Report on second selection of resources, revising selection in D2.1

Deliverable D2.5

Version 1.5

2012-01-31

Editors: 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	Jan 23, 2012	A. Moreno	UPC	First dratf
V1.1	Jan 25, 2012	A. Moreno	UPC	1st contribution of partners
V1.2	Jan 27, 2012	A. Moreno	UPC	2nd contribution of partners
V1.3	Feb 2, 2012	A. Moreno	UPC	3rd contribution of partners
V1.4	Feb 3, 2012	A. Moreno	UPC	Pre-final for QC
V1.5	Feb 7, 2012	A. Moreno	UPC	Final

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise.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

METANET4U

Report on second selection of resources, revising selection in D2.1

Document METANET4U-2012-D2.5
EC CIP project #270893

Deliverable
Number: D2.5
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

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 *Deliverable D2.1 Report on first seletion of resources* listed the resources that were prepared in WP3 and delivered in month M10 as a first batch of resources (Batch 1) as well as a selecion of resources and tools to be delivered later on.

The objective of *Deliverable D2.5 Report on second seletion of resources* is to describe the remaining language resources that 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 list includes a **new selection of resources** that will be available though Metashare. The new list of resources **focuses on exogeneous resources,** as we believe that this will have the effect of extending the community around META-NET,

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 2 and 3. Section 4 contains a list and a short description of the new language resources. This list complements the Language resources described in Deliverable 2.1.

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.

Next section shows the list of resources that every partner is willing to deliver. For an easy understandability of the tables showed in section 3, we include here the definition of the descriptors as they were defined in D2.1

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. Deliverable D2.1 includes, 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…). **The same information is included in this deliverable for the new language resources**

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 ones are chosen. Therefore it is important to have a sensible perspective in this respect from the offset. 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 able to gather them. Even for endogenous resources, it is very likely that not all are immediately ready to be distributed, as some of them are indicated as requiring 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 following table summarizes the number of resources that each partner is going to deliver in the remaining part of the project.

	1. ULX	2. IST	3. UNIMAN	4. UAIC	5. RACAI	6. UOM	7. UPC	8. UPF	Total
Data	35	12	0	7	9	17	35	24	139
Batch 2	31	6	0	0	5	8	33	12	95
Endogenous resources	11	5			4	2	15	2	39
Restricted exogenous resources	13	1			1	5	18	6	44
Unrestricted exogenous resources	7					1		4	12
Batch 3	4	6	0	7	4	9	2	12	44
Endogenous resources	4	6		2	3	2	2	3	22
Restricted exogenous resources	0			5	1	7		8	21
Unrestricted exogenous resources	0							1	1
Software	11	1	21	19	17	4	5	4	82
Batch 2	6	0	13	8	9	0	1	4	41
Endogenous resources			3	8	9			4	24
Restricted exogenous resources			6				1		7
Unrestricted exogenous resources	7		4						11
Batch 3	4	1	8	11	8	4	4	0	40
Endogenous resources	4	1	3	10	6	4			28
Restricted exogenous resources			5	1	1		3		10
Unrestricted exogenous resources					1		1		2
Total	46	13	21	26	26	21	40	28	221

The following tables show, for each partner, the resources they plan to deliver in Batches 2 (Month 18) and 3 (Month 24), as well as the effort in PMs for each resource.

3.1 ULX - University of Lisbon

Data/Soft	Where	Delivery	Designation	Total
Data	Endogenous resources	Batch 2	Abbreviations	0,1
			C-ORAL-ROM Portuguese Corpus	6
			CINTIL-Internacional Corpus of Portuguese	5
			Lexicon of multiword expressions	6
			MWN.PT	0,15
			PAROLE corpus	0,2
			PAROLE lexicon	0,2
			PropBank	12
			SIMPLE lexicon	0,2
			Stopwords	0,1
			Treebank	12
		Total Batch 2		41,95
		Batch 3	DependencyBank	12
			LogicalFormBank	8
			LT Corpus	7
			Technical Corpus	8
		Total Batch 3		35
	Total Endogenous resources			76,95
	Restricted exogenous resources	Batch 2	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		2
	Total Restricted exogenous resources			2
	Unrestricted exogenous resources	Batch 2	CETEMPúblico	0,15
			Corpus NILC	0,15
			CorpusTCC	0,15
			PLN-BR Gold	0,15
			RHETALHO	0,15
			Summ-it	0,15
			TeMário 2006	0,15
		Total Batch 2		1,05
	Total Unrestricted exogenous resources			1,05

Data/Soft	Where	Delivery	Designation	Total
Total Data				80
Software	Endogenous resources	Batch 3	Chunker	0,1
			LX-Service	4
			Tagger	0,1
			Tokenizer	0,1
		Total Batch 3		4,3
	Total Endogenous resources			4,3
	Unrestricted exogenous resources	Batch 2	DiZer 2.0	0,15
			Forma	0,15
			GistSumm	0,15
			NILC Taggers	0,15
			Ontolp Plugin	0,15
			Stemmer	0,15
			Text Aligners	0,15
		Total Batch 2		1,05
	Total Unrestricted exogenous resources			1,05
Total Software				5,35
TOTAL ULX				85,35

3.2. IST - Instituto Superior Técnico

Data/Soft	Where	Delivery	Designation	Total
Data	Endogenous resources	Batch 2	TED Talks (3)	4
			CALL	4
			PTSTAR Golden Collection	1
		Total Batch 2		9
		Batch 3	LECTRA (6)	4
		Total Batch 3		4
	Total Endogenous resources			13
	Restricted exogenous resources	Batch 2	TAP	4
		Total Batch 2		4
	Total Exogenous resources			4
Total Data				17
Software	Endogenous resources	Batch 3	Etxt2DB	8
		Total Batch 3		8
	Total Endogenous resources			8
Total Software				8
Total IST				25

3.3. UNIMAN-University of Manchester

Data/Soft	Where	Delivery	Designation	Total
Software	Endogenous resources	Batch 2	U-Compare NaCTeM Sentence Detector	0,15
			U-compare platform	0,15
			U-Compare Workbench	0,15
		Total Batch 2		0,45
		Batch 3	NEMINE	0,15
			STEPP	0,15
			U-Compare STEPP PoS Tagger	0,15
		Total Batch 3		0,45
	Total Endogenous resources			0,9
	Restricted exogenous resources	Batch 2	U-Compare GENIA Sentence Detector	0,10
			U-Compare GENIA Tokenizer	0,10
			U-Compare OpenNLP PoSTagger	0,10
			U-Compare OpenNLP Sentence Detector	0,10
			U-Compare OpenNLP Tokenizer	0,10
			U-Compare Type System	0,10
		Total Batch 2		0,7
		Batch 3	Enju	0,10
			Genia tagger/chunker and NER	0,15
			U-Compare Enju Parser	0,10
			U-Compare GENIA PoS Tagger	0,10
			U-Compare Stanford Parser	0,10
		Total Batch 3		0,55
	Total Restricted exogenous resources			1,25
	Unrestricted exogenous resources	Batch 2	Apertium Morphological Analyser	0,10
			Apertium POS tagger	0,10
			Apertium MT transfer	0,10
			Apertium Mophological generator	0,10
	Total Unrestricted Exogenous Resources			0,4
Total Software				2,55
Total UNIMAN				2,55

3.4. UAIC - University Alexandru Ioan Cuza

Data/Soft	Where	Delivery	Designation	Total
Data	Endogenous resources	Batch 3	eDTLR-sources	0,5
			RomMorph-UAIC	0,5
		Total Batch 3		1
	Total Endogenous resources			1
	Restricted exogenous resources	Batch 3	DEA	1,5
			DLPE	1,5
			eDTLR	2,4
			RoWN-eDTLR	2,7
			SRoL – Sounds of the Romanian Language	0,5
		Total Batch 3		8,6
	Total Restricted exogenous resources			8,6
Total Data				9,6
Software	Endogenous resources	Batch 2	Categorizer-UAIC	2,5
			DP-UAIC	3
			Lemmatizer-UAIC	2
			NP-chunker-UAIC	3
			RARE-RO-UAIC	4
			Splitter-UAIC	2,5
			Summarizer-UAIC	2
			Tokenizer-UAIC	1
		Total Batch 2		20
		Batch 3	ALPE-UAIC	4
			AnaMorph-UAIC	3
			Diacritics-UAIC	0,5
			FDGparser-UAIC	2
			Language identifier-UAIC	1
			Occurrence Finder-UAIC	2
			OntologyBuilder-UAIC	2
			QA-UAIC	3
			SRL-UAIC	3
			TE-UAIC	3
		Total Batch 3		23,5
	Total Endogenous resources			43,5
	Restricted exogenous resources	Batch 3	ANNIE	0,5
		Total Batch 3		0,5
	Total Restricted exogenous resources			0,5
Total Software				44
Total UAIC				53,6

3.5. RACAI - Romanian Academy

Data/Soft	Where	Delivery	Designation	Total
Data	Endogenous resources	Batch 2	Mapping list from PWN2.0 to PWN3.0	1
			NAACL 2003	0.5
			ROMORPH	1
			RO-WORDNET (part 2)	5.5
		Total Batch 2		8
		Batch 3	Strongly comparable corpus En-FR-RO, tagged, lemmatized and aligned	3.5
			Romanian-French conversation dictionary	1
			RO-EN JRC-Acquis	1.5
		Total Batch 3		6
	Total Endogenous resources			14
	Restricted exogenous resources	Batch 2	Romanian WEB 1T 5	2
		Total Batch 2		2
		Batch 3	CoDII-NPI.ro	0.5
		Total Batch 3		0.5
	Total Restricted exogenous resources			2.5
Total Data				16.5

Data/Soft	Where	Delivery	Designation	Total
Software	Endogenous resources	Batch 2	COLLOC	1
			LangId	1
			LexChain	1
			LexPar	1
			RO-HYPHEN	1
			TTL-lemmatizer	1
			TTL-Tagger	2
			TTL-Tokenizer	2
			TTL-chunker	1
		Total Batch 2		9
		Batch 3	VoiceForge	2
			Lexacc	3
			YAWA	2
			WN-Builder	1
			DIAC+	1
			SynWSD	1
		Total Batch 3		10
	Total Endogenous resources			19
	Restricted exogenous resources	Batch 3	VISL Multilingual Dependency-Parser	1
	Total Restricted exogenous resources			1
	Unrestricted exogenous resources	Batch 3	LUCON	1
	Total unrestricted exogenous resources			1
	Total exogenous resources			2
Total Software				21
Total RACAI				39.5

3.6. UOM - University of Malta

Data/Soft	Where	Delivery	Designation	Total
Data	Endogenous resources	Batch 2	F_MONA_1	0,4
			MaIToBI Corpus	1
		Total Batch 2		1,4
		Batch 3	MaltiWordNet of Maltese – Preliminary version – 15,000 entries	3
			MLRS Corpus	1
		Total Batch 3		4
	Total Endogenous resources			5,4
	Restricted exogenous resources	Batch 2	Local Government documentation	1,5
			Malta Online Dictionary	1
			Maltese Speech Engine Corpus	1
			Maltese Wikipedia	1
			SPAN	1
		Total Batch 2		4,7
		Batch 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
			Maltese Fiction	0,1
			MFSA_Companies_Register	0,2
		Total Batch 3		6,5
	Total Restricted exogenous resources			11,2
	Unrestricted exogenous resources	Batch 2	Basic English-Maltese Dictionary	0,2
		Total Batch 2		1
	Total Unrestricted exogenous resources			1
Total Data				17,6
Software	Endogenous resources	Batch 3	MLRS API	3
			MLRS API - POS Tagger	1
			MLRS1 Corpus Manager	0,5
			MLRS1 Lexicon Editor	0,5
		Total Batch 3		5
	Total Endogenous resources			5
Total Software				5
Total UOM				22,6

3.7. UPC – Universitat Politècnica de Catalunya

Data/Soft	Where	Delivery	Designation	Total
Data	Endogenous resources	Batch 2	ALBAYZIN	0,1
			Catalan-SpeechDat(I)	0,1
			SALA-Mexico	0,1
			SALA-Venezuela	0,1
			Spanish SpeechDat (II)	0,1
			SpeechDat-Car Spain	0,1
			Speecon Catalan	0,1
			Interface expressive database	0,1
			LC-STAR Catalan Phonetic Lexicon	0,1
			LC-STAR Spanish Phonetic Lexicon	0,1
			UPC-ESMA	0,1
			TC-STAR Spanish Baseline Female	2
			TC-STAR Spanish Baseline Male	2
			TC-STAR Bilingual Expressive Speech	0,1
			TC-STAR Bilingual VC	4
		Total Batch 2		9,2
		Batch 3	CHIL UPC Interactive Seminars	20
			CHIL UPC Seminars	0,1
		Total Batch 3		20,1
	Total Endogenous resources			29,3
	Restricted exogenous resources	Batch 2	Ahosyn: Large Bilingual Speech Database for Synthesis	0,2
			Bizkaifon: speech and video database for the Western dialects of the Basque Language	0,2
			EL_PERIODICO_97-07	7,9
			EmodB_EU1: Emotional speech and video database in Standard Basque	0,2
			EmodB_EU2: Emotional speech database in Standard Basque	0,2
			Galician SpeechDat FDB	0,2
			LAS CORTES	0,1
			SPANISH EPPS	0,1
			Speech Rate database for Basque	0,2
			Speech-Dat like database for Basque	0,2
			Speech-dat like database for Basque (Mobile).	0,2
			Transgrigal DB	0,2
			DOGalicia: Parallel Galician-Spanish Corpus	0,2
			GCG: GrupoCorreoGalego	0,2
		Total Batch 2		10,3
	Total Restricted exogenous resources			10,3
Total Data				39,6

Data/Soft	Where	Delivery	Designation	Total
Software	Endogenous resources	Batch 3	Gaia	0,1
			NannyRecord	0,1
			Saga	0,1
		Total Batch 3		0,3
	Total Endogenous resources			0,3
	Restricted exogenous resources	Batch 2	Cotovia Transcriber	0,1
		Total Batch 2		0,1
		Batch 3	Segre	0,1
		Total Batch 3		0,1
	Total Restricted exogenous resources			0,2
Total Software				0,5
Total UPC				40,1

3.8. UPF - University Pompeu Fabra

Data/Soft	Where	Delivery	Designation	Total pm
Data	Endogenous resources	Batch 2	News paper headlines corpus	0,2
			TRL V-Subcat Lexicon	0,2
		Total Batch 2		0,4
		Batch 3	IULA Technical Corpus	11
			Parallel IULA Technical Corpus	19
			Spanish - Translated Penn Tree Bank chapter 23	4
		Total Batch 3		34
	Total Endogenous resources			34,4
	Restricted exogenous resources	Batch 2	CESS_EU: The Basque Dependency Treebank	0,1
			Corpus CLUVI	0,1
			Corpus Técnico do Galego	0,1
			Dicionario CLUVI inglés-galego	0,1
			Euskal Wordnet 3.0	0,05
			Termoteca	0,1
		Total Batch 2		0,55
		Batch 3	AnCora-Ca	0,05
			AnCora-Co-CA	0,05
			AnCora-Co-ES	0,05
			AnCora-Es	0,05
			Computer Science Tri-lingual Corpus	0,1
			SenSem Corpus	0,1
			SenSem Database	0,1
			VOLEM	0,1
		Total Batch 3		0,6
	Total Restricted exogenous resources			1,15
	Unrestricted exogenous resources	Batch 2	Apertium English dictionary	0,1
			Apertium French dictionary	0,1
			Apertium Italian dictionary	0,1
			WikiCorpus	6
		Total Batch 2		6,3
		Batch 3	Spanish Wordnet 3.0	0,1
		Total Batch 3		0,1
	Total Unrestricted exogenous resources			6,4
Total Data				46,5

Data/Soft	Where	Delivery	Designation	Total
Software	Endogenous resources	Batch 2	Converters to LMF 2	2
			Tools for automatic UTF-8 conversion	6
			Tools for Catalan Corpus Processing	0
			Tools for Spanish Corpus Processing	12
		Total Batch 2		20
	Total Endogenous resources			20
Total Software				20
Total UPF				46,5

4 List of New Resources

Section 4 of Deliverable D2.1 has a short description of the resources that were chosen, at the beginning of the project, to be delivered during the lifetime of the project. This section contains a description of the new resources selected by the partners and that were not described in D2.1. This section, therefore, complements Deliverable 2.1

4.1. Partner: ULX

4.1.1 Data resources

Endogenous resources

Name of the resource: Lexicon of multiword expressions
Type: lexicon
Upgrade: full documentation: corpus used, metadata, criteria, format.
Annotation: MWE typology, frequency and lexical association measure, concordances
Covered languages: Portuguese
Effort: 6 pm

Name of the resource: LT Corpus
Type: corpus, raw text corpus
Upgrade: adapt to a TEI-compliant format; complete documentation, metadata decription and secure the copyrights; update POS annotation, lemmatisation
Covered languages: Portuguese
Effort: 7pm

Name of the resource: LogicalFormBank
Type: corpus, annotated corpus
Rationale: 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
Effort: 8pm

4.1.2 Software and LR tools

Endogenous resources

Name of the resource: LX-Service
Type: LT tool
Upgrade: documentation and metadata; upgrading to compliance with standards; update methods and functionalities
Covered languages: Portuguese
Effort: 4pm

4.2. Partner: UNIMAN

The resources Enju, Genia tagger/chunker and NER, U-Compare Enju Parser, U-Compare GENIA PoS Tagger, U-Compare GENIA Sentence Detector , U-Compare GENIA Tokenizer, U-Compare OpenNLP PoSTagger, U-Compare OpenNLP Sentence Detector, U-Compare OpenNLP Tokenizer, U-Compare Stanford Parser, were presented and described in deliverable 2.1, although 0 person-month effort was associated with them. The resources NEMINE, STEPP, U-Compare NaCTeM Sentence Detector, U-Compare STEPP PoS Tagger, and U-Compare Workbench were assigned 0 pm but were not described. UNIMAN will however allocate efforts in batch 2 and batch 3 for all these resources, and they will be delivered via MetaShare. The following text describes the last mentioned six resources plus the new resources.

4.2.1 Software and LR tools

Endogenous resources

Name of the resource: NEMINE
Type: LR tools
Content: named entity recogniser
Language: English
Upgrade: documentation and metadata
Effort: 0,15

Name of the resource: STEPP
Type: LR tools
Content: Tagger
Language: English
Upgrade: documentation and metadata
Effort: 0,15

Name of the resource: U-Compare NaCTeM Sentence Detector
Type: LR tools
Content: Sentence splitter
Language: English
Upgrade: documentation and metadata
Effort: 0,15

Name of the resource: U-Compare STEPP PoS Tagger
Type: LR tools
Content: PoS Tagger
Language: English
Upgrade: documentation and metadata
Effort: 0,15

Name of the resource: U-Compare Workbench
Type: LR tools
Content: Workflow management tool
Language: Any

Upgrade: documentation and metadata
Effort: 0,15

Unrestricted Exogenous resources

Name of the resource: U-Compare Apertium morphological analyser
Type: NLP tool
Content: Morphological analyser – module of the Aperium machine translation system
Covered languages: English, Catalan, Basque, Spanish, Portuguese
Upgrade: Documentation and metadata
Effort: 0.10PM

Name of the resource: U-Compare Apertium POS tagger
Type: NLP tool
Content: Part of speech tagger – module of the Apertium machine translation system
Covered languages: English, Catalan, Basque, Spanish, Portuguese
Upgrade: Documentation and metadata
Effort: 0.10PM

Name of the resource: U-Compare Apertium MT transfer
Type: NLP tool
Content: Machine translation transfer – module of the Apertium machine translation system
Covered languages: English, Catalan, Basque, Spanish, Portuguese
Upgrade: Documentation and metadata
Effort: 0.10PM

Name of the resource: U-Compare Apertium morphological generator
Type: NLP tool
Content: Mophological genenerator – module of the Apertium machine translation system
Covered languages: English, Catalan, Basque, Spanish, Portuguese
Upgrade: Documentation and metadata
Effort: 0.10 pm

4.3. Partner: UAIC

The endogenous resources Categorizer-UAIC, Splitter-UAIC, Diacritics-UAIC , LanguageIdentifier-UAIC and the restricted exogenous resources ANNIE, DEA, DLPE and SRoL – Sounds of the Romanian language, were presented and described in deliverable 2.1, although 0 person-month effort was associated with them. UAIC will however allocate efforts in batch 2 and batch 3 for these resources, and they will be delivered via MetaShare.

4.4. Partner: RACAI

4.4.1 Language resources

Endogenous resources

Name of the resource: NAACL 2003
Type: parallel annotated corpus
Upgrade detailed description document, including various statistics; conversion of SGML entities into UNICODE characters; updating the text to the 2005 morphological rules; adding standardized metadata.
Covered languages: English, Romanian
Effort: 0.5 pm

Name of the resource: Romanian-French conversation dictionary
Type: bilingual dictionary
Upgrade detailed description document, including various statistics; standardized character encoding; re-encoding of the dictionary according to a standard specification (META-NET recommended one)
Covered languages: English, Romanian
Effort: 1.0 pm

Name of the resource: ROPMORPH
Type: unification-based morphology
Upgrade improving lemmatization of the out of vocabulary words ; writing documentation
Covered languages: Romanian
Effort: 1.0 pm

Name of the resource: Mapping list from WN2.0 to WN3.0
Type: lexical ontology
Upgrade improving coverage
Covered languages: Romanian
Effort: 1.0 pm

Name of the resource: RO-EN JRC-Acquis
Type: 30 Million token annotated corpus
Upgrade 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; additionally, we intend to finer-grain align it at sentence and word level with English (with this information reflected into the metadata)
Covered languages: English, Romanian
Effort: 1.5 pm

Name of the resource: News Multilingual Annotated Strongly Comparable Corpus (Corpus En-Fr-Ro)
Type: corpus, tagged, lemmatized and document aligned
Upgrade: this is a new resource, extracted by a specially designed crawler that continuously watch the New site of European Parliament. The document are aligned,

and each document in the three languages is tagged and lemmatized; metadata description is generated at the crawling time;
Covered languages: English, French, Romanian
Effort: 3.5 pm

Name of the resource: RO-WordNet (version 2)
Type: lexical ontology
Upgrade: adding 30,000 more synsets (that is doubling version 1); fully aligned (including the first 30,000 synsets of version 1) to PWN3.0; SUMO, DOMAIN3.0 explicitly added; massive corrections in the definitions, updated synsets (more literals added), typos corrections, etc.
Covered languages: Romanian
Effort: 5.5 pm

4.4.2 Software and LR tools

Name of the resource: Lexacc
Type: lexical processing
Upgrade:
Covered languages: Romanian
Effort: 3.0 pm

4.5. Partner: UOM

4.5.1 Data resources

Endogenous resources

Name of the resource: MaIToBI Corpus
Type: Speech, multimodal corpus
Rationale: Data is incompletely and inconsistently annotated and has limited online accessibility. Documentation, both narrative and formal, is currently fragmentary. However, much of it is not easily accessible to potential beneficiaries because of incomplete coverage, those in a position to benefit from it.
Upgrade: Improve formal and narrative documentation of corpus. Upgrade corpus to a representation to standard agreed with META-NET concerning multimodal encoding.
Extend: Aim will be to ensure that all existing information in corpus is upgraded in a consistent way.
Licensing: none
Covered languages: Standard Maltese
Effort: 1pm

Restricted exogenous resources

Name of the resource: Maltese Wikipedia
Type: written

Rationale: Reasons for extension/linking: development of thesaurus entries.
Upgrade: Linking: Links based on Wikipedia structure made that link to thesaurus entries
Extend: None
Covered languages: Maltese
Effort: 1 pm

Name of the resource: SPAN
Type: speech, part of 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
Covered languages: Standard Maltese
Effort: 1 pm

Name of the resource: Local Government documentation
Type: written
Licensing: under negociation
Covered languages: Maltese
Effort: 1.5 pm

4.6. Partner: UPC

Some research work was carried out to determine the interest of the owners of the exogenous resources listed and described in D2.1 to deliver their resources though Meta-share and to investigate about the quality of their specific resources.

As a result, some databases included in D2.1 are not included in D2.5: SpeechDat (M), Speecon Spanish, Festival, and HTS. The database El_Periódico has poor documentation, only a small part of the texts are aligned, and UPC didn't succeed in gaining a desired 5 year extension to extend/upgrade that resource. As a result, UPC successfully contacted other exogenous resources whose owners agreed to deliver their data though Meta-share. UPC will improve the current documentation of 'El_Periodico' and will align some of the available text. The description of the new resources follows:

4.6.1 Data resources

Endogenous resources

Name of the resource: Catalan-SpeechDat (I)
Type: speech, speech database
Content: sentences, application words, digits, numbers, telephone numbers, money amounts, dates, hours, cities, names...
Covered languages: Catalan (5 dialect from Catalonia)

Restricted exogenous resources

Name of the resource: EmodB_EU1: Emotional speech and video database in Standard Basque
Type: multimodal, speech and video
Content: 170 items (sentences, words) in standard Basque repeated for the 6 MPEG emotions (happiness, sadness, fear, anger, surprise and disgust) and neutral style, for a female voice. The audio was registered at 32kHz, 16bits, professional studio, 1 microphone and laryngograph included. Two cameras were used for video recordings (one frontal, one lateral). Face markers were used. The female voice has been segmented at phone level and manually revised for the neutral style.
Covered languages: Basque

Name of the resource: EmodB_EU2: Emotional speech database in Standard Basque
Type: multimodal, speech and video
Content: 702 phonetically balanced sentences in standard Basque repeated for the 6 MPEG (happiness, sadness, fear, anger, surprise and disgust) emotions and neutral style, for one male and one female voice. It also contains a continuous read speech of 8 min, in the same 7 styles. It was registered at 48kHz, 16bits, semi-professional room, 2 microphones and laryngograph included. Automatically segmented, half of the emotions have been manually checked for the Female voice.
Covered languages: Basque

Name of the resource: Speech-Dat like database for Basque
Type: speech, speech database
Content: The Basque FDB-1060 database contains the recordings of 1,060 speakers of Basque recorded over the fixed telephone network. Each speaker uttered around 43 read and spontaneous items
Covered languages: Basque

Name of the resource: Bizkaifon: speech and video database for the Western dialects of the Basque Language
Type: multimodal, speech and video
Content: Bizkaifon is a multimodal (speech and video) database containing recordings of the many different western dialects of Basque. Most of them are transcribed to Standard Basque. It is accessible via web.
Covered languages: Basque

Name of the resource: Speech Rate database for Basque
Type: speech, speech database
Content: This database contains 12 sentences written in standard Basque and uttered by 9 male speakers at slow, normal and fast rate, each repeated 7 times (total of 2268 sentences).
Covered languages: Basque

Name of the resource: Ahosyn: Large Bilingual Speech Database for Synthesis
Type: speech, speech database
Content: 3799 (Basque) and 3995 (Spanish) phonetically balanced sentences recorded by one male and one female voice talents in neutral style. It was registered

at 48kHz, 16bits, semi-professional room, 2 microphones and laryngograph included.
Covered languages: Basque

Name of the resource: Speech-dat like database for Basque (Mobile).
Type: speech, speech database
Content: The Basque MDB-600 database contains the recordings of 660 speakers of Basque recorded over the mobile telephone network. This database is partitioned into 4 CDs. The database complies with the common specifications created in the SpeechDat project
Covered languages: Basque

Name of the resource: Galician SpeechDat FDB
Type: speech, speech database
Content: sentences, application words, digits, numbers, telephone numbers, money amounts, dates, hours, cities, names...
Covered languages: Galician

Name of the resource: Transgrigal DB
Type: multimodal, speech and video
Content: Transcrigal DB is a database of TV News shows in Galician. The news shows have been broadcasted by the Regional Public TV Station (TVG). Currently it consists of 31 shows. Each audio-video file is accompanied by a transcription text file.
Covered languages: Galician

Name of the resource: DOGalicia: Parallel Galician-Spanish Corpus
Type:text, parallel corpus
Content: A parallel Galician-Spanish corpus designed for statistical translation purposes. The original text comes from the "Diario Oficial de Galicia (DOG)" (Official Regional Gazette of Galicia) http://www.xunta.es/diario-oficial-galicia/
Covered languages: Galician-Spanish

Name of the resource: GCG: GrupoCorreoGalego
Type: text, annotated corpus
Content: A morphologically and syntactically annotated corpus for Galician language
Covered languages: Galician

4.6.2 Software and LR tools

Endogenous resources

Name of the resource: Segre
Type: LR Tool,
Content: A phonetic transcription software for Catalan.
Covered languages: Catalan

Restricted exogenous resources

Name of the resource: Cotovia Transcriber
Type: LR Tool,
Content: A phonetic transcription software for Galician.
Covered languages: Galician

4.7. Partner: UPF

4.7.1 Data resources

Endogenous resources
Name of the resource: TRL V-Subcat Lexicon
Type: lexica, lexicon
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Spanish

Name of the resource: News paper headlines corpus
Type: text, corpus
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Spanish

Name of the resource: Spanish - Translated Penn Tree Bank chapter 23
Type: text, corpus
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Spanish

Unrestricted exogenous resources
Name of the resource: WikiCorpus
Type: text, corpus
Upgrade: Metadata annotation, conversion to UTF
Covered languages: Spanish

4.7.2 Software and LR tools

Endogenous resources
Name of the resource: Converters to LMF 2
Type: LT Tool
Upgrade: Metadata annotation, conversion to UTF