helpdesk email addresses, which were frequently used by the partners:

> helpdesk-technical@meta-share.eu for any technical questions regarding setting up a local repository with the META-SHARE software,
> helpdesk-metadata@meta-share.eu for questions regarding the metadata in META-SHARE,
> helpdesk-legal@meta-share.eu for questions regarding Intellectual Property Rights and licencing issues.

Not all questions were cleared by the time of writing this deliverable, nevertheless some answers provided allowed partners to move on with the tasks so that the planned upload of resources could be achieved in time.

The answers provided were immediately forwarded among the partners and posted to the METANET4U wiki. As work was progressing in updating the META-SHARE software, project partners received news about updated versions and added features every week. This on the one hand brought clarification to unclear points but on the other hand made work along the agreed schedule slightly more difficult as it was now work towards a moving target.

Another issue, which came up shortly before the deadline for the upload of Batch 1, was that version 1 of the META-SHARE software did not restrict user access. Thus anyone entering META-SHARE could download every

resource provided on the platform without logging in. It was therefore decided to nevertheless upload all resources, but to protect the directories with restricted resources on the META-SHARE nodes by a password (which is only revealed in the present report, for partners in META-SHARE and possibly project officers and project reviewers:

The user name is `metanet4u`, the password is `batch14u`.).

Furthermore, there were discussions on the legal status of the latest versions of the licences, which were released by T4ME on Thursday, November 24th. Since these were announced by T4ME as "proposals", the question was raised whether they were legally stable enough to be used for the clearing of IPRs. T4ME assured the legal validity of these licences. Therefore they could and should be used, but could still be adjusted in upcoming versions if some of the signing parties were not happy with them.

2.5 Achievements

This section lists the projects partners' contributions for the upload of Batch 1 of resources. Apart from the individual preparation of resources and solving of legal issues, the partners had mainly to deal with setting up the META-SHARE software locally. The following list summarises the results of setting up the respective repositories.

Progress Report of Individual Partners on Installation of META-SHARE V1 Software

1. FCUL - The META-SHARE software has been successfully installed and working since the 25th of October. As of November 14th, the ULX META-SHARE node is available at http://META-SHARE.metanet4u.eu.

2. CLUL- A local version of META-SHARE is installed. The 36 resources from METANET4U.zip has been imported and can be browsed but not edited. Creation of new metadata records is not working. They decided to use FCUL's local server at http://META-SHARE.metanet4u.eu.

3. IST - Initially, a local version META-SHARE (debug mode) was installed and running from 26th of October and can be accessed at http://metanet4u.l2f.inesc-id.pt.

4. UNIMAN - A local version of META-SHARE was installed. There were some issues with metadata editing. Eventually, UNIMAN decided to use FCUL's local server at http://META-SHARE.metanet4u.eu.

5. UAIC - A local version of META-SHARE was installed and can be accessed at http://85.122.23.81/.

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

7. UOM - decided to use FCUL's repository at http://META-SHARE.metanet4u.eu. Negotiations for full support by UOM's IT Services are under way to install and maintain a local node.

8. UPC - installed version 1, but eventually did no use it, since version 1 of the metadata did not include metadata for speech resources yet. Therefore it was decided to use a repository by ELDA for uploading the resources and metadata. The URL to the repository will follow.

9. UPF: - A local version of META-SHARE was installed and can be accessed at http://193.145.50.68.

3 Upload of Batch 1 of Resources

3.1 Procedures

After installing the local repositories, partners were to send their resources, metadata and narrative descriptions to RACAI for validation. After completion of this process, partners were expected to send ZIP files with their metadata XML files to UOM. UOM collected these files and forwarded them as a whole to the central META-SHARE repository at DFKI. This step is necessary, since in the current version of the META-SHARE software, synchronisation of the metadata between all repositories has not yet been implemented in version 1.

Updated versions of Licences and Depositor's Agreements (following the discussions in Athens and Berlin) were sent around by T4ME (the latest update was received on 24/11/2011. However, it is to be noted that fine-tuning of the licence texts in order to cater for the needs of individual partners is still be done in the next update.

Filling out the metadata forms in META-SHARE was done mainly by using the web interface provided by META-SHARE. Other partners created the metadata through the use of imported XSD files provided by T4ME in their last batch of metadata schema files (for further details on these, see D4.1).

All uploaded language resources were accompanied by their narrative descriptions in pdf format in order to provide more information in human-readable form to the end user, who would browse for the resources in META-SHARE.

3.2 Resources planned to be uploaded for Batch 1

Table 1 lists the resources originally planned to be uploaded for Batch 1. It states the names the project partners, their promised resources, the kind

of each resource and the languages included in the respective resource. Note that there are some differences between the originally promised resources (Table 1) and the actually uploaded resources (Table 2 in Section 3.3). This is due to several reasons:

1. For some exogenous resources, licence negotiations between the resource owners and the project partners were not finished by the time of the upload of Batch 1. Their upload was thus postponed for Batch 2 (in the case of UOM for the Basic English-Maltese Dictionary).
2. Some resources were split up to several "sub-resources" (in the case of UPC and UPF)
3. Some partners decided to upload more resources than they originally planned for Batch 1 (in the case of UPF for Apertium Bilingual dictionary Italian-Catalan and IST for both their resources uploaded).
4. In some cases, resources were not uploaded directly into local META-SHARE nodes since they already existed in ELDA and were included in the local META-SHARE node hosted by ELDA (in the case of UNIMAN for the BioLexicon and several resources by UPC, marked in Table 2 by a footnote).

Table 1: Resources planned to be uploaded for Batch 1

Partner	Name of Resource	Resource Type	Languages covered
1. ULX - University of Lisbon	DEF Corpus	corpus, annotated corpus	Portuguese
	Multifunctional Computational Lexicon of Contemporary Portuguese - CORLEX	lexicon, frequency lexicon	Portuguese
	PF Corpus	corpus, speech	Portuguese
	Spoken Portuguese	speech database	Portuguese
	Corpus NILC	raw text corpus	Portuguese
	CorpusTCC	corpus, annotated corpus	Portuguese
	NILC Taggers	grammar, training models for tagger	Portuguese
	PLN-BR Gold	corpus, annotated	Portuguese

		corpus	
	RHETALHO	corpus, annotated corpus	Portuguese
	Summ-it	corpus, annotated corpus	Portuguese
	TeMário 2006	corpus, annotated corpus	Portuguese
2. IST - Instituto Superior Técnico			
3. UNIMAN-University of Manchester	BioLexicon	lexicon, large-scale terminological resource	English, medical
	GREC	annotated corpus	English, medical
	SemLink Resources		
	GENIA event corpus	annotated corpus, event annotation	English, medical
	GENIA	merged corpus, part-of-speech annotation and terms	English, medical
4. UAIC - University Alexandru Ioan Cuza	1984_NP	manually annotated NP chunks corpus	Romanian
	1984AnaphoraRo	annotated corpus	Romanian
	FrRoMWE	annotated French-Romanian corpus	Romanian, French
	QA-corpus-UAIC	annotated Question Answering corpus	Romanian
	RO-FDGBank	syntactic annotated	Romanian

		corpus	
	RO-FN	wordnet, FrameNet-based English-Romanian parallel corpus of semantic roles	English, Romanian
	RoSemClass	semantic classes for lexicals for political discourse analysis	Romanian
	TE-pairsResource-UAIC	grammars, collection of Text-Hypothesis pairs	Romanian
	TE-rules	grammars, collection of rules for classification of textual inferences	Romanian, English
5. RACAI - Romanian Academy	Multilingual News Corpus	written, comparable corpora	all official EU languages except Irish
	RO-Acquis	annotated corpus	Romanian
	Romanian Balanced Corpus	annotated corpus	Romanian
	RO-SemCor	parallel sense-annotated corpus	Romanian, English
	RO-WordNet (first version)	lexical ontology, semantic dictionary	Romanian
	WEB-DEX	dictionary, Romanian reference explanatory dictionary	Romanian
	Wordform lexicons	tagged and lemmatized wordform lists, lexicon	Romanian, English, French, German

	Multilingual Subjectivity Analysis: Gold Standard and Training Data	Gold Standard / Training Data Corpus	English, Romanian, Spanish
	TimeBank parallel corpus	annotated, written corpus	Romanian, English
	RO-SAM EUROM	audio corpus (WAV) and transcriptions (XML)	Romanian
6. UOM - University of Malta	F_MONA_1/Maltese Spoken Newspaper	speech data	Maltese
	Laws of Malta	raw written text corpus	Maltese, English
	Maltese Acquis Communautaire EN	raw written text corpus	English
	Maltese Acquis Communautaire MT	raw written text corpus	Maltese
	Maltese Wordlist	lexicon, wordlist	Maltese
	Basic English-Maltese Dictionary	dictionary	English, Maltese
	Illum_Corpus	raw written text corpus	Maltese
7. UPC - Technical University of Catalonia	AGORA	annotated speech database	Catalan, Spanish
	Bilingual Speech synthesis	synthetic speech, text-to-speech synthesis	Spanish, English
	CatalanBN	speech database	Catalan
	Catalan-SpeechDat	speech database	Catalan
	EUROM.1	speech database	Spanish (one dialect)
	FESTCAT	speech database	Catalan (central)
	FESTCAT-SEL	speech database	Catalan (one dialect)
	FREE-SPEECH	speech database	Catalan (one dialect)
	LC-STAR Dialogues	speech database	Spanish,

			Catalan
	Spanish Festival models	synthetic speech, text-to-speech synthesis	Spanish
	Spanish Festival voices	synthetic speech, text-to-speech synthesis	Spanish
	SpeechDat-Car Catalan	speech database	Spanish from Spain (5 dialects)
8. UPF - University Pompeu Fabra	Basic Vocabulary on the Human Genome	lexica, lexical resource - equivalents	Spanish, Catalan, English, Galician, Basque
	Corpus PAAU 92	raw text written corpus	Spanish
	Genoma corpus	raw text written corpus	Spanish, Catalan
	Multilingual Vocabulary of Economics	lexica, lexical resource - equivalents	Spanish, Catalan, English, Galician, Basque
	Neologisms of the year: Bank of Spanish and Catalan Neologisms	lexica, lexical resource	Spanish, Catalan
	UPF_Term	lexica, terminology bank	Spanish, Catalan, English, French
	PAROLE lexicon	lexica, lexicon	Catalan, Spanish
	SIMPLE lexicon	lexica, lexicon	Catalan, Spanish
	Apertium Basque dictionary	lexica, lexicon	Basque
	Apertium Bilingual dictionary Basque-Spanish	lexica, lexicon	Basque, Spanish
	Apertium Bilingual dictionary CA-ES	lexica, lexicon	Spanish, Catalan
	Apertium Bilingual dictionary	lexica, lexicon	English,

	English-Catalan		Catalan
	Apertium Bilingual dictionary English-Galician	lexica, lexicon	English, Galician
	Apertium Bilingual dictionary English-Spanish	lexica, lexicon	English, Spanish
	Apertium Bilingual dictionary French-Catalan	lexica, lexicon	French, Catalan
	Apertium Bilingual dictionary French-Spanish	lexica, lexicon	French, Spanish
	Apertium Bilingual dictionary Occitan-Catalan	lexica, lexicon	Occitan, Catalan
	Apertium Bilingual dictionary Occitan-Spanish	lexica, lexicon	Occitan, Spanish
	Apertium Bilingual dictionary Portuguese-Catalan	lexica, lexicon	Portuguese, Catalan
	Apertium Bilingual dictionary Portuguese-Galician	lexica, lexicon	Portuguese, Galician
	Apertium Bilingual dictionary Spanish-Asturian	lexica, lexicon	Spanish, Asturian
	Apertium Bilingual dictionary Spanish-Galician	lexica, lexicon	Spanish, Galician
	Apertium Bilingual dictionary Spanish-Portuguese	lexica, lexicon	Spanish, Portuguese
	Apertium Bilingual dictionary Spanish-Romanian	lexica, lexicon	Spanish, Romanian
	Apertium Catalan dictionary	lexica, lexicon	Catalan
	Apertium Galician dictionary	lexica, lexicon	Galician
	Apertium Spanish dictionary	lexica, lexicon	Spanish
	FreeLing Asturian dictionary	lexica, lexicon	Asturian
	FreeLing Catalan dictionary	lexica, lexicon	Catalan
	FreeLing Catalan sense dictionary	lexica, lexicon	Catalan
	FreeLing Galician dictionary	lexica, lexicon	Galician
	FreeLing Spanish dictionary	lexica, lexicon	Spanish
	FreeLing Spanish sense dictionary	lexica, lexicon	Spanish

3.3 Actually uploaded resources

Resources that were originally promised but eventually not uploaded in Batch 1 due to IPR issues (i.e., licence agreements were still under negotiation at the time of the upload deadline for Batch 1) were postponed for Batch 2.

Table 2: Resources actually uploaded in Batch 1

Partner	Name of Resource	Resource Type	Languages covered
1. ULX - University of Lisbon	DEF Corpus	corpus, annotated corpus	Portuguese
	Multifunctional Computational Lexicon of Contemporary Portuguese - CORLEX	lexicon, frequency lexicon	Portuguese
	PF Corpus	corpus, speech	Portuguese
	Spoken Portuguese	speech database	Portuguese
2. IST - Instituto Superior Técnico	CorpusNE	corpus	Portuguese
	CorpusParallelo	parallel corpus	Portuguese, ?
3. UNIMAN-University of Manchester	BioLexicon[1]	lexicon, large-scale terminological resource	English, medical
	GREC	annotated corpus	English, medical
	SemLink Resources		
	GENIA event corpus	annotated corpus, event annotation	English, medical
	GENIA	merged corpus, part-of-speech annotation and terms	English, medical
4. UAIC - University Alexandru Ioan	1984_NP	manually annotated NP	Romanian

[1] hosted at ELDA

Cuza		chunks corpus	
	1984AnaphoraRo	annotated corpus	Romanian
	FrRoMWE	annotated French-Romanian corpus	Romanian, French
	QA-corpus-UAIC	annotated Question Answering corpus	Romanian
	RO-FDGBank	syntactic annotated corpus	Romanian
	RO-FN	wordnet, FrameNet-based English-Romanian parallel corpus of semantic roles	English, Romanian
	RoSemClass	semantic classes for lexicals for political discourse analysis	Romanian
	TE-pairsResource-UAIC	grammars, collection of Text-Hypothesis pairs	Romanian
	TE-rules	grammars, collection of rules for classification of textual inferences	Romanian, English
5. RACAI - Romanian Academy	Multilingual News Corpus	written, comparable corpora	all official EU languages except Irish
	RO-Acquis	annotated corpus	Romanian
	Romanian Balanced Corpus	annotated corpus	Romanian
	RO-SemCor	parallel sense-annotated	Romanian,

		corpus	English
	RO-WordNet (first version)	lexical ontology, semantic dictionary	Romanian
	WEB-DEX	dictionary, Romanian reference explanatory dictionary	Romanian
	Wordform lexicons	tagged and lemmatized wordform lists, lexicon	Romanian, English, French, German
	Multilingual Subjectivity Analysis: Gold Standard and Training Data	Gold Standard / Training Data Corpus	English, Romanian, Spanish
	TimeBank parallel corpus	annotated, written corpus	Romanian, English
	RO-SAM EUROM	audio corpus (WAV) and transcriptions (XML)	Romanian
6. UOM - University of Malta	Laws of Malta	raw written text corpus	Maltese, English
	Maltese Acquis Communautaire EN	raw written text corpus	English
	Maltese Acquis Communautaire MT	raw written text corpus	Maltese
	Maltese Wordlist	lexicon, wordlist	Maltese
	Illum_Corpus	raw written text corpus	Maltese
7. UPC - Technical University of Catalonia	AGORA	annotated speech database	Catalan, Spanish
	Bilingual Speech synthesis[2]	synthetic speech, text-to-speech synthesis	Spanish, English
	CatalanBN	speech database	Catalan

[2] hosted at ELDA

	Catalan-SpeechDat[2]	speech database	Catalan
	EUROM.1[2]	speech database	Spanish (one dialect)
	FESTCAT[2]	speech database	Catalan (central)
	FESTCAT-SEL[2]	speech database	Catalan (one dialect)
	FREE-SPEECH	speech database	Catalan (one dialect)
	LC-STAR Dialogues[3]	speech database	Spanish, Catalan
	Spanish Festival models[2]	synthetic speech, text-to-speech synthesis	Spanish
	Spanish Festival voices[2]	synthetic speech, text-to-speech synthesis	Spanish
	SpeechDat-Car Catalan[2]	speech database	Spanish from Spain (5 dialects)
8. UPF - University Pompeu Fabra	Basic Vocabulary on the Human Genome	lexica, lexical resource - equivalents	Spanish, Catalan, English, Galician, Basque
	Corpus PAAU 92	raw text written corpus	Spanish
	Genoma corpus	raw text written corpus	Spanish, Catalan
	Multilingual Vocabulary of Economics	lexica, lexical resource - equivalents	Spanish, Catalan, English, Galician, Basque
	Neologisms of the year: Bank of Spanish and Catalan Neologisms	lexica, lexical resource	Spanish, Catalan
	UPF_Term	lexica, terminology	Spanish, Catalan,

[3] split up into the three databases *TALP Tourism Dialogues Spanish, Catalan* and *Translation*

		bank	English, French
	PAROLE lexicon	lexica, lexicon	Catalan, Spanish
	SIMPLE lexicon	lexica, lexicon	Catalan, Spanish
	Apertium Basque dictionary	lexica, lexicon	Basque
	Apertium Bilingual dictionary Basque-Spanish	lexica, lexicon	Basque, Spanish
	Apertium Bilingual dictionary CA-ES	lexica, lexicon	Spanish, Catalan
	Apertium Bilingual dictionary English-Catalan	lexica, lexicon	English, Catalan
	Apertium Bilingual dictionary English-Galician	lexica, lexicon	English, Galician
	Apertium Bilingual dictionary English-Spanish	lexica, lexicon	English, Spanish
	Apertium Bilingual dictionary French-Catalan	lexica, lexicon	French, Catalan
	Apertium Bilingual dictionary French-Spanish	lexica, lexicon	French, Spanish
	Apertium Bilingual dictionary Occitan-Catalan	lexica, lexicon	Occitan, Catalan
	Apertium Bilingual dictionary Occitan-Spanish	lexica, lexicon	Occitan, Spanish
	Apertium Bilingual dictionary Portuguese-Catalan	lexica, lexicon	Portuguese, Catalan
	Apertium Bilingual dictionary Portuguese-Galician	lexica, lexicon	Portuguese, Galician
	Apertium Bilingual dictionary Spanish-Asturian	lexica, lexicon	Spanish, Asturian
	Apertium Bilingual dictionary Spanish-Galician	lexica, lexicon	Spanish, Galician
	Apertium Bilingual dictionary Spanish-Portuguese	lexica, lexicon	Spanish, Portuguese
	Apertium Bilingual dictionary Spanish-Romanian	lexica, lexicon	Spanish, Romanian

	Apertium Catalan dictionary	lexica, lexicon	Catalan
	Apertium Galician dictionary	lexica, lexicon	Galician
	Apertium Spanish dictionary	lexica, lexicon	Spanish
	FreeLing Asturian dictionary	lexica, lexicon	Asturian
	FreeLing Catalan dictionary	lexica, lexicon	Catalan
	FreeLing Catalan sense dictionary	lexica, lexicon	Catalan
	FreeLing Galician dictionary	lexica, lexicon	Galician
	FreeLing Spanish dictionary	lexica, lexicon	Spanish
	FreeLing Spanish sense dictionary	lexica, lexicon	Spanish

As the comparison of the two tables shows, most of the project partners delivered exactly the foreseen resources. Slight deviations are as follows.

ULX postponed the following resources for deliverable in Batch 2 due to IPR negotiations that were still in progress at the time of the deadline for Batch 1:

Corpus NILC
CorpusTCC
NILC Taggers
PLN-BR Gold
RHETALHO
Summ-it
TeMário 2006

IST originally had no resources planned for Batch 1 but delivered two corpora: CorpusNE and CorpusParallelo.

UOM postponed two resources to delivery for Batch 2:

F_MONA_1/Maltese Spoken Newspaper (due to some additional work on the resource undertaken by its author)
Basic English-Maltese Dictionary (due to ongoing IPR negotiations)

UPC split up its database *LC-STAR Dialogues* into the three databases called

TALP Tourism Dialogues - Spanish
TALP Tourism Dialogues - Catalan
TALP Tourism Dialogues – Translation

The following resources by UPC were hosted at ELDA, but nevertheless count as resources delivered for Batch 1:

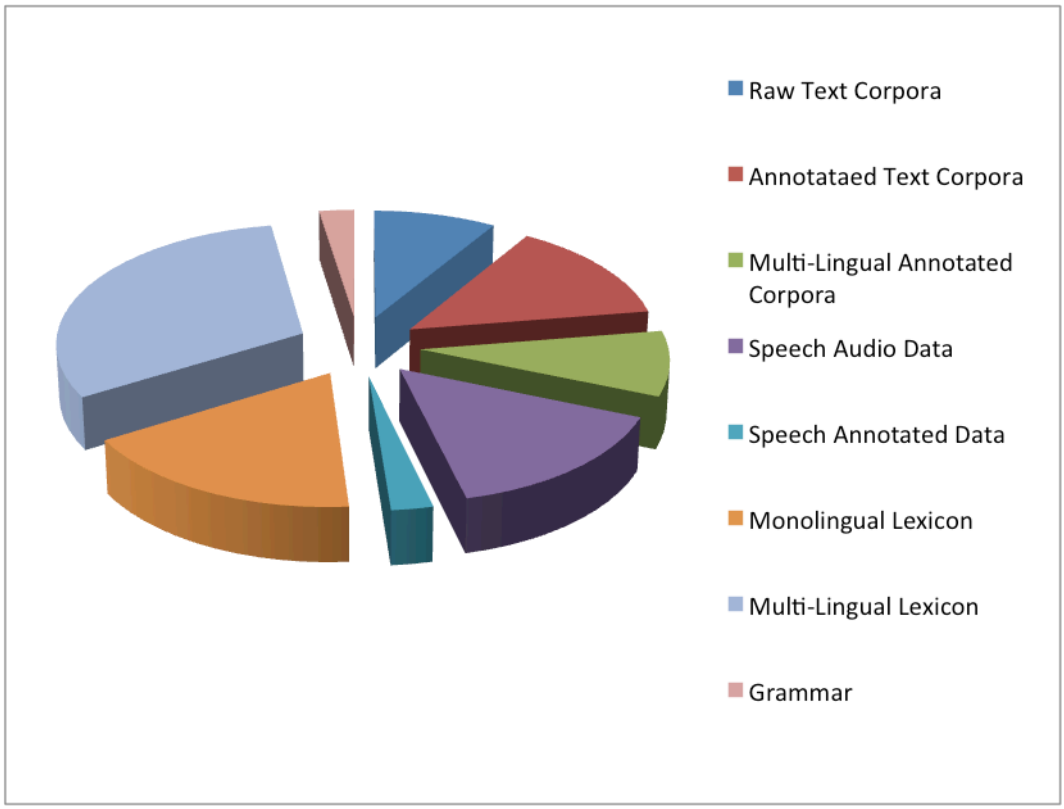
Bilingual (Spanish English) Speech synthesis HTS models.
Catalan-SpeechDat For the Fixed Telephone Network Database
Catalan-SpeechDat for the Mobile Telephone Network Database
Spanish EUROM.1
FESTCAT Catalan TTS Baseline male 10h
FESTCAT Catalan TTS Baseline female 10h
FESTCAT-SEL Catalan TTS Baseline 8 spks x 1h
Spanish Festival HTS models male speech
Spanish Festival HTS models female speech
Spanish Festival voice male
Spanish Festival voice female
SpeechDat-Car Catalan

UNIMAN's BioLexicon was hosted at ELDA, but nevertheless counts as resources delivered for Batch 1.

In order to visualise the composition by resource type of the present upload, we originally planned to present a pie-chart based on the corresponding attribute appearing in deliverable 2.4. However, with 41 different types, the result was difficult to interpret. We therefore distilled the original types into the following eight categories for the purposes of visualisation:

1. raw text corpora
2. annotated text corpora
3. multlilingual annotated text corpora
4. speech audio data
5. speech annotated data
6. monolingual lexicons
7. multilingual lexicons
8. grammars

Collapsed to 8 main categories, the distribution of resource types can be summarised in the following pie chart.



4 Conclusions

Even though there were slight changes in the actual uploads of Batch 1 (mostly for IPR reasons mentioned earlier), the overall aims were fulfilled. Installing and working with version 1 of the META-SHARE software was a "training" phase for the project partners. Due to updates in the software and IPR licences during the preparation phase, a working routine was developing while the upload for Batch 1 was prepared. The process was fruitful, and experience gained from it will no doubt be useful for the preparation and upload of the resources for Batch 2.

5 References

Moreno, Asunción (2011): *D2.4 – Report on methodology and criteria followed for the selection of resources.*

Tufiş, Dan (2011): *Narrative Descriptions for the Resources delivered as BATCH 1.*

Tufiş, Dan (2011): *D3.1 - WP3: Delivery of the BATCH 1 of Resources Validation Report*

6 Appendix

6.1 Appendix 1: Narrative Description Template

DEAR PARTNERS,

Below please find suggestions for the structuring of the resource documentation (per type). These suggestions follow the suggested SR documentation sent yesterday.

LEXICA DOCUMENTATION

1. BASIC INFORMATION
   - *1.1 Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc),*
   - *1.2 Representation of the lexicon (flat files, database, markup)*
   - *1.3 Character encoding*
2. ADMINISTRATIVE INFORMATION
   - *2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)*
   - *2.2 Delivery medium (if relevant; description of the content of each piece of medium)*
   - *2.3 Copyright statement and information on IPR*
3. TECHNICAL INFORMATION
   - *3.1 Directories and files*
   - *3.2 Data structure of an entry*
   - *3.3 Lexicon size (nmb. of lexical items, KB occupied on disk)*
4. CONTENT INFORMATION
   - *4.1 The natural language(s) of the lexicon*
   - *4. 2 Entry Type*
   - *4.3 Attributes and their values*
   - *4.4 Coverage of the lexicon*
   - *4.5 Intended application of the lexicon*
   - *4.6 POS assignment*
   - *4.7 Reliability (automatically/manually constructed)*
5. RELEVANT REFERENCES AND OTHER INFORMATION

CORPORA DOCUMENTATION (including parsed corpora)

1 BASIC INFORMATION
   - *1.1 Corpus composition*
   - *1.2 Representation of the corpora (flat files, database, markup)*
   - *1.3 Character encoding*

2 ADMINISTRATIVE INFORMATION
   - *2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)*
   - *2.2 Delivery medium (if relevant; description of the content of each piece of medium)*
   - *2.3 Copyright statement and information on IPR*

3 TECHNICAL INFORMATION
   - *3.1 Directories and files*
   - *3.2 Data structure of an entry*
   - *3.3 Corpora size (nmb. of tokens, MB occupied on disk)*

4 CONTENT INFORMATION
   - *4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)*
   - *4.2 The natural language(s) of the corpus*
   - *4. 3 Domain(s)/register(s) of the corpus*
   - *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)*
     - *4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),*
     - *4.4.3Alignment information (if the corpus contain aligned documents: level of alignment, how it was achieved)*
     - *4.4.4 Attributes and their values (if annotated)*
   - *4.5 Intended application of the corpus*
   - *4.6 Reliability of the annotations (automatically/manually assigned) – if any*

5 RELEVANT REFERENCES AND OTHER INFORMATION

For other types of resources (e.g. grammars, TE resources) one may use something similar to corpora.

1 BASIC INFORMATION
- *1.1 Resource composition*
- *1.2 Representation of the resource (flat files, database, markup)*
- *1.3 Character encoding*

2 ADMINISTRATIVE INFORMATION
- *2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)*
- *2.2 Delivery medium (if relevant; description of the content of each piece of medium)*
- *2.3 Copyright statement and information on IPR*

3 TECHNICAL INFORMATION
- *3.1 Directories and files*
- *3.2 Data structure of an entry*
- *3.3 Resource size (nmb. of rules, MB occupied on disk)*

4 CONTENT INFORMATION
- *4.1 Type of the resource (language (in(dependant)*
- *4.2 The natural language(s) for the resource is applicable (if language dependent)*
- *4. 3 Domain(s)/register(s) of the corpus*
- *4.4 Annotations in the corpus (if an annotated corpus)*
  - *4.4.1 Types of annotations*
  - *4.4.2 Tags (if POS/MSD/TIME/discourse/etc –tagged or parsed),*
  - *4.4.3 Attributes and their values (if annotated)*
- *4.5 Intended application of the resource*
- *4.6 Reliability of the annotations (automatically/manually assigned) – if any*

5 RELEVANT REFERENCES AND OTHER INFORMATION

6.2 Appendix 2: Complete Narrative Descriptions

WP3: Narrative Descriptions for the Resources delivered as BATCH 1

Document METANET4U-2011-D3.1

EC CIP project #270893

Responsible: Dan Tufiș (WP3 coordinator)

Contributors: all the WP3 participants

28 November 2011

Abstract

This document is a collation of the resource descriptions provided by the partners involved in WP3. These descriptions were used as a basis for the quick quality check described in the D3.1 Report.

The narratives are presented clustered by the institutions involved in the WP3.

Contents

Partner ULX

Português Fundamental Corpus

1 BASIC INFORMATION

1.1 Corpus composition

Português Fundamental is a corpus of spoken language, collected between 1970 and 1974, composed of 1800 recordings (500 hours) made in Continental Portugal and the Islands. Of these 1800 conversations, a sample was selected and transcribed.

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. The plain text files also have automatically assigned POS-tag information.

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 (pos) tagged with POS information.

2.3 Copyright statement and information on IPR

The resource is free license-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

Português Fundamental has 137 wav files, 137 exb files, 137 txt files and 137 pos files, corresponding to the audio files, the transcriptions and the plain text files tagged with POS information.

3.2 Data structure of an entry

This is not relevant as the corpus is provided as an EXMARaLDA file. Each transcription file is structured in utterances, containing one or more sentences.

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

The corpus contains 153,588 tokens and needs about 4.7 GB for disk storage for the wav, the exb, and the txt files, and about 1.32 MB for the html 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

The language of the corpus is European Portuguese.

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

The corpus was recorded in a situation of spontaneous oral communication, on different themes of everyday life, with speakers of different ages and social and professional backgrounds.

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 annotation of the corpus was made at the utterance level and includes discourse mark-up, like pauses, hesitations, reformulations, extralinguistic elements, speaker overlapping, etc. 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.

4.4.3Alignment information (if the corpus contain 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 a brief pause and/or a tonal unit boundary. This symbol is also used to separate hesitation indicators and extralinguistic elements within an 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.

The following tags were applied in the POS-tagging.

POS codification	
Tag	**Category**
ADJ	Adjectives
ADV	Adverbs
CARD	Cardinals
CJ	Conjunctions
CL	Clitics
CN	Common Nouns
DA	Definite Articles
DEM	Demonstratives
DFR	Denominators of Fractions
DGTR	Roman Numerals
DGT	Digits
DM	Discourse Marker
EADR	Electronic Addresses
EOE	End of Enumeration
EXC	Exclamatives
GER	Gerunds
GERAUX	Gerunds as auxiliary verbs
IA	Indefinite Articles

IND	Indefinites
INF	Infinitive
INFAUX	Infinitive auxiliary verb
INT	Interrogatives
ITJ	Interjection
LTR	Letters
LADV1…LADVn	Latin Multi-Word Adverbs
MGT	Magnitude Classes
MTH	Months
NP	Noun Phrases
ORD	Ordinals
PADR	Part of Address
PNM	Part of Name
PNT	Punctuation Marks
POSS	Possessives
PPA	Past Participles not in compound tenses
PP	Prepositional Phrases
PPT	Past Participle in compound tenses
PREP	Prepositions
PRS	Personals
QNT	Quantifiers
REL	Relatives
STT	Social Titles
SYB	Symbols
TERMN	Optional Terminations
UM	"um" or "uma"

UNIT	Measurement units in abbreviated form
VAUX	Finite "ter" or "haver" in compound tenses
V	Verbs (other than PPA, PPT, INF or GER)
WD	Week Days
LADV1…LADVn	Multi-Word Adverbs
Contracted forms	**Combinations of :**
CL+CL	Two clitics
PREP+ADV	Preposition and Adverb
PREP+DA	Preposition and Definite Articles
PREP+DEM	Preposition and Demonstratives
PREP+IND	Preposition and Indefinite
PREP+INT	Preposition and Interrogative
PREP+PRS	Preposition and Personal pronoun
PREP+QNT	Preposition and Quantifier
PREP+REL	Preposition and Relative
PREP+UM	Preposition and "um" or "uma"

Regarding the metadata, each file has meta information 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	Values
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	**Values**
Name	
Age	
Geographical origin	
Residence	
Education	Illiterate Primary school Middle school High school University students Graduated
Profession	
Linguistic influence	
Role	Interviewer Informant

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

AA.VV. (1984), *Português Fundamental, Vocabulário e Gramática*, tomo 1, Vocabulário, Lisboa, INIC.

Bacelar do Nascimento, M. F., M. L. Garcia Marques e M. L: Segura da Cruz (1987), *Português Fundamental, Métodos e Documentos Português Fundamental, Vocabulário e Gramática,* tomo 1, Inquérito de Frequência, Lisboa, INIC, CLUL.

Bacelar do Nascimento, M. F., P. Rivenc, M.L. Segura da Cruz (1987), *Português Fundamental, Métodos e Documentos*, tomo 2, Inquérito de Disponibilidade, Lisboa, INIC, CLUL.

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 5th International Conference on Language Resources and Evaluation (LREC2006), Genoa, Italy.

Spoken Portuguese Corpus

1 BASIC INFORMATION

1.1 Corpus composition

The Spoken Portuguese corpus was collected among sociolinguistically diverse speakers having Portuguese as mother tongue or as second language. In a total of 86 recordings, the texts exemplify the Portuguese spoken in Portugal (30), in Brazil (20), in the African countries with Portuguese as its official language: Angola, Cape Verde, Guinea-Bissau, Mozambique and Sao Tome and Principe (5 each), in Macao (5), in Goa (3) and in East-Timor (3), corresponding to a total of 8h44m of recording. The recordings cover a period that goes from 1970 to 2001, and approximately 70% of them fall within the nineties.

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. The plain text files also have automatically assigned POS-tag information.

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 (pos) tagged with POS information.

2.3 Copyright statement and information on IPR

The resource is free license-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 Spoken Portuguese corpus has 86 wav files, 86 exb files, 86 txt files, and 86 pos files, corresponding to the audio files, the transcriptions and the plain text files 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 (available at http://www.exmaralda.org/en_downloads.html) for details.

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

The corpus contains 128,542 tokens and needs about 1.29 GB for disk storage for the wav, the exb, and the txt 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 informal conversations between acquaintances, friends or relatives as well as formal acts as, for instance, radio programs or conferences.

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. 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 contain 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.

The following tags were applied in the POS-tagging.

POS codification	
Tag	**Category**
ADJ	Adjectives
ADV	Adverbs
CARD	Cardinals
CJ	Conjunctions
CL	Clitics
CN	Common Nouns
DA	Definite Articles
DEM	Demonstratives
DFR	Denominators of Fractions

DGTR	Roman Numerals
DGT	Digits
DM	Discourse Marker
EADR	Electronic Addresses
EOE	End of Enumeration
EXC	Exclamatives
GER	Gerunds
GERAUX	Gerunds as auxiliary verbs
IA	Indefinite Articles
IND	Indefinites
INF	Infinitive
INFAUX	Infinitive auxiliary verb
INT	Interrogatives
ITJ	Interjection
LTR	Letters
LADV1…LADVn	Latin Multi-Word Adverbs
MGT	Magnitude Classes
MTH	Months
NP	Noun Phrases
ORD	Ordinals
PADR	Part of Address
PNM	Part of Name
PNT	Punctuation Marks
POSS	Possessives
PPA	Past Participles not in compound tenses
PP	Prepositional Phrases

PPT	Past Participle in compound tenses
PREP	Prepositions
PRS	Personals
QNT	Quantifiers
REL	Relatives
STT	Social Titles
SYB	Symbols
TERMN	Optional Terminations
UM	"um" or "uma"
UNIT	Measurement units in abbreviated form
VAUX	Finite "ter" or "haver" in compound tenses
V	Verbs (other than PPA, PPT, INF or GER)
WD	Week Days
LADV1…LADVn	Multi-Word Adverbs
Contracted forms	**Combinations of :**
CL+CL	Two clitics
PREP+ADV	Preposition and Adverb
PREP+DA	Preposition and Definite Articles
PREP+DEM	Preposition and Demonstratives
PREP+IND	Preposition and Indefinite
PREP+INT	Preposition and Interrogative
PREP+PRS	Preposition and Personal pronoun
PREP+QNT	Preposition and Quantifier
PREP+REL	Preposition and Relative
PREP+UM	Preposition and "um" or "uma"

Regarding the metadata, each exb file has meta information 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

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, F. (2001), (coord.) *Português Falado, Documentos Autênticos*, Gravações áudio com transcrições alinhadas, em CD-ROM, Lisboa, Centro de Linguística da Universidade de Lisboa e Instituto Camões.

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.

Bacelar do Nascimento, M. F. (2001), "Les études portugaises sur la langue parlée", in CARREIRA, M. H. A. (org.) *Travaux et Documents, Les langues romanes en dialogue(s)*, 11-2001, Université Paris 8, Vincennes Saint-Denis, pp. 209-221.

Bacelar do Nascimento, M. F. et alii (2001), Poster "Português Falado", in *Feira de Projectos*, promovida pela Comissão Nacional do Ano Europeu das Línguas, Lisboa, Centro Cultural Casapiano, 27-30 de Setembro.

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 5th International Conference on Language Resources and Evaluation (LREC2006), Genoa, Italy.

Bettencourt Gonçalves, J. (2000), "Português Falado: variedades geográficas e sociais", in *Estudos de gramática portuguesa (1)* Eberhard Gärtner, Christine Hundt, Axel Schönberger (eds.), Frankfurt am Main.

Bettencourt Gonçalves, J. e R. Veloso (2000), "Spoken Portuguese: Geographic and Social Varieties", in *Proceedings of the Second International Conference on Language Resources and Evaluation, Volume II*, National technical University of Athens Press, Athens, Greece, pp. 905-908

Pereira, L. A. S. (2004), "The use of concordancing in Portuguese teaching". In Sinclair, J. M. (ed.) *How to Use Corpora in Language Teaching*. Amsterdam, John Benjamins P. C.: 109-122.

Portuguese Definitions Corpus

I. Basic Information

1.1. Corpus information

The corpus presented here is a collection of several tutorials and scientific papers in the field of Information Technology with 603 annotated definitions from Portuguese. The texts were collected from the Web at the beginning of the 2006 and they are organized in three folders with 32 files of three different sub-domains with 268,064 tokens: Information Society (91,825), Information Technology (80,483), and e-Learning (94,756).

In this corpus, a definition is assumed to be a sentence containing an expression (the *definiendum*) and its definition (the *definiens*) and a connector between them. We identify three different tipology definitions corresponding to three different connectors, that is the verb "to be" ("ser"), all other verbs other than "to be" and punctuation mark such as ":", finally a last class, covering all definitions not covered by the previous classification (see Del Gaudio, 2007c and 2009a). The following table displays the distribution of the different types of definitions in the corpus.

Type	*IS*	*IT*	*e-Learning*	***Total***
is_def	68	40	17	125
verb_def	80	77	66	223
punct_def	9	89	35	133
other_def	31	52	39	122
Total	**188**	**258**	**157**	**603**

This corpus was collected in the context of Language Technologies for eLearning project (www.lt4el.eu) founded by European Union whose main goal is to improve e-Learning systems by using multilingual language technology tools and semantic web techniques.

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

The corpus is represented in a variant of the XCES format described by DTD file (see LT4ELAnaProjectv3.4.dtd).

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

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

III. Technical Information

3.*1. Directories and files*

The archive that can be uploaded on the Meta-Share is a .zip file with 33 files: 32 XML and 1 DTD.

3.2. Data structure of an entry

For each text file with a set of sentences, the data is divided into paragraphs, which has the respective sentences segmented by tokens.

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

The corpus is composed by 268,064 tokens with 2.7 *MB* compressed (25.6 *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 annotated corpus.

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

The language of the corpus is Portuguese with pre-spelling reform of 1990[1].

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

Concerning the Information domain, there are three sub-domains in this corpus: Information Society, Information Technology for non-experts, and e-Learning.

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 corpus was pre-processed in order to convert it into a common XML format, conforming to a DTD derived from the XCES DTD for linguistically annotated corpora (see Ide and Suderman, 2002). The following example shows more detailed the structure of a DTD format, containing the lemma of each word, the attribute *ctag*, the POS information, and the *msd* with the morpho-syntactic inflection.

1 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.

```
−<s id="s205">
 −<definingText def="k1" def_type1="is_def" id="d1">
  −<markedTerm dt="y" id="k1" kw="y">
     <tok base="xml" class="word" ctag="PNM" id="t1724" sp="y">XML</tok>
   </markedTerm>
  −<connector>
     <tok base="ser" class="word" ctag="V" id="t1725" msd="pi-3s" sp="y">é</tok>
   </connector>
   <tok base="um" class="word" ctag="UM" id="t1726" msd="ms" sp="y">um</tok>
   <tok base="mecanismo" class="word" ctag="CN" id="t1727" msd="ms" sp="y">mecanismo</tok>
   <tok base="ou" class="word" ctag="CJ" id="t1728" sp="y">ou</tok>
   <tok base="_" class="word" ctag="PNT" id="t1729" msd="?" sp="y">"</tok>
  −<markedTerm id="z114" kw="y">
     <tok base="metalinguagem" class="word" ctag="CN" id="t1730" msd="fs">metalinguagem</tok>
   </markedTerm>
   <tok class="punctuation" ctag="PNT" id="t1731" sp="y">"</tok>
   <tok base="para" class="word" ctag="PREP" id="t1732" sp="y">para</tok>
   <tok base="criar" class="word" ctag="V" id="t1733" msd="inf-nInf" sp="y">criar</tok>
   <tok base="linguagem" class="word" ctag="CN" id="t1734" msd="fp" sp="y">linguagens</tok>
   <tok base="marcar,marcado" class="word" ctag="PPA" id="t1735" msd="fp" sp="y">marcadas</tok>
   <tok base="com" class="word" ctag="PREP" id="t1736" sp="y">com</tok>
   <tok base="finalidade" class="word" ctag="CN" id="t1737" msd="fp" sp="y">finalidades</tok>
   <tok base="especial" class="word" ctag="ADJ" id="t1738" msd="fp">especiais</tok>
   <tok class="punctuation" ctag="PNT" id="t1739" sp="y">.</tok>
  </definingText>
 </s>
```

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

The corpus was automatically annotated with morpho-syntactic information using the LX-Suite[2] (see Silva, 2007). This is a set of tools for the shallow processing of Portuguese with state of the art performance. This pipeline of modules comprises several tools, namely a sentence chunker (99.94% F-score), a tokenizer (99.72%), a POS tagger (98.52%), and nominal and verbal featurizers (99.18%), and lemmatizers (98.73%).

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)

In each sentence with a definition structure, the term defined, the definition, and the connector were manually annotated using a different XML tag, *markedTerm*, *connector* (only for copula definitions), and *markedTerm*, respectively, with the assignment of the information about the type of definition. The definition typology is made of four different classes whose members were tagged with *is_def*, for copula definitions, *verb_def*, for verbal non copula definitions, *punct_def*, for definitions whose connector is a punctuation mark, and finally *other_def*, for all the remaining definitions.

4.5. Intended application of the corpus

Concerning that the main goal of this corpus was to test a tool for supporting glossary construction in a automatic way in e-Learning management systems for Portuguese (see Del Gaudio, 2009b and 2009c), it also

2 Available at http://lxcenter/services/en/LXServicesSuite.html.

compose a reference corpus for various comparative analysis in specialized language for Portuguese and between languages. This definitions corpus is also important in the context of Question Answering (QA), ontology learning, dictionary and glossary construction, among others.

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

Firstly, the corpus was automatically annotated with LX-Suite tools with high accuracy (see 4.2.). In a second phase, human experts annotators marked the definitions structures.

V. Relevant References and Other Information

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.

Ide, Nancy and Keith Suderman, XML, Corpus Encoding Standard, Document XCES 0.2. Technical Report, Department of Computer Science, Vassar College and Equipe Langue ed Dialogue, New York, USA and LORIA/CNRS, Vandouvre-les-Nancy, France, 2002.

Del Gaudio, Rosa and António Branco, 2007b, "Automatic Extraction of Definitions in Portuguese: A Rule-Based Approach". In J. Neves, M. Santos and J. Machado (eds.), *EPIA2007 - 13th Portuguese Conference on Artificial Intelligence*, LNAI 4874, Berlin, Springer, pages 659-670.

Del Gaudio, Rosa and António Branco, 2007c, "Supporting e-Learning with Automatic Glossary Extraction: Experiments with Portuguese". In *Proceedings of the Workshop on Natural Language Processing and Knowledge Representation for e-Learning Environments*, RANLP2007 - International Conference on Recent Advances in Natural Language Processing.

Del Gaudio, Rosa and António Branco, 2009a, "Evaluating a Learning Management System improved with Language Technology". In *Proceeding of the 12th International Conference Interactive Computer Aided Learning (ICL)*. Villach, Austria.

Del Gaudio, Rosa and António Branco, 2009b, "Extraction of Definitions in Portuguese: An Imbalanced Data Set Problem". In *Proceedings of Text Mining and Applications (TEMA 2009)*, pages 501-512.

Del Gaudio, Rosa and António Branco, 2009c, "Improving e-Learning Experience with Language Technology: Evaluation Results". In *Proceeding of the International Conference Interactive Computer Aided Blended Learning (ICBL)*. Florianopolis, Brazil.

Multifunctional Computational Lexicon of Contemporary Portuguese

1. BASIC INFORMATION

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

The Multifunctional Computational Lexicon of Contemporary Portuguese (MCL) was extracted from CORLEX. CORLEX is a subcorpus of the Reference Corpus of Contemporary Portuguese (CRPC) which contains a spoken subcorpus (of 856.195 words) and a written subcorpus (of 15.354.243 words). CORLEX contains written and spoken texts of several genres, this diversity is a characteristic of this corpus. CORLEX is composed mainly by journalistic texts (56% of the written subcorpus and 53% of the whole corpus). The spoken corpus contains orthographic transcriptions of informal conversations and more formal productions like conferences, interviews in the radio and TV, etc. Regarding the written subcorpus, CORLEX is also composed by literary texts (20%), Techno-scientific texts (20%) and varia (4%).The MCL has 26.443 lemma and 140.315 tokens, with the minimum lemma frequency of 6.

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

The corpus is represented in txt and pdf 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 MCL is available in two different formats, txt (for evaluation purposes) and pdf (for consultation). The resource contains 2 files in txt format, one corresponding to an indexation, with numerical frequency, by alphabetical order, and the other corresponding to an indexation, with numerical frequency, by decreasing frequency order; 26 pdf files with the corpus organized by alphabetical order, and 12 pdf files with the corpus organized by decreasing frequency order.

3.2 Data structure of an entry

Each lemma and wordform corresponds to an entry in the database and it is marked with information regarding the POS categorization and frequency. As shown in the examples below, in the txt and the pdf files, each lemma entry is introduced by the symbol "@", and every lemma and wordform has the POS category information between brackets. In the txt files the frequency information is preceded by the symbol "#". In the pdf files, on the other hand, the frequency of lemma and wordforms was calculated using a logarithmic scale ($\log_{10}/2$), in order to obtain a more homogeneous distribution of the quantitative data, and is coded as follows:

Frequency level ($\log_{10}/2$):

Lemma:		**Tokens:**	
6 - 10	◪□□□□□	0 - 5	○○○○○○
11 - 31	■□□□□□	6 - 10	◐○○○○○
32 - 100	■◪□□□□	11 - 31	●○○○○○
101 - 316	■■□□□□	32 - 100	●◐○○○○
317 - 1.000	■■◪□□□	101 - 316	●●○○○○
1.001 - 3.162	■■■□□□	317 - 1.000	●●◐○○○
3.163 - 10.000	■■■◪□□	1.001 - 3.162	
10.001 - 31.622	■■■■□□	3.163 - 10.000	●●●◐○○
31.623 - 100.000	■■■■◪□	10.001 - 31.622	●●●●○○
100.001 - 316.227	■■■■■□	31.623 - 100.000	●●●●◐○
316.228 - 1.000.000	■■■■■◪	100.001 - 316.227	●●●●●○
1.000.001 - 3.162.277	■■■■■■	316.228 - 1.000.000	●●●●●◐

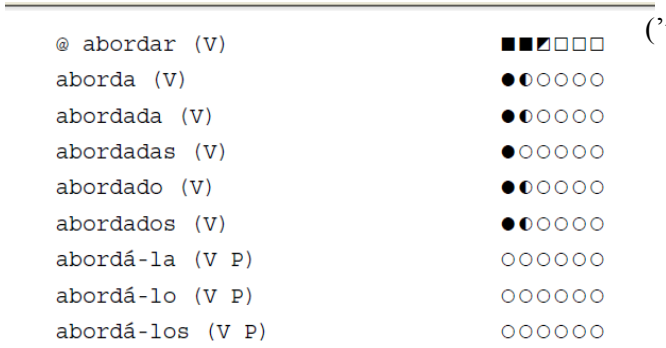
The structure of an entry in MCL is exemplified below:

Txt file (organized by alphabetical order)

@ abordar (V) # 724 ('to approach')
aborda (V) # 77
abordada (V) # 56

abordadas (V) # 31
abordado (V) # 83
abordados (V) # 60
abordá-la (V P) # 1
abordá-lo (V P) # 1
abordá-los (V P) # 2

Pdf file (organized by alphabetical order)



('to approach')

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

The lexicon has 26.443 lemmas and 140.315 tokens, and needs 6,38 Mb for disk storage.

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

The language of the corpus is European Portuguese.

4. 2 Entry Type

There are two types of entries, both of them having the same structure: entries for lemmas and entries for wordforms.

4.3 Attributes and their values

The lemmas and tokens are marked with the following codes:

POS codification	
Noun	N
Verb	V
Adjective	A
Pronoun and Adjunct Pronoun	P
Article	T
Adverb	R
Adposition	S
Conjunction	C
Numeral	M

Interjection	I
Foreign word	F
Abbreviation	X
Acronym / Sigla	G
Symbol	B
Mediopassive *Se*	U
Locution element	L
Emphatic particle	E
Other codification	
Displaced element	_d
Non-conventional orthography	*
Contraction	+
Lemma	@
Reconstructed lemma	[]
Reconstructed token	<>

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

The lexicon can be used in linguistic research and for improving and developing numerous kinds of Natural Language Processing tools and applications, such as morphological and syntactic taggers, lemmatization, word-sense disambiguation or machine translation.

4.6 POS assignment

All wordforms included in the MCL were automatically tagged (morphosyntactic tagging) and lemmatized, and then a manual verification of all the tags attributed to each wordform and lemma (with the minimum lemma frequency of 6) was made. The criteria followed in this verification were the same used in the Português Fundamental project (cf. Bacelar do Nascimento et al., 1987).

4.7 Reliability (automatically/manually constructed)

In order to extract the lexicon from CORLEX, all different lexical forms occurring in the corpus were indexed. All wordforms were then automatically tagged (morphosyntactic tagging) and lemmatized by PALAVROSO (an automatic analyzer developed by INESC). The next task consisted on a manual verification of all the tags attributed to each wordform and lemma (with the minimum lemma frequency of 6).

5. RELEVANT REFERENCES AND OTHER INFORMATION

Bacelar do Nascimento, M. F., M. L. Garcia Marques e M. L: Segura da Cruz (1987), *Português Fundamental, Métodos e Documentos*, Vol. I, *Inquérito de Frequência*, INIC-CLUL, Lisboa: 358-391.

Bacelar do Nascimento, M. F. (2001), "Um novo léxico de frequências do português" in Volume de *Homenagem ao Professor Herculano de Carvalho* (no prelo).

Bacelar do Nascimento, M. F. et alii (2001), Poster "Léxico Multifuncional Computorizado do Português Contemporâneo" in *Feira de Projectos*, promovida pela Comissão Nacional do Ano Europeu das Línguas, Lisboa, Centro Cultural Casapiano, 27-30 de Setembro.

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.

Amaro, R. e F. Barreto (2004), "Multifunctional Computational Lexicon of Contemporary Portuguese: an available resource for multitype applications", *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (LREC 2004): 1075-1078.

Partner IST

Questions + NE CORPUS

1 BASIC INFORMATION

1.1 Corpus composition

This corpus consists of a set of nearly 5,500 manually annotated questions to be used as training corpus in machine learning based NER systems and 500 annotated questions for testing. Named entities in these questions were identified and classified according to the categories: Person, Location and Organization. More details about the process of building these corpora can be consulted in [1].
The original corpus of 6000 questions in English can be found in http://cogcomp.cs.illinois.edu/Data/QA/QC/. Details about it are presented in [2].

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

The corpus is a txt file.

1.3 Character encoding

The characters are 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 archive that will be uploaded on the MetaShare platform will contain one folder with two files with .txt extension (the train and the test file).

3.2 Data structure of an entry

This is not relevant as the corpus is provided as a text file, where each line contains a single question. In each question, all the named entities are identified: each named entity is between symbols "<" and ">".

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

The corpus for training has 5452 questions, 55620 tokens and occupies about 315 KB. The corpus for testing has 500 questions, 3758 tokens and occupies about 25 KB.

4 CONTENT INFORMATION

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

This corpus is in monolingual.

4.2 The natural language(s) of the corpus

The language of the corpus is English.

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

The corpus has questions from different types: factoids, definitional, lists.

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 corpus is annotated according with the following categories: Person, Location and Organization. An example can be seen in the following:

How did serfdom develop in and then leave <LOC>Russia</LOC> ?
What films featured the character <PER>Popeye Doyle</PER> ?
When did <ORG>CNN</ORG> begin broadcasting ?

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

Tags in use are:
a) <LOC> for Locations;
b) <PER> for Persons;
c) <ORG> for organization.

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)

Not relevant

4.5 Intended application of the corpus

This corpus can be used to train and test Named Entity Recognition in questions. As questions are different from declarative sentences, they need an appropriate corpus for training, as showed in [1]. In addition, the corpus where we have identified the named entities is widely used by the machine learning community, because each of its questions is labeled based on one of the most widely known taxonomies for question classification: Li and Roth's two-layer taxonomy [2]. This taxonomy consists of a set of six coarse-grained categories and fifty fined-grained ones. This fact makes this corpus a very valuable resource for training and testing machine learning models in question classification, and, more generally, making it a very valuable resource for question answering. By identifying the named entities in that corpus, these can be used to improve the attained models, as named entities can also be used as features.

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

The annotations were automatically built and manually checked, being the annotation process described in [1].

5 RELEVANT REFERENCES AND OTHER INFORMATION

[1] Ana Cristina Mendes, Luísa Coheur, Paula Vaz Lobo. Named Entity Recognition in Questions: Towards a Golden Collection in LREC'10. May, 2010.

[2] Xin Li, Dan Roth, Learning Question Classifiers. COLING'02. August, 2002.

Questions EN+PT CORPUS

1 BASIC INFORMATION

1.1 Corpus composition

This parallel corpus (Portuguese and English) consists of two sets of nearly 5,500 plus 500 questions each, to be used as training/testing corpora, respectively. Details on the translation and some experiments regarding statistical machine translation of questions can be found in [1]
The original corpus of 6000 questions in English can be found in http://cogcomp.cs.illinois.edu/Data/QA/QC/. Details about it are presented in [2].
The parallel corpus has some corrections regarding the original corpus.

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

The corpus is a txt file.

1.3 Character encoding

The characters are 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 archive that will be uploaded on the MetaShare platform will contain one folder with four files with .txt extensions (the train and the test files for Portuguese and English).

3.2 Data structure of an entry

This is not relevant as the corpus is provided as a text file, where each line contains a single question.

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

In what concerns the Portuguese language, the corpus for training has 5457 questions, 62977 tokens and occupies about 360 KB. The corpus for testing has 499 questions, 4576 tokens and occupies about 28 KB.
In what concerns the English language, the corpus for training has 5457 questions, 60954 tokens and occupies about 336 KB. The corpus for testing has 499 questions, 4258 tokens and occupies about 23 KB.

4 CONTENT INFORMATION

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

This corpus is a parallel bilingual corpus.

4.2 The natural language(s) of the corpus

The language of the corpus are European Portuguese and English.

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

The corpus has questions from different types: factoids, definitional, lists.

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)

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

---

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

Both corpora have the same number of lines, and each line is the translation of the line with the same number. Thus, both documents can be use to create a sentence by sentence alignement.

4.4.4 Attributes and their values (if annotated)

4.7 Intended application of the corpus

This corpus can be applied to improve statistical machine translation on questions, as showed in [1]. In addition, the original corpus is widely used by the machine learning community, because each of its questions is labeled based on one of the most widely known taxonomies for question classification: Li and Roth's two-layer taxonomy [2]. This taxonomy consists of a set of six coarse-grained categories and fifty fined-grained ones. This fact makes this corpus a very valuable resource for training and testing machine learning models in question classification, and, more generally, making it a very valuable resource for question answering. Thus, by translating it to Portuguese, we have now a Portuguese corpus where each question is labeled with a category from Li & Roth taxonomy, which can be used to train and test models for classifying questions in Portuguese.

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

All the translations were hand made by an expert and the used guidelines are described in [1].

5 RELEVANT REFERENCES AND OTHER INFORMATION

[1] Ângela Costa, Tiago Luís, Joana Ribeiro, Ana Cristina Mendes e Luísa Coheur. An English-Portuguese parallel corpus of questions: translation guidelines and application in SMT. Paper submitted to LREC'12.

[2] Xin Li, Dan Roth, Learning Question Classifiers. COLING'02. August, 2002.

Partner UNIMAN

BioLexicon

1. BASIC INFORMATION

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

The BioLexicon (Thompson et al, 2011) is a large-scale, wide-coverage computational lexicon covering the biomedical domain. A large part of the lexicon is concerned with covering biomedical terms and their variants. Entries for domain-specific verbs include syntactic and semantic information. The lexicon includes entries that correspond to biomedical-specific vocabulary, as well as general language words.

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

The BioLexicon is modelled on an XML DTD, based on the Lexical Markup Framework (LMF) model (Quochi et al, 2009). The model is implemented as a MySQL relational database. The contents of the BioLexicon can also be queried via a web interface:
http://wiki.ilc.cnr.it/BootStrep/searchPanel.action

1.3 Character encoding

The characters have been encoded in UTF8

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 will be available from ELRA via the MetaShare platform.

2.3 Copyright statement and information on IPR

The resource is available license-based and attracts a fee for both research and commercial usage. Different fees are charged according to the type of usage.

3. TECHNICAL INFORMATION

3.1 Directories and files

There are large number of tables that constitute the BioLexicon database, which are fully documented in the BioLexiconHandBook.pdf document included in the distribution. The tables can be split into 3 different categories, i.e.

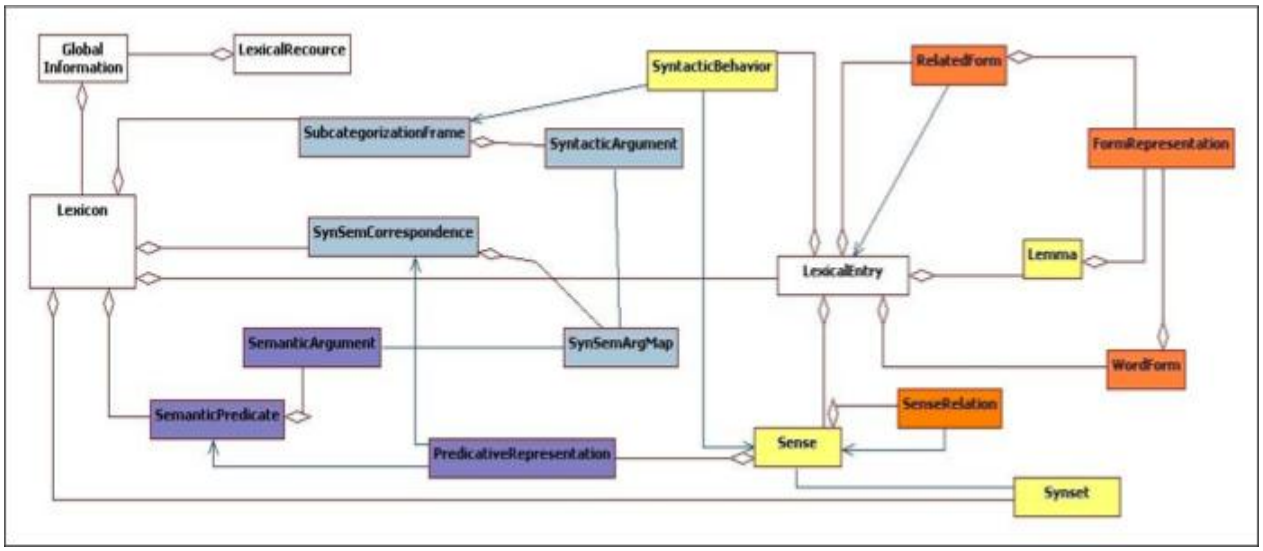
i. Morphological tables, to store lemmas, parts of speech, variants, related forms and inflections
ii. Syntactic tables, to store syntactic subcategorisation frames for frames, together with information about the different types of syntactic arguments
iii. Semantic tables, to store senses of entries, predicative representations of entries, semantic predicates, semantic arguments of predicates, and correspondences between syntactic and semantic frames

Further information about the structure of the database can be found here:

http://www.bootstrep.eu/pub/Intern/DocumentStore/BOOTStrep_Deliverable_D.2.2.pdf

3.2 Data structure of an entry

The general structure of a lexical entry in the BioLexicon is shown in Fig. 1



The *Lexical Entry* class is used to represent the lexemes of lag domain language, i.e. terms and verb entries as abstract units of vocabulary, which, in the case of verbs, comprise also the set of different forms of the same lemma. The *Lexical Entry* is a means for managing the *Lemma,* the *SyntacticBehaviour* and *Sense* objects. Therefore, the *Lexical Entry* manages the

relationship between the forms, full terms or short forms, their lexical meanings and their syntactic behaviours.

The BioLexicon model consists of a number of independent lexical objects (or classes) and a set of Data Categories (DCs), i.e., attribute–value pairs, which represent the main building blocks of lexical representation, especially tuned to the design goals of the lexicon. The set of DCs used in the BioLexicon consists of both categories drawn from the standard sets of the ISO Data Category Registry [77, 78], and categories created specifically for the biomedical domain.

More information can be found here:

http://www.bootstrep.eu/pub/Intern/DocumentStore/BOOTStrep_Deliverable_D.2.1_final.pdf

An example of an entry in XML format that conforms to the BioLexicon DTD is as follows:

```
<LexicalEntry ID="LE_interleukin-2">
<feat att="pos" val="Noun"/>
<Lemma>
<FormRepresentation>
<feat att="writtenform" val="interleukin-2"/>
<feat att="VariantType"" val="FullForm"/>
<feat att="source" val="database">
</FormRepresentation>
<FormRepresentation>
<feat att="writtenform" val="IL2"/>
<feat att="VariantType" val="orthographic"/>
<feat att="source" val="database">
</FormRepresentation>
<FormRepresentation>
<feat att="writtenform" val="IL-2"/>
<feat att="VariantType" val="orthographic"/>
<feat att="source" val="corpus">
<feat att="confScore" val="0.89999456">
</FormRepresentation>
</Lemma>
</LexicalEntry>
```

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

The BioLexicon contains 2236706 lexical entries

Nouns	Verbs	Adjectives	Adverbs	Total
2231574	1154	3428	550	2236706

The uncompressed tar file containing the BioLexicon data is approximately 3 GB.

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

The language of the BioLexicon is English.

4. 2 Entry Type

There are four types of entries: nouns, verbs, adjectives and adverbs.

4.3 Attributes and their values

Referring to the XML example in section 3.2, each *feat* element represents a data category. A *LexicalEntry* element can contain a number of data categories, encoding part of speech, grammatical gender, etc. A *Lemma* can have a number of variant forms, each represented by a *FormRepresentation* element, whose data categories include the written form of the variant, the type of variant (e.g., *orthographic*) and the *source*, which could b a biomedical database or a corpus (obtained by applying text mining techniques to the corpus). In the latter case, a confidence score shows with what confidence the variant has been extracted.

Other types of elements that can appear in *LexicalEntry* elements include *WordForm*, to encode different inflections of words, *RelatedForm*, to encode derived words (e.g. *abolishment* is derived from *abolish*), *SyntacticBehaviour*, to encode the syntactic behaviour of a verb. Detailed information about syntactic and semantic frames, as well as links between them, is also included in the lexicon.

More information about the BioLexicon model is available here:

http://www.bootstrep.eu/pub/Intern/DocumentStore/BOOTStrep_Deliverable_D.2.1_final.pdf

4.4 Coverage of the lexicon

The BioLexicon aims to achieve maximum, general-purpose coverage of biomedical language, together with detailed coverage of words used within the gene regulation domain.

- General language vocabulary used in biomedical texts is obtained from the MedPost dictionary(Smith et al, 2004).
- Additional biomedical terms have been extracted from 14 different biomedical resources. Since each resource has a different focus, the BioLexicon constitutes a unified, consolidated resource of biomedical terminology. 12 different semantic categories of terms have been extracted, including those that are common across all types of biomedical texts, e.g. gene/protein and species names, together with more focussed sets of terms rather are more closely related to gene regulation topics, such as operon names or gene ontology terms.
- 70,000 variants of genes and protein names have been automatically extracted by applying text mining techniques to 15 million MEDLINE abstracts, to ensure that real usage of these terms in biomedical texts is accounted for.
- Semantic categories of terms covered are: *Gene/Protein, Chemical, Organism, Enzyme, Protein Domain, Protein Complex, Disease, Molecular Role, Cell, Transcription Factor, Operon, Sequence.*

- Verbs extracted from the MedPost dictionary were augmented with 658 manually selected verbs that were considered to be either highly relevant or specific to the biomedical domain. Orthographic variations are included (e.g. British vs. American spellings), as are related entries (e.g. the adjective *absorbent* is derived from verb *absorb*. Each domain specific verb includes syntactic subcategorisation frames. For 168 of these verbs, sematic frame is available, and there are 668 links between syntactic and semantic frames.

4.5 Intended application of the lexicon

The BioLexicon has already been integrated into a number of different domain- specific tools, i.e. a part-of-speech tagger (Sasaki et al, 200, a lemmatizer, an information extraction system(Sasaki et al, 2010) and a fact extraction system. It is also intended to support a wide range of other tasks relating to biomedical text mining and information retrieval, including: dynamic query term completion during search input, detection of protein-protein interactions via co-occurrence, etc.

4.6 POS assignment

The POS assignment in the entries extracted from the MedPost dictionary is based on the MedPost tagger. Additional entries are Biomedical terms, manually selected verbs, or words derived from these verbs, which have also been manually curated.

4.7 Reliability (automatically/manually constructed)

Different types of information in the BioLexicon have been collected using different methods. Care has been taken to ensure that the quality of the entries in the BioLexicon is as high as possible, as explained in the following points"

- The MedPost tagger (Smith et al, 2004) achieves an accuracy of 97% on biomedical texts.
- Automatic term variant extraction works in two phrases: a named entity recognition tool (Sasaki et al, 2008), which performs with an accuracy of 73.78 F-Score, which constitutes state-of-the-art performance for gene/protein recognition. In the second phase, recognized terms are mapped to similar entries through normalization (Tsuruoka et al, 2008). The automatic normalization method acheived a precision of 76.7, and recall of 63.3. This compares favourably with the performance of manually constructed normalisation rules.
- Syntactic subcategorisation frames were acquired automatically, based on the output of the Enju parser tuned to the domain (achieving an F-Score of 86.87 on biomedical texts) (Hara et al, 2005). Extracted frames were filtered according to a manually-determined threshold based on their frequency of occurrence, given the verb. This aimed to filter out "noisy" frames as much as possible.
- Semantic frames were determined based on manual annotation of biomedical events carried out by domain experts (Thompson et al, 2008; Thompsom et al, 2009). Inter-annotator agreement scores were calculated, and reached levels of up to 0.89.
- Linking syntactic and semantic frames. This processes has been carried out manually be a linguistics expert (Venturi et al, 2009)

5. RELEVANT REFERENCES AND OTHER INFORMATION

Hara T, Miyao Y, Tsujii J: **Adapting a probabilistic disambiguation model of an HPSG parser to a new domain.** In *Proceedings of IJCNLP 2005*:199- 210.

ISO-12620: **Terminology and other content language resources – Data Categories – Specifications of data categories and management of a Data Category Registry for language resources:** ISO/TC37/SC3/WG4; 2006.

Quochi V, del Gratta R, Sassolini E, Bartolini R, Monachini M, Calzolari N: **A Standard Lexical-Terminological Resource for the Bio Domain.** In *Human Language Technology Challenges of the Information Society: Third Language and Technology Conference (LTC).* Springer-Verlag; 2009: 325-335.

Sasaki Y, McNaught J, Ananiadou S: **The value of an in-domain lexicon in genomics QA.** *J Bioinform Comput Biol* 2010, **8(1):147-161.**

Sasaki Y, Thompson P, McNaught J, Ananiadou S: **Three BioNLP tools powered by a biological lexicon.** In *Proceedings of EACL: Demonstrations Session* 2009:61-64.

Sasaki Y, Tsuruoka Y, McNaught J, Ananiadou S: **How to make the most of named entity dictionaries in statistical NER**. BMC Bioinformatics 2008, 9 (Suppl 11):S5.

Smith L, Rindflesch T, Wilbur WJ: **MedPost: a part-of-speech tagger for Biomedical text.** *Bioinformatics* 2004, 20(14):2320-2321.

Thompson P, Cotter P, McNaught J, Ananiadou S, Montemagni S, Trabucco A, Venturi G: **Building a bio-event annotated corpus for the acquisition of semantic frames from biomedical corpora.** In *Proceedings of LREC* 2008:2159-2166.

Thompson P, Iqbal SA, McNaught J, Ananiadou S: **Construction of an annotated corpus to support biomedical information extraction.** *BMC Bioinformatics* 2009, **10:349.**

Tsuruoka Y, McNaught J, Ananiadou S: **Normalizing biomedical terms by minimizing ambiguity and variability.** *BMC Bioinformatics* 2008, **9 Suppl 3:S2.**

Venturi, G., Montemagni, S., Marchi, S., Sasaki, Y., Thompson, P., McNaught, J. and Ananiadou, S.: **Bootstrapping a Verb Lexicon for Biomedical Information Extraction**. In: Gelbukh, A.(Ed.) Proceedings of the 10th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing) 2009: 137-148, Springer

Wright SE: **A global data category registry for interoperable language resources**. In *Proceedings of LREC* 2004:123-126.

GENIA CORPUS

1 BASIC INFORMATION

1.1 Corpus composition

The corpus consists of 2,000 MEDLINE abstracts, collected using the three MeSH terms *human*, *blood cells* and *transcription factors*.

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

The corpus is available in three formats

- A text file containing part-of-speech (POS) annotation, based on the Penn Treebank format
- An XML file containing inline POS annotation
- A "merged" XML format, containing inline annotations, corresponding to both POS and term annotations

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 as an archive.

2.3 Copyright statement and information on IPR

The resource is freely available for research purposes

3 TECHNICAL INFORMATION

3.1 Directories and files

The archive contains the directory *GENIAcorpus3.02p.* It contains the following files (NOTE: each file contains the *complete* corpus in the format specified).

- *GENIAcorpus3.02.pos.txt* - a plain text file containing the POS-tagged corpus, with formatting based on the Penn TreeBank (PTB) formatting.
- *GENIAcorpus3.02.pos.xml* – an XML encoded file containing the abstracts with inline POS tags.
- *GENIAcorpus3.02.merged.xml* – an XML file containing the abstracts with inline annotation corresponding both to POS tags and term annotations
- *GENIAontology.daml* – a file containing the GENIA term ontology which has been used to annotate the terms.
- *gpml.merged.dtd* – The DTD to which the 2 XML files conform.
- *gpml.css* – A stylesheet to highlight the annotated terms
- *gpml.readme.html* – a file providing brief details of the structure of the file and the term annotation format
- *gpml.css.legend.html* – a file explaining the color-coding used in the css file

3.2 Data structure of an entry

This is not relevant as the corpus is a set of text files.

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

The corpus contains approximately 500,000 tokens. The PTB text-based POS corpus format requires 5.1MB of disk space, while the XML version of the POS corpus requires 10.6MB of disk space. The merged XML corpus, containing both POS and term annotations, requires 16.2 MB of disk space.

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

The language of the corpus is English.

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

The corpus contains abstracts of biomedical research 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)

The corpus is tokenized, and token is assigned a POS tag. The merged version of the corpus additionally contains annotations corresponding to biomedical terms.

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

Each abstract was automatically tokenized using the Penn tokenizer (http://www.cis.upenn.edu/~treebank/tokenization.htm) and assigned POS tags using the JunK tagger (Kazama et al, 2001). Corrections to the automatically assigned annotations were then made by annotators (see the POS annotation guidelines: http://www-tsujii.is.s.u-tokyo.ac.jp/%7Ejdkim/publications/GENIA_Guidelines_POS.pdf). POS tags generally follow the Pen TreeBank (PTB) format (Santorini, 1990), with some modifications:

- The NNP and NNPS (proper name) tags are not used, except for the names of journals, authors, research institutes, and initials of patients. Especially, (discoverers') names in technical terms (e.g. Epstein-Barr virus, Southern blotting) are not tagged as NNP.
- The SYM tag has been eliminated as far as possible.

The POS annotation of the GENIA corpus is reported in detail in Tateisi & Tsujii (2004). See also http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=Part%2Dof%2DSpeech+Annotation.

Term annotations were manually annotated, and assigned categories from the GENIA ontology (35 categories corresponding to the terminal ontology nodes). The term annotation takes care of semi-structured coordinated clauses by recovering ellipsis. As an example, the phrase *CD2 and CD 25 receptors* refers to two terms, i.e. *CD2 receptors* and *CD25 receptors*, but *CD2 receptors* doesn't appear in the text. The annotation is carried out in such a way to allow these separate terms to be identified. Term annotation is described in more detail in Kim et al. (2003). See also http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=Technical+Term+Annotation.

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

Not applicable – this is a monolingual corpus.

4.4.4 Attributes and their values (if annotated)

PTB-style text format

The file contains one token/POS pair per line, and a "==========" line (20 equal signs) is put between sentences.

For example:

```
==========
These/DT
findings/NNS
should/MD
be/VB
useful/JJ
for/IN
therapeutic/JJ
strategies/NNS
and/CC
the/DT
development/NN
of/IN
immunosuppressants/NNS
targeting/VBG
the/DT
CD28/NN
costimulatory/NN
pathway/NN
./.
==========
```

XML-format

An example of the XML format of the merged corpus (*GENIAcorpus3.02.merged.xml*) is shown below. Note that the format of the XML file containing only POS tags (*GENIAcorpus3.02.pos.xml*) is the same, except for that the *cons* tags (corresponding to term annotations) are not present.

```
<article>
<articleinfo>
<bibliomisc>MEDLINE:95369245</bibliomisc>
</articleinfo>
<title>
<sentence><cons lex="IL-2_gene_expression" sem="G#other_name"><cons
lex="IL-2_gene" sem="G#DNA_domain_or_region"><w c="NN">IL-2</w>
<w c="NN">gene</w></cons> <w c="NN">expression</w></cons> <w
```

c="CC">and</w> <cons lex="NF-kappa_B_activation"
sem="G#other_name"><cons lex="NF-kappa_B"
sem="G#protein_molecule"><w c="NN">NF-kappa</w> <w
c="NN">B</w></cons> <w c="NN">activation</w></cons> <w
c="IN">through</w> <cons lex="CD28" sem="G#protein_molecule"><w
c="NN">CD28</w></cons> <w c="VBZ">requires</w> <w
c="JJ">reactive</w> <w c="NN">oxygen</w> <w c="NN">production</w>
<w c="IN">by</w> <cons lex="5-lipoxygenase"
sem="G#protein_molecule"><w c="NN">5-lipoxygenase</w></cons><w
c=".">.</w></sentence>
</title>
<abstract> …
</abstract>

The tags and attributes are as follows:

- *article* tag – surrounds each abstract in the corpus
- *articleinfo* tag – contains information about the article
- *bibliomisc* tag– contains the MEDLINE id of the article
- *title* tag– contains the title of the article
- *abstract* tag – contains the main text of the abstract
- *sentence* tag – surrounds the text of each sentence in the title/abstract
- *w* tag – surrounds each token
  - *c* attribute – the part of speech assigned to the token
- *cons* tag – corresponds to a term annotation – surrounds one of more *w* tags. These can be embedded, as is the case with *IL-2 gene* and *IL-2 gene expression* in the example above.
  - *lex* attribute – the complete lexical representation of the term, with spaces replaced by underscores
  - *sem* attribute - the semantic category assigned to the term. The *G* prefix denotes that the category has been assigned from the GENIA term ontology.

Co-ordinated structures containing 2 or more terms that are not both completely specified in the conjunction (due to ellipsis) are annotated as shown below. The coordinated phrase in this case is *hematopoietic and trophoblast cells*, which refers to the two terms *hematopoietic cells* and *trophoblast cells*.

<cons lex="(AND hematopoietic_cell trophoblast_cell)" sem="(AND
G#cell_type G#cell_type)"><cons lex="hematopoietic*"><w
c="JJ">hematopoietic</w></cons> <w c="CC">and</w> <cons
lex="trophoblast*"><w c="NN">trophoblast</w></cons> <cons
lex="*cell"><w c="NNS">cells</w></cons></cons><w
c=".">.</w></sentence>

A *cons* tag is placed around the whole phrase, and the use of *AND* in both the *lex* and *sem* attribute values shows that there are 2 distinct terms involved. An embedded *cons* tag is placed around both the unique and common parts of each of the two terms. The use of the * symbol in

lex attribute of these embedded *cons* elements indicates that each *cons* element contains only part of the term, with the remainder in another *cons* element.

4.9 Intended application of the corpus

The corpus is intended to allow to facilitate the building of domain-specific term recognition systems, and to help to adapt existing POS taggers and other applications to the biomedical domain.

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

In terms of the manually-corrected POS tags, inter-annotator agreement scores were 0.985 Kappa.

5 RELEVANT REFERENCES AND OTHER INFORMATION

Kazama, J., Y. Miyao, and J. Tsujii, 2001. A maximum entropy tagger with unsupervised hidden markov models. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium.*

Kim, J-D., T. Ohta, Y. Tateisi, and J. Tsujii (2003). Genia corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19:180i–182i.

Santorini, B., 1990. Part-of-speech tagging guidelines for the Penn Treebank project. Technical Report MS-CIS- 90-47, Department of Computer and Information Science, University of Pennsylvania.

Tateisi, Yuka and Jun'ichi Tsujii. Part-of-Speech Annotation of Biology Research Abstracts. In *Proceedings of 4th International Conference on Language Resource and Evaluation (LREC2004).* pp. 1267-127

GENIA EVENT CORPUS

1 BASIC INFORMATION

1.1 Corpus composition

The corpus consists of 1000 MEDLINE abstracts. It is a subset of the original GENIA corpus, which was selected using the three MeSH terms *human*, *blood cells* and *transcription factors*.

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

The corpus is provided as a set of XML files, which contain both the abstract text and annotations.

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 as an archive.

2.3 Copyright statement and information on IPR

The resource is freely available for research purposes

3 TECHNICAL INFORMATION

3.1 Directories and files

The archive contains the directory *Genia_Metaknowledge_Corpus*. It contains the following sub-directories and files:

- *Corpus* – directory containing 1,000 XML files. Each file contains one Medline abstract.
- *ModifiedGENIAtypes* – directory containing the DTD and CSS files for the XML files constituting the GENIA event corpus. All the XML files in the corpus are validated against the DTD. The CSS well works with the Opera web browser.
- *GENIAontologies* – directory containing two GENIA ontologies encoded in OWL. The GENIAterm40.owl defines the term classes on which the GENIA term annotation is based. The GENIAevent.owl defines the event classes on which the GENIA event annotation is based.

- *Guidelines_for_event_annotation.pdf* – Annotation guidelines used to produce the event annotations (see below for more details)
- *Meta-knowledge_annotation_guidelines.pdf* – Annotation guidelines used to add meta-knowledge annotation to the events (see below for more details)
- *README.html* – Provides basic introductory information about the corpus.

3.2 Data structure of an entry

This is not relevant as the corpus is a set of text files.

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

The corpus contains approximately 220,000 tokens. It requires approximately 21.7 MB of disk space.

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

The language of the corpus is English.

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

The corpus contains abstracts of biomedical research 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)

The corpus (Kim et al, 2008) contains structural annotations obtained from MEDLINE, e.g. to identify the abstract title and main body of the abstract. The text is split into sentences. In each sentence, three types of information are annotated:

- Firstly, biomedical terms are identified and assigned categories from the GENIA term ontology.
- Secondly, event structures are identified and assigned categories from the GENIA event ontology.

- Thirdly, detailed information is annotated about how the event should be interpreted, according to its textual context. We call this information *meta-knowledge*.

Consider the following sentence:

The results suggest that LMP1 activates NF-kappa B

Term annotation identifies *LMP1* and *NF-kappa B* as terms, while event annotation identifies that there is a relationship between these terms, i.e. they participate in an event of type *Positive_Regulation*, in which *activates* is the *event trigger, LMP1* is the CAUSE of the event and *NF-kappa B* is the THEME, i.e. what is affected by the event. Finally, meta-knowledge annotation encodes that fact the event is a slightly speculated analysis of results rather than, e.g., a definite fact.

Meta-knowledge is classified according to 5 different dimensions:

- *Knowledge Type* – general information content of the event. Does it represent an investigation, observation, analysis, etc.
- *Certainly level* – the level of certainty associated with the occurrence of the event
- *Polarity* – whether or not the event is negated
- *Manner* - the rate, level, strength or intensity of the event (in biological terms)
- *Source* - the source or origin of the knowledge being expressed by the event. Specifically, we distinguish between events that can be attributed to the current study, and those that are attributed to other studies

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

Term annotations were manually annotated as part of the original GENIA annotation (see Kim et al., 2003). Terms were assigned categories from the GENIA ontology (35 biologically categories corresponding to terminal nodes in the ontology). The term annotation takes care of semi-structured coordinated clauses by recovering ellipsis. As an example, the phrase *CD2 and CD 25 receptors* refers to two terms, i.e. *CD2 receptors* and *CD25 receptors*, but *CD2 receptors* doesn't appear in the text. The annotation is carried out in such a way to allow these separate terms to be identified. Term annotations are inline, within eaxh sentence annotation. Term annotation is described in more detail in Kim et al. (2003). See also http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=Technical+Term+Annotation.

Event annotation was carried manually out on top of term annotation to identify the biological processes in which the terms participate. Event

annotations are attached to each sentence, providing information about the structure of each event. More information about the event annotation can be found in Kim et al (2008), http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=Event+Annotation and associated annotation guidelines: http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/release/GENIA_event_annotation_guidelines.pdf

Meta-knowledge annotation was carried out manually on top of the event annotations. More information about the meta-knowledge annotation can be found in Thompson et al. (2011). Also see: and http://www.nactem.ac.uk/meta-knowledge/.

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

Not applicable – this is a monolingual corpus.

4.4.4 Attributes and their values (if annotated)

An extract from a typical annotated file is as follows:

```
<sentence id="S2">We have examined the effect of <term id="T6"
lex="leukotriene_B4" sem="Organic_compound_other">leukotriene
B4</term> (<term id="T7" lex="LTB4"
sem="Organic_compound_other">LTB4</term>), a potent lipid <term
id="T9" lex="proinflammatory_mediator"
sem="Protein_family_or_group">proinflammatory mediator</term>, on the
expression of the <cons id="T10" lex="(AND proto-oncogene_c-jun proto-
oncogene_c-fos)" sem="(AND DNA_domain_or_region
DNA_domain_or_region)"><frag id="F4">proto-oncogenes</frag> <frag
id="F5"><term id="A3" sem="DNA_domain_or_region">c-
jun</term></frag> and <frag id="F6"><term id="A4"
sem="DNA_domain_or_region">c-fos</term></frag></cons>.</sentence>

<event KT="Investigation" id="E7" uncertainty="doubtful">
<type class="Regulation"/>
<theme idref="E9"/>
<cause idref="T7"/>
<clue>We have <clueKT>examined</clueKT> the
<clueType>effect</clueType> <linkCause>of</linkCause> leukotriene B4
(LTB4), a potent lipid proinflammatory mediator,
<linkTheme>on</linkTheme> the expression of the proto-oncogenes c-jun
and c-fos.</clue>
</event>
```

Each *sentence* tag contains term annotations. These are either indicated using *term* tags for simple, or additionally using *cons* and *frag* tags for

more complex terms. An example of a complex term is a co-ordination in which two terms are contained, but ellipsis is involved. In the above example, this is exemplified by the phrase *proto-oncogenes c-jun and c-fos*, in which there are 2 terms, *proto-oncogene c-jun* and *proto-oncogene c-fos*. The *cons* tag surrounds the whole phrase, while *frag* is used to denote the individual parts of the terms, i.e., the common part of the two terms, *proto-oncogenes* and the unique parts, i.e., *c-jun* and *c-fos*. All of the above-mentioned tags have *id* attributes containing unique ids. *Term* and *cons* have the following additional attributes:

- *lex* attribute – the lexical representation of the term, which spaces replaced by underscores. In the case of the *cons* tag, this may include "AND" to show that two or more terms are contained within the text surrounded by the tag.
- *sem* – the semantic class assigned to the term from the GENIA term ontology. In the case of the *cons* tag, there may be multiple categories (one for each term within the enclosed phrase), indicated using "AND".

Tags of type *event* follow the *sentence* annotation and encode the events that have been annotated within the sentence. Each event has an *id* attribute to assign a unique id. Other attributes are optional, but can include the following (if non default values are assigned).

- *uncertainty* – one of the basic types of information relating to event interpretation annotated as part of the original event annotation. Can have the following values: *certain* (there is no doubt that the event took place; default value), *probable* (there is some level of speculation surrounding the event), *doubtful* (the event is under investigation). Note that the combination of meta-knowledge annotation dimensions (see below) is intended to provide more detailed information relating to event interpretation, and hence largely supersedes this attribute. However, it is retained for historical purposes.
- *assertion* – another one of the basic types of interpretation added as part of the original event annotation. Can have the following values: *non-exist* (the event is explicitly negated), *exist* (there is no explicit negation of the event; default value). This is somewhat similar to the *Polarity* meta-knowledge dimension. However, meta-knowledge is meant to encode more subtle differences, and so largely supersedes this attribute. However, it is retained for historical purposes.
- *KT [Knowledge Type]*, *CL [Certainty Level], Polarity, Manner, Source* – these correspond to the newly added meta-knowledge dimensions that add greater detail about the intended interpretation of events. Each has its own set of possible values. Further details can be found in Thompson et al. (2011) and the meta-knowledge

annotation guidelines (http://www.nactem.ac.uk/meta-knowledge/Annotation_Guidelines.pdf)

The *clue* tag includes the text of the sentence in which the event is contained. Several clue expressions may be annotated, which are envisaged to help with the automatic recognition of events and associated meta-knowledge. The possible tags that can occur within the *clue* tag are as follows:

- *clueType* – The event trigger word or phrase
- *clueLoc* – the location in which the event took place
- *clueExperiment* – experimental techniques specified for the event.
- *clueTime* – corresponds to when the event happened or will happen.
- *linkCause* – used to indicate words that are used in the text link between and event and its CAUSE. They can be seen as words that "introduce" the CAUSE of the event, e.g. ***effect*** *of X on Y,* where *of* is the linkCause.
- *linkTheme* – used to indicate words used in the text to link the event and its THEME. They can be seen as words that introduce the THEME of the event, e.g. e.g. ***effect*** *of X on Y,* where *on* is the linkTheme.
- *coRefCause* – annotated when the CAUSE of the event is an expression such as *it* or *this protein*, referring to a previously introduced (or coreferent) NE, either in the current sentence or in a previous sentence.
- *coRefTheme* – annotated when the THEME of the event contains an expression such as *it* or *this protein*, referring to a previously introduced (or coreferent) NE, either in the current sentence or in a previous sentence.
- *clueKT* – clue expression used to determine the chosen value of the *Knowlegde Type (KT)* meta-knowledge dimension.
- *clueCL* – clue expression used to determine the chosen value of the *Certainty Level* (CL) meta-knowledge dimension.
- *cluePolarity* - clue expression used to determine the chosen value of the *Polarity* meta-knowledge dimension.
- *clueManner* - clue expression used to determine the chosen value of the *Manner* meta-knowledge dimension.
- *clueSource* – clue expression used to determine the chosen value of the *Manner* meta-knowledge dimension.

4.11 Intended application of the corpus

The corpus is intended to allow to facilitate the advanced information semantic search systems in the biomedical domain, that allow events to be located using structured queries that can take into account semantic

roles, NEs, etc. Systems trained to recognise meta-knowledge can allow extra sets of search criteria to be specified. This can allow, e.g., the isolation of new experimental knowledge to facilitate biomedical database curation, enable textual inference to detect entailments and contradictions, etc.

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

In terms of the meta-knowledge annotations, inter-annotator agreement rates in the range 0.84 – 0.92 Kappa were achieved, according to the different dimensions (Thompson et al., 2011)

5 RELEVANT REFERENCES AND OTHER INFORMATION

Kim, J-D., T. Ohta, Y. Tateisi, and J. Tsujii (2003). Genia corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19:180i–182i.

Kim, J.-D., Ohta, T. and Tsujii, J.. (2008). Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9:10

Thompson, P., Nawaz, R., McNaught, J. and Ananiadou, S. (2011). Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinformatics*, 12:393

GREC CORPUS

1 BASIC INFORMATION

1.1 Corpus composition

The corpus consists of 240 MEDLINE abstracts on the subject of gene regulation. 167 of these abstracts concern the *E. coli* species, while the remaining 73 abstracts concern the *Human* species.

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

The corpus is available in two formats

- standoff annotation
- XML-encoded annotation

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 as an archive.

2.3 Copyright statement and information on IPR

The resource is freely available for research purposes.

3 TECHNICAL INFORMATION

3.1 Directories and files

The corpus is available in 2 different formats, contained within 2 archives:

- GREC_Standoff.zip - contains plain text files (.txt) containing abstract texts and associated standoff annotations files (.a1 and .a2). Split into two sub-directories:
  - Ecoli – cooresponds to *E. coli* abstracts
  - Human – corresponds to Human abstracts
- GREC_XML.zip – contains the annotated abstracts in XML format. There are three subdirectories:
  - GRECResources – contains the DTD file to which the XML files conform
  - Ecoli – contains the annotated *E. coli* abstracts in XML format
  - Human – contains the Human abstracts in XML format

3.2 Data structure of an entry

This is not relevant as the corpus is a set of text files.

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

The corpus contains approximately 52,000 tokens. The standoff version of the corpus requires approximately 3 MB on disk, while the XML version of the corpus requires approximately 2.3 MB on disk.

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

The language of the corpus is English.

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

The corpus contains abstracts of biomedical research 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)

The standoff version of the corpus is annotated with biomedical event structures and named entities associated with these event structures. The XML version of the corpus is additionally annotated with sentence boundaries. Further information is provided at: http://www.nactem.ac.uk/GREC/.

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

Each abstract was automatically split into sentences using the GENIA tagger (Tsuruoka et al., 2005). 6 biologist annotators then annotated the following information (see Thompson et al, 2009)

- Verbs and nominalised verbs describing gene regulation events (event triggers)
- Semantically-related arguments of these event triggers within the same sentence; each argument was assigned a semantic role from a set of 13 possible roles
- Named entities occurring within sematic arguments were annotated and assigned appropriate named entity types. The hierarchy contains around 70 NE categories, arranged with 5

supertypes,i.e. PROTEINS, NUCLEIC_ACIDS, LIVING_SYSTEMS, PROCESSES and EXPERIMENTAL. Further details, with the full set of NEs used, can be found in the annotation guidelines: http://www.nactem.ac.uk/download.php?target=GREC/Event_annotation_guidelines.pdf

As a simple example, consider the following sentence:

*The narL gene product **activates** the nitrate reductase operon*

The sentence contains a single event, with the trigger *activates*. There are two arguments:

- The narL gene product
- the nitrate reductase operon

The argument *The narL gene product* is assigned the semantic role *AGENT* and the biological concept *Protein*, whilst the argument *the nitrate reductase operon* is assigned the semantic role *THEME* and the biological concept *Operon*.

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

Not applicable – this is a monolingual corpus.

4.4.4 Attributes and their values (if annotated)

Standoff annotations

For the standoff version of the corpus, each text file (with extension ".txt") is accompanied by two files (with extensions containing the annotations. The format is based largely on the one used in the BioNLP Shared Task'09 (Kim et al., 2009) (see also http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/SharedTask/) with some modifications.

".a1" files

These files contains text spans that constitute event arguments and named entities occurring within these spans. The format of these files is as follows:

T1	Activator 0 4	EnvZ
T2	SPAN 15 27	through OmpR
T3	Regulator 23 27	OmpR

Each span/NE is assigned as id beginning with "T". The NE type (or "SPAN" – for event argument text spans that do correspond directly to NEs) is accompanied by start and end offsets in the abstract and the corresponding text.

".a2" files

These files contain text spans that constitute event triggers, and their associated event triggers. The format is as follows:

```
T13     Gene_Activation 263 273         activation
T14     GRE 296 304             requires
T15     GRE 309 317             function
E1      Gene_Activation:T13 Theme:T1,T4
E2      GRE:T14 Agent:E1 Theme:E3
```

Lines with an id starting with "T" correspond to event triggers, and follow the same format as the lines in the ".a1" files. Instead of NE types, a category of the event is shown in addition to offsets and text spans.

Lines with an id stating with "E" correspond to the event structures. The type of the event is separated by a colon from the id of the event trigger. This is followed by a list of the semantic arguments associated with the event. For each argument, the semantic role assigned to the argument is separated by a colon from the id of the argument. The argument can correspond to either:

- one or more simple text spans (contained within the associated ".a1" file), with ids begins with "T". Event arguments can consist of discontinuous text spans. In this case, each part of the argument is identified separately in the ".a1" file, and the event argument in the ".a2" file is specified using a comma-separated list of ids.
- another event structure described within the same ".a2" file. In this case, the id will start with "E", and will refer to (starting with "E"): an event argument can be another event.

Further information about the format of the standoff annotations can be found here: http://www.nactem.ac.uk/GREC/standoff.html

XML annotations

In the XML version of the corpus, a single file is present for each abstract, containing both the abstract text and the annotations. This format is based on the one used to represent the GENIA event corpus (Kim et al, 2008), with a small number of additions/modifications. See also:

http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=Event+Annotation

An example of the annotation is shown below:

<sentence id="S7">In contrast, <term sem="SPAN" id="T15" lex="sites_upstream_of_the_promoters">sites upstream of the promoters</term> did not appear to be necessary for repression, but were required for activation by <term sem="Activator" id="T16" lex="Lrp">Lrp</term> plus <term sem="Amino_Acids" id="T17" lex="alanine">alanine</term> or <term sem="Amino_Acids" id="T18" lex="leucine">leucine</term> of one of the major dad promoters, <term sem="Promoter" id="T19" lex="P2">P2</term>.</sentence>

<event id="E10">
<type class="Gene_Repression" />
<Condition idref="T15" />
<clue>In contrast, sites upstream of the promoters did not appear to be necessary for <clueType>repression</clueType>, but were required for activation by Lrp plus alanine or leucine of one of the major dad promoters, P2.</clue>
</event>

The following tags and attributes are used:

- *sentence* tag - represents a sentence
  - *id* attribute – the id of the sentence
- *term* tag – represents a NE or event argument within a sentence
  - *id* attribute – an id assigned to the term tag
  - *sem* attribute – the NE category assigned to the term span, or *SPAN* for event arguments that do constitute NEs.
  - *lex* – the textual representation of the contents of the term tag, with spaces replaced by underscores.
- *event* tag - represents an event structure
  - *id* attribute – an id assigned to the event
- *type* tag– the type of the event
  - *class* attribute – semantic class assigned to the event
- *Agent, Theme, Manner, Instrument, Location, Source, Destination, Temporal, Condition, Rate, Descriptive-Theme, Descriptive-Agent, Purpose* tags – One or more of these will be present within the event tag, according to the semantic roles assigned to the identified event arguments (see Thompson et al. (2009) for more details about the semantic roles used).
  - *Idref* attribute – stores the id of the event argument – either a *term* tag id or an *event* tag id. As mentioned above, arguments may consist of multiple, discontinuous spans. The attributes *idref1*, *idref2,* etc. may also be present if the argument consists of a discontinuous text span, to store the ids of other *term* tags that constitute the complete argument.
- *clue* tag – contains the complete sentence in which the event is contained. Expressions that constitute clues for identifying the

event in the text are annotated within the sentence. In the current version of the corpus, only *clueType* is annotated.

- *clueType* tag – surrounds the event trigger within the *clue* text.

Further information about the format of the standoff annotations can be found here: http://www.nactem.ac.uk/GREC/xml.html

4.13 Intended application of the corpus

The corpus is intended to allow to the training of advanced, domain-specific semantic search systems that allow searches to be carried out over documents using structured semantic queries using named entities, semantic roles etc. as search constraints.

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

The reliability of the annotations was verified by calculating inter-annotator agreement rates for a portion of the corpus. Since the event annotation task consists of a number of sub-tasks, agreement rates were calculated for each of these. These are shown in Table 1.

Agreement Type	**F-Score**	
	E.coli	**Human**
Event identification	72.27%	76.37%
Argument identification (relaxed span match)	90.23%	91.27%
Argument identification (exact span match)	75.10%	77.48%
Semantic role assignment	88.96%	88.30%
Biological concept identification	82.55%	82.03%
Bio-concept category assignment (exact)	71.02%	66.03%
Bio-concept assignment (considering parent)	75.38%	68.97%
Bio-concept supercategory assignment	95.52%	94.75%

Table 1: Inter-annotator agreement figures for the GREC corpus

5 RELEVANT REFERENCES AND OTHER INFORMATION

Kim, J.-D., Ohta, T., Pyysalo, S., Kano, Y. and Tsujii, J.. (2009). Overview of BioNLP'09 Shared Task on Event Extraction. *In Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pp. 19.

Kim, J.-D., Ohta, T. and Tsujii, J.. (2008). Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9:10

Thompson, P., Iqbal, S. A., McNaught, J. and Ananiadou, S.. (2009). Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics*, 10:349

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. In *Advances in Informatics - 10th Panhellenic Conference on Informatics*, pages 382-392, Springer-Verlag

SemLink

1 BASIC INFORMATION

1.1 Resource composition

SemLink (Palmer, 2009) provides a mapping between complementary lexical resources. In the current release, two mappings are available:

- a mapping between VerbNet (Kipper et al, 2008) and PropBank (Palmer et al, 2005)
- a mapping between VerbNet and FrameNet (Fillmore et al, 2008). The version of FrameNet used is v1.2

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

The mappings are encoded within a set of XML and plain text files

1.3 Character encoding

The characters are UTF8 encoded.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Martha Palmer

Address: Department of Linguistics, University of Colorado at Boulder
295 UCB Boulder, Colorado 80309-0295
Affiliation: Department of Linguistics, University of Colorado at Boulder
Position: Professor
Telephone: + 1 (303) 492-1300
Fax: + 1 (303) 492-4416
e-mail: martha.palmer@colorado.edu

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

The resource is available on the MetaShare platform as an archive.

2.3 Copyright statement and information on IPR

The resource is freely available for research purposes, according to the terms of the licence included within the archive.

3 TECHNICAL INFORMATION

3.1 Directories and files

The archive contains the directory *semlink1.1.* It contains the following sub-directories and files:

- *vn-fn* – directory containing mappings between VerbNet and FrameNet. The following files are contained within this directory:
  - *VNclass-FNframeMappings.xml* – provides mappings between VerbNet verbs and FrameNet lexical entries
  - *VN-FN_roleMapping.xml* – contains the possible role correspondences between VerbNet thematic roles and FrameNet frame elements.
  - *README.txt* - a file providing more details about the mappings contained within the above two files and their XML encoding
- *vn-pb* - directory containing mappings between VerbNet and PropBank. The following files are contained within this directory:
  - *type_map.xml* – specifies potential mappings between PropBank rolesets and VerbNet classes for a given lemma
  - *vnprop.txt* – specifies the correct mapping between PropBank roleset and VerbNet class for each predicate in the PropBank corpus. Contains only VerbNet role labels.
  - *vnpbprop.txt* – same as *vnprop.txt* except that PropBank role labels are also included
  - *mapping_stats.txt* – Provides general statistics regarding mappings between PropBank and VerbNet argument types
  - *README.TXT* – Provides information about the VerbNet-PropBank mapping files.
- LICENSE.TXT – provides details of the licence that must be followed when using SemLink
- README.TXT – Provides general information about SemLink

3.2 Data structure of an entry

This is not relevant as the corpus is a set of text files.

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

The SemLink directory requires 27.2MB of disk space. The following statistics provide information about the size of the resource:

- The mapping between VerbNet verbs and FrameNet frames considers 4755 VerbNet verbs.
    - Mappings between 2168 of these VerbNet verbs and FrameNet lexical units have been found.
- According to VerbNet/FrameNet mappings identified, 598 mappings are provided between sets of thematic roles in VerbNet classes and FrameNet frame elements
- The VerbNet-PropBank mappings consider 1881 lemmas from the PropBank corpus, leading to a total of possible 2665 mappings between PropBank rolesets and VerbNet classes.
- 112,917 occurrences of predicates appearing in the PropBank corpus have links to appropriate VerbNet classes.

4 CONTENT INFORMATION

4.1 Type of the resource (language (in)dependent)

The resource is language-dependent

4.2 The natural language(s) of the corpus

The language of the corpus is English.

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

The resource is concerned with language behavior in general English texts.

4.4 Annotations in the resource (if an annotated resource)

4.4.1 Types of annotations

Mappings exist at several levels, i.e. classes of verbs/predicates, individual members of these classes, and at the level of their arguments. See next section for further details.

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

The VerbNet-FrameNet mapping consists of two types of information, in separate files.

1. Possible mappings between individual verbs belonging to VerbNet classes and individual lexical entries belonging to FrameNet frames (*VNclass-FNframeMappings.xml*). The mapping is many-many: VerbNet verbs can map to multiple lexical entries in FrameNet, and vice versa.
2. For all VerbNet classes and FrameNet frames that are linked according to the mappings above, mappings are provided between the thematic roles in the VerbNet classes and the frame elements in the FrameNet frames (*VN-FN_roleMapping.xml*)

The VerbNet-PropBank mapping also consists of 2 types of information, in separate files.

1. For each lemma that occurs in PropBank, each possible roleset is listed, together a mapping to an appropriate class in VerbNet (*type_map.xml*) Appropriate mappings between the arguments of the PropBank roleset and corresponding thematic roles in VerbNet are also shown.
2. For each verbal occurrence in the PropBank corpus, the appropriate PropBank roleset-VerbNet class mapping, according to the possible mappings listed in *type_map.xml*. This information is contained in *vnprop.txt* and *vnpbprop.txt.* These files also include mappings between PropBank argument labels and VerbNet thematic role labels.

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

Not applicable – this is a monolingual resource.

4.4.4 Attributes and their values (if annotated)

VerbNet-FrameNet mapping

VNclass-FNframeMappings.xml

This file lists elements that include the following attributes:

class -- VerbNet class ID (numeric)
vnmember -- VerbNet class member (string, the verb lemma)
fnframe -- FrameNet Frame (string)
fnlexent -- FrameNet lexical entry ID (numeric)
versionID -- VerbNet version ID (either 1.5 or 2.0)

The structure of an element can be demonstrated by this example:

<vncls class='9.1-2' vnmember='put' fnframe='Placing' fnlexent='5355' versionID='vn2.0' />

Note that the string values for 'fnframe' and 'vnmember' attributes can include hyphens and underscores.

There are two special values possible for the 'fnframe' attribute:

fnframe='DS' -- Different Sense
fnframe='NA' -- Not Available

"Different Sense" covers cases where a particular VerbNet lemma exists as the word form of one or more Lexical Units in FrameNet, but none that share the lexical semantics of the VerbNet Class member closely enough.

"Not Available" covers cases where a VerbNet lemma doesn't exist as a word form at all in FrameNet.

The 'fnlexent' attribute provides the lexical entry ID number FrameNet assigned to the verb.

VN-FN_roleMapping.xml

The thematic role / frame element mapping file includes the possible role correspondences for the VerbNet Classes and FrameNet Frames that have been mapped. The number of role mappings depends on the particular Class and Frame, and so will vary in number.

The two new attributes included here are the following:

fnrole -- FrameNet frame element (string)
vnrole -- VerbNet thematic role (string)

The structure of an element can be demonstrated by this example:

```
<vncls class='9.1' fnframe='Placing'>
  <roles>
    <role fnrole='Agent' vnrole='Agent'/>
    <role fnrole='Cause' vnrole='Agent'/>
    <role fnrole='Goal' vnrole='Destination'/>
    <role fnrole='Theme' vnrole='Theme'/>
  </roles>
</vncls>
```

Note that the string values for the 'fnframe' attribute can include hyphens and underscores.

VerbNet-PropBank mapping

type_map.xml

The type mapping is provided as a single xml file, containing entries of the form:

```
<predicate lemma="muzzle">
  <argmap pb-roleset="muzzle.01" vn-class="9.9">
    <role pb-arg="1" vn-theta="Destination" />
    <role pb-arg="0" vn-theta="Agent" />
    <role pb-arg="2" vn-theta="Theme" />
  </argmap>
  <argmap pb-roleset="muzzle.01" vn-class="22.4">
    <role pb-arg="1" vn-theta="Patient1" />
    <role pb-arg="0" vn-theta="Agent" />
    <role pb-arg="2" vn-theta="Patient2" />
  </argmap>
</predicate>
```

Each <predicate> entry describes a single verb lemma, and contains one or more <argmap> entries. Each <argmap> entry describes the mapping between arguments for a specific (PropBank roleset, VerbNet class) pair, using one or more <role> entries. Each <role> entry describes the mapping between PropBank ARGn labels and VerbNet thematic roles for a single argument role.

vnprop.txt / vnpbprob.txt

These files provide the mappings between Propbank rolesets and VerbNet classes at the level of individual predicates occurring within the PropBank corpus.

Both files use the same format as PropBank's prop.txt file. In particular, each line describes a single predicate and its arguments.
The columns are as follows:

wsj-filename sentence terminal tagger verb inflection arguments...

Where:

- 'wsj-filename' is the name of the file in merged penn treebank, wsj section

- 'sentence' is the number of the sentence in the file (starting with 0)

- 'terminal' is the number of the terminal in the sentence that the location of the verb. Note that the terminal number counts empty constituents as terminals and starts with 0. This will hold for all references to terminal number in this description.

- 'tagger' is the name of the annotator who performed the mapping.

- 'verb' is a token identifying the verb's PropBank roleset and VerbNet class. It has the form <roleset>;VN=<vncls> where <roleset> is a PropBank roleset and <vncls> is a VerbNet class number.

- 'inflection' consists of 5 characters representing person, tense, aspect, voice, and form of the verb, respectively. See the PropBank documentation for details.

- 'arguments...' is a string representing the annotation associated with a particular argument or adjunct of the proposition. Each proplabel is dash '-' delimited and has the following columns

    1. column for the 'syntactic relation'. See the PropBank documentation for details.

    2. column for the 'label'. In vnprop.txt, this will consist of a VerbNet thematic role label (Agent, Patient, etc); or a PropBank role (ARG0, ARG1, etc) if the role does not have an appropriate mapping target in VerbNet. In vnpbprop.txt, this will have the form "ARG<n>[<theta>]", where <n> is a PropBank role number and <theta> is a VerbNet thematic role; or simply "ARG<n>" if there is no appropriate mapping target. The label "rel" is used to mark the position of the relation word (i.e., the verb).

    3. column for feature. See the PropBank documentation for details.

4.15 Intended application of the resource

The SemLink mappings provide a means to make future improvements to semantic role labeling systems, and will benefit Question-Answering, Information Extraction, inferencing and other NLP applications.

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

The type-to-type mappings are manually assigned. Token to token mappings were created automatically using the type-to-type mappings and are in the process of being hand-corrected.

5 RELEVANT REFERENCES AND OTHER INFORMATION

Fillmore, Charles J., Christopher R. Johnson, and Miriam R.L. Petruck (2003) Background to FrameNet. *International Journal of Lexicography,* 16:235– 250.

Kipper, Karin, Anna Korhonen, Neville Ryant, and Martha Palmer (2008) A large-scale classification of English verbs. *Language Resources and Evaluation Journal,* 42(1):21–40.

Palmer, Martha (2009) Semlink: Linking PropBank, VerbNet and FrameNet In *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon.* Sept. 2009, Pisa, Italy: GenLex-09, 2009

Palmer, Martha, Daniel Gildea, and Paul Kingsbury (2005). The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics,* 31(1):71–106.

Further information about FrameNet is available at: https://framenet.icsi.berkeley.edu/fndrupal/

Further information about SemLink is available at: http://verbs.colorado.edu/semlink/. The unified verb index makes use of these mappings: http://verbs.colorado.edu/verb-index/

Further information about VerbNet is available at: http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

Partner UAIC

COREFERENCE ANNOTATED CORPUS - 1984

1 BASIC INFORMATION

1.1 Corpus composition

The corpus contains noun phrase and coreference annotation for the Romanian version of the 1984 novel of George Orwell.

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

The corpus is represented in XCES format.

1.3 Character encoding

The characters are UTF8 encoded.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Dan CRISTEA and/or Eugen IGNAT
Address: General Berthelot, 16, Iași 700483, Romania
Affiliation: "Al. I. Cuza" University of Iasi, Faculty of Computer Science
Position: Research Assistant
Telephone: + 40 232 201542
e-mail: dcristea@info.uaic.ro or eugen.ignat@info.uaic.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.

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

The archive that will be uploaded on the MetaShare platform contains two XML files.

3.2 Data structure of an entry

The resource is structured into paragraphs, containing one or more sentences. Each sentence is segmented into tokens (equivalent to words or named entities) that can be embedded or not in noun phrase structures. Noun phrases from the original corpus, which are referenced across several sentences, are marked as discourse entities (DEs in the file 1984_RARE.xml or COREF in the file 1984.xml).

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

The corpus contains 6520 sentences and 118328 tokens. There are two versions included. In the file 1984_RARE.xml, for all the noun phrases in the "1984" corpus, discourse entities are automatically annotated using the RARE software.

The file 1984.xml contains a manually annotated corpus of only 2611 tokens. Unlike the previous version, the nested structure of the noun phrases is much poorer in this version, but the corectness of the information it contains is guaranteed.

4 CONTENT INFORMATION

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

This corpus is in Romanian, annotated at discourse entities. However, the text of the resource (the novel "1984" of George Orwell) exists is parallel in another 10 languages (English, Bulgarian Czech, Estonian, Hungarian, Slovene, Latvian, Lithuanian, Serbo-Croatian and Russian), available on the MULTEXT-East project site (not provided in this resource).

4.2 The natural language(s) of the corpus

The language of the corpus is Romanian.

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

The corpus is extracted from a dystopian novel about a fictive society, originally published in 1944.

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 corpus is annotated at part-of-speech, noun phrases and coreference.

The file 1984_RARE.xml follows the format used in the MULTEXT-East project, additionally containing one xml tag:

- **<DE> -** It contains a list of noun phrases ID's that are found in a coreference chain.

```
<DE ID="906" reList="Oro.3.7.38.1.31" />
<DE ID="907" reList="Oro.3.7.38.1.37,Oro.3.7.38.1.42,Oro.3.7.38.1.44,Oro.3.7.38.1.46" />
<DE ID="908" reList="Oro.3.7.38.1.39" />
```

The file 1984.xml contains the coreference informations as an additional attribute for the <NP> element:

- **COREF** – It contains the noun phrases ID which the NP reffers to.

```
<NP ID="NP17" HEADID="TOK55" COREF="NP10">
  <W ID="W65" root="el" pv="Pronoun" Type="pers" Person="third" Gender="masculine" Number="singular" Case="direct" RO="TOK55">el</W>
</NP>
<W ID="W66" type="PERIOD" RO="PTERM_P0">.</W>
<NP ID="NP18" HEADID="TOK56">
  <W ID="W67" root="hol" pv="Noun" Type="common" Gender="masculine" Number="singular" Definiteness="yes" RO="TOK56">Holul</W>
  <NP ID="NP19" HEADID="TOK57" COREF="NP13">
    <W ID="W68" root="bloc" pv="Noun" Type="common" Gender="masculine" Number="singular" Definiteness="yes" RO="TOK57">blocului</W>
  </NP>
</NP>
```

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

A thorough description of the part of speech tags used in the annotation of the "1984" corpus can be found in "MULTEXT-East Morphosyntactic Specifications, Version 3.0"[3].

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

The annotations related strictly to noun phrases are not multi language. But the original corpus is: it is sentence aligned and contains 11 languages. These alignments were manually verified for en-xx language pairs. The rest of the alignments were automatically generated. The current description does not contain these parallel versions.

[3] http://nl.ijs.si/ME/V3/msd/html/msd.html#SECTION05200000000000000000 – section 5 („Application to Romanian")

4.4.4 Attributes and their values (if annotated)

An example of the annotation of the file 1984_RARE.xml is explained below:

```
<p xml:id="Oro.3.7.4">
<s>…</s>
<s xml:id="Oro.3.7.4.5">
  <w>…</w>
  <NP>
    <HEAD>
      <w Case="direct" Definiteness="yes" Gender="masculine" MSD="Npmsry" Number="singular" POS="NOUN" Type="proper" ana="#Npmsry"
lemma="Oceania" xml:id="Oro.3.7.4.5.5">Oceania</w>
    </HEAD>
  </NP>
  <w Case="direct" MSD="Px3--r" POS="PRONOUN" Person="third" Type="reflexive" ana="#Px3--r" lemma="sine"
xml:id="Oro.3.7.4.5.6">se</w>
  <w MSD="Vmii3s" Mood="indicative" Number="singular" POS="VERB" Person="third" Tense="imperfect" Type="predicative" ana="#Vmii3s"
lemma="afla" xml:id="Oro.3.7.4.5.7">afla</w>
  <w MSD="Sp" POS="ADPOSITION" ana="#Sp" lemma="în" xml:id="Oro.3.7.4.5.8">în</w>
  <NP>
    <HEAD>
      <w Case="direct" Definiteness="no" Gender="masculine" MSD="Ncmsrn" Number="singular" POS="NOUN" Type="common"
ana="#Ncmsrn" lemma="război" xml:id="Oro.3.7.4.5.9">război</w>
    </HEAD>
    <w MSD="Sp" POS="ADPOSITION" ana="#Sp" lemma="cu" xml:id="Oro.3.7.4.5.10">cu</w>
    <NP>
      <HEAD>
        <w Case="direct" Definiteness="yes" Gender="masculine" MSD="Npmsry" Number="singular" POS="NOUN" Type="proper"
ana="#Npmsry" lemma="Eurasia" xml:id="Oro.3.7.4.5.11">Eurasia</w>
      </HEAD>
    </NP>
  </NP>
  <w MSD="SCOLON" ana="#SCOLON" xml:id="Oro.3.7.4.5.12">;</w>
  <NP>
    <HEAD>
      <w Case="direct" Definiteness="yes" Gender="masculine" MSD="Npmsry" Number="singular" POS="NOUN" Type="proper" ana="#Npmsry"
lemma="Oceania" xml:id="Oro.3.7.4.5.13">Oceania</w>
    </HEAD>
  </NP>
  <w MSD="Vmii3s" Mood="indicative" Number="singular" POS="VERB" Person="third" Tense="imperfect" Type="predicative" ana="#Vmii3s"
lemma="fi" xml:id="Oro.3.7.4.5.14">era</w>
  <w MSD="Rg" POS="ADVERB" ana="#Rg" lemma="dintotdeauna" xml:id="Oro.3.7.4.5.15">dintotdeauna</w>
  <w MSD="Sp" POS="ADPOSITION" ana="#Sp" lemma="în" xml:id="Oro.3.7.4.5.16">în</w>
  <NP>
    <HEAD>
      <w Case="direct" Definiteness="no" Gender="masculine" MSD="Ncmsrn" Number="singular" POS="NOUN" Type="common"
ana="#Ncmsrn" lemma="război" xml:id="Oro.3.7.4.5.17">război</w>
    </HEAD>
    <w MSD="Sp" POS="ADPOSITION" ana="#Sp" lemma="cu" xml:id="Oro.3.7.4.5.18">cu</w>
    <NP>
      <HEAD>
        <w Case="direct" Definiteness="yes" Gender="masculine" MSD="Npmsry" Number="singular" POS="NOUN" Type="proper"
ana="#Npmsry" lemma="Eurasia" xml:id="Oro.3.7.4.5.19">Eurasia</w>
      </HEAD></s>
<s> …</s>
</p>
<DE ID="52" reList="Oro.3.7.4.5.5,Oro.3.7.4.5.13,Oro.3.7.4.9.20" />
```

Each paragraph is annotated in a <p> tag. Inside each paragraph, the sentences are annotated in <s> tags, with each word in a <w> tag. The <w> tags have MSD-like part of speech attributes annotated in the MULTEXT-East Project Each word has an unique ID in the xml:id attribute. The words are grouped into noun phrases in <NP> tags. Each NP's head is marked in the HEAD tag.

The coreference information is annotated in the DE tag at the end of the file, each DE tag containing an ID and a list of the words that are coreferenced (their unique tags).

For the file 1984.xml, the annotation if detailed below:

```
<NP ID="NP8" HEADID="TOK24">
  <W ID="W30" root="vânt" pv="Noun" Type="common" Gender="masculine" Number="singular" Definiteness="yes" RO="TOK24">vântul</W>
 </NP>
 <NP ID="NP9" HEADID="TOK25" COREF="NP8">
  <W ID="W31" root="care" pv="Pronoun" Person="third" Case="direct" RO="TOK25">care</W>
 </NP>
```

The paragraph and sentence levels are not annotated. The corpus is annotated at word level, each word tag <W> containing part of speech attributes. Words are grouped in noun phrases in NP tags, and if noun groups are coreferenced, they contain a COREF attribute stating the ID of the NP group it refers to.

4.17 Intended application of the corpus

The corpus can be used for testing an anaphora resolution system. The file 1984.xml, being a manually validated file, can also be used to train an anaphora resolution system.

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

The file 1984_RARE.xml was automatically annotated by the RARE system RARE, a rule-based anaphora engine. The file 1984.xml was manually annotated by Cecilia Bolea and Aura Condurache from the Institute of Computer Science, Romanian Academy, Iasi Branch.

5 RELEVANT REFERENCES AND OTHER INFORMATION

Erjavec Tomaž, *MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora,* In Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'2004, ELRA, 2004.

Erjavec Tomaž Harmonised *Morphosyntactic Tagging for Seven Languages and Orwell's 1984*, in Proceedings of the 6th Natural Language Processing Pacific Rim Symposium, Tokyo, pp. 487-492, 2001.

Cristea, D. and Postolache, O. , *How to deal with wicked anaphora*. In Antonio Branco, Tony McEnery and Ruslan Mitkov (eds.): Anaphora Processing: Linguistic, Cognitive and Computational Modelling, Benjamin Publishing Books, 2005.

Orasan, C., Cristea, D., Mitkov, R., Branco, A., *Anaphora Resolution Exercise - An Overview*. LREC-2008. Marakesh, 2008.

Cristea, D., Postolache, O. *Anaphora Resolution: Framework, Creation of Resources, and Evaluation*. In Proceedings of the Fifth International Conference Formal Approaches to South Slavic and Balkan Languages, FASSBL, 18-20 October, 2006, Sofia, Bulgaria, 2006.

ROMANIAN CORPUS 1984 ANNOTATED WITH NOUN PHRASES

1 BASIC INFORMATION

1.1 Corpus composition

This corpus is based on Multext-East cesAna: Nineteen Eighty-Four (Romanian) from MULTEXT-East, Version 4 edition. The original corpus contains 6520 sentences and 118328 tokens annotated with part of speech tags (manually corrected). The delivered corpus contains noun phrase chunks annotations. A part of these are deep noun phrases annotated with an automatic tool (grammar based) and another part is shallow noun and prepositional phrases annotated by human experts.

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

The corpus is represented in XML format.

1.3 Character encoding

The characters are UTF8 encoded.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Radu Simionescu,
Address: General Berthelot, 16, Iași 700483, Romania
Affiliation: "Al. I. Cuza" University of Iasi, Faculty of Computer Science
Position: Research Assistant
Telephone: + 40 740 172558
e-mail: radu.simionescu@info.uaic.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.

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

The archive that will be uploaded on the MetaShare platform will contain two files with .xml extension.

3.2 Data structure

The backbone of the noun phrases annotations is represented by the tokens in the original corpus. Each token can have a parent phrase (noun/prepositional phrase). Each phrase is nested under another phrase or is a top level phrase.

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

The corpus contains 6520 sentences and 118328 tokens. There are two versions included. **One** contains a deep noun phrase structure for all the sentences in the original "1984" corpus, automatically annotated using a recursive grammar.

The other version contains a shallow noun phrase structure, manually annotated for only 1537 sentences (27753 tokens). Unlike the previous version, the nested structure is much poorer in this version, but the corectness of the information it contains is guaranteed.

The files occupy about 20MB on disk.

4 CONTENT INFORMATION

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

This corpus is in Romanian, but the sentences it contains are parallel with another 10 languages from the MULTEXT-East project (English, Bulgarian Czech, Estonian, Hungarian, Slovene, Latvian, Lithuanian, Serbo-Croatian and Russian).

4.2 The natural language(s) of the corpus

The original version of "1984" was written by George Orwell in English.

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

The text is a dystopian novel about a fictive society. It was published in 1944.

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 two files contain two different formats.
The deep NP chunked version is the original 1984 corpus format merged with another two types of xml tags:

- **<NP> -** delimits the boundaries of a noun phrase. It can contain a sequence of words and other nested noun phrases.

```
<w MSD="Sp" POS="ADPOSITION" ana="#Sp" lemma="întru" type="lsplit"…>
<NP>
      <w Case="direct" Gender="feminine" MSD="Tifsr" Number="singular"…>
      <HEAD>
            <w Case="direct" Definiteness="no" Gender="feminine" MSD="Ncfs…>
      </HEAD>
      <w Case="direct" Definiteness="no" Gender="feminine" MSD="Afpfsr…>
      <w MSD="Cc" POS="CONJUNCTION" Type="coordinating" ana="#Cc" lemm…>
      <w Case="direct" Definiteness="no" Gender="…>
      <w MSD="Sp" POS="ADPOSITION" ana="#Sp" lemma="de" xml:id="Oro.1. …>
      <NP>
            <HEAD>
                  <w Case="direct" Definiteness="yes" Gender=…>
            </HEAD>
      </NP>
</NP>
<w MSD="COMMA" ana="#COMMA" xml:id="Oro.1.2.2.1.9">,</w>
<w MSD="Cs" POS="CONJUNCTION" ana="#Cs" lemma="pe_când" type="comp…>
```

- **<HEAD> -** identifies the head word of a noun phrase. Each phrase has a head word.

```
<NP>
      <HEAD>
            <w Case="direct" Definiteness="yes" Gender=…>
      </HEAD>
</NP>
```

The shallow NP chunked version contains only some sentences from the original files. The <c> tag marks a punctuation mark. Each word token contains a „pos" argument representing the part of speech tag. Each token can contain an extra argument, „chunk", which represents a list of parent phrases ids. The margins of a phrase can be found by looking for the sequence of words which all have the phrase's id as parent.

```
<w lemma="un" pos="Tifsr" chunk="Np#2">o</w>
<w lemma="voce" pos="Ncfsrn" chunk="Np#2">voce</w>
<w lemma="melodios" pos="Afpfsrn" chunk="Np#2,Ap#2">melodioasă</w>
<w lemma="citi" pos="Vmii3s" chunk="Vp#1">citea</w>
<w lemma="un" pos="Tifsr" chunk="Np#3">o</w>
<w lemma="listă" pos="Ncfsrn" chunk="Np#3">listă</w>
<w lemma="de" pos="Spsa" chunk="Pp#2">de</w>
<w lemma="cifră" pos="Ncfp-n" chunk="Pp#2,Np#4">cifre</w>
<w lemma="legat" pos="Afpfp-n" chunk="Pp#2,Np#4,Ap#3">legate</w>
<w lemma="de" pos="Spsa" chunk="Pp#3">de</w>
<w lemma="producţie" pos="Ncfsry" chunk="Pp#3,Np#5">producţia</w>
<w lemma="de" pos="Spsa" chunk="Pp#4">de</w>
<w lemma="fontă" pos="Ncfsrn" chunk="Pp#4,Np#6">fontă</w>
```

Identifying the nested structure of the phrases implies detecting their margins first. The smaller phrase is nested direcly under the next bigger one.

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

A thorough description of the part of speech tags used in the original corpus (and in the deliverables described here) can be found in "MULTEXT-East Morphosyntactic Specifications, Version 3.0"[4].

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

The annotations related strictly to noun phrases are not multi language. But the original corpus is: it is sentence aligned and contains 11 languages. These alignments were manually verified for en-xx language pairs. The rest of the alignments were generated automatically. The current deliverable does not contain these parallel versions.

4.4.4 Attributes and their values (if annotated)

This subsection describes the "chunk" attribute used within a word tag <w> in the shallow chunked version. Each phrase is identified in a list by a phrase type (Pp, Np etc) and a number(e.g. Pp#3 or Np#4). The types of chunks are:

- Np – noun phrase
- Pp – prepositional phrase (usually start with a preposition, and contain an Np
- NpCc – a conjunction of noun phrases (usually contains two or more noun phrases and conjunctions in between)
- PpCc – a conjunction of prepositional phrases

4.19 Intended application of the corpus

[4] http://nl.ijs.si/ME/V3/msd/html/msd.html#SECTION05200000000000000000 – section 5 („Application to Romanian")

The corpus can be used for training and testing a noun phrase chunking System.

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

The deep noun phrase version was automatically annotated by a tool based on a recursive grammar. The shallow noun phrase version was manually annotated by Cecilia Bolea and Aura Condurache from the Institute of Computer Science, Romanian Academy

5 RELEVANT REFERENCES AND OTHER INFORMATION

Erjavec Tomaž MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora [Lucrare]. - Paris : In Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'2004, ELRA, 2004.

Erjavec Tomaž Harmonised Morphosyntactic Tagging for Seven Languages and Orwell's 1984 [Report]. - [s.l.] : In Proceedings of the 6th Natural Language Processing Pacific Rim Symposium, Tokyo, pp. 487-492, 2001.

FrRoMWE

1 BASIC INFORMATION

1.1 Corpus composition

The corpus consists of the French and Romanian version of the novel "Madame Bovary" by Gustave Flaubert; the Romanian translation version used is that of Demostene Botez (see references). The Romanian and French versions are aligned at the sentence and word level, the words are morphologically analyzed, and the multi-word expressions in both languages are annotated and aligned.

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

The corpus is represented in XML and text format.

1.3 Character encoding

The characters are UTF8 encoded.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Alex Moruz,
Address: General Berthelot, 16, Iași 700483, Romania
Affiliation: "Al. I. Cuza" University of Iasi, Faculty of Computer Science
Position: Lecturer
Telephone: + 40 232 201549
Fax: + 40 232 201490
e-mail: mmoruz@infoiasi.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.

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

The archive that will be uploaded on the MetaShare platform will contain three different folders (corresponding to the 3 parts of the novel). Each folder will contain different numbers of files with .xml extension and a number of text files. In addition, there is another text file outside the directories that gives the word level alignment.

3.2 Data structure of an entry

Each XML file contains a series of translation units (*tu*), which contain a French and Romanian segment; each segment contains a set of words and MWEs. The text documents in each directory contain a list of all the French and Romanian MWEs in the xml files. The text file outside the directories contains three tab separated columns: TU_ID, FRWord_ID and RoWord_ID.

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

The corpus contains 7405 translation aligned sentences, and needs about 12MB on disk.

4 CONTENT INFORMATION

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

This corpus is a parallel corpus, annotated at morphological and MWE level.

4.2 The natural language(s) of the corpus

The languages of the corpus are standard Romanian and standard French.

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

The text register of the corpus is literary.

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 corpus is annotated at translation unit, sentence, multi word expression and word levels, providing morpho-lexical information. The following example shows the detailed structure with all tags and attributes used in the annotation.

```
<tu id="1">
<seg lang="fr">
<s id="bovary.P1.I.1.1.fr.fr">
    <W ana="Pp1" id="1" lemma="il">Nous</W>
    <W ana="Vmii1p" id="2" lemma="être">étions</W>
    <W ana="Spa" id="3" lemma="à">à</W>
    <W ana="Da-fs" id="4" lemma="le">l'</W>
    <W ana="Ncfs" id="5" lemma="étude">étude</W>
    <c/>
    <W ana="Cs" id="6" lemma="quand">quand</W>
    <W ana="Da-ms" id="7" lemma="le">le</W>
    <W ana="Ncms" id="8" lemma="proviseur">Proviseur</W>
    <W ana="Vmis3s" id="9" lemma="entrer">entra</W>
    <W ana="Vmps-s" id="10" lemma="suivre">suivi</W>
    <W ana="Spd" id="11" lemma="de">d'</W>
    <W ana="Da-ms" id="12" lemma="un">un</W>
    <W ana="Af-ms" id="13" lemma="nouveau">nouveau</W>
    <W ana="Vmps-s" id="14" lemma="habiller">habillé</W>
    <W ana="Sp" id="15" lemma="en">en</W>
    <W ana="Ncms" id="16" lemma="bourgeois">bourgeois</W>
    <W ana="Cc" id="17" lemma="et">et</W>
    <W ana="Spd" id="18" lemma="de">d'</W>
    <W ana="Da-ms" id="19" lemma="un">un</W>
    <MWE DEF="" ID="0" OBS="" OTHER_TYPE="" TYPE="COLLOCATION">
    <HEAD ID="107">
    <W ana="Ncms" id="20" lemma="garçon">garçon</W>
    </HEAD>
    <W ana="Spd" id="21" lemma="de">de</W>
    <W ana="Ncfs" id="22" lemma="classe">classe</W>
    </MWE>
    <W ana="Pr" id="23" lemma="qui">qui</W>
    <W ana="Vmii3s" id="24" lemma="porter">portait</W>
    <W ana="Da-ms" id="25" lemma="un">un</W>
    <W ana="Af-ms" id="26" lemma="grand">grand</W>
    <W ana="Ncms" id="27" lemma="pupitre">pupitre</W>
    <c/>
```

```
</s>
</seg>
<seg lang="ro">
<s id="bovary.P1.I.1.1.ro.ro">
    <W ana="Vmii1" id="1" lemma="fi">Eram</W>
    <W ana="Spsa" id="2" lemma="în">în</W>
    <W ana="Ncfsry" id="3" lemma="sală">sala</W>
    <W ana="Spsa" id="4" lemma="de">de</W>
    <W ana="Ncfsrn" id="5" lemma="meditaţie">meditaţie</W>
    <W ana="Rw" id="6" lemma="când">când</W>
    <W ana="Ncmsry" id="7" lemma="director">directorul</W>
    <W ana="Vmis3s" id="8" lemma="intra">intră</W>
    <W ana="Vmp--sm" id="9" lemma="urma">urmat</W>
    <W ana="Spsa" id="10" lemma="de">de</W>
    <W ana="Timsr" id="11" lemma="un">un</W>
    <W ana="Ncms-n" id="12" lemma="elev">elev</W>
    <W ana="Afpms-n" id="13" lemma="nou">nou</W>
    <c/>
    <W ana="Afpms-n" id="14" lemma="îmbrăcat">îmbrăcat</W>
    <W ana="Rgp" id="15" lemma="orăşeneşte">orăşeneşte</W>
    <c/>
    <W ana="Crssp" id="16" lemma="şi">şi</W>
    <W ana="Spsa" id="17" lemma="de">de</W>
    <W ana="Timsr" id="18" lemma="un">un</W>
    <MWE DEF="" ID="0" OBS="" OTHER_TYPE="" TYPE="COLLOCATION">
    <W ana="Ncms-n" id="19" lemma="băiat">băiat</W>
    <W ana="Spsa" id="20" lemma="de">de</W>
    <W ana="Ncms-n" id="21" lemma="serviciu">serviciu</W>
    </MWE>
    <c/>
    <W ana="Pw3--r" id="22" lemma="care">care</W>
    <W ana="Vmii3s" id="23" lemma="aduce">aducea</W>
    <W ana="Timsr" id="24" lemma="un">un</W>
    <W ana="Ncms-n" id="25" lemma="pupitru">pupitru</W>
    <W ana="Spsa" id="26" lemma="din">din</W>
    <W ana="Pd3fpr" id="27" lemma="acela">cele</W>
    <W ana="Afp-p-n" id="28" lemma="mare">mari</W>
    <c/>
</s>
</seg>
</tu>
```

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

The corpus contains morpho-syntactic information (MSD) which has been assigned automatically with RACAI's high accuracy TTL tagger (Ion, 2007; Tufis et al., 2008). It also contains information about multi-word expressions.

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

The corpus is a bilingual version of the novel "Madame Bovary" by Gustave Flaubert. It is automatically aligned at the sentence and word

level using the RACAI word aligner, and approximately 20% of the word level alignment is manually validated.

4.4.4 Attributes and their values (if annotated)

The *tu* tag, which can contain *seg* elements, has one attribute:
- *id* specifies the position of the textual unit in corpus

The *seg* tag can have a *lang* attribute and an *s* element. An *s* element has an *id* attribute, which gives the sentence position in the novel, and may have *W* and *MWE* elements. A *MWE* element contains *W* and *HEAD* (the *HEAD* element contains the word that is the head of the MWE) elements, and has the attributes *ID*, *DEF*, *TYPE* (the type of the MWE), OTHER_TYPE (the secondary type of the *MWE*, if any) and *OBS* (used by annotators for internal observations).
The *W* element has the attributes *id*, *lemma* (which specifies the lemma form of the word) and *ana* (the MSD tag for the word). The MSDs follows the Multext-East specifications (Erjavec, 2004). For Romanian there are 614 different MSDs (Tufis et al. 1997). They have been slightly modified (new tags for named entities have been added) are largely described in (Tufis and Ion, 2006)

4.21 Intended application of the corpus

Due to the mark-up accuracy, the corpus can be used for building robust statistical language models. It can also be used as a reference corpus for Romanian in various corpus specific types of investigation: quantitative analysis, collocation extraction, grammar induction, etc.

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

The annotations are reliable. The word alignment has been manually validated for 20% of the text (particularly the beginning of the text) and the MWE elements have been manually extracted for the French version and automatically determined and manually checked (by means of word alignment) for Romanian.

5 RELEVANT REFERENCES AND OTHER INFORMATION

Tomaz Erjavec. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the 4th LREC Conference*, LREC'04, Lisabona, pp. 1535 - 1538,

Dan Tufiş, Barbu A.M., Pătraşcu V., Rotariu G., Popescu C. 1997.”Corpora and Corpus-Based Morpho-Lexical Processing”. In Dan Tufiş, P. Andersen (eds.) “Recent Advances in Romanian Language Technology”, Editura Academiei, pp. 35-56.

Maria Husarciuc, 2009. *Echivalarea în limba română a unităţilor frazeologice infinitivale din limba franceză*, în Lucrările Atelierului Resurse lingvistice şi instrumente pentru prelucrarea limbii române (Iaşi, 19-21 noiembrie 2008), Editura Universităţii „Alexandru Ioan Cuza" Iaşi, 2009, p. 115-124.

Maria Husarciuc, Anca-Diana Bibiri, 2010. *Phraseological units between national specificity and universality. Contrastive approach applied to French and Romanian languages*, in Jarmo Korhonen et al. (Hg.), Phraseologie. global -- areal – regional, Akten der Konferenz EUROPHRAS 2008 vom 13.-16.8.2008 in Helsinki, Gunter Narr Verlag, 2010, p. 333-338 (ISBN 978-3-8233- 6508-2).

Diana Trandabăţ, Maria Husarciuc, 2008. *Romanian Semantic Role Resource*, in Proceedings of LREC 2008, May 26 - June 1, ELRA - European Language Resources Association, Marrakech, Morocco, 2008, p. 2806-2810 (ISBN: 2-9517408-4-0).

Sources of the corpora:

Gustave Flaubert, 1857. *Madame Bovary*, electronic format available within the Gutenberg Project at http://www.gutenberg.org/etext/14155.

Gustave Flaubert, 1913. *Dictionnaire des idées reçues*, electronic version realized within the Gutenberg project, available at http://www.gutenberg.org/etext/14156.

Gustave Flaubert, 1915. *Doamna Bovary*, translation by Ludovic Dauş, 2nd revised edition, Ed. Minerva, Bucharest, 1915.

Gustave Flaubert, 1968. *Doamna Bovary*, translated in Romanian by Demostene Botez, Ed. for Universal Literature, Bucharest, 1968.

ROMANIAN QUESTION ANSWERING CORPUS

1 BASIC INFORMATION

1.1 Corpus composition

The corpus consists of 200 questions of type: FACTOID, DEFINITION and LIST.

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

The corpus is represented in XML format.

1.3 Character encoding

The characters are UTF8 encoded.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Adrian Iftene,
Address: General Berthelot, 16, Iași 700483, Romania
Affiliation: "Al. I. Cuza" University of Iasi, Faculty of Computer Science
Position: Lecturer
Telephone: + 40 232 201549
Fax: + 40 232 201490
e-mail: adiftene@infoiasi.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.

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

The archive that will be uploaded on the MetaShare platform will contain one file with .xml extension.

3.2 Data structure of an entry

The XML file is structured in question groups. In a group is one question in Romanian. For each question we have five tags: the *question string*, the *focus*, the *keywords*, the *question type* and the *answer type*.

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

The corpus contains 200 questions and needs about 48 kB for disk storage.

4 CONTENT INFORMATION

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

This corpus has questions in Romanian.

4.2 The natural language(s) of the corpus

The languages of the corpus are standard Romanian, orthography being compliant with the current Romanian Academy norms.

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

The questions have the answers in Romanian Wikipedia or in JRC-Acquis domains.

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 corpus is annotated at questions group. The following example shows the detailed structure with all tags and attributes used in the annotation.

```
<GOLD_QA_RO>
        <q q_id="001">
          <string>Ce procent al populaţiei din România foloseşte Internetul?</string>
          <focus>procent</focus>
          <keywords>populaţie România Internetul</keywords>
          <questionType>FACTOID</questionType>
          <answerType>MEASURE</answerType>
        </q>
        <q q_id="002">
          <string>Numiţi trei scriitori români.</string>
          <focus>scriitor</focus>
          <keywords>numi român trei</keywords>
          <questionType>LIST</questionType>
          <answerType>PERSON</answerType>
        </q>
        <q q_id="003">
          <string>Ce se înţelege prin reducerea TVA?</string>
          <focus>reducerea</focus>
          <keywords>TVA</keywords>
          <questionType>DEFINITION</questionType>
          <answerType>OTHER</answerType>
        </q>
…
</GOLD_QA_RO>
```

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

The corpus is simple text and it is not parsed with any tool.

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

The texts of the questions are in Romanian.

4.4.4 Attributes and their values (if annotated)

The *q* tag corresponds to a question and it has five tags:
- *string* which represents the question.
- *focus* specifies the most relevant word from the question.
- *keywords* has the important words from current question (verbs, nouns, adjectives and named entities).
- *questionType* which specify the question type (can be FACTOID, DEFINITION or LIST).
- *answerType* is the type of answer (can be OTHER, LOCATION, PERSON, ORGANIZATION, OBJECT, etc.).

4.23 Intended application of the corpus

The corpus can be used for testing a question processing component from a question answering system for Romanian.

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

The corpus was manual annotated by a native Romanian speaker and it was used with success in CLEF competitions from 2006 to 2011.

5 RELEVANT REFERENCES AND OTHER INFORMATION

Iftene, A., Gînscă, A. L., Moruz, A., Trandabăţ, D., Husarciuc, M. 2011. Question Answering for Machine Reading Evaluation on Romanian and English Languages. Notebook Paper for the CLEF 2011 LABs Workshop, ISBN 978-88-904810-1-7, ISSN 2038-4726, 19-22 September, Amsterdam, Netherlands.

Iftene, A., Trandabăţ, D., Moruz, A., Husarciuc, M. 2010. Question Answering on Romanian, English and French Languages. Notebook Paper for the CLEF 2010 LABs Workshop, ISBN 978-88-904810-0-0, ISSN 2038-496322-23, 22-23 September, Padua, Italy.

Iftene, A., Trandabăţ, D., Pistol, I., Moruz, A., Husarciuc, M., Cristea, D. 2009. UAIC Participation at QA@CLEF2008. In Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers. Lecture Notes in Computer Science. ISBN: 978-3-642-04446-5. Vol. 5706/2009, pp. 448-451, ISBN 978-3-540-74998-1, ISSN 0302-9743 (Print) 1611-3349 (Online).

Iftene, A., Trandabăţ, D., Pistol, I., Moruz, A., Balahur-Dobrescu, A., Cotelea, D., Dornescu, I., Drăghici, I., Cristea, D. 2008. UAIC Romanian Question Answering system for QA@CLEF. In Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers. Editors Peters,C., Jijkoun, V., Mandl, T., Muller, H., Oard, D.W., Penas, A., Petras, V., Santos, D. Lecture Notes in Computer

Science, LNCS 5152, ISBN: 978-3-540-85759-4. Pp. 336-343, Springer-Verlag Berlin Heidelberg, May 2008

Puşcaşu, G., Iftene, A., Pistol, I., Trandabăţ, D., Tufiş, D., Ceauşu, A., Ştefănescu, D., Ion, R., Dornescu, I., Moruz, A., Cristea, D. 2007. Cross-Lingual Romanian to English Question Answering at CLEF 2006. In Evaluation of Multilingual and Multi-modal Information Retrieval,7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 20-22, 2006, Revised Selected Papers. Editors Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., Rijke, M., Stempfhuber, M. ISSN: 0302-9743, ISBN (10): 3-540-74998-5, ISBN (13): 978-3-540-74998-1. Lecture Notes in Computer Science, Vol. 4730, pp. 385-394, Springer.

ROMANIAN FRAMENET

1 BASIC INFORMATION

1.1 Corpus composition

The corpus contains multilingual (English and Romanian) semantic role annotations for 1508 sentences. Using semantic role analysis, we are able to answer questions such as: “What roles do entities play in different contexts?” or “When, why, where or how an event takes place?” A semantic role represents the relationship between a predicate (either verbal or nominal, annotated as the target) and an argument (annotated with the specific semantic role).

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

The corpus is represented in one file in XML format.

1.3 Character encoding

The characters are UTF8 encoded.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Diana Trandabăț
Address: Alexandru Lăpușneanu 14, room 201
Affiliation: Faculty of Computer Science, University AL. I. Cuza of Iasi, Romania
Position: postdoctoral researcher
Telephone: +40 232 201771
e-mail: dtrandabat@info.uaic.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.

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

The archive that will be uploaded on the MetaShare platform contains one XML file.

3.2 Data structure of an entry

An entry contains the English sentence and its translation into Romanian. The resource contains 754 entries, each with one English and one Romanian sentence, summing up to 1508 sentences. Beside the text of the sentences, each entry contains the annotation of the sentences at three levels: semantic role (named frame elements as in FrameNet – see Fillmore et al., 2002), grammatical function and phrase type. The English sentences are extracted from the English FrameNet with their annotations for the three levels. For the annotation of the Romanian sentences, the annotation import program described in (Trandabăț 2010) and (Trandabăț 2011a) was used. After the automatic transfer of semantic roles from English to Romanian, the Romanian obtained roles were manually validated, so that the semantic role annotation for the Romanian sentences can be also considered a gold annotation. The annotations for the other two levels (grammatical functions and phrase type) were automatically imported as well from English to Romanian, but are not yet validated.

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

The corpus contains 754 entries, summing up to 1508 sentences. The sentences are not tokenized, so the total number of tokens is not known.

4 CONTENT INFORMATION

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

This corpus is a parallel multilingual (English-Romanian) corpus, gold annotated for semantic roles.

4.2 The natural language(s) of the corpus

The corpus' languages are English and Romanian.

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

The English sentences are extracted from FrameNet, a semantic role resource whose examples are attestations taken from naturalistic corpora, rather than constructed by a linguist or lexicographer. The main corpus FrameNet uses is the 100-million-word British National CorpusBritish National Corpus (BNC), which is both large and balanced across genres (editorials, textbooks, advertisements, novels, sermons, etc.). FrameNet also uses U.S. newswire texts provided by the Linguistic Data Consortium, and recently the newly released initial part of the American National Corpus. The Romanian text is a translation of the English sentences.

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 corpus is annotated at sentence level and, for each sentence, semantic roles are annotated. The following example shows the detailed structure with all tags and attributes used in the annotation.

```
<annotationSet ID="615 " status="AUTO">
 <layers>
  <layer name="FE" rank="1" >
   <labels>
    <label name="Protagonist"  words="Partenerii de pescuit John Armstrong , 52 ,
și James Scrimgeour , 62 ," />
   <label_en name="Protagonist"  words="FISHING pals John Armstrong , 52 , and
James Scrimgeour , 62 ," />
   <label name="Time"  words="când un val neobișnuit le- a scufundat barca lângă
mal la Cresswell" />
   <label_en name="Time"  words="when a freak wave sank their boat near the
shore at Cresswell" />
   </labels>
  </layer>
   <layer name="GF" >
   <labels>
    <label name="Ext"  words="Partenerii de pescuit John Armstrong , 52 , și James
Scrimgeour , 62 ," />
    <label_en name="Ext"  words="FISHING pals John Armstrong , 52 , and James
Scrimgeour , 62 ," />
    <label name="Dep"  words="când un val neobișnuit le- a scufundat barca lângă
mal la Cresswell" />
    <label_en name="Dep"  words="when a freak wave sank their boat near the
shore at Cresswell" />
   </labels>
  </layer>
   <layer name="PT" >
   <labels>
```

<label name="NP" words="Partenerii de pescuit John Armstrong , 52 , şi James Scrimgeour , 62 ," />
<label_en name="NP" words="FISHING pals John Armstrong , 52 , and James Scrimgeour , 62 ," />
<label name="Sinterrog" words="când un val neobişnuit le- a scufundat barca lângă mal la Cresswell" />
<label_en name="Sinterrog" words="when a freak wave sank their boat near the shore at Cresswell" />
</labels>
</layer>
<layer name="Sent" /> <layer name="Other" />
<layer name="Target" >
<labels>
<label name="Target" words="au înecat" />
<label_en name="Target" words="drowned" />
</labels>
</layer>
<layer name="Verb" />
<layer name="FE" rank="2" >
<labels>
<label name="Cause" words="când un val neobişnuit le- a scufundat barca lângă mal la Cresswell" />
<label_en name="Cause" words="when a freak wave sank their boat near the shore at Cresswell" />
</labels>
</layer>
</layers>
<sentence>
<text>Partenerii de pescuit John Armstrong , 52 , şi James Scrimgeour , 62 , s-au înecat când un val neobişnuit le- a scufundat barca lângă mal la Cresswell , Tyneside .</text>
<text_en>Fishing pals John Armstrong , 52 , and James Scrimgeour , 62 , drowned when a freak wave sank their boat near the shore at Cresswell , Tyneside </text_en>
</sentence>
</annotationSet>

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

The corpus contains semantic role tags which are manually validated for both languages; for grammatical functions and phrase type tags validated only for English. The tagset's description can be found on the FrameNet project's webpage - https://framenet.icsi.berkeley.edu/.

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

The corpus contains sentences in English and Romanian aligned at sentence level (each entry contains an aligned pair of sentences). The semantic roles, the grammatical functions and the phrase types are also aligned between the two languages.

4.4.4 Attributes and their values (if annotated)

The *annotationSet* tag encloses each pair of annotated sentences, and can have several enclosed tags:

- *layers* which encloses the annotations for the three different layers
- *sentence* which enclosed the two sentences: the English sentence in the *text_en* tag and the Romanian sentence in the *text* tag.

The *layers* tag contains the different annotation levels in separate *layer* tags. Each *layer* tag has a *name* of the layer as its attribute, and contains a *labels* tag with several *label* tags (for the Romanian annotation) and *label_en* tags (for the English annotation) inside it. Each *label* tag has a *name* and a list of *words*. The labels inside the layer(s) named FE represent the semantic roles, the label(s) inside the Target layer represent the predicate with whom the semantic roles are linked.

4.25 Intended application of the corpus

Due to the mark-up accuracy, the corpus can be used for building robust statistical language models for semantic role labeling - see (Trandabăţ, 2011b).

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

The corpus contains manual annotations for the English sentences. For the Romanian sentences, the English annotation was automatically imported using the method described in (Trandabăţ, 2010) and (Trandabăţ, 2011a). For the semantic role level, the automatically transferred annotation was manually validated, so it is highly reliable. For the other two levels (grammatical functions and phrase type) the automatically imported annotation was not validated.

5 RELEVANT REFERENCES AND OTHER INFORMATION

Charles J. Fillmore, Collin F. Baker, and Hiroaki Sato. (2002) *The FrameNet Database and Software Tools* Proceedings of the Third International Conference on Languag Resources and Evaluation. Vol. IV. Las Palmas: LREC-2002.

Diana Trandabăţ. (2011a) *Towards automatic cross-lingual transfer of semantic annotation*, in 6e Rencontres Jeunes Chercheurs en Recherche d'Information (RJCRI) 2011, 16-18 March, Avignon, France.

Diana Trandabăţ. (2010) *Natural Language Processing Using Semantic Frames*, PhD Thesis, University Al. I. Cuza Iasi, Romania.

Diana Trandabăţ, Maria Husarciuc (2008) *Romanian Semantic Role Resource*, Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco, May, 28 - 30, 2008, ISBN 2-9517408-4-0, pp. 2806 - 2810.

Diana Trandabăţ (2007) *Semantic Frames in Romanian Natural Language Processing*, Proceedings of the North American Chapter of the Association for Computational Linguistics NAACL-HLT 2007, Companion Volume: Doctoral Consortium, ACL, April 2007, Rochester, New York, USA, pp. 29-32, ISBN 1-932432-92-2.

Diana Trandabăţ (2011b) *Extracting Semantic Information from Texts*, in Proceedings of the 13th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC2011, Timisoara, Romania.

ROMANIAN SYNTACTIC ANNOTATED CORPUS - RO-FDGBank

1 BASIC INFORMATION

1.1 Corpus composition

The corpus consists of 3501 sentences from different genres: literature (literary texts from textbook exercises = 521; "1984" by George Orwell = 975); FrameNet (1094 texts translated from the English FrameNet); Wikipedia (396 texts) and Aquis Communitaire (515 texts). The total number of annotated words in the entire corpus is 82882.

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

The corpus is represented in XML format.

1.3 Character encoding

The characters are UTF8 encoded.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Cenel-Augusto PEREZ,
Address: General Berthelot, 16, Iași 700483, Romania
Affiliation: "Al. I. Cuza" University of Iasi, Faculty of Computer Science
Position: Associated Teacher
Telephone: + 40 232 201030
Fax: + 40 232 201490

e-mail: augusto.perez@infoiasi.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.

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

The archive that will be uploaded on the MetaShare platform will contain four different folders (corresponding to the 4 types of texts it covers): Literature, FrameNet, Wikipedia and Aquis Communitaire. Each folder will contain different numbers of files with .xml extension.

3.2 Data structure of an entry

This is not relevant as the corpus is provided as a text file. It is structured in paragraphs, containing one or more sentences. Each sentence is segmented into tokens (equivalent to word or named entities) that can be embedded or not in chunk structures.

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

The corpus contains 3501 sentences and needs about 7,70 MB for disk storage.

4 CONTENT INFORMATION

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

This corpus is a balanced monolingual, heavily annotated corpus.

4.2 The natural language(s) of the corpus

The languages of the corpus are standard Romanian, orthography being compliant with the current Romanian Academy norms.

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

The text registers represented into the corpus are: journalistic language as used in the daily newspapers, official language as used in legal documents, artistic language as used in literary fiction, Romanian texts from Wikipedia.

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 corpus is annotated at paragraph, sentence, constituent group and word levels, providing morpho-lexical and syntactic information. The following example shows the detailed structure with all tags and attributes used in the annotation.

```
<?xml version='1.0' encoding='UTF-8'?>
<treebank id="FrameNetRo">
 <sentence id="1" parser="" user="Augusto" date="2010-00-15">
  <word id="1" form="Blîndeţea" lemma="Blîndeţea" postag="Ncfsry" head="5" chunk="" deprel="sbj."/>
  <word id="2" form="lui" lemma="lui" postag="Tf-so" head="3" chunk="" deprel="det."/>
  <word id="3" form="april" lemma="April" postag="Np" head="1" chunk="" deprel="a.subst."/>
  <word id="4" form="ne" lemma="eu" postag="Pp1-pa--------w" head="5" chunk="" deprel="c.i."/>
  <word id="5" form="umplea" lemma="umple" postag="Vmii3s" head="0" chunk=""/>
  <word id="6" form="tuturor" lemma="tot" postag="Pi3-po" head="5" chunk="" deprel="c.i."/>
  <word id="7" form="sufletele" lemma="suflet" postag="Ncfpry" head="5" chunk="" deprel="c.d."/>
  <word id="8" form="ca" lemma="ca" postag="Rc" head="5" chunk="" deprel="c.c.m."/>
  <word id="9" form="un" lemma="un" postag="Timsr" head="10" chunk="" deprel="det."/>
  <word id="10" form="abur" lemma="abur" postag="Ncms-n" head="8" chunk="" deprel="prep."/>
  <word id="11" form="de" lemma="de" postag="Spsa" head="13" chunk="" deprel="a.subst."/>
  <word id="12" form="melancolie" lemma="melancolie" postag="Ncfsrn" head="11" chunk="" deprel="prep."/>
  <word id="13" form="şi" lemma="şi" postag="Crssp" head="10" chunk="" deprel="coord."/>
  <word id="14" form="de" lemma="de" postag="Spsa" head="13" chunk="" deprel="a.subst."/>
  <word id="15" form="doruri" lemma="dor" postag="Ncfp-n" head="14" chunk="" deprel="prep."/>
  <word id="16" form="nedesluşite" lemma="nedesluşit" postag="Afpfp-n" head="15" chunk="" deprel="a.adj."/>
  <word id="17" form="." lemma="." postag="PERIOD" head="5" chunk="" deprel="punct."/>
 </sentence>
```

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

The corpus contains: a)morpho-syntactic information (MSD-style) which has been assigned with the TTL tagger (Ion, 2007; Tufis et al., 2008) and b)syntactic relations which have been assigned with TreeAnnotator (http://students.info.uaic.ro/~mmoruz/FDAnnotator/TreeAnnotator.tgz.).

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)

The are two main tags in this corpus: *sentence* and *word* tags.

The *sentence* tag marks all the sentences from the treebank and has as attributes the *id* of the respective sentence and the *name* of the annotator that modified the respective sentence.
The word tag marks the individual words from the sentence and has the following attributes: 1. *id* of the word; 2. *form* of the respective word; 3. *lemma*; 4. part of speech (*postag*); 5. *head*, which shows the id of the head of the respective word and *deprel*, which shows the name of the functional dependency relation between the word and its head.

4.27 Intended application of the corpus

The corpus can be used for building robust statistical language models. It can also be used as a reference corpus for Romanian in various corpus specific types of investigation: quantitative analysis, collocation extraction, grammar induction, etc.

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

The annotations are highly reliable. The sentences mark-up has been fully validated. The MSD tagging accuracy is at least 98%. The syntactic annotation has been achieved based on the *Romanian Academy's Grammar* rules and some conventions defined over the group's agreement.

5 RELEVANT REFERENCES AND OTHER INFORMATION

Perez, Cenel-Augusto, Linguistic resources for the processing of the Romanian lamnguage (in Romanian), PhD thesis, University Al. I. Cuza of Iasi, to be released.

Radu Ion. *Word Sense Disambiguation Methods Applied to English and Romanian*. PhD thesis (in Romanian). Romanian Academy, Bucharest, 2007.

Dan Tufiş, Elena Irimia, Radu Ion, and Alexandru Ceauşu. *Unsupervised Lexical Acquisition for Part of Speech Tagging*. In Proceedings of the 6th Language Resources and Evaluation Conference - LREC 2008, Marrakech, Morocco, May 2008. ELRA – European Language Ressources Association.

RoSemClass

1. BASIC INFORMATION

1.1.Lexicon type: semantic classes of lexical items for political discourse analysis.

The current approaches in analysing the political language are based on Natural Language Processing (NLP) techniques designed to investigate lexical-semantic aspects of the discourse. An important NLP problem is the text categorization. One of the important steps in our research was the classification task of the political lexicon in electoral context.

The lexicon contains a collection of lemmas and stems from various POS categories (explicitly marked only in the file classes_eminescu.xml): verb, noun, adjective and adverb. In the context of the lexical-semantic analysis, the pronouns, numerals, prepositions and conjunctions, considered to be semantically empty, have been left out. Our version includes 30 semantic classes, chosen to fit optimally with the necessities of interpreting the political discourse. The semantic classes are partially placed in a hierarchy, to be found in the file semantic_classes_hierarchy.xml

1.2.Representation of the lexicon

RoSemClass is distributed as XML files.

1.3.Character encoding: the characters have been encoded in UTF8.

2. ADMINISTRATIVE INFORMATION

*2.1.*Contact persons

Name: Daniela GÎFU and Mihaela MOCANU
Address: Gral Berthelot no. 16, 700483
Affiliation: Faculty of Computer Science, “Alexandru Ioan Cuza” University
Position: Postdoctoral Researcher
Telephone: +40-232-201 724 (Daniela GÎFU) or +40-232-202 345 (Mihaela MOCANU)
e-mail: daniela.gifu@info.uaic.ro or mocanu.mihaela@uaic.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

The resource contains three XML files: the file semantic_classes_hierarchy.xml presents the hierarchy between the semantic classes used, the file classes_general.xml contains the semantic classes of general words, and the file classes_eminescu.xml contains the semantic classes of the frequent words found in political discourses of Eminescu (the greatest Romanian poet).

3.2. Data structure of an entry

An entry of the lexicon contains a word form, a lemma or a stem, with its associated semantic class(es).

3.3. Lexicon size

The file classes_general.xml contains 5730 stems/lemmas, while the file classes_eminescu.xml contains 29829 lexical items, with the following distribution:

5338 lexical items – social class (2)
323 lexical items – family class (3)
14 lexical items – friend class (4)
6436 lexical items – human class (5)
2137 lexical items – positive emotional (7)
4820 lexical items – negative emotional (8)
449 lexical items – cognitive (12)
209 lexical items – insight (13)
1389 lexical items – cause (14)
479 lexical items – doubt (15)
720 lexical items – inhibition (17)
573 lexical items – sight (19)
366 lexical items – hearing (20)
20 lexical items – feeling (21)
505 lexical items – work (23)
2315 lexical items – achievement (24)
1107 lexical items – failure (25)
242 lexical items – leisure (26)
926 lexical items – money (28)
1363 lexical items – religion (29)

The needed disk space is about 2.10MB.

4. CONTENT INFORMATION

4.1. The natural language of the lexicon

The language of the lexicon is Romanian.

4.2. Entry Type

For the file classes_general.xml, an entry has the form:

<word stem="acasă" classes="30,7,27"/> or*

<word lemma="complot" classes="30,10"/>

For the file classes_eminescu.xml, each entry has the form

<word form="România" lemma="România" freq="1759" classes="2" />

4.3. Attributes and their values

For the file classes_general.xml, an *word* entry contains attributes for each lemma or stem, its corresponding semantic class(es) are marked. For the file classes_eminescu.xml, an *word* entry contains the word form, its lemma, its frequency in the corpus of political discourses of Eminescu, and the semantic class.

4.4. Coverage of the lexicon

The file classes_general.xml covers the actual political language, while the file classes_eminescu.xml covers the Romanian political language in the second half of the XIX century.

4.5. Intended application of the lexicon

RoSemClass has been used for socio-political discourse analysis. We were mainly interested to determine those political attitudes which were able to influence the voting decision of the electorate.

4.6. POS assignment

The resource contains only word forms that classify in different categories, with no POS or other processing.

4.7.Reliability

The lexicon has used the lemmatization service offered by RACAI as webservice, followed by manual semantic annotation.

5. RELEVANT REFERENCES AND OTHER INFORMATION

Dan Cristea, Dan Tufiş: "Linguistic resources and information technologies applied to the Romanian language" (in Romanian), in O. Ichim şi, F.-T. Olariu (eds.): "Identitatea limbii şi literaturii române în perspectiva globalizării", Academia Română, Institutul de Filologie Română „A. Philippide", Ed. Trinitas, Iaşi, 2002.

Daniela Gîfu, Dan Cristea: "Computational Techniques in Political Language Processing: AnaDiP-2011" in J.J. Park, L.T. Yang, and C. Lee (Eds.): "FutureTech 2011", Part II, CCIS 185, pp. 188–195, Springer-Verlag Berlin Heidelberg 2011.

Mocanu, Mihaela. "Semiotic analysis of the political language of Eminescu" (in Romanian), PhD Thesis, Faculty of Philosophy and Socio-politics sciences

ROMANIAN TEXTUAL ENTAILMENT CORPUS

1 BASIC INFORMATION

1.1 Corpus composition

The corpus consists of 200 pairs of text-hypothesis in Romanian for which we know the human decision about entailment relation. The text-hypothesis pairs are uniform distributed into three classes of decisions: 72 Entailment, 60 Contradiction and 78 Unknown.

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

The corpus is represented in XML format.

1.3 Character encoding

The characters are UTF8 encoded.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Adrian Iftene,
Address: General Berthelot, 16, Iași 700483, Romania
Affiliation: "Al. I. Cuza" University of Iasi, Faculty of Computer Science
Position: Lecturer
Telephone: + 40 232 201549
Fax: + 40 232 201490
e-mail: adiftene@infoiasi.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.

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

The archive that will be uploaded on the MetaShare platform will contain one file with .xml extension.

3.2 Data structure of an entry

The XML file is structured in entailment groups. In a group are one text and 10 hypotheses. For every hypothesis we have also the entailment relation between it and the text from current group.

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

The corpus contains 20 texts with 200 corresponding hypotheses and needs about 35 kB for disk storage.

4 CONTENT INFORMATION

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

This corpus is in Romanian.

4.2 The natural language(s) of the corpus

The languages of the corpus are standard Romanian, orthography being compliant with the current Romanian Academy norms.

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

The texts are from Romanian Wikipedia and Romanian newspapers and the hypotheses are related to these 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)

The corpus is annotated at entailment group, where for every text we can found 10 hypotheses. The following example shows the detailed structure with all tags and attributes used in the annotation.

```
<entailment-corpus>
    <entailment-group id_group="1">
      <text>Mi-e dor de Ștefan Iordache. Golul rămas prin plecarea lui e uriaș. A fost atât de important el, lucru știut
      și în viață, dar mai ales acum când nu mai e; eu raportez tot ce mi se întâmplă la acest gol. Radu Beligan,
      parcimonios foarte în aprecieri, dar care nu se joacă însă cu valoarea, spune despre el: „Știu că oamenii sunt
      reticenți la auzul unei declarații absolute și totuși nu ezit să spun că îl consider pe Ștefan Iordache cel mai mare
      actor al unei generații de talente uimitoare, care apare o dată la 50 de ani.</text>
      <hypothesis id_hypothesis="1" entailment="Yes">Ștefan Iordache a murit.</hypothesis>
      <hypothesis id_hypothesis="2" entailment="Yes">Ștefan Iordache a fost actor.</hypothesis>
      <hypothesis id_hypothesis="3" entailment="Yes">Beligan face o afirmație despre Iordache.</hypothesis>
      <hypothesis id_hypothesis="4" entailment="No">Radu Beligan nu-l cunoaște pe Iordache.</hypothesis>
      <hypothesis id_hypothesis="5" entailment="No">Ștefan Iordache a împlinit 50 de ani.</hypothesis>
      <hypothesis id_hypothesis="6" entailment="No">Ștefan Iordache este un actor în viață.</hypothesis>
      <hypothesis id_hypothesis="7" entailment="Unknown">Ștefan Iordache a fost cel mai mare actor al unei
      generații de talente uimitoare.</hypothesis>
      <hypothesis id_hypothesis="8" entailment="Unknown">Beligan este președintele României.</hypothesis>
      <hypothesis id_hypothesis="9" entailment="Unknown">Iordache are 50 de milioane de lei.</hypothesis>
      <hypothesis id_hypothesis="10" entailment="Unknown">Radu Beligan a fost cel mai mare actor al unei
      generații de talente uimitoare.</hypothesis>
    </entailment-group>
    …
</entailment-corpus>
```

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

The corpus is simple text and it is not parsed with any tool.

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

The contents of the texts and of the hypotheses are in Romanian.

4.4.4 Attributes and their values (if annotated)

The *entailment-group* tag corresponds to a group of text with ten hypotheses and it has one attribute:

- *id_group* which identifies unique the current entailment group.

Inside of *entailment-group* tag we have *text* tag and *hypotheses* tags. The *text* tag corresponds to the text string. Every *hypothesis* tag has the hypothesis string and has two attributes:

- *id_hypothesis* which identify unique the current hypothesis.
- *entailment* represents the relation between current text and the current hypothesis (can be "YES" – entailment, "NO" – contradiction and "UNKNOWN" – unknown).

4.29 Intended application of the corpus

The corpus can be used for testing a textual entailment system for Romanian.

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

The corpus was manual annotated by three native Romanian speakers and it was used with success in order to test our Romanian Textual Entailment system.

5 RELEVANT REFERENCES AND OTHER INFORMATION

Iftene, A. 2009. Textual Entailment (Ph.D. Thesis) Technical Report. "Al. I. Cuza" University. ISSN 1224-9327. 169 pages. October, 2009. Iasi, Romania.

Iftene, A. 2008. Textual Entailment on Romanian. In CD Proceedings of the International Workshop "Tools for Computer-Aided Translation" . ISBN 978-9-291220-37-3. Romanian Academy, February 28-29, Bucharest, Romania.

Iftene, A., Balahur-Dobrescu, A. 2008. A Language Independent Approach for Recognizing Textual Entailment. In journal "Research in Computing Science". Vol. 334, Pp. 3-14. Instituto Politecnico Nacional, Centro de Investigacion en Computacion, Mexico 2007. ISSN: 1870-4069. Poster at 9th International Conference on Intelligent Text Processing and Computational Linguistics (CICling 2008). 17-23 February. Haifa, Israel.

Iftene, A., Balahur-Dobrescu, A. 2007. Textual Entailment on Romanian. The third Workshop on Romanian Linguistic Resources and Tools for Romanian Language Processing. ISSN 1843-911X. Pp. 109-118, 14-15 December. Iaşi, România.

TE-RULES

1 BASIC INFORMATION

1.1 Corpus composition

The corpus consists of 20 entailment rules.

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

The corpus is represented in XML format.

1.3 Character encoding

The characters are UTF8 encoded.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Alex Moruz,
Address: General Berthelot, 16, Iași 700483, Romania
Affiliation: "Al. I. Cuza" University of Iasi, Faculty of Computer Science
Position: Lecturer
Telephone: + 40 232 201549
Fax: + 40 232 201490
e-mail: mmoruz@infoiasi.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.

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

The archive that will be uploaded on the MetaShare platform will contain one file with .xml extension.

3.2 Data structure of an entry

The XML file contains a set of *rule* tags, each representing an entailment rule. Each rule has a *type* attribute, which refers to the external resource the rule uses for tests.
The *T* and *H* elements contain the specific context for rule application, and contain elements of type *node*, which represent information nodes in the text and hypothesis. Node elements have attributes of type *tag* (part of speech), *relation* (syntactic relation) and *parent* (the syntactic parent of the node).
The *test* element refers to the test that needs to be performed (which varies depending on the external resource and defined externally) and the *value* element contains the result of the rule application. The *type* attribute of the *value* element is used for choosing between numerical results for rule applications or final results for rule application.

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

The corpus contains 20 questions and needs about 10 kB for disk storage.

4 CONTENT INFORMATION

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

This corpus contains language independent (but resource dependent) entailment rules.

4.2 The natural language(s) of the corpus

The corpus has no natural language.

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

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 corpus is annotated at the rule level. The following example shows the detailed structure with all tags and attributes used in the annotation.

```
<rules>
  <rule id="3" type="DIRT">
        <T>
              <node id="1" tag="V" lemma="var1"/>
              <node id="2" parent="1" relation="rel1"/>
              <node id="3" parent="1" relation="rel2"/>
        </T>
        <H>
```

```
                <node id="1" tag="V" lemma="var2"/>
                <node id="2" parent="1" relation="rel1"/>
        </H>
        <value type="decimal">DIRT.similarity(var1,var2)</value>
  </rule>
  <rule id="4" type="VerbOcean">
        <T>
                <node id="1" tag="V" lemma="var1"/>
        </T>
        <H>
                <node id="1" tag="V" lemma="var2"/>
        </H>
        <test>antonym(var1,var2)</test>
        <value type="solution">CONTRADICTION</value>
  </rule>
</rules>
```

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

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

4.4.4 Attributes and their values (if annotated)

The *rule* tag corresponds to a rule and it has four elements:

- The *type* attribute refers to the external resource the rule uses for tests.
- *T* contains the text content.
  - *node* elements represent information nodes in the text and hypothesis
    - *tag* attribute (part of speech)
    - *relation* attribute (syntactic relation)
    - *parent* attribute (the syntactic parent of the node)
- *H* contains the hypothesis content.
  - Same as for the T element
- *test* contains the resource specific test necessary for determining the value of the rule application.
- *value* specifies the result of the rule application
  - The *type* attribute is used for choosing between numerical results for rule applications or final results for rule application

4.31 Intended application of the corpus

The corpus can be integrated in rule-based TE systems for improving entailment detection.

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

The corpus was manually created on the basis of the rules used in the UAIC TE system.

5 RELEVANT REFERENCES AND OTHER INFORMATION

Iftene, A., Balahur-Dobrescu, A. 2007. Hypothesis Transformation and Semantic Variability Rules Used in Recognising Textual Entailment. In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. Pages 125-130. 28-29 June, Prague, Czech Republic.

Iftene, A., Balahur-Dobrescu, A. 2008. Named Entity Relation Mining Using Wikipedia. In Pro-ceedings of the Sixth International Language Resources and Evaluation (LREC'08). pp: 763-766. 28-30 May, Marrakech, Morocco.

Iftene, A. 2008. UAIC Participation at RTE4. In Text Analysis Conference (TAC 2008) Workshop - RTE-4 Track. National Institute of Standards and Technology (NIST). November 17-19, 2008. Gaithersburg, Maryland, USA.

Iftene, A. 2009. Textual Entailment, PhD Thesis, "Al. I. Cuza" University, Iasi

Iftene, A., Moruz, M.-A. 2010. UAIC Participation at RTE5, Proceedings of the Second Text Analysis Conference (TAC 2009) November 16-17, 2009 National Institute of Standards and Technology Gaithersburg, Maryland, USA

Iftene, A., Moruz, M.-A. 2011. UAIC Participation at RTE6, The Sixth PASCAL Recognizing Textual Entailment Challenge, Proceedings of the Third Text Analysis Conference (TAC 2010) November 15-16, 2010 National Institute of Standards and Technology Gaithersburg, Maryland, USA

Moruz, M. A. 2011. Predication Driven Textual Entailment, PhD Thesis, "Al. I. Cuza" University, Iasi

Partner RACAI

RO-Wordnet

1. BASIC INFORMATION

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

Ro-WordNet (RWN) is a lexical ontology following the Princeton WordNet (PWN) organizational principles. The synsets in RWN are aligned with PWN3.0 and, additionally, they are associated with SUMO/MILO concepts and labeled with DOMAINS3.0 categories.

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

RWN is distributed as an XML file, observing the structure of BalkaNet wordnets. The file can be loaded and browsed in VisDic (as well as in its descendant versions), the official editor and browser of the BalkaNet project.

1.6 Character encoding

The characters have been encoded in UTF8.

2. ADMINISTRATIVE INFORMATION

2.1 Contact person

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 (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 license-based for research purposes and fee license-based for commercial purposes.

3. TECHNICAL INFORMATION

3.1 Directories and files

WNROM – the directory containing the following files:

wnrom.xml – the proper Romanian WordNet file

wnrom.cfg – the VisDic configuration file for RWN

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

The current (validated) version contains 30,006 synsets, with the following distribution:

Noun synsets	Verb synsets	Adj. synsets	Adv. synsets	Total
21158	7163	851	834	30006

The needed disk space is about 14 000 Kb.

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

The language of the lexical ontology is Romanian. Via alignment with PWN, it is virtually a bilingual English-Romanian dictionary.

4. 2 Entry Type

There are four types of entries, all of them having the same structure: entries for nouns, for verbs, for adjectives and for adverbs.

4.3 Attributes and their values

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, R, A.

The value of the <POS> is one of the N, V, R, A (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 Tufiş 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.

4.4 Coverage of the lexicon

The design procedure of the RWN followed the ***conceptual density principle*** (Tufiş 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 lexicon

The lexical ontology has been used in practically all NLP-enhanced applications developed at RACAI: tagging, lemmatization, word-sense disambiguation, word alignment, collocation extraction, document classification, question-answering, machine translation.

4.6 POS assignment

The part of speech assignment is the one in the Explanatory Dictionary of Romanian.

4.7 Reliability (automatically/manually constructed)

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 was possible) under the validation of a lexicographer.

5. RELEVANT REFERENCES AND OTHER INFORMATION

References on the Romanian WordNet

1. Dan Tufiş. Radu Ion, Luigi Bozianu, Alexandru Ceauşu, Dan Ştefănescu: RO-Wordnet. In *Proceedings of the 4th Global WordNet Association Conference*, January 22-25, 2008, Szeged, Hungary
2. Dan Tufiş (ed.) *Special Issue on BalkaNet, Romanian Academy*, vol7, no. 2-3, 2004, ISSN 1453-8245
3. Dan Tufiş, D. Cristea, S. Stamou. 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
4. Tufiş, D., Barbu, E., Barbu Mititelu, V., Ion, R., Bozianu, L. (2004b). The Romanian Wordnet. *Romanian Journal on Information Science and Technology*, vol. 7, no. 2-3 (pp. 107-124).
5. Dan Tufiş, Radu Ion, Nancy Ide. Word sense disambiguation as a wordnets validation method in Balkanet. In *Proceedings of the 4th LREC Conference*, Lisbon, 2004, 741-744; 1071-1074
6. Tufiş, D. & Barbu, E. (2004). A Methodology and Associated Tools for Building Interlingual Wordnets. In *Proceedings of LREC2004*, Lisbon, Portugal (pp. 1067-1070).

A larger version (not entirely validated) of Romanian WordNet can be browsed at the web address www.racai.ro/wnbrowser.

References to Princeton WordNet, DOMAINS taxonomy and Sumo/MILO ontology

1. 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).
2. 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).
3. 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).
4. Fellbaum, Ch. (Ed.) (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.
5. Vossen, P. (Ed.) (1998). *A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.

The Princeton WordNet 3.0 can be freely downloaded from the address:
http://wordnet.princeton.edu/wordnet/download/ and can be used on-line at the address:
http://wordnetweb.princeton.edu/perl/webwn

WEB-DEX

1. BASIC INFORMATION

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

WEB-DEX is an explanatory dictionary based on the 1996 edition of the standard explanatory dictionary of Romanian published by the Romanian Academy

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

WEB-DEX is distributed as an XML file.

1.9 Character encoding

The characters have been encoded in UTF8

2. ADMINISTRATIVE INFORMATION

2.1 Contact person

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

The resource will be uploaded on the MetaShare platform as an archive with the structure as described under 3.1.

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

The archive contains an XML file (dex.utf8.xml).

3.2 Data structure of an entry

The structure of an entry in WEB-DEX is exemplified below:

```
<entry id="JACHETĂ">
   <hw>JACHETĂ</hw><stress>JACH`ETĂ</stress>
      <alt> <brack>  <gram>nominativ_feminin_singular_indefinit</gram>
                     <orth>jachetă;</orth></brack>
            <brack>  <gram>nominativ_feminin_plural_indefinit</gram>
                     <orth>jachete</orth></brack></alt>
      <pos>substantiv</pos>
      <gen>feminin</gen>
         <struc>
```

```
            <def>Haină ( tricotată) femeiască încheiată în față, care acoperă partea de sus a corpului și care
                se poartă peste bluză sau peste rochie </def>
             <struc type="Sec">
                <def>Haină bărbătească de ceremonii, croită pe talie, lungă până aproape de genunchi.
                 </def>
             </struc>
          </struc>
   <etym>Din limba  <lang>fr.</lang>jaquette</etym>
</entry>
```

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

54.222 entries, requiring 44 MB of hard disk

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

The language of the lexicon is Romanian.

4.2 Entry Type

Each entry in the printed dictionary is represented as described at item 3.2.

4.3 Attributes and their values

See section 3.2:

4.4 Coverage of the lexicon

The lexical stock covers the basic general language vocabulary of Romanian

4.5 Intended application of the lexicon

General lexicographic studies, NLP applications, IR tasks, etc.

4.6 POS assignment

The part of speech assignment is the one in the printed Explanatory Dictionary of Romanian

4.7 Reliability (automatically/manually constructed)

Manually constructed based on the printed form.

5. RELEVANT REFERENCES AND OTHER INFORMATION

6. Tufiş, D., Rotariu, G.,Barbu A.M. “Data Sampling, Lemma Selection and a Core Explanatory Dictionary of Romanian”. În Proceedings of the 5th International Workshop on Computational Lexicography COMPLEX, Pecs, Ungaria, 1999, pp. 219-228
7. Erjavec, T., Tufiş, D., Varadi T. “Developing TEI-Conformant Lexical Databases for CEE Languages”. In Proceedings of the 4th International Workshop on Computational Lexicography COMPLEX, Pecs, Ungaria, 1999, pp.205-210.

8. Dimitrova L., Erjavec T., Ide N., KaalepH.J., Petkevic V., Dan Tufiş: Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages, COLING, Montreal 1998

RO-TblWordForm

1. BASIC INFORMATION

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

This is a wordform lexicon containing statistical information extracted from the Romanian Balanced Corpus

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

The lexicon is a flat file, one entry per line, fields being tab separated

1.12 Character encoding

The characters are UTF8 encoded

2. ADMINISTRATIVE INFORMATION

2.1 Contact person

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

The resource will be uploaded on the MetaShare platform as zip 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

There is only one file named **tblwordform.ro**

3.2 Data structure of an entry

Each entry is a four fields line, tab separated:
<wordform><tab>lemma<tab><msd><tab><frequency> where:

- <wordform> is the occurrence form in the ROMBAC corpus
- <lemma> is the lemma of the wordform or "=", if the word form is the lemma form
- <msd> is a morpho-syntactic tag compliant with the Multext-East specification
- <frequency> is the of the wordform in the ROMBAC corpus;
  Only word forms that at least 5 occurrences have been retained in the lexicon

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

There are 111462 entries and the lexicon requires 3,1 MB disk space

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

The language of the lexicon is Romanian

4. 2 Entry Type

Each of the 14 grammatical types defined by the Multext-East specifications are represented in the wordform lexicon.

As Multext-East specifications are supposed to cover many languages, some attributes might be irrelevant for a specific language. In the linear encoding of morpho-syntactic information for a wordform, the position of the attribute that is irrelevant is filled in with the special character '-'.

4.3 Attributes and their values

The MSD is a linear attribute value representation with fixed positions for each part of speech. Each position corresponds to a specific attribute and it is filled in by one character code. If the respective attribute is not relevant for the combination of the other attribute-values the position of the attribute that is irrelevant is filled in with the special character '-'.
For instance, a singular (s) masculine (m) common (c) noun (N) definite form (y) and in an oblique case –genitive or dative (o) will be encoded as **Ncmsoy**; the code **Vmip2s** describes a main (m) verb (V) indicative mode (i), present tense (p) second person (2) singular (s).

4.4 Coverage of the lexicon

The lexical entries cover general language as reflected in ROMBAC (Romanian Balanced Corpus)

4.5 Intended application of the lexicon

The lexicon is meant for all types of basic language processing (tokenization, tagging, lemmatization)

4.6 POS assignment

The MSD (extended POS) have been manually assigned by trained linguists

4.7 Reliability (automatically/manually constructed)

Highly reliable

5. RELEVANT REFERENCES AND OTHER INFORMATION

Tomaz Erjavec. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the 4th LREC Conference*, LREC'04, Lisabona, pp. 1535 - 1538,

Dan Tufiş, Radu Ion - Specificaţii pentru clasa de etichete folosite în adnotarea morfo-lexicală a limbii române. Research Report, RACAI, June 2007

Dan Tufiş, Barbu A.M., Pătraşcu V., Rotariu G., Popescu C. 1997."Corpora and Corpus-Based Morpho-Lexical Processing". In Dan Tufiş, P. Andersen (eds.) "Recent Advances in Romanian Language Technology", Editura Academiei, pp. 35-56.

Multilingual News Corpus

1 BASIC INFORMATION

1.1 Corpus composition

4622 documents.

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

The corpus is represented in XCES format.

1.3 Character encoding

The documents are UTF8 encoded.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

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 (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, license-based, for research purposes

3 TECHNICAL INFORMATION

3.1 Directories and files

The corpora contains 5 sets of data grouped in separate folders ("ec.europa.eu", "euronews", "europarl1", "europarl2", "europarl3"). Each folder has 3 subfolders named "en-xces", "ro-xces" and "fr-xces" for english, romanian and french documents (in xces format). The Xml Schema Definitions can be found in the folder "XCES-Schema" located in the Root folder. The file "mling-news-general-metadata.xml" contains some general information about the corpora (license, author, etc.) and "mling-news-text-metadata.xml" contains annotation metadata (languages, number of tokens, annotation mode etc.).

3.2 Data structure of an entry

The documents are plain text UTF8 encoded. They are grouped together by their language. The en-xces folder contains documents in English, fr-xces contains the French documents and ro-xces contains the Romanian documents. The filenames for comparable entries start with the same unique identifier (either a numeric value or a randomly generated GUID) and end with the character '_' and their language code (e.g. 1_EN.xml). Examples:

euronews\en-xces\1_EN.xml euronews\ro-xces\1_RO.xml euronews\fr-xces\1_FR.xml
europarl1\en-xces\1_EN.xml europarl1\ro-xces\1_RO.xml europarl1\fr-xces\1_FR.xml

The unique identifier is relative to each set (europarl1, europarl2, euronews etc.) meaning that "euronews\en-xces\1_EN.xml" is not the same document as "europarl1\en-xces\1_EN.xml".

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

- ec.europa.eu (set 1 of files): 137 documents for each language (total 411 documents)
- Euronews (set 2 of files): 506 documents for each language (total 1518 documents)
- europarl1 (set 3 of files): 492 documents for each language (total 1476 documents)
- europarl2 (set 4 of files): 500 documents for each language (total 1500 documents)
- europarl3 (set 5 of files): 212 documents for each language (total 636 documents)

The number of tokens (words) is 1334942 for English, 659031 for Romanian and 1480103 for French.

The size on disk is 277 MB.

4 CONTENT INFORMATION

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

This is a multilingual comparable corpus.

4.2 The natural language(s) of the corpus

The languages for the corpus are: romanian, english, and french

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

The text registers represented into the corpus are: journalistic language as used in the daily newspapers and official language as used in legal documents.

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)

See XCES documentation for details.

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

The corpus is POS tagged.

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

The corpus is aligned at document level.make

4.4.4 Attributes and their values (if annotated)

See XCES documentation for details.

4.33 Intended application of the corpus

Multilingual applications (MT, CLIR)

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

The annotations are automatically generated.

5 RELEVANT REFERENCES AND OTHER INFORMATION

1. http://www.xces.org/
2. http://www.statmt.org/europarl/
3. Ide, N., Romary, L. (2007). Towards International Standards for Language Resources. In Dybkjaer, L., Hemsen, H., Minker, W. (Eds.), *Evaluation of Text and Speech Systems*, Springer, 263-84.
4. Ide, N., Baker, C., Fellbaum, C., Fillmore, C., Passonneau, R. (2008). MASC: The Manually Annotated Sub-Corpus of American English. *Proceedings of the Sixth Language Resources and Evaluation Conference (LREC)*, Marrakech, Morocco.
5. R.Ion, D.Tufiş, T. Boroş, A. Ceauşu, D. Ştefănescu: On-line Compilation of Comparable Corpora and their Evaluation. In Proceedings of the 7th Conference on Formal Approaches to South Slavic and Balkan Languages (FASSBL 2010), Dubrovnic, October, 2010, pp 29-34.

6. Dan Tufiş, Radu Ion, Alexandru Ceauşu, and Dan Ştefănescu. RACAI's Linguistic Web Services. In *Proceedings of the 6th Language Resources and Evaluation Conference - LREC 2008*, Marrakech, Morocco, May 2008. ELRA - European Language Resources Association.

RO-JRC-ACQUIS

1 BASIC INFORMATION

1.1 Corpus composition

The corpus consists of the Romanian version of the Acquis Communautaire, the common set of laws of the European Union member states. There are 10704 documents in which 34234437 tokens occur. Out of these, 27968652 are words and the rest, punctuation.

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

The corpus is represented in XML Corpus Encoding Standard (XCES) format which is compliant with the XCES Schema revision 0.4 (2003)

1.3 Character encoding

The characters are UTF-8 encoded in the Latin 2 character set. A special mention is to the Romanian diacritics “ş” and “ţ” with their upper case variants ”Ş” and ”Ţ” which are not the (incorrect) ones from the Latin 2 character set (“ş” and “ţ” and “Ş” and “Ţ” respectively).

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: +4021 3188103
Fax: +40 21 3188142
e-mail: tufis@racai.ro

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

The resource will be uploaded on the RACAI's 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

The archive that will be uploaded on the MetaShare platform will contain 47 different folders out of which 46 will contain XCES XML files of the respective laws grouped by year in the interval 1958-2006 except 1959-1961. One folder called 'XCES-Schema' contains the XCES schemas against which the validation of the XML files is ensured.

3.2 Data structure of an entry

An entry is a XCES encoded XML file.

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

The corpus contains 34234437 tokens including punctuation and 27968652 words. Out of the archive it needs about 2.8 GB for disk storage on a Windows 7 computer with the NTFS file system in place.

4 CONTENT INFORMATION

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

This corpus is a monolingual, POS tagged, lemmatized, chunked (shallow parsed) corpus and word sense disambiguated (for selected words – words from the domain)

4.2 The natural language(s) of the corpus

The language of the corpus is standard Romanian, orthography being compliant with the current Romanian Academy norms. The diacritical signs are in place (Tufiș and Ceaușu, 2008).

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

The text register represented into the corpus is the official language as used in legal documents.

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 corpus is annotated at paragraph, sentence, constituent group and word levels, providing morpho-lexical and syntactic information. The following example shows the detailed structure with all tags and attributes used in the annotation. For more details about the XCES format, see www.xces.org.

```
- <xces:cesDoc version="0.1">
  - <xces:text complete="y" id="RO_JRC_Acquis">
    - <xces:body>
      - <xces:p id="jrc32006D0313_ro_1">
        - <xces:s id="jrc32006D0313_ro_1_1">
            <xces:tok type="word" msd="Ncfsry;Np#1;ili:ENG20-05500743-n,ENG20-04628484-n" base="decizie">Decizia</xces:tok>
            <xces:tok type="word" msd="Ncmsoy;Np#1;ili:ENG20-07808337-n,ENG20-07809840-n" base="consiliu">Consiliului</xces:tok>
          </xces:s>
        </xces:p>
      - <xces:p id="jrc32006D0313_ro_2">
        - <xces:s id="jrc32006D0313_ro_2_1">
            <xces:tok type="word" msd="Spsa;Pp#1" base="din">din</xces:tok>
            <xces:tok type="word" msd="Mc;Pp#1,Np#1" base="10">10</xces:tok>
            <xces:tok type="word" msd="Ncms-n;Pp#1,Np#1" base="aprilie">aprilie</xces:tok>
            <xces:tok type="word" msd="Mc;Pp#1,Np#1" base="2006">2006</xces:tok>
          </xces:s>
        </xces:p>
      - <xces:p id="jrc32006D0313_ro_3">
        - <xces:s id="jrc32006D0313_ro_3_1">
            <xces:tok type="word" msd="Vmg;Vp#1" base="privi">privind</xces:tok>
            <xces:tok type="word" msd="Ncfsry;Np#1;ili:ENG20-06001364-n,ENG20-00199187-n" base="încheiere">încheierea</xces:tok>
            <xces:tok type="word" msd="Timso;Np#1" base="un">unui</xces:tok>
            <xces:tok type="word" msd="Ncms-n;Np#1" base="acord_de_cooperare">acord_de_cooperare</xces:tok>
            <xces:tok type="word" msd="Crssp;Np#1" base="și">și</xces:tok>
            <xces:tok type="word" msd="Ncfsrn;Np#1" base="asistență">asistență</xces:tok>
            <xces:tok type="word" msd="Spsa;Pp#1" base="între">între</xces:tok>
            <xces:tok type="word" msd="Ncfsry;Pp#1,Np#2" base="Curtea_Penală_Internațională">Curtea_Penală_Internațională</xces:tok>
            <xces:tok type="word" msd="Crssp;Pp#1,Np#2" base="și">și</xces:tok>
            <xces:tok type="word" msd="Np;Pp#1,Np#2" base="Uniunea_Europeană">Uniunea_Europeană</xces:tok>
          </xces:s>
        </xces:p>
```

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

The corpus contains morpho-syntactic information (MSD) which has been assigned automatically with our high accuracy TTL tagger (Ion, 2007; Tufis et al., 2008) which implements the tiered tagging methodology (Tufiș, 1999; Tufiș & Dragomirescu, 2006). About 20% of the MSD have been manually checked, validated and, where the case, corrected (Tufiș and Irimia, 2006).

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)

The *xces:p*, *xces:s* and *xces:tok* tags identify the level of the text under the tag: paragraph, sentence and token. *id* specifies the position of the textual unit in corpus:

- 'jrc32006D0313_ro_1' for the paragraph level
- 'jrc32006D0313_ro_1_1': the first part (jrc32006D0313) is the document identifier in the JRC Acquis corpus (the CELEX code). Then the language code follows (,ro'), the id of the paragraph (the first integer) and the id of the sentence (the second integer);

Under each <xces:tok> tag can be found three attributes and a word form:

- base, whose value is the dictionary form of the word form;

- msd: which combine the MSD code associated to the word form, the chunk information, separated by semicolon; (ex: msd="Np;Np#1") and list of Princeton WordNet synset identifiers which are the most likely senses of that word; the WSD procedure is described in Ion (2010b).

- type: which values can be either "word" or "punctuation"; (ex: type="word">*mult*</xces:tok> or type="punctuation">.</xces:tok>)

The MSDs follows the Multext-East specifications (Erjavec, 2004). For Romanian there are 614 different MSDs (Tufis et al. 1997). They have been slightly modified (new tags for named entities have been added) are largely described in (Tufis and Ion, 2006).

5.5 Intended application of the corpus

Due to the mark-up accuracy, the corpus can be used for building robust statistical language models. It can also be used as a reference corpus for Romanian in various corpus specific types of investigation: quantitative analysis, collocation extraction, grammar induction, etc.

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

The annotations are highly reliable. The paragraph and sentence mark-up has been fully validated. The MSD tagging accuracy is at least 98%. The chunking annotation has been achieved based on a regular grammar defined over the MSD tags. The reliability of chunking mark-up is therefore similar to the tagging accuracy (cca. 98%). The WSD annotation is around 80% accurate given the fact that the most 2 labels have been assigned (to selected words).

6 RELEVANT REFERENCES AND OTHER INFORMATION
http://www.xces.org/

Alexandru Ceauşu. 2008. Colectarea şi procesarea documentelor româneşti ale corpusului JRC-Acquis. In Diana Maria Trandabăţ, Dan Cristea, Dan Tufiş (eds.), *Lucrările atelierului Resurse Lingvistice şi Instrumente pentru Prelucrarea Limbii Române*, Editura Universităţii „Al. I Cuza", Iaşi.

Tomaz Erjavec. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the 4th LREC Conference*, LREC'04, Lisabona, pp. 1535 - 1538,

Radu Ion, Dan Ştefănescu, Alexandru Ceauşu, Dan Tufiş, Elena Irimia and Verginica Barbu Mititelu. 2010. *A Trainable Multi-factored QA System*. In Carol Peters, Giorgio Di Nunzio, Mikko Kurimo, Thomas Mandl, Djamel Mostefa, Anselmo Peñas, Giovanna Roda (eds.) *Multilingual Information Access Evaluation*, Vol. I Text Retrieval Experiments, pp. 257—264, Lecture Notes in Computer Science, Volume 6241/2010, Springer-Verlag.

Radu Ion, and Dan Ştefănescu. 2010b. *RACAI: Unsupervised WSD Experiments @ SemEval-2, Task 17*. In Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval-2, pp. 411—416, Uppsala, Sweden, July 2010. (C) Association for Computational Linguistics. ISBN: 978-1-932432-70-1.

Radu Ion. 2007. Word Sense Disambiguation methods applied to English and Romanian. Ph.D. thesis, Research Institute for Artificial Intelligence (RACAI), Romanian Academy, 153 pages.

Dan Tufiş and Alexandru Ceauşu. 2008. DIAC+: A Professional Diacritics Recovering System. In *Proceedings of the 6th LREC Conference*, Marrakech.

Tufiş, D. 1999."Tiered Tagging and Combined Classifiers". In F. Jelinek, E. Nöth (eds) Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence 1692, Springer, 1999, pp. 28-33

Dan Tufiş, Liviu Dragomirescu. 2004. Tiered Tagging Revisited. In *Proceedings of the 4th LREC'04 Conference*, Lisabona, pp. 39-42

Dan Tufiş, Barbu A.M., Pătraşcu V., Rotariu G., Popescu C. 1997."Corpora and Corpus-Based Morpho-Lexical Processing". In Dan Tufiş, P. Andersen (eds.) "Recent Advances in Romanian Language Technology", Editura Academiei, pp. 35-56.

Dan Tufiş, Radu Ion. 2007. Specificaţii pentru clasa de etichete folosite în adnotarea morfo-lexicală a limbii române. Raport de cercetare, iunie, Institutul de Cercetări pentru inteligenţă artificială, 24 pages.

Dan Tufiş, Elena Irimia. 2006. RoCo_News - A Hand Validated Journalistic Corpus of Romanian. In *Proceedings of the 5th LREC Conference*, Genoa, pp. 869-872

Dan Tufiş, Radu Ion, Alexandru Ceauşu, and Dan Ştefănescu. 2008. RACAI's Linguistic Web Services. In *Proceedings of the 6th LREC Conference – LREC'08*, Marrakech.

ROMANIAN BALANCED CORPUS

1 BASIC INFORMATION

1.1 Corpus composition

The corpus consists of equal shares of texts from 5 different genres: journalism, legalese, fiction, medicine and biographical data for Romanian literary personalities. For each genre, texts have been selected containing around 7,000,000 words, so that the entire corpus counts around 41,000,000 words, including punctuation.

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

The corpus is represented in XCES format.

1.3 Character encoding

The characters are UTF8 encoded.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

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 (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, 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 five different folders (corresponding to the 5 genres it covers): Literature, Law, Academic, Medicine and Journalism. Each folder will contain different numbers of files with .xml extension.

3.2 Data structure of an entry

This is not relevant as the corpus is provided as a text file. It is structured in paragraphs, containing one or more sentences. Each sentence is segmented into tokens (equivalent to word or named entities) that can be embedded or not in chunk structures.

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

The corpus contains 41.534.961 tokens, including punctuation, and needs about 4 GB for disk storage. There are around 660.000 different word forms.

4 CONTENT INFORMATION

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

This corpus is a balanced monolingual, heavily annotated corpus.

4.2 The natural language(s) of the corpus

The language of the corpus is standard Romanian, orthography being compliant with the current Romanian Academy norms. The diacritical signs are in place (Tufiş &Ceauşu, 2008).

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

The text registers represented into the corpus are: journalistic language as used in the daily newspapers, official language as used in legal documents, artistic language as used in literary fiction, technical (medical) language as used in medical treatise and academic language as used in the literary anthologies.

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 corpus is annotated at paragraph, sentence, constituent group and word levels, providing morpho-lexical and syntactic information. The following example shows the detailed structure with all tags and attributes used in the annotation. For more details about the XCES format, see www.xces.org.

```
<?xml version="1.0" encoding="UTF-8"?>
<xces:cesCorpus xmlns:xces=http://www.xces.org/schema/2003
                xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
          xsi:schemaLocation="http://www.xces.org/schema/2003
          http://www.xces.org/schema/2003/xcesDoc.xsd"
<xces:p id="p1545">
 <xces:s id="Calinescu_George_Bietul_Ioanide_1545">
    <xces:tok base="Conțescu" msd="Np;Np#1" type="word">Conțescu</xces:tok>
   <xces:tok base="întreba" msd="Vmis3s;Vp#1" type="word">întrebă</xces:tok>
   <xces:tok base="pe" msd="Spsa;Pp#1" type="word">pe</xces:tok>
   <xces:tokbase="Ioanide" msd="Np;Pp#1,Np#2" type="word">Ioanide</xces:tok>
   <xces:tok base="dacă" msd="Csssp;Pp#1,Np#2" type="word">dacă</xces:tok>
   <xces:tok base="transport"msd="Ncmsry;Pp#1,Np#2" type="word">transportul</xces:tok>
   <xces:tok base="material" msd="Ncmsoy;Pp#1,Np#2" type="word">materialului</xces:tok>
   <xces:tok base="-" msd="DASH" type="punctuation">-</xces:tok>
   <xces:tok base="cărămidă" msd="Ncfsrn;Np#3" type="word">cărămidă</xces:tok>
   <xces:tok base="," msd="COMMA" type="punctuation">,</xces:tok>
   <xces:tok base="var" msd="Ncms-n;Np#4" type="word">var</xces:tok>
   <xces:tok base="-" msd="DASH" type="punctuation">-</xces:tok>
   <xces:tok base="nu" msd="Qz;Vp#2" type="word">nu</xces:tok>
   <xces:tok base="avea" msd="Va--3;Vp#2" type="word">ar</xces:tok>
   <xces:tok base="costa" msd="Vmnp;Vp#2" type="word">costa</xces:tok>
   <xces:tok base="prea" msd="Rp;Vp#2,Ap#1" type="word">prea</xces:tok>
   <xces:tok base="mult" msd="Rgp;Ap#1" type="word">mult</xces:tok>
   <xces:tok base="." msd="PERIOD" type="punctuation">.</xces:tok>
 </xces:s>
</xces:p>
```

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

The corpus contains morpho-syntactic information (MSD) which has been assigned automatically with our high accuracy TTL tagger (Ion, 2007; Tufis et al., 2008) which implements the tiered tagging methodology (Tufiș, 1999; Tufiș & Dragomirescu, 2006). About 20% of the MSD have been manually checked, validated and, where the case, corrected (Tufiș and Irimia, 2006).

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

Not relevant

4.4.5 Attributes and their values (if annotated)

The *xces:p*, *xces:s* and *xces:tok* tags identify the level of the text under the tag: paragraph, sentence, , token.
id specifies the position of the textual unit in corpus:

- p1545 for the paragraph level
- *Calinescu_George_Bietul_Ioanide*_1545: the first part (italic) is a slightly modified version of the title of the document containing the sentence and the number (1545) is the number of the sentence in the

document (in our interpretation, each paragraph has only one sentence, therefore the sentence number coincides with the paragraph number)
- In the case of the chunk structures, the *id* attribute specifies the type of chunk and its position in the sentence (ex: Np#1).

Under each <xces:tok> tag can be found three attributes and a word form (see italics in the example under 4.4.1):
- base, whose value is the dictionary form of the word form;
- msd: which combine the MSD code associated to the word form and the chunk information, separated by semicolon; (ex: msd="Np;Np#1")
- type: which values can be either "word" or "punctuation"; (ex: type="word">*mult*</xces:tok> or type="punctuation">.</xces:tok>)
The MSDs follows the Multext-East specifications (Erjavec, 2004). For Romanian there are 614 different MSDs (Tufis et al. 1997). They have been slightly modified (new tags for named entities have been added) are largely described in (Tufis and Ion, 2006)

6.5 Intended application of the corpus

Due to the mark-up accuracy, the corpus can be used for building robust statistical language models. It can also be used as a reference corpus for Romanian in various corpus specific types of investigation: quantitative analysis, collocation extraction, grammar induction, etc.

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

The annotations are highly reliable. The paragraph and sentence mark-up has been fully validated. The MSD tagging accuracy is at least 98%. The chunking annotation has been achieved based on a regular grammar defined over the MSD tags. The reliability of chunking mark-up is therefore similar to the tagging accuracy (cca. 98%).

5 RELEVANT REFERENCES AND OTHER INFORMATION

http://www.xces.org/

Tomaz Erjavec. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the 4th LREC Conference*, LREC'04, Lisabona, pp. 1535 - 1538,

Tufiş, D. 1999."Tiered Tagging and Combined Classifiers". In F. Jelinek, E. Nöth (eds) Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence 1692, Springer, 1999, pp. 28-33

Dan Tufiş, Liviu Dragomirescu. 2004. Tiered Tagging Revisited. In *Proceedings of the 4th LREC'04 Conference*, Lisabona, pp. 39-42

Dan Tufiş, Barbu A.M., Pătraşcu V., Rotariu G., Popescu C. 1997."Corpora and Corpus-Based Morpho-Lexical Processing". In Dan Tufiş, P. Andersen (eds.) "Recent Advances in Romanian Language Technology", Editura Academiei, pp. 35-56.

Dan Tufiş, Radu Ion. 2007. Specificaţii pentru clasa de etichete folosite în adnotarea morfo-lexicală a limbii române. Raport de cercetare, iunie, Institutul de Cercetări pentru inteligenţă artificială, 24 pages.

Dan Tufiş, Elena Irimia. 2006. RoCo_News - A Hand Validated Journalistic Corpus of Romanian. In *Proceedings of the 5th LREC Conference*, Genoa, pp. 869-872

Radu Ion. 2007. Word Sense Disambiguation methods applied to English and Romanian. Ph.D. thesis, Research Institute for Artificial Intelligence (RACAI), Romanian Academy, 153 pages.

Dan Tufiş, Radu Ion, Alexandru Ceauşu, and Dan Ştefănescu. 2008. RACAI's Linguistic Web Services. In *Proceedings of the 6th LREC Conference – LREC'08*, Marrakech.

Dan Tufiş and Alexandru Ceauşu. 2008. DIAC+: A Professional Diacritics Recovering System. In *Proceedings of the 6th LREC Conference,* Marrakech.

SemCor CORPUS

1 BASIC INFORMATION

1.1 Corpus composition

SemCor En-Ro corpus (Lupu et al., 2005; Ion, 2007) is an English-Romanian parallel corpus which was developed starting from the English SemCor (Mihalcea and Pedersen, 2003), a sense-tagged corpus created at Princeton University by the WordNet Project[5] research team, which itself was originally a subpart of the Brown balanced corpus (Kučera and Francis, 1967), containing news articles, literature, scientific and religious texts. In spite of its small dimension, SemCor has been extensively used both as training and testing data in various Word-Sense Disambiguation experiments and competitions, as word-sense annotated resources are scarce (Ng, 1997; Stetina et al., 1998; de Loupy, 1998; Mihalcea & Moldovan , 1999; Mihalcea & Moldovan, 2001).
En-Ro SemCor contains a total of 178,499 words for English and 175,603 words for Romanian (Ion, 2007).

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

[5] http://wordnet.princeton.edu/

The corpus is represented in XCES format.

1.3 Character encoding

The characters are UTF8 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: +4021 3188103
Fax: +40 21 3188142
e-mail: tufis@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.

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

The archive that will be uploaded on the MetaShare platform will contain 2 xml files for both English and Romanian content. Every sentence (e.g. *<xces:s id="br_a01_2_2_en">*) in any of the files has a unique identifier which corresponds to the parallel sentence in the other file.

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 containing syntactic and semantic information about its grammatical *meta-category*, *lemma*, *morpho-sintactic descriptor* (*msd)* – *tag*[6], *syntactic constituent membership* (NP – Noun Phrase; VP – Verb Phrase; AP – Adjectival Phrase; PP – Prepositional Phrase), associated Princeton WordNet 3.0 *word-sense* and *syntactic lexical attracted word* as a 0-based position in the current sentence.

[6] http://nl.ijs.si/ME/V4/msd/html/index.html

The meta-categories are hand-made clusters created taking into consideration the empirical evidence of POS translation affinities (see the annexes of this file).

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

The corpus contains 354,102 tokens (including punctuation): 178,499 for English and 175,603 for Romanian.
The Space requires on Disk is about 31Mb.

4 CONTENT INFORMATION

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

This corpus is a balanced parallel, heavily annotated corpus.

4.2 The natural language(s) of the corpus

The languages of the corpus are English and standard Romanian. The diacritics and all special characters are encoded as SGML entities.

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

The text registers represented into the corpus are: news articles, literature, scientific and religious 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)

The corpus is annotated at paragraph, sentence, constituent group and word levels, providing morpho-lexical, syntactic and semantic information. Each token has a descriptor attribute containing syntactic and semantic information about its grammatical *meta-category*, *lemma*, *morpho-sintactic descriptor* (*msd*) – *tag*[7], *syntactic constituent membership* (NP – Noun Phrase; VP – Verb Phrase; AP – Adjectival Phrase; PP – Prepositional Phrase), associated Princeton WordNet 3.0 *word-sense* and *syntactic lexical attracted word* as a 0-based position in the current sentence. The meta-categories are hand-made clusters created taking into consideration the empirical evidence of POS translation affinities (see the annexes of this file).

The following example shows the detailed structure with all tags and attributes used in the annotation. For more details about the XCES format, see www.xces.org.

[7] http://nl.ijs.si/ME/V4/msd/html/index.html

```
<xces:cesDoc version="0.1">
    <xces:text id="SemCor3_0_en_ro" complete="y">
        <xces:body>
            <xces:p id="p1">
                <xces:s id="br_a01_1_1_en">
                    <xces:tok base="the" msd="2+,Dd;Np#1" type="word">The</xces:tok>
                    <xces:tok base="Fulton_County_Grand_Jury" msd="8+,Np;Np#1;ili:ENG30-00031264-n;0"
                            type="word">Fulton_County_Grand_Jury</xces:tok>
                    <xces:tok base="say" msd="1+,Vmis;Vp#1;ili:ENG30-01009240-v;1"
                            type="word">said</xces:tok>
                    <xces:tok base="Friday" msd="1+,Ncns;Np#2;ili:ENG30-15164463-n;2"
                            type="word">Friday</xces:tok>
                    <xces:tok base="a" msd="21+,Ti-s;Np#3;5" type="word">an</xces:tok>
                    <xces:tok base="investigation" msd="1+,Ncns;Np#3;ili:ENG30-05800611-n;3"
                            type="word">investigation</xces:tok>
                    <xces:tok base="of" msd="5+,Sp;Pp#1;7" type="word">of</xces:tok>
                    <xces:tok base="Atlanta" msd="8+,Np;Pp#1,Np#4;ili:ENG30-09076675-n;5"
                            type="word">Atlanta</xces:tok>
                    <xces:tok base="&amp;apos;s" msd="21+,St;Pp#1,Np#4;7" type="word">&amp;apos;s</xces:tok>
                    <xces:tok base="recent" msd="1+,Afp;Pp#1,Np#4,Ap#1;ili:ENG30-01730444-s;10"
                            type="word">recent</xces:tok>
                    <xces:tok base="primary_election" msd="1+,Ncns;Pp#1,Np#4;ili:ENG30-00182571-n;3"
                            type="word">primary_election</xces:tok>
                    <xces:tok base="produce" msd="1+,Vmis;Vp#2;ili:ENG30-02141146-v;10"
                            type="word">produced</xces:tok>
                    <xces:tok base="&amp;quot;" msd="DBLQ" type="punctuation">&amp;quot;</xces:tok>
                    <xces:tok base="no" msd="22+,Dz3;Np#5;14" type="word">no</xces:tok>
                    <xces:tok base="evidence" msd="1+,Ncns;Np#5;ili:ENG30-05823932-n;11"
                            type="word">evidence</xces:tok>
                    <xces:tok base="&amp;quot;" msd="DBLQ" type="punctuation">&amp;quot;</xces:tok>
                    <xces:tok base="that" msd="31+,Cs;19" type="word">that</xces:tok>
                    <xces:tok base="any" msd="22+,Di3;Np#6;18" type="word">any</xces:tok>
                    <xces:tok base="irregularity" msd="1+,Ncnp;Np#6;ili:ENG30-00737188-n;19"
                            type="word">irregularities</xces:tok>
                    <xces:tok base="take_place" msd="1+,Vmis;Vp#3;ili:ENG30-00339934-v;14"
                            type="word">took_place</xces:tok>
                    <xces:tok base="." msd="PERIOD" type="punctuation">.</xces:tok>
        </xces:s>
```

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

The corpus contains ***morpho-syntactic*** information (MSD) which has been assigned automatically with RACAI's high accuracy TTL tagger (Ion, 2007; Tufiş et al., 2008). The ***grammatical meta-categories*** are also marked using TTL as an unsigned integer pointing to a cluster of morpho-syntactic descriptors. The clusters were manually created based on empirical evidence of POS translation affinities (see the annexes of this file).

Another annotation is the ***syntactic constituent membership*** (shallow parsing info) also added by the TTL.

PWN 3.0 ***WSD tags*** are also present. They were manually assigned to each token representing a content word.

The last tag represents the *syntactic lexical attracted word* as a 0-based position in the current sentence. They were automatically annotated using LexPar (Ion and Barbu-Mititelu, 2006), an application using Lexical Attraction Models.

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

The alignment is encoded in the sentence ids. Sentences having the same id are reciprocal translation. The translation was performed manually by the NLP group at FII-UAIC[8]. Only 81 out of 352 original Englsih SemCor files have been translated:

- br-a01, br-a02, br-a11 to br-a15 (including),
- br-b13, br-b20,
- br-c01, br-c02, br-c04,
- br-d01 to br-d04 (including),
- br-e01, br-e02, br-e04, br-e21, br-e24, br-e29,
- br-f03, br-f10, br-f19, br-f43,
- br-g01, br-g11, br-g15,
- br-h01,
- br-j01 to br-j20 (including), br-j23, br-j37, br-j52 to br-j60 (including), br-j70,
- br-k01 to br-k19 (including).

4.4.6 Attributes and their values (if annotated)

The *xces:p*, *xces:s* and *xces:tok* tags identify the level of the text under the tag: paragraph, sentence, token.

- text *id* (i.e. *SemCor3_0_en_ro*): specifies the name of the corpus and the languages it contains;
- paragraph *id* (e.g. p1545): specifies the position of the paragraph unit in corpus;
- sentence id (e.g. *br_a01_49_58_en*): represents the original English SemCor file the sentence belongs to;
- In the case of the chunk structures, the *id* attribute specifies the type of chunk and its position in the sentence (e.g. *Np#1* – Noun Phrase no. 1).

Under each *<xces:tok>* tag can be found three attributes and a word form (see italics in the example under 4.4.1):

- base, whose value is the dictionary form of the word form;

- msd: which contains: the grammatical meta Category (see the anexes), the *MSD* tag associated to the wordform, the *chunk* information, the associate PWN 3.0 *word-sense* and the *syntactic lexical attracted word* as a 0-based position in the current sentence, all separated by semicolon; (e.g. *msd="1+,Vmis;Vp#1;ili:ENG30-01009240-v;1"*)

[8] Faculty of Informatics – Al. I. Cuza University, Iaşi

- *type:* whose value can be either "word" or "punctuation"; (e.g. type="word">*mult*</xces:tok> or type="punctuation">.</xces:tok>)
The MSDs follows the Multext-East specifications (Erjavec, 2004)[9]. For Romanian there are 614 different MSDs (Tufiş et al. 1997). They have been slightly modified (new tags for named entities have been added) are largely described in (Tufiş and Ion, 2006)

6.7 Intended application of the corpus

The primary purpose of the corpus is to be used as training or testing data for WSD tools. Because it is a small corpus, it CANNOT be used as a reference corpus for English or Romanian. However, due to the fact that it is a balanced corpus, it can be used for building small language models for word prefixes or suffixes.

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

The annotations are highly reliable. The MSD tagging accuracy is at least 98%. The chunking annotation has been achieved based on a regular grammar defined over the MSD tags. The word-sense labels have been manually assigned and the syntactic lexical attracted word was annotated using a state of the art theoretical model.

5 RELEVANT REFERENCES AND OTHER INFORMATION

1. de Loupy, C. and El-Beze, M. and Marteau, P-F. (1998). *Word Sense Disambiguation using HMM Tagger*. In Proceedings of the First International Conference on Language Resources and Evaluation, pages 1255–1258, Grenade, Spain;
2. Erjavec, T. (2004). *MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora*. In *Proceedings of the 4th LREC Conference*, LREC'04, Lisabona, pp. 1535 – 1538;
3. Ion, R. (2007). *Word Sense Disambiguation methods applied to English and Romanian*. Ph.D. thesis, Research Institute for Artificial Intelligence (RACAI), Romanian Academy, 153 pages;
4. Ion, R. and Barbu-Mititelu, V. (2006). *Constrained Lexical Attraction Models*. In Proceedings of the Nineteenth International Florida Artificial Intelligence Research Society Conference, pages 297–302, Menlo Park, Calif., USA. AAAI Press;
5. Kučera, H. and Francis, N.W. (1967). *Computational analysis of present-day American English*. Brown University Press, Providence, Rhode Island;
6. Lupu, M., Trandabăţ, D. and Husarciuc, M. (2005). *A Romanian SemCor Aligned to the English and Italian MultiSemCor*. In Proceedings of the Romance FrameNet Workshop and Kick-off Meeting, EuroLAN 2005, pages 20–27, Babes-Bolyai University, Cluj-Napoca, Romania;

[9] http://nl.ijs.si/ME/V4/msd/html/index.html

7. Mihalcea, R. and Moldovan, D. (1999). *A method for word sense disambiguation of unrestricted text*. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999), College Park, MA;
8. Mihalcea, R. and Moldovan, D. (2001). *A highly accurate bootstrapping algorithm for word sense disambiguation*. International Journal on Artificial Intelligence Tools, 10(1–2);
9. Mihalcea, R. and Pedersen, T. (2003). *An Evaluation Exercise for Word Alignment*. In Proceedings of the HLT-NAACL 2003 Workshop: Building and Using Parallel Texts Data Driven Machine Translation and Beyond, pages 1–10, Edmonton, Canada;
10. Ng, H.T. (1997). *Getting serious about word sense disambiguation*. In Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?, pages 1–7, Washington, D.C., USA;
11. Stetina, J., Kurohashi, S. and Nagao, M. (1998). *General word sense disambiguation method based on a full sentential context*. In Proceedings of the Coling-ACL'98 Workshop "Usage of WordNet in Natural Language Processing Systems", pages 1–8, Montreal;
12. Tufiș, D. and Ion, R. (2007). *Specificații pentru clasa de etichete folosite în adnotarea morfo-lexicală a limbii române*. Raport de cercetare, iunie, Institutul de Cercetări pentru inteligență artificială, 24 pages;
13. Tufiş, D., Barbu, A.M., Pătraşcu, V., Rotariu, G. and Popescu, C. (1997). *Corpora and Corpus-Based Morpho-Lexical Processing*. In Dan Tufiş, P. Andersen (eds.) "Recent Advances in Romanian Language Technology", Editura Academiei, pp. 35-56;
14. Tufiş, D., Ion, R., Ceaușu, A. and Ştefănescu, D. (2008). *RACAI's Linguistic Web Services*. In *Proceedings of the 6th LREC Conference – LREC'08*, Marrakech.

Annexes

English meta-categories clusters:

0	Pt3	1	Ncfs--y	1	Vmis-p
1	Af	1	Ncm	1	Vmis1s
1	Afc	1	Ncm---y	1	Vmis2s
1	Afp	1	Ncmp	1	Vmis3s
1	Afs	1	Ncmp--y	1	Vmn
1	M	1	Ncms	1	Vmnp
1	Mc	1	Ncms--y	1	Vmpp
1	Mc-p-d	1	Ncn	1	Vmps
1	Mo	1	Ncn---y	1	Vo
1	Mo-s-r	1	Ncnp	1	Voip
1	Nc	1	Ncnp--y	1	Voip3s
1	Nc----y	1	Ncns	1	Vois
1	Nc-p	1	Ncns--y	1	Von
1	Nc-p--y	1	Vm	1	Vops
1	Nc-s	1	Vmcs	2	Dd
1	Nc-s--y	1	Vmip	2	Dd--p
1	Ncf	1	Vmip-p	2	Dd--s
1	Ncf---y	1	Vmip1s	2	Dd3
1	Ncfp	1	Vmip2s	2	Dd3-p
1	Ncfp--y	1	Vmip3s	2	Dd3-s
1	Ncfs	1	Vmis	2	Dd3n

2	Dd3np	8	Npns	13	Pp3msa
2	Dd3ns	8	Y	13	Pp3msn
2	Pd--p	8	Yn	13	Pp3ns
2	Pd--s	10	Ds	14	R
2	Pd3	10	Ds----p	14	R-p---q
2	Pd3-p	10	Ds----s	14	Rm
2	Pd3-s	10	Ds1---p	14	Rmc
2	Pd3n	10	Ds1---s	14	Rmp
2	Pd3np	10	Ds2	14	Rmp---q
2	Pd3ns	10	Ds3---p	14	Rmp---r
3	Va	10	Ds3---sf	14	Rms
3	Vacs	10	Ds3---sm	14	Rsc
3	Vaip	10	Ds3---sn	14	Rsp
3	Vaip-p	10	Ps	14	Rss
3	Vaip1p	10	Ps----p	15	Qn
3	Vaip1s	10	Ps----s	16	I
3	Vaip2s	10	Ps1---p	21	St
3	Vaip3s	10	Ps1---s	21	Ti-s
3	Vais	10	Ps2	22	Di
3	Vais-p	10	Ps3	22	Di--p
3	Vais1s	10	Ps3---p	22	Di--s
3	Vais2s	10	Ps3---sf	22	Di3
3	Vais3s	10	Ps3---sm	22	Di3-p
3	Van	12	Px	22	Di3-s
3	Vapp	12	Px1-p	22	Di3n
3	Vaps	12	Px1-s	22	Di3np
4	Dw	12	Px2-p	22	Dz
4	Dw--------q	12	Px2-s	22	Dz--s
4	Dw--------r	12	Px3-p	22	Dz3
4	Dw3-p	12	Px3-s	22	Dz3-s
4	Dw3-s	12	Px3fs	22	Pi
4	Pw	12	Px3ms	22	Pi3
4	Pw----------q	12	Px3ns	22	Pi3-p
4	Pw----------r	13	Pp	22	Pi3-s
4	Pw---a----------q	13	Pp--pn	22	Pi3fs
4	Pw---a----------r	13	Pp--sn	22	Pi3ms
4	Pw3----------q	13	Pp1	22	Pi3n
4	Pw3----------r	13	Pp1-pa	22	Pi3np
4	Pw3-p	13	Pp1-pn	22	Pi3ns
4	Pw3-s	13	Pp1-sa	22	Pz3
4	Pw3n	13	Pp1-sn	22	Pz3-s
5	S	13	Pp2	22	Pz3ms
5	Sp	13	Pp2-p	22	Pz3n
7	Qz	13	Pp3	22	Pz3ns
8	Np	13	Pp3-pa	31	Cc
8	Np-p	13	Pp3-pn	31	Cc-i
8	Np-s	13	Pp3fs	31	Cc-n
8	Npfs	13	Pp3fsa	31	Cs
8	Npms	13	Pp3fsn	50	DATE
8	Npnp	13	Pp3ms	100	X

Romanian meta-categories clusters:

1	Af	1	Af--pry	1	Af--sry
1	Af----n	1	Af--s-n	1	Af--svn
1	Af--p-n	1	Af--son	1	Af--svy
1	Af--poy	1	Af--soy	1	Afcfp-n

1	Afcfpoy	1	Afsmpoy	1	Mmmpo-yy
1	Afcfpry	1	Afsmpry	1	Mmmpr-n
1	Afcfson	1	Afsms-n	1	Mmmpr-ny
1	Afcfsoy	1	Afsmsoy	1	Mmmpr-y
1	Afcfsrn	1	Afsmsry	1	Mmmpr-yy
1	Afcfsry	1	Afsmsvy	1	Mmmso-y
1	Afcmp-n	1	M	1	Mmmso-yy
1	Afcmpoy	1	Mc	1	Mmmsr-n
1	Afcmpry	1	Mc-p-d	1	Mmmsr-ny
1	Afcms-n	1	Mc-p-l	1	Mmmsr-y
1	Afp	1	Mc-p-r	1	Mmmsr-yy
1	Afp-p-n	1	Mc-s-d	1	Mo
1	Afp-p-ny	1	Mc-s-r	1	Mo---l
1	Afp-poy	1	Mcfp-l	1	Mo---ln
1	Afpf--n	1	Mcfp-ln	1	Mo---lny
1	Afpf--ny	1	Mcfp-rn	1	Mo-s-r
1	Afpfp-n	1	Mcfpoly	1	Mofp-ln
1	Afpfp-ny	1	Mcfprln	1	Mofpoly
1	Afpfpon	1	Mcfprly	1	Mofpolyy
1	Afpfpoy	1	Mcfs-l	1	Mofprly
1	Afpfpoyy	1	Mcfsoln	1	Mofprlyy
1	Afpfpry	1	Mcfsoly	1	Mofs-l
1	Afpfpryy	1	Mcfsrln	1	Mofsoln
1	Afpfson	1	Mcfsrly	1	Mofsoly
1	Afpfsony	1	Mcmp-l	1	Mofsolyy
1	Afpfsoy	1	Mcms-ln	1	Mofsrln
1	Afpfsoyy	1	Mcmsoly	1	Mofsrly
1	Afpfsrn	1	Mcmsrl	1	Mofsrlyy
1	Afpfsrny	1	Mcmsrly	1	Momp-ln
1	Afpfsry	1	Mffpoly	1	Mompoly
1	Afpfsryy	1	Mffprln	1	Mompolyy
1	Afpfsvn	1	Mffprly	1	Momprly
1	Afpfsvy	1	Mffsoln	1	Momprlyy
1	Afpm--n	1	Mffsoly	1	Moms-l
1	Afpmp-n	1	Mffsrln	1	Moms-ln
1	Afpmp-ny	1	Mffsrly	1	Momsoly
1	Afpmpoy	1	Ml-po	1	Momsolyy
1	Afpmpoyy	1	Ml-pr	1	Momsrly
1	Afpmpry	1	Mlfpo	1	Momsrlyy
1	Afpmpryy	1	Mlfpr	1	Nc
1	Afpms-n	1	Mlmpo	1	Nc---n
1	Afpms-ny	1	Mlmpr	1	Nc-p-n
1	Afpmsoy	1	Mmfp--n	1	Nc-poy
1	Afpmsoyy	1	Mmfp--ny	1	Nc-pry
1	Afpmsry	1	Mmfpo-y	1	Nc-pvy
1	Afpmsryy	1	Mmfpo-yy	1	Nc-s
1	Afpmsvn	1	Mmfpr-y	1	Nc-s-ny
1	Afpmsvy	1	Mmfpr-yy	1	Nc-son
1	Afs	1	Mmfso-n	1	Nc-soy
1	Afsfp-n	1	Mmfso-ny	1	Nc-sry
1	Afsfpoy	1	Mmfso-y	1	Ncf--n
1	Afsfpry	1	Mmfso-yy	1	Ncf--ny
1	Afsfson	1	Mmfsr-n	1	Ncfp-n
1	Afsfsoy	1	Mmfsr-ny	1	Ncfp-ny
1	Afsfsrn	1	Mmfsr-y	1	Ncfpoy
1	Afsfsry	1	Mmfsr-yy	1	Ncfpoyy
1	Afsmp-n	1	Mmmpo-y	1	Ncfpry

1	Ncfpryy	1	Vmil2p----y	2	Dd3-po---e
1	Ncfpvy	1	Vmil2s	2	Dd3-po---o
1	Ncfs-n	1	Vmil2s----y	2	Dd3-pr
1	Ncfson	1	Vmil3p	2	Dd3-so
1	Ncfsony	1	Vmil3p----y	2	Dd3-sr
1	Ncfsoy	1	Vmil3s	2	Dd3fpo
1	Ncfsoyy	1	Vmil3s----y	2	Dd3fpr
1	Ncfsrn	1	Vmip1p	2	Dd3fpr---e
1	Ncfsrny	1	Vmip1p----y	2	Dd3fpr---o
1	Ncfsry	1	Vmip1s	2	Dd3fpr--y
1	Ncfsryy	1	Vmip1s----y	2	Dd3fso
1	Ncfsvn	1	Vmip2p	2	Dd3fso---e
1	Ncfsvy	1	Vmip2p----y	2	Dd3fso---o
1	Ncm	1	Vmip2s	2	Dd3fsr
1	Ncm--n	1	Vmip2s----y	2	Dd3fsr---e
1	Ncmp-n	1	Vmip3	2	Dd3fsr---o
1	Ncmp-ny	1	Vmip3-----y	2	Dd3fsr--ye
1	Ncmpoy	1	Vmip3p	2	Dd3fsr--yo
1	Ncmpoyy	1	Vmip3p----y	2	Dd3mpo
1	Ncmpry	1	Vmip3s	2	Dd3mpr
1	Ncmpryy	1	Vmip3s----y	2	Dd3mpr---e
1	Ncmpvy	1	Vmis1p	2	Dd3mpr---o
1	Ncms-n	1	Vmis1p----y	2	Dd3mpr--y
1	Ncms-ny	1	Vmis1s	2	Dd3mpr--yo
1	Ncms-y	1	Vmis1s----y	2	Dd3mso
1	Ncmsoy	1	Vmis2p	2	Dd3mso---e
1	Ncmsoyy	1	Vmis2p----y	2	Dd3mso---o
1	Ncmsrn	1	Vmis2s	2	Dd3msr
1	Ncmsrny	1	Vmis2s----y	2	Dd3msr---e
1	Ncmsry	1	Vmis3p	2	Dd3msr---o
1	Ncmsryy	1	Vmis3p----y	2	Dd3msr--y
1	Ncmsvn	1	Vmis3s	2	Dd3msr--yo
1	Ncmsvny	1	Vmis3s----y	2	Pd3-po
1	Ncmsvy	1	Vmm-2p	2	Pd3-pr
1	Vm--1	1	Vmm-2p----y	2	Pd3-so
1	Vm--2	1	Vmm-2s	2	Pd3-sr
1	Vm--3	1	Vmm-2s----y	2	Pd3fpo
1	Vmg	1	Vmnp	2	Pd3fpr
1	Vmg-------y	1	Vmnp------y	2	Pd3fpr--y
1	Vmii1	1	Vmp	2	Pd3fso
1	Vmii1-----y	1	Vmp--pf	2	Pd3fsr
1	Vmii1p	1	Vmp--pf---y	2	Pd3fsr--y
1	Vmii1s	1	Vmp--pm	2	Pd3mpo
1	Vmii2p	1	Vmp--pm---y	2	Pd3mpr
1	Vmii2p----y	1	Vmp--sf	2	Pd3mpr--y
1	Vmii2s	1	Vmp--sf---y	2	Pd3mso
1	Vmii2s----y	1	Vmp--sm	2	Pd3msr
1	Vmii3p	1	Vmp--sm---y	2	Pd3msr--y
1	Vmii3p----y	1	Vmsp1p	3	Qf
1	Vmii3s	1	Vmsp1s	3	Va
1	Vmii3s----y	1	Vmsp2p	3	Va--1
1	Vmil1	1	Vmsp2s	3	Va--1-----y
1	Vmil1p	1	Vmsp3	3	Va--1p
1	Vmil1p----y	1	Vmsp3-----y	3	Va--1s
1	Vmil1s	1	Vmsp3s	3	Va--1s----y
1	Vmil1s----y	1	Vmsp3s----y	3	Va--2p
1	Vmil2p	2	Dd3-po	3	Va--2p----y

3	Va--2s	4	Pw3--o	8	Ypfso
3	Va--2s----y	4	Pw3--r	8	Ypfsr
3	Va--3	4	Pw3-po	8	Ypmpr
3	Va--3-----y	4	Pw3-so	8	Ypms
3	Va--3p	4	Pw3fpr	8	Ypmso
3	Va--3p----y	4	Pw3fso	8	Ypmsr
3	Va--3s	4	Pw3fsr	8	Yr
3	Va--3s----y	4	Pw3mpr	8	Yv
3	Vag	4	Pw3mso	10	Dh--p
3	Vag-------y	4	Pw3msr	10	Dh--s
3	Vaii1	5	Sp	10	Dh-fso
3	Vaii2p	5	Spca	10	Dh-fsr
3	Vaii2s	5	Spcg	10	Dh1fp
3	Vaii3p	5	Spsa	10	Dh1fs
3	Vaii3s	5	Spsay	10	Dh1fso
3	Vail1p	5	Spsd	10	Dh1fsr
3	Vail1s	5	Spsg	10	Dh1mp
3	Vail2p	5	Spsgy	10	Dh1ms
3	Vail2s	7	Qz	10	Dh2fp
3	Vail3p	7	Qz-y	10	Dh2fs
3	Vail3s	8	Np	10	Dh2fso
3	Vaip1p	8	Npfp-n	10	Dh2fsr
3	Vaip1s	8	Npfpoy	10	Dh2mp
3	Vaip1s----y	8	Npfpry	10	Dh2ms
3	Vaip2p	8	Npfs-n	10	Dh3fp
3	Vaip2s	8	Npfson	10	Dh3fs
3	Vaip3p	8	Npfsoy	10	Dh3fso
3	Vaip3p----y	8	Npfsrn	10	Dh3fsr
3	Vaip3s	8	Npfsry	10	Dh3mp
3	Vaip3s----y	8	Npfsvy	10	Dh3ms
3	Vais1p	8	Npmp-n	10	Ds--p
3	Vais1s	8	Npmpoy	10	Ds--s
3	Vais2p	8	Npmpry	10	Ds1fp-p
3	Vais2s	8	Npms-n	10	Ds1fp-s
3	Vais3p	8	Npms-y	10	Ds1fsop
3	Vais3s	8	Npmsoy	10	Ds1fsos
3	Vam-2p	8	Npmsry	10	Ds1fsos-y
3	Vam-2s	8	Npmsvn	10	Ds1fsrp
3	Vanp	8	Npmsvy	10	Ds1fsrs
3	Vap--sm	8	Y	10	Ds1fsrs-y
3	Vap--sm---y	8	Ya	10	Ds1mp-p
3	Vasp1p	8	Yn	10	Ds1mp-s
3	Vasp1s	8	Ynfpvy	10	Ds1ms-p
3	Vasp2p	8	Ynfsoy	10	Ds1ms-s
3	Vasp2s	8	Ynfsry	10	Ds1msrs-y
3	Vasp3	8	Ynmpoy	10	Ds2---s
4	Dw3--o	8	Ynmpry	10	Ds2fp-p
4	Dw3--r	8	Ynmpvy	10	Ds2fp-s
4	Dw3--r---e	8	Ynmsoy	10	Ds2fsop
4	Dw3-po	8	Ynmsry	10	Ds2fsos
4	Dw3-po---e	8	Ynmsvy	10	Ds2fsos-y
4	Dw3fpr	8	Yp	10	Ds2fsrp
4	Dw3fso---e	8	Yp-p	10	Ds2fsrs
4	Dw3fsr	8	Yp-so	10	Ds2fsrs-y
4	Dw3mpr	8	Yp-sr	10	Ds2mp-p
4	Dw3mso---e	8	Ypfpr	10	Ds2mp-s
4	Dw3msr	8	Ypfs	10	Ds2ms-p

10	Ds2ms-s
10	Ds2msrs-y
10	Ds3---p
10	Ds3---s
10	Ds3fp-s
10	Ds3fsos
10	Ds3fsos-y
10	Ds3fsrs
10	Ds3fsrs-y
10	Ds3mp-s
10	Ds3ms-s
10	Ds3msrs-y
10	Ps--p
10	Ps--s
10	Ps1fp-p
10	Ps1fp-s
10	Ps1fsrp
10	Ps1fsrs
10	Ps1mp-p
10	Ps1mp-s
10	Ps1mprp
10	Ps1mprs
10	Ps1ms-p
10	Ps1ms-s
10	Ps2---s
10	Ps2fp-p
10	Ps2fp-s
10	Ps2fsrp
10	Ps2fsrs
10	Ps2mp-p
10	Ps2mp-s
10	Ps2mprp
10	Ps2mprs
10	Ps2ms-p
10	Ps2ms-s
10	Ps2msrs-y
10	Ps3---p
10	Ps3---s
10	Ps3fp-s
10	Ps3fsrs
10	Ps3mp-s
10	Ps3mprs
10	Ps3ms-s
12	Px3--a
12	Px3--a--------s
12	Px3--a--------w
12	Px3--a--y-----w
12	Px3--d
12	Px3--d--------s
12	Px3--d--------w
12	Px3--d--y-----w
13	Pp--pa
13	Pp--pd
13	Pp--po
13	Pp--pr
13	Pp--sa
13	Pp--sd
13	Pp--sn
13	Pp--so
13	Pp--sr
13	Pp1-pa--------w
13	Pp1-pa--y-----w
13	Pp1-pd--------s
13	Pp1-pd--------w
13	Pp1-pd--y-----w
13	Pp1-pr--------s
13	Pp1-sa--------s
13	Pp1-sa--------w
13	Pp1-sa--y-----w
13	Pp1-sd--------s
13	Pp1-sd--------w
13	Pp1-sd--y-----w
13	Pp1-sn--------s
13	Pp1-sr--------s
13	Pp2
13	Pp2----------s
13	Pp2-pa--------w
13	Pp2-pa--y-----w
13	Pp2-pd--------s
13	Pp2-pd--------w
13	Pp2-pd--y-----w
13	Pp2-po--------s
13	Pp2-pr--------s
13	Pp2-s---------s
13	Pp2-sa--------s
13	Pp2-sa--------w
13	Pp2-sa--y-----w
13	Pp2-sd--------s
13	Pp2-sd--------w
13	Pp2-sd--y-----w
13	Pp2-sn--------s
13	Pp2-so--------s
13	Pp2-sr--------s
13	Pp3-p
13	Pp3-p---------s
13	Pp3-pd--------w
13	Pp3-pd--y-----w
13	Pp3-po--------s
13	Pp3-pr--------s
13	Pp3-s
13	Pp3-sd--------w
13	Pp3-sd--y-----w
13	Pp3-so--------s
13	Pp3-sr--------s
13	Pp3fpa--------w
13	Pp3fpa--y-----w
13	Pp3fpo--------s
13	Pp3fpr--------s
13	Pp3fpr--y-----s
13	Pp3fs---------s
13	Pp3fsa--------w
13	Pp3fsa--y-----w
13	Pp3fso--------s
13	Pp3fsr--------s
13	Pp3fsr--y-----s
13	Pp3mpa--------w
13	Pp3mpa--y-----w
13	Pp3mpo--------s
13	Pp3mpr--------s
13	Pp3mpr--y-----s
13	Pp3ms---------s
13	Pp3msa--------w
13	Pp3msa--y-----w
13	Pp3mso--------s
13	Pp3msr--------s
13	Pp3msr--y-----s
14	R
14	Rc
14	Rgc
14	Rgp
14	Rgpy
14	Rgs
14	Rp
14	Rp-y
14	Rw
14	Rw-y
14	Rz
15	Qn
15	Qn-y
15	Qs
16	I
21	T--p
21	T--po
21	T--pr
21	T--s
21	T--so
21	T--sr
21	Td-po
21	Tdfpr
21	Tdfso
21	Tdfsr
21	Tdmpr
21	Tdmso
21	Tdmsr
21	Tf-s-y
21	Tf-so
21	Tffpoy
21	Tffpry
21	Tffs-y
21	Tffsoy
21	Tfmpoy
21	Tfmpry
21	Tfms-y
21	Tfmsoy
21	Tfmsry
21	Ti-po
21	Tifp-y
21	Tifso
21	Tifsoy
21	Tifsr
21	Tifsry

21	Timp-y
21	Timso
21	Timsr
21	Timsry
21	Ts-po
21	Tsfp
21	Tsfs
21	Tsmp
21	Tsms
22	Di3
22	Di3------e
22	Di3-----y
22	Di3--r
22	Di3--r---e
22	Di3-po
22	Di3-po---e
22	Di3-s----e
22	Di3-sr
22	Di3-sr---e
22	Di3-sr--y
22	Di3fp
22	Di3fpo
22	Di3fpr
22	Di3fpr---e
22	Di3fso
22	Di3fso---e
22	Di3fsr
22	Di3fsr---e
22	Di3mp
22	Di3mpo
22	Di3mpr
22	Di3mpr---e
22	Di3ms
22	Di3ms----e
22	Di3mso---e
22	Di3msr
22	Di3msr---e
22	Di3msr--y
22	Dz3
22	Dz3-po---e
22	Dz3fso---e
22	Dz3fsr---e
22	Dz3mpr---e
22	Dz3mso---e
22	Dz3msr---e
22	Pi3
22	Pi3--r
22	Pi3-po
22	Pi3-pr
22	Pi3-so
22	Pi3-sr
22	Pi3fpo
22	Pi3fpr
22	Pi3fso
22	Pi3fsr
22	Pi3mpo
22	Pi3mpr
22	Pi3mso
22	Pi3msr
22	Pi3msr--y
22	Pz3
22	Pz3-po
22	Pz3-so
22	Pz3-sr
22	Pz3fpr
22	Pz3fso
22	Pz3fsr
22	Pz3mpr
22	Pz3mso
22	Pz3msr
31	C
31	Cccsp
31	Ccssp
31	Ccsspy
31	Crssp
31	Cscsp
31	Csssp
31	Cssspy
50	DATE
100	X

Ro-TimeBank corpus

1 BASIC INFORMATION

1.1 Corpus composition

The corpus consists of:

- 183 files with Romanian news texts (translated from English), with ISO-TimeML and other (name-entities, header, sentence) mark-ups.
- 181 alignment files and 181 parallel English-Romanian files, XCES format, including POS, lemma and chunk attributes.

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

The Ro-TimeBank corpus is represented in XML format.
Each line of an .align file has the format:

TU_ID index_ro index_en S|P

where:

<TU_ID> is the ID of the translation unit (TU) the TEQs belong to;
<index_ro> is the token's ID in the Romanian segment of the TU;
<index_en> is the ID of the English token that is TEQ with the Romanian token;
S | P indicates that the alignment is Sure or Probable.

1.3 Character encoding

UTF8.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Dan Tufiş
Affiliation: Research Institute for Artificial Intelligence, Romanian Academy
e-mail: tufis@racai.ro

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

The first version of the corpus will be available on the MetaShare platform as an archive. Improved versions will be available upon request.

2.3 Copyright statement and information on IPR

The resource is free, license-based for research purposes and fee license-based for commercial purposes.

The TimeBank corpus (description, IPR and copyright) is detailed in [4].
The principles and methodology to obtain the alignments and XCES files are detailed in [2] and [3]. Therefore, the copyright & IPR are governed by these authors.

3 TECHNICAL INFORMATION

3.1 Directories and files

The archive to be uploaded on the MetaShare platform contains:
Ro-TimeBank/data/

Ro-TimeBank/data/align

It contains the 181 files with the English-Romanian alignments. The alignments were automatically obtained and then validated and corrected.

Ro-TimeBank/data/en-ro-msd

The 181 files with the parallel English-Romanian texts, XCES format.

Ro-TimeBank/data/ro

The 183 Romanian files with temporal (TimeML) and other (NE, sentence, header) markups.

3.2 Data structure of an entry

Please see 1.2. and 3.1. above.

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

The Ro-TimeBank corpus contains 4715 sentences, 65375 lexical units; the whole corpus (with annotations) needs about 3.00 MB for disk storage.
There are 125625 words in the parallel /en-ro subfolder, with 18720 sentences.
The .align and the XCES files need 1.36 MB and 7.55 MB.

4 CONTENT INFORMATION

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

The corpus is parallel (XCES format, with alignments) [2, 3].

4.2 The natural language(s) of the corpus

The languages of the corpus are standard Romanian, orthography being compliant with the current Romanian Academy norms and English.

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

The Romanian texts are translations of English news 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)

The Ro-TimeBank corpus [1] is annotated according to TimeML standard [5], including also mark-ups for header, sentence and named-entities information.
The XCES corpus includes mark-ups for POS, lemma, chunks [2].

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

The corpus is (automatically) POS-tagged and (semi-automatically) TIME tagged [1].

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

The alignments were obtained automatically using the YAWA toolkit and then manually checked [2, 3].

4.4.4 Attributes and their values (if annotated)

```
<!ELEMENT ISO-TimeML-Ro ( #PCDATA | s | EVENT | TIMEX3 | SIGNAL | TLINK | SLINK | ALINK )* >
<!ATTLIST ISO-TimeML-Ro xsi:noNamespaceSchemaLocation CDATA #IMPLIED >
<!ATTLIST ISO-TimeML-Ro xmlns:xsi CDATA #IMPLIED >
<!ATTLIST TimeML comment CDATA #IMPLIED >

<!ELEMENT s ( #PCDATA | EVENT | TIMEX3 )* >

<!ELEMENT EVENT ( #PCDATA ) >
<!ATTLIST EVENT eid ID #REQUIRED >
<!ATTLIST EVENT mainevent (YES | NO) #IMPLIED >
<!ATTLIST EVENT pred CDATA #IMPLIED >
<!ATTLIST EVENT class ( OCCURRENCE | PERCEPTION | REPORTING | ASPECTUAL | I_STATE |
I_ACTION | STATE ) #REQUIRED >
<!ATTLIST EVENT pos ( ADJECTIVE | NOUN | VERB | PREPOSITION | OTHER ) #REQUIRED >
<!ATTLIST EVENT tense ( FUTURE | PAST | SIM_PAST | PLUS_PAST | PRESENT | NONE ) #REQUIRED >
<!ATTLIST EVENT aspect ( NONE | PERFECTIVE | IMPERFECTIVE ) #REQUIRED >
<!ATTLIST EVENT polarity ( POS | NEG ) #REQUIRED >
<!ATTLIST EVENT mood ( SUBJUNCTIVE | CONDITIONAL | IMPERATIVE | NONE ) #REQUIRED >
```

<!ATTLIST EVENT vform (INFINITIVE | GERUNDIVE | PARTICIPLE | NONE) #REQUIRED >
<!ATTLIST EVENT modality (NECESSITY | POSSIBILITY | OBLIGATION | PERMISSION) #IMPLIED >
<!ATTLIST EVENT comment CDATA #IMPLIED >

<!ELEMENT TIMEX3 (#PCDATA) >
<!ATTLIST TIMEX3 tid ID #REQUIRED >
<!ATTLIST TIMEX3 type (DATE | DURATION | SET | TIME) #REQUIRED >
<!ATTLIST TIMEX3 value NMTOKEN #REQUIRED >
<!ATTLIST TIMEX3 anchorTimeID IDREF #IMPLIED >
<!ATTLIST TIMEX3 beginPoint IDREF #IMPLIED >
<!ATTLIST TIMEX3 endPoint IDREF #IMPLIED >
<!ATTLIST TIMEX3 freq NMTOKEN #IMPLIED >
<!ATTLIST TIMEX3 functionInDocument (CREATION_TIME | EXPIRATION_TIME | MODIFICATION_TIME | PUBLICATION_TIME | RELEASE_TIME | RECEPTION_TIME | NONE) #IMPLIED>
<!ATTLIST TIMEX3 mod (BEFORE | AFTER | ON_OR_BEFORE | ON_OR_AFTER | LESS_THAN | MORE_THAN | EQUAL_OR_LESS | EQUAL_OR_MORE | START | MID | END | APPROX) #IMPLIED >
<!ATTLIST TIMEX3 quant CDATA #IMPLIED >
<!ATTLIST TIMEX3 temporalFunction (false | true) #IMPLIED >
<!ATTLIST TIMEX3 valueFromFunction IDREF #IMPLIED >
<!ATTLIST TIMEX3 comment CDATA #IMPLIED >

<!ELEMENT SIGNAL (#PCDATA) >
<!ATTLIST SIGNAL sid ID #REQUIRED >
<!ATTLIST SIGNAL comment CDATA #IMPLIED >

<!ELEMENT TLINK EMPTY >
<!ATTLIST TLINK lid ID #REQUIRED >
<!ATTLIST TLINK relType (BEFORE | AFTER | INCLUDES | IS_INCLUDED | DURING | DURING_INV | SIMULTANEOUS | IAFTER | IBEFORE | IDENTITY | BEGINS | ENDS | BEGUN_BY | ENDED_BY) #REQUIRED >
<!ATTLIST TLINK eventID IDREF #IMPLIED >
<!ATTLIST TLINK timeID IDREF #IMPLIED >
<!ATTLIST TLINK relatedToEvent IDREF #IMPLIED >
<!ATTLIST TLINK relatedToTime IDREF #IMPLIED >
<!ATTLIST TLINK signalID IDREF #IMPLIED >
<!ATTLIST TLINK origin CDATA #IMPLIED >
<!ATTLIST TLINK syntax CDATA #IMPLIED >
<!ATTLIST TLINK comment CDATA #IMPLIED >

<!ELEMENT SLINK EMPTY >
<!ATTLIST SLINK lid ID #REQUIRED >
<!ATTLIST SLINK relType (CONDITIONAL | COUNTER_FACTIVE | EVIDENTIAL | FACTIVE | INTENSIONAL | NEG_EVIDENTIAL) #REQUIRED >
<!ATTLIST SLINK eventID NMTOKEN #REQUIRED >
<!ATTLIST SLINK subordinatedEvent NMTOKEN #REQUIRED >
<!ATTLIST SLINK signalID NMTOKEN #IMPLIED >

<!ATTLIST SLINK syntax CDATA #IMPLIED >
<!ATTLIST SLINK comment CDATA #IMPLIED >

<!ELEMENT ALINK EMPTY >
<!ATTLIST ALINK lid ID #REQUIRED >
<!ATTLIST ALINK relType (CONTINUES | CULMINATES | INITIATES | REINITIATES | TERMINATES) #REQUIRED >
<!ATTLIST ALINK eventID IDREF #REQUIRED >
<!ATTLIST ALINK relatedToEvent IDREF #REQUIRED >
<!ATTLIST ALINK signalID IDREF #IMPLIED >
<!ATTLIST ALINK syntax CDATA #IMPLIED >
<!ATTLIST ALINK comment CDATA #IMPLIED >
-

6.9 Intended application of the corpus

The corpus can be used for NLP applications using temporal information, temporal parsing, machine learning, machine translation, summarization.

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

The corpus was checked for alignments and temporal markups.

5 RELEVANT REFERENCES AND OTHER INFORMATION

[1] Corina Forăscu. Contributions to Romanian language processing through discourse analysis methods. (in Romanian). PhD thesis. Romanian Academy, Bucharest. 2011.

[2] Radu Ion.Word Sense Disambiguation Methods Applied to English and Romanian. (in Romanian). PhD thesis. Romanian Academy, Bucharest. 2007.

[3] Dan Tufiș. A Cheap and Fast Way to Build Useful Translation Lexicons. In Proceedings of the 19th Intl. Conf. On Computational Linguistics, Taipei, 2002, pp. 1030-1036.

[4] Pustejovsky, James, Marc Verhagen, Roser Sauri, Jessica Littman, Robert Gaizauskas, Graham Katz, Inderjeet Mani, Robert Knippen, and Andrea Setzer. (2006). TimeBank 1.2. Linguistic Data Consortium, Philadelphia, ISBN: 1-58563-386-0.

[5] ISO: Language Resource Management – Semantic Annotation Framework (SemAF) – Part 1: Time and Events. Secretariat KATS, 2009. ISO Report ISO/TC37/SC4 N269 (ISO/WD 24617-1).

RO-SAM EUROM Sample

Author(s): Babel project

Institute: University “Politehnica” of Timisoara

Address: Vasile Parvan 2, 1900 Timisoara, Romania

Email:boldea@cs.utt.ro

Date: 1997-09-29 (created) 2004-05-10 (updated)

Version: 3

1. INTRODUCTION

This is a small portion of the Romanian speech data built within the framework of the Copernicus project BABEL. The XML encoding of the speech transcription has been achieved within the „Multext-East” Copernicus project. The entire speech corpus can be aquired from ELRA (ELRA-S0170). The ELRA description says: “The BABEL Romanian Database is a speech database that was produced by a research consortium funded by the European Union under the COPERNICUS programme (COPERNICUS Project 1304). The project began in March 1995 and was completed in December 1998. The objective was to create a database of languages of Central and Eastern Europe in parallel to the EUROM1 databases produced by the SAM Project (funded by the ESPRIT programme).

The BABEL consortium included six partners from Central and Eastern Europe (who had the major responsibility of planning and carrying out the recording and labelling) and six from Western Europe (whose role was mainly to advise and in some cases to act as host to BABEL researchers). The five databases collected within the project concern the Bulgarian, Estonian, Hungarian, Polish, and Romanian languages.

The Romanian database consists of the basic "common" set which is:

* The Many Talker Set: 50 males, 50 females; each to read 4 connected passages, 1 block of 2-3 "filler" sentences, 4 phonemically compact sentences, 3-7 individual sentences, and 26 numbers.
* The Few Talker Set: 5 males, 5 females from the Many Talker Set; each to read additionally 3 blocks of syllables and, in 4 supplemental sessions, 16 connected passages, 4 blocks of 2-3 "filler" sentences, 4 repetitions of the 26 numbers.

* The Very Few Talker Set: 1 male, 1 female from the Few Talker Set; each to read additionally 5 pairs of context words and the syllables in these 5 contexts.”

Part of the translation of the BABEL texts from English into Romanian was carried on at RACAI. Below is the description of the sample we contribute to MetaShare. The XML encoding, done within the Multext-East Project (1995-1998) is due to Tomaz Erjavec of Josef Stefan Institute.

SPEECH FILE FORMATS

The audio data is stored in separate WAVE files, which are PCM encoded at 22Khz, 16 bits per sample, mono.

1.1 DIRECTORY STRUCTURE

All files are stored in a single Directory.

1.2 FILE NAMING CONVENTIONS

Each filename starts with the string „spch„ followed by the two digit file number and the string „-ro.wav” (eg. „spch01-ro.wav”). The file number is 0-based.

1.3 LABEL FILES

The file „spch-ro.xml” is an XML file containing the TEI header (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-teiHeader.html) for this resource and separate descriptions for each audio file.

An example of one audio file description is explained below:

```
   <text id="mtes-ro." lang="ro">
     <body lang="ro">
       <div id="sro.1" n="O0" type="block">
         <head>*BLOCK: O0</head>
         <p id="sro.1.2">
<s id="sro.1.2.1"> Sentence 1 </s>
<s id="sro.1.2.2"> Sentence 2 </s>
<s id="sro.1.2.3"> Sentence 3 </s>
<s id="sro.1.2.4"> Sentence 4 </s>
<s id="sro.1.2.5"> Sentence 5 </s>
         </p>

         <ab>[<xref url="spch00-ro.wav">speech file</xref>]</ab>
       </div>
.
.
.
     </body>
   </text>
```

Each tag refers to one audio file and contains the following tags: <head>, <p> and <ab>. The <p> tag has an <s> tag for each sentence in the audio file (each audio file contains 5 sentences). The <ab> tag is the reference to the audio file on the disk (filename).

2. DATA

The database is split into 40 audio files, each file containing a number of 5 sentences, with a total number of 200 sentences.

2.1 PHONETIC TRANSCRIPTION (YES/NO)

No phonetic transcription was provided and the file transcription.txt was automatically generated with our phonetic transcription tool (no text normalization present).

3. SEGMENTATION

3.1 SEGMENTATION TYPE

The audio files are segmented at sentence level.

3.2 TOKEN COUNT

The database is composed of 200 sentences with a total number 2203 words (1026 distinct).
The occurance count of each allophone in the corpus is as follows:

Allophone	Count	Allophone	Count
@	5	l	380
ch	155	m	320
o@	58	n	615
z	102	o	345
e@	120	a@	190
p	337	a	1491
r	707	b	92
s	604	d	334
t	770	e	1013
u	516	f	117
v	143	g	89
w	43	je	5
h	14	ij	32
i	742	pau	444
j	183	zh	27
k	387	dz	21

The database contains the following isolated digits:

DIGIT	COUNT
1	1
2	5
3	9
4	3
5	5
6	6
7	2

8	3
9	2

No utterances of digit “0” where found.

The database contains the following natural numbers:

Number	Count	Number	Count
10	3	60	2
13	1	62	1
14	1	63	1
15	3	80	1
16	3	84	1
17	1	89	1
20	3	500	1
23	1	584	1
30	4	700	1
36	1	762	1
38	2	900	1
40	2	989	1
46	1	1000	1
		1989	1

The database contains one spontanous date: 6 March 1989 and three occurances of spontanous time strings where found: 06:15, 05:30 and 10:30

The total number of diphones is 10212 with 518 distinct values and the total number of triphones is 10023 with 2705 distinct values.

4. LEXICON

A lexicon file (LEXICON.TBL) was automatically generated. Each line contains one word and its occurence count separated by <tab>:

word<tab>count

5. SPEAKER DEMOGRAPHIC INFORMATION

Unknown.

5.1 ACCENT/REGIONS

Unknown.

5.2 SPEAKER AGES

Unknown.

5.3 SPEAKER OVERLAP

Unknown.

6. RECORDING CONDITIONS

6.1 SOFTWARE

Unknown.

6.2 HARDWARE

Unknown.

7. TEST MATERIAL

None provided

Multilingual Subjectivity Analysis: Gold Standard Data SET

1 BASIC INFORMATION

1.1 The data set composition

The data represents a set of 1590 of language quotations (reported speech) manually annotated for sentiment (POSitive, NEGative, OBJective/neutral) towards entities mentioned inside the quotation. Objective is the default, meaning that the absence of a label can be interpreted as OBJ. Each of the quotations is characterized by:

- the news ID (e.g. dailymail-be691eeeedee8e7d24eca8299e84937c)
- the quotation itself (note that the sentiment value refers to the entity mentioned inside the quotation, and not to the entire text of the quotation)
- source name
- source ID
- target ID
- target name
- the sentiment mark-up by a pair of annotators (Ann1...Ann4)
- An agreement label (TRUE or FALSE) between the annotators

Source Name refers to the person who issued the quotation (e.g. Angela Merkel said: "..."). Target Name refers to the entity mentioned inside the quotation, i.e. the entity whose sentiment value we are interested in (e.g. " ... Tony Blair ..."). Ann1 to Ann4 refers to the four different human annotators. The Agreement column simply shows whether the pairs of annotators agreed or not.

The data set is accompanied by the annotation guidelines the authors used to annotate the examples.

1.2 Representation of the data set (flat files, database, markup)

The data set is provided as an Excel file with three sheets (the first sheet contains reference information, the second contains the data set itself and the third sheet contains the annotation guidelines.

1.3 Character encoding

The characters are UTF8 encoded.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Ralf Steinberger
Address: Joint Research Centre, ISPRA, Italy
Affiliation: Institute for the Protection and Security of the Citizen (IPSC)
Position: Head of Language Technology Group
Telephone: +39 - 0332 78 6271 + 5648
Fax: +39 - 0332 78 5154
e-mail: Ralf.Steinberger@jrc.ec.europa.eu

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

The data set can be freely downloaded from the address below:
http://langtech.jrc.ec.europa.eu/JRC_Resources.html .

2.3 Copyright statement and information on IPR

The resource is free, with requested citation of the relevant papers (see below)

3 TECHNICAL INFORMATION

3.1 Directories and files

Not relevant

3.2 Data structure of an entry

The data set is represented in a Excel sheet, one quotation snippet per line, followed by the annotation information (see above section 1.1)

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

1590 annotated snippets, 0.6 MB on disk

4 CONTENT INFORMATION

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

Monolingual annotated snippets

4.2 The natural language(s) of the corpus

English

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

Journalism

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)

Human labeling on positivity or negativity of the snippet (towards the target person or organization)
Agreement between annotators, labeled automatically

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

POS(itive), NEG(ative), empty (objective) and TRUE or FALSE (annotators agreement)

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)

Not relevant

6.11 Intended application of the corpus

The purpose of the annotation was to produce a gold standard collection of news snippets to be used in training/evaluation opinion mining crawlers, sentiment classifiers.

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

Manual mark-up of two annotators (most of the time)

5 RELEVANT REFERENCES AND OTHER INFORMATION

Balahur-Dobrescu Alexandra & Ralf Steinberger (2009). Rethinking sentiment analysis in the news: from theory to practice and back. 'Workshop on Opinion Mining and Sentiment Analysis' (WOMSA), held at the 2009 CAEPIA-TTIA 13th Conference of the Spanish Association for Artificial Intelligence, pp. 1-12. Sevilla, Spain, 13.11.2009. Available from: http://langtech.jrc.ec.europa.eu/Documents/09_WOMSA-WS-Sevilla_Sentiment-Def_printed.pdf

Balahur Alexandra, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik van der Goot, Matina Halkia, Bruno Pouliquen & Jenya Belyaeva (2010). *Sentiment Analysis in the News*. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010), pp. 2216-2220. Valletta, Malta, 19-21 May 2010. Available from: http://langtech.jrc.it/Documents/2010_03_LREC_Sentiment-analysis.pdf

Balahur Alexandra, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik van der Goot, Matina Halkia, Bruno Pouliquen & Jenya Belyaeva (2010). Sentiment Analysis in the News. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010), pp. 2216-2220. Valletta, Malta, 19-21 May 2010. Available from: http://langtech.jrc.ec.europa.eu/Documents/2010_03_LREC_Sentiment-analysis.pdf

Partner UOM

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: Grazio Falzon % Toni Sant
Address: t.sant@hull.ac.uk
Affiliation:
Position: Falzon is retired and "out of touch with the internet" (Toni Sant)
e-mail: http://www.aboutmalta.com/ % Toni Sant: t.sant@hull.ac.uk

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

2.3 Copyright statement and information on IPR
On the website: "Copyright © 1997, Grazio Falzon" Negotiations are being undertaken via Toni Sant (Consulting editor of aboutmalta.com).

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

Illum Corpus

1 BASIC INFORMATION

1.1 *Corpus composition*

The full editions of ILLUM from 12/11/2006 to 30/05/2010 (185 issues).

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

XML files with paragraph marking (<paragraph> … </paragraph>) and each word on a separate line.

1.3 Character encoding

UTF-8

2 ADMINISTRATIVE INFORMATION

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

Name: Saviour Balzan
Affiliation: MediaToday Co. Ltd
Address: Vjal ir-Rihan, San Gwann SGN SGN 9016, Malta
Telephone: +39 0332 78-5648 or 78-9478
Fax: +39 0332 78-5154
e-mail: illum@mediatoday.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

 Managing Editor Saviour Balzan agreed to the texts being used. We are currently working out which licence schema he is going to choose.

3 TECHNICAL INFORMATION

3.1 *Directories and files*

1 folder containing 5,269 XML files (with one article each)

3.2 Data structure of an entry

<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<article>

<source>ILLUM</source>
<url>http://www.illum.com.mt/2006/11/12/emmanuel_micallef.html</url>
<date>2006/11/12</date>
<text>
<paragraph>
…
</paragraph>
…
</text>
</article>

3.3 Corpora size (nmb. of tokens, MB occupied on disk)
2,249,294 tokens
39.7 MB on disk

4 CONTENT INFORMATION

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

4.2 The natural language(s) of the corpus
Maltese

4.3 Domain(s)/register(s) of the corpus
News

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)

Paragraph mark-up

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

--

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

--

4.4.4 Attributes and their values (if annotated)

--

4.5 Intended application of the corpus
Text corpus to be tagged for linguistic research

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

5 RELEVANT REFERENCES AND OTHER INFORMATION

Laws of Malta MT

1 BASIC INFORMATION

1.1 *Corpus composition*

The corpus contains the Laws of Malta in Maltese from the official government website. The unannotated raw text files were extracted from the pdf files that can be found on the website.

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

Folder structure is: laws_mt/txt/FILE1.txt, … FILEn.txt
There are no subfolders in the txt folder.

1.3 Character encoding

UTF-8

2 ADMINISTRATIVE INFORMATION

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

Name: Ministry for Justice and Home Affairs
Address: 32/33 House of Catalunya, Marsamxett Road, Valletta VLT1955
Affiliation: http://www.justiceservices.gov.mt/lom.aspx?pageid=24
Telephone: (+356) 2295 7000
Fax: (+356) 2295 7348
e-mail: lawdrafting.unit@gov.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 as a folder with text files.

2.3 Copyright statement and information on IPR

The resource is freely available for personal and non-commercial use. Government of Malta must be referred to as the source of the original documents.

3 TECHNICAL INFORMATION

3.1 *Directories and files*

A flat directory containing all the text files.

3.2 Data structure of an entry

Just text and line breaks taken over from the PDF files.

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

69,9 MB on disk

4 CONTENT INFORMATION

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

The corpus is monolingual, but it is planned to align it with the respective English language corpus of Maltese legislation to achieve a parallel corpus at a later stage. The text consists of non-annotated raw data.

4.2 The natural language(s) of the corpus
Maltese

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

4.4 Annotations in the corpus (if an annotated corpus)
The corpus is not annotated.

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
Machine translation

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

5 RELEVANT REFERENCES AND OTHER INFORMATION
The law texts were downloaded as pdf files by a web crawler and the text extracted from them. The website of the Laws of Malta is here: http://www.justiceservices.gov.mt/lom.aspx?pageid=24

Laws of Malta EN

1 BASIC INFORMATION

1.1 *Corpus composition*
The corpus contains the Laws of Malta in English from the official government website. The unannotated raw text files were extracted from the pdf files that can be found on the website.

1.2 Representation of the corpora (flat files, database, markup)
Folder structure is: laws_en/txt/FILE1.txt, … FILEn.txt
There are no subfolders in the txt folder.

1.3 Character encoding
UTF-8

2 ADMINISTRATIVE INFORMATION

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

Name: Ministry for Justice and Home Affairs
Address: 32/33 House of Catalunya, Marsamxett Road, Valletta VLT1955
Affiliation: http://www.justiceservices.gov.mt/lom.aspx?pageid=24
Telephone: (+356) 2295 7000
Fax: (+356) 2295 7348
e-mail: lawdrafting.unit@gov.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 as a folder with text files.

2.3 Copyright statement and information on IPR
The resource is freely available for personal and non-commercial use. Government of Malta must be referred to as the source of the original documents.

3 TECHNICAL INFORMATION

3.1 *Directories and files*
A flat directory containing all the text files

3.2 Data structure of an entry
Just text and line breaks taken over from the PDF files.

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

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)
The corpus is monolingual, but it is planned to align it with the respective Maltese language corpus of Maltese legislation to achieve a parallel corpus at a later stage. The text consists of non-annotated raw data.

4.2 The natural language(s) of the corpus
English

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

Legalese

4.4 Annotations in the corpus (if an annotated corpus)
There are no annotations in the 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
Machine translation

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

5 RELEVANT REFERENCES AND OTHER INFORMATION
The law texts were downloaded as pdf files by a web crawler and the text extracted from them. The website of the Laws of Malta is here: http://www.justiceservices.gov.mt/lom.aspx?pageid=24

Maltese Wordlist

1. BASIC INFORMATION

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

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

1.3 Character encoding
UTF-8

2. ADMINISTRATIVE INFORMATION

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

Name: Ramon Casha
Address: 65, Triq Stagno, Qormi
Affiliation: Malta Linux User Group
Position: Lead Developer
Telephone: (+356) 99477331
Fax: +
e-mail: rcasha@gmail.com

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

3. TECHNICAL INFORMATION

3.1 Directories and files

one single .txt file containing all entries.

3.2 Data structure of an entry

one word form per line

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

824,839 unique words, 10.3MB uncompressed

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

Maltese

4. 2 Entry Type

one word form per line

4.3 Attributes and their values

none

4.4 Coverage of the lexicon

not known

4.5 Intended application of the lexicon

Spell checkers, Linux OS

4.6 POS assignment

No

4.7 Reliability (automatically/manually constructed)

automatically constructed

5. RELEVANT REFERENCES AND OTHER INFORMATION

The wordlist has been developed by the author/copyright holder for the Linux user group http://linux.org.mt.

Partner UPC

AGORA

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

The Agora database contains the recordings of 34 TV shows of Catalan public broadcast TV3. The shows are highly moderated debates with a high variation in topics and invited speakers. The database consists of 68 files with a total audio time of 43h. Each file corresponds to half show of an airing day with an average duration of 38 min. The transcription follows the general guideline generated within the TC-STAR project for European Parliament Plenary Sessions but it was extended to include additional information as the language, background condition, silence/voice segmentation, speaker segmentation and acoustic events. The transcriptions have four layers. Transcriptions follow the TRS format produced by the Transcriber transcribing tool.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: José A. R. Fonollosa
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 4016439
Fax: +34 93 401 6447
e-mail: jose.fonollosa@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 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, …), video and audio files, and transcription files in the same directory.

3.2. Encoding

Documentation is encoded in word/pdf/plain text.

64 video are stored as MPEG-2. Their coresponding audio files are stored as 16-bit 32 kHz uncompressed speech samples . Files have an average duration of 38 min.

Each audio file has an accompanying XML transcription file.
The XML transcription files contain information about the database, speakers, turns, segmentation, background sounds, channel and literal transcriptions.

3.3. Resource size (size of recorded speech/MB occupied on disk)

The corpus contains about 43 hours 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 Catalan for broadcast news and TV debates applications

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 recordings of 34 TV shows of Catalan public broadcast TV3

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

The transcription follows the general guideline generated within the TC-STAR project for European Parliament Plenary Sessions but it was extended to include additional information as the language, background condition, silence/voice segmentation, speaker segmentation and acoustic events. The transcriptions have four layers. 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 871 adult Catalan speakers (441 male, 113 female, 317 unknown), and 157 adult Spanish speakers (83 male, 29 female, 45 unknown). Speakers may originate from different accents. Speakers are unbalanced in gender favouring male speakers in total duration.

4.6. Recording platform

4.6.1.Domain(s), environments,

All the shows were performed in a closed TV studio

4.6.2. Recording platform

Not known

Bilingual (Spanish English) Speech synthesis HTS models

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

This database contains Bilingual (English and Spanish) Hidden Markov Models (HMM) for their use in Festival TTS using the HTS toolkit. Speaker dependent models were trained with more than

4h30m of speech (2h15m for each language) from 2 female bilingual speakers and 2 male bilingual speakers. The speech data can be found in the TC-STAR Bilingual Voice-Conversion Spanish Speech Database and in the TC-STAR Bilingual Expressive Spanish Speech Database that are also available in the ELRA catalogue and Meta-Share. The Bilingual (Spanish English) Speech synthesis HTS models were created within the scope of the METANET4U project funded by the European Commission

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 6437
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

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), and all the files required to use the models either with the HTS Toolkit or with the Festival TTS system.

3.2. Encoding

Documentation is encoded in plain text.

All the data files are compatible with the HTS-2.2 and Festival 2.1 (available version in Nov. 2011).

3.3. Resource size (size of recorded speech/MB occupied on disk)

For each speaker (4) and language (English/Spanish) a festival-compatible voice folder of about 3MB is provided

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This resource includes statistical models for producing synthetic speech in the framework of statistic parametric synthesis. The Hidden Markov Models (HMM) are compatible with the HTS-2.2 toolkit. This toolkit generates speech based on statistical models. It requires that the input text is represented by phonetic-prosodic labels. These labels can be generated using the Festival 2.1 system. The hts-engine is integrated in Festival so that it can be used as a stand-alone TTS system.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

The voices were trained from speech data produced to support research on voice conversion and on expressive speech synthesis, in the framework of speech-to-speech translation.
For each speaker, and for each language, three different speech corpora are available:

- C33: mimick sentences. The corpus consist on approximately 15 minutes of short sentences pronounced mimicking the prosody of a acoustic template.
- C11: paralel corpus. The duration of the data is approximately 1 hour. The speakers read short paragraphs derived from transcriptions of the European Parliament. The Spanish text is the translation of the English selection.

C40: expressive corpus. The duration of the data is also approximately 1 hour. As in the previous case, the corpus consist on short paragraphs from the European Parliament and the Spanish and English text are paralel corpus. The difference is that 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 Lexicon. Description of the lexicon (if applicable)

The input text is transcribed phonetically and prosodically using the Festival front-end. 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

4.4. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

The resource includes two male and two female speakers. Each speaker is bilingual (English/Spanish). In the selection process several candidates were considered. The selection criteria included language profile (native Spanish and English), profession (professional speakers or radio speakers were preferred) and suitability of the voice for speech synthesis.

4.5. Recording environment

The speech resources were recorded in a in-house recording studio. Three channels were recorded at 96kHz and 24bits/sample: high-quality membrane microphone, close microphone and laryngograph. The speakers, located in an isolated room read the prompt text from a screen. An operator in the recording room controlled the computer (prompts and signal levels) while an expert check the pronunciation and style.

The models included in this resource are derived from the membrane microphone with signals downsampled to 16kHz and 16bits/sample.

3/24 BN (Catalan BN)

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

The 3/24 BN database contains 80 hours of recordings of the Catalan news television channel 3/24. 19 hours of this database are fully transcribed while the remaining data are solely segmented and annotated with respect to speaking style, recording condition and

speaker. The transcription follows the general guideline generated within the TC-STAR project for European Parliament Plenary Sessions but it was extended to include additional information as the language, background condition, silence/voice segmentation, speaker segmentation and acoustic events featuring additional time stampsThe transcriptions have four layers. Transcriptions follow the TRS format produced by the Transcriber transcribing tool. .

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: José A. R. Fonollosa
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 4016439
Fax: +34 93 401 6447
e-mail: jose.fonollosa@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

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, …), 20 speech files (4 complete hours per file), each having an accompanying video, and transcription file) in the same directory.

3.2. Encoding

Documentation is encoded in word/pdf/plain text.

The recorded video files are stored as MPEG-2. Whereas the extracted speech files are stored as 16-bit 48 kHz uncompressed speech samples .

Each speech file has an accompanying transcription file.
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. translator, guest), 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 20 speech files (audio and video)/80 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 ASR systems in Catalan, 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 contents of the database are: Complete broadcast news sessions including 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 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 recordings contain segments from 1599 adult Catalan speakers of unknown accent (981 male, 421 female, 197 unknown), and 351 adult Spanish speakers (217 male, 82 female, 52 unknown) Speakers are unbalanced in gender favouring male speakers in total duration.

4.6. Recording platform

4.6.1.Domain(s), environments,

Four recording environments were defined:

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 originate from DVB-T video streams, whereas the audio channel is provided with 48 kHz sample rate, 16 bit uncompressed samples. The video streams are MPEG-2 encoded.

Catalan-SpeechDat For the Fixed Telephone Network Database

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

This speech database contains the recordings of 2000 Catalan speakers who called from Fixed telephones and who are recorded over the fixed PSTN using and ISDN-BRI interface. Each speaker uttered around 50 read and spontaneous items. The speech database follows the specifications made within the SpeechDat (II) project. The database was validated by UVIGO. The Catalan-SpeechDat for the Fixed Telephone Network 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

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 1.000.00 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,…

The contents of the database are: application words, digits, numbers, currency amounts, dates, times, word spotting phrases, spellings, forenames, surnames, cities, company names, yes/no, phonetically rich sentences. 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 2000 adult speakers. Speakers were selected from four different accents from Catalonia plus speakers from Valencia and Balearic islands. Speakers are balanced in sex and gender and distributed in three age groups.

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.

Catalan-SpeechDat for the Mobile Telephone Network Database

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

This speech database contains the recordings of 2000 Catalan speakers who called from GSM telephones and who are recorded over the fixed PSTN using and ISDN-BRI interface. Each speaker uttered around 50 read and spontaneous items. The speech database follows the specifications made within the SpeechDat (II) project. The database was validated by UVIGO. The Catalan-SpeechDat for the Mobile Telephone Network 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

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, labeling session, transcriptions and annotations.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The corpus contains about 100000 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,…

The contents of the database are: application words, digits, numbers, currency amounts, dates, times, word spotting phrases, spellings, forenames, surnames, cities, company names, yes/no, phonetically rich sentences. 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 2000 adult speakers. Speakers were selected from four different accents from Catalonia plus speakers from Valencia and Balearic islands. Speakers are balanced in sex and gender and distributed in three age groups.

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. Recording software is ADA.

Spanish EUROM.1

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

EUROM1 is a multilingual European speech database. 60 speakers per language who pronounced numbers, passages, sentences, CVCV words. using close talking microphone in an anecoic room. Equivalent corpora for each of the European languages: same number of speakers selected in the same way, and recorded in the same conditions with common file formats.

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

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 in a nested structured directory.

3.2. Encoding

Documentation is encoded in word text. It contains the read text and the canonical phonetic transcription is Castilian Spanish

Speech files are stored as sequences 16 kHz Each prompted utterance is stored within a separate file.

Each speech file has an accompanying file with the orthographic transcription.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The corpus contains about 13000 speech files.

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database for training and assessment of ASR systems in Spanish. The database fulfills the EUROM1 specifications.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

Many Talker Corpus (60 speakers):
- 100 numbers, each spoken once
- 5 sentences, each spoken once
- 3 passages.

Few Talker Corpus (10 speakers):
- 100 Numbers, each spoken 5 times
- 25 sentences
- 15 passages
- CVCV words, each spoken 5 times.

Very Few Talker Corpus (2 speakers):
- 82 CVCV words embedded in 5 different carrier phrases, spoken once
- 10 carrier phrase words, each spoken five times.

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

All the recordings were supervised, at recording time, by a technician who supervised the quality of the signal and one linguistic who supervised pronuntiation errors. Both persons were located in a room close to the anecoic room were the speaker was talking. In case of errors, the speaker was requested to repeat the utterance. For this reason, no pronuntiation errors were expected and no other labelling was performed.
All the prompted text was manually phonetically transcribed in canonical Castillian.

4.4. Lexicon. Description of the lexicon (if applicable)

n.a.

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

The database contains recordings from 60 adult speakers. Speakers were selected from Castilian accent. Speakers are balanced in sex and gender.

4.6. Recording platform

4.6.1.Domain(s), environments, Recording mode.

All the recordings were performed in an anecoic room.
The recordings contain no speaking errors.

Recording Mode 1: A take was recorded in one complete segment, the sampling and transfer process was started at the beginning of the take, all the acoustic signal was recorded and the sampling and transfer process was only stopped at the end of the take.
Prompting Style 1: ABORT TAKE and re-record. The subject was instructed that if he/she made a speaking error the prompting and recording systems were stopped by an escape mechanism. This situation was indicated to the subject and the prompting system was started to re-record that take.

Mixed Timing Strategy: The timing of the prompt was controlled by a logical combination of apredetermined interval and the endpoint of an utterance. The display of each new promt was triggered by whichever was later of the predetermined interval or the endpoint.

Manual Timing: Used only for passages.

The relevant parameters selected for Mixed Timing were as follows:
Extinction level: -40 dB (level the signal must cross down to be considered as silence)

End Signal Silence: 500 ms (duration of silence that determine the end of recordings)

4.6.2 Recording platform

The recording platform was based in a PC with a software generated in the EUROM project.

FESTCAT Catalan 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 Catalan 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 FESTCAT Catalan TTS Baseline Male Speech Database was created within the scope of the FESTCAT/Linkat project funded by the Catalan Government

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 free, license-based, for research purposes 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) in a nested structured directory..

3.2. Encoding

Documentation is encoded in plain text.

Speech files are stored as sequences of 24-bit 96kHz uncompressed speech samples. Each prompted utterance is stored within three separate files, one per channel.

Each speech file has three accompanying label files, identified by the file extesion:

- .caL: ascii file with time of the closure glottal instants (pitch marks). The file was derived automatically from the signal file using praat.
- .caP: ascii file with phonetic transcription and phonetic segmentation. The segmentation has been derived automatically by our in-house HMM-based segmentation tooklkit.
  .caS: 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 5,300 speech files/ 11 hours of recorded speech and needs about 30 GB for disk storage. A simplified version of one channel (membrane microphone), 16kHz and 16bits is also provided (1.4GB).

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database for training TTS systems in Catalan. 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. Some of them are general and the goal is provide high phonetic and prosodic coverage in several domains: novels, dialogs and monologs, news, software user manuals, transcription of the Catalan parliament, teaching book, phonetically rich sentences and additional questions. Other subcorpora are designed for specific applications: numbers, dates, villages and cities, URLs, spellings, screen reading commands, IVR commands, company names and brands, etc. Furthermore, a small Spanish corpus is also included so that the Spanish phonemes and diphones are represented.

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

All the speech data is labeled with orthographic and phonetic transcription.
The phonetic segmentation and the pitch labeling was created automaticall 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

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 male professional speaker, adult, from the central Catalan 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 simulatenously 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..

FESTCAT Catalan 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 Catalan 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 FESTCAT Catalan TTS Baseline Female Speech Database was

created within the scope of the FESTCAT/Linkat project funded by the Catalan Government

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

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 plain text.

Speech files are stored as sequences of 24-bit 96kHz uncompressed speech samples. Each prompted utterance is stored within three separate files, one per channel.

Each speech file has three accompanying label files, identified by the file extension:

.caL: ascii file with time of the closure glottal instants (pitch marks). The file was derived automatically from the signal file using Praat.

.caP: ascii file with phonetic transcription and phonetic segmentation. The segmentation has been derived automatically by our in-house HMM-based segmentation tooklkit.

.caS: 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 5,300 speech files/ 11 hours of recorded speech and needs about 30 GB for disk storage. A simplified version of one channel (membrane microphone), 16kHz and 16bits is also provided (1.4GB).

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database for training TTS systems in Catalan. 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. Some of them are general and the goal is provide high phonetic and prosodic coverage in several domains: novels, dialogs and monologs, news, software user manuals, transcription of the Catalan parliament, teaching book, phonetically rich sentences and additional questions. Other subcorpora are designed for specific applications: numbers, dates, villages and cities, URLs, spellings, screen reading commands, IVR commands, company names and brands, etc. Furthermore, a small Spanish corpus is also included so that the Spanish phonemes and diphones are represented

4.3. Transcriptions, annotations:

All the speech data is labeled with orthographic and phonetic transcription.

The phonetic segmentation and the pitch labeling was created automaticall 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

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 was recorded by a female professional speaker, adult, from the central catalan 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

FESTCAT-SEL Catalan TTS Baseline 8 spks x 1h

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

This database contains the recordings of four female and four male Catalan 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 1 hours of speech per speaker. The FESTCAT Catalan TTS Baseline 8 spks x 1h Speech Database was created within the scope of the FESTCAT/LINKAT project funded by the Catalan Government. In the project 10 speakers were recorded in order to select two speakers. This resource comprises the eight speakers which were not selected and that can be used for creating synthetic voices with medium-size databases

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 free, license-based, for research purposes 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), speech files (one per utterance) and label files (each speech file has accompanying label files) in a nested structured directory.

3.2. Encoding

Documentation is encoded in plain text.

Speech files are stored as sequences of 24-bit 96kHz uncompressed speech samples. Each prompted utterance is stored within three separate files, one per channel.

Each speech recording has three accompanying label files, identified by the file extesion:

.caL: ascii file with time of the closure glottal instants (pitch marks). The file was derived automatically from the signal file using praat.

.caP: ascii file with phonetic transcription and phonetic segmentation. The segmentation has been derived automatically by our in-house HMM-based segmentation tooklkit.

.caS: 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)

Each speaker recorded approximately 500 sentences / 70 min. of recorded speech. It needs about 20 GB for disk storage (for the eight speakers). A simplified version of one channel (membrane microphone), 16kHz and 16bits is also provided (1.2GB).

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database to be used for building multispeaker voices for text-to-speech in Catalan. For each speaker, a medium-size database was produced. It allows to create synthetic voices using both unit selection technology or statistical synthesis.

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 reads approximately 500 short paragraphs selected from Catalan novels. To increase the inter-speaker phonetic variability, five different (text) corpus were produced (each corpus is read at most by one male speaker and one female speaker). At each corpus, the paragraphs were selected from a large paragraph collection so that the phonetic and prosodic variability is maximized

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

The database include orthographic transcription, phonetic transcription, phonetic segmentation and labelling of the closure glottal instants. All the information is derived automatically from the speech signals and the prompts (text corpora). The text corpus were carefully produced so that there are no pronuntiation ambiguity in the text. Furthermore, during the recording sessions, the operators check that the speakers read properly the corpus. Minor annotations about pronunciation are registered and used in the label files

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. The phonetic transcription was derived automatically using the tools and lexica available at the UPC laboratory.

4.5. Speakers.

The database contains recordings from 10 adult speakers, five male and five female. Speakers were selected by a casting agency with the requirements that they are professional speakers, from the central dialect.

4.6. Recording platform

The database was recorded in an 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 consist 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

Catalan FreeSpeech Database

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

The FreeSpeech Catalan database was recorded in 1999 for automatic dictation purposes. 148 speakers (100 adult, 48 children) belonging to the 4 main Catalan dialects read texts from several domains. The signals were recorded at 16 kHz with a headset microphone and a FreeSpeech (IBM) mouse, which has a built-in microphone. The Catalan FreeSpeech Database was funded by the Catalan Government.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Javier Hernando
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 4016433
Fax: +34 93 401 6447
e-mail: javier.hernando@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 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), 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 plain text.

Speech files are stored as sequences of 16-bit 16 kHz uncompressed speech samples. Each prompted utterance is stored within a separate file.

Each speech file has an accompanying XML/ASCII SAM label file
Label files contain information about the database, speech signal coding, speakers, turns, segmentation, labeling session, transcriptions and annotations.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The corpus contains about 120 hours of recorded speech (about 15 GB)

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database for training ASR systems in Catalan for dictation applications.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

The text read by the speakers was provided by Philips, and includes a wide diversity of topics and writing styles.

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

Only the texts that were read by the speakers are provided. The recordings were not posteriorly checked for consistency with the read text.

4.4. Lexicon. Description of the lexicon (if applicable)

N.A.

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

The database contains recordings from 148: 50 women, 50 men and 48 children. Speakers were selected from the four main Catalan dialects.

4.6. Recording platform

4.6.1.Domain(s), environments,

Recordings were performed in a sound proof room

4.6.2 Recording platform

The recording platform is based on equipment provided by Philips. The signals were recorded at 16 kHz with a headset microphone and a FreeSpeech (IBM) mouse, which has a built-in microphone.

TALP Tourism Dialogues - Spanish

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

The "TALP Tourism Dialogues - Spanish" comprises the recordings of 210 dialogues (10,600 turns) in the touristic domain. Each participant took the role of either customer or agent and talked spontaneously to achieve a predefined goal defined in a given scenario. The scenarios include hotel, travel agency, tourism information office and railway/airline company. The data was recorded over the telephone (a-law, 8kHz) using a platform that imposed half-duplex communication: there are not turn overlapping. The total recordings time is 30 hours. The database includes the orthographic transcription enriched with additional labels to indicate external noises, speaker noises and disfluences in spontaneous speech

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Nuria Castells
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 4137856
Fax: +34 93 401 6447
e-mail: castell@lsi.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

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 file for each turn) and label files (each speech file has an accompanying label file) in a nested structured directory.

3.2. Encoding

Documentation is encoded in plain text.

Speech files are stored in a riff (wave) file with a header and as sequences of 8-bit 8 kHz A-law uncompressed speech samples (CCITT G.711 recommendation). Each dialogue is in a separate directory and each turn is store in a separate file.

Each speech file has an accompanying SAM label file.
Label files contain information about the database, speech signal coding, speaker code, segmentation and enhanced orthographic transcription

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The corpus contains about 10.600 files (turns) of recorded speech (30 hours)

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

The database was created for collecting spontaneous dialogue data in the tourism domain. The data can be used to study dialog and spontaneous speech. It can also be used for train domain-specific acoustic and language models for ASR systems in 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 contents of the database are spontaneous dialogs in the tourism domain. Two volunteers talk to each other using a telephone plataform. Therefore, all the iteration is by voice. One speaker acts as the agent while the other plays the role of client. They are provided with an scenario (goal) and some tools related with the task.

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 annotation also marks some disfluences as re-start, end of word missing, etc. 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.

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. The pronunciation information has been produced automatically using the in-house UPC grapheme-to-phoneme system. The number of entries is around 106k

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

There were no specific speaker selection strategy but the accent and gender of the speakers was registered. The number of speakers is 76.

4.6. Recording platform

The database was recorded using a telephone platform called GAIA, developed at UPC. The client called the platform to ask for an agent. The agent is connected with the client and the speech and other information is logged in the platform.

The platform is half-duplex: each speaker needs to push '#' to give the turn to the other speaker.

Most of the calls with telephones at UPC. Therefore, they represent the internal network.

The platform used digital telephone lines: the speech is recorded directly in the original A-LAW encoding.

TALP Tourism Dialogues - Catalan

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

The "TALP Tourism Dialogues - Catalan" comprises the recordings of 160 dialogues (8,600 turns) in the touristic domain. Each participant took the role of either customer or agent and talked spontaneously to achieve a predefined goal defined in a given scenario. The scenarios include hotel, travel agency, tourism information office and railway/airline company. The data was recorded over the telephone (a-law, 8kHz) using a platform that imposed half-duplex communication: there are not turn overlapping. The total recordings time is 22 hours. The database includes the orthographic transcription enriched with additional labels to indicate external noises, speaker noises and disfluences in spontaneous speech.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Nuria Castells
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 4137856
Fax: +34 93 401 6447
e-mail: castell@lsi.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

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 file for each turn) and label files (each speech file has an accompanying label file) in a nested structured directory..

3.2. Encoding

Documentation is encoded in plain text.

Speech files are stored in a riff (wave) file with a header and as a sequences of 8-bit 8 kHz A-law uncompressed speech samples (CCITT G.711 recommendation). Each dialogue is in a separate directory and each turn is store in a separate file.

Each speech file has an accompanying SAM label file.
Label files contain information about the database, speech signal coding, speaker code, segmentation and enhanced orthographic transcription

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The corpus contains about 8.600 files (turns) of recorded speech (22 hours)

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

The database was created for collecting spontaneous dialogue data in the tourism domain. The data can be used to study dialog and spontaneous speech. It can also be used for train domain-specific acoustic and language models for ASR systems in Catalan.

4.2. Content description.

The contents of the database are spontaneous dialogs in the tourism domain. Two volunteers talk to each other using a telephone platform. Therefore, all the iteration is by voice. One speaker acts as the agent while the other plays the role of client. They are provided with an scenario (goal) and some tools related with the task

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 annotation also marks some disfluences as re-start, end of word missing, etc. 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.

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. The pronunciation information has been produced semi- automatically using the in-house UPC grapheme-to-phoneme system. Special care was put in proper names to identifiy the language of the names and transcribe according to that language. The number of entries is around 106k

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

There were no specific speaker selection strategy but the accent and gender of the speakers was registered. The number of speakers is 58.

4.6. Recording platform

The database was recorded using a telephone platform called GAIA, developed at UPC. The *client* called the platform to ask for an *agent.* The *agent* is connected with the *client* and the speech and other information is logged in the platform.

The platform is half-duplex: each speaker needs to push '#' to give the turn to the other speaker.

Most of the calls with telephones at UPC. Therefore, they represent the internal network.

The platform used digital telephone lines: the speech is recorded directly in the original A-LAW enconding..

TALP Tourism Dialogues - Translation

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

This corpus consist of the translation into English and either Spanish or Catalan of the transcriptions in the ”TALP Tourism Dialogues – Spanish” and in the ”TALP Tourism Dialogues – Catalan” databases. The dialogues were recorded over the telephone network. Each participant took the role of either customer or agent and were provided with a scenario describing the goal of the conversation. The spoken dialogs were transcribed and the translated into two other languages so that a tri-lingual corpus was produced.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Nuria Castells
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 4137856
Fax: +34 93 401 6447
e-mail: castell@lsi.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

3. TECHNICAL INFORMATION

3.1. Directories and files

The archive uploaded on the MetaShare platform contain two files containing documentation (copyright, readme), and for each language (English, Catalan, Spanish) there are two files, one with the transcriptions or translations of the Spanish dialogs and the other with the transcriptions or translations of the Catalan dialogs. Each file contains one line per turn, with a code to identify dialog, turn in the dialog, scenario code, speaker code, language and source language.

3.2. Encoding

Documentation is encoded in text.

Translation files are encoded using ISO-8859.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The corpus contains about 19,000 turns translated into three languages. The size of the uncompressed database is less than 10MB.

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

The primary use of the database is for supporting the research on machine translation of spoken language and for training statistical machine translation models.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

The contents of the database are spontaneous dialogs in the tourism domain. Two volunteers talk to each other using a telephone plataform. Therefore, all the iteration is by voice. One speaker acts as the agent while the other plays the role of client. They are provided with an scenario (goal) and some tools related with the task.

Before translation, the transcriptions were cleaned up. However, the spontaneous style was preserved. The translation tried to be literal but correct in the target language. It was done by professional translators

4.3. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

There were no specific speaker selection strategy but the accent and gender of the speakers was registered. The total number of speakers recording either Spanish or Catalan dialogs was 122.

Spanish Festival HTS models male speech

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

This database contains Hidden Markov Models (HMM) for their use in Festival TTS using the HTS toolkit. The HMM models are trained with 10h of speech from the TC-STAR Spanish Baseline Male Speech Database that is also available in Meta-Share. The speech data can be found in the TC-STAR Male Baseline Voice Spanish Database that are also available in the ELRA catalogue and Meta-Share. The Spanish Festival HTS models male speech database was created within the scope of the METANET4U project funded by the European Commission..

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

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, install), statistical models to be used by HTS to generate the synthetic speech and scheme scripts to link the Festival TTS with HTS.

3.2. Encoding

Documentation is encoded in plain text.

All the data files are compatible with the HTS-2.2 and Festival 2.1 (available version in Nov. 2011).

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The resource consist of a festival-compatible voice folder of about 3MB is provided.

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This resource includes statistical models for producing synthetic speech in the framework of statistic parametric synthesis. The Hidden Markov Models (HMM) are compatible with the HTS-2.2 toolkit. This toolkit generates speech based on statistical models. It requires that the input text is represented by phonetic-prosodic labels. These labels can be generated using the Festival 2.1 system. The hts-engine is integrated in Festival so that it can be used as a stand-alone TTS system.

4.2. Training data

The voices were trained from speech data produced in the TC-STAR project to create high-quality synthetic voices. The specifications of the corpus content, recording procedure, labelling, speaker selection is described in the TC-STAR deliverable D9 (www.tcstar.org).

The training data contain more than 10 hours of speech recorded in a sound-proof room with professional equipment. The speaker is a professional speaker carefully selected from five candidates. The corpus consist on transcriptions of speech, generic text material (newspapers, novels) and specific data to increase the general coverage (sentences for triphones, questions) and for specific

applications (cities and villages, IVR commands, URLs, spellings, etc.).

The orthographic transcription was supervised by an expert. Although in the original database the phonetic transcription was also reviewed by an expert, in this resource, the phonetic transcription has been produced by the Festival front-end which has not been updated. This was needed in order to make voices compatible with the TTS

Spanish Festival HTS models female speech

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

This database contains Hidden Markov Models (HMM) for their use in Festival TTS using the HTS toolkit. The HMM models are trained with 10h of speech from the TC-STAR Spanish Baseline Female Speech Database that is also available in Meta-Share. The speech data can be found in the TC-STAR female Baseline Voice Spanish Database that are also available in the ELRA catalogue and Meta-Share. The Spanish Festival HTS models female speech database was created within the scope of the METANET4U project funded by the European Commission.

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 free, license-based, for research purposes 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, install), statistical models to be used by HTS to generate the synthetic speech and scheme scripts to link the Festival TTS with HTS.

3.2. Encoding

Documentation is encoded in plain text.

All the data files are compatible with the HTS-2.2 and Festival 2.1 (available version in Nov. 2011).

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The resource consist of a festival-compatible voice folder of about 3MB is provided.

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This resource includes statistical models for producing synthetic speech in the framework of statistic parametric synthesis. The Hidden Markov Models (HMM) are compatible with the HTS-2.2 toolkit. This toolkit generates speech based on statistical models. It requires that the input text is represented by phonetic-prosodic labels. These labels can be generated using the Festival 2.1 system. The hts-engine is integrated in Festival so that it can be used as a stand-alone TTS system.

4.2. Training data

The voices were trained from speech data produced in the TC-STAR project to create high-quality synthetic voices. The specifications of

the corpus content, recording procedure, labelling, speaker selection is described in the TC-STAR deliverable D9 (www.tcstar.org).

The training data contain more than 10 hours of speech recorded in a sound-proof room with professional equipment. The speaker is a professional speaker carefully selected from five candidates. The corpus consist on transcriptions of speech, generic text material (newspapers, novels) and specific data to increase the general coverage (sentences for triphones, questions) and for specific applications (cities and villages, IVR commands, URLs, spellings, etc.).

The orthographic transcription was supervised by an expert. Although in the original database the phonetic transcription was also reviewed by an expert, in this resource, the phonetic transcription has been produced by the Festival front-end which has not been updated. This was needed in order to make voices compatible with the TTS

Spanish Festival voice male

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

This database contains a unit-selection voice for their use in Festival TTS. The voice has been built from 10h of speech from the TC-STAR Spanish Baseline Male Speech Database. The speech data can be found in the TC-STAR Male Baseline Voice Spanish Database that are also available in the ELRA catalogue and Meta-Share. The Spanish Festival voice - male was created within the scope of the METANET4U project funded by the European Commission.

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, both for research purposes 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. The data is included in the folder structure which is typically used for Festival voices.

3.2. Encoding

Documentation is encoded in plain text.

Speech files are stored as rif files (wav), 16-bits, 16kHz. Each prompted utterance is stored within a separate file.

The voice includes the prompts, utterance files, clustering trees and other additional information needed for the data to be used in Festival.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The corpus contains about 3600 speech files/11 hours of recorded speech and requires about 1.5 GB for disk storage.

4. CONTENT INFORMATION

4.1 Type of the resource

This resource includes data, and indexing information for producing synthetic speech in the framework of unit-selection concatenative TTS. The voice is compatible with Festival 2.1, the most recent version available on Dec. 2011.

4.2. Training data

The voices were trained from speech data produced in the TC-STAR project to create high-quality synthetic voices. The training data and derivative files (utterances) are included as part of the resource.

The specifications of the corpus content, recording procedure, labelling, speaker selection is described in the TC-STAR deliverable D9 (www.tcstar.org).

The training data contain more than 10 hours of speech recorded in a sound-proof room with professional equipment. The speaker is a professional speaker carefully selected from five candidates. The corpus consist on transcriptions of speech, generic text material (newspapers, novels) and specific data to increase the general coverage (sentences for triphones, questions) and for specific applications (cities and villages, IVR commands, URLs, spellings, etc.).

The orthographic transcription was supervised by an expert. Although in the original database the phonetic transcription was also reviewed by an expert, in this resource, the phonetic transcription has been produced by the Festival front-end which has not been updated. This was needed in order to make voices compatible with the TTS.

Spanish Festival voice female

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

This database contains a unit-selection voice for their use in Festival TTS. The voice has been built from 10h of speech from the TC-STAR Spanish Baseline Female Speech Database. The speech data can be found in the TC-STAR Female Baseline Voice Spanish Database that are also available in the ELRA catalogue and Meta-Share. The Spanish Festival voice - female was created within the scope of the METANET4U project funded by the European Commission.

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, both for research purposes 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. The data is included in the folder structure which is typically used for Festival voices.

3.2. Encoding

Documentation is encoded in plain text.

Speech files are stored as rif files (wav), 16-bits, 16kHz. Each prompted utterance is stored within a separate file.

The voice includes the prompts, utterance files, clustering trees and other additional information needed to be used in Festival.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The corpus contains about 3600 speech files/11 hours of recorded speech and requires about 1.5 GB for disk storage.

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This resource includes data, and indexing information for producing synthetic speech in the framework of unit-selection concatenative TTS. The voice is compatible with Festival 2.1, the most recent version available on Dec. 2011.

4.2. Training data

The voices were trained from speech data produced in the TC-STAR project to create high-quality synthetic voices. The training data and derivative files (utterances) are included as part of the resource.

The specifications of the corpus content, recording procedure, labelling, speaker selection is described in the TC-STAR deliverable D9 (www.tcstar.org).

The training data contain more than 10 hours of speech recorded in a sound-proof room with professional equipment. The speaker is a professional speaker carefully selected from five candidates. The corpus consist on transcriptions of speech, generic text material (newspapers, novels) and specific data to increase the general coverage (sentences for triphones, questions) and for specific applications (cities and villages, IVR commands, URLs, spellings, etc.).

The orthographic transcription was supervised by an expert. Although in the original database the phonetic transcription was also reviewed by an expert, in this resource, the phonetic transcription has been produced by the Festival front-end which has not been updated. This was needed in order to make voices compatible with the TTS

SpeechDat-Car Catalan

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

The Catalan SpeechDat-Car database contains the in-car recordings of 300 speakers who uttered from around 120 read and spontaneous items. Each speaker recorded two sessions. Recordings have been made through 4 different channels, of which 4 were in-car microphones (1 close-talk microphone, 3 far-talk microphones). The 300 Catalan speakers were selected from 5 different dialectal regions, are balanced in gendre and age groups. The database was validated

by UVIGO. The Catalan-SpeechDat-Car 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 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, …), 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 plain text.

Four high quality audio channels were recorded in a car in a mobile platform and were stored as sequences of 16bit, 16 kHz uncompressed 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 70.000 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 Catalan for in-car applications. The database fulfills the Speechdat Car (www.speechdat.org) specifications except for the encoding of the signal files. Signal files encoding follows Speecon specifications (www.speechdat.org). No transmission though the GSM network was simultaneously recorded

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

The contents of the database are: application words, digits, numbers, currency amounts, dates, times, word spotting phrases, spellings, forenames, surnames, cities, company names, yes/no, phonetically rich sentences. 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 301 adult speakers who recorded two sessions. Speakers were selected from two different Catalan accents. Speakers are balanced in sex and gender and distributed in three age groups.

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, CEQ: no restrictions.
2. car in town traffic, CEQ: everything set to off or close
3. car in town traffic, CEQ: with noisy conditions
4. car moving at a low speed with rough road conditions, CEQ: everything set to off or close
5. car moving at a low speed with rough road conditions and CEQ: with noisy conditions
6. car moving at a high speed with good road conditions CEQ: no restrictions.

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

The recording platform is a ‘mobile’ recording platform (PltM) installed inside the car, recording multi-channel speech utterances in a high bandwidth mode (60-7000 Hz, 16 kHz sample frequency).
Multi-channel recordings are performed in the car. 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-talk microphone, a Lavalier microphone and a remote noise cancelling microphone with 2 Handsfree microphones placed at different locations in the car

The mobile recording platform in the car (PltM) uses a PC to drive the recording process. 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, one close signal and 2 far-talk signals). The positions for the far-talk microphones are:

- A_Column: at the ceiling of the car near the A-pillar
- Center: at the ceiling of the car over the mid-console (near the rear mirror)

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
- Multi-channel board recording API
- User Interface (MMI)

Prompt file management

Partner UPF

Corpus92 CORPUS

Corpus92 CORPUS

5 BASIC INFORMATION

5.1 Corpus composition

The corpus consists of a number of texts corresponding to Access to University examinations held on June 1992 in several Spanish universities (see Battaner et al., 2005). It contains about 350,000 words in 3 documents.

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

The corpus is represented in XML stand-off files (GraF format).

5.3 Character encoding

The characters are UTF8 encoded.

6 ADMINISTRATIVE INFORMATION

6.1 Contact person

Name: Sergi Torner,
Address: Roc Boronat, 138, 08018
Affiliation: Applied Linguistic Institute, Universitat Pompeu Fabra
Position: Researcher
Telephone: +34 935422369
Fax: +34 935422321
e-mail: sergi.torner@upf.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 free, license-based with restrictions, for research purposes and fee license-based for commercial purposes.

7 TECHNICAL INFORMATION

7.1 Directories and files

The archive that will be uploaded on the MetaShare platform will contain a single compressed file. Each document is divided in five different files that have a single base name; each file contains the following information:

1. basename+".anc" : header base document
2. basename+"-plain.txt": plain text document
3. basename+"-pos.xml": IULA tagging info
4. basename+"-seg.xml": document segmentation
5. basename+"-s.xml": text layout

7.2 Data structure of an entry

Each document of the corpus contains a number of sections holding textual elements: heads, paragraphs, notes, lists, tables and figures. Each of the above mentioned textual elements contains one or more sentences. Each sentence is segmented into tokens (equivalent to word or named entities) that can be formed by single words, proper names, numbers, dates, locutions, and other multiword structures.

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

The corpus contains about 350,000 tokens and needs about 113 MB for disk storage.

8 CONTENT INFORMATION

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

This corpus is a balanced monolingual annotated corpus.

8.2 The natural language(s) of the corpus

The language of the corpus is standard Spanish.

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

The text register represented into the corpus is scientific language as used by students accessing to the university.

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 corpus is annotated at paragraph, sentence, constituent group and word levels, providing part-of-speech information. The following example (document g00159) shows the set of files with all tags and attributes used in the annotation.

m00159.anc (header base document):

```
<cesHeader xmlns="http://www.xces.org/ns/GrAF/1.0/"creator="UPF"
date.created="2011/11/7 10:59:9" version="1.0.4">
<fileDesc>
    <titleStmt>
        <title>Corpus92 Ciencias</title>
    </titleStmt>
    <extent wordCount=">108586"></extent>
    <sourceDesc>
        <title>Corpus92 Ciencias</title>
        <publisher type="org">UPF</publisher>
        <edition></edition>
        <pubDate>01/06/1992</pubDate>
        <pubPlace>Espa&ntilde;a</pubPlace>
    </sourceDesc>
</fileDesc>
<profileDesc>
    <langUsage>
    <language iso639="spa">Spanish (Castilian)</language>
    <textClass catRef="">
    <keywords>
        <keyterm>gsl</keyterm>
        <keyterm>parea</keyterm>
    </keywords>
        <domain></domain>
        <subdomain></subdomain>
        <subject></subject>
        <audience></audience>
        <medium></medium>
    </textClass>
    <primaryData loc="g00159" medium="xml"></primaryData>
    <annotations>
        <annotation ann.loc="g00159-plain.txt" type="original">Original
Document</annotation>
        <annotation ann.loc="g00159-seg.xml" type="s">Base
segmentation</annotation>
        <annotation ann.loc="g00159-pos.xml" type="Iula Tagging">IULA
Part of Speech tags annotations</annotation>
        <annotation ann.loc="g00159-s.xml" type="layout">Text
layout</annotation>
    </annotations>
</profileDesc>
<revisionDesc>
    <change>
        <changeDate value="2011/11/7"></changeDate>
        <respName>IULA</respName>
```

```
            <item>item</item>
        </change>
        <change>
            <changeDate value="04/09/99"</changeDate>
            <respName>torner</respName>
            <item>marcatge nivel 2</item>
        </change>
</revisionDesc>
</cesHeader>
```

m00159-plain.txt (plain text document):

Biología

Las enzimas son unas moléculas que su función principal es la de bajar la energía de las reacciones químicas (energía de activación).

…

m00159-pos.xml (IULA tagging info):

```
<?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="tok" occurs="108585"/>
            </tagsDecl>
            <dependencies>
                <dependsOn loc="m00159-seg.xml"/>
            </dependencies>
            <annotationSets>
                <annotationSet name="xces"
type="http://www.xces.org/schema/2003"/>
            </annotationSets>
        </header>s
<node xml:id="iula-n0">
        <link targets="seg-r0"/>
</node>
<a label="TOK" ref="iula-n0" as="xces">
        <fs>
            <f name="base" value="biologí"/
            <f name="msd" value="NCFS-"/>
        </fs>
</a>
<node xml:id="iula-n1">
        <link targets="seg-r1"/>s
</node>
<a label="TOK" ref="iula-n1" as="xces">
        <fs>
            <f name="base" value="="/>
            <f name="msd" value="DELS"/>
        </fs>
</a>
<node xml:id="iula-n2">
        <link targets="seg-r2"/>
```

```
</node>
<a label="TOK" ref="iula-n2" as="xces">
    <fs>
        <f name="base" value="el"/>
        <f name="msd" value="D..FP.--"/>
    </fs>
</a>
<node xml:id="iula-n3">
    <link targets="seg-r3"/>
</node>
<a label="TOK" ref="iula-n3" as="xces">
    <fs>
        <f name="base" value="enzima"/>
        <f name="msd" value="NC.P-"/>;
    </fs>
</a>
...
```

m00159-seg.xml (document segmentation):

```
<?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 9"/>
<region xml:id="seg-r1" anchors="10 10"/>
<region xml:id="seg-r2" anchors="12 15"/>
<region xml:id="seg-r3" anchors="16 23"/>
<region xml:id="seg-r4" anchors="24 27"/>
<region xml:id="seg-r5" anchors="28 32"/>
<region xml:id="seg-r6" anchors="33 43"/>
<region xml:id="seg-r7" anchors="44 47"/>
<region xml:id="seg-r8" anchors="48 50"/>
<region xml:id="seg-r9" anchors="51 59"/>
...
```

m00159-s.xml (text layout):

```
<?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="6097"/>
<tagUsage gi="p" occurs="3445"/>
<tagUsage gi="div1" occurs="282"/>
<tagUsage gi="head" occurs="373"/>
<tagUsage gi="name" occurs="852"/>
<tagUsage gi="item" occurs="839"/>
<tagUsage gi="na" occurs="3308"/>
<tagUsage gi="foreign" occurs="7"/>
<tagUsage gi="loc" occurs="861"/>
```

```
<tagUsage gi="list" occurs="230"/>
<tagUsage gi="abbr" occurs="23"/>
<tagUsage gi="num" occurs="1050"/>
<tagUsage gi="gap" occurs="20"/>
<tagUsage gi="table" occurs="2"/>
<tagUsage gi="hi" occurs="34"/>
<tagUsage gi="gap" occurs="20"/>
</tagsDecl>
<annotationSets><annotationSet name="xces"
type="http://www.xces.org/schema/2003"/> </annotationSets>
</header>
<region xml:id="div1-r1" anchors="0 2569"/>
<node xml:id="div1-n1">
      <link targets="div1-r1"/>
</node>
<a label="div" ref="div1-n1" as="xces">
      <fs>
            <f name="id" value="div-1"/>
      </fs>
</a>
<region xml:id="head-r1" anchors="0 10"/>
<node xml:id="head-n1">
      <link targets="head-r1"/>
</node>
<a label="head" ref="head-n1" as="xces">
      <fs>
            <f name="id" value="div1-1head1"/>
      </fs>
</a>
<region xml:id="p-r1" anchors="12 151"/>
<node xml:id="p-n1">
      <link targets="p-r1"/>
</node>
<a label="p" ref="p-n1" as="xces">
      <fs>
            <f name="id" value="div11-p1"/>
      </fs>
</a>
...
```

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

The corpus contains morpho-syntactic information (MSD) which has been assigned automatically with our locally trained TreeTagger.

4.4.3Alignment 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)

The MSDs follows the IULA tagset (Morel et al., 1998) adapted to PAROLE/MULTEXT specifications (http://aune.lpl.univ-aix.fr/projects/multext/LEX/LEX.LangSpec.sp.html)

6.13 Intended application of the corpus

The corpus can be used for building statistical language models for this language usage.

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

The annotations are highly reliable. The paragraph and sentence mark-up has been fully syntactically validated as well as the MSD tagging.

5 RELEVANT REFERENCES AND OTHER INFORMATION

Vivaldi Palatresi, Jorge (2009). "Corpus and exploitation tool: IULACT and bwanaNet" in Cantos Gómez, Pascual; Sánchez Pérez, Aquilino (ed.) A survey on corpus-based research = Panorama de investigaciones basadas en corpus [Proceedings of the I Congreso Internacional de Lingüística de Corpus (CICL-09), 7-9 May 2009]. Murcia: Asociación Española de Lingüística del Corpus. Pp. 224-239.

Cabré, M. T.; Bach, C.; Vivaldi, J. (2006). "10 anys del Corpus de l'IULA". Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra

Battaner, P.; Torner, S. (2005). "El Corpus PAAU 1992 Estudios descriptivos, textos y vocabulario". Papers de l'IULA, Sèrie Monografiess n. 9. Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra.

Cabré, M. Teresa; Bach, Carme (2004). "El corpus tècnic del IULA: corpus textual especializado plurilingüe" in Panace@ - Boletín de Medicina y Traducción 5(16). [s.l]: MedTrad. Pp. 173-176.

Morel, J.; Torner, S.; Vivaldi, J., de Yzaguirre, L.; Cabré, M.T. (1998). "El corpus de l'IULA: etiquetaris". Papers de l'IULA, Sèrie Informes n. 18. Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra.

GENOME CATALAN CORPUS

GENOME CATALAN CORPUS

9 BASIC INFORMATION

9.1 Corpus composition

The corpus consists of a number of specialized texts of Genome domain. This is LSP corpus has been created with articles from specialized publications, PhD theses, etc. It contains about 950 K words in 133 documents.

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

The corpus is represented in XML stand-off files (GraF format).

9.3 Character encoding

The characters are UTF8 encoded.

10 ADMINISTRATIVE INFORMATION

10.1 Contact person

Name: Jorge Vivaldi,
Address: Roc Boronat, 138, 08018
Affiliation: Applied Linguistic Institute, Universitat Pompeu Fabra
Position: Researcher
Telephone: +34 935422332
Fax: +34 935422321
e-mail: jorge.vivaldi@upf.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 free, license-based with restrictions, for research purposes and fee license-based for commercial purposes.

11 TECHNICAL INFORMATION

11.1 Directories and files

The archive that will be uploaded on the MetaShare platform will contain a single compressed file. Each document is divided in five different files that have a single base name; each file contains the following information:

6. basename+".anc" : header base document
7. basename+"-plain.txt": plain text document
8. basename+"-pos.xml": IULA tagging info
9. basename+"-seg.xml": document segmentation
10. basename+"-s.xml": text layout

11.2 Data structure of an entry

Each document of the corpus contains a number of sections holding textual elements: heads, paragraphs, notes, lists, tables and figures. Each of the above mentioned textual elements contains one or more sentences. Each sentence is segmented into tokens (equivalent to word or named entities) that can be formed by single words, proper names, numbers, dates, locutions, and other multiword structures.

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

The corpus contains 950,000 tokens and needs about 315 MB for disk storage.

12 CONTENT INFORMATION

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

This corpus is a balanced monolingual annotated corpus.

12.2 The natural language(s) of the corpus

The language of the corpus is standard Catalan.

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

The text register represented into the corpus is specialized language as used in scientific literature.

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 corpus is annotated at paragraph, sentence, constituent group and word levels, providing part-of-speech information. The following

example (document m00520) shows the set of files with all tags and attributes used in the annotation.

m00520.anc (header base document):

```
<cesHeader xmlns="http://www.xces.org/ns/GrAF/1.0/"creator="UPF"
date.created="2011/11/4 15:43:55" version="1.0.4">
<fileDesc>
    <titleStmt>
        <title>La prehist&ograve;ria i la gen&egrave;tica d'Europa</title>
    </titleStmt>
    <extent wordCount=">5737"></extent>
    <sourceDesc>
        <title>La prehist&ograve;ria i la gen&egrave;tica d'Europa</title>
        <publisher type="org"> Universitat de Val&egrave;ncia </publisher>
        <edition>1</edition>
        <pubDate>1/11/1995</pubDate>
        <pubPlace>Val&egrave;ncia</pubPlace>
    </sourceDesc>
</fileDesc>
<profileDesc>
    <textClass catRef="">
    <keywords>
        <keyterm>mcb</keyterm>
        <keyterm>parea</keyterm>
    </keywords>
        <domain></domain>
        <subdomain></subdomain>
        <subject></subject>
        <audience></audience>
        <medium></medium>
    </textClass>
    <primaryData loc="m00520" medium="xml"></primaryData>
    <annotations>
        <annotation ann.loc="m00520-plain.txt" type="original">Original
Document</annotation>
        <annotation ann.loc="m00520-seg.xml" type="s">Base
segmentation</annotation>
        <annotation ann.loc="m00520-pos.xml" type="Iula Tagging">IULA Part of
Speech tags annotations</annotation>
        <annotation ann.loc="m00520-s.xml" type="layout">Text layout</annotation>
    </annotations>
</profileDesc>
<revisionDesc>
    <change>
        <changeDate value="2011/11/4"</changeDate>
        <respName>IULA</respName>
        <item>item</item>
    </change>
    <change>
        <changeDate value="07/30/02"</changeDate>
        <respName>giraldo</respName>
        <item>marcatge nivel 2</item>
    </change>
</revisionDesc>
</cesHeader>
```

m00520-plain.txt (plain text document):

LA PREHISTÒRIA I LA GENÉTICA D' EUROPA

La història i els orígen

Si parlem de recuperar el passat, reconstruir-lo, conéixer-lo, usualment pensem que hem de recòrrer a la història.
....

m00520-pos.xml (IULA tagging info):

```
<?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="tok" occurs="5736"/>
        </tagsDecl>
        <dependencies>
            <dependsOn loc="m00520-seg.xml"/>r
        </dependencies>
        <annotationSets>
            <annotationSet name="xces"
type="http://www.xces.org/schema/2003"/>
        </annotationSets>
    </header>
<node xml:id="iula-n0">
    <link targets="seg-r0"/>
</node>
<a label="TOK" ref="iula-n0" as="xces">
    <fs>
        <f name="base" value="el"/>
        <f name="msd" value="D..FS.--"/>
    </fs>
</a>
<node xml:id="iula-n1">
    <link targets="seg-r1"/>
</node>
<a label="TOK" ref="iula-n1" as="xces">
    <fs>
        <f name="base" value="prehistòria"/>
        <f name="msd" value="NCFS-"/>
    </fs>
</a>
<node xml:id="iula-n2">
    <link targets="seg-r2"/>
</node>
<a label="TOK" ref="iula-n2" as="xces">
    <fs>
        <f name="base" value="i"/>
        <f name="msd" value="CS"/>
    </fs>
</a>
...
```

m00520-seg.xml (document segmentation):

```
<?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 2"/>
<region xml:id="seg-r1" anchors="3 15"/>
<region xml:id="seg-r2" anchors="16 17"/>
<region xml:id="seg-r3" anchors="18 20"/>
<region xml:id="seg-r4" anchors="21 30"/>
<region xml:id="seg-r5" anchors="31 32"/>
<region xml:id="seg-r6" anchors="32 33"/>
...
```

m00520-s.xml (text layout):

```
<?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="165"/>
                <tagUsage gi="p" occurs="40"/>
                <tagUsage gi="div1" occurs="1"/>
                <tagUsage gi="head" occurs="8"/>
                <tagUsage gi="name" occurs="78"/>
                <tagUsage gi="item" occurs="7"/>
                <tagUsage gi="foreign" occurs="3"/>
                <tagUsage gi="date" occurs="1"/>
                <tagUsage gi="loc" occurs="70"/>
                <tagUsage gi="list" occurs="1"/>
                <tagUsage gi="abbr" occurs="3"/>
                <tagUsage gi="num" occurs="50"/>
                <tagUsage gi="hi" occurs="14"/>
        </tagsDecl>
        <annotationSets>
        <annotationSet name="xces"
        type="http://www.xces.org/schema/2003"/> </annotationSets>
</header>
<region xml:id="div1-r1" anchors="0 30682"/>
<node xml:id="div1-n1">
    <link targets="div1-r1"/>
</node>
<a label="div" ref="div1-n1" as="xces">
    <fs>
        <f name="id" value="div-1"/>
    </fs>
</a>
<region xml:id="head-r1" anchors="0 40"/>
<node xml:id="head-n1">
```

```
        <link targets="head-r1"/>
</node>
<a label="head" ref="head-n1" as="xces">
        <fs>
                <f name="id" value="div1-1head1"/>
        </fs>
</a>
<region xml:id="head-r2" anchors="43 70"/>
<node xml:id="head-n2">
        <link targets="head-r2"/>
</node>
<a label="head" ref="head-n2" as="xces">
        <fs>
                <f name="id" value="div1-1head2"/>
        </fs>
</a>
...
```

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

The corpus contains morpho-syntactic information (MSD) which has been assigned automatically with our locally trained TreeTagger.

4.4.3Alignment 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)

The MSDs follows the IULA tagset (Morel et al., 1998) adapted to PAROLE/MULTEXT specifications (http://aune.lpl.univ-aix.fr/projects/multext/LEX/LEX.LangSpec.sp.html)

6.15 Intended application of the corpus

The corpus can be used for building robust statistical language models for this language and domain.

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

The annotations are reliable. The paragraph and sentence mark-up has been fully syntactically validated. The MSD tagging accuracy is at least 95%.

5 RELEVANT REFERENCES AND OTHER INFORMATION

Vivaldi Palatresi, Jorge (2009). "Corpus and exploitation tool: IULACT and bwanaNet" in Cantos Gómez, Pascual; Sánchez Pérez, Aquilino (ed.) A survey on corpus-based research

= Panorama de investigaciones basadas en corpus [Proceedings of the I Congreso Internacional de Lingüística de Corpus (CICL-09), 7-9 May 2009]. Murcia: Asociación Española de Lingüística del Corpus. Pp. 224-239.

Cabré, M. T.; Bach, C.; Vivaldi, J. (2006). 10 anys del Corpus de l'IULA. Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra

Cabré, M. Teresa; Bach, Carme (2004). "El corpus tècnic del IULA: corpus textual especializado plurilingüe" in Panace@ - Boletín de Medicina y Traducción 5(16). [s.l]: MedTrad. Pp. 173-176.

Morel, J.; Torner, S.; Vivaldi, J., de Yzaguirre, L.; Cabré, M.T. (1998). "El corpus de l'IULA: etiquetaris". Papers de l'IULA, Sèrie Informes n. 18. Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra.

GENOME SPANISH CORPUS

GENOME SPANISH CORPUS

13 BASIC INFORMATION

13.1 Corpus composition

The corpus consists of a number of specialized texts of Genome domain. This is LSP corpus has been created with articles from specialized publications, PhD theses, etc. It contains about 1,650 K words in 276 documents.

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

The corpus is represented in XML stand-off files (GraF format).

13.3 Character encoding

The characters are UTF8 encoded.

14 ADMINISTRATIVE INFORMATION

14.1 Contact person

Name: Jorge Vivaldi,
Address: Roc Boronat, 138, 08018
Affiliation: Applied Linguistic Institute, Universitat Pompeu Fabra
Position: Researcher

Telephone: +34 935422332
Fax: +34 935422321
e-mail: jorge.vivaldi@upf.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 free, license-based, for research purposes and fee license-based for commercial purposes.

15 TECHNICAL INFORMATION

15.1 Directories and files

The archive that will be uploaded on the MetaShare platform will contain a single compressed file. Each document is divided in five different files that have a single base name; each file contains the following information:

11. basename+".anc" : header base document
12. basename+"-plain.txt": plain text document
13. basename+"-pos.xml": IULA tagging info
14. basename+"-seg.xml": document segmentation
15. basename+"-s.xml": text layout

15.2 Data structure of an entry

Each document of the corpus contains a number of sections holding textual elements: heads, paragraphs, notes, lists, tables and figures. Each of the above mentioned textual elements contains one or more sentences. Each sentence is segmented into tokens (equivalent to word or named entities) that can be formed by single words, proper names, numbers, dates, locutions, and other multiword structures.

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

The corpus contains 1,650,000 tokens and needs about 511 MB for disk storage.

16 CONTENT INFORMATION

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

This corpus is a balanced monolingual annotated corpus.

16.2 The natural language(s) of the corpus

The language of the corpus is standard Spanish.

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

The text register represented into the corpus is specialized language as used in scientific literature.

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 corpus is annotated at paragraph, sentence, constituent group and word levels, providing part-of-speech information. The following example (document m00203) shows the set of files with all tags and attributes used in the annotation.

m00203.anc (header base document):

```
<cesHeader xmlns="http://www.xces.org/ns/GrAF/1.0/"creator="UPF"
date.created="2011/11/4 15:43:48" version="1.0.4">
<fileDesc>
      <titleStmt>
            <title>GenÃ©tica del comportamiento</title
      </titleStmt>
      <extent wordCount=">4984"></extent>
      <sourceDesc>
            <title>Genética del comportamiento</title
            <publisher type="org"> Investigación y ciencia </publisher
            <edition>6</edition>
            <pubDate>6/1/1995</pubDate>
            <pubPlace>Barcelona</pubPlace>
      </sourceDesc>
</fileDesc>
<profileDesc>
      <textClass catRef="">
      <keywords>
            <keyterm>mcb</keyterm>
            <keyterm>parea</keyterm>
      </keywords>
            <domain></domain>
            <subdomain></subdomain>
            <subject></subject>
            <audience></audience>
            <medium></medium>
      </textClass>
      <primaryData loc="m00203" medium="xml"></primaryData>
      <annotations>
            <annotation ann.loc="m00203-plain.txt" type="original">Original
Document</annotation>
```

```
            <annotation ann.loc="m00203-seg.xml" type="s">Base
segmentation</annotation>
            <annotation ann.loc="m00203-pos.xml" type="Iula Tagging">IULA Part of
Speech tags annotations</annotation>
            <annotation ann.loc="m00203-s.xml" type="layout">Text layout</annotation>
      </annotations>
</profileDesc>
<revisionDesc>
      <change>
            <changeDate value="2011/11/4"</changeDate>
            <respName>IULA</respName>
            <item>item</item>
      </change>
      <change>
            <changeDate value="03/15/01"</changeDate>
            <respName>suarez</respName>
            <item>marcatge nivel 2</item>
      </change>
</revisionDesc>
</cesHeader>
```

m00203-plain.txt (plain text document):

Genética del comportamiento

El estudio del cortejo y la có³pula en la mosca de la fruta permite comprender la influencia de los genes en el despliegue de comportamientos complejos.

Las pistas iniciales sobre el funcionamiento de los mecanismos hereditarios se obtuvieron en los primeros quince años de nuestro siglo.
…

m00203-pos.xml (IULA tagging info):

```
<?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="tok" occurs="4983"/>
            </tagsDecl>
            <dependencies>
                  <dependsOn loc="m00203-seg.xml"/>
            </dependencies>
            <annotationSets>
                  <annotationSet name="xces"
type="http://www.xces.org/schema/2003"/>
            </annotationSets>
      </header>
<node xml:id="iula-n0">
      <link targets="seg-r0"/>
</node>
<a label="TOK" ref="iula-n0" as="xces">
      <fs>
            <f name="base" value="genético"/>
            <f name="msd" value="AF.FS-"/>
```

```
.       </fs>
</a>
<node xml:id="iula-n1">
        <link targets="seg-r1"/>
</node>
<a label="TOK" ref="iula-n1" as="xces">
        <fs>
                <f name="base" value="de"/>e
                <f name="msd" value="SP^M"/>o
        </fs>
</a>
<node xml:id="iula-n2">
        <link targets="seg-r2"/>
</node>
<a label="PGR" ref="iula-n2" as="xces">
        <fs>
                <f name="base" value="l"/>e
                <f name="msd" value="D..MP.--"/>
        </fs>
</a>
<node xml:id="iula-n3">
        <link targets="seg-r3"/>
</node>
<a label="TOK" ref="iula-n3" as="xces">
        <fs>
                <f name="base" value="comportamiento"/>
                <f name="msd" value="NCMS-"/>
        </fs>
</a>
...
```

m00203-seg.xml (document segmentation):

```
<?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 9"/>
<region xml:id="seg-r1" anchors="10 12"/>
<region xml:id="seg-r2" anchors="12 13"/>
<region xml:id="seg-r3" anchors="14 28"/>
<region xml:id="seg-r4" anchors="30 32"/>
<region xml:id="seg-r5" anchors="33 40"/>
<region xml:id="seg-r6" anchors="41 43"/>
<region xml:id="seg-r7" anchors="43 44"/>
<region xml:id="seg-r8" anchors="45 52"/>
<region xml:id="seg-r9" anchors="53 54"/>
<region xml:id="seg-r10" anchors="55 57"/>
<region xml:id="seg-r11" anchors="58 65"/>
<region xml:id="seg-r12" anchors="66 68"/>
<region xml:id="seg-r13" anchors="69 71"/>
<region xml:id="seg-r14" anchors="72 77"/>
...
```

m00203-s.xml (text layout):

```
<?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="175"/>
<tagUsage gi="p" occurs="50"/>
<tagUsage gi="div1" occurs="1"/>
<tagUsage gi="head" occurs="6"/>
<tagUsage gi="name" occurs="82"/>
<tagUsage gi="item" occurs="6"/>
<tagUsage gi="na" occurs="8"/>
<tagUsage gi="foreign" occurs="10"/>
<tagUsage gi="loc" occurs="16"/>
<tagUsage gi="list" occurs="1"/>
<tagUsage gi="abbr" occurs="2"/>
<tagUsage gi="num" occurs="38"/>
<tagUsage gi="hi" occurs="30"/>
</tagsDecl>
<annotationSets><annotationSet name="xces"
type="http://www.xces.org/schema/2003"/> </annotationSets>
</header>
<region xml:id="div1-r1" anchors="0 27840"/>
<node xml:id="div1-n1">
    <link targets="div1-r1"/>
</node>
<a label="div" ref="div1-n1" as="xces">
    <fs>
        <f name="id" value="div-1"/>
    </fs>
</a>
<region xml:id="head-r1" anchors="0 28"/>
<node xml:id="head-n1">
    <link targets="head-r1"/>
</node>
<a label="head" ref="head-n1" as="xces">
    <fs>
        <f name="id" value="div1-1head1"/>
    </fs>
</a>
<region xml:id="p-r1" anchors="30 182"/>
<node xml:id="p-n1">
    <link targets="p-r1"/>
</node>
<a label="p" ref="p-n1" as="xces">
    <fs>
        <f name="id" value="div11-p1"/>
    </fs>
</a>
...
```

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

The corpus contains morpho-syntactic information (MSD) which has been assigned automatically with our locally trained TreeTagger.

4.4.3Alignment 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)

The MSDs follows the IULA tagset (Morel et al., 1998) adapted to PAROLE/MULTEXT specifications (http://aune.lpl.univ-aix.fr/projects/multext/LEX/LEX.LangSpec.sp.html)

6.17 Intended application of the corpus

The corpus can be used for building robust statistical language models for this language and domain.

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

The annotations are reliable. The paragraph and sentence mark-up has been fully syntactically validated. The MSD tagging accuracy is at least 95%.

5 RELEVANT REFERENCES AND OTHER INFORMATION

Vivaldi Palatresi, Jorge (2009). "Corpus and exploitation tool: IULACT and bwanaNet" in Cantos Gómez, Pascual; Sánchez Pérez, Aquilino (ed.) A survey on corpus-based research = Panorama de investigaciones basadas en corpus [Proceedings of the I Congreso Internacional de Lingüística de Corpus (CICL-09), 7-9 May 2009]. Murcia: Asociación Española de Lingüística del Corpus. Pp. 224-239.

Cabré, M. T.; Bach, C.; Vivaldi, J. (2006). 10 anys del Corpus de l'IULA. Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra

Cabré, M. Teresa; Bach, Carme (2004). "El corpus tècnic del IULA: corpus textual especializado plurilingüe" in Panace@ - Boletín de Medicina y Traducción 5(16). [s.l]: MedTrad. Pp. 173-176.

Morel, J.; Torner, S.; Vivaldi, J., de Yzaguirre, L.; Cabré, M.T. (1998). “El corpus de l’IULA: etiquetaris”. Papers de l’IULA, Sèrie Informes n. 18. Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra.

Apertium

Basque LMF Apertium Dictionary

This is the LMF version of the Basque Apertium dictionary. Monolingual dictionaries for Spanish, Catalan, Gallego and Euskera have been generated from the Apertium expanded lexicons of the es-ca (for both Spanish andCatalan) es-gl (for Galician) and eu-es (for Basque). 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.

BASIC INFORMATION

IdentificationInfo

- resourceName: Basque LMF Apertium Dictionary
- resourceShortName: LMF Apertium Eu
- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: Basque LMF Apertium Dictionary
- resourceShortName: LMF Apertium Eu
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF

        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: GPL
    - restrictionsOfUse: other
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)
            - CommunicationInfo:
                - address: Roc Boronat, 138
                - zipCode: 08018
                - city: Barcelona
                - country: Spain
                - telephoneNumber: +34 93 5421207
                - faxNumber: +34 93 5422321
                - email: jorge.vivaldi@upf.edu
                - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- FundingInfo:
    - ProjectInfo:
        - projectName: METANET4U
        - projectShortName: METANET4U
        - fundingType: ownFunds
        - fundingType: euFunds
        - fundingCountry: Spain
        - funder: UPF
        - funder: EU
        - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
    - lingualityType: monolingual
- LanguageInfo:
    - languageCoding: eus
    - languageName: Basque
- SizeInfo:
    - size: 6258

    - sizeUnitMultiplier: unit
    - sizeUnit: entries
- TextFormatInfo:
    - mime-type: text/xml
- CharacterEncodingInfo:
    - characterEncoding: UTF-8
- DomainInfo:
    - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Basque Apertium dictionary. Monolingual dictionaries for Spanish, Catalan, Gallego and Euskera have been generated from the Apertium expanded lexicons of the es-ca (for both Spanish andCatalan) es-gl (for Galician) and eu-es (for Basque). 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.
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: lexicon
- LexicalConceptualResourceCreationInfo:
    - originalSource: Apertium Monolingual dictionary as described in http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
    - creationMode: automatic
    - creationModeDetails: This lexicon was created with the Apertium2LMP.xsl
    - creationTool: Apertium2LMP.xsl
- LexicalConceptualResourceEncodingInfo:
    - encodingLevel: morphology
    - conformanceToStandardsBestPractice: LMF
    - linguisticInformation: lemma
    - linguisticInformation: semantics-CrossReferences
    - linguisticInformation: partOfSpeech

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
- publication: 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.

Basque-Spanish LMF Apertium Bilingual dictionary

This is the LMF version of the Apertium bilingual dictionary for Basque and Spanish languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.

BASIC INFORMATION

IdentificationInfo

- resourceName: Basque-Spanish LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium Eu-Es
- url: http://.es

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: Basque-Spanish LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium Eu-Es
- url: http://.es

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138

    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: GPL
    - restrictionsOfUse: other
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)

        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- FundingInfo:
    - ProjectInfo:
        - projectName: METANET4U
        - projectShortName: METANET4U
        - fundingType: ownFunds
        - fundingType: euFunds
        - fundingCountry: Spain
        - funder: UPF
        - funder: EU
        - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
  - lingualityType: bilingual
- LanguageInfo:
  - languageCoding: spa
  - languageName: Spanish
- LanguageInfo:
  - languageCoding: eus
  - languageName: Basque
- SizeInfo:
  - size: 16694
  - sizeUnitMultiplier: unit
  - sizeUnit: entries
- TextFormatInfo:
  - mime-type: text/xml
- CharacterEncodingInfo:
  - characterEncoding: UTF-8
- DomainInfo:
  - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Apertium bilingual dictionary for Basque and Spanish languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: lexicon
- LexicalConceptualResourceCreationInfo:

    - originalSource: Apertium Bilingual dictionary as described in http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
    - creationMode: automatic
    - creationModeDetails: This lexicon was created with the Apertium2LMP.xsl
    - creationTool: Apertium2LMP.xsl
- LexicalConceptualResourceEncodingInfo:
    - encodingLevel: morphology
    - conformanceToStandardsBestPractice: LMF
    - linguisticInformation: lemma
    - linguisticInformation: semantics-CrossReferences
    - linguisticInformation: partOfSpeech

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
- publication: 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.

English-Catalan LMF Apertium Bilingual dictionary

This is the LMF version of the Apertium bilingual dictionary for English and Catalan languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.

BASIC INFORMATION

IdentificationInfo

- resourceName: English-Catalan LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium En-Ca
- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: English-Catalan LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium En-Ca
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321

            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: GPL
    - restrictionsOfUse: other
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)
            - CommunicationInfo:
                - address: Roc Boronat, 138
                - zipCode: 08018
                - city: Barcelona
                - country: Spain
                - telephoneNumber: +34 93 5421207
                - faxNumber: +34 93 5422321
                - email: jorge.vivaldi@upf.edu
                - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- FundingInfo:
    - ProjectInfo:
        - projectName: METANET4U
        - projectShortName: METANET4U
        - fundingType: ownFunds

- fundingType: euFunds
- fundingCountry: Spain
- funder: UPF
- funder: EU
- url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
  - lingualityType: bilingual
- LanguageInfo:
  - languageCoding: eng
  - languageName: English
- LanguageInfo:
  - languageCoding: cat
  - languageName: Catalan
- SizeInfo:
  - size: 28145
  - sizeUnitMultiplier: unit
  - sizeUnit: entries
- TextFormatInfo:
  - mime-type: text/xml
- CharacterEncodingInfo:
  - characterEncoding: UTF-8
- DomainInfo:
  - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Apertium bilingual dictionary for English and Catalan languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and

target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.

- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: lexicon
- LexicalConceptualResourceCreationInfo:
  - originalSource: Apertium Bilingual dictionary as described in http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
  - creationMode: automatic
  - creationModeDetails: This lexicon was created with the Apertium2LMP.xsl
  - creationTool: Apertium2LMP.xsl
- LexicalConceptualResourceEncodingInfo:
  - encodingLevel: morphology
  - conformanceToStandardsBestPractice: LMF
  - linguisticInformation: lemma
  - linguisticInformation: semantics-CrossReferences
  - linguisticInformation: partOfSpeech

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
- publication: 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.

English-Galician LMF Apertium Bilingual dictionary

This is the LMF version of the Apertium bilingual dictionary for English and Galician languages. 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.

BASIC INFORMATION

IdentificationInfo

- resourceName: English-Galician LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium En-Gl
- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: English-Galician LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium En-Gl
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:

    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: GPL
    - restrictionsOfUse: other
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)
            - CommunicationInfo:
                - address: Roc Boronat, 138
                - zipCode: 08018
                - city: Barcelona
                - country: Spain
                - telephoneNumber: +34 93 5421207
                - faxNumber: +34 93 5422321
                - email: jorge.vivaldi@upf.edu
                - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018

        - city: Barcelona
        - country: Spain
        - telephoneNumber: +34 93 5421207
        - faxNumber: +34 93 5422321
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/
- FundingInfo:
    - ProjectInfo:
        - projectName: METANET4U
        - projectShortName: METANET4U
        - fundingType: ownFunds
        - fundingType: euFunds
        - fundingCountry: Spain
        - funder: UPF
        - funder: EU
        - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
    - lingualityType: bilingual
- LanguageInfo:
    - languageCoding: spa
    - languageName: Spanish
- LanguageInfo:
    - languageCoding: glg
    - languageName: Galician
- SizeInfo:
    - size: 29636
    - sizeUnitMultiplier: unit
    - sizeUnit: entries
- TextFormatInfo:
    - mime-type: text/xml
- CharacterEncodingInfo:
    - characterEncoding: UTF-8
- DomainInfo:
    - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Apertium bilingual dictionary for English and Galician languages. 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.
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: lexicon
- LexicalConceptualResourceCreationInfo:
    - originalSource: Apertium Bilingual dictionary as described in http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
    - creationMode: automatic
    - creationModeDetails: This lexicon was created with the Apertium2LMP.xsl
    - creationTool: Apertium2LMP.xsl
- LexicalConceptualResourceEncodingInfo:
    - encodingLevel: morphology
    - conformanceToStandardsBestPractice: LMF
    - linguisticInformation: lemma
    - linguisticInformation: semantics-CrossReferences
    - linguisticInformation: partOfSpeech

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
- publication: 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.

English-Spanish LMF Apertium Bilingual dictionary

This is the LMF version of the Apertium bilingual dictionary for English and Spanish languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.

BASIC INFORMATION

IdentificationInfo

- resourceName: English-Spanish LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium En-Es
- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: English-Spanish LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium En-Es
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada

        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: GPL
    - restrictionsOfUse: other
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)
            - CommunicationInfo:
                - address: Roc Boronat, 138
                - zipCode: 08018
                - city: Barcelona
                - country: Spain
                - telephoneNumber: +34 93 5421207
                - faxNumber: +34 93 5422321
                - email: jorge.vivaldi@upf.edu
                - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- FundingInfo:
    - ProjectInfo:
        - projectName: METANET4U
        - projectShortName: METANET4U
        - fundingType: ownFunds
        - fundingType: euFunds
        - fundingCountry: Spain
        - funder: UPF
        - funder: EU
        - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
    - lingualityType: bilingual
- LanguageInfo:
    - languageCoding: eng
    - languageName: English
- LanguageInfo:
    - languageCoding: spa

    - languageName: Spanish
- SizeInfo:
    - size: 26076
    - sizeUnitMultiplier: unit
    - sizeUnit: entries
- TextFormatInfo:
    - mime-type: text/xml
- CharacterEncodingInfo:
    - characterEncoding: UTF-8
- DomainInfo:
    - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Apertium bilingual dictionary for English and Spanish languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: lexicon
- LexicalConceptualResourceCreationInfo:
    - originalSource: Apertium Bilingual dictionary as described in http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
    - creationMode: automatic
    - creationModeDetails: This lexicon was created with the Apertium2LMP.xsl
    - creationTool: Apertium2LMP.xsl
- LexicalConceptualResourceEncodingInfo:
    - encodingLevel: morphology
    - conformanceToStandardsBestPractice: LMF
    - linguisticInformation: lemma

- linguisticInformation: semantics-CrossReferences
- linguisticInformation: partOfSpeech

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
- publication: 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.

French-Catalan LMF Apertium Bilingual dictionary

This is the LMF version of the Apertium bilingual dictionary for French and Catalan languags. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.

BASIC INFORMATION

IdentificationInfo

- resourceName: French-Catalan LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium Fr-Ca
- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: French-Catalan LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium Fr-Ca
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: GPL
    - restrictionsOfUse: other
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:

        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)
            - CommunicationInfo:
                - address: Roc Boronat, 138
                - zipCode: 08018
                - city: Barcelona
                - country: Spain
                - telephoneNumber: +34 93 5421207
                - faxNumber: +34 93 5422321
                - email: jorge.vivaldi@upf.edu
                - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- FundingInfo:
    - ProjectInfo:
        - projectName: METANET4U
        - projectShortName: METANET4U
        - fundingType: ownFunds
        - fundingType: euFunds
        - fundingCountry: Spain
        - funder: UPF
        - funder: EU
        - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
  - lingualityType: bilingual
- LanguageInfo:
  - languageCoding: fra
  - languageName: Catalan
- LanguageInfo:
  - languageCoding: cat
  - languageName: Catalan
- SizeInfo:
  - size: 7902
  - sizeUnitMultiplier: unit
  - sizeUnit: entries
- TextFormatInfo:
  - mime-type: text/xml
- CharacterEncodingInfo:
  - characterEncoding: UTF-8
- DomainInfo:
  - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Apertium bilingual dictionary for French and Catalan languags. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: lexicon
- LexicalConceptualResourceCreationInfo:
  - originalSource: Apertium Bilingual dictionary as described in http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
  - creationMode: automatic
  - creationModeDetails: This lexicon was created with the Apertium2LMP.xsl
  - creationTool: Apertium2LMP.xsl
- LexicalConceptualResourceEncodingInfo:
  - encodingLevel: morphology
  - conformanceToStandardsBestPractice: LMF
  - linguisticInformation: lemma
  - linguisticInformation: semantics-CrossReferences
  - linguisticInformation: partOfSpeech

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
- publication: 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.

French-Spanish LMF Apertium Bilingual dictionary

This is the LMF version of the Apertium bilingual dictionary for French and Spanish languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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

BASIC INFORMATION

IdentificationInfo

- resourceName: French-Spanish LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium Fr-Es

- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: French-Spanish LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium Fr-Es
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018

            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: GPL
    - restrictionsOfUse: other
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)
            - CommunicationInfo:
                - address: Roc Boronat, 138
                - zipCode: 08018
                - city: Barcelona
                - country: Spain
                - telephoneNumber: +34 93 5421207
                - faxNumber: +34 93 5422321
                - email: jorge.vivaldi@upf.edu
                - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- FundingInfo:

- ProjectInfo:
  - projectName: METANET4U
  - projectShortName: METANET4U
  - fundingType: ownFunds
  - fundingType: euFunds
  - fundingCountry: Spain
  - funder: UPF
  - funder: EU
  - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
  - lingualityType: monolingual
- LanguageInfo:
  - languageCoding: spa
  - languageName: Spanish
- LanguageInfo:
  - languageCoding: fra
  - languageName: French
- SizeInfo:
  - size: 19023
  - sizeUnitMultiplier: unit
  - sizeUnit: entries
- TextFormatInfo:
  - mime-type: text/xml
- CharacterEncodingInfo:
  - characterEncoding: UTF-8
- DomainInfo:
  - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Apertium bilingual dictionary for French and Spanish languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: lexicon
- LexicalConceptualResourceCreationInfo:
    - originalSource: Apertium Bilingual dictionary as described in http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
    - creationMode: automatic
    - creationModeDetails: This lexicon was created with the Apertium2LMP.xsl
    - creationTool: Apertium2LMP.xsl
- LexicalConceptualResourceEncodingInfo:
    - encodingLevel: morphology
    - conformanceToStandardsBestPractice: LMF
    - linguisticInformation: lemma
    - linguisticInformation: semantics-CrossReferences
    - linguisticInformation: partOfSpeech

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
- publication: 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.

Italian-Catalan LMF Apertium Bilingual dictionary

This is the LMF version of the Apertium bilingual dictionary for Italian and Catalan languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries

(LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.

BASIC INFORMATION

IdentificationInfo

- resourceName: Italian-Catalan LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium It-Ca
- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: Italian-Catalan LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium It-Ca
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu

        - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: GPL
    - restrictionsOfUse: other
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)
            - CommunicationInfo:
                - address: Roc Boronat, 138
                - zipCode: 08018
                - city: Barcelona
                - country: Spain
                - telephoneNumber: +34 93 5421207
                - faxNumber: +34 93 5422321
                - email: jorge.vivaldi@upf.edu
                - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
    - OrganizationInfo:

        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- FundingInfo:
    - ProjectInfo:
        - projectName: METANET4U
        - projectShortName: METANET4U
        - fundingType: ownFunds
        - fundingType: euFunds
        - fundingCountry: Spain
        - funder: UPF
        - funder: EU
        - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
    - lingualityType: bilingual
- LanguageInfo:
    - languageCoding: itl
    - languageName: Italian
- LanguageInfo:
    - languageCoding: cat
    - languageName: Catalan
- SizeInfo:
    - size: 7537
    - sizeUnitMultiplier: unit
    - sizeUnit: entries
- TextFormatInfo:
    - mime-type: text/xml
- CharacterEncodingInfo:

    - characterEncoding: UTF-8
- DomainInfo:
    - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Apertium bilingual dictionary for Italian and Catalan languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: lexicon
- LexicalConceptualResourceCreationInfo:
    - originalSource: Apertium Bilingual dictionary as described in http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
    - creationMode: automatic
    - creationModeDetails: This lexicon was created with the Apertium2LMP.xsl
    - creationTool: Apertium2LMP.xsl
- LexicalConceptualResourceEncodingInfo:
    - encodingLevel: morphology
    - conformanceToStandardsBestPractice: LMF
    - linguisticInformation: lemma
    - linguisticInformation: semantics-CrossReferences
    - linguisticInformation: partOfSpeech

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
- publication: 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.

Occitan-Catalan LMF Apertium Bilingual dictionary

This is the LMF version of the Apertium bilingual dictionary for Occitan and Catalan languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.

BASIC INFORMATION

IdentificationInfo

- resourceName: Occitan-Catalan LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium Oc-Ca
- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: Occitan-Catalan LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium Oc-Ca
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu

- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: GPL
    - restrictionsOfUse: other
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)
            - CommunicationInfo:
                - address: Roc Boronat, 138
                - zipCode: 08018
                - city: Barcelona

        - country: Spain
        - telephoneNumber: +34 93 5421207
        - faxNumber: +34 93 5422321
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- FundingInfo:
    - ProjectInfo:
        - projectName: METANET4U
        - projectShortName: METANET4U
        - fundingType: ownFunds
        - fundingType: euFunds
        - fundingCountry: Spain
        - funder: UPF
        - funder: EU
        - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:

    - lingualityType: bilingual
- LanguageInfo:
    - languageCoding: oci
    - languageName: Occitan
- LanguageInfo:
    - languageCoding: cat
    - languageName: Catalan
- SizeInfo:
    - size: 14915
    - sizeUnitMultiplier: unit
    - sizeUnit: entries
- TextFormatInfo:
    - mime-type: text/xml
- CharacterEncodingInfo:
    - characterEncoding: UTF-8
- DomainInfo:
    - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Apertium bilingual dictionary for Occitan and Catalan languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: lexicon
- LexicalConceptualResourceCreationInfo:
    - originalSource: Apertium Bilingual dictionary as described in http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
    - creationMode: automatic

    - creationModeDetails: This lexicon was created with the Apertium2LMP.xsl
    - creationTool: Apertium2LMP.xsl
- LexicalConceptualResourceEncodingInfo:
    - encodingLevel: morphology
    - conformanceToStandardsBestPractice: LMF
    - linguisticInformation: lemma
    - linguisticInformation: semantics-CrossReferences
    - linguisticInformation: partOfSpeech

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
- publication: 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.

Occitan-Spanish LMF Apertium Bilingual dictionary

This is the LMF version of the Apertium bilingual dictionary for Occitan and Spanish languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.

BASIC INFORMATION

IdentificationInfo

- resourceName: Occitan-Spanish LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium Oc-Es
- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: Occitan-Spanish LMF Apertium Bilingual dictionary

- resourceShortName: LMF Apertium Oc-Es
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:

- license: GPL
- restrictionsOfUse: other
- distributionAccessMedium: downloadable
- downloadLocation: http://www..com/
- distributor:
  - OrganizationInfo:
    - organizationName: Institut Universitari Lingüística Aplicada - UPF
    - organizationShortName: IULA - UPF
    - departmentName: Institut Universitari Lingüística Aplicada (UPF)
    - CommunicationInfo:
      - address: Roc Boronat, 138
      - zipCode: 08018
      - city: Barcelona
      - country: Spain
      - telephoneNumber: +34 93 5421207
      - faxNumber: +34 93 5422321
      - email: jorge.vivaldi@upf.edu
      - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
  - OrganizationInfo:
    - organizationName: Institut Universitari Lingüística Aplicada - UPF
    - organizationShortName: IULA - UPF
    - departmentName: Institut Universitari Lingüística Aplicada (UPF)
    - CommunicationInfo:
      - address: Roc Boronat, 138
      - zipCode: 08018
      - city: Barcelona
      - country: Spain
      - telephoneNumber: +34 93 5421207
      - faxNumber: +34 93 5422321
      - email: jorge.vivaldi@upf.edu
      - url: http://www.iula.upf.edu/
- FundingInfo:
  - ProjectInfo:
    - projectName: METANET4U
    - projectShortName: METANET4U
    - fundingType: ownFunds
    - fundingType: euFunds
    - fundingCountry: Spain
    - funder: UPF

        - funder: EU
        - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
    - lingualityType: bilingual
- LanguageInfo:
    - languageCoding: oci
    - languageName: Occitan
- LanguageInfo:
    - languageCoding: spa
    - languageName: Spanish
- SizeInfo:
    - size: 11606
    - sizeUnitMultiplier: unit
    - sizeUnit: entries
- TextFormatInfo:
    - mime-type: text/xml
- CharacterEncodingInfo:
    - characterEncoding: UTF-8
- DomainInfo:
    - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Apertium bilingual dictionary for Occitan and Spanish languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.

- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: lexicon
- LexicalConceptualResourceCreationInfo:
  - originalSource: Apertium Bilingual dictionary as described in http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
  - creationMode: automatic
  - creationModeDetails: This lexicon was created with the Apertium2LMP.xsl
  - creationTool: Apertium2LMP.xsl
- LexicalConceptualResourceEncodingInfo:
  - encodingLevel: morphology
  - conformanceToStandardsBestPractice: LMF
  - linguisticInformation: lemma
  - linguisticInformation: semantics-CrossReferences
  - linguisticInformation: partOfSpeech

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
- publication: 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.

Portuguese-Catalan LMF Apertium Bilingual dictionary

This is the LMF version of the Apertium bilingual dictionary for Portugues and Catalan languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.

BASIC INFORMATION

IdentificationInfo

- resourceName: Portuguese-Catalan LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium Pt-Ca
- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: Portuguese-Catalan LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium Pt-Ca
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:

    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: GPL
    - restrictionsOfUse: other
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)
            - CommunicationInfo:
                - address: Roc Boronat, 138
                - zipCode: 08018
                - city: Barcelona
                - country: Spain
                - telephoneNumber: +34 93 5421207
                - faxNumber: +34 93 5422321
                - email: jorge.vivaldi@upf.edu
                - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018

        - city: Barcelona
        - country: Spain
        - telephoneNumber: +34 93 5421207
        - faxNumber: +34 93 5422321
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/
- FundingInfo:
    - ProjectInfo:
        - projectName: METANET4U
        - projectShortName: METANET4U
        - fundingType: ownFunds
        - fundingType: euFunds
        - fundingCountry: Spain
        - funder: UPF
        - funder: EU
        - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
    - lingualityType: bilingual
- LanguageInfo:
    - languageCoding: por
    - languageName: Portuguese
- LanguageInfo:
    - languageCoding: cat
    - languageName: Catalan
- SizeInfo:
    - size: 6310
    - sizeUnitMultiplier: unit
    - sizeUnit: entries
- TextFormatInfo:
    - mime-type: text/xml
- CharacterEncodingInfo:
    - characterEncoding: UTF-8
- DomainInfo:
    - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Apertium bilingual dictionary for Portugues and Catalan languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: lexicon
- LexicalConceptualResourceCreationInfo:
    - originalSource: Apertium Bilingual dictionary as described in http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
    - creationMode: automatic
    - creationModeDetails: This lexicon was created with the Apertium2LMP.xsl
    - creationTool: Apertium2LMP.xsl
- LexicalConceptualResourceEncodingInfo:
    - encodingLevel: morphology
    - conformanceToStandardsBestPractice: LMF
    - linguisticInformation: lemma
    - linguisticInformation: semantics-CrossReferences
    - linguisticInformation: partOfSpeech

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
- publication: 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.

Portuguese-Galician LMF Apertium Bilingual dictionary

This is the LMF version of the Apertium bilingual ditionary for Portugues and Galician languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.

BASIC INFORMATION

IdentificationInfo

- resourceName: Portuguese-Galician LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium Pt-Gl
- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: Portuguese-Galician LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium Pt-Gl
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada

        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: GPL
    - restrictionsOfUse: other
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)
            - CommunicationInfo:
                - address: Roc Boronat, 138
                - zipCode: 08018
                - city: Barcelona
                - country: Spain
                - telephoneNumber: +34 93 5421207
                - faxNumber: +34 93 5422321
                - email: jorge.vivaldi@upf.edu
                - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- FundingInfo:
    - ProjectInfo:
        - projectName: METANET4U
        - projectShortName: METANET4U
        - fundingType: ownFunds
        - fundingType: euFunds
        - fundingCountry: Spain
        - funder: UPF
        - funder: EU
        - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
    - lingualityType: bilingual
- LanguageInfo:
    - languageCoding: por
    - languageName: Portuguese
- LanguageInfo:
    - languageCoding: glg

    - languageName: Galician
- SizeInfo:
    - size: 10387
    - sizeUnitMultiplier: unit
    - sizeUnit: entries
- TextFormatInfo:
    - mime-type: text/xml
- CharacterEncodingInfo:
    - characterEncoding: UTF-8
- DomainInfo:
    - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Apertium bilingual ditionary for Portugues and Galician languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: lexicon
- LexicalConceptualResourceCreationInfo:
    - originalSource: Apertium Bilingual dictionary as described in http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
    - creationMode: automatic
    - creationModeDetails: This lexicon was created with the Apertium2LMP.xsl
    - creationTool: Apertium2LMP.xsl
- LexicalConceptualResourceEncodingInfo:
    - encodingLevel: morphology
    - conformanceToStandardsBestPractice: LMF
    - linguisticInformation: lemma

- linguisticInformation: semantics-CrossReferences
- linguisticInformation: partOfSpeech

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
- publication: 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.

Spanish-Asturian LMF Apertium Bilingual dictionary

This is the LMF version of the Apertium bilingual dictionary for Spanish and Asturian languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.

BASIC INFORMATION

IdentificationInfo

- resourceName: Spanish-Asturian LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium Es-Ast
- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: Spanish-Asturian LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium Es-Ast
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi

- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: GPL
    - restrictionsOfUse: other
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:

        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- FundingInfo:
    - ProjectInfo:
        - projectName: METANET4U
        - projectShortName: METANET4U
        - fundingType: ownFunds
        - fundingType: euFunds
        - fundingCountry: Spain
        - funder: UPF
        - funder: EU
        - url: http://metanet4u.eu/

ValidationInfo

- validated: true

- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
  - lingualityType: bilingual
- LanguageInfo:
  - languageCoding: spa
  - languageName: Spanish
- LanguageInfo:
  - languageCoding: ast
  - languageName: Asturian
- SizeInfo:
  - size: 37217
  - sizeUnitMultiplier: unit
  - sizeUnit: entries
- TextFormatInfo:
  - mime-type: text/xml
- CharacterEncodingInfo:
  - characterEncoding: UTF-8
- DomainInfo:
  - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Apertium bilingual dictionary for Spanish and Asturian languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: lexicon
- LexicalConceptualResourceCreationInfo:
  - originalSource: Apertium Bilingual dictionary as described in http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
  - creationMode: automatic
  - creationModeDetails: This lexicon was created with the Apertium2LMP.xsl
  - creationTool: Apertium2LMP.xsl
- LexicalConceptualResourceEncodingInfo:
  - encodingLevel: morphology
  - conformanceToStandardsBestPractice: LMF
  - linguisticInformation: lemma
  - linguisticInformation: semantics-CrossReferences
  - linguisticInformation: partOfSpeech

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
- publication: 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.

Spanish-Catalan LMF Apertium Bilingual dictionary

This is the LMF version of the Apertium bilingual dictionary for Spanish and Catalan languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.

BASIC INFORMATION

IdentificationInfo

- resourceName: Spanish-Catalan LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium Es-Ca

- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: Spanish-Catalan LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium Es-Ca
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018

        - city: Barcelona
        - country: Spain
        - telephoneNumber: +34 93 5421207
        - faxNumber: +34 93 5422321
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: GPL
    - restrictionsOfUse: other
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)
            - CommunicationInfo:
                - address: Roc Boronat, 138
                - zipCode: 08018
                - city: Barcelona
                - country: Spain
                - telephoneNumber: +34 93 5421207
                - faxNumber: +34 93 5422321
                - email: jorge.vivaldi@upf.edu
                - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- FundingInfo:

- ProjectInfo:
  - projectName: METANET4U
  - projectShortName: METANET4U
  - fundingType: ownFunds
  - fundingType: euFunds
  - fundingCountry: Spain
  - funder: UPF
  - funder: EU
  - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
  - lingualityType: bilingual
- LanguageInfo:
  - languageCoding: cat
  - languageName: Catalan
- LanguageInfo:
  - languageCoding: spa
  - languageName: Spanish
- SizeInfo:
  - size: 37687
  - sizeUnitMultiplier: unit
  - sizeUnit: entries
- TextFormatInfo:
  - mime-type: text/xml
- CharacterEncodingInfo:
  - characterEncoding: UTF-8
- DomainInfo:
  - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Apertium bilingual dictionary for Spanish and Catalan languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: lexicon
- LexicalConceptualResourceCreationInfo:
    - originalSource: Apertium Bilingual dictionary as described in http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
    - creationMode: automatic
    - creationModeDetails: This lexicon was created with the Apertium2LMP.xsl
    - creationTool: Apertium2LMP.xsl
- LexicalConceptualResourceEncodingInfo:
    - encodingLevel: morphology
    - conformanceToStandardsBestPractice: LMF
    - linguisticInformation: lemma
    - linguisticInformation: semantics-CrossReferences
    - linguisticInformation: partOfSpeech

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
- publication: 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.

Spanish-Galician LMF Apertium Bilingual dictionary

This is the LMF version of the Apertium bilingual dictionary for Spanish Galician languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries

(LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.

BASIC INFORMATION

IdentificationInfo

- resourceName: Spanish-Galician LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium Es-Gl
- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: Spanish-Galician LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium Es-Gl
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu

        - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: GPL
    - restrictionsOfUse: other
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)
            - CommunicationInfo:
                - address: Roc Boronat, 138
                - zipCode: 08018
                - city: Barcelona
                - country: Spain
                - telephoneNumber: +34 93 5421207
                - faxNumber: +34 93 5422321
                - email: jorge.vivaldi@upf.edu
                - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
    - OrganizationInfo:

        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- FundingInfo:
    - ProjectInfo:
        - projectName: METANET4U
        - projectShortName: METANET4U
        - fundingType: ownFunds
        - fundingType: euFunds
        - fundingCountry: Spain
        - funder: UPF
        - funder: EU
        - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
    - lingualityType: bilingual
- LanguageInfo:
    - languageCoding: spa
    - languageName: Spanish
- LanguageInfo:
    - languageCoding: glg
    - languageName: Galician
- SizeInfo:
    - size: 9263
    - sizeUnitMultiplier: unit
    - sizeUnit: entries
- TextFormatInfo:
    - mime-type: text/xml
- CharacterEncodingInfo:

    - characterEncoding: UTF-8
- DomainInfo:
    - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Apertium bilingual dictionary for Spanish Galician languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: lexicon
- LexicalConceptualResourceCreationInfo:
    - originalSource: Apertium Bilingual dictionary as described in http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
    - creationMode: automatic
    - creationModeDetails: This lexicon was created with the Apertium2LMP.xsl
    - creationTool: Apertium2LMP.xsl
- LexicalConceptualResourceEncodingInfo:
    - encodingLevel: morphology
    - conformanceToStandardsBestPractice: LMF
    - linguisticInformation: lemma
    - linguisticInformation: semantics-CrossReferences
    - linguisticInformation: partOfSpeech

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
- publication: 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.

Spanish-Portuguese LMF Apertium Bilingual dictionary

This is the LMF version of the Apertium bilingual dictionary for Spanish and Portuguese languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.

BASIC INFORMATION

IdentificationInfo

- resourceName: Spanish-Portuguese LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium Es-Pt
- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: Spanish-Portuguese LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium Es-Pt
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu

- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: GPL
    - restrictionsOfUse: other
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)
            - CommunicationInfo:
                - address: Roc Boronat, 138
                - zipCode: 08018
                - city: Barcelona

        - country: Spain
        - telephoneNumber: +34 93 5421207
        - faxNumber: +34 93 5422321
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- FundingInfo:
    - ProjectInfo:
        - projectName: METANET4U
        - projectShortName: METANET4U
        - fundingType: ownFunds
        - fundingType: euFunds
        - fundingCountry: Spain
        - funder: UPF
        - funder: EU
        - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:

    - lingualityType: bilingual
- LanguageInfo:
    - languageCoding: spa
    - languageName: Spanish
- LanguageInfo:
    - languageCoding: por
    - languageName: Portuguese
- SizeInfo:
    - size: 13332
    - sizeUnitMultiplier: unit
    - sizeUnit: entries
- TextFormatInfo:
    - mime-type: text/xml
- CharacterEncodingInfo:
    - characterEncoding: UTF-8
- DomainInfo:
    - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Apertium bilingual dictionary for Spanish and Portuguese languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: lexicon
- LexicalConceptualResourceCreationInfo:
    - originalSource: Apertium Bilingual dictionary as described in http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
    - creationMode: automatic

    - creationModeDetails: This lexicon was created with the Apertium2LMP.xsl
    - creationTool: Apertium2LMP.xsl
- LexicalConceptualResourceEncodingInfo:
    - encodingLevel: morphology
    - conformanceToStandardsBestPractice: LMF
    - linguisticInformation: lemma
    - linguisticInformation: semantics-CrossReferences
    - linguisticInformation: partOfSpeech

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
- publication: 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.

Spanish-Romanian LMF Apertium Bilingual dictionary

This is the LMF version of the Apertium bilingual dictionary for Spanish Romanian languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.

BASIC INFORMATION

IdentificationInfo

- resourceName: Spanish-Romanian LMF Apertium Bilingual dictionary
- resourceShortName: LMF Apertium Es-Ro
- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: Spanish-Romanian LMF Apertium Bilingual dictionary

- resourceShortName: LMF Apertium Es-Ro
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:

- license: GPL
- restrictionsOfUse: other
- distributionAccessMedium: downloadable
- downloadLocation: http://www..com/
- distributor:
  - OrganizationInfo:
    - organizationName: Institut Universitari Lingüística Aplicada - UPF
    - organizationShortName: IULA - UPF
    - departmentName: Institut Universitari Lingüística Aplicada (UPF)
    - CommunicationInfo:
      - address: Roc Boronat, 138
      - zipCode: 08018
      - city: Barcelona
      - country: Spain
      - telephoneNumber: +34 93 5421207
      - faxNumber: +34 93 5422321
      - email: jorge.vivaldi@upf.edu
      - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
  - OrganizationInfo:
    - organizationName: Institut Universitari Lingüística Aplicada - UPF
    - organizationShortName: IULA - UPF
    - departmentName: Institut Universitari Lingüística Aplicada (UPF)
    - CommunicationInfo:
      - address: Roc Boronat, 138
      - zipCode: 08018
      - city: Barcelona
      - country: Spain
      - telephoneNumber: +34 93 5421207
      - faxNumber: +34 93 5422321
      - email: jorge.vivaldi@upf.edu
      - url: http://www.iula.upf.edu/
- FundingInfo:
  - ProjectInfo:
    - projectName: METANET4U
    - projectShortName: METANET4U
    - fundingType: ownFunds
    - fundingType: euFunds
    - fundingCountry: Spain
    - funder: UPF

- funder: EU
- url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
  - lingualityType: bilingual
- LanguageInfo:
  - languageCoding: spa
  - languageName: Spanish
- LanguageInfo:
  - languageCoding: ron
  - languageName: Romanian
- SizeInfo:
  - size: 14579
  - sizeUnitMultiplier: unit
  - sizeUnit: entries
- TextFormatInfo:
  - mime-type: text/xml
- CharacterEncodingInfo:
  - characterEncoding: UTF-8
- DomainInfo:
  - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Apertium bilingual dictionary for Spanish Romanian languages. Bilingual LMF dictionaries were generated from Apertium bilingual dix files. For each Apertium bilingual correspondence, the corresponding source and target monolingual entries (LexicalEntry) were generated in addition to the bilingual correspondence (SenseAxis) element. 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.

- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: lexicon
- LexicalConceptualResourceCreationInfo:
  - originalSource: Apertium Bilingual dictionary as described in http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
  - creationMode: automatic
  - creationModeDetails: This lexicon was created with the Apertium2LMP.xsl
  - creationTool: Apertium2LMP.xsl
- LexicalConceptualResourceEncodingInfo:
  - encodingLevel: morphology
  - conformanceToStandardsBestPractice: LMF
  - linguisticInformation: lemma
  - linguisticInformation: semantics-CrossReferences
  - linguisticInformation: partOfSpeech

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
- publication: 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.

Catalan LMF Apertium Dictionary

This is the LMF version of the Catalan Apertium ditionary. Monolingual dictionaries for Spanish, Catalan, Gallego and Euskera have been generated from the Apertium expanded lexicons of the es-ca (for both Spanish and Catalan) es-gl (for Galician) and eu-es (for Basque). 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.

BASIC INFORMATION

IdentificationInfo

- resourceName: Catalan LMF Apertium Dictionary
- resourceShortName: LMF Apertium Ca
- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: Catalan LMF Apertium Dictionary
- resourceShortName: LMF Apertium Ca
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF

        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: GPL
    - restrictionsOfUse: other
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)
            - CommunicationInfo:
                - address: Roc Boronat, 138
                - zipCode: 08018
                - city: Barcelona
                - country: Spain
                - telephoneNumber: +34 93 5421207
                - faxNumber: +34 93 5422321
                - email: jorge.vivaldi@upf.edu
                - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207

        - faxNumber: +34 93 5422321
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/
- FundingInfo:
    - ProjectInfo:
        - projectName: METANET4U
        - projectShortName: METANET4U
        - fundingType: ownFunds
        - fundingType: euFunds
        - fundingCountry: Spain
        - funder: UPF
        - funder: EU
        - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
    - lingualityType: monolingual
- LanguageInfo:
    - languageCoding: cat
    - languageName: Catalan
- SizeInfo:
    - size: 37644
    - sizeUnitMultiplier: unit
    - sizeUnit: entries
- TextFormatInfo:
    - mime-type: text/xml
- CharacterEncodingInfo:
    - characterEncoding: UTF-8
- DomainInfo:
    - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Catalan Apertium ditionary. Monolingual dictionaries for Spanish, Catalan, Gallego and Euskera have been generated from the Apertium expanded lexicons of the es-ca (for both Spanish and Catalan) es-gl (for Galician) and eu-es (for Basque). 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.
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: lexicon
- LexicalConceptualResourceCreationInfo:
    - originalSource: Apertium Monolingual dictionary as described in http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
    - creationMode: automatic
    - creationModeDetails: This lexicon was created with the Apertium2LMP.xsl
    - creationTool: Apertium2LMP.xsl
- LexicalConceptualResourceEncodingInfo:
    - encodingLevel: morphology
    - conformanceToStandardsBestPractice: LMF
    - linguisticInformation: lemma
    - linguisticInformation: semantics-CrossReferences
    - linguisticInformation: partOfSpeech

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
- publication: 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.

Galician LMF Apertium Dictionary

This is the LMF version of the Galician Apertium dictionary. Monolingual dictionaries for Spanish, Catalan, Galician and Euskera have been generated from the Apertium expanded lexicons of the es-ca (for both Spanish and Catalan) es-gl (for Galician) and eu-es (for Basque). 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.

BASIC INFORMATION

IdentificationInfo

- resourceName: Galician LMF Apertium Dictionary
- resourceShortName: LMF Apertium Gl
- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: Galician LMF Apertium Dictionary
- resourceShortName: LMF Apertium Gl
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: GPL
    - restrictionsOfUse: other
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)
            - CommunicationInfo:
                - address: Roc Boronat, 138
                - zipCode: 08018
                - city: Barcelona
                - country: Spain
                - telephoneNumber: +34 93 5421207
                - faxNumber: +34 93 5422321
                - email: jorge.vivaldi@upf.edu
                - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF

        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- FundingInfo:
    - ProjectInfo:
        - projectName: METANET4U
        - projectShortName: METANET4U
        - fundingType: ownFunds
        - fundingType: euFunds
        - fundingCountry: Spain
        - funder: UPF
        - funder: EU
        - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
    - lingualityType: monolingual
- LanguageInfo:
    - languageCoding: glg
    - languageName: Galician
- SizeInfo:
    - size: 9134
    - sizeUnitMultiplier: unit
    - sizeUnit: entries
- TextFormatInfo:
    - mime-type: text/xml
- CharacterEncodingInfo:
    - characterEncoding: UTF-8
- DomainInfo:
    - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Galician Apertium dictionary. Monolingual dictionaries for Spanish, Catalan, Galician and Euskera have been generated from the Apertium expanded lexicons of the es-ca (for both Spanish and Catalan) es-gl (for Galician) and eu-es (for Basque). 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.
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: lexicon
- LexicalConceptualResourceCreationInfo:
    - originalSource: Apertium Galician Monolingual dictionary as described in http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
    - creationMode: automatic
    - creationModeDetails: This lexicon was created with the Apertium2LMP.xsl
    - creationTool: Apertium2LMP.xsl
- LexicalConceptualResourceEncodingInfo:
    - encodingLevel: morphology
    - conformanceToStandardsBestPractice: LMF
    - linguisticInformation: lemma
    - linguisticInformation: semantics-CrossReferences
    - linguisticInformation: partOfSpeech

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
- publication: 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.

Spanish LMF Apertium Dictionary

This is the LMF version of the Apertium Spanish dictionary. Monolingual dictionaries for Spanish, Catalan, Gallego and Euskera have been generated from the Apertium expanded lexicons of the es-ca (for both Spanish and Catalan) es-gl (for Galician) and eu-es (for Basque). 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.

BASIC INFORMATION

IdentificationInfo

- resourceName: Spanish LMF Apertium Dictionary
- resourceShortName: LMF Apertium Es
- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: Spanish LMF Apertium Dictionary
- resourceShortName: LMF Apertium Es
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138

        - zipCode: 08018
        - city: Barcelona
        - country: Spain
        - telephoneNumber: +34 93 5421207
        - faxNumber: +34 93 5422321
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: GPL
    - restrictionsOfUse: other
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)
            - CommunicationInfo:
                - address: Roc Boronat, 138
                - zipCode: 08018
                - city: Barcelona
                - country: Spain
                - telephoneNumber: +34 93 5421207
                - faxNumber: +34 93 5422321
                - email: jorge.vivaldi@upf.edu
                - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- FundingInfo:
    - ProjectInfo:
        - projectName: METANET4U
        - projectShortName: METANET4U
        - fundingType: ownFunds
        - fundingType: euFunds
        - fundingCountry: Spain
        - funder: UPF
        - funder: EU
        - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
    - lingualityType: monolingual
- LanguageInfo:
    - languageCoding: spa
    - languageName: Spanish
- SizeInfo:
    - size: 38604
    - sizeUnitMultiplier: unit
    - sizeUnit: entries

- TextFormatInfo:
  - mime-type: text/xml
- CharacterEncodingInfo:
  - characterEncoding: UTF-8
- DomainInfo:
  - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Apertium Spanish dictionary. Monolingual dictionaries for Spanish, Catalan, Gallego and Euskera have been generated from the Apertium expanded lexicons of the es-ca (for both Spanish and Catalan) es-gl (for Galician) and eu-es (for Basque). 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.
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: lexicon
- LexicalConceptualResourceCreationInfo:
  - originalSource: Apertium Monolingual dictionary as described in http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
  - creationMode: automatic
  - creationModeDetails: This lexicon was created with the Apertium2LMP.xsl
  - creationTool: Apertium2LMP.xsl
- LexicalConceptualResourceEncodingInfo:
  - encodingLevel: morphology
  - conformanceToStandardsBestPractice: LMF
  - linguisticInformation: lemma
  - linguisticInformation: semantics-CrossReferences
  - linguisticInformation: partOfSpeech

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf
- publication: 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.

Freeling

Asturian LMF Freeling Lexicon

This is the LMF version of the Asturian Freeling lexicon. FreeLing is a developer-oriented library providing language analysis services. If you want to develop, say, a machine translation system, and you need some kind of linguistic processing of the source text, your MT application can call FreeLing modules to do the required analysis.

BASIC INFORMATION

IdentificationInfo

- resourceName: Asturian LMF Freeling Lexicon
- resourceShortName: Asturian LMF Freeling Lexicon
- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: Asturian LMF Freeling Lexicon
- resourceShortName: Asturian LMF Freeling Lexicon
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018

    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: GPL
    - restrictionsOfUse: other
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)
            - CommunicationInfo:

        - address: Roc Boronat, 138
        - zipCode: 08018
        - city: Barcelona
        - country: Spain
        - telephoneNumber: +34 93 5421207
        - faxNumber: +34 93 5422321
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- FundingInfo:
    - ProjectInfo:
        - projectName: METANET4U
        - projectShortName: METANET4U
        - fundingType: ownFunds
        - fundingType: euFunds
        - fundingCountry: Spain
        - funder: UPF
        - funder: EU
        - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
  - lingualityType: monolingual
- LanguageInfo:
  - languageCoding: ast
  - languageName: Asturian
- SizeInfo:
  - size: 40048
  - sizeUnitMultiplier: unit
  - sizeUnit: entries
- TextFormatInfo:
  - mime-type: text/xml
- CharacterEncodingInfo:
  - characterEncoding: UTF-8
- DomainInfo:
  - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Asturian Freeling lexicon. FreeLing is a developer-oriented library providing language analysis services. If you want to develop, say, a machine translation system, and you need some kind of linguistic processing of the source text, your MT application can call FreeLing modules to do the required analysis.
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: wordList
- LexicalConceptualResourceCreationInfo:
  - originalSource: Asturian Freeling lexicon as described in http://nlp.lsi.upc.edu/freeling/
  - creationMode: automatic
  - creationModeDetails: This lexicon was created with the Freeling2LMF.pl
  - creationTool: Freeling2LMF.pl
- LexicalConceptualResourceEncodingInfo:
  - encodingLevel: morphology
  - conformanceToStandardsBestPractice: LMF
  - linguisticInformation: lemma

- linguisticInformation: partOfSpeech

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: Freeling User Manual at http://nlp.lsi.upc.edu/freeling/doc/userman/userman.pdf

Catalan LMF Freeling Lexicon

This is the LMF version of the Catalan Freeling lexicon. FreeLing is a developer-oriented library providing language analysis services. If you want to develop, say, a machine translation system, and you need some kind of linguistic processing of the source text, your MT application can call FreeLing modules to do the required analysis.

BASIC INFORMATION

IdentificationInfo

- resourceName: Catalan LMF Freeling Lexicon
- resourceShortName: Catalan LMF Freeling Lexicon
- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: Catalan LMF Freeling Lexicon
- resourceShortName: Catalan LMF Freeling Lexicon
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
  - address: Roc Boronat, 138
  - zipCode: 08018
  - city: Barcelona
  - country: Spain
  - email: jorge.vivaldi@upf.edu
- affiliation:
  - OrganizationInfo:

        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: GPL
    - restrictionsOfUse: other
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)
            - CommunicationInfo:
                - address: Roc Boronat, 138
                - zipCode: 08018
                - city: Barcelona
                - country: Spain
                - telephoneNumber: +34 93 5421207

        - faxNumber: +34 93 5422321
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- FundingInfo:
    - ProjectInfo:
        - projectName: METANET4U
        - projectShortName: METANET4U
        - fundingType: ownFunds
        - fundingType: euFunds
        - fundingCountry: Spain
        - funder: UPF
        - funder: EU
        - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
    - lingualityType: monolingual
- LanguageInfo:

    - languageCoding: cat
    - languageName: Catalan
- SizeInfo:
    - size: 71862
    - sizeUnitMultiplier: unit
    - sizeUnit: entries
- TextFormatInfo:
    - mime-type: text/xml
- CharacterEncodingInfo:
    - characterEncoding: UTF-8
- DomainInfo:
    - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Catalan Freeling lexicon. FreeLing is a developer-oriented library providing language analysis services. If you want to develop, say, a machine translation system, and you need some kind of linguistic processing of the source text, your MT application can call FreeLing modules to do the required analysis.
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: wordList
- LexicalConceptualResourceCreationInfo:
    - originalSource: Catalan Freeling lexicon as described in http://nlp.lsi.upc.edu/freeling/
    - creationMode: automatic
    - creationModeDetails: This lexicon was created with the Freeling2LMF.pl
    - creationTool: Freeling2LMF.pl
- LexicalConceptualResourceEncodingInfo:
    - encodingLevel: morphology
    - conformanceToStandardsBestPractice: LMF
    - linguisticInformation: lemma
    - linguisticInformation: semantics-CrossReferences
    - linguisticInformation: partOfSpeech

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://nlp.lsi.upc.edu/freeling/doc/userman/userman.pdf

Spanish LMF Freeling Lexicon

This is the LMF version of the Spanish Freeling lexicon. FreeLing is a developer-oriented library providing language analysis services. If you want to develop, say, a machine translation system, and you need some kind of linguistic processing of the source text, your MT application can call FreeLing modules to do the required analysis.

BASIC INFORMATION

IdentificationInfo

- resourceName: Spanish LMF Freeling Lexicon
- resourceShortName: Spanish LMF Freeling Lexicon
- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: Spanish LMF Freeling Lexicon
- resourceShortName: Spanish LMF Freeling Lexicon
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138

        - zipCode: 08018
        - city: Barcelona
        - country: Spain
        - telephoneNumber: +34 93 5421207
        - faxNumber: +34 93 5422321
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: GPL
    - restrictionsOfUse: other
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)
            - CommunicationInfo:
                - address: Roc Boronat, 138
                - zipCode: 08018
                - city: Barcelona
                - country: Spain
                - telephoneNumber: +34 93 5421207
                - faxNumber: +34 93 5422321
                - email: jorge.vivaldi@upf.edu
                - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- FundingInfo:
    - ProjectInfo:
        - projectName: METANET4U
        - projectShortName: METANET4U
        - fundingType: ownFunds
        - fundingType: euFunds
        - fundingCountry: Spain
        - funder: UPF
        - funder: EU
        - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
    - lingualityType: monolingual
- LanguageInfo:
    - languageCoding: spa
    - languageName: Spanish
- SizeInfo:
    - size: 76318
    - sizeUnitMultiplier: unit
    - sizeUnit: entries

- TextFormatInfo:
  - mime-type: text/xml
- CharacterEncodingInfo:
  - characterEncoding: UTF-8
- DomainInfo:
  - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Spanish Freeling lexicon. FreeLing is a developer-oriented library providing language analysis services. If you want to develop, say, a machine translation system, and you need some kind of linguistic processing of the source text, your MT application can call FreeLing modules to do the required analysis.
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: wordList
- LexicalConceptualResourceCreationInfo:
  - originalSource: Spanish Freeling lexicon as described in http://nlp.lsi.upc.edu/freeling/index.php
  - creationMode: automatic
  - creationModeDetails: This lexicon was created with the Freeling2LMF.pl.
  - creationTool: Freeling2LMF.pl
- LexicalConceptualResourceEncodingInfo:
  - encodingLevel: morphology
  - conformanceToStandardsBestPractice: LMF
  - linguisticInformation: lemma
  - linguisticInformation: semantics-CrossReferences
  - linguisticInformation: partOfSpeech

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://nlp.lsi.upc.edu/freeling/index.php?option=com_content
- publication: http://nlp.lsi.upc.edu/freeling/doc/userman/userman.pdf

Galician LMF Freeling Lexicon

This is the LMF version of the Galician Freeling lexicon. FreeLing is a developer-oriented library providing language analysis services. If you want to develop, say, a machine translation system, and you need some kind of linguistic processing of the source text, your MT application can call FreeLing modules to do the required analysis.

BASIC INFORMATION

IdentificationInfo

- resourceName: Galician LMF Freeling Lexicon
- resourceShortName: Galician LMF Freeling Lexicon
- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: Galician LMF Freeling Lexicon
- resourceShortName: Galician LMF Freeling Lexicon
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207

        - faxNumber: +34 93 5422321
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: GPL
    - restrictionsOfUse: other
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)
            - CommunicationInfo:
                - address: Roc Boronat, 138
                - zipCode: 08018
                - city: Barcelona
                - country: Spain
                - telephoneNumber: +34 93 5421207
                - faxNumber: +34 93 5422321
                - email: jorge.vivaldi@upf.edu
                - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30

- resourceCreator:
  - OrganizationInfo:
    - organizationName: Institut Universitari Lingüística Aplicada - UPF
    - organizationShortName: IULA - UPF
    - departmentName: Institut Universitari Lingüística Aplicada (UPF)
    - CommunicationInfo:
      - address: Roc Boronat, 138
      - zipCode: 08018
      - city: Barcelona
      - country: Spain
      - telephoneNumber: +34 93 5421207
      - faxNumber: +34 93 5422321
      - email: jorge.vivaldi@upf.edu
      - url: http://www.iula.upf.edu/
- FundingInfo:
  - ProjectInfo:
    - projectName: METANET4U
    - projectShortName: METANET4U
    - fundingType: ownFunds
    - fundingType: euFunds
    - fundingCountry: Spain
    - funder: UPF
    - funder: EU
    - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
  - lingualityType: monolingual
- LanguageInfo:
  - languageCoding: glg
  - languageName: Galician
- SizeInfo:
  - size: 49898
  - sizeUnitMultiplier: unit
  - sizeUnit: entries
- TextFormatInfo:
  - mime-type: text/xml
- CharacterEncodingInfo:
  - characterEncoding: UTF-8

- DomainInfo:
    - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Galician Freeling lexicon. FreeLing is a developer-oriented library providing language analysis services. If you want to develop, say, a machine translation system, and you need some kind of linguistic processing of the source text, your MT application can call FreeLing modules to do the required analysis.
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: wordList
- LexicalConceptualResourceCreationInfo:
    - originalSource: Galician Freeling lexicon as described in http://nlp.lsi.upc.edu/freeling/
    - creationMode: automatic
    - creationModeDetails: This lexicon was created with the Freeling2LMF.pl
    - creationTool: Freeling2LMF.pl
- LexicalConceptualResourceEncodingInfo:
    - encodingLevel: morphology
    - conformanceToStandardsBestPractice: LMF
    - linguisticInformation: lemma
    - linguisticInformation: semantics-CrossReferences
    - linguisticInformation: partOfSpeech

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://nlp.lsi.upc.edu/freeling/index.php?option=com_content
- publication: http://nlp.lsi.upc.edu/freeling/doc/userman/userman.pdf

Catalan LMF Freeling Sense

This is the LMF version of the Catalan Freeling Sense. FreeLing is a developer-oriented library providing language analysis services. If you want to develop, say, a machine translation system, and you need some kind of linguistic processing of the source text, your MT application can call FreeLing modules to do the required analysis.

BASIC INFORMATION

IdentificationInfo

- resourceName: Catalan LMF Freeling Sense
- resourceShortName: Catalan LMF Freeling Sense
- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: Catalan LMF Freeling Sense
- resourceShortName: Catalan LMF Freeling Sense
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207

        - faxNumber: +34 93 5422321
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: GPL
    - restrictionsOfUse: other
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)
            - CommunicationInfo:
                - address: Roc Boronat, 138
                - zipCode: 08018
                - city: Barcelona
                - country: Spain
                - telephoneNumber: +34 93 5421207
                - faxNumber: +34 93 5422321
                - email: jorge.vivaldi@upf.edu
                - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30

- resourceCreator:
  - OrganizationInfo:
    - organizationName: Institut Universitari Lingüística Aplicada - UPF
    - organizationShortName: IULA - UPF
    - departmentName: Institut Universitari Lingüística Aplicada (UPF)
    - CommunicationInfo:
      - address: Roc Boronat, 138
      - zipCode: 08018
      - city: Barcelona
      - country: Spain
      - telephoneNumber: +34 93 5421207
      - faxNumber: +34 93 5422321
      - email: jorge.vivaldi@upf.edu
      - url: http://www.iula.upf.edu/
- FundingInfo:
  - ProjectInfo:
    - projectName: METANET4U
    - projectShortName: METANET4U
    - fundingType: ownFunds
    - fundingType: euFunds
    - fundingCountry: Spain
    - funder: UPF
    - funder: EU
    - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
  - lingualityType: monolingual
- LanguageInfo:
  - languageCoding: cat
  - languageName: Catalan
- SizeInfo:
  - size: 43561
  - sizeUnitMultiplier: unit
  - sizeUnit: entries
- SizeInfo:
  - size: 65380
  - sizeUnitMultiplier: unit
  - sizeUnit: semanticUnits

- TextFormatInfo:
  - mime-type: text/xml
- CharacterEncodingInfo:
  - characterEncoding: UTF-8
- DomainInfo:
  - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Catalan Freeling Sense. FreeLing is a developer-oriented library providing language analysis services. If you want to develop, say, a machine translation system, and you need some kind of linguistic processing of the source text, your MT application can call FreeLing modules to do the required analysis.
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: wordnet
- LexicalConceptualResourceCreationInfo:
  - originalSource: Catalan Freeling Sense as described in http://nlp.lsi.upc.edu/freeling/
  - creationMode: automatic
  - creationModeDetails: This lexicon was created with the Freeling2LMF.pl.
  - creationTool: Freeling2LMF.pl
- LexicalConceptualResourceEncodingInfo:
  - encodingLevel: morphology
  - conformanceToStandardsBestPractice: LMF
  - linguisticInformation: lemma
  - linguisticInformation: semantics-CrossReferences
  - linguisticInformation: semantics-Relations

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://nlp.lsi.upc.edu/freeling/index.php?option=com_content
- publication: http://nlp.lsi.upc.edu/freeling/doc/userman/userman.pdf

Spanish LMF Freeling Sense

This is the LMF version of the Spanish Freeling Sense. FreeLing is a developer-oriented library providing language analysis services. If you want to develop, say, a machine translation system, and you need some kind of linguistic processing of the source text, your MT application can call FreeLing modules to do the required analysis.

BASIC INFORMATION

IdentificationInfo

- resourceName: Spanish LMF Freeling Sense
- resourceShortName: Spanish LMF Freeling Sense
- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: Spanish LMF Freeling Sense
- resourceShortName: Spanish LMF Freeling Sense
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207

        - faxNumber: +34 93 5422321
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: GPL
    - restrictionsOfUse: other
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)
            - CommunicationInfo:
                - address: Roc Boronat, 138
                - zipCode: 08018
                - city: Barcelona
                - country: Spain
                - telephoneNumber: +34 93 5421207
                - faxNumber: +34 93 5422321
                - email: jorge.vivaldi@upf.edu
                - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30

- resourceCreator:
  - OrganizationInfo:
    - organizationName: Institut Universitari Lingüística Aplicada - UPF
    - organizationShortName: IULA - UPF
    - departmentName: Institut Universitari Lingüística Aplicada (UPF)
    - CommunicationInfo:
      - address: Roc Boronat, 138
      - zipCode: 08018
      - city: Barcelona
      - country: Spain
      - telephoneNumber: +34 93 5421207
      - faxNumber: +34 93 5422321
      - email: jorge.vivaldi@upf.edu
      - url: http://www.iula.upf.edu/
- FundingInfo:
  - ProjectInfo:
    - projectName: METANET4U
    - projectShortName: METANET4U
    - fundingType: ownFunds
    - fundingType: euFunds
    - fundingCountry: Spain
    - funder: UPF
    - funder: EU
    - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
  - lingualityType: monolingual
- LanguageInfo:
  - languageCoding: spa
  - languageName: Spanish
- SizeInfo:
  - size: 6213
  - sizeUnitMultiplier: unit
  - sizeUnit: entries
- SizeInfo:
  - size: 17359
  - sizeUnitMultiplier: unit
  - sizeUnit: semanticUnits

- TextFormatInfo:
    - mime-type: text/xml
- CharacterEncodingInfo:
    - characterEncoding: UTF-8
- DomainInfo:
    - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Spanish Freeling Sense. FreeLing is a developer-oriented library providing language analysis services. If you want to develop, say, a machine translation system, and you need some kind of linguistic processing of the source text, your MT application can call FreeLing modules to do the required analysis.
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: wordnet
- LexicalConceptualResourceCreationInfo:
    - originalSource: Spanish Freeling Sense as described in http://nlp.lsi.upc.edu/freeling/
    - creationMode: automatic
    - creationModeDetails: This lexicon was created with the Freeling2LMF.pl
    - creationTool: Freeling2LMF.pl
- LexicalConceptualResourceEncodingInfo:
    - encodingLevel: morphology
    - conformanceToStandardsBestPractice: LMF
    - linguisticInformation: lemma
    - linguisticInformation: semantics-CrossReferences
    - linguisticInformation: semantics-Relations

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://nlp.lsi.upc.edu/freeling/doc/userman/userman.pdf

Parole

Spanish LMF Parole Lexicon

This is the LMF version of the Spanish Parole 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.

BASIC INFORMATION

IdentificationInfo

- resourceName: Spanish LMF Parole Lexicon
- resourceShortName: Spanish LMF Parole Lexicon
- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: Spanish LMF Parole Lexicon
- resourceShortName: Spanish LMF Parole Lexicon
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada

        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: AGPL
    - restrictionsOfUse: academic-nonCommercialUse
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)
            - CommunicationInfo:
                - address: Roc Boronat, 138
                - zipCode: 08018
                - city: Barcelona
                - country: Spain
                - telephoneNumber: +34 93 5421207
                - faxNumber: +34 93 5422321
                - email: jorge.vivaldi@upf.edu
                - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- FundingInfo:
    - ProjectInfo:
        - projectName: METANET4U
        - projectShortName: METANET4U
        - fundingType: ownFunds
        - fundingType: euFunds
        - fundingCountry: Spain
        - funder: UPF
        - funder: EU
        - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
    - lingualityType: monolingual
- LanguageInfo:
    - languageCoding: spa
    - languageName: Spanish
- SizeInfo:
    - size: 64594

    - sizeUnitMultiplier: unit
    - sizeUnit: entries
- TextFormatInfo:
    - mime-type: text/xml
- CharacterEncodingInfo:
    - characterEncoding: UTF-8
- DomainInfo:
    - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Spanish Parole 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.
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: lexicon
- LexicalConceptualResourceCreationInfo:
    - originalSource: Spanish PAROLE lexicon as described in http://www.ub.edu/gilcub/lascosas/pubYreps/index_par.html. 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: ParoleSimple2LMF.xsl
- LexicalConceptualResourceEncodingInfo:
    - encodingLevel: morphology
    - conformanceToStandardsBestPractice: LMF
    - theoreticModel: Parole
    - theoreticModel: Genelex
    - linguisticInformation: lemma
    - linguisticInformation: morpho-Inflection

- linguisticInformation: partOfSpeech

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://www.ub.edu/gilcub/lascosas/pubYreps/index_par.html
- publication: Villegas, M., Brosa, I. and Bel, N. (1998). 'El léxico PAROLE del español'. Actas del XIV Congreso de la SEPLN, Septiembre 1998.

Spanish LMF ParoleSimple Lexicon

This is the LMF version of the Spanish ParoleSimple 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.

BASIC INFORMATION

IdentificationInfo

- resourceName: Spanish LMF ParoleSimple Lexicon
- resourceShortName: Spanish LMF ParoleSimple Lexicon
- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: Spanish LMF ParoleSimple Lexicon
- resourceShortName: Spanish LMF ParoleSimple Lexicon
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:

    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: AGPL
    - restrictionsOfUse: academic-nonCommercialUse
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF

        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- FundingInfo:
    - ProjectInfo:
        - projectName: METANET4U
        - projectShortName: METANET4U
        - fundingType: ownFunds
        - fundingType: euFunds
        - fundingCountry: Spain
        - funder: UPF
        - funder: EU
        - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic

- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
  - lingualityType: monolingual
- LanguageInfo:
  - languageCoding: spa
  - languageName: Spanish
- SizeInfo:
  - size: 7698
  - sizeUnitMultiplier: unit
  - sizeUnit: entries
- SizeInfo:
  - size: 9861
  - sizeUnitMultiplier: unit
  - sizeUnit: semanticUnits
- TextFormatInfo:
  - mime-type: text/xml
- CharacterEncodingInfo:
  - characterEncoding: UTF-8
- DomainInfo:
  - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Spanish ParoleSimple 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.
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: lexicon
- LexicalConceptualResourceCreationInfo:
  - originalSource: Spanish PAROLE lexicon as described in http://www.ub.edu/gilcub/lascosas/pubYreps/index_par.html. 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: ParoleSimple2LMF.xsl
- LexicalConceptualResourceEncodingInfo:
  - encodingLevel: morphology
  - encodingLevel: syntax
  - encodingLevel: semantics
  - conformanceToStandardsBestPractice: LMF
  - theoreticModel: Parole
  - theoreticModel: Genelex
  - linguisticInformation: lemma
  - linguisticInformation: morpho-Inflection
  - linguisticInformation: partOfSpeech
  - linguisticInformation: syntax-SubcatFrame
  - linguisticInformation: syntacticoSemanticLinks

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://www.ub.edu/gilcub/lascosas/pubYreps/index_par.html
- publication: http://www.ub.edu/gilcub/SIMPLE/reports/simple/SIMPLE_FGuidelines.rtf.zip
- publication: Villegas, M., Brosa, I. and Bel, N. (1998). 'El léxico PAROLE del español'. Actas del XIV Congreso de la SEPLN, Septiembre 1998.

Neologisms of the year Bank of Spanish &Catalan Neologisms

Bank of Catalan Neologisms

This is the LMF version of the Spanish Bank of Neologisms at the Observatori de Neologia (UPF). The Observatori de Neologia (OBNEO), under the direction by Dr. M. Teresa Cabré, is a public-funded consolidated group within the Institut Universitari de Lingüística Aplicada at Universitat Pompeu Fabra. This project analyzes the phenomenon of the appearance of new words or neologisms in the usage, both for Catalan and Spanish. Since 1996 has been recognized as a consolidated research group of Universitat Pompeu Fabra.

BASIC INFORMATION

IdentificationInfo

- resourceName: Bank of Catalan Neologisms
- resourceShortName: Bank of Catalan Neologisms
- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: Bank of Catalan Neologisms
- resourceShortName: Bank of Catalan Neologisms
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona

        - country: Spain
        - telephoneNumber: +34 93 5421207
        - faxNumber: +34 93 5422321
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: AGPL
    - restrictionsOfUse: academic-nonCommercialUse
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)
            - CommunicationInfo:
                - address: Roc Boronat, 138
                - zipCode: 08018
                - city: Barcelona
                - country: Spain
                - telephoneNumber: +34 93 5421207
                - faxNumber: +34 93 5422321
                - email: jorge.vivaldi@upf.edu
                - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- FundingInfo:
    - ProjectInfo:
        - projectName: METANET4U
        - projectShortName: METANET4U
        - fundingType: ownFunds
        - fundingType: euFunds
        - fundingCountry: Spain
        - funder: UPF
        - funder: EU
        - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
    - lingualityType: monolingual
- LanguageInfo:
    - languageCoding: cat
    - languageName: Catalan
- SizeInfo:
    - size: 4622
    - sizeUnitMultiplier: unit
    - sizeUnit: entries
- TextFormatInfo:
    - mime-type: text/xml

- CharacterEncodingInfo:
    - characterEncoding: UTF-8
- DomainInfo:
    - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Spanish Bank of Neologisms at the Observatori de Neologia (UPF). The Observatori de Neologia (OBNEO), under the direction by Dr. M. Teresa Cabré, is a public-funded consolidated group within the Institut Universitari de Lingüística Aplicada at Universitat Pompeu Fabra. This project analyzes the phenomenon of the appearance of new words or neologisms in the usage, both for Catalan and Spanish. Since 1996 has been recognized as a consolidated research group of Universitat Pompeu Fabra.
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: lexicon
- LexicalConceptualResourceCreationInfo:
    - originalSource: Bank of Catalan Neologisms http://obneo.iula.upf.edu/bobneo/index.php
- LexicalConceptualResourceEncodingInfo:
    - encodingLevel: morphology
    - conformanceToStandardsBestPractice: LMF
    - linguisticInformation: lemma
    - linguisticInformation: partOfSpeech

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://www.iula.upf.edu/obneo/obpresca.htm

Bank of Spanish Neologisms

This is the LMF version of the Spanish Bank of Neologisms at the Observatori de Neologia (UPF). The Observatori de Neologia (OBNEO), under the direction by Dr. M. Teresa Cabré, is a public-funded consolidated group within the Institut Universitari de Lingüística Aplicada at Universitat Pompeu Fabra. This project analyzes the phenomenon of the appearance of new words or neologisms in the usage, both for Catalan and Spanish. Since 1996 has been recognized as a consolidated research group of Universitat Pompeu Fabra.

BASIC INFORMATION

IdentificationInfo

- resourceName: Bank of Spanish Neologisms
- resourceShortName: Bank of Spanish Neologisms
- url: http://www.uila.upf.edu

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: Bank of Spanish Neologisms
- resourceShortName: Bank of Spanish Neologisms
- url: http://www.uila.upf.edu

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona

        - country: Spain
        - telephoneNumber: +34 93 5421207
        - faxNumber: +34 93 5422321
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: AGPL
    - restrictionsOfUse: academic-nonCommercialUse
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)
            - CommunicationInfo:
                - address: Roc Boronat, 138
                - zipCode: 08018
                - city: Barcelona
                - country: Spain
                - telephoneNumber: +34 93 5421207
                - faxNumber: +34 93 5422321
                - email: jorge.vivaldi@upf.edu
                - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- FundingInfo:
    - ProjectInfo:
        - projectName: METANET4U
        - projectShortName: METANET4U
        - fundingType: ownFunds
        - fundingType: euFunds
        - fundingCountry: Spain
        - funder: UPF
        - funder: EU
        - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
    - lingualityType: monolingual
- LanguageInfo:
    - languageCoding: spa
    - languageName: Spanish
- SizeInfo:
    - size: 3053
    - sizeUnitMultiplier: unit
    - sizeUnit: entries
- TextFormatInfo:
    - mime-type: text/xml

- CharacterEncodingInfo:
    - characterEncoding: UTF-8
- DomainInfo:
    - domain: general

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Spanish Bank of Neologisms at the Observatori de Neologia (UPF). The Observatori de Neologia (OBNEO), under the direction by Dr. M. Teresa Cabré, is a public-funded consolidated group within the Institut Universitari de Lingüística Aplicada at Universitat Pompeu Fabra. This project analyzes the phenomenon of the appearance of new words or neologisms in the usage, both for Catalan and Spanish. Since 1996 has been recognized as a consolidated research group of Universitat Pompeu Fabra.
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: lexicon
- LexicalConceptualResourceCreationInfo:
    - originalSource: Bank of Spanish Neologisms http://obneo.iula.upf.edu/bobneo/index.php
- LexicalConceptualResourceEncodingInfo:
    - encodingLevel: morphology
    - conformanceToStandardsBestPractice: LMF
    - linguisticInformation: lemma
    - linguisticInformation: partOfSpeech

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://www.iula.upf.edu/obneo/obpresca.htm

Multilingual Vocabulary of Economics

This is the LMF version of the Multilingual Vocabulary of Economics Resource developed within the frame of the research project RICOTERM2. Financed by the Ministry of Science and Technology, within the Programa Nacional de Promoción General del Conocimiento (2004-2007).

BASIC INFORMATION

IdentificationInfo

- resourceName: Multilingual Vocabulary of Economics
- resourceShortName: Multilingual Vocabulary of Economics
- url: http://.es

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: Multilingual Vocabulary of Economics
- resourceShortName: Multilingual Vocabulary of Economics
- url: http://.es

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona

        - country: Spain
        - telephoneNumber: +34 93 5421207
        - faxNumber: +34 93 5422321
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: AGPL
    - restrictionsOfUse: academic-nonCommercialUse
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)
            - CommunicationInfo:
                - address: Roc Boronat, 138
                - zipCode: 08018
                - city: Barcelona
                - country: Spain
                - telephoneNumber: +34 93 5421207
                - faxNumber: +34 93 5422321
                - email: jorge.vivaldi@upf.edu
                - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- FundingInfo:
    - ProjectInfo:
        - projectName: METANET4U
        - projectShortName: METANET4U
        - fundingType: ownFunds
        - fundingType: euFunds
        - fundingCountry: Spain
        - funder: UPF
        - funder: EU
        - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
    - lingualityType: multilingual
- LanguageInfo:
    - languageCoding: spa
    - languageName: Spanish
    - sizePerLanguage:
        - size: 1180
        - sizeUnitMultiplier: unit
        - sizeUnit: entries
- LanguageInfo:
    - languageCoding: eus

    - languageName: Basque
    - sizePerLanguage:
        - size: 1180
        - sizeUnitMultiplier: unit
        - sizeUnit: entries
- LanguageInfo:
    - languageCoding: glg
    - languageName: Galician
    - sizePerLanguage:
        - size: 1180
        - sizeUnitMultiplier: unit
        - sizeUnit: entries
- LanguageInfo:
    - languageCoding: cat
    - languageName: Catalan
    - sizePerLanguage:
        - size: 1180
        - sizeUnitMultiplier: unit
        - sizeUnit: entries
- SizeInfo:
    - size: 5900
    - sizeUnitMultiplier: unit
    - sizeUnit: entries
- TextFormatInfo:
    - mime-type: text/xml
- CharacterEncodingInfo:
    - characterEncoding: UTF-8
- DomainInfo:
    - domain: economy

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Multilingual Vocabulary of Economics Resource developed within the frame of the research project RICOTERM2. Financed by the Ministry of Science and Technology, within the Programa Nacional de Promoción General del Conocimiento (2004-2007).
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: lexicon
- LexicalConceptualResourceCreationInfo:
    - originalSource: Multilingual Vocabulary of Economics from IULA UPF located at http://www.iula.upf.edu/rec/ricoterm/docums/vocecon/eng/index.html
    - creationMode: automatic
- LexicalConceptualResourceEncodingInfo:
    - encodingLevel: morphology
    - conformanceToStandardsBestPractice: LMF
    - linguisticInformation: lemma
    - linguisticInformation: semantics-CrossReferences

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://www.iula.upf.edu/rec/ricoterm/3/index_eng.htm

Basic Vocabulary of Human Genome

This is the LMF version of the Basic Vocabulary of Human Genome from the IULA UPF located at http://www.iula.upf.edu/rec/vbgenoma/esp/index.html The project Vocabulary of the Human Genome (Biotechnology 2), approved in the 2003 plenary meeting of REALITER, incorporates the basic terminology most used in texts about genomics. The vocabulary presents selected entries in English and their equivalents in peninsular and Latin American Spanish, French, Italian, Galician, Portuguese and Catalan. A long with the information on equivalents, the users can find grammatical information and synonyms documented as variants in each of the languages.

BASIC INFORMATION

IdentificationInfo

- resourceName: Basic Vocabulary of Human Genome
- resourceShortName: Basic Vocabulary of Human Genome
- url: http://.es

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: Basic Vocabulary of Human Genome

- resourceShortName: Basic Vocabulary of Human Genome
- url: http://.es

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:

- license: AGPL
- restrictionsOfUse: academic-nonCommercialUse
- distributionAccessMedium: downloadable
- downloadLocation: http://www..com/
- distributor:
  - OrganizationInfo:
    - organizationName: Institut Universitari Lingüística Aplicada - UPF
    - organizationShortName: IULA - UPF
    - departmentName: Institut Universitari Lingüística Aplicada (UPF)
    - CommunicationInfo:
      - address: Roc Boronat, 138
      - zipCode: 08018
      - city: Barcelona
      - country: Spain
      - telephoneNumber: +34 93 5421207
      - faxNumber: +34 93 5422321
      - email: jorge.vivaldi@upf.edu
      - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30
- resourceCreator:
  - OrganizationInfo:
    - organizationName: Institut Universitari Lingüística Aplicada - UPF
    - organizationShortName: IULA - UPF
    - departmentName: Institut Universitari Lingüística Aplicada (UPF)
    - CommunicationInfo:
      - address: Roc Boronat, 138
      - zipCode: 08018
      - city: Barcelona
      - country: Spain
      - telephoneNumber: +34 93 5421207
      - faxNumber: +34 93 5422321
      - email: jorge.vivaldi@upf.edu
      - url: http://www.iula.upf.edu/
- FundingInfo:
  - ProjectInfo:
    - projectName: METANET4U
    - projectShortName: METANET4U
    - fundingType: ownFunds
    - fundingType: euFunds
    - fundingCountry: Spain
    - funder: UPF

        - funder: EU
        - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
    - lingualityType: multilingual
- LanguageInfo:
    - languageCoding: spa
    - languageName: Spanish
    - sizePerLanguage:
        - size: 991
        - sizeUnitMultiplier: unit
        - sizeUnit: entries
- LanguageInfo:
    - languageCoding: eng
    - languageName: English
    - sizePerLanguage:
        - size: 991
        - sizeUnitMultiplier: unit
        - sizeUnit: entries
- LanguageInfo:
    - languageCoding: cat
    - languageName: Catalan
    - sizePerLanguage:
        - size: 847
        - sizeUnitMultiplier: unit
        - sizeUnit: entries
- SizeInfo:
    - size: 3303
    - sizeUnitMultiplier: unit
    - sizeUnit: entries
- TextFormatInfo:
    - mime-type: text/xml
- CharacterEncodingInfo:
    - characterEncoding: UTF-8
- DomainInfo:
    - domain: medicine

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the Basic Vocabulary of Human Genome from the IULA UPF located at http://www.iula.upf.edu/rec/vbgenoma/esp/index.html The project Vocabulary of the Human Genome (Biotechnology 2), approved in the 2003 plenary meeting of REALITER, incorporates the basic terminology most used in texts about genomics. The vocabulary presents selected entries in English and their equivalents in peninsular and Latin American Spanish, French, Italian, Galician, Portuguese and Catalan. A long with the information on equivalents, the users can find grammatical information and synonyms documented as variants in each of the languages.
- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: lexicon
- LexicalConceptualResourceCreationInfo:
    - originalSource: Basic Vocabulary of Human Genome from IULA UPF located at http://www.iula.upf.edu/rec/vbgenoma/esp/index.html
    - creationMode: automatic
- LexicalConceptualResourceEncodingInfo:
    - encodingLevel: morphology
    - conformanceToStandardsBestPractice: LMF
    - linguisticInformation: lemma
    - linguisticInformation: partOfSpeech
    - linguisticInformation: semantics-CrossReferences

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://www.iula.upf.edu/publi076.htm

LMF UPF Term

This is the LMF version of the LMF UPF Term located at http://www.iula.upf.edu/rec/upfterm/cat/index.htm The UPF_TERM terminological data bank was created with the purpose of establishing an electronic resort for the consultation and

diffusion of terminological projects elaborated by students of the Facultat de Traducció i Interpretació, by the IULA and other work and research centres in the Universitat Pompeu Fabra. The UPF_TERM bank, placed in a database server devoted to this end, is free access for students, lecturers and researchers, from the Universitat Pompeu Fabra as well as outside users.
UPF_TERM is the name that receives the set of global resorts of this terminology bank, which is divided in several databases. Each database gathers terminological papers with common features; therefore, the questions and answers can be more productive. Each register of each database has the author, the title and the origin of each project.

BASIC INFORMATION

IdentificationInfo

- resourceName: LMF UPF Term
- resourceShortName: LMF UPF Term
- url: http://www.iula.upf.edu/rec/upfterm/cat/index.htm

ADMINISTRATIVE INFORMATION

IdentificationInfo

- resourceName: LMF UPF Term
- resourceShortName: LMF UPF Term
- url: http://www.iula.upf.edu/rec/upfterm/cat/index.htm

contactPerson

- surname: Vivaldi
- givenName: Jorge
- CommunicationInfo:
    - address: Roc Boronat, 138
    - zipCode: 08018
    - city: Barcelona
    - country: Spain
    - email: jorge.vivaldi@upf.edu
- affiliation:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada (Universitat Pompeu Fabra)
        - organizationShortName: IULA UPF
        - departmentName: Institut Universitari Lingüística Aplicada
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207

        - faxNumber: +34 93 5422321
        - email: jorge.vivaldi@upf.edu
        - url: http://www.iula.upf.edu/

DistributionInfo

- availability: available-unrestrictedUse
- iprHolder:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- LicenseInfo:
    - license: AGPL
    - restrictionsOfUse: academic-nonCommercialUse
    - distributionAccessMedium: downloadable
    - downloadLocation: http://www..com/
    - distributor:
        - OrganizationInfo:
            - organizationName: Institut Universitari Lingüística Aplicada - UPF
            - organizationShortName: IULA - UPF
            - departmentName: Institut Universitari Lingüística Aplicada (UPF)
            - CommunicationInfo:
                - address: Roc Boronat, 138
                - zipCode: 08018
                - city: Barcelona
                - country: Spain
                - telephoneNumber: +34 93 5421207
                - faxNumber: +34 93 5422321
                - email: jorge.vivaldi@upf.edu
                - url: http://www.iula.upf.edu/

TECHNICAL INFORMATION

ResourceCreationInfo

- creationStartDate: 2011-10-10
- creationEndDate: 2011-10-30

- resourceCreator:
    - OrganizationInfo:
        - organizationName: Institut Universitari Lingüística Aplicada - UPF
        - organizationShortName: IULA - UPF
        - departmentName: Institut Universitari Lingüística Aplicada (UPF)
        - CommunicationInfo:
            - address: Roc Boronat, 138
            - zipCode: 08018
            - city: Barcelona
            - country: Spain
            - telephoneNumber: +34 93 5421207
            - faxNumber: +34 93 5422321
            - email: jorge.vivaldi@upf.edu
            - url: http://www.iula.upf.edu/
- FundingInfo:
    - ProjectInfo:
        - projectName: METANET4U
        - projectShortName: METANET4U
        - fundingType: ownFunds
        - fundingType: euFunds
        - fundingCountry: Spain
        - funder: UPF
        - funder: EU
        - url: http://metanet4u.eu/

ValidationInfo

- validated: true
- validationType: formal
- validationMode: automatic
- validationModeDetails: The lexicon validates against the LMF DTD v.16

TextInfo

- LingualityInfo:
    - lingualityType: multilingual
- LanguageInfo:
    - languageCoding: spa
    - languageName: Spanish
- LanguageInfo:
    - languageCoding: cat
    - languageName: Catalan
- LanguageInfo:
    - languageCoding: eng
    - languageName: English
- LanguageInfo:
    - languageCoding: fre

    - languageName: French
- LanguageInfo:
    - languageCoding: lat
    - languageName: Latin
- LanguageInfo:
    - languageCoding: ita
    - languageName: Italian
- LanguageInfo:
    - languageCoding: ger
    - languageName: German
- LanguageInfo:
    - languageCoding: por
    - languageName: Portuguese
- LanguageInfo:
    - languageCoding: rum
    - languageName: Rumanian
- LanguageInfo:
    - languageCoding: swe
    - languageName: Swedish
- LanguageInfo:
    - languageCoding: dut
    - languageName: Dutch
- SizeInfo:
    - sizeUnitMultiplier: unit
    - sizeUnit: entries
- TextFormatInfo:
    - mime-type: text/xml
- CharacterEncodingInfo:
    - characterEncoding: UTF-8

VersionInfo

- version: 1
- lastDateUpdated: 2011-10-30

CONTENT INFORMATION

ContentInfo

- description: This is the LMF version of the LMF UPF Term located at http://www.iula.upf.edu/rec/upfterm/cat/index.htm The UPF_TERM terminological data bank was created with the purpose of establishing an electronic resort for the consultation and diffusion of terminological projects elaborated by students of the Facultat de Traducció i Interpretació, by the IULA and other work and research centres in the Universitat Pompeu Fabra. The UPF_TERM bank, placed in a database server devoted to this end, is free access for students, lecturers and researchers, from the Universitat Pompeu Fabra as well as outside users. UPF_TERM is the name that receives the set of

global resorts of this terminology bank, which is divided in several databases. Each database gathers terminological papers with common features; therefore, the questions and answers can be more productive. Each register of each database has the author, the title and the origin of each project.

- resourceType: lexicalConceptualResource
- mediaType: text

LexicalConceptualResourceInfo

- lexicalConceptualResourceType: lexicon
- LexicalConceptualResourceCreationInfo:
  - originalSource: UPF Term located at http://www.iula.upf.edu/rec/upfterm/cat/index.htm
  - creationMode: automatic
- LexicalConceptualResourceEncodingInfo:
  - encodingLevel: morphology
  - conformanceToStandardsBestPractice: LMF
  - linguisticInformation: lemma
  - linguisticInformation: semantics-CrossReferences

RELEVANT REFERENCES AND OTHER INFORMATION

ResourceDocumentationInfo

- publication: http://www.iula.upf.edu/publi076.htm

6.3 Appendix 3: Quick Validation Report

WP3: Delivery of the BATCH 1 of Resources Validation Report (D3.1)

Document METANET4U-2011-D3.1

EC CIP project #270893

Deliverable

Number: D3.1

Completion: Final

Status: Submitted

Dissemination level: Restricted to project participants

Responsible: Dan Tufiş (WP3 coordinator)

30 November 2011

Colour code:

Corpus received and validated

Corpus not received but validated externally

Exogenous Corpus not received (both restricted and unrestricted)

Endogenous Corpus not received

** [U-EX] = Unrestricted Exogenous Resource*

** [R-EX] = Restricted Exogenous Resource*

Contents

Validation Methodology

This validation report briefly describes the WP3 activities related to the delivery of the first batch of resources.

Before transferring the resources to RACAI for the validation purposes, there have been concluded individual transfer agreements between each partner and RACAI concerning the confidentiality of the respective resources. The model of the individual agreement is shown below:

Addendum to the Consortium Agreement

In addition to the provisions of the Consortium Agreement of the project METANET4U - Enhancing the European Linguistic Infrastructure, number 270893, under the Information and Communication Technology Policy Support Programme (ICT PSP), Competitiveness and Innovation Framework Programme (CIP), in fourth call for proposals, of June 1, 2010, under the Theme 6 Multilingual Web, Objective 6.1 Open Linguistic Infrastructure, with the funding scheme of "Pilot Type B Project", the two partners

PARTNER X (Short name), and
INSTITUTUL DE CERCETARI PENTRU INTELIGENTIA ARTIFICIALA (ICPIA)

Are the parties of the present addendum to the referred Consortium Agreement, and they wish to specify or supplement binding commitments among themselves.

NOW, THEREFORE, IT IS HEREBY AGREED AS FOLLOWS:

For the sole purpose of executing the plan of work of the workpackage WP3:

Clause 1: Partner X pass to ICPIA copies of their language resources or tools.
Clause 2: ICPIA has access to and will use these copies for the sole purpose of producing the quick evaluation check and respective report, required by the execution of WP3.
Clause 3: ICPIA will destroy all the copies of the above mentioned language resources and tools immediately after the execution of the task referred to in clause 2, and gets or retains no rights whatsoever over those resources or tools.

This agreement is signed by the contact persons of Partner X and ICPIA as these are determined in the above mentioned Consortium Agreement :

Partner X
(signature and date)

RACAI
(signature and date)

The validation consisted in verifying if the provided documentation files (Microsoft Word document format) correspond to the corpus files received.

The corpora have many different formats, from xml to text to wav files, containing any number of files.

For corpus validation we have used the following tools/methods:

XML Validation

To validate xml files quickly we have developed an XML validator application that is able to be run from the command line to output xml validation errors. The tool has the following options: a. Validate xml files against an externally provided XSD schema or DTD;
b. Validate xml files loading internal XSD schemas;
c. Validate xml files loading internal DTDs;
d. Check for xml well-formed files if the files do not have any attached DTD or XSD schemas.

Ex: „xmlValidate.exe ApertiumBatch1 d" will validate all xml files in folder ApertiumBatch1, considering only internally referenced DTDs:

```
xmlValidator will not use an external DTD, only internal referenced DTDs.
Source: All 20 xml files in [.]
   File 1 [Apertium-ca-it.ca-it-LMF.xml]:Document is OK.
   File 2 [Apertium-en-ca.en-ca-LMF.xml]:Document is OK.
   File 3 [Apertium-en-es.en-es-LMF.xml]:Document is OK.
   File 4 [Apertium-en-gl.en-gl-LMF.xml]:Document is OK.
   File 5 [Apertium-es-ast.es-ast-LMF.xml]:Document is OK.
   File 6 [Apertium-es-ca.ca-LMF.xml]:Document is OK.
   File 7 [Apertium-es-ca.es-ca-LMF.xml]:Document is OK.
   File 8 [Apertium-es-ca.es-LMF.xml]:Document is OK.
   File 9 [Apertium-es-gl.es-gl-LMF.xml]:Document is OK.
   File 10 [Apertium-es-gl.gl-LMF.xml]:Document is OK.
   File 11 [Apertium-es-pt.es-pt-LMF.xml]:Document is OK.
   File 12 [Apertium-es-ro.es-ro-LMF.xml]:Document is OK.
   File 13 [Apertium-eu-es.eu-es-LMF.xml]:Document is OK.
   File 14 [Apertium-eu-es.eu-LMF.xml]:Document is OK.
   File 15 [Apertium-fr-ca.fr-ca-LMF.xml]:Document is OK.
   File 16 [Apertium-fr-es.fr-es-LMF.xml]:Document is OK.
   File 17 [Apertium-oc-ca.oc-ca-LMF.xml]:Document is OK.
   File 18 [Apertium-oc-es.oc-es-LMF.xml]:Document is OK.
   File 19 [Apertium-pt-ca.pt-ca-LMF.xml]:Document is OK.
   File 20 [Apertium-pt-gl.pt-gl-LMF.xml]:Document is OK.

All files are VALID.
DONE.
```

Linux native utilities

Because of the variety of files to be inspected, we have used standard linux native tools, like sed, grep, cat, tr, sort, wc.

For example, to count the number of tokens in all xml files in a folder, the following command was used:
„cat *.xml | grep –o „<token" | wc –l"

A command to list the number of xml files in all subfolders (recursively) is:

„find . -name "*.xml" | wc –l”

This type of command chaining allowed us be flexible in counting, sorting entitis, text splitting, etc.

Visual inspection of files

The same variety of file formats has required visual inspection of files, especially because in some documentation files the file format is specified and must match the format that acutally exists in that file. This is true for both xml files and text files that adhere to a specific format.

Also, every file that is listed in the documentation file must exist in the respective corpus.

Corpus size is also another value that is inspected visually, by selecting all the files and seeing how much space they occupy.

Unicode validation

The main issue with older corpora is the existence of SGML encodings of characters that could not be represented in standard ASCII format (English character set). However, Unicode allows for native non-English character representation.

As all files should be in Unicode format (UTF representation, usually UTF8 is most commonly used), a tool was developed to search for files encoding SGML entities. The tool was written in c# and takes as input a directory path and outputs all the files that contain SGML entites and the list of the identified entities. Like the xml validation tool presented at item 1, this tool allows for a number of different configuration parameters, like specifying file search patterns, forcing XML special characters match or not (like > or "), etc.

For example, in Romanian, a SGML encoded sentence would look like:

„Afar**ă** plou**ă**.”

while in Unicode format looks like:

„Afar**ă** plou**ă**.”

The tool will identify all SGML entities found in files and mark those files as not-Unicode compliant.

*An addition to this tool is the mapping between SGML entities and Unicode character codes. This allows direct translation of SGML entities to Unicode characters. This addition to the validation tool was very useful for Romanian corpus preparation, allowing fast UTF8 conversion as well as directly inserting custom rules in the mapping table (for example, a useful rule is to change from the old format of the letter **ş** to the new format – SGML entity **Ş** was translated in the Unicode correspondent of **&Scomma;**, etc.)

Overall, the validation process consists of:

a) verifying the existence of files, file structure and occupied space,
b) inspecting for format correspondence with the documentation file.

c) if files are xml they are checked to see if they are:
    1. well-formed (correct xml structure) and
    2. validate against provided or internally referenced XSD/DTD schemas.
d) check every value/measure in the documentation file to see if it corresponds (ex: counting the number of tokens, of sentences, of xml units, etc).
e) finally, files are inspected for UTF8 compatibility.

The partners' resources have been checked for the conformity with their narrative descriptions (according to the procedure described above) in the order they have been transferred to RACAI. All the problems discovered during the validation procedure have been sent to the respective partners. The resources that were transferred earlier have been re-validated and the corrections were acknowledged. The resources that arrived late could not be re-validated, but the responsible partners informed RACAI that all the signalled problems were removed.

Some resources were already validated by external reviewers (e.g. ELRA) and as such they were not sent to RACAI for a new validation.

We should mention that due to the uncertain status of the IPR for the exogenous resources, some of them have not been delivered.

All the partners informed RACAI (responsible for WP3) and UoM (responsible for WP4) that the metadata for the resources subject to the Batch 1 delivery have been uploaded on MetaShare platforms (some on the local repositories and some on the central repositories).

The updated narrative descriptions of the resources delivered in Batch 1 were collected in a separate document (more than 380 pages) uploaded on the intranet server of the project.

In the next sections we present the latest status of the validation (and revalidation) process for each partner. The colour codes used for quick visualisation of the Batch 1 status are the following:

Corpus received and successfully validated
Corpus not received but validated externally
Exogenous Corpus not received (both restricted and unrestricted)
Endogenous Corpus not received

** [U-EX] = Unrestricted Exogenous Resource*
** [R-EX] = Restricted Exogenous Resource*

Partner 1. ULX - University of Lisbon

1.1. DEF Corpus

The corpus presented is a collection of several tutorials and scientific papers in the field of Information Technology with 603 annotated definitions from Portuguese. The texts were collected from the Web at the beginning of the 2006 and they are organized in three folders with 32 files of three different sub-domains with 268,064 tokens: Information Society (91,825), Information Technology (80,483), and e-Learning (94,756).

There are 12 files in the E-Learning subfolder, 12 files in the IS subfolder and 7 files in the IT subfolder.

All files validate against the provided DTD file.

The corpus contains a documentation file that is accurate.

1.2. PF Corpus (transcribed speech)

This corpus contains 137 conversations in wav format and associated files:

1. Wav File, 16bit stereo PCM encoding, 44KHz.
2. Exb file that align sound with the transcriptions using EXMARaLDA format (xml format)
3. Txt file with plain-text transcription
4. POS tagged file of the transcription (source – plain text file)

The corpus size is 4.62GB. The exb files have been tested for xml structure and are well-formed.

The corpus contains a documentation file that is accurate.

1.3. Multifunctional Computational Lexicon of Contemporary Portuguese – CORLEX

This corpus is a lexicon extracted from CORLEX. It is composed of 2 text files and 38 pdf files. The two text files contain lemmas and surface word forms sorted alphabetically and by frequency, respectively. The pdf files contain a "graphical" representation of the frequency of the lemmas and word forms.

Example from the txt files:

@ abalizado (A) # 8
abalizada (A) # 6
abalizadas (A) # 1
abalizado (A) # 1

@ abalo (N) # 96
abalo (N) # 59
abalos (N) # 37

The corpus contains a documentation file that is accurate.

1.4. Spoken Portuguese

The corpus contains 86 recordings in the same format as corpus 3 (PF Corpus). For each recording the following files exist:

1. Wav File, 16bit stereo PCM encoding, 44KHz.
2. Exb file that align sound with the transcriptions using EXMARaLDA format (xml format)
3. Txt file with plain-text transcription
4. POS tagged file of the transcription (source – plain text file)

The exb files have been tested for xml structure and are well-formed.

The corpus contains a documentation file that is accurate.

1.5. Corpus NILC [R-EX]

Not received.

1.6. CorpusTCC [R-EX]

Not received.

1.7. PLN-BR Gold (corpus) [R-EX]

Not received.

1.8. NILC Taggers (grammar, training models for tagger) [R-EX]

Not received.

Partner 2. IST

2.1. CorpusNE

The folder contains 2 text files, train.txt and test.txt. The files contain questions with annotated Named Entities (categories: Person, Location, Organization). We have counted 3092 annotated entities.

The corpus contains a documentation file that is accurate.

2.2. CorpusParalelo

The folder contains 4 text files, representing a training-test parallel corpus (English-Portuguese). The training files contain 5457 sentences and the test files contain 500 sentences.

The format for the files is the following:

DESC:manner How can I find a list of celebrities ' real names ?
ENTY:animal What fowl grabs the spotlight after the Chinese Year of the Monkey ?
ABBR:exp What is the full form of .com ?

The corpus contains a documentation file that is accurate, with the small exception that in point 3.3 it is said that the test files contain 499 questions, and we have found 500.

Partner 3. UNIMAN - University of Manchester

3.1. BioLexicon GREC

BioLexocon was a collaborative work between a number of partners on the BOOTStrep project, and since it was difficult to obtain permission from all parties in the creation of the lexicon, UNIMAN did its own validation, as written in the documentation file.

3.2. SemLink Resources [R-EX]

SemLink provides a mapping between complementary lexical resources: VerbNet-PropBank and VerbNet – FrameNet.

The corpus is split in two folders, each containing one of the mappings.

Folder vn-fn contains 3 files, out of which 2 are xml. One file, *VN-FN_roleMapping.xml* does not validate :

```
VN-FN_roleMapping.xml:2412: element roles: validity error : Element roles content does not follow the DTD,
expecting (role)+, got ()
   </roles>
         ^
VN-FN_roleMapping.xml:3646: element roles: validity error : Element roles content does not follow the DTD,
expecting (role)+, got ()
   </roles>
```

Folder vn-pb contains 5 files, out of which one is a well-formed xml file.

The corpus contains a documentation file that is accurate. The parsing problem has been reported to the author (this is an exogenous restricted resource).

3.3. GENIA [U-EX]

The corpus consists of 2,000 MEDLINE abstracts.

There are 8 files present, out of which 3 represent the corpus: a POS file, the xml POS file and a merged format. All the files are accurately described in the documentation file.

The xml files validate against the provided DTD.

The corpus contains a documentation file that is accurate.

3.4. GENIA Meta-knowledge[U-EX]

The corpus consists of 1000 MEDLINE abstracts, having 1000 xml files.

The xml files validate against the provided DTD.

The corpus contains a documentation file that is accurate.

3.5. GREC[U-EX]

The corpus consists of 240 MEDLINE abstracts, out of which 167 regard *E. Coli* (/Ecoli), and 73 regard Humans (/Human).

The corpus is present in txt and xml formats.

For the txt versions, each abstract is accompanied by two files with the extensions .a1 and .a2. (annotation files)

For the xml version of the corpus, the abstracts are simple xml files. They validate against the provided DTD.

The corpus contains a documentation file that is accurate.

Partner 4. UAIC - University Alexandru Ioan Cuza

4.1. 1984_NP

This folder is based on the previous folder, containing two files, NP-annotated.

Both files are well-formed.

The folder contains a documentation file that is accurate.

4.2. 1984AnaphoraRo

The folder contains 2 xml files: 1984.xml and 1984_RARE.xml. Both are well-formed xml files. The first file does not contain a schema, so cannot be validated against a standard. The structure looks like:

<DOCUMENT><NP ID="NP0" HEADID="TOK0"><W ID="W0" root="parte" pv="Noun" Type="common" Gender="feminine" Number="singular" Definiteness="yes" RO="TOK0">PARTEA</W><W ID="W1" root="ÎNTÂI" pv="Numeral" Type="ordinal" Form="letter" Definiteness="no" RO="TOK1">ÎNTÂI</W></NP><W ID="W2" root="1" pv="Numeral" Type="cardinal" Number="singular" Form="digit" RO="DIG0">1</W><W ID="W3" root="întru" pv="Adposition" Type="preposition" Formation="simple" RO="LSPLIT0">Într-</W>

The folder contains a documentation file that is accurate.

4.3. FrRoMWE

The corpus consists of the French and Romanian version of the novel “Madame Bovary”. It contains three folders, one for each part of the novel. Each folder contains xml and text files. The xml files contain translation units. The text files contain references to all the multi-word expressions in both Romanian and French. There is also a word-level alignment file.

All the xml files have been tested and are well-formed (they do not have a schema).

The corpus contains a documentation file that is accurate.

4.4. QA-corpus-UAIC

The Question-Answering corpus contains a single xml file containing 200 questions in Romanian.

The file is a well-formed xml.

The folder contains a documentation file that is accurate.

4.5. RO-FDGBank

The corpus contains 4 balanced sets of annotated Romanian sentences in 4 subfolders.

File 1.xml from /Aquis Comunitar/Wikipedia+Aquis is identical (duplicate) of /Wikipedia/ Wikipedia+Aquis.

In total the corpus contains 28 files (counting the duplicate file as one file only), all being well-formed xml files.

The folder contains a documentation file that is accurate.

4.6. RO-FN

The corpus contains multilingual (English and Romanian) semantic role annotations, structured in one xml file.

The xml file is well-formed (no referenced schema).

The folder contains a documentation file that is accurate.

4.7. RoSemClass

The folder contains 3 xml files about semantic classes of lexical items for political discourse analysis:

1. The main file “classes_eminescu.xml”, with the structure

```
<lexic name="Eminescu" lang="ro">
<word form="vor" lemma="vrea" freq="2398" classes="14" />
<word form="RomÃ¢nia" lemma="RomÃ¢nia" freq="1759" classes="2" />
<word form="oameni" lemma="om" freq="1700" classes="5" />
<word form="statului" lemma="stat" freq="1330" classes="2" />
```

2. The class hierarchy “semantic_classes_hierarchy.xml”

```
<classes>
<class name="swear" id="1"/>
<class name="social" id="2"/>
<class name="family" id="3" parent="2"/>
```

3. The lexicon “lexicon Daniela Gifu.xml”

```
<lexic name="Public" lang="ro">
<word stem="ï»¿acasÄƒ*" classes="30,7,27"/>
<word lemma="complot" classes="30,10"/>
<word lemma="apatrid" classes="30,1,5,25"/>
<word stem="apatrizi*" classes="30,1,5,25"/>
<word stem="hitler*" classes="30,1,5"/>
```

The folder contains a documentation file that is accurate.

4.8. TE-pairs

This Textual Entailment folder contains one xml file, containing 20 groups of 10 hypotheses each, having the format:

```
<entailment-group id_group="1">
<text> Mi-e dor de Ştefan Iordache. Golul rămas prin plecarea lui e uriaş. A fost atât de important el, lucru ştiut şi în viaţă, dar mai ales acum când nu mai e; eu raportez tot ce mi se întâmplă la acest gol. Radu Beligan, parcimonios foarte în aprecieri, dar care nu se joacă însă cu valoarea, spune despre el: „Ştiu că oamenii sunt reticenţi la auzul unei declaraţii absolute şi totuşi nu ezit să spun că îl consider pe Ştefan Iordache cel mai mare actor al unei generaţii de talente uimitoare, care apare o dată la 50 de ani.
</text>
<hypothesis id_hypothesis="1" entailment="Yes">Ştefan Iordache a murit.</hypothesis>
<hypothesis id_hypothesis="2" entailment="Yes">Ştefan Iordache a fost actor.</hypothesis>
```

The xml file is well formed.

The folder contains a documentation file that is accurate.

4.9. TE-rules

The TE-Rules folder contains one xml file, encoding 20 entailment rules in the format:

```
<rule id="1" type="DIRT">
    <T>
      <node id="1" tag="V" lemma="var1"/>
      <node id="2" parent="1" relation="rel1"/>
      <node id="3" parent="1" relation="rel2"/>
    </T>
    <H>
      <node id="1" tag="V" lemma="var2"/>
      <node id="2" parent="1" relation="rel1"/>
      <node id="3" parent="1" relation="rel2"/>
    </H>
    <value type="decimal">DIRT.similarity(var1,var2)</value>
  </rule>
```

The xml file is well formed.

The folder contains a documentation file that is accurate.

Partner 5. RACAI - Romanian Academy

5.1. Multilingual News Corpus

The corpus consists of news in English, French and Romanian. There are 1847 xml files representing articles for each of the languages.

The format of the files is XCES:

<xces:p id="p15"><xces:s id="s634472976784083750_en_15"><xces:tok base="""" msd="DBLQ" type="punctuation">"</xces:tok><xces:tok base="le" msd="Ncnp;Np#1" type="word">Les</xces:tok><xces:tok base="(0.57)neig" msd="Vmip3s;Vp#1" type="word">neiges</xces:tok><xces:tok base="(WF)du" msd="Afp;Vp#1,Np#2,Ap#1" type="word">du</xces:tok><xces:tok base="Kilimandjaro" msd="Np;Np#2" type="word">Kilimandjaro</xces:tok>

All files validate against the provided xsd schema.

The corpus contains a documentation file that is accurate.

5.2. RO-Acquis

The corpus consists of the Romanian version of the Acquis Communautaire, the common set of laws of the European Union member states. There are 10704 documents in which 34234437 tokens occur. Out of these, 27968652 are words and the rest, punctuation.

The corpus has 42 sub-folders containing xml files. All files validate against the referenced schema.

The corpus contains a documentation file that is accurate.

5.3. Romanian Balanced Corpus

The corpus consists of equal shares of texts from 5 different genres: journalism, legalese, fiction, medicine and biographical data for Romanian literary personalities.

All folders contain validated XCES compliant xml files.

The Agenda folder contains 10294016 tokens, including punctuation.

The EMEA folder contains 10950271 tokens, including punctuation.

The JRC folder contains 9067516 tokens, including punctuation.

The DGLR folder contains 5802961 tokens, including punctuation.

The Literatura folder contains 8002596 tokens, including punctuation.

The corpus contains a documentation file that is accurate.

5.4. RO-SemCor

SemCor En-Ro corpus is an English-Romanian parallel corpus which was developed starting from the English SemCor.

The folder contains two xml files: SemCor30-ro-xces.xml and SemCor30-en-xces.xml. Both validate against the provided schema.

The corpus contains 354,102 tokens (including punctuation): 178,499 for English and 175,603 for Romanian.

The corpus contains a documentation file that is accurate.

5.5. RO-WordNet (first version)

Ro-WordNet (RWN) is a lexical ontology following the Princeton WordNet (PWN) organizational principles. The synsets in RWN are aligned with PWN3.0 and, additionally, they are associated with SUMO/MILO concepts.

One xml file is present, containing 30006 synsets with the format:

<SYNSET><ID>ENG30-15032376-n</ID><POS>n</POS><SYNONYM><LITERAL>otravă<SENSE>1</SENSE></LITERAL></SYNONYM><DEF>Substanţă chimică toxică, care, introdusă sau formată în organism, provoacă tulburări importante, leziuni grave etc. şi uneori moartea.</DEF><BCS>1</BCS><ILR>ENG30-02754756-n<TYPE>hyponym</TYPE></ILR><ILR>ENG30-03554795-n<TYPE>hyponym</TYPE></ILR><ILR>ENG30-15034074-n<TYPE>hyponym</TYPE></ILR><DOMAIN>chemistry</DOMAIN><SUMO>BiologicallyActiveSubstance<TYPE>+</TYPE></SUMO></SYNSET>

The corpus contains a documentation file that is accurate.

5.6. WEB-DEX

WEB-DEX is an explanatory dictionary based on the 1996 edition of the standard explanatory dictionary of Romanian published by the Romanian Academy.

It contains an xml file (45MB), with the entry format as:

```
<entry type="homonym" id="A.4">
  <hw>A</hw>
  <orth>A</orth>
  <pos>prepoziţie</pos>
  <struc>
    <usg>Formează numerale distributive</usg>
    <def>
     De câte...
    </def>
    <struc type="phrase">
      <orth>3 saci a 80 de kg. </orth>
    </struc>
  </struc>
  <etym>
   Din limba
    <lang>fr.</lang>
   á.
  </etym>
</entry>
```

The corpus contains a documentation file that is accurate.

5.7. Wordform lexicons

This is a wordform lexicon containing statistical information extracted from the Romanian Balanced Corpus. The lexicon is a flat file, one entry per line, fields being tab separated (text format):

gemeni		geamăn	Afpmp-n	17	
gemeni		geamăn	Ncmp-n		52
gemenii	geamăn	Ncmpry		88	

The file size is 3.1MB, containing 111462 entries.

The corpus contains a documentation file that is accurate.

5.8.Multilingual Subjectivity Analysis: Gold Standard & Training Data [U-EX]

The corpus consists of a single Excel (.xls) document that contains a sheet with 1590 lines. Each line is a language quotation (reported speech) manually annotated for sentiment towards entities mentioned inside it.

The corpus contains a documentation file that is accurate.

5.9. TimeBank parallel corpus [U-EX]

The corpus consists of Romanian news texts and English-Romanian aligned files. There are 183 Romanian files in /ro folder, 181 annotated xml parallel files in /en-ro-msd with the 181 alignment files in text format in /align.

The corpus contains well-formed xml files.

The corpus contains a documentation file.

5.10. RO-SAM EUROM 0,5 (spoken Romanian) [U-EX]

The corpus is part of the „Multext-East multilingual corpus of text and speech data" that covers six languages: Bulgarian, Czech, Estonian, Hungarian, Romanian, and Sloven. The audio data is stored in separate WAVE files, which are PCM encoded at 22Khz, 16 bits per sample, mono.

There are 40 Romanian speech wav files.

The corpus also contains an xml file that lists all spoken sentences (well-formed xml) (*spch-ro.xml*).

The corpus contains a documentation file that is accurate.

Partner 6. UOM - University of Malta

6.1. F_MONA_1_Maltese Spoken Newspapers

Not received.

6.2. Laws of Malta

The corpus contains raw text files representing Maltese laws extracted from the government website, both in English and Maltese.

There is a documentation file but it contains duplicate information, one time for the English subfolder (containing laws in English) and one time for Maltese subfolder (containing laws in Maltese). It should contain just a single file discussing both subfolders, otherwise, the folders should be split in two separate corpora, each with its own documentation file.

We have counted 3135 files in the English subfolder and 3140 files in the Maltese subfolders, an information that is not present in the documentation file.

6.3. Maltese Acquis Communautaire EN

The standard Aquis Corpus, English monolingual version. There are 47 folders, each folder representing one year, containing a variable number of valid xml files.

The folder contains a documentation file that is accurate, with the minor error that there are only 23547 files (not 23551 as specified at item 3.1), excluding supporting xml, html and dtd files (5 other files).

6.4. Maltese Acquis Communautaire MT

The standard Aquis Corpus, Maltese monolingual version. There are 48 folders, each folder representing one year, containing a variable number of valid xml files.

The folder contains a documentation file that is accurate, with the minor error that there are only 10547 files (not 10551 as specified at item 3.1), excluding supporting xml, html and dtd files (5 other files).

6.5. Maltese Wordlist

The folder contains a text file consisting of Maltese words, one per line. There are 824839 words.

The folder contains a documentation file that is accurate.

6.6. Basic English-Maltese Dictionary [U-EX]

The corpus consists of Bilingual wordlist, consisting of alphabetically ordered English lemmas with their Maltese translation and Maltese pronunciation. There is a single xml file containing the bilingual word list, with the format:

```
<entry>
        <form>
                <orth>FOOTBALL</orth>
        </form>
        <sense>
```

```
                                        <cit xml:lang="mt">
                                                <quote>futbol</quote>
                                                <gramGrp>
                                                        <pos>n</pos>
                                                        <gen>M</gen>
                                                </gramGrp>
                                                <pron>fut'bol</pron>
                                        </cit>
                                </sense>
                        </entry>
```

The xml fails validation with :

File 1 [Basic Dictionary.xml]:

ERROR: The element 'name' has invalid child element 'email'. List of possible elements expected: 'abbr address date dateRange dateStruct expan geogName lang measure name num orgName persName placeName rs time timeRange timeStruct add app corr damage del orig reg restore sic space supplied unclear oRef oVar pRef pVar formula handShift distinct emph foreign gloss hi mentioned soCalled term title ptr ref xptr xref caesura c cl m phr s seg w att gi tag val anchor addSpan delSpangap figure alt altGrp certainty fLib fs fsLib fvLib index interp interpGrp joinjoinGrp link linkGrp respons span spanGrp timeline cb fw lb milestone pb'.

ERROR: The 'email' element is not declared.

as well as with other element definition errors.

The parsing error has been communicated to the owners who will remove it. The corpus contains a documentation file that is accurate.

6.8. Illum Corpus [U-EX]

The corpus consists of the full editions of ILLUM from 12/11/2006 to 30/05/2010 (185 issues).

There are 5267 well-formed xml files present, having the format:

```
<text>
 <paragraph>
                        Pajjiż
                        demokratiku
                        ma
                        jfittixx
                        biss
                        dak
                        li
                        jaqbel
                        għall-maġġoranza
                        iżda
                        jħares
                        ukoll
                        id-drittijiet
                        tal-minoranzi
                        .
        </paragraph>
```

We have counted 60167 paragraphs.

The corpus contains a documentation file that is accurate.

Partner 7. UPC - Technical University of Catalonia

7.1. AGORA

Validated by ELRA.

7.2. Bilingual Speech synthesis

Validated by ELRA.

7.3. CatalanBN

Validated by ELRA.

7.4. Catalan-SpeechDat (dialects)

Validated by ELRA.

7.5. Spanish EUROM.1

Validated by ELRA.

7.6. FESTCAT (Catalan)

Validated by ELRA.

7.7. FESTCAT-SEL (Catalan)

Validated by ELRA.

7.8. FREE-SPEECH (Catalan)

Validated by ELRA.

7.9. LC-STAR Dialogues (Spanish, Catalan)

Validated by ELRA.

7.10. Spanish Festival models

Validated by ELRA.

7.11. Spanish Festival voices

Validated by ELRA.

7.12. SpeechDat-Car Catalan

Validated by ELRA.

Partner 8. UPF - University Pompeu Fabra

8.1. Basic Vocabulary on the Human Genome

In the received documentation the following link is specified:
http://www.iula.upf.edu/rec/vbgenoma/esp/index.html

This address points to a web application that allows access to the resources. They have been randomly tested for correctness.

8.2. Multilingual Vocabulary of Economics

In the received documentation the following link is specified:
http://www.iula.upf.edu/rec/ricoterm/docums/vocecon/eng/index.html

This address points to a web application that allows access to the resources. They have been randomly tested for correctness.

8.3. Neologisms of the year: Bank of Spanish & Catalan Neologisms

In the received documentation the following link is specified:
http://obneo.iula.upf.edu/bobneo/index.php

This address points to a web application that allows access to the resources. They have been randomly tested for correctness.

8.4. UPF_Term

In the received documentation the following link is specified:
http://www.iula.upf.edu/rec/upfterm/cat/index.htm

This address points to a web application that allows access to the resources. They have been randomly tested for correctness.

8.5. Corpus PAAU 92

- ✓ The documentation file is present and correct.
- × The files, even though they are valid xml documents (tested for structure correctness without schema) fail xml validation:
  g00160-pos.xml:2: validity error : Validation failed: no DTD found !
  <graph xmlns="http://www.xces.org/ns/GrAF/1.0/">

  The second line in each file defines a namespace but no schema location, so xml validators fail.

8.6. Genoma corpus (both catalan and spanish)

Catalan:

- ✓ The documentation file is present and correct.
- × Out of the 133 documents, document m00913 is empty (a zero-bytes file).

- × Valid xml structure but failing xml validation similar to corpus92 (error: no DTD found)

Spanish:

- ✓ The documentation file is present and correct.
- × Valid xml structure but failing xml validation similar to corpus92 (error: no DTD found)

8.7. PAROLE lexicon (Spanish) [R-EX]

- ✓ For the lexicon file there is an xml file and a html file that describe it.

This folder contains one xml file (Parole-LMF.xml) and a DTD (DTD_LMF_REV_16.dtd).

The file has the following format:

<LexicalEntry id="UMN003" morphologicalPatterns="MFGN05"><feat att="id" val="UMN003"/><feat att="gramcat" val="NOUN"/><feat att="gramsubcat" val="COMMON"/><feat att="autonomy" val="YES"/><feat att="synulist" val="USABADA"/><Lemma><feat att="writtenForm" val="abada"/></Lemma>
</LexicalEntry>

The file contains 64594 Lexical Entries.

The file passes validation against the provided DTD.

8.8. SIMPLE lexicon (Spanish) [R-EX]

The LMF version of the Spanish ParoleSimple lexicon has 7698 entries covering 9861 semantic units. It is UTF-8 encoded according to LMF specifications.

- ✓ For the Simple lexicon file there is an xml file and a html file that describe it.

This folder contains one xml file (ParoleSimple-LMF.xml) and a DTD (DTD_LMF_REV_16.dtd).

The file has the following format:

<LexicalEntry id="UMN001" morphologicalPatterns="MFGN05">
<feat att="id" val="UMN001"/>
<feat att="gramcat" val="NOUN"/>
<feat att="gramsubcat" val="COMMON"/>
<feat att="autonomy" val="YES"/>
<feat att="synulist" val="USABABA"/>
<Lemma><feat att="writtenForm" val="ababa"/></Lemma>
</LexicalEntry>

The file passes validation against the provided DTD.

8.9. Apertium Bilingual dictionaries [U-EX] (14): (we have found 16 dictionaries, not 14!)

Basque-Spanish, Catalan-Spanish, English-Catalan, English-Galician, English-Spanish, French-Catalan, French-Spanish, Occitan-Catalan, Occitan-Spanish,

Portuguese-Catalan, Portuguese-Galician, Spanish-Asturian, Spanish-Galician, Spanish-Portuguese, Spanish-Romanian

We found: ca-it, en-ca, en-es, en-gl, es-ast, es-ca, es-gl, es-pt, es-ro, eu-es, fr-ca, fr-es, oc-ca, oc-es, pt-ca, pt-gl.

- ✓ For every dictionary there is an xml file containing meta-data documentation.

This folder contains 20 xml files and a DTD. The files are bilingual lexicons in the LMF format output by Apertium.

The files have the following format:

```
<LexicalEntry id="id57792-l">
        <feat att="partOfSpeech" val="cnjadv"/>
        <Lemma>
                <feat att="writtenForm" val="amb la finalitat que"/>
        </Lemma>
        <Sense id="id57792-l-s"/>
</LexicalEntry>
```

In total, they contain 672258 Lexical Entries, and 580618 sense entries.

All files pass validation against provided DTD.

8.10. Apertium monolingual dictionaries [U-EX] (4): Basque, Catalan, Galician, Spanish

These files are included in 8.9 in a single folder, with the same mentions.

8.11. FreeLing dictionaries [U-EX] (6):

Asturian dictionary, Catalan dictionary, Catalan sense dictionary, Galician dictionary, Spanish dictionary, Spanish sense dictionary

- ✓ For every corpus file there is an xml file and detailed meta-data documentation.

diccAS_freeling2LMF.xml – has 40048 lemmas, 157731 word forms
diccES_freeling2LMF.xml – has 76318 lemmas, 669121 word forms
diccCA_freeling2LMF.xml – has 71862 lemmas, 878126 word forms
diccGL_freeling2LMF.xml – has 49898 lemmas, 577235 word forms
diccES-Sense_freeling2LMF.xml – has 6213 entries, 17359 senses
diccCA-Sense_freeling2LMF.xml – has 43561 entries, 65380 senses