METANET4U

Report on first selection of resources

Deliverable D2.1

Version 2.4

2011-03-31

Editor: Asunción Moreno

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
V1.0	15 Feb 2011	A. Moreno	UPC	First draft
V2.0	20 Feb 2011	A. Moreno	UPC	1st contribution from partners
V2.1	27 Feb 2011	A. Moreno	UPC	Updates on descriptions of LR
V2.2	8 March 2011	A. Moreno	UPC	Pre-final version
V2.3	15 March 2011	A. Moreno	UPC	Internal quality check review
V2.4	31 March 2011	A. Moreno	UPC	Final version

METANET4U

Report on first selection of resources

Document METANET4U-2011-D2.1
EC CIP project #270893

Deliverable
Number: D2.1
Completion: Final
Status: Submitted
Dissemination level: Restricted to project participants

Responsible: Asunción Moreno (WP2 coordinator)

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

Authors: António Branco, Amalia Mendes, Isabel Trancoso, Hugo Meinedo, Sophia Ananiadou, Paul Thompson, John McNaught, Dan Cristea, Diana Trandabat, Dan Tufis, Mike Rosner, Asunción Moreno, Núria Bel, Jorge Vivaldi, Eva Revilla

Reviewer: Paul Thompson

Contents

1 Introduction

During the preparation of the METANET4U proposal, partners elaborated a list of language resources, including both data and software, which could be made available for the purposes of the project. The data was prepared by all partners and was included in Appendix A of the Annex 1 of the contract.

The objective of this *Deliverable D2.1 Report on first seletion of resources* is to select the resources that will be prepared in WP3 and delivered in month M10 as a first batch of resources (Batch 1). The remaining language resources will be delivered either in the second batch during month 18 (M18) (Batch 2), or in the third batch during month 24 (M24) (Batch 3). The decision of whether to deliver in Batch2 or Batch 3 will be taken in Task 2.3 in deliverables 2.4 and 2.5.

This deliverable is organized as follows: section 2 explains the way in which language resources are described and categorised, while section 3 shows, for each partner, the resources that will be delivered in Batch 1 and the resources that will be delivered in subsequent batches. Section 4 contains a list of the language resources that, at this stage of the project, partners are willing to deliver. This list is an updated version of the original list of the Annex 1 of the contract.

2 Language resource descriptors

The main purpose of the list of resources is to identify, assess and select language resources and tools (LR&T) that could be interesting for the project. The identification and selection of LR&T has been carried out by taking into account their potential utility, according to the following key features: multilinguality, easy to attract users, popular or well known LR, fitness to purpose, IPR clear between the owner institutions, perennity, and maturity. The following types of resources have been considered as top priority: parallel corpora (raw, aligned, annotated,...), large monolingual corpora, tools (lemmatizers, tokenizers, irrespective of compatibility with datasets), and bilingual lexica.

A first grouping of the LR&T is as Data and Software:

Data
This group includes among others: lexica, wordnets, thesauri, annotated corpora, parallel corpora, speech recognition databases and speech synthesis databases. The data sets listed in section 4 include comments on: whether or not they cover a complete range of data **types** (e.g., from POS tagged corpora to propbanks); identification of possible gaps in that range; the type or types of **licences** under which they are being made available (where appropriate); etc.

Software and tools
This group includes, among others, language identifiers, hyphenizers, tokenizers, lemmatizers, sentence splitters, pos taggers, NP-chunkers, parsers, semantic role

labellers, summarisers, word aligners, lexicon editors, linguistic web services, workflow platforms, grapheme to phoneme converters, etc. The software and tools listed in section 4 include information and comments on: the **language**(s) or dialect(s) dealt with; whether they can be hooked together into pipelines; whether or not they cover a complete range of functionalities (e.g., from tokenization to deep linguistic representation); identification of possible gaps in that range; the type or types of **licences** under which they are being made available (where appropriate); etc.

The identified LR&T have been grouped, depending of the conditions under which they are being made available to the consortium as:

Endogenous resources
These resources are owned or controlled by the relevant partner, do not depend on third parties to be released and will be made available at the appropriate milestone of this project.

Restricted exogenous resources (restricted availability)
These resources are not owned or controlled by the relevant partner, are not freely available, and depend on third parties to be released and for which arrangements will be sought to possibly make them available.

Unrestricted exogenous resources (*un*restricted availability)
These resources are not owned or controlled by the relevant partner, are freely available, and do not depend on third parties to be released/distributed.

Each resource to be delivered (data or software, endogenous or exogenous, restricted or unrestricted) needs to be documented. A specific number of resources will be cleaned and upgraded according to linguistic & interoperability standards. For each of these selected resources, a **rationale** of why the resource has to be **upgraded** (improve documentation, remove bugs and inconsistencies, improve portability…) or **extended** (improve coverage, increase suitability for research and development, bring together to create cross-lingual resources…) is included

3 Selection of resources

The main purpose of the selection of resources is to make a work distribution proposal for WP3, and to choose expected dates for the delivery of language resources, software and tools. For this purpose, the following aspects were taken into account:

Given that most of our project effort is concentrated on gathering, preparing and enhancing LRs, overall workload will be heavily influenced by how we chose the LRs,, and which are chosen. Therefore it is important to have a sensible perspective in this respect from the offset.

The LRs are ultimately gathered to be distributed through the Meta-Share platform, and Meta-Share will have two versions:

- in v1 (due by July 2011), which will be capable of distributing datasets;
- in v2 (due by Feb 2012), which will have the additional capabilities of distributing software tools (and possibly web services). This automatically introduces a prioritisation of the resources to be made available.

V2 of Meta-Share (but not v1) will have a billing system: this also introduces also a prioritisation of the resources to be made available, since those that will be distributed for a fee can be only released when v2 is available.

Whilst endogenous resources belong to partners, exogenous resources require more time and effort to make them available, e.g., certain resources are restricted and their owners need to be contacted, licenses need to be obtained, etc. Even for the unrestricted exogenous resources, time is required in order to be gather them. Additionally, the owners will be more easily convinced to deliver their resources if they see Meta-Share working. Therefore, given that only version v2 will be "ready for dissemination outside the Meta-Net network", this will put a serious constraint on which exogenous resources we will be targeting and able to attract before v2 is ready and running (due by Feb 2012).

Last but not least, even for endogenous resources, it is very likely that not all are immediately ready to be distributed, as some of them are indicated as needing to undergo considerable enhancements, etc. This circumstance also needs to be taken into account by the different project partners, in order to prioritize the resources that they plan to deliver.

The first Metanet4u batch of resources (Batch1) is expected by M10. By that time, only Meta-Share platform v1 will be available. Consequently, most of the resources delivered will consist of Endogenous data. The second and third Metanet4u batches of resources are expected for M18 and M24. By that time, Meta-Share platform v2 will be available and Exogenous resources as well as software can be easily managed.

The following tables show, for each partner, the resources they plan to deliver in Batch 1 or in Batches 2 and 3, as well as the effort in PMs for each resource. Please note that the total effort per partner is the 80% of the total effort to be devoted in WP3. There is a remaining 20% effort to be applied, possibly on other non-listed resources, during the second or third batch.

Partner	Data/Soft	Where	Delivery	Designation	Total pm
1. ULX - University of Lisbon	Data	Endogenous resources	Batch 1	DEF Corpus	0,15
				Multifunctional Computational Lexicon of Contemporary Portuguese - CORLEX	0,5
				PF Corpus	8
				Spoken Portuguese	8
			Total Batch 1		16,65
			Batch 2/3	Abbreviations	0,1
				CINTIL-Internacional Corpus of Portuguese	5
				C-ORAL-ROM Portuguese Corpus	6
				DependencyBank	12
				MWN.PT	0,15
				PAROLE corpus	0,2
				PAROLE lexicon	0,2
				PropBank	12
				SIMPLE lexicon	0,2
				Stopwords	0,1
				Technical Corpus	8
				Treebank	12
			Total Batch 2/3		55,95
		Total Endogenous resources			72,6
		Restricted exogenous resources	Batch 2/3	Clássicos LP/Porto Editora	0,15
				COMPARA	0,15
				Corpus NILC	0,15
				Dicionário de Verbos do Português Medieval (DVPM)	0,2
				European Parliament Parallel Corpus	0,15
				Geo-Net-PT01	0,15
				Glossário	0,15
				MorDebe	0,15
				Norma Urbana Culta (NURC)	0,15
				Ontologia de Nanociência e Nanotecnologia	0,15
				Panorama do Português Oral de Maputo (PPOM)	0,15
				PORLEX	0,15
				The JRC-Acquis Multilingual Parallel Corpus	0,15
			Total Batch 2/3		2
		Total Restricted exogenous resources			2
		Unrestricted exogenous resources	Batch 1	Corpus NILC	0,15
				CorpusTCC	0,15
				NILC Taggers	0,15
				PLN-BR Gold	0,15
				RHETALHO	0,15
				Summ-it	0,15
				TeMário 2006	0,15
			Total Batch 1		1,05
			Batch 2/3	CETEMPúblico	0,15
			Total Batch 2/3		0,15
		Total Unrestricted exogenous resources			1,2
	Total Data				75,8
	Software	Endogenous	Batch 2/3	Chunker	0,1

		resources		Tagger	0,1
				Tokenizer	0,1
			Total Batch 2/3		0,3
		Total Endogenous resources			0,3
		Unrestricted exogenous resources	Batch 2/3	DiZer 2.0	0,15
				Forma	0,15
				GistSumm	0,15
				Ontolp Plugin	0,15
				Stemmer	0,15
				Text Aligners	0,15
			Total Batch 2/3		0,9
		Total Unrestricted exogenous resources			0,9
	Total Software				1,2
Total 1. ULX - University of Lisbon					77

Partner	Data/Soft	Where	Delivery	Designation	Total pm
2. IST - Instituto Superior Técnico	Data	Endogenous resources	Batch 2/3	CORAL	5
				LECTRA	9
				Named entities tagged in natural language questions	9
				PTSTAR golden collection	9
			Total Batch 2/3		32
		Total Endogenous resources			32
	Total Data				32
Total 2. IST - Instituto Superior Técnico					32

Partner	Data/Soft	Where	Delivery	Designation	Total pm
3. UNIMAN-University of Manchester	Data	Endogenous resources	Batch 1	BioLexicon	**0,15**
				GREC	**0,15**
			Total Batch 1		**0,3**
		Total Endogenous resources			**0,3**
		Restricted exogenous resources	Batch 1	SemLink Resources	**0**
			Total Batch 1		**0**
			Batch 2/3	GENIA event corpus	**8,55**
			Total Batch 2/3		**8,55**
		Total Restricted exogenous resources			**8,55**
		Unrestricted exogenous resources	Batch 1	GENIA	**0,3**
				GREC	**0,15**
			Total Batch 1		**0,45**
		Total Unrestricted exogenous resources			**0,45**
	Total Data				**9,3**
	Software	Endogenous resources	Batch 2/3	NEMINE	**0**
				STEPP	**0**
				U-Compare NaCTeM Sentence Detector	**0**
				U-compare platform	**0,15**
				U-Compare STEPP PoS Tagger	**0**
				U-Compare Workbench	**0**
			Total Batch 2/3		**0,15**
		Total Endogenous resources			**0,15**
		Restricted exogenous resources	Batch 2/3	Enju	**0**
				Genia tagger/chunker and NER	**0**
				U-Compare Enju Parser	**0**
				U-Compare GENIA PoS Tagger	**0**
				U-Compare GENIA Sentence Detector	**0**
				U-Compare GENIA Tokenizer	**0**
				U-Compare OpenNLP PoSTagger	**0**
				U-Compare OpenNLP Sentence Detector	**0**
				U-Compare OpenNLP Tokenizer	**0**
				U-Compare Stanford Parser	**0**
				U-Compare Type System	**0,15**
			Total Batch 2/3		**0,15**
		Total Restricted exogenous resources			**0,15**
	Total Software				**0,3**
Total 3. UNIMAN-University of Manchester					**9,6**

Partner	Data/Soft	Where	Delivery	Designation	Total pm
4. UAIC - University Alexandru Ioan Cuza	Data	Endogenous resources	Batch 1	1984_NP	0,2
				1984AnaphoraRo	0,2
				FrRoMWE	0
				QA-corpus-UAIC	0
				RO-FDGBank	3
				RO-FN	3
				RoSemClass	1
				TE-pairsResource-UAIC	1
				TE-rules	1
			Total Batch 1		9,4
			Batch 2/3	eDTLR-sources	0,2
				RomMorph-UAIC	0,1
			Total Batch 2/3		0,3
		Total Endogenous resources			9,7
		Restricted exogenous resources	Batch 2/3	DEA	0
				DLPE	0
				eDTLR	2
				RoWN-eDTLR	1,7
				SRoL – Sounds of the Romanian Language	0
			Total Batch 2/3		3,7
		Total Restricted exogenous resources			3,7
	Total Data				13,4
	Software	Endogenous resources	Batch 2/3	ALPE-UAIC	4
				AnaMorph-UAIC	3
				Categorizer-UAIC	0
				Diacritics-UAIC	0
				DP-UAIC	3
				FDGparser-UAIC	2
				Language identifier-UAIC	0
				Lemmatizer-UAIC	2
				NP-chunker-UAIC	3
				Occurrence Finder-UAIC	2
				OntologyBuilder-UAIC	2
				QA-UAIC	3
				RARE-RO-UAIC	4
				Splitter-UAIC	0
				SRL-UAIC	3
				Summarizer-UAIC	2
				TE-UAIC	3
				Tokenizer-UAIC	1
			Total Batch 2/3		37
		Total Endogenous resources			37
		Restricted exogenous resources	Batch 2/3	ANNIE	0
			Total Batch 2/3		0
		Total Restricted exogenous resources			0
	Total Software				37
Total 4. UAIC - University Alexandru Ioan Cuza					50,4

Partner	Data/Soft	Where	Delivery	Designation	Total pm
5. RACAI - Romanian Academy	Data	Endogenous resources	Batch 1	Multilingual News Corpus	2
				RO-Acquis	2
				Romanian Balanced Corpus	2
				RO-SemCor	1
				RO-WordNet (first version)	4
				WEB-DEX	2
				Wordform lexicons	2
			Total Batch 1		15
		Total Endogenous resources			15
		Exogenous resources	Batch 1	Multilingual Subjectivity Analysis: Gold Standard and Training Data	1
				TimeBank parallel corpus	1
			Total Batch 1		2
			Batch 2/3	CoDII-NPI.ro	0,5
			Total Batch 2/3		0,5
		Total Exogenous resources			2,5
		Restricted exogenous resources	Batch 2/3	WEB 1T 5-gram	2
			Total Batch 2/3		2
		Total Restricted exogenous resources			2
		Unrestricted exogenous resources	Batch 1	RO-SAM EUROM	0,5
			Total Batch 1		0,5
		Total Unrestricted exogenous resources			0,5
	Total Data				20
	Software	Endogenous resources	Batch 2/3	COLLOC	1
				DIAC+	1
				LangId	1
				LexChain	1
				LexPar	1
				RO-HYPHEN	1
				SynWSD	1
				TTL-chunker	1
				TTL-lemmatizer	1
				TTL-Tagger	2
				TTL-Tokenizer	2
				VoiceForge	2
				WN-Builder	1
				YAWA	2
			Total Batch 2/3		18
		Total Endogenous resources			18
		Exogenous resources	Batch 2/3	Lucon	1
			Total Batch 2/3		1
		Total Exogenous resources			1
		Restricted exogenous resources	Batch 2/3	VISL Dependency-Parser	1
			Total Batch 2/3		1
		Total Restricted exogenous resources			1
	Total Software				20
Total 5. RACAI - Romanian Academy					40

Partner	Data/Soft	Where	Delivery	Designation	Total pm
6. UOM - University of Malta	Data	Endogenous resources	Batch 1	F_MONA_1	0,2
				Laws of Malta	3
				Maltese Acquis Communautaire EN	0,1
				Maltese Acquis Communautaire MT	0,1
				Maltese Spoken Newspaper	0,2
				Maltese Wordlist	1
			Total Batch 1		4,6
			Batch 2/3	MaltiWordNet of Maltese – Preliminary version – 15,000 entries	3
				MLRS Corpus	1
			Total Batch 2/3		4
		Total Endogenous resources			8,6
		Restricted exogenous resources	Batch 2/3	Acquilina Dictionary MT/EN	1
				Busuttil Dictionary EN/MT	1
				Busuttil Dictionary MT/EN	1
				Combined Maltese/English Bilingual Lexicon	3
				Eurowordnet	0,2
				Malta Online Dictionary	1
				Maltese Fiction	0,1
				Maltese Speech Engine Corpus	0,2
				MFSA_Companies_Register	0,2
			Total Batch 2/3		7,7
		Total Restricted exogenous resources			7,7
		Unrestricted exogenous resources	Batch 1	Basic English-Maltese Dictionary	1
				Illum_Corpus	0,2
			Total Batch 1		1,2
		Total Unrestricted exogenous resources			1,2
	Total Data				17,5
	Software	Endogenous resources	Batch 2/3	MLRS API	3
				MLRS API - POS Tagger	1
				MLRS1 Corpus Manager	0,5
				MLRS1 Lexicon Editor	0,5
			Total Batch 2/3		5
		Total Endogenous resources			5
	Total Software				5
Total 6. UOM - University of Malta					22,5

Partner	Data/Soft	Where	Delivery	Designation	Total pm
7. UPC - Technical University of Catalonia	Data	Endogenous resources	Batch 1	AGORA	0,1
				Bilingual Speech synthesis	4
				CatalanBN	0,1
				Catalan-SpeechDat	0,1
				EUROM.1	0,1
				FESTCAT	0,1
				FESTCAT-SEL	0,1
				FREE-SPEECH	0,1
				LC-STAR Dialogues	0,1
				Spanish Festival models	4
				Spanish Festival voices	4
				SpeechDat-Car Catalan	0,1
			Total Batch 1		12,9
			Batch 2/3	ALBAYZIN	0,1
				CHIL UPC Interactive Seminars	20
				CHIL UPC Seminars	0,1
				INTERFACE	0,1
				LC-STAR CATALAN	0,1
				LC-STAR SPANISH	0,1
				SALA-Mexico	0,1
				SALA-Venezuela	0,1
				Spanish SpeechDat (M) and SpeechDat (II)	0,1
				SpeechDat-Car Spain	0,1
				Speecon Catalan	0,1
				TALP TTS0 BASELINES	0,1
				TC-STAR TTS BASELINES	4
				TC-STAR TTS Expressive	0,1
				TC-STAR VC	4
			Total Batch 2/3		29,2
		Total Endogenous resources			42,1
		Restricted exogenous resources	Batch 2/3	BN RadioBCN	0,1
				EL_PERIODICO_97-07	0
				LAS CORTES	0,1
				SPANISH EPPS	0,1
				Speecon Spanish (SVOX)	0,1
			Total Batch 2/3		0,4
		Total Restricted exogenous resources			0,4
	Total Data				42,5
	Software	Endogenous resources	Batch 2/3	Gaia	0,1
				NannyRecord	0,1
				Saga	0,1
			Total Batch 2/3		0,3
		Total Endogenous resources			0,3
		Unrestricted exogenous resources	Batch 2/3	Festival	0,1
				HTS	0,1
			Total Batch 2/3		0,2
		Total Unrestricted exogenous resources			0,2
	Total Software				0,5
Total 7. UPC –Universitat Politècnica de Catalunya					43

Partner	Data/Soft	Where	Delivery	Designation	Total pm
8. UPF - University Pompeu Fabra	Data	Endogenous resources	Batch 1	Basic Vocabulary on the Human Genome	0,1
				Corpus PAAU 92	1
				Genoma corpus	0,1
				Multilingual Vocabulary of Economics	0,1
				Neologisms of the year: Bank of Spanish and Catalan Neologisms	0,2
				UPF_Term	0,1
			Total Batch 1		1,6
			Batch 2/3	IULA Technical Corpus	11
				Parallel IULA Technical Corpus	19
			Total Batch 2/3		30
		Total Endogenous resources			31,6
		Restricted exogenous resources	Batch 1	PAROLE lexicon	0,25
				SIMPLE lexicon	0,2
			Total Batch 1		0,45
			Batch 2/3	6305-QC	0,1
				AnCora-Ca	0,05
				AnCora-Co-CA	0,05
				AnCora-Co-ES	0,05
				AnCora-Es	0,05
				CESS_EU: The Basque Dependency Treebank	0,1
				Computer Science Tri-lingual Corpus	0,1
				Corpus CLUVI	0,1
				Corpus Técnico do Galego	0,1
				Dicionario CLUVI inglés-galego	0,1
				DOGC CAT-SPA Parallellized Corpus	0,1
				Electronic Corpus of Academic Materials – University of Zaragoza (ECAM-UZ)	0,1
				European Community Law Catalan Glossary mapped to EUROVOC	0,1
				Euskal Wordnet 3.0	0,05
				SenSem Corpus	0,1
				SenSem Database	0,1
				Spanish FrameNet	0,1
				Termoteca	0,1
				VOLEM	0,1
			Total Batch 2/3		1,65
		Total Restricted exogenous resources			2,1
		Unrestricted exogenous resources	Batch 1	Apertium Basque dictionary	0,1
				Apertium Bilingual dictionary , Basque-Spanish,	0,1
				Apertium Bilingual dictionary CA-ES	0,1
				Apertium Bilingual dictionary English-Catalan	0,1
				Apertium Bilingual dictionary English-Galician	0,1
				Apertium Bilingual dictionary English-Spanish	0,1
				Apertium Bilingual dictionary French-Catalan	0,1
				Apertium Bilingual dictionary French-Spanish	0,1
				Apertium Bilingual dictionary Occitan-Catalan	0,1
				Apertium Bilingual dictionary Occitan-Spanish	0,1
				Apertium Bilingual dictionary Portuguese-Catalan	0,1
				Apertium Bilingual dictionary Portuguese-Galician	0,1

				Apertium Bilingual dictionary Spanish-Asturian	0,1
				Apertium Bilingual dictionary Spanish-Galician	0,1
				Apertium Bilingual dictionary Spanish-Portuguese	0,1
				Apertium Bilingual dictionary Spanish-Romanian	0,1
				Apertium Catalan dictionary	0,1
				Apertium Galician dictionary	0,1
				Apertium Spanish dictionary	0,1
				FreeLing Asturian dictionary	0,1
				FreeLing Catalan dictionary	0,1
				FreeLing Catalan sense dictionary	0,1
				FreeLing Galician dictionary	0,1
				FreeLing Spanish dictionary	0,1
				FreeLing Spanish sense dictionary	0,1
			Total Batch 1		2,5
			Batch 2/3	Spanish Wordnet 3.0	0,1
			Total Batch 2/3		0,1
		Total Unrestricted exogenous resources			2,6
	Total Data				36,3
	Software	Endogenous resources	Batch 2/3	Tools for automatic UTF-8 conversion	6
				Tools for Catalan Corpus Processing	0
				Tools for Spanish Corpus Processing	12
			Total Batch 2/3		18
		Total Endogenous resources			18
	Total Software				18
Total 8. UPF - University Pompeu Fabra					54,3
Total general					328,4

4 List of Resources

4.1. Partner: ULX

4.1.1 Data resources

Endogenous resources

Name of the resource: Abbreviations
Type: lexicon, list of abbreviations
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: C-ORAL-ROM Portuguese Corpus
Type: corpus, speech
Rationale: the annotation adheres to a specific internal format; missing information
Upgrade: updating with regards to alignment format, to make it compatible with widely used programs; documentation and metadata
Covered languages: Portuguese

Name of the resource: CINTIL-Internacional Corpus of Portuguese
Type: corpus, speech
Rationale: this is a manually verified annotated corpus, but there have been some users' reports of inconsistencies
Upgrade: check annotation consistency with regard to ambiguous classes (e.g. relative and interrogative pronouns, multiword units).
Covered languages: Portuguese

Name of the resource: DEF Corpus
Type: corpus, annotated corpus
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: DependencyBank
Type: corpus, annotated corpus
Rationale: given current state of the art and applications that build on this resource (SMT, parsing, etc.), this treebank is lacking adaptation
Upgrade: updating to metashare formats, documentation and metadata, upgrading annotation (10K sentences), cross-linking with other treebanks
Covered languages: Portuguese

Name of the resource: Multifunctional Computational Lexicon of Contemporary Portuguese - CORLEX
Type: lexicon, frequency lexicon
Rationale: incomplete documentation
Upgrade: full documentation: corpus used, metadata, criteria, format.
Covered languages: Portuguese

Name of the resource: MWN.PT
Type: wordnet, wordnet
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: PAROLE corpus
Type: corpus, annotated corpus
Rationale: incomplete documentation
Upgrade: full documentation of metadata and annotation manual
Covered languages: Portuguese

Name of the resource: PAROLE lexicon
Type: lexicon, lexicon
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: PF Corpus
Type: corpus, speech
Rationale: the transcriptions of the recordings follow a specific internal format; the corpus is not POS annotated
Upgrade: transcriptions will be updated to the C-ORAL-ROM format (based mainly on CHAT) and converted to XML; corpus will be updated to a TEI-compliant format; documentation and metadata; tagged with POS information
Covered languages: Portuguese

Name of the resource: PropBank
Type: corpus, annotated corpus
Rationale: given current state of the art and applications that build on this resource (SMT, parsing, etc.), this treebank is lacking adaptation
Upgrade: documentation and metadata, cross-linking with other treebanks, upgrading annotation of 10K sentences, adaptation to metashare formats
Covered languages: Portuguese

Name of the resource: SIMPLE lexicon
Type: lexicon, lexicon
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: Spoken Portuguese
Type: speech, speech database
Rationale: the annotation adheres to a specific internal format; missing information
Upgrade: adapt this corpus to a TEI-compliant format; complete its documentation and metadata description, and secure the copyrights.
Covered languages: Portuguese

Name of the resource: Stopwords
Type: lexicon, list of stopwords
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: Technical Corpus
Type: corpus, raw text corpus
Rationale: the annotation adheres to a specific internal format; no POS and lemmatisation
Upgrade: complete the metadata documentation and adapt this corpus to a TEI-compliant format; update POS annotation, lemmatisation, and the annotation of simple and MW terminological units.
Covered languages: Portuguese

Name of the resource: Treebank
Type: corpus, annotated corpus
Rationale: given current state of the art and applications that build on this resource (SMT, parsing, etc.), this treebank is lacking adaptation
Upgrade: adaptation to metashare formats, upgrading annotation of 10K sentences, cross-linking with other treebanks, documentation and metadata
Covered languages: Portuguese

Restricted exogenous resources

Name of the resource: Clássicos LP/Porto Editora
Type: corpus, raw text corpus
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: COMPARA
Type: corpus, raw text corpus
Upgrade: documentation and metadata
Covered languages: Portuguese/English

Name of the resource: Corpus NILC
Type: corpus, raw text corpus
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: Dicionário de Verbos do Português Medieval (DVPM)
Type: dictionary, dictionary
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: European Parliament Parallel Corpus
Type: corpus, raw text corpus
Upgrade: documentation and metadata
Covered languages: Portuguese/English

Name of the resource: Geo-Net-PT01
Type: grammar, ontology
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: Glossário

Type: lexicon, lexical database
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: MorDebe
Type: lexicon, lexical database
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: Norma Urbana Culta (NURC)
Type: corpus, speech
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: Ontologia de Nanociência e Nanotecnologia
Type: ontology, ontology
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: Panorama do Português Oral de Maputo (PPOM)
Type: corpus, speech
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: PORLEX
Type: lexicon, lexical database with information on word structure
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: The JRC-Acquis Multilingual Parallel Corpus
Type: raw text corpus, multilingual corpus
Upgrade: documentation and metadata
Covered languages: Portuguese (22 languages)

Unrestricted exogenous resources

Name of the resource: CETEMPúblico
Type: corpus, raw text corpus
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: Corpus NILC
Type: corpus, annotated corpus
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: CorpusTCC
Type: corpus, annotated corpus
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: NILC Taggers
Type: grammar, training models for tagger
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: PLN-BR Gold
Type: corpus, annotated corpus
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: RHETALHO
Type: corpus, annotated corpus
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: Summ-it
Type: corpus, annotated corpus
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: TeMário 2006
Type: corpus, raw text corpus
Upgrade: documentation and metadata
Covered languages: Portuguese

4.1.2 Software and LR tools

Endogenous resources

Name of the resource: Chunker
Type: LT tool, sentence splitter
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: Tagger
Type: LT tool, part-of-speech tagger
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: Tokenizer
Type: LT tool, tokenizer
Upgrade: documentation and metadata
Covered languages: Portuguese

Unrestricted exogenous resources

Name of the resource: DiZer 2.0
Type: LT tool, discourse parser
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: Forma
Type: LT tool, tagger
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: GistSumm
Type: LT tool, summarizer
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: Ontolp Plugin
Type: LT tool, ontology buidling
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: Stemmer
Type: LT tool, stemmer
Upgrade: documentation and metadata
Covered languages: Portuguese

Name of the resource: Text Aligners
Type: LT tool, sentence and word aligners
Upgrade: documentation and metadata
Covered languages: Portuguese/other language

4.2. Partner: IST

4.2.1 Data resources

Endogenous resources

Name of the resource: PTSTAR golden collection
Type: annotated corpus, multilingual word alignments
Rationale: Only a subset of the corpus has been transcribed and annotated
Upgrade: no
Extend: yes
Licensing: n.a. since this is an endogenous resource
Covered languages: 6 european

Name of the resource: Named entities tagged in natural language questions
Type: raw text corpus, Text
Upgrade: yes
Extend: no
Licensing: n.a. since this is an endogenous resource
Covered languages: English

Name of the resource: LECTRA
Type: speech, Speech database
Rationale: Only a subset of the corpus has been transcribed and annotated
Upgrade: no

Extend: yes
Licensing: n.a. since this is an endogenous resource
Covered languages: Portuguese

Name of the resource: CORAL
Type: speech, Speech database
Upgrade: yes
Extend: no
Licensing: n.a. since this is an endogenous resource
Covered languages: Portuguese

4.3. Partner: UNIMAN

4.3.1 Data resources

Endogenous resources

Name of the resource: BioLexicon
Type: Lexicon, large-scale terminological resource
Rationale: large unique biomedical lexical resource
Upgrade: documentation and metadata
Extend: add verbs from SemLink resources
Licensing: endogenous; ELRA licences
Covered languages: English/biomedical domain

Name of the resource: GREC
Type: annotated corpus, annotated corpus
Upgrade: documentation and metadata
Covered languages: English/ biomedical

Unrestricted exogenous resources

Name of the resource: GENIA
Type: annotated corpus, event annotation
Upgrade: documentation and metadata
Covered languages: English/ biomedical

Name of the resource: GENIA
Type: , part-of-speech annotation and terms (merged corpus)
Upgrade: documentation and metadata
Covered languages: English/ biomedical

Name of the resource: GREC
Type: annotated corpus, annotated corpus
Rationale: make the corpus compliant to Metanet/metashare standards, currently part of U-compare infrastructure
Upgrade: documentation and metadata
Extend: Extension: annotate the GREC event corpus with new layers of annotation to allow advanced information extraction involving opinion mining, inconsistency and contradiction checking, entailment and hedging.

In addition, we will wrap the extended resource
Licensing: endogenous resource
Covered languages: English/ biomedical

Restricted exogenous resources

Name of the resource: «Designation»
Type: «Type», «Type1»
Rationale: «Rational»
Upgrade: «Upgrade»
Extend: «Extend»
Licensing: «Licencing»
Covered languages: «Language»

Name of the resource: «Designation»
Type: «Type», «Type1»
Rationale: «Rational»
Upgrade: «Upgrade»
Extend: «Extend»
Licensing: «Licencing»
Covered languages: «Language»

4.3.2 Software and LR tools

Restricted exogenous resources

Name of the resource: Enju
Type: LR tools, Deep Parser
Upgrade: documentation and metadata
Covered languages: English

Name of the resource: Genia tagger/chunker and NER
Type: LR tools, tagger, chunker and NERfor biomedical text
Upgrade: documentation and metadata
Covered languages: English, biomedical

Name of the resource: U-Compare Enju Parser
Type: LR tools, HPSG Parser
Upgrade: documentation and metadata
Covered languages: English

Name of the resource: U-Compare GENIA PoS Tagger
Type: LR tools, PoS Tagger
Upgrade: documentation and metadata
Covered languages: English

Name of the resource: U-Compare GENIA Sentence Detector
Type: LR tools, Sentence splitter
Upgrade: documentation and metadata
Covered languages: English

Name of the resource: U-Compare GENIA Tokenizer
Type: LR tools, Tokenizer
Upgrade: documentation and metadata
Covered languages: English

Name of the resource: U-Compare OpenNLP PoSTagger
Type: LR tools, PoS Tagger
Upgrade: documentation and metadata
Covered languages: English, Spanish, German, Thai

Name of the resource: U-Compare OpenNLP Sentence Detector
Type: LR tools, Sentence splitter
Upgrade: documentation and metadata
Covered languages: English

Name of the resource: U-Compare OpenNLP Tokenizer
Type: LR tools, Tokenizer
Upgrade: documentation and metadata
Covered languages: English, Spanish

Name of the resource: U-Compare Stanford Parser
Type: LR tools, PCFG Parser
Upgrade: documentation and metadata
Covered languages: English

Name of the resource: U-Compare Type System
Type: Services, Specification of linguistic annotations
Upgrade: documentation and metadata
Covered languages: Generic, with English-specific extensions

4.4. Partner: UAIC

4.4.1 Data resources

Endogenous resources

Name of the resource: RoSemClass
Type: semantic classes of lexicals for political discourse analysis
Rationale: extend the semantic classes with sentiment information
Upgrade: align the existing lexicals to RoWN and add new sentiment classes.
Extend: We see three ways of upgrading NLP resources. The first is to make them all compatible with a consistent set of standards agreed upon by the project consortium. The second refers to their quality. We think that all tools should be validated by using generally agreed evaluation criteria. And the third refers to making the tools accessible as Web resources.
Licensing: GNU-OPL
Covered languages: Romanian

Name of the resource: TE-rules
Type: Grammars, collection of rules for classification of textual inferences

Rationale: valuable semantic information can be obtained with proper symbolic rules, improving the performance of the TE-UAIC system
Upgrade: include more rules of the following forms: if (X sells Y to Z) then "Z has Y", if (X is a SPORT champion) then "X plays SPORT".
Extend: see above
Licensing: GNU-OPL
Covered languages: Romanian and English

Name of the resource: TE-pairsResource-UAIC
Type: Grammars, collection of Text-Hypothesis pairs
Rationale: the TE-UAIC system accuracy is directly linked to the existing number of training pairs (only 200 in this set)
Upgrade: raise the size of the training set for Romanian to 2000 pairs (same as the English set)
Extend: see above
Licensing: GNU-OPL
Covered languages: Romanian

Name of the resource: RomMorph-UAIC
Type: Grammars, paradigmatic morphology rules in symbolic form
Rationale: update to META-SHARE compliance: Metadata, UTF-8
Upgrade: Metadata annotation, conversion to UTF
Extend: see above
Licensing: internal
Covered languages: modern Romanian

Name of the resource: RO-FN
Type: wordnet, FrameNet-based English-Romanian parallel corpus of semantic roles.
Rationale: the corpus only contains around 1500 sentences, and is relatively small for a usage in machine learning. Further annotations are relatively simple to obtain automatically, but must be validated.
Upgrade: translate English sentences, automatically import annotation, and validate the import
Extend: see above
Licensing: GNU-OPL
Covered languages: English and Romanian

Name of the resource: RO-FDGBank
Type: syntactic annotated corpus
Rationale: the corpus has been developed in two separate stages, not all trees were verified for consistency across annotators
Upgrade: validate the sentences and correct them where needed, increase the number of trees
Extend: see above
Licensing: GNU-OPL
Covered languages: Romanian

Name of the resource: QA-corpus-UAIC

Type: annotated corpus, annotated Question Answering corpus
Rationale: this corpus is specific to the Wikipedia and JRC-Acquis domains
Upgrade: expansion using the corpora and resources from the previous years of the CLEF competition
Extend: see above
Licensing: GNU-OPL
Covered languages: Romanian

Name of the resource: FrRoMWE
Type: annotated corpus, annotated French-Romanian corpus
Rationale: the corpus must be extended in order to obtain different contexts for each phraseological unit; the annotation must be verified and completed.
Upgrade: extending the bilingual corpus, refining the annotation of phraseological units
Extend: see above
Licensing: GNU-OPL
Covered languages: Romanian, French

Name of the resource: eDTLR-sources
Type: raw text corpus, collection of scanned and OCR-ed books
Rationale: update to META-SHARE compliance: Metadata, UTF-8 and LMF
Upgrade: Metadata annotation, conversion to UTF
Extend: see above
Licensing: to be decided by the members of the eDTLR project consortium
Covered languages: Romanian

Name of the resource: 1984AnaphoraRo
Type: annotated corpus, annotated corpus
Rationale: update to META-SHARE compliance: Metadata, UTF-8
Upgrade: Metadata annotation, conversion to UTF
Extend: see above
Licensing: freely available
Covered languages: Romanian

Name of the resource: 1984_NP
Type: annotated corpus, NP chunks manually annotated corpus
Rationale: update to META-SHARE compliance: Metadata, UTF-8
Upgrade: Metadata annotation, conversion to UTF
Extend: see above
Licensing: freely available
Covered languages: Romanian

Restricted exogenous resources

Name of the resource: RoWN-eDTLR
Type: lexical ontology, Semantic dictionary
Rationale: extend Ro-WN resource and its alignment with the English WN, link Ro-WN to eDTLR
Upgrade: align the existent synsets with word senses listed in eDTLR, extract new synsets from eDTLR, align them with En-WN
Licensing: not available

Covered languages: Romanian

Name of the resource: SRoL – Sounds of the Romanian Language
Type: speech, speech corpus: annotated and documented speech resource
Rationale: update to META-SHARE compliance: Metadata, UTF-8
Upgrade: Metadata annotation, conversion to UTF
Licensing: copyrighted
Covered languages: Romanian

Name of the resource: eDTLR
Type: lexica, lexicography
Rationale: not coupled with a morphology module
Upgrade: use the morphological modules of UAIC to develop a words flexing interface
Licensing: to be decided by the members of the eDTLR project consortium
Covered languages: Romanian

Name of the resource: DLPE
Type: dictionary, Dictionary of poetic language in Eminescu's work
Rationale: update to META-SHARE compliance: Metadata, UTF-8
Upgrade: Metadata annotation, conversion to UTF
Licensing: for owner's group only
Covered languages: Romanian

Name of the resource: DEA
Type: dictionary, Dictionary of adult education
Rationale: update to META-SHARE compliance: Metadata, UTF-8
Upgrade: Metadata annotation, conversion to UTF
Licensing: for owner's group only
Covered languages: Romanian

4.4.2 Software and LR tools

Endogenous resources

Name of the resource: OntologyBuilder-UAIC
Type: LR tools, Ontology Builder
Rationale: a source of errors is the inconsistent use of diacritics in Romanian. More tests are needed, as well as an upgrade of its resources.
Upgrade: a component that restores diacritics could be inserted as a pre-processing step
Extend: see above
Licensing: GNU-OPL
Covered languages: Romanian and English

Name of the resource: AnaMorph-UAIC
Type: LR tools, Word flexing system
Rationale: unable to handle words belonging to multiple flexing paradigms, obsolete user interface

Upgrade: improve the set of lexicons/models, rewrite resource code and the user interface, improve documentation
Extend: see above
Licensing: GNU-OPL
Covered languages: Romanian

Name of the resource: Categorizer-UAIC
Type: LR tools, Document category/domain identification
Rationale: not standard input-output formats; low performance because of the small size of the training corpus
Upgrade: for the moment we have an established fixed set of categories, but the user should be able to define its own categories as long as training data are also provided. We intend to collect new training corpora in order to diversify the categorisation.
Extend: see above
Licensing: GNU-OPL
Covered languages: English

Name of the resource: Diacritics-UAIC
Type: LR tools, Diacritics Recovery System
Rationale: precision still low
Upgrade: work on the model, improve the attached resources
Extend: see above
Licensing: GNU-OPL
Covered languages: Romanian

Name of the resource: DP-UAIC
Type: LR tools, Discourse Parser system
Rationale: unknown performance because of the small size of the training corpus
Upgrade: rewrite tool's code or implement wrappers, improve the attached resources (set of discourse markers, etc.); improve accuracy
Extend: see above
Licensing: GNU-OPL
Covered languages: Romanian and English

Name of the resource: FDGparser-UAIC
Type: LR tools, Functional Dependency Parser
Rationale: not standard input-output formats; low performance because of the small size of the training corpus
Upgrade: retrain the system using a larger training set; improve the language model for determining syntactic relations; measure and improve accuracy
Extend: see above
Licensing: GNU-OPL
Covered languages: Romanian

Name of the resource: Language identifier-UAIC
Type: LR tools, Language identifier
Rationale: now recognizes only English and Romanian, not known accuracy

Upgrade: collect and train on more data and additional languages; rewrite resource code; measure and improve accuracy
Extend: see above
Licensing: GNU-OPL
Covered languages: Romanian and English

Name of the resource: Lemmatizer-UAIC
Type: LR tools, Lemmatizer
Rationale: the system accuracy is still poor on words which can have several possible lemmas. Standardize input and output formats as XML
Upgrade: the use of contexts to decide ambiguous lemmas. Rewrite resource code or implement wrappers for compatibility reasons
Extend: see above
Licensing: GNU-OPL
Covered languages: Romanian and English

Name of the resource: ALPE-UAIC
Type: LR tools, NLP workflow builder
Rationale: beta-version only, limited functionality,
Upgrade: finalize development, supplement with a core hierarchy of annotation schemes
Extend: see above
Licensing: GNU-OPL
Covered languages: Romanian and English

Name of the resource: Occurrence Finder-UAIC
Type: LR tools, Occurrence Finder
Rationale: fixed output format, does not use a lemmatiser
Upgrade: rewrite resource code to allow more flexible visualisation of results, incorporate a lemmatiser
Extend: see above
Licensing: GNU-OPL
Covered languages: Romanian

Name of the resource: QA-UAIC
Type: LR tools, Question Answering
Rationale: performance under the top-most systems in QA
Upgrade: add additional semantic resources to the information retrieval module
Extend: see above
Licensing: GNU-OPL
Covered languages: Romanian and English

Name of the resource: RARE-RO-UAIC
Type: LR tools, Robust rule-based Anaphora Resolution system
Rationale: rule-based, therefore with potential for being improved, non standard input-output
Upgrade: transform the system into a mixed one (rules + statistical), rewrite resource code or implement wrappers, improve performance, evaluate
Extend: see above

Licensing: GNU-OPL
Covered languages: Romanian and English

Name of the resource: Splitter-UAIC
Type: LR tools, Sentence Splitter
Rationale: non standard input and output formats, unknown accuracy
Upgrade: rewrite resource code or implement wrappers, measure and improve accuracy
Extend: see above
Licensing: GNU-OPL
Covered languages: Romanian and English

Name of the resource: SRL-UAIC
Type: LR tools, Semantic Role Labeling
Rationale: unexplored possibilities to improve performance, non standard input&output
Upgrade: improve performance and standardize input and output formats, rewrite resource code or implement wrappers
Extend: see above
Licensing: GNU-OPL
Covered languages: Romanian, English

Name of the resource: Summarizer-UAIC
Type: LR tools, Summarization system
Rationale: unknown performance
Upgrade: rewrite resource code; measure and improve accuracy
Extend: see above
Licensing: GNU-OPL
Covered languages: Romanian and English

Name of the resource: TE-UAIC
Type: LR tools, Textual Entailment
Rationale: although the system is ranked among the best known, it is still unable to handle correctly Text-Hypothesis pairs that contain similar sets of words with different word order
Upgrade: SRL-UAIC could be used in order to identify differences in the distribution of semantic roles in the Text and Hypothesis
Extend: see above
Licensing: GNU-OPL
Covered languages: Romanian and English

Name of the resource: Tokenizer-UAIC
Type: LR tools, Tokenizer
Rationale: output format non standard, not known accuracy
Upgrade: rewrite resource code or implement wrappers, measure and improve accuracy
Extend: see above
Licensing: GNU-OPL
Covered languages: Romanian and English

Name of the resource: NP-chunker-UAIC
Type: LR tools, NP-chunker
Rationale: not standard input-output formats; limited set of regular expressions; not known accuracy
Upgrade: rewrite resource code or implement wrappers; upgrade the set of regular expressions, measure and improve accuracy
Extend: see above
Licensing: GNU-OPL
Covered languages: Romanian

Restricted exogenous resources

Name of the resource: ANNIE
Type: LR tools, Named Entities Recognizer
Rational: update to META-SHARE compliance: Metadata, UTF-8
Upgrade: Metadata annotation, conversion to UTF
Extend:
Licensing: GPL
Covert languages: Romanian and English

4.5. Partner: RACAI

4.5.1 Data resources

Endogenous resources

Name of the resource: Wordform lexicons
Type: tagged and lemmatized wordform lists, lexicon
Rationale: training new ML tools
Upgrade: UTF8 encoding;
Extend: each entry in this lexicon would be associated with relative frequencies computed from our corpora and/or the web.
Licensing: not available
Covered languages: Romanian, English, French, German

Name of the resource: WEB-DEX
Type: dictionary, Romanian reference explanatory dictionary
Rationale: None
Upgrade: It is encoded using a proprietary schema (developed within CONCEDE European project) and it needs reformatting according to a wider used dictionary format (TEI or LMF).It is encoded using a proprietary schema (developed within CONCEDE European project)
Extend: Upgrading actions: there will be provided: a detailed description document, including various statistics; re-encoding of the dictionary according to a standard specification (META-NET recommended one), updating the orthography, modifying the outdated parts
Licensing: not available
Covered languages: Romanian

Name of the resource: Romanian Balanced Corpus

Type: annotated corpus, largest annotated corpus for Romanian
Rationale: training new ML tools
Upgrade: the preprocessing of the corpus components has been done using different tools improved over the time. Several texts are written with the old orthography, sometimes without diacritics, and, there are too many unknown words
Extend: about 10 mio new tokens
Licensing: not available
Covered languages: Romanian

Name of the resource: RO-WordNet
Type: lexical ontology, Semantic dictionary
Rationale: 30,000 synsets only, not all aligned with thesauri of senses and with English WN
Upgrade: a detailed description document, including various statistics; error corrections and orthography unification , adding new synsets (doubling the actual synset number), updating the interlingual alignment to Princeton WordNet 3.0
Extend: add 30,000 synsets, aligned them to WN2.0 and WN3.0, add SUMO/MILO., DOMAINS, SENTIWORDNET and connotation annotations
Licensing: not available
Covered languages: Romanian

Name of the resource: RO-SemCor
Type: parallel sense-annotated corpus, parallel corpus
Rationale: further development of research in supervized WSD
Upgrade: writing documentation, ; conversion of sgml entities into Unicode characters; needs updating the text to the 2005 morphological rules;
Licensing: not available
Covered languages: Romanian-English

Name of the resource: RO-Acquis
Type: annotated corpus, Romanian Corpus
Rationale: building specialized language model
Upgrade: mark-up the juridical multiword terms as single tokens
Extend: Upgrading actions: a detailed description document, including various statistics; cleaning up the texts, adding the missing diacritics, updating the text to the 2005 morphological rules; the metadata should be drastically revised;
Licensing: OPOCE (EU Publication Office) and JRC-Ispra for the raw texts; for the pre-processed version (our delivery) we do not impose any restrictions other than the academic ones (citations).
Covered languages: Romanian

Name of the resource: Multilingual News Corpus
Type: Written, comparable corpora
Rationale: Reasons for extension/linking:
Upgrade: xml encoding, documentation
Licensing: not available
Covered languages: all EU official languages but Irish.

Exogenous resources

Name of the resource: TimeBank parallel corpus
Type: annotated corpus, written corpus
Rationale: train TimeML taggers
Upgrade: documenting, adding metadata, extending the corpus and annotating the extended texts
Covered languages: Romanian, English

Name of the resource: Multilingual Subjectivity Analysis: Gold Standard and Training Data
Type: Gold Standard / Training Data Corpus. The Spanish and Romanian parts of the corpus are machine translations of the English part,
Upgrade: : at least adding diacritics and correcting the machine translations (for Romanian)
Covered languages: English, Romanian, Spanish

Name of the resource: CoDII-NPI.ro
Type: , written
Upgrade: unknown
Covered languages: Romanian

Unrestricted exogenous resources

Name of the resource: RO-SAM EUROM
Type: wav file and their transcription as XML files, a collection of sentences spoken by professional speakers
Rationale: development of speech technology for Romanian
Upgrade: UTF8 encoding, ortography adaptation, documentation
Licensing: freely available
Covered languages: Romanian

Restricted exogenous resources

Name of the resource: WEB 1T 5-gram
Type: 1-, 2-, 3-, 4- and 5-grams, a collection of n-grams in 10 European languages (including Romanian),
Rationale: Reasons for extension/linking: possibly mixed erroneous statistics
Upgrade: diacritics restoration, new corrected statistics will be computed
Licensing: LDC for the initial resource, n.a. for the new resource
Covered languages: Czech, Dutch, French, German, Italian, Polish, Portuguese, Romanian, Spanish and Swedish

4.5.2 Software and LR tools

Endogenous resources

Name of the resource: YAWA
Type: LR tools, word aligner
Rationale: works with SGML entities encoding for the non-ASCII caracters and should work with Unicode; this extension would require the modification of the

cognate function scoring; YAWA contains one language-pair specific (Romanian-English) rule-based module which
Upgrade: writing a proper user manual, modification of the code to work with Unicode character encoding and replacing the rule-based language-pair dependent module with a trainable ML language-independent module (based on LexPar for instance).
Extend: n.a. since this is an endogenous resource
Licensing: n.a.
Covered languages: Romanian, English

Name of the resource: WN-Builder
Type: LR tools, wordnet editor
Rationale: works with sgml entities for Romanian diacritical characters; some formatting facilities are missing, but repeatedly asked for by the lexicographers using it
Upgrade: writing a proper user manual, Unicode character set encoding, adding new formatting facilities; removing some recently discovered bugs
Extend: n.a. since this is an endogenous resource
Licensing: n.a.
Covered languages: language independent

Name of the resource: VoiceForge
Type: audio segmentation and speech synthesis,
Rationale: this is a PhD project, the program is already functional, but needs training, evaluations and further performance improvements
Upgrade: creating resources for testing and evaluation, writing a manual and a benchmark
Licensing: n.a.
Covered languages: Language-Independent

Name of the resource: TTL-Tokenizer
Type: LR tools, Tokenizer
Rationale: works with sgml entities for Romanian diacritical characters, requires language resources for recognizing compounds, including named entities. Need to fix a bug where when recognizing named entities, some replacements with system-dependent strings are no
Upgrade: writing a proper user manual, Unicode character set encoding, updating the tokenization lists, adding a gazetteer and an enhanced NER module, fixing bugs.
Extend: n.a. since this is an endogenous resource
Licensing: n.a.
Covered languages: Romanian, English, French

Name of the resource: TTL-Tagger
Type: LR tools, Morpho-syntactic tagger
Rationale: works with sgml entities for Romanian diacritical characters,
Upgrade: writing a proper user manual, Unicode character set encoding, retraining the HMM model on much larger hand tagged data (it has been trained on Orwell's "1984" -120,000 tokens, now we can tag it on "ROCO/AGENDA"+"1984" corpora, altogether 8,000,000 tokens)

Extend: n.a. since this is an endogenous resource
Licensing: n.a.
Covered languages: Romanian, English, French

Name of the resource: TTL-lemmatizer
Type: LR tools, lemmatizer
Rationale: works with sgml entities for Romanian diacritical characters, requires language resources for recognizing compounds, including named entities. Loads a big lemmatization table into the memory which is a source for memory leaks with different Perl versions.
Upgrade: writing a proper user manual, Unicode character set encoding, updating the lemmatization lists to handle compounds (including NEs) and revising the module for lemmatization the out of vocabulary words. Also it needs an optimization in order not to store t
Extend: n.a. since this is an endogenous resource
Licensing: n.a.
Covered languages: Romanian, English, French

Name of the resource: TTL-chunker
Type: LR tools, chunker
Rationale: extending the chunking rules for the verbal chunk: currently it does not recognize the negative compound tenses and passive diathesis as a single chunk; we would like to enhance the resources for chunking English and French as well as to build chunking re
Upgrade: writing a proper user manual, updating the chunking rules for Romanian, English, French in order to the negative compound tenses and passive diathesis, as well as to build chunking resources for some new languages. We are also studying the possibility of
Extend: n.a. since this is an endogenous resource
Licensing: n.a.
Covered languages: Romanian, English, French

Name of the resource: SynWSD
Type: LR tools, Word Sense Disambiguation
Rationale: taking advantage of the enhanced LexPar (directed and labelled links). Studying the supervised WSD training that SynWSD natively supports.
Upgrade: writing a proper user manual, extending the code for dealing with both unlabeled & undirected and directed & labelled dependency links for Romanian (and maybe for French). Make it train on sense-annotated corpora and perform like a supervised WSD algorit
Extend: n.a. since this is an endogenous resource
Licensing: n.a.
Covered languages: any language for which a BalkaNet compliant wordnet is available

Name of the resource: RO-HYPHEN
Type: LR tools, hyphenator
Rationale: currently this program is an offline utility taking a list of Romanian words, as input and provides as output a list of hyphenated words; the diacritical

characters are encoded as sgml entities; it should word with Unicode encoding and should work also on
Upgrade: writing a proper user manual, updates the program for Unicode character encoding
Extend: n.a. since this is an endogenous resource
Licensing: n.a.
Covered languages: Romanian

Name of the resource: LexPar
Type: LR tools, Dependency linker
Rationale: LexPar builds only dependency undirected links; however, for richly inflectional languages (especially case-marked languages) tagged with lexical tags (as tiered tagging does) some of the undirected links may be turned into directed ones and rough syntact
Upgrade: writing a proper user manual, extending the code for creating directed and labelled dependency links for Romanian (and maybe for French); rewriting the filtering rules to take advantage of direction and the label (when available) of a link. Also, we need
Extend: n.a. since this is an endogenous resource
Licensing: n.a.
Covered languages: language independent

Name of the resource: LexChain
Type: LR tools, Lexical Chain
Rationale: currently it does not take into account the sequence and the types of the semantic relations that make up a chain. Recently we realized that we need to investigate the paths that make-up a lexical chain and choose those paths that resemble human judgement
Upgrade: writing a proper user manual, implementing the human-judged paths into lexical chains, reassessing the value of each semantic path as the the semantic similarity goal.
Extend: n.a. since this is an endogenous resource
Licensing: n.a.
Covered languages: language independent for which a wordnet is available

Name of the resource: LangId
Type: LR tools, language identification
Rationale: currently it handles 25 languages (all EU languages but Irish and three endangered languages) but we would like to extend it to handle much more languages and different writing scripts
Upgrade: writing a proper user manual, modifying the code to handle Unicode encoded characters, training the language recognizer on some major new languages (Japanese, Chinese, Hindi etc)
Extend: n.a. since this is an endogenous resource
Licensing: n.a.
Covered languages: language independent

Name of the resource: DIAC+
Type: LR tools, diacritics restorations for Romanian texts

Rationale: works only for MS Word 2007;
Upgrade: writing a proper user manual, Unicode character set encoding; extending it to other MSOffice programs (e.g. PowerPoint, Excel) and to raw text files; building a language model based on much larger training data; retraining the system with the new language
Extend: n.a. since this is an endogenous resource
Licensing: n.a.
Covered languages: Romanian

Name of the resource: COLLOC
Type: LR tools, collocation extractor
Rationale: currently it works on raw texts and we want to turn it into understanding and taking advantage of the xml annotated (in a standard way) corpora
Upgrade: writing a proper user manual, modifying the code to use the xml token annotations in a language-aware tokenized text; the annotations should be XCES compliant
Extend: n.a. since this is an endogenous resource
Licensing: n.a.
Covered languages: language independent

Exogenous resources

Name of the resource: Lucon
Type: concordancer
Upgrade: creating resources for testing and evaluation, writing a manual
Licensing: N/A
Covered languages: Language-Independent

Restricted exogenous resources

Name of the resource: VISL Dependency-Parser
Type: LR tools, a dependency parser for several languages (including Romanian) texts
Covered languages: Danish, English, Esperanto, French, German, Italian, Norwegian,

4.6. Partner: UOM

4.6.1 Data resources

Endogenous resources

Name of the resource: MLRS Corpus
Type: raw text corpus, written corpus
Rationale: increase coverage of the corpus
Upgrade: increase coverage (size increased from 9M to 1G) by acquiring text from sources mentioned earlier, particularly news, fiction, government documentation; POS annotation at state-of-the-art accuracy. Corpus representation and documentation to be brought int
Extend: An absolute essential first step is POS tagging, which will be achieved using the tagger component currently being developed under the MLRS toolkit umbrella.

Licensing: none
Covered languages: Maltese

Name of the resource: Maltese Wordlist
Type: lexica, Lexicon/Knowledge Sourcc
Rationale: Current wordlist has limited coverage (approximately 100K word types), contains errors, and is only documented in an informal way.
Upgrade: Primarily these actions will be to (i) increase coverage (ii) increase quality and accuracy of entries (iii) create appropriate documentation and meta-data for the system. implement system for checking and maintaining accuracy.
Extend: Size increased from 100K to c. 1M. Initially aim is to perfect the list. Related goal is to develop relation between wordlist and lexicon.
Licensing: none
Covered languages: Maltese

Name of the resource: Maltese Spoken Newspaper
Type: speech, spoken corpus
Upgrade: Improve formal and narrative documentation of corpus. Upgrade corpus to a representation to the standard agreed with META-NET concerning monolingual corpus encoding.
Covered languages: Maltese

Name of the resource: Maltese Acquis Communautaire MT
Type: raw text corpus, Written corpus
Upgrade: Improve formal and narrative documentation of corpus. Upgrade corpus to a representation to the standard agreed with META-NET concerning monolingual corpus encoding.
Covered languages: MT

Name of the resource: Maltese Acquis Communautaire EN
Type: raw text corpus, Written corpus
Upgrade: Improve formal and narrative documentation of corpus. Upgrade corpus to a representation to the standard agreed with META-NET concerning monolingual corpus encoding.
Covered languages: EN

Name of the resource: Laws of Malta
Type: raw text corpus, written corpus
Rationale: Data is currently available as a text tile. Aim is to translate it to a consistent representation standard compatible with META-NET. The reason for upgrading is to prepare it for subsequent alignment with the EN version of the same document.
Extend: The aim here is to use the EN and MT corpora to create a substantial aligned bilingual corpus MT/EN in a standardised format since this can be used as a basis for a number
of linking activities including a bilingual lexicon. In this case the terminology i
Licensing: none, since the material is endogenous to Malta and freely available from the Maltese Government website
Covered languages: Maltese/EN

Name of the resource: Laws of Malta
Type: raw text corpus, written corpus
Rationale: Data is currently available as a text tile. Aim is to translate it to a consistent representation standard compatible with META-NET. The reason for upgrading is to prepare it for subsequent alignment with the EN version of the same document.
Upgrade: Improve formal and narrative documentation of corpus. Upgrade corpus to a representation to the standard agreed with META-NET concerning monolingual corpus encoding.
Extend: Aim will be to ensure that all existing information in corpus is upgraded in a consistent way.
Licensing: none, since the material is endogenous to Malta and freely available from the Maltese Government website
Covered languages: Maltese/EN

Name of the resource: F_MONA_1
Type: speech, speech data
Upgrade: Improve formal and narrative documentation of corpus. Upgrade corpus to a representation to the standard agreed with META-NET concerning monolingual corpus encoding.
Covered languages: Maltese

Name of the resource: MaltiWordNet of Maltese – Preliminary version – 15,000 entries
Type: Wordnet,
Rationale: Reasons for extension/linking: insufficient number of concepts included given the current state of the art for lexical semantic networks
Upgrade: Linking: Linking will be carried out with a subset of entries in English wordnet
Extend: None
Licensing: n.a. since this is an endogenous resource
Covered languages: MT

Unrestricted exogenous resources

Name of the resource: Illum_Corpus
Type: raw text corpus, written
Upgrade: Improve formal and narrative documentation of corpus. Upgrade corpus to a representation to the standard agreed with META-NET concerning monolingual corpus encoding.
Covered languages: Maltese

Name of the resource: Basic English-Maltese Dictionary
Type: dictionary,
Rationale: The dictionary is currently represented as an online text file. Aim is to extract entries for words represented according to a standard compatible to META-NET such as LMF. This is an essential first step to shortcut manual construction of Maltese Wordnet.

Upgrade: Develop semi-automatic software for translation of entries to compatible standard
Extend: Upgrading extent Complete
Licensing: this is an exogenous resource so that a licence for the use of digital version will need to be obtained from the owner Carlo Farrugia.
Covered languages: EN/MT

Restricted exogenous resources

Name of the resource: MFSA_Companies_Register
Type: raw text corpus, Maltese Company Data
Upgrade: Investigate upgrade of corpus to a representation to standard agreed with META-NET concerning multimodal encoding.
Covered languages: Maltese/english

Name of the resource: Maltese Speech Engine Corpus
Type: speech, Annotated speech corpus
Upgrade: Improve formal and narrative documentation of corpus. Upgrade corpus to a representation to the standard agreed with META-NET concerning monolingual corpus encoding.
Covered languages: Maltese

Name of the resource: Maltese Fiction
Type: raw text corpus, Written
Upgrade: Improve formal and narrative documentation of corpus. Upgrade corpus to a representation to the standard agreed with META-NET concerning monolingual corpus encoding.
Covered languages: Maltese

Name of the resource: Eurowordnet
Type: lexica, Lexicon/Database
Upgrade: Improve formal and narrative documentation of corpus. Upgrade corpus to a representation to the standard agreed with META-NET concerning monolingual corpus encoding.
Covered languages: Maltese

Name of the resource: Acquilina Dictionary MT/EN
Type: lexica, Lexicon
Rationale: The dictionary is currently represented as series of word documents that are essentially page images. Aim is to extract entries for words represented according to a standard compatible to META-NET such as LMF.
Upgrade: : (i) segmentation into separate entries (ii) representation of each entry
Extend: Upgrading extent Approximately 30% of the current content, which means 10,000 entries in the first instance. Of particular importance is the translation field since this will be used for the semi automated construction of the Maltese Wordnet
Licensing: this is an exogenous resource so that a licence for the use of digital version will need to be obtained. Such a licence has already been granted to the University of Arizona for research pruposes. The copyright is currently owned by Midsea Books, Valletta
Covered languages: Maltese/English

Name of the resource: Malta Online Dictionary
Type: dictionary,
Rationale: The dictionary is currently represented as an online text file. Aim is to extract entries for words represented according to a standard compatible to META-NET such as LMF. This is an essential first step to shortcut manual construction of Maltese Wordnet.
Upgrade: Develop semi-automatic software for translation of entries to compatible standard
Extend: Upgrading extent Complete
Licensing: this is an exogenous resource so that a licence for the use of digital version will need to be obtained from the owner Grazio Falzon
Covered languages: MT

Name of the resource: Combined Maltese/English Bilingual Lexicon
Type: Lexicon,
Rationale: Reasons for extension/linking: Aquilina and Busuttil dictionaries have various shortcomings including a lack of recent terminology and the presence of archaic forms. Other online dictionaries (e.g. Falzon) are either specialised or have restricted coverage
Upgrade: Extension: Extend coverage to form a useful subset of the union of all entries
Extend: None
Licensing: Depends on licence obtained with the various stakeholders of all Midsea books who own the copyright for Aquilina dictionary.
Covered languages: EN/MT

Name of the resource: Busuttil Dictionary MT/EN
Type: dictionary,
Rationale: Dictionary is currently a text file segmented by entry. Aim is to upgrade to a standard compatible to META-NET such as LMF. Semi automatic software for carrying out this task will be developed.
Upgrade: : (i) segmentation into separate entries (ii) representation of each entry
Extend: Upgrading extent Depends on the proportion of useful entries since many entries are likely to be outdated so aim is to select useful entries only.
Licensing: this is an exogenous resource so that a licence for the use of digital version will need to be obtained. Currently available through Google Books.
Covered languages: EN/MT

Name of the resource: Busuttil Dictionary EN/MT
Type: dictionary,
Rationale: Dictionary is currently a text file segmented by entry. Aim is to upgrade to a standard compatible to META-NET such as LMF. Semi automatic software for carrying out this task will be developed.
Upgrade: : (i) segmentation into separate entries (ii) representation of each entry
Extend: Upgrading extent Depends on the proportion of useful entries since many entries are likely to be outdated so aim is to select useful entries only.
Licensing: this is an exogenous resource so that a licence for the use of digital version will need to be obtained. Currently available through Google Books.
Covered languages: EN/MT

4.6.2 Software and LR tools

Endogenous resources

Name of the resource: MLRS1 Lexicon Editor
Type: Services, Web Service
Rationale: user interface currently runs only under .NET framework and has a number of missing functionalities in particular regarding extraction of an initial wordlist and morphological analysis of wordforms.
Upgrade: Reimplementation using open source tools. Turn into a web service. Integration with corpus management tools.
Extend: Extension as appropriate. Will not include external database that is currently used for storing textual content.
Licensing: None at present, since software is open source
Covered languages: Any

Name of the resource: MLRS1 Corpus Manager
Type: Services, Web Service
Rationale: user interface currently runs only under .NET framework and has a number of missing functionalities for corpus maintenance, annotation and management of users.
Upgrade: Reimplementation using open source tools and upgrading of current functionality. Handling of bilingual corpora. Incorporation of facilities for managing spoken as well as text resources
Extend: Extension as appropriate. Will not include external database that is currently used for storing textual content.
Licensing: None at present, since software is open source
Covered languages: Any

Name of the resource: MLRS API
Type: Services, Web Service
Rationale: current API includes low-level functionality up to POS tagging. Higher levels of functionality is simply missing. When developed, the act as enablers for other language sensitive tools and services
Upgrade: Development of various higher level functionlaity e.g. (i) shallow parsing modules (ii) deep parsing modules (iii) names entity recognition (iv) stemmer (v) morphological analyser (vi) NP Chunker
Extend: Initially POS tags. Subsequently, development of PNE recognition, coreference resolution, semantic annotation
Licensing: None at present, since software is open source
Covered languages: Any

Name of the resource: MLRS API - POS Tagger
Type: POS Tagger
Rationale: current level of performance is insufficient in terms of accuracy (for POS). Other essential levels of functionality (see below) are simply missing. If developed, they act as enablers for other language-sensitive tools and services.
Upgrade: increase accuracy of POS tagging; develop higher levels of annotation

Extend: Initially, POS tags. Subsequently, development of modules for named-entity recognition, coreference resolution, semantic annotation.
Licensing: None at present, since software is open source
Covered languages:

4.7. Partner: UPC

4.7.1 Data resources

Endogenous resources

Name of the resource: AGORA
Type: annotated speech, annotated speech database
Covered languages: Mainly Catalan ; some Spanish

Name of the resource: ALBAYZIN
Type: speech, annotated speech database
Covered languages: Spanish (1 dialect)

Name of the resource: Bilingual Speech synthesis
Type: synthetic speech, text to speech synthesis
Rationale: Spanish English Bilingual Festival voices
Upgrade: Train and improve the system
Covered languages: Spanish English

Name of the resource: CatalanBN
Type: speech, speech database
Covered languages: Catalan

Name of the resource: Catalan-SpeechDat
Type: speech, speech database
Covered languages: Catalan (5 dialect from Catalonia)

Name of the resource: CHIL UPC Interactive Seminars
Type: speech, multimodal speech database
Rationale: Annotations are partial
Upgrade: The aim is to upgrade that multimodal database by updating the current annotation so that: 1) it covers the whole audiovisual documents, 2) it can also be used with other speech technologies, and 3) it corresponds to a full-scene description that allows to carry out research on the integration of the technology outputs for a multi-level based scene analysis. The final annotation will include labels and time stamps for person gestures, spatial relations, prosody information, emotions, and other features. The annotation will be carried out for five one-hour-long multimodal recordings in order to have audiovisual documents that are multilingual (English, Spanish and Catalan).
Extend: None
Covered languages: European English, non-native speakers. This multimodal and multilingual resource, will be shared with a free-of-charge non-commercial license, mainly focused to the research work on the various involved technologies, and in particular on the integration of their outputs for a multi-level based scene analysis.

Name of the resource: CHIL UPC Seminars
Type: speech, speech database
Covered languages: Spanish

Name of the resource: EUROM.1
Type: speech, speech database
Covered languages: Spanish (1 dialect)

Name of the resource: FESTCAT
Type: speech, speech database
Covered languages: Catalan (1 dialect: central)

Name of the resource: FESTCAT-SEL
Type: speech, speech database
Covered languages: Catalan (1 dialect: central)

Name of the resource: FREE-SPEECH
Type: speech, speech database
Covered languages: Catalan (1 dialect)

Name of the resource: INTERFACE
Type: speech, speech database
Covered languages: Spanish

Name of the resource: LC-STAR CATALAN
Type: lexica, lexicon
Covered languages: Catalan

Name of the resource: LC-STAR Dialogues
Type: speech, speech database
Covered languages: Spanish & Catalan

Name of the resource: LC-STAR SPANISH
Type: lexica, lexicon
Covered languages: Spanish

Name of the resource: SALA-Mexico
Type: speech, speech database
Covered languages: 5 dialects of Spanish Mexican

Name of the resource: SALA-Venezuela
Type: speech, speech database
Covered languages: 5 dialects of Spanish Venezuelan

Name of the resource: Spanish Festival models
Type: synthetic speech, Text to speech synthesis
Rationale: Spanish ready to use HTS voices
Upgrade: Train and improve system

Covered languages: Spanish

Name of the resource: Spanish Festival voices
Type: synthetic speech, text to speech synthesis
Rationale: Spanish ready to use high quality Festival voices
Upgrade: Train and improve system
Covered languages: Spanish

Name of the resource: Spanish SpeechDat (M) and SpeechDat (II)
Type: speech, speech database
Covered languages: Spanish from Spain (5 dialects)

Name of the resource: SpeechDat-Car Catalan
Type: speech, speech database
Covered languages: 5 dialects of Catalan (in Catalonia)

Name of the resource: SpeechDat-Car Spain
Type: speech, speech database
Covered languages: 5 dialects of Spanish (in Spain)

Name of the resource: Speecon Catalan
Type: speech, speech database
Covered languages: 5 dialects of Catalan (in Catalonia)

Name of the resource: TALP TTS0 BASELINES
Type: speech, speech database
Covered languages: Spanish

Name of the resource: TC-STAR TTS BASELINES
Type: speech, speech database
Rationale: Easy to use in standard systems (i.e. Festival)
Upgrade: Add Festival marks
Extend: The full database
Covered languages: Spanish (1 dialect)The same as the original database

Name of the resource: TC-STAR TTS Expressive
Type: speech, speech database
Covered languages: Spanish and English (bilingual corpus and bilingual speakers)

Name of the resource: TC-STAR VC
Type: speech, speech database
Rationale: Easy to use in standard systems (i.e. Festival)
Upgrade: Add Festival marks
Extend: The full database
Covered languages: Spanish and English (bilingual corpus and bilingual speakers). The same as the original database

Restricted exogenous resources

Name of the resource: BN RadioBCN

Type: speech, speech database
Covered languages: Spanish

Name of the resource: EL_PERIODICO_97-07
Type: raw text corpus, newspaper
Rationale: Texts are old (97-07)
Upgrade: Extend the database
Extend: 5 (2008-2012) additionnel years
Covered languages: bilingual Spanish/Catalan

Name of the resource: LAS CORTES
Type: speech, speech database
Covered languages: Spanish

Name of the resource: SPANISH EPPS
Type: speech, speech database
Covered languages: Spanish

Name of the resource: Speecon Spanish (SVOX)
Type: speech, speech database
Covered languages: 5 dialects of Spanish (in Spain)

4.7.2 Software and LR tools

Endogenous resources

Name of the resource: Gaia
Type: Services, Platform to integrate speech-to-speech translation components
Covered languages: n.a. (charset iso-latin1, interface, English)

Name of the resource: NannyRecord
Type: LR tools, Multichannel Speech Recording Plataform
Covered languages: n.a. (charset iso-latin1, interface, English)

Name of the resource: Saga
Type: LR tools, Rule-based phonetic transcription
Covered languages: Spanish (Spain, and Latin-american dialects).

Unrestricted exogenous resources

Name of the resource: Festival
Type: LR tools, Text to speech synthesis system
Covered languages: Spanish, English, Catalan, and other

Name of the resource: HTS
Type: LR tools, Synthetic waveform generation
Covered languages: English, Japanese, Chinese, and other

4.8. Partner: UPF

4.8.1 Data resources

Endogenous resources

Name of the resource: UPF_Term
Type: lexica, Terminology bank
Rationale: update to META-SHARE compliance: described by Metadata and UTF-8
Upgrade: Metadata annotation, conversion to UTF and LMF
Covered languages: Spanish / Catalan / English / French / …

Name of the resource: Neologisms of the year: Bank of Spanish and Catalan Neologisms
Type: lexica, Lexical resource - neologisms
Rationale: update to META-SHARE compliance: Metadata, UTF-8 and LMF
Upgrade: Metadata annotation, conversion to UTF and LMF
Covered languages: Spanish

Name of the resource: Neologisms of the year: Bank of Spanish and Catalan Neologisms
Type: lexica, Lexical resource - neologisms
Rationale: update to META-SHARE compliance: Metadata, UTF-8 and LMF
Upgrade: Metadata annotation, conversion to UTF and LMF
Covered languages: Catalan

Name of the resource: Multilingual Vocabulary of Economics
Type: lexica, Lexical Resource - equivalents
Rationale: update to META-SHARE compliance: Metadata, UTF-8 and LMF
Upgrade: Metadata annotation, conversion to UTF and LMF
Covered languages: Spanish / Catalan / English / Galician / Basque

Name of the resource: IULA Technical Corpus
Type: annotated corpus, Written annotated corpus (POS-tagged)
Rationale: Reasons for extension/linking: given the current state of the art and possible applications that build on this resource (parsing, SMT, etc.), to actually support such usage, a treebank needs to be associated to this corpus both monolingually and for the parallel, aligned parts of it, and by linking it to other treebanks of the project by adding WSJ translations.
Upgrade: PoS conversion to agreed standards and extension to syntactic information (dependency and constituency information) by means of enlarging to a treebank
Extend: 40,000 sentences treebanked
Licensing: PoS conversion to agreed standards and extension to syntactic information (dependency and constituency information) by means of enlarging to a treebank. The resulting treebank will be publicly available resources, most probably under a GNU-GPL license or similar.
Covered languages: Spanish / Catalan / English

Name of the resource: Genoma corpus
Type: raw text corpus, Written corpus

Rationale: update to META-SHARE compliance: described by Metadata and UTF-8
Upgrade:
Covered languages: Spanish / Catalan

Name of the resource: Corpus PAAU 92
Type: raw text corpus, Written corpus
Upgrade:
Covered languages: Spanish

Name of the resource: Basic Vocabulary on the Human Genome
Type: lexica, Lexical Resource - equivalents
Rationale: update to META-SHARE compliance: Metadata, UTF-8 and LMF
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Spanish / Catalan / English / Galician / Basque

Name of the resource: Parallel IULA Technical Corpus
Type: parallel corpus, Written aligned
Rationale: For new models of SMT it is important to have aligned treebanks.
Upgrade: Metadata annotation
Extend: 20,000 sentences treebanked
Licensing: Treebanking of 20.000 sentences (aprox.). The resulting treebank will be publicly available resources, most probably under a GNU-GPL license or similar.
Covered languages: Spanish / Catalan / English

Unrestricted exogenous resources

Name of the resource: Apertium Bilingual dictionary Spanish-Galician
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata and LMF
Upgrade: Metadata annotation
Covered languages: Spanish-Galician

Name of the resource: Apertium Bilingual dictionary Basque-Spanish
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata and LMF
Upgrade: Metadata annotation
Covered languages: Basque-Spanish

Name of the resource: Apertium Bilingual dictionary CA-ES
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata and LMF
Upgrade: Metadata annotation
Covered languages: Spanish-Catalan

Name of the resource: Apertium Bilingual dictionary English-Catalan
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata and LMF
Upgrade: Metadata annotation
Covered languages: English-Catalan

Name of the resource: Apertium Bilingual dictionary English-Galician
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata and LMF
Upgrade: Metadata annotation
Covered languages: English-Galician

Name of the resource: Apertium Bilingual dictionary English-Spanish
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata and LMF
Upgrade: Metadata annotation
Covered languages: English-Spanish

Name of the resource: Apertium Bilingual dictionary French-Catalan
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata and LMF
Upgrade: Metadata annotation
Covered languages: French-Catalan

Name of the resource: Apertium Bilingual dictionary French-Spanish
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata and LMF
Upgrade: Metadata annotation
Covered languages: French-Spanish

Name of the resource: Apertium Bilingual dictionary Occitan-Catalan
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata and LMF
Upgrade: Metadata annotation
Covered languages: Occitan-Catalan

Name of the resource: Apertium Bilingual dictionary Occitan-Spanish
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata and LMF
Upgrade: Metadata annotation
Covered languages: Occitan-Spanish

Name of the resource: Apertium Bilingual dictionary Portuguese-Catalan
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata and LMF
Upgrade: Metadata annotation
Covered languages: Portuguese-Catalan

Name of the resource: Apertium Basque dictionary
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata, UTF-8 and LMF
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Basque

Name of the resource: Apertium Bilingual dictionary Spanish-Asturian

Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata and LMF
Upgrade: Metadata annotation
Covered languages: Spanish-Asturian

Name of the resource: Spanish Wordnet 3.0
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata, UTF-8 and LMF
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Spanish

Name of the resource: Apertium Bilingual dictionary Spanish-Portuguese
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata and LMF
Upgrade: Metadata annotation
Covered languages: Spanish-Portuguese

Name of the resource: Apertium Bilingual dictionary Spanish-Romanian
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata and LMF
Upgrade: Metadata annotation
Covered languages: Spanish-Romanian

Name of the resource: Apertium Catalan dictionary
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata, UTF-8 and LMF
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Catalan

Name of the resource: Apertium Galician dictionary
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata, UTF-8 and LMF
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Galician

Name of the resource: Apertium Spanish dictionary
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata, UTF-8 and LMF
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Spanish

Name of the resource: FreeLing Asturian dictionary
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata, UTF-8 and LMF
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Asturian

Name of the resource: FreeLing Catalan dictionary
Type: lexica, lexicon

Rationale: update to META-SHARE compliance: Metadata, UTF-8 and LMF
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Catalan

Name of the resource: FreeLing Catalan sense dictionary
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata, UTF-8 and LMF
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Catalan

Name of the resource: FreeLing Galician dictionary
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata, UTF-8 and LMF
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Galician

Name of the resource: FreeLing Spanish dictionary
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata, UTF-8 and LMF
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Spanish

Name of the resource: FreeLing Spanish sense dictionary
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata, UTF-8 and LMF
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Spanish

Name of the resource: Apertium Bilingual dictionary Portuguese-Galician
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata and LMF
Upgrade: Metadata annotation
Covered languages: Portuguese-Galician

Restricted exogenous resources

Name of the resource: Electronic Corpus of Academic Materials – University of Zaragoza (ECAM-UZ)
Type: annotated corpus, Written + oral corpus
Rationale: update to META-SHARE compliance: described by Metadata and UTF-8
Upgrade: Metadata annotation, conversion to UTF and LMF
Covered languages: English, Spanish

Name of the resource: AnCora-Co-CA
Type: corpus, corpus
Rationale: update to META-SHARE compliance: Metadata, UTF-8
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Catalan

Name of the resource: AnCora-Co-ES

Type: corpus, corpus
Rationale: update to META-SHARE compliance: Metadata, UTF-8
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Spanish

Name of the resource: AnCora-Es
Type: corpus, corpus
Rationale: update to META-SHARE compliance: Metadata, UTF-8
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Spanish

Name of the resource: Euskal Wordnet 3.0
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata, UTF-8
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Basque

Name of the resource: 6305-QC
Type: raw text corpus, Questions for Question Answering Classification
Rationale: update to META-SHARE compliance: Metadata, UTF-8 and LMF
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Spanish

Name of the resource: CESS_EU: The Basque Dependency Treebank
Type: raw text corpus, Written Corpus
Rationale: update to META-SHARE compliance: described by Metadata and UTF-8
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Basque

Name of the resource: Corpus CLUVI
Type: raw text corpus, Written Corpus
Rationale: update to META-SHARE compliance: described by Metadata and UTF-8
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Basque, Catalan, English, French, Galician, German, Portuguese, Spanish

Name of the resource: Corpus Técnico do Galego
Type: annotated corpus, Written Corpus
Rationale: update to META-SHARE compliance: described by Metadata and UTF-8
Upgrade: Metadata annotation, conversion to UTF and LMF
Covered languages: Galician

Name of the resource: AnCora-Ca
Type: corpus, corpus
Rationale: update to META-SHARE compliance: Metadata, UTF-8
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Catalan

Name of the resource: DOGC CAT-SPA Parallellized Corpus

Type: annotated corpus, Aligned Corpus
Rationale: update to META-SHARE compliance: described by Metadata and UTF-8
Upgrade: Metadata annotation, conversion to UTF and LMF
Covered languages: Catalan, Spanish

Name of the resource: VOLEM
Type: lexica, Lexicon / Knowledge Source
Rationale: update to META-SHARE compliance: described by Metadata and UTF-8
Upgrade: Metadata annotation, conversion to UTF and LMF
Covered languages: Spanish, Catalan

Name of the resource: European Community Law Catalan Glossary mapped to EUROVOC
Type: lexica, Terminological Resource
Rationale: update to META-SHARE compliance: described by Metadata and UTF-8
Upgrade: Metadata annotation, conversion to UTF and LMF
Covered languages: Catalan

Name of the resource: PAROLE lexicon
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata, UTF-8 and LMF
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Catalan

Name of the resource: PAROLE lexicon
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata, UTF-8 and LMF
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Spanish

Name of the resource: SenSem Corpus
Type: raw text corpus, Written Corpus
Rationale: update to META-SHARE compliance: described by Metadata and UTF-8
Upgrade: Metadata annotation, conversion to UTF and LMF
Covered languages: Spanish

Name of the resource: SenSem Database
Type: lexica, Lexicon / Knowledge Source
Rationale: update to META-SHARE compliance: Metadata, UTF-8 and LMF
Upgrade: Metadata annotation, conversion to UTF and LMF
Covered languages: Spanish

Name of the resource: SIMPLE lexicon
Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata, UTF-8 and LMF
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Catalan

Name of the resource: SIMPLE lexicon

Type: lexica, lexicon
Rationale: update to META-SHARE compliance: Metadata, UTF-8 and LMF
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Spanish

Name of the resource: Spanish FrameNet
Type: lexica, Lexicon / Knowledge Source
Rationale: update to META-SHARE compliance: described by Metadata and UTF-8
Upgrade: Metadata annotation, conversion to UTF and LMF
Covered languages: Spanish

Name of the resource: Termoteca
Type: lexica, Terminological Resource
Rationale: update to META-SHARE compliance: described by Metadata and UTF-8
Upgrade: Metadata annotation, conversion to UTF and LMF
Covered languages: English, French, Galician, Spanish

Name of the resource: Dicionario CLUVI inglés-galego
Type: lexica, Lexicon / Knowledge Source
Rationale: update to META-SHARE compliance: described by Metadata and UTF-8
Upgrade: Metadata annotation, conversion to UTF and LMF
Covered languages: English, Galician

Name of the resource: Computer Science Tri-lingual Corpus
Type: corpus, corpus
Rationale: update to META-SHARE compliance: Metadata, UTF-8 and TEI/CES conformant
Upgrade:
Covered languages: English-Spanish-Catalan

4.8.2 Software and LR tools

Endogenous resources

Name of the resource: Tools for Spanish Corpus Processing
Type: LR tools, Corpus Processing
Rationale: no stand-off annotation format possible as input or output. Necessary for extensions to treebank
Upgrade: rewrite resource code or implement wrappers
Licensing: The resulting version with xml based stand-off markup input/output of the different modules will be made publicly available, most probably under a GNU-GPL license or similar
Covered languages: Spanish

Name of the resource: Tools for Catalan Corpus Processing
Type: LR tools, Corpus Processing
Rationale: no stand-off annotation format possible as input or output. Necessary for extensions to treebank
Upgrade: rewrite resource code or implement wrappers

Licensing: The resulting version with xml stand-off markup input/output of the different modules will be made publicly available, most probably under a GNU-GPL license or similar
Covered languages: Catalan

Name of the resource: Tools for automatic UTF-8 conversion
Type: LR infrastructure tools, Cleaning and pre-processing
Rationale: automatize the integration into META-SHARE
Upgrade:
Licensing: The resulting software will be publicly available resources without restrictions
Covered languages: n/a