



Readers: Evaluation And DEvelopment of Reading Systems

Anselmo Peñas

CHIST-ERA meeting, 27 March 2013, Brussels

Overview

1. Starting date
2. Scientific background
3. Challenges, objectives and expected results
4. Partners and their value to the consortium
5. Workplan
6. Management
7. Outputs
8. Valorisation

Starting date

- **January 1st, 2013**
 - Kick-off meeting, Madrid, January 22-23, 2013



- Different starting dates of partners ...
 - U. Edinburgh, UK (EPSRC EP/K017845/1)
 - October 1st, 2012
 - Synapse Développement, France (ANR-12-CHRI-0004-03)
 - November 1st, 2012
 - UPV/EHU, Spain (TBC)
 - UNED, Spain (TBC)



Scientific Background

- Learning by Reading (DARPA 2005)
 - Pilot studies
 - Semantic representation of documents
 - Aligned to Knowledge Bases
 - Bottleneck: Knowledge Integration
- Machine Reading Program (DARPA 2009)
 - “Develop technology enabling formal processing over naturally occurring text”
 - Decouple Reading and Reasoning machines
 - Limitation of classic supervised Information Extraction
- Leverage redundancy: from local to global
 - Knowledge Bases cannot grow on their own
 - Develop semi-supervised and unsupervised approaches
 - Macro-reading
 - Open Information Extraction



Scientific Background

Attached to a KB

- Micro-reading take advantage of the KB
 - Known instances and their categories
 - Subset /superset relations
 - Typed relations
 - Constrains
- Macro-reading populates the KB
 - Semi-supervised relation extraction + Entity Linking
 - Leverage redundancy
 - Beyond the sum of micro-readings
 - Self-growing Relational Knowledge Bases
- At the same time is continuously increasing its ability to map
 - More seeds -> more patterns
 - More patterns -> more seeds

Scientific Background

- But imagine you are able to get all patterns...
- Would it be enough?
- A single message is never self-contained
- There is always implicit information

$$\frac{\text{New_Information}}{\text{Background_Knowledge_required}} \sim 0$$



Scientific Background

- Is this implicit information expressable by our KBs?
 - Default answer is: No
- Then we can't use our KB to make explicit the implicit information

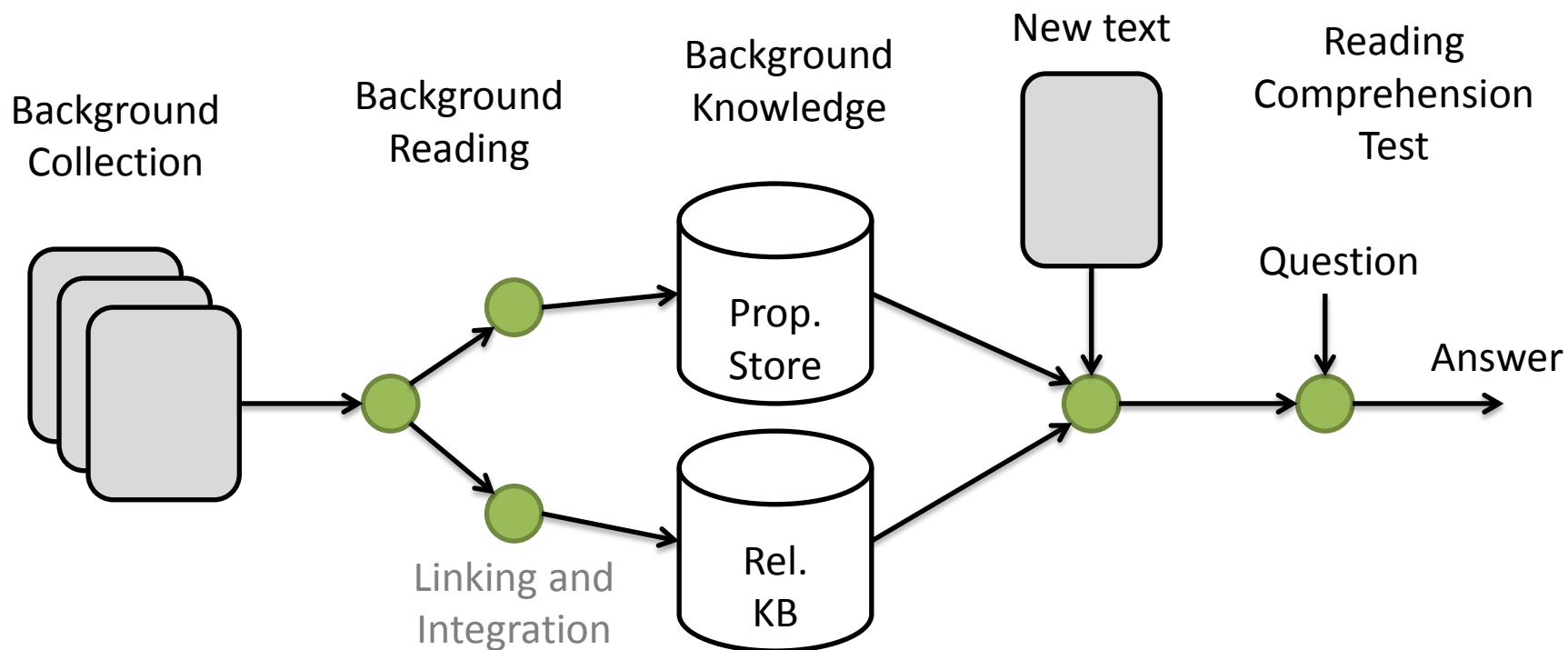
And what is missing...

... can't be used by a pattern to extract information

- We need to explore additional alternatives
 - Not attached to predefined relational KBs
 - Unsupervised

Scientific Background

● Reading Technology (READERS)





Challenges

- Background Reading
 - Induction of semantic roles ->
 - Build propositional background knowledge bases
- Knowledge Linking and Integration
 - Acquired background knowledge
 - Existing Knowledge Bases
- Read a single document
 - Bringing the background knowledge
 - Required to answer questions about the doc.
- Develop evaluation methodologies for Machine Reading

Objectives

1. Background Reading

- Find natural representations of background knowledge after processing large amounts of texts
- Explore unsupervised and distant-supervised algorithms
 - Induction of semantic roles
 - Discovery of class-instance relations
 - etc.
- Create propositional background knowledge bases

Objectives

2. Knowledge Linking and Integration

- Integrate the acquired background knowledge with existing knowledge repositories
- Develop distant supervised techniques for Entity Linking and Relation Extraction
- Extend existing knowledge bases with new taxonomical relations and instances

Objectives

3. Machine Reading system development and application

- Develop a Machine Reading system
 - English and French
 - Able to solve reading comprehension tests
- Apply the Machine Reading technology to assist the assessment of children reading abilities



Objectives

4. Evaluation of Machine Reading systems

- Develop evaluation methodology for Machine Reading systems
- Coordinate the generation of benchmarks
- Coordinate an international comparative/competitive evaluation task

Expected Results

- Advance the state of the art in Machine Reading and Natural Language Understanding
- New unsupervised and distant supervised methods for:
 - background knowledge acquisition
 - semantic role induction
 - automatic taxonomy creation
 - discovery of class-instance relations.
- New propositional knowledge bases linking them to existing knowledge repositories

Expected Results

- Algorithms for efficient storage and use of the knowledge resulting from processing large amounts of unstructured text
- A real industrial Machine Reading system
- Protocols for measuring the progress in the development of Machine Reading technology
- Organization of an international campaign for Machine Reading evaluation in several languages, with 100% reusable benchmarks by research groups and companies.

Consortium

Country	Institution/Department	Principal Investigator	Co-investigators
Spain	UNED (coordination)	Anselmo Peñas	Post-doc
United Kingdom	U. Edinburgh	Maria Mirella Lapata	Kristian Woodsend
France	Synapse Développement	Dominique Laurent	Patrick Seguela
Spain	UPV/EHU	Eneko Agirre	German Rigau



Consortium

UNED NLP & IR Group, Spain

- 10 senior PhD
- 12 researchers/students
- Research lines:
 - Multilingual Multimodal Information Access
 - Text Mining
 - Online Reputation Management
 - Information Extraction and Textual Inference
 - **Information Access Evaluation**

Consortium

Institute for Language, Cognition, and
Computation (ILCC), U. Edinburgh, UK

- 27 academic faculty
- 78 researchers/students
- Research lines:
 - Linguistic resources and tools
 - Information Extraction
 - Statistical Machine Translation
 - Large-scale robust NLP
 - **Probabilistic learning for NL Understanding**
 - **Extracting semantic information from large volumes of text**
 - **Unsupervised methods**



Consortium

Synapse Développement, France

- Company specialized in linguistic software development
- Founded in 1994
- B2C: more than 30,000 clients
- B2B
- Research lines:
 - Linguistic analyzers
 - Opinion detection
 - **Question Answering**

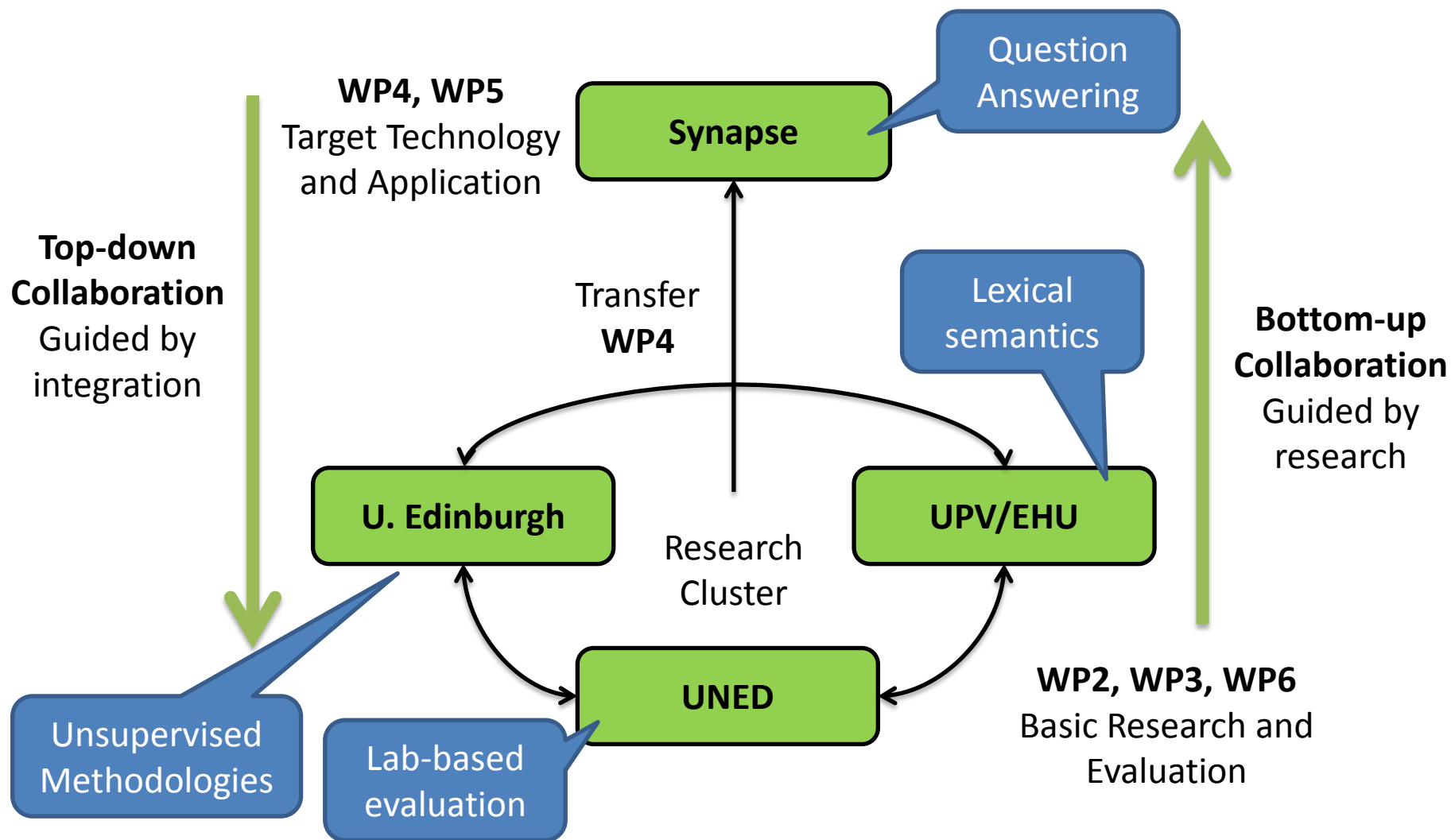
Consortium

IXA NLP group, Spain

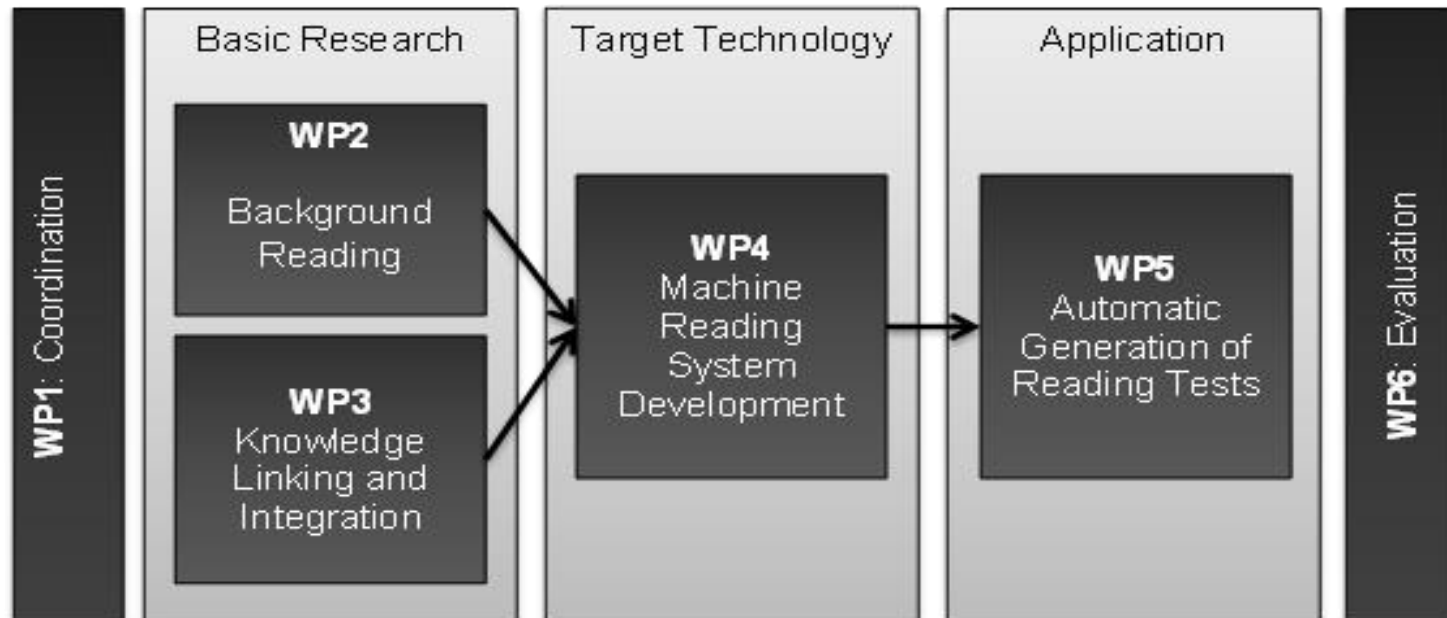
Interdisciplinary team

- 33 CS and 15 linguists
- 28 PhD
- 26 students/research assistants
- Research lines:
 - Building NLP resources and tools for Basque
 - Also English and Spanish
 - Applications: IR, IE, MT, education
 - Commercial: spell-checker, IR, dictionaries, etc.
 - ... **lexico-semantic resources**
 - ... **semantic processing**

Consortium



Workplan



person-month	WP1	WP2	WP3	WP4	WP5	WP6	total
1 (UNED)	8	12	1	2	1	12	36
2 (U. Edinburgh)	2	20	6	4	2	2	36
3 (Synapse Développement)	2	0.5	0.5	66	20	3	92
4 (UPV/EHU)	2	1	18	1	1	1	24
total	14	33.5	25.5	73	24	18	188



WP1 Coordination

	Task leader
WP1 Coordination	UNED
1.1 Work plan monitoring	UNED
1.2 Reporting and Dissemination	UNED
1.3 Collaboration with third parties and Community building	UNED

Deliverable	Month of delivery	Title of deliverable
D1.1	M12	1 st annual report
D1.2	M24	2 nd annual report
D1.3	M36	Final report

WP2 Background Reading

	Task leader
WP2 Background Reading	ED
2.1 Processing of document collections	UNED
2.2 Generation of Proposition Stores	UNED
2.3 Robust Semantic Role Labeling (SRL)	ED
2.4 SRL for multiple languages	ED
2.5 SRL without parsing	ED
2.7 Learning to read	UPV/EHU
2.8 Background Knowledge Base creation	ED

Deliverable	Month of delivery	Title of deliverable
D2.1	M12	Initial Proposition Store
D2.1	M24	Semantic Role Labelling component
D2.2	M36	Open IE and knowledge base construction

WP3 Knowledge Linking and Integration

	Task leader
WP3 Knowledge Linking and Integration	UPV/EHU
3.1 Preparation of Knowledge Bases	UPV/EHU
3.2 KB interface for disambiguation	UPV/EHU
3.3 KB interface for relation Extraction	UPV/EHU
3.4 KB interface for similarity, relatedness and expansion	UPV/EHU
3.5 Knowledge Base population	UPV/EHU
3.6 Taxonomy Induction	ED

Deliverable	Month of delivery	Title of deliverable
D3.1	6	Design of knowledge linking and integration
D3.2	30	Results of knowledge linking and integration

WP4 Machine Reading System Development

	Task leader
WP4 Machine Reading System Development	SD
4.1 Development of Multilingual platform	SD
4.2 Development of French Reading Machine	SD
4.3 Development of English Reading Machine	SD
4.4 Software release	ED

Deliverable	Month of delivery	Title of deliverable
D4.1	12	Storing format in Knowledge Base
D4.2	24	Multilingual platform with beta-version
D4.3	24	SRL Software release
D4.4	36	Multilingual platform with final version
D4.5	36	Taxonomy induction software release

WP5 Application to Automatic Generation of Reading Tests

	Task leader
WP5 Application to Automatic Generation of Reading Tests	SD
5.1 Reading tests elaboration and knowledge levels	SD
5.2 Automatic generation of French reading tests	SD
5.3 Automatic generation of English reading tests	SD

Deliverable	Month of delivery	Title of deliverable
D5.2	24	QCM generator in French
D5.3	32	QCM generator in English

WP6 Evaluation of Machine Reading

	Task leader
WP6 Evaluation of Machine Reading	UNED
6.1 Definition of the campaign exercises and goals	UNED
6.2 Evaluation methodology and resources	UNED
6.3 Activation of the campaign exercises	UNED
6.4 Assessment and comparison of systems results	UNED
6.5 Workshop	UNED

Deliverable	Month of delivery	Title of deliverable
D6.2	M24	2 nd Report on the lab-based evaluation
D6.3	M36	3 rd Report on the lab-based evaluation

Milestones

No of Milestone	Delivery month	WP involved	Title
M1	M12	WP2	Initial Proposition Store
M2	M24	WP2, WP4	Semantic Role Labelling Software release
M3	M36	WP3, WP4, WP6	Machine Reading system
M4	M36	WP4, WP5	Automatic Generator of Reading Comprehension tests

Management

- Kick-off meeting, Madrid, January 22-23, 2013
- Monthly phone conference of the Technical Coordination Committee (TCC)
 - Integrated by Work Package Leaders (WPLs)
- Exchange
 - Bernardo Cabaleiro (UNED)
 - Internship in Edinburgh: 10/9/2012 – 10/12/2012
 - Oier Lopez de Lacalle (UPV(EHU))
 - Postdoc at Edinburgh until 2014
- Coordination with the Japanese National Institute of Informatics (Prof. Noriko Kando) to co-organize the “Entrance Exams” challenge

Management

Website: nlp.uned.es/readers-project/



Readers: Evaluation And DEvelopment of Reading Systems

[Summary](#) - [Objectives](#) - [Partners](#)

Summary

The READERS project proposes new unsupervised computational models to automatically extract background knowledge after reading large amounts of unstructured text. This knowledge will be in the form of classes, categorized entities and predicates whose arguments are typified by probability distributions over classes. Classes themselves will be automatically organized into taxonomies related to the predicates in which they participate.

In this way, new methods and models based on extensional definitions of concepts will be developed and deployed for the automatic creation of knowledge bases. Importantly, these will be closely related to textual representations and instrumental in enabling textual inferences. The extracted knowledge will be also linked to external human-made resources such as Freebase, DBPedia and WordNet, and the knowledge bases will be interfaced with several engines for performing disambiguation, relation extraction, term expansion, and measuring relatedness.

A key part of the project will be the development of a reading matching that will use all these resources and tools. The purpose of our reading machine is to answer queries about a given text. Texts are never self-contained and their interpretation always requires recovering large amounts of background knowledge. Thus, the Machine Reading technology under development must incorporate not only language processing but also the recovery and use of large amounts of background knowledge.

This Machine Reading technology will be evaluated through Multiple-Choice Reading Comprehension tests (MRC) developed by humans over unseen documents. MRC tests enable objective and reproducible evaluation experiments, and will be 100% reusable as benchmarks available for the international community. Interestingly, the industrial partner in charge of the Machine Reading system development will apply the reverse technology to automatically generate MRC tests for the automatic assessment of children's reading abilities. This reading machine will work with at least two languages, English and French.

Outputs

Entrance Exams at CLEF

- Evaluate systems under the same conditions humans are evaluated to enter Japanese University
- In collaboration with NTCIR and Todai Robot project (Japanese National Institute of Informatics)
- Exams are created by the Japanese National Center for University Admissions Tests
- In QA4MRE 2013 Lab -> Entrance Exams for English
2013-2014: Reading Comprehension Tests
More info at: <http://nlp.uned.es/entrance-exams/>



Valorisation

- Industrial partner
 - Generation of tests to assess children reading
 - Potential B2B new markets in future
 - Summarization
 - Question Answering
 - Information synthesis
 - E.g. Answer questions to build auditory reports