This project is partially funded by the European Commission under the Seventh (FP7-2007-2013) Framework Programme for Research and Technological Development. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

D7.5: EVALUATION OF LANGUAGE TECHNOLOGY: ERROR ANALYSIS

Authors: Richard Evans, Iustin Dornescu, Mijail Kabadjov

Affiliation: UW

Date: 30 September 2013

Document Number: FIRST_D7.5_20130930

Status/Version: Final

Distribution Level: Public

Project Reference	287607
Project Acronym	FIRST
Project Full Title	A Flexible Interactive Reading Support Tool
Distribution Level	Public
Contractual Date of Delivery	30 October 2013
Actual Date of Delivery	30 October 2013
Document Number	FIRST_D7.5_20131030
Status & Version	Approved v.1
Number of Pages	32
WP Contributing to the Deliverable	WP7 – Testing and Evaluation
WP Task responsible	Vesna Jordanova
Authors	Richard Evans, Iustin Dornescu, Mijail Kabadjov
Other Contributors	Paloma Moreda, Elena Lloret, Eduard Barbu
Reviewer	Constantin Orasan
EC Project Officer	Magdalena Szwochertowska
Keywords:	Evaluation, LT, error analysis

Abstract:

Deliverable *D7.5* presents evaluation and error analysis of language technologies developed in the first 24 months of the project. The evaluation is made with reference to user requirements specified in this period (Deliverable D2.2). This deliverable includes background on the evaluation process and quantitative evaluation and error analysis of the LT developed in WP3, WP4, and WP5 (Deliverables D6.2 and D6.3 and internal deliverables D3.i, D3.ii, D4.ii, D4.iii, D5.ii, and D5.iii). The deliverable concludes with a summary of actions derived from the error analysis to steer development of the prototype in months 24-30.

Evaluation of Language Technology: Error Analysis

Topic **Page**

1. Executive Summary

This report describes the analysis of errors made by the software prototypes developed in WP3, WP5, and WP5 (described in internal deliverables D3.ii, D4.iii, and D5.iii). Section 2 of the report summarises the context of the evaluation experiments from the perspective of WP2 (User Requirements), WP3 (Processing Structural Complexity), WP4 (Processing Ambiguity in Meaning), and WP7 (Testing and Evaluation). This section also describes the general methodology adopted for error analysis. Sections 3-5 of the report describe the specific approach taken to error analysis in each of the work packages concerned with development of LT prototypes in the FIRST project. Finally, Section 6 concludes the report with recommendations for improvements of the prototypes to be undertaken in the next six months.

2. Background and methodology

Error analysis of the FIRST prototype builds on RTD activities undertaken in the derivation of user requirements that serve as a basis for evaluation (WP2), the development of LT implementing the functionality of the prototype (WP3-5), and the development of annotated corpora and tools to enable automatic evaluation of the LT. These aspects of background are briefly summarised in Sections 2.1-2.5.

2.1 Background: WP2 User Requirements

Reports on specific user requirements of the LT developed in the FIRST project were delivered in months 12 (D2.1) and 20 (D2.2). The second of these deliverables has superseded the first. The FIRST prototype is evaluated with respect to its ability to meet these user requirements. The specific user requirements relating to each of the three broad areas of language technology are presented in Sections 2.2.1, 2.2.2., and 2.2.3, which provide background information on each of the LT work packages.

In addition to user requirements related to language technology, the user surveys reported on in D2.2 revealed that end users of the prototypes developed in this project demand access to different types of document.

Children[1] seek access to documents in the *informative: arts/leisure* domain (e.g. games, movies, and music reviews). There is little demand for access to:

- Informative texts in the medium of periodicals (newspapers).
- Texts belonging to the genre of press, though exceptions are made for informative texts related to their own particular interests.
- Imaginative texts (fiction).

By contrast, adults demand access to:

- informative texts on different types of science,
- Informative text on world affairs/news in the medium of periodicals (newspapers),
- Informative text on a wide range of topics from generic to specific,
- Informative text on ASD,
- Imaginative text (including books and “written-to-be-spoken” content such as scripts and screenplays,
- Social networking and communication.

These findings helped to steer selection of the test data used to evaluate the LT components.

2.2 Background: Language Technology

Error analysis of the FIRST prototype integrates LT components developed in three work packages.

[1] For BG, end users are children, for EN, end users are adults, and for ES, end users are children and adults.

2.2.1 WP3 Processing Structural Complexity

The implemented structural complexity processor currently converts syntactically complex sentences (known grammatically as compound sentences and complex sentences) into a more accessible form. It makes a contribution to the user requirements printed in bold font in Table 1.

WP	Obstacle	Assistive element	URCode
WP3	**Multiple copulative coordinated clauses**	**Substitute with sentences divided by periods.**	**UR301**
	Subordinate adjective clause	**Substitute by adjective or remove and divide sentences by period.**	**UR302**
	Explicative clauses	**Remove explicative clause**	**UR303**
	Final / Illative	**Substitute rare conjunctions by more common ones.**	**UR304**
	Comparative	**Substitute rare conjunctions by more common ones.**	**UR305**
	Coordinate Adversative conjunction. Except when followed by the left subordination boundary "when"	**Substitute rare conjunctions by more common ones.**	**UR306**
	Adverbial clause after main clause	Place adverbial clauses before main clause.	UR307
	Conditional clause after main clause.	Place conditional clauses before main clause.	UR308
	Long sentences	**Sentences < 15 words**	**UR309**
	Semicolon and suspension points	**Avoid the use of semicolon and suspension points**	**UR310**
	Brackets and non common punctuation marks (&,%, /…)	Avoid non common punctuation marks	UR311
	Improper grammar	Correct grammar	UR312
	Sentences in passive voice	Use active voice	UR313
	Sentences with double negative	Avoid negative and double negative sentences	UR314
	Paragraphs cut at the end of page	**Avoid cutting paragraph at the end of page**	**UR315**
	Words cut at the end of line	**Avoid words cut at the end of line.**	**UR316**
	Long words	**Avoid adverbs ending in "-ly" (English, "-mente", -o Bulgarian.**	**UR317**
		Avoid words with more than 7 characters	**UR318**
		Use shorter synonym for words with more than 7 characters and adverbs ending in "-ly" (Spanish: "mente", Bulgarian: "o")	**UR319**

Table 1: User requirements related to WP3 Processing Structural Complexity.

The syntactic simplification module is a rule-based system, described in more detail in D3.ii. The rules applied by the system were evaluated in terms of precision and recall. The Structural Complexity Processor generates XML with SENTENCE elements that contain information about the original form of the sentence (ORIGINAL element) and an optional SIMPLIFIED element that contains information about the converted form of the text of the ORIGINAL element. SIMPLIFIED elements contain one or more S elements, each of which contains one maximally simplified sentence derived from ORIGINAL. An example SENTENCE element is displayed in Fig. 1.

```
<SENTENCE>
     <ORIGINAL>The libel jury was told this week that the words used
     had been ruled to be a libel and they had only to agree damages.
     </ORIGINAL>
     <SIMPLIFIED>
          <S DERIVATION="CCV-12a">The libel jury was told this week
          that the words used had been ruled to be a libel.</S>
          <S DERIVATION="CCV-12b">The libel jury was told this week
          that they had only to agree damages.</S>
     </SIMPLIFIED>
</SENTENCE>
```

Figure 1: Example SENTENCE element.

Using the same structure, linguistic experts created a gold standard for the purposes of evaluation. In the gold standard, elements within SIMPLIFIED were specified manually.

User requirement UR301 is met by rules applied in an iterative process that, on each iteration, convert a string containing a conjoint clause into two strings in which the conjoint clause is substituted by each of its conjoins (the conjoins may themselves be conjoint clauses). The system was run in two modes: one in which the classification of signs of syntactic complexity was obtained by consulting an oracle and one in which an automatic tagging approach to sign classification (see D3.ii) was applied. User requirements UR302 and UR 303 are met by rules applied in an iterative process that, on each iteration, convert a complex sentence into two strings, one of which is produced by deleting one subordinate clause from the complex sentence, the other produced by linking the subordinate clause with the noun phrase that it modifies. Semicolons are triggers for many of the rules developed to meet UR301, UR302, and UR303. Their application means that the syntactic processor goes some way toward meeting UR310 (semicolons). User requirements UR304-306 have been addressed by the development of gazetteers of these conjunctions and conjuncts and simple transliteration rules. In this project, methods to address user requirements UR315 and UR316 are implemented in the user interface, falling within the remit of WP6. User requirements UR317-319 are addressed via web services developed in WP4 exploiting resources (gazetteers) developed in WP3. The degree to which the structural complexity processor is able to meet these user requirements is not evaluated in this report due to an absence of test data.

2.2.2 Background: WP4 Processing Ambiguity in Meaning

The implemented meaning disambiguator currently converts semantically ambiguous sentences into a more accessible form. It makes a contribution to the user requirements printed in bold font in Table 2. Specifically, the components developed for this purpose include a coreference resolution system, a word sense disambiguation system, and gazetteers of difficult words such as mental verbs, infrequent words, acronyms/abbreviations, specialised terminology, and slang expressions.

WP4	**Polysemy**	**Provide easier synonyms**	**UR401**
		Detect and highlight polysemy	**UR425**
	Phraseological units (idioms, lexicalized metaphors)	**Replace by a simple word.**	**UR402**
		Detect and highlight when replacement is not possible	**UR425**
		Provide simple definitions to explain phraseological units	**UR410**
	Less common words	**Replace infrequent words with simpler synonym**	**UR405**
		Replace places of origin for less frequent adjectives of nationality	**UR406**
		Provide simple definitions to explain mental verbs	**UR411**
		Provide simple definitions to explain infrequent words	**UR413**
	Emotional language	**Replace complicated emotional adjectives with simpler synonym**	**UR404**
		Provide simple definitions to explain emotional adjectives	**UR412**
		Replace complicated mental verbs with simpler synonym	**UR403**
	Slang	**Normalize infrequent slang**	**UR407**
		Provide simple definitions to explain infrequent slang	**UR414**
		Detect specialized slang belonging to a domain	**UR423**
		Provide simple definitions to explain specialized slang	**UR424**
	Infrequent acronyms and abbreviations	**Expand infrequent acronyms and abbreviations**	**UR415**
	Temporal expressions	Resolve temporal expressions	UR416
	Long numerical expressions	Express long numerical expressions with words	UR417
	Anaphors	Detect and leave unresolved anaphors with low resolution confidence level.	UR418
		Resolve pronominal anaphora	**UR419**
		Resolve definite descriptions	**UR420**
		Resolve ellipsis	**UR421**
	Non lexicalized metaphors	**Provide idea of inferred meaning when possible and highlight**	**UR422**

Table 2: User requirements related to WP4 Processing Ambiguity in Meaning

2.2.3 Background: WP5 Generation of Personalised Documents

The user requirements corresponding to WP5 were reported in D2.2. LT has been developed in WP5 to address the requirements printed in bold in Table 3.

WP5	Understanding of general meaning	Give relevant idea on top of text	UR501
		Show key words.	**UR502**
		Post questions in or after the text to help monitor comprehension	UR503
		Give information on key concepts before reading text	UR504
		Support the overall meaning of the text with images.	**UR505**
		Provide text with summaries	**UR506**
	Phraseological unit	Support the understanding of phraseological units with images	UR507
	Non lexicalized metaphors	Support the understanding of metaphorical language with images	UR508
	Less common words	**Support the understanding of less common words with images**	**UR509**
	Emotional language	Support the understanding emotional adjectives with images	UR510
	Polysemy	**Support the understanding of polysemy with images**	**UR511**
	Long paragraphs	**Divide long paragraphs**	**UR513**

Table 3. User requirementscorresponding to WP 5 Generation of Personalised Documents.

Given that most of the user requirements in WP5 are concerned with image retrieval, two modules were developed in order to meet them:

1. An Online Image Retrieval Module that uses Google, Bing and ImageNet data-sources to retrieve images for the user selected concepts.
2. An Offline Image retrieval module that disambiguates previously identified terms against Wikipedia and retrieves the associated images from the corresponding Wikipedia pages.

The Offline Image Retrieval Module is agnostic about the terms being disambiguated: they may be tagged in any annotation set.

2.3 Background: WP7 Testing and Evaluation

RTD activity in the first 12 months of the project led to development and acquisition of annotated resources that have since been exploited for the purpose of evaluation and error analysis of the structural complexity processor, the meaning disambiguator, and the personalised document generator.

With regard to processing structural complexity (WP3), in the first twelve months of the project, collections of three different categories of documents (news articles, patient healthcare information, and literature) were collected. A subset of this was annotated with information about the syntactic bounding and linking functions of various signs of syntactic complexity. To summarise, coordinators (conjunctions and punctuation marks) were annotated with information about their conjoins (the types of constituents

that they link together). Subordination boundaries (complementisers, wh-words, and punctuation marks) were annotated with information about the constituents that they bound. This data has been used to evaluate the performance of the automatic classifier (see D3.i) used to predict the class labels of signs of syntactic complexity. Subsequently, a second subset of the document collection was annotated to encode information about more accessible forms of complex sentences in these documents. This data constitutes a gold standard against which the output of the structural complexity processor can be compared. It is also used to evaluate the rule sets applied by the structural complexity processor.

Evaluation of services processing ambiguity in meaning (WP4) was supported by the availability of two annotated resources. Coreference resolution in English and Spanish was evaluated by comparison of system output with human-annotated test data used in the SemEval-2010 *Coreference Resolution in Multiple Languages* task (Recasens et al., 2010). Evaluation of the automatic word sense disambiguation method employed by the meaning disambiguator was derived by use of a corpus developed for use in the SEMEVAL 2013, Multilingual Word Sense Disambiguation task.

Data collected for subsequent use in reading comprehension testing (D7.4) was annotated for the purpose of evaluating the LT developed in the project. This data set consists of six texts written in English, six in Bulgarian, and twelve in Spanish. Reference to these texts was used to support manual evaluation of LT services developed to process ambiguity in meaning (WP4) and to generate personalised documents (WP5)

2.4 A methodology for evaluation and error analysis of LT components

The methodology employed is based on examination of the degree to which LT presented in D3.ii. D4.iii, and D5.iii meets the user requirements specified in D2.2. The LT prototypes are evaluated with regard to the particular sets of functions that they implement in order to address user requirements. The evaluation is based on test sets available in the LT background and also developed in WP7 (Testing and Evaluation). In most cases, the empirical approach to automatic evaluation, typical of the field of natural language processing (NLP), has been adopted. Here, the annotated corpora described in D7.3 are used in combination with standard evaluation metrics forming part of the LT background. In general, the LT components are evaluated with respect to their ability to detect and remove obstacles to reading comprehension in input documents. Unexpected findings obtained via the empirical approach have been followed up with a qualitative assessment of individual LT functions. Sections 3.1-3.3 describe the specific approach taken to evaluation and error analysis of the different LT prototypes and functions.

The prototypes developed in the FIRST project are intended to meet the user requirements presented in D2.1 and D2.2 (summarised in Section 2.3). The methodology described in this section uses a range of resources to quantify the degree to which the LT components meet these requirements:

a) Linguistic analysis of the output of the structural complexity processor (D3.ii), the meaning disambiguator (D4.iii), and the personalised document generator (D5.iii).
b) Quantitative assessment of the readability of personalised documents output by prototypes D6.3

3. Evaluation and Error Analysis: Structural Complexity Processor (v2)

3.1 Method

As noted in D3.ii, processing structural complexity in this project consists of automatic syntactic simplification. The method employed is a two step process:

a) Detection and classification of the function of certain reliable signs of syntactic complexity.
b) Application of an iterative rule-based sentence rewriting algorithm.

Signs of syntactic complexity occur in Bulgarian, English, and Spanish. Signs of specific functional types co-occur with particular types of syntactic complexity that were highlighted in D2.2 as posing obstacles to reading comprehension for readers with ASD. Specifically, signs of type CEV and SSEV are obstacles to reading comprehension to be detected by the automatic classification process and removed by the sentence rewriting rules developed in WP3 (presented in D3.ii). They are signs of conjoint clauses and subordinate clauses.

Process (a) can be evaluated via standard methods in LT based on comparing classifications made by the system with classifications made by linguistic experts. As noted in Section 3.1., processing of structural complexity depends upon accurate identification of the bounding and linking functions of signs of syntactic complexity.

Evaluation of process (b) is similar in complexity to evaluation of machine translation systems. Potential methodologies for evaluation include comparison of system output with human simplification of a given text, analysis of the post-editing effort required to convert an automatically simplified text into a suitable form for end users, comparisons using experimental methods such as eye tracking, extrinsic evaluation via NLP applications such as information extraction, all of which have weaknesses in terms of adequacy and expense.

Due to the disadvantages of these previously established methods, an alternative evaluation method is proposed for use in the FIRST project. Rather than evaluating the final output of the rule-based syntactic processor, the accuracy of the two rule sets employed by the syntactic processor is reported. Of these rule sets, the first was developed for use with an iterative algorithm to convert compound sentences (containing conjoint clauses) into a more accessible form. The second was developed for use with the same algorithm for the purpose of converting complex sentences (containing subordinate clauses) into a more accessible form.

Evaluation of the rule set developed for conversion of compound sentences focuses on rules that generate duplicate sentences in which the conjoint clause in the original is replaced by a different conjoin of that coordinate constituent. The number of duplicate sentences generated is equal to the number of conjoins detected by the rules in the conjoint clauses. Evaluation of the rule set developed for conversion of complex sentences focuses on rules used to identify subordinate clauses modifying nouns and to generate new sentences that link modified nouns with their clausal modifiers. The evaluation method is based on comparing sets of simplified sentences derived from an original sentence by linguistic experts and by the structural complexity processor.[2] It reports the error rates and total numbers of errors made by individual rules that comprise the rule sets. The evaluation experiments assess the use of the developed rule sets in

[2] Structural complexity processor v2, presented in D3.ii.

two modes: one in which the system consults an oracle for classification of signs of syntactic complexity and one in which the system consults the output of the tagging approach to sign classification described in D3.i. The gold standards developed for use in this evaluation are a subset of the sentences annotated with information about signs of syntactic complexity.

The accuracy of rules developed to convert compound sentences and complex sentences into a more accessible form was obtained. In this context, SENTENCE elements generated by the Structural Complexity Processor are compared with those annotated by linguistic experts in the gold standard (key). These XML elements, which encode information about the automatic syntactic processing of sentences, are complex. They contain two daughter nodes. ORIGINAL is an obligatory leaf node containing textual data. The textual data is the original form of the sentence that has been processed syntactically. The second node, SIMPLIFIED, is optional, and has one or more daughter nodes, S, which contain textual data. The textual data contained in each S element comprises one of the more accessible sentences derived from ORIGINAL by the rules applied by the syntactic processor.

Care was taken by the linguistic experts creating the gold standard to ensure that all S elements within SIMPLIFIED are grammatical. The validity of simplified sentences in the response is determined by comparison with this grammatically correct gold standard. The accuracy scores obtained by the syntactic processor thus reflect the degree to which it is able to meet user requirement UR312. The process of highlighting or substituting rare conjunctions listed in gazetteers developed to meet UR304-306 was not evaluated due to the absence of sufficient test data.

3.2 Results

With regard to process (a), automatically labelling the syntactic function of signs of syntactic complexity, Table 4 displays F1 scores for the automatic prediction of class labels for the most prevalent signs in the collection of documents.

Class label	*F1*	*Cumulative frequency*
SSEV	0.9412	25.1%
CMV1	0.8325	33.5%
CMN1	**0.6942**	**41.7%**
SSMN	0.8294	49.0%
CEV	0.7991	55.9%
SSCM	0.9673	60.4%
ESEV	***0.5014***	**64.9%**
SSMA	0.9264	69.1%
ESMP	***0.5096***	**72.7%**
SSMP	0.8304	76.2%
CLN	0.7269	79.8%
ESMN	***0.4917***	**83.0%**
SSMV	0.8092	85.9%
ESCM	0.9072	88.5%

Table 4: F1 scores for prediction of class labels for most prevalent classes of syntactic complexity. Information about class labels used in trigger patterns is printed in bold

In Table 4, rows printed are classes of sign that are labelled with F1 < 0.7. Some consequences of the inaccurate classification of signs will be discussed later in this section.

Table 5 presents the accuracy of the rules implemented to convert conjoint clauses to a more accessible form. The columns News, Health, and Literature provide information relevant to the evaluation of the rules when applied to texts from different categories of text. The row *#Compound sentences* displays the number of ORIGINAL sentences in the test data that contain signs of conjoint clauses (signs of class CCV). The row *Accuracy* displays the ratio of sentences containing conjoint clauses that are correctly converted into a more accessible form by the rules implemented by the syntactic processor. Computation of accuracy is based on the mean Levenshtein similarity[3] between S elements within SIMPLIFIED elements automatically generated by the system and the most similar S elements in SIMPLIFIED elements manually generated by linguistic experts. This expression is denoted LS. Once the most similar S element in the key has been found for an S element in the response, that element is no longer considered when obtaining most similar S elements for sister Ss in the response. In this evaluation, sentences are considered to be converted correctly if their LS > 0.95. The row *ΔFlesch* displays the change in readability, assessed using the Flesch Reading Ease metric, after conversion of input sentences into a more accessible form by the structural complexity processor. The figures in parenthesis are the reading ease score of the converted sentences. The row *Long ORIGINAL* displays the proportion of SENTENCEs in the test data whose ORIGINAL elements are more than 15 words long. The row *Long ORIGINAL for which SIMPLIFIED have been generated* displays the proportion of SENTENCEs in *Long Original* to which automatic conversion rules were applied. *Long S elements* displays the proportion of S elements generated by the structural complexity processor that are more than 15 words long.

As noted in D3.ii, the structural complexity processor exploits information about the functions of signs of syntactic complexity. For each text category (columns *News*, *Health*, and *Literature*), results are provided for two version of the processor. In one, the functions of these signs are obtained by consulting an oracle (*Oracle*). In the other, the tagging approach to sign classification presented in D3.i and D3.ii is used to obtain the classes of signs of syntactic complexity (*Classifier*). The Flesch Reading Ease scores of the original documents are 68.1 (plain English) for News, 63.0 (plain English) for Health, and 82.4 for Literature.

Text category	News		Health		Literature	
	Oracle	Classifier	Oracle	Classifier	Oracle	Classifier
#Compound sentences	698		325		418	
Accuracy	0.7579	0.3137	0.6123	0.4431	0.2464	0.1148
ΔFlesch	+11.1	+9.9	+8.2	+10.2	+15.3	+13.6
Long ORIGINAL	0.6413		0.2269		0.4903	
Long ORIGINAL for which SIMPLIFIED have been generated	0.8539	0.8792	0.6771	0.9051	0.8317	0.8586
Long S elements	0.2958	0.3496	0.0909	0.0663	0.2934	0.2624

Table 5: Evaluation of rules to improve accessibility of compound sentences. Here, a sentence is described as *long* if it contains more than 15 words (UR309)

[3] Defined as 1 minus the ratio of Levenshtein distance between the two sentences to the length in characters of the longest of the two sentences being compared.

Text category	News		Health		Literature	
	Oracle	Classifier	Oracle	Classifier	Oracle	Classifier
#Complex sentences	369		335		379	
R	0.4526	0.4336	0.2925	0.2269	0.4749	0.2586
ΔFlesch	+2.5	+2.3	+0.8	+0.9	+2.3	+2.3
Long ORIGINAL	0.6413		0.2271		0.4903	
Long ORIGINAL for which SIMPLIFIED have been generated	0.9100	0.8987	0.6325	0.6824	0.8394	0.8329
Long S	0.7828	0.7790	0.5766	0.6201	0.7632	0.7625

Table 6: Evaluation of rules to improve accessibility of complex sentences. Here, a sentence is described as *long* if it contains more than 15 words (UR309)

Table 6 presents the same information about rules implemented to convert sentences containing subordinate clauses to a more accessible form. In this table, *#Complex sentences* denotes the number of sentences in the test data that contain subordinate clauses.

The patterns used to simplify sentences containing signs of class CEV exploit information about other signs within the same sentence. The patterns are sensitive to accurate detection of the broad class subordination boundaries (left and right boundaries).

The patterns used to simplify sentences containing signs of class SSEV exploit information about other signs within the same sentence. The patterns are sensitive to accurate detection of the broad classes of left and right subordination boundaries, and specific classes of sign that are the right boundaries of subordinate clauses (ESEV), noun phrases (ESMN), adjective phrases (ESMA), and coordinators of noun phrases (CMN1). Accurate conversion of sentences containing signs of class SSEV thus depends on a sign classifier that is accurate over a relatively large number of classes. Statistics presented in Table gg (Section 3.1) show that F1 scores for the automatic prediction of class labels for several of these classes is below 0.7. Comparison of the two columns *Oracle* and *Classifier* in Table 6 reveals that, despite the relatively high F-score obtained for automatic prediction of other class labels, the impact of errors made by the sign classifier is large.

The error rate of rules used by the syntactic processor to convert sentences containing conjoint clauses into an more accessible form was also recorded. The most error-prone trigger patterns are listed in Table 7, together with information on the conjoin that they are intended to detect (left or right), their error rate, and the number of number of errors made. In the patterns, words with particular parts of speech are denoted by the symbol *w* with the relevant Penn Treebank tag appended as a subscript. Verbs with clause complements[4] are denoted v_{CCV}. Signs of syntactic complexity are denoted by the symbol *s* with the abbreviation of the functional class appended as a subscript. Specific words are printed in italics. In the patterns, thee clause coordinator is denoted '_' and upper case letters are used to denote stretches of contiguous text. Finally, continuous sequences of text occurring in patterns are denoted using '...' It should be borne in mind that sentences containing *n* conjoint clauses will trigger the application of *n-1* rules, only one of which may be erroneous in its application. The statistics presented in Table 7 and Table 8 are based on the number of times that each rule was seen to apply in the derivation of an incorrect sentence. In this experiment, signs of syntactic complexity were classified using an oracle, in order to isolate the influence of the rules in the system output. The statistics are derived for rules applied to texts of all three

[4] Listed in Deliverable D3.i (Figure 2)

categories/genres. In this context, the accuracy which with the syntactic processor converts sentences containing conjoint clauses into a more accessible form is 0.5767. The accuracy of this task with regard to subordinate clauses is 0.4109.

ID	Conjoin	Trigger Pattern	Error rate	#Errors
CCV-24b	B	A _ B	0.1314	59
CCV-24a	A	A _ B	0.1189	54
CCV-12b	A that C	A *that* B _ C	0.5952	25
CCV-25a	NA	NA	0.9565	22
CCV-26a	A v_{CCV} B : " C	A v_{CCV} B : " C _ D	0.2133	16
CCV-26b	A v_{CCV} B : " D	A v_{CCV} B : " C _ D	0.2029	14
CCV-27b	A v_{CCV} B " D	A v_{CCV} B " C _ D	0.5333	8
CCV-9b	A $w_{\{JJ\|JJS\|VBN\}}$ C	A $w_{\{JJ\|JJS\|VBN\}}$ B _ C	0.5555	5
CCV-12a	A that B	A *that* B _ C	0.1136	5
CCV-15b	A *said* B	A *said* $w_{\{NNS\|NNP\|NN\|PRP\|DT\}}$ _ B	0.2857	4

Table 7: Error rates for rules converting sentences containing conjoint clauses into a more accessible form

Rule CCV-25a is applied when the input sentence triggers none of the other implemented patterns. Errors of this type quantify the number of sentences containing conjoint clauses that cannot be converted into a more accessible form by rules included in the structural complexity processor. These errors can only be addressed via the addition of new rules or the adjustment of already implemented rules.

ID	Matrix clause / subordinate clause	Trigger Pattern	Error rate	#Errors
SSCCV-78a	NA	NA	0.5172	45
SSCCV-72a	A , _ C $w_{\{VBG\|VBN\|VBP\|VBD\|VB\|VBZ\}}$ D	A *s* B _ C $w_{\{VBG\|VBN\|VBP\|VBD\|VB\|VBZ\}}$ D	0.3333	4
SSCCV-36a	NA	A *told* $w_{\{NN\|NNP\|NNS\|PRP\|DT\|IN\}}$* _ B	0.1667	4
SSCCV-13b	w_{VBN} w_{IN} {$w_{\{DT\|PRP\$\|NNP\|NNPS\|NNS\|NN\|CD\}}$|-|,}* $w_{\{NN\|NNP\|NNS\}}$ B	A w_{VBN} w_{IN} {$w_{\{DT\|PRP\$\|NNP\|NNPS\|NNS\|NN\|CD\}}$|-|,}* $w_{\{NN\|NNP\|NNS\}}$ _ B	1	3
SSCCV-61a	A w_{IN} w_{DT}* $w_{\{NN\|NNP\|NNS\}}$ {$w_{\{NN\|NNP\|NNPS\}}$|*of*}* {.|?|!}	A w_{IN} w_{DT}* $w_{\{NN\|NNP\|NNS\}}$ {$w_{\{NN\|NNP\|NNPS\}}$|*of*}* _ w_{VBD} B {.|?|!}"	0.5	2
SSCCV-62a	A w_{IN} w_{DT}* $w_{\{NN\|NNP\|NNS\}}$ {$w_{\{NN\|NNP\|NNPS\}}$|*of*}*	A w_{IN} w_{DT}* $w_{\{NN\|NNP\|NNS\}}$ {$w_{\{NN\|NNP\|NNPS\}}$|*of*}* _ w_{VBD} $w_{\{RB\|VBN\}}$* B	0.6667	2
SSCCV-20b	$w_{\{DT\|NN\|NNP\|NNS\|JJS\|JJ\|VBN\}}$* $w_{\{NN\|NNP\|NNPS\|NNS\}}$ B	A $w_{\{IN\}}$ $w_{\{DT\|NN\|NNP\|NNS\|JJS\|JJ\|VBN\}}$* $w_{\{NN\|NNP\|NNPS\|NNS\}}$ _ B	1	2
SSCCV-61b	w_{DT}* $w_{\{NN\|NNP\|NNS\}}$ {$w_{\{NN\|NNP\|NNPS\}}$|*of*}* w_{VBD} B	A w_{IN} w_{DT}* $w_{\{NN\|NNP\|NNS\}}$ {$w_{\{NN\|NNP\|NNPS\}}$|*of*}* _ w_{VBD} B {.|?|!}"	0.5	2
SSCCV-28a	NA	w_{VBD} w_{RP}* *that*	0.0555	2
SSCCV-33b	*that* $w_{\{NN\|NNP\|NNS\}}$ w_{VBD} w_{DT} $w_{\{NN\|NNS\|NNP\}}$*	{*an*|*a*} $w_{\{NN\|NNP\|NNS\}}$ _ w_{VBD} w_{DT} $w_{\{NN\|NNS\|NNP\}}$* {.|*!*|?}	1	1

Table 8: Error rates for rules converting sentences containing subordinate clauses into a more accessible form

While numerous patterns have an error rate of 1, the vast majority of rules are triggered by two patterns. At present, we consider that a rule should only be deactivated if both patterns have error rates greater than 0.5. Experimental deactivation of rules triggered by patterns SSCCV-13b, SSCCV-62a, SSCCV-20b, and SSCCV-33b did not lead to improved accuracy of the conversion process.

The development of gold standard data sets and the automatic evaluation method described in this section has enabled filtration of error-prone rules from the prototype.

SSCCV-36a is a pattern used to prevent processing of sentences that contain verbs with clause complements. Using the sentence rewriting algorithm to process sentences containing these subordinate clauses generates ungrammatical output. SSCCV-28a is a pattern used to prevent processing of relative clauses bounded on the left by the complementiser *that*. These subordinate clauses are usually clause complements of the first verb in the trigger pattern.

Overall, it was noted that conversion of complex sentences into a more accessible form is more error prone (accuracy=41.09%) than conversion of compound sentences (accuracy=57.67%).

4. Evaluation and Error Analysis: Meaning Disambiguator (v2)

The different modules for addressing the semantic obstacles to reading comprehension are described in D4.iii. With respect to the evaluation of these obstacles, two types can be distinguished:

- **Quantitative evaluation** for which gold-standard corpora were available (coreference resolution for Spanish and word sense disambiguation).
- **Qualitative evaluation** of modules (such as those detecting and removing polysemous, rare, and specialised words) for which gold-standard corpora are not available and quantitative evaluation is not possible. An analysis of the obstacles detected and removed by the tools developed in WP4 was also carried out with reference to the texts selected for use in the reading comprehension tests delivered to end users.

The following sections present the results obtained for each of these types of evaluation.

Quantitative evaluation

Gold-standard corpora were available to support quantitative evaluation of Spanish coreference resolution (ES), including anaphor resolution (ES), and word sense disambiguation (EN, ES).

- **Coreference resolution**

The Spanish section of the SemEval-2010 corpus [5] for *Task 1: Coreference Resolution in Multiple Languages* was employed for evaluating Spanish coreference. The reason for using this corpus was: i) it provides annotated data with coreference information (pronominal anaphora, definite description and ellipsis); and ii) it is also available for other languages, and therefore, this will allow us to test the multilinguality of our approach, or the feasibility to adapt our ML to other languages.

Concerning **anaphor detection,** for pronominal anaphora and definite descriptions, we relied on the output provided by Freeling (Padró, 2011) with respect to different types of language analysis, either POS tagger or shallow parsing, respectively. Based on these outputs, we defined different rules in order to allow us to identify pronouns in the first case, and noun phrases in the second one. For ellipsis detection, we carried out machine learning experiments. These results were then compared with the annotations available in the SemEval-2010 training dataset.

Type of coreference	Best algorithm for Coreference Detection	Precision
Pronominal anaphora detection	Information provided by Freeling (POS tagger)	96.85%
Definite descriptions detection	Information provided by Freeling (Shallow parsing)	74.50%
Ellipsis – Detection	Naïve Bayes	98.18%
	AVERAGE	**89.84%**

Table 9. Results on coreference detection

Regarding **anaphor resolution**, we experimented with a wide range of machine learning algorithms. The one that obtained the best precision results for pronominal anaphora and ellipsis resolution was the *Voting*

[5]The corpus is available at: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2011T01

Feature Intervals (VFI) algorithm. For definite descriptions, *Part* was the best algorithm. The three types of coreference processed in the project were automatically resolved with average precision > 82% (Table 10).

Type of coreference	Best algorithm for Coreference Resolution	Precision
Pronominal anaphora resolution	VFI	73.59%
Definite descriptions resolution	PART	99.81%
Ellipsis – Resolution	VFI	73.53%
	AVERAGE	**82.31%**

Table 10. Results for coreference resolution (best ML algorithm)

Comparing our results for detection and resolution with respect to the state of the art, we find that the best performing system in Semeval-2010 task (Recasens et al., 2010) was TANL-1 (Attardi et al., 2010), which achieved 84.1% precision for detection and 79% for resolution of coreference in the test dataset. The approach adopted in the FIRST project achieves superior performance results (+7% precision for detection and +4% for resolution) than the best performing systems in Semeval-2010. More information about the Spanish coreference resolution process can be found in D4.ii and D4.iii.

Coreference resolution in English exploits the Stanford Coreference system. It is reported to perform with an F1 score of around 60%. It was the top-rated system at the CoNLL-2011 shared task on coreference resolution.

Two coreference resolution systems for Bulgarian were also evaluated. The approach distributed with the LINGUA text processing system for Bulgarian (Tanev and Mitkov, 2002) was tested by its authors on a corpus of texts from two domains: software manuals (221 pronouns) and tourist guides (116 pronouns). Their system achieves a precision of 72.6% (a more detailed account of performance is shown in Table 11).

Type of scoring measure	Standard	Optimised	Baseline
Success rate	72.6	75.7	60.4
Critical succ. Rate	67.7	70	55.7
Non trivial succ. Rate	72.3	75.4	60.4

Table 11.Success rate of pronoun resolution

LINGUA's pronoun resolver was also used to process the texts selected by clinical partners for the purpose of reading comprehension testing. Its performance results are displayed in Table 12.

File name	Recognised	Resolved	Correct
BalansiranoHranene	3	3	2
ChereshaSAleniSartza	16	15	6
MorskiObitateli	1	1	0
Pcheli	2	2	1
StraniOtES	2	2	1
ZlatnoSxkrovishte	4	4	2
TOTAL	**28**	**27**	**12**

Table 12.Pronoun resolution on FIRST's pilot data set

Error analysis was conducted with regard to this small test sample, with errors categorised as follows:

- No Matches (NM): missed anaphors (anaphoric pronouns for which no antecedent was proposed).
- Wrong Matches (WM): wrong antecedent chosen.
- Spurious Matches (SM): non-anaphoric expression was resolved.

	NM	WM	SM	Examples
BalansiranoHranene	2	1	0	(NM's: организмани; коитопринатрупванетоси)
ChereshaSAleniSartza	4	9	0	(NM's: ние, децата; катоонзи, койтотежитристакилограма; тяхнатачерница; Хоратагикачиха)
MorskiObitateli	0	1	0	NA
Pcheli	1	1	0	(NM's: техния живот;)
StraniOtES	1	1	0	(NM's: даопределисобственитесицели)
ZlatnoSxkrovishte	0	2	0	(NM's:)
TOTAL	8	15	0	

Table 13: Categories of error in Bulgarian pronoun resolution.

The categorization of errors used in this experiment enables an evaluation of the meaning disambiguator with resect to pronoun resolution:

Precision (correct/(correct+WMs+SMs)): 12/27= 44.44%
Recall (correct/(correct+WMs+NMs)): 12/35= 34.3%

The statistics presented in table 12 and Table 13 indicate an overall precision score of 44.4%. If the results derived from processing ChereshaSAleniSartza, which is a difficult literary text, are ignored, overall precision is at the level of 50%. This is a reasonable baseline for a system that has not been tuned and optimised for the domains at hand.

The main source of errors for the wrong matches is ambiguous syntactic agreement and possibly lack of grammatical function detection, such as subject and object, and/or higher level syntactic information, for example appositions (e.g., "ние, децата", which translates to "us, the children"). This can be addressed by using a more sophisticated morpho-syntactic parser with higher precision for recognising syntactic features, grammatical functions and more complex syntactic structure.

One wrong match made in the text MorskiObitateli was due to an overly complicated link in which a coordinate NP was taken as antecedent of a 3rd person, plural pronoun. Generally, anaphoric links between plural anaphors and coordinate NPs are relatively rare (compared with the linking of plural anaphors to plural NPs). For this reason, the resolution algorithm may be modified to exclude coordinate NPs from the sets of possible antecedents from which it will select when resolving plural pronouns.

Possessive and reflexive pronouns are not handled by the system and account for most of the no matches.

Spurious matches are more relevant in the resolution of definite descriptions, not pronouns.

- **Word sense disambiguation**

Evaluation of the automatic word sense disambiguation method employed by the meaning disambiguator was derived by use of a corpus developed for use in the SEMEVAL 2013, Multilingual Word Sense Disambiguation task.[6] The corpus was created for evaluating the task of word sense disambiguation by reference to the senses (synsets) of different words in the text, by means of semantic resources developed by the Natural Language Processing community: Balkanet (Stamou et al., 2002) for Bulgarian, Wordnet (Fellbaum, 1998) for English and Multiwordnet (Atserias et al., 1997) for Spanish. Currently, the SEMEVAL corpus is only available for English and Spanish, which prevented automatic quantitative evaluation of word sense disambiguation for Bulgarian.

The evaluation of WSD for English and Spanish was performed by checking whether the methods developed were able to assign the correct synset to each word in the test data. Specifically, we tested the following two methods that were integrated in the Freeling tool: the Most Frequent Sense (MFS) and the UKB approach based on application of Personalised PageRank to a Lexical Knowledge Base (more information about these methods can be found in D4.ii).

The following table shows the results of both methods for English and Spanish.

Language	WSD method	WSD Precision
English	MFS	52.10%
	UKB	45.80%
Spanish	MFS	55.20%
	MFS-Freeling	17.90%
	UKB	42.30%

Table 14. Evaluation of the performance of Word Sense Disambiguation methods

Performance of the Spanish WSD methods was worse than expected (see D4.ii). For this reason, the system was evaluated under two different settings:

- Relying on the synset annotations provided by Freeling.
- Obtaining the MFS directly from the resources employed (i.e., for a word we considered its MFS, the first returned sense of MultiWordnet, in the same manner we did for English).

[6] http://www.cs.york.ac.uk/semeval-2013/task12/index.php?id=data Only recently made available, this resource was acquired after production of D7.3.

As is evident from Table 15, performance of the MFS method is much improved when annotations provided by Freeling or those obtained directly from MultiWordNet are exploited. In order to be sure that the process did not have any errors, the MFS for Spanish was tested on documents in which words had been manually annotated with their senses. These documents were part of the annotated corpora we are developing for evaluating all the obstacles in a quantitative manner. The results obtained are shown below:

Language	WSD method	WSD Precision
Spanish	MFS	30.20%
	MFS-Freeling	57.45%

Table 15. Evaluation of the performance of MFS methods for the annotated documents

In this case, we observed that the annotations provided by Freeling led to better performance than annotations obtained via the first sense of each word provided in MultiWordnet. A detailed analysis of the SEMEVAL corpus indicated that the Spanish fragment of that corpus had been disambiguated in accordance with the most frequent senses of words derived from the English WordNet. It is likely that the most frequent senses of words in English do not directly match the most frequent senses in Spanish. Such discrepancies affected the results of WSD in Spanish (obtaining 17.90%). For this reason, the results are not considered meaningful and the system relies instead on the annotations provided by Freeling.

Qualitative evaluation (difficult words)

Due to the lack of suitable test data, it was necessary to analyse system output by hand in order to evaluate modules detecting and removing obstacles to reading comprehension caused by the occurrence of difficult words in input texts. These obstacles include polysemous words, complicated mental verbs, infrequent words, long words and adverbs with *–ly* (EN) and *–mente* (ES) suffixes, infrequent acronyms/abbreviations, specialised words, and infrequent slang.

In each case, we analysed:

- The number of obstacles that could be detected in a text (detection),
- The number of definitions that could be obtained for the detected obstacles (one step toward removal), and
- The number of synonyms that could be obtained for the detected obstacles (one step toward removal).

Where possible, these studies were made for all three languages (BG, EN, and ES). The statistical results of this analysis are presented in Tables 16-22. In these tables, "-" indicates that no methods have been employed to detect or remove a particular obstacle.

For this evaluation, two different corpora were used:

- A subset of 500 newswire texts that were randomly selected from the newswire corpora described in D4.i and D4.ii. (Corpora available for Bulgarian, English, and Spanish)
- The set of texts collected by clinical partners for use in reading comprehension testing (to be presented in D7.6 (January 2014).

Tables 16-18 present results obtained by analysing the output of the meaning disambiguator when processing the 500 newswire texts.

Obstacle	Detection (# of words detected as an obstacle)	Number of definitions	Number of synonyms
Polysemous words	7366	7366	7366
Complicated mental verbs	12	4	4
Infrequent words	51594	5935	5935
Long words & -ly and -mente adverbs	45935	6451	6451
Infrequent acronyms/abbreviations	-	N/A	N/A
Specialised	4880	4880	4880
Infrequent slang	0	N/A	N/A

Table 16. Global average results for the 500 documents (Bulgarian)

Obstacle	Detection (# of words detected as an obstacle)	Number of definitions	Number of synonyms
Polysemous words	52966	52966	52966
Complicated mental verbs	1355	1355	1355
Infrequent words	23423	22298	22298
Long words & -ly and -mente adverbs	23853	23085	23085
Infrequent acronyms/abbreviations	129	N/A	N/A
Specialised	15644	15644	15644
Infrequent slang	0	N/A	N/A

Table 17. Global average results for the 500 documents (English)

Obstacle	Detection (# of words detected as an obstacle)	Number of definitions	Number of synonyms
Polysemous words	32321	12839	32321
Complicated mental verbs	0	0	0
Infrequent words	19376	5353	10699
Long words & -ly and -mente adverbs	20775	7122	15197
Infrequent acronyms/abbreviations	248	N/A	N/A
Specialised	7564	3884	7564
Infrequent slang	0	N/A	N/A

Table 18. Global average results for the 500 documents (Spanish)

Two observations were made as a result of the cross-language analysis of system performance:

1. While the modules detect a large number of obstacles, the numbers of definitions and synonyms provided differs according to the language being processed. Relatively few obstacles were removed from Bulgarian texts. This is due to the poorer level of coverage provided by the semantic resources available for that language (Balkanet). Analysis of the system performance when processing Spanish texts led to a similar observation: for many of the detected obstacles (words), the resources used lack sufficient information to provide definitions or synonyms. Furthermore, the current implementation of the LT services does not yet exploit lemmatisers for Bulgarian, English, and Spanish. In subsequent months, the feasibility of acquiring and exploiting lemmatisers to improve the performance of the meaning disambiguator will be investigated. The problem may also be mitigated by the use of alternate semantic resources that offer better coverage.
2. The test data consists of newswire, which is quite a formal category of text. As a result, no instances of infrequent slang (UR407, UR414, UR423, and UR424) were detected in this data.

Tables 18-20 present statistical results derived by analysis of system output when processing texts selected by clinical partners for use in reading comprehension tests (6 texts in Bulgarian for children; 6 texts in English for adults, and 12 texts in Spanish – 6 for children and 6 for adults). The tables below show the global results for each language.

Obstacle	Detection (# of words detected as an obstacle)	Number of definitions	Number of synonyms
Polysemous words	79	79	79
Complicated mental verbs	0	0	0
Infrequent words	746	62	62
Long words & -ly and -mente adverbs	567	48	48
Infrequent acronyms/abbreviations0	-	N/A	N/A
Specialised	42	42	42
Infrequent slang	0	N/A	N/A

Table 18. Global results for the Bulgarian clinical texts

Obstacle	Detection (# of words detected as an obstacle)	Number of definitions	Number of synonyms
Polysemous words	1877	1877	1877
Complicated mental verbs	67	67	67
Infrequent words	735	624	624
Long words & -ly and -mente adverbs	789	700	70
Infrequent acronyms/abbreviations	0	N/A	N/A
Specialised	385	385	385
Infrequent slang	0	N/A	N/A

Table 19. Global results for the English clinical texts

Obstacle	Detection (# of words detected as an obstacle)	Number of definitions	Number of synonyms
Polysemous words	1369	496	1369
Complicated mental verbs	0	0	0
Infrequent words	767	199	454
Long words & -ly and -mente adverbs	878	295	662
Infrequent acronyms/abbreviations	1	N/A	N/A
Specialised	317	150	317
Infrequent slang	0	N/A	N/A

Table 20. Global results for the Spanish clinical texts

These analyses are based on system output derived by processing a relatively small number of texts. It was noted that some of the potential obstacles to reading comprehension are very scarce (e.g. acronyms). For some of the obstacles, detection is achieved by consultation of specific word lists (e.g. lists of acronyms, lists of mental verbs). The consequence of this is that detection is restricted only to elements included in these gazetteers. It is therefore necessary to analyse the degree of coverage provided by each list. Linguistic ambiguity means that not every occurrence of an item in the word lists may be a genuine instance of the phenomena listed in Table 2, Section 2.2.2. It is therefore also necessary to evaluate the extent to which the occurrence of particular entries in the word lists indicates the occurrence of difficult words in a particular text. These types of evaluation will begin after completion of the manual annotation of the corpus.

As in the case of the 500 newswire texts, the language of the texts forming the basis of reading comprehension tests is quite formal. As a result, slang is not used in these documents, as it is normally used in less formal texts.

Table 21 displays the percentages of different types of *potential* obstacle to reading comprehension that were successfully detected *and* removed by the meaning disambiguator that are real obstacles for people with ASD. The clinical partners based in Spain linguistically analysed the texts used in the reading

comprehension tests. Manual comparison was then made between the genuine obstacles identified by the clinicians and the potential obstacles detected and removed by the system.

Detection	Definition-Resolution	Synonym - Resolution
95.83%	25%	50%

Table 21. Percentage of genuine obstacles to reading comprehension detected and removed by our tool

These results reveal the limitations of the semantic resources employed. While the tool is able to detect almost all obstacles, definitions and/or alternate synonyms of difficult words cannot always be provided. It can be noted that the semantic resources exploited (i.e. MultiWordNet) enable more effective access to synonyms of words than to their definitions. Since the framework and the modules developed in WP4 have been designed to enable exploitation of alternate semantic resources, future work will include identification and acquisition of useful alternatives to integrate.

5. Evaluation and Error Analysis: Personalised Document Generator (v2)

5.1 Method

In the Sections 5.2-5.3, evaluation of both the Online and Offline Image Retrieval modules is described with reference to a set of documents simplified by the Software prototype, named *Open Book*. The documents for which these evaluation experiments were run were those selected for use in the reading comprehension tests to be presented in D7.6 (January 2014).

The test set is composed of 12 documents: 6 for adults in English and 6 for adults in Spanish. Evaluation of the image retrieval process is made with respect to a set of query terms. The terms for which evaluation will be performed were randomly chosen from the Specialized Term Set, constructed by the Term Identification module. Table 22 displays the set of terms randomly chosen for each document.

Document Name	Selected Terms	Language
Document 1: An Unlikely Muse	chaos ,plunder, inspiration, embezzlement ,atrocity extravagance architecture infrastructure, heroine homosexual	English
Document 2: Skara Brae	monument, encroachment ,necklace ,seaweed ,flint ,whale, radiocarbon, excavation, landowner, tide	English
Document 3: Wind Power in the US	turbine, pollution telecommunication habitat, inventor, developer, megawatt, researcher, funding, projection	English
Document 4: Camberwell College Swimming Pools	tadpole, swimmer, stamina, fitness	English
Document 5: Gateway Academy Pre-Sessional Courses	coursework, gateway, tutor, tourist, lecturer, tuition	English
Document 6: The Shock of the Truth	churchman, playwright, mathematician, theologian, envelope, astronomer	English
Document 1	ignorancia, perdón, partidario, bahía, exhibición, taquilla, diseñador, lanzamiento, dictador	Spanish
Document 2	chimenea, hierba, sedimento, bahía, depósito, trigo, joya habitante, comodidad, alga	Spanish
Document 3	Instantánea, carbón, electricidad, carbono, archivo	Spanish
Document 4	alumno, adulto, piscina, monitor	Spanish
Document 5	herramienta, estudiante, informática, asignatura,instrucción, certificado, expulsión	Spanish
Document 6	incredulidad, editor, esfera, navegante, órbita	Spanish

Table 22. Selected Terms for Image Evaluation for Each Text.

The evaluation procedure is the following:

1. Each annotator evaluates the suitability of the retrieved image with regard to terns considered (i) intrinsically and (ii) in context. Both types of evaluation are necessary because the retrieved image may be appropriate for illustrating the intrinsic concept, but it might not be appropriate for illustrating the concept in particular contexts.
2. The annotators assign to each concept a label (*Abstract* or *Concrete*) denoting the level of abstractness of concepts. This information will be used to guide future development of the tool.

3. The Kappa agreement (Siegel and Castellan, 1988) between annotators is then computed and reported.
4. The precision of image retrieval is computed for both types of evaluation (terms considered intrinsically and in context) and reported.

5.2 Evaluation of Online Image Retrieval Module

The Online Image Retrieval module was used to retrieve image URLs for the concepts listed in table 22. For each concept, the first two images retrieved by each engine were evaluated. Two annotators read the texts, identified the specialized concepts and annotated the appropriateness of the retrieved image for concepts considered both intrinsically (CI) and in context (CC).

The quality of the image retrieval modules was measured using Kappa score to assess the degree to which two human annotators considered the images returned by the systems to be appropriate. These results are displayed in Table 23.

Language	Google		Bing		Global	
	CI	CC	CI	CC	CI	CC
English	0.57	0.62	0.76	0.72	0.68	0.68
Spanish	0.62	0.61	0.52	0.53	0.57	0.57

Table 23. The inter-annotator agreement for Online Image Retrieval Module

According to the interpretations of K-Score proposed by Landis and Koch (1977), any agreement score between 0.41 and 0.60 is moderate. Agreement scores between 0.61 and 0.80 are considered substantial. Given that the annotators performed the task with little training, the obtained K-Scores, which vary between 0.53 and 0.76 (from moderate to substantial agreement on the Landis and Koch scale) can be considered acceptable for the task. For English the annotators showed a preference for images retrieved by the Bing search engine but for Spanish they showed a preference for those returned by the Google search engine. The annotators show roughly the same level of agreement regardless of whether concepts are considered intrinsically or in context.

The accuracy of the retrieval procedure is computed in the customary way as the proportion of correctly identified images divided by the total number of images. The computation is done both when concepts are considered intrinsically and in context, and for each annotator. The reported accuracy is the average accuracy score for the two annotators. The accuracy results are reported for each annotator and overall in tables 24, 25 and 26 respectively:

Language	Google		Bing		Global	
	CI	CD	CI	CD	CI	CD
English	0.73	0.67	0.58	0.51	0.66	0.59
Spanish	0.65	0.50	0.74	0.54	0.70	0.52

Table 24. Accuracy according to the first annotator of the Online Image Retrieval Module

Language	Google		Bing		Global	
	CI	CD	CI	CD	CI	CD
English	0.57	0.52	0.49	0.44	0.53	0.48
Spanish	0.56	0.42	0.64	0.5	0.6	0.46

Table 25. Accuracy according to the second annotator of the Online Image Retrieval Module

Language	Google		Bing		Global	
	CI	CD	CI	CD	CI	CD
English	0.65	0.60	0.54	0.48	0.59	0.54
Spanish	0.60	0.46	0.69	0.52	0.65	0.50

Table 26. The overall accuracy of the Online Image Retrieval Module

As expected, the accuracy with which modules retrieve suitable images for concepts considered in context is lower than that achieved when concepts are considered intrinsically. This indicates it is necessary, in future development, to improve the Online Image Retrieval web service by incorporating contextual information.

5.3 Evaluation of Offline Image Retrieval Module

The Offline Image Retrieval Module is based on the Wikipedia Disambiguation Web Service. Unfortunately, not all concepts listed in Table 22 can be mapped onto Wikipedia. For English and Spanish, 77% of the concepts can be mapped. The average accuracy with which concepts can be mapped is relatively high: 95% for English and 88% for Spanish.

The extent to which suitable images could be retrieved for successfully mapped concepts, both when considered intrinsically and in context, was measured. The evaluation task was performed by two human annotators. As in the case of the online image retrieval module, inter-annotator agreement (Kappa) between the human annotators and the systems was obtained. The results are displayed in table 27:

Language	CI (Intrinsic)	CC (In Context)
English	0.77	0.70
Spanish	0.79	0.71

Table 27. The inter-annotator agreement for Offline Image Retrieval Module.

The agreement figures are higher than was the case of the Online Image Retrieval module. These figures indicate substantial agreement, according to the scale proposed by Landis and Koch.

The accuracy with which the systems are able to retrieve suitable images for concepts, both intrinsically and in context, for each annotator and overall, are displayed in Tables 28-30.

Language	CI (Intrinsic)	CC (In Context)
English	0.88	0.81
Spanish	0.63	0.54

Table 28. The precision for the first annotator for Offline Image Retrieval Module

Language	CI (Intrinsic)	CD(In Context)
English	0.88	0.88
Spanish	0.72	0.68

Table 29. The precision for the second annotator for Offline Image Retrieval Module

Language	CI (Intrinsic)	CD(In Context)
English	0.88	0.85
Spanish	0.68	0.61

Table 30. The overall precision for Online Image Retrieval Module

Accuracy scores for image retrieval of concepts considered both intrinsically and in context are greater than those obtained for the Online Image Retrieval Module. As in the case of the Online Image Retrieval Web Service, slightly reduced levels of accuracy were obtained for image retrieval of concepts considered in context.

6. Conclusions and Recommendations for RTD in the Next Six Months

User requirements such as UR312 and UR418 suggest that end users have low tolerance for ungrammatical and erroneous output generated by the LT services. The error analysis conducted in this report suggests that end users should not yet be provided with direct access to many of the services developed so far in the FIRST project. Instead, it is anticipated that most of the services will be applied in scenarios in which intermediaries convert documents into a more accessible form in a semi-automatic process exploiting output from the Open Book prototype developed in the FIRST project. Structural complexity processing, word sense disambiguation, and coreference resolution should only be applied in use cases in which system output will be curated by a human intermediary. It is hypothesised that manual curation of automatically generated output will be more rapid and less onerous than fully manual conversion of documents into an accessible form for end users. This hypothesis will be tested in the final year of the project (to be reported in D7.8). If made accessible via the GUI for end users, these users should be made aware of the low reliability of the services. In the next six months, the efficacy of the LT components will be improved on the basis of the error analysis described in this report. Complementary information via user feedback will be required in order to judge the tolerance of all types of user to the errors reported here.

Research in the final year of the project will evaluate the extent to which the LT components facilitate conversion of documents into a more accessible form by intermediaries. This evaluation will include an assessment of the relative effort involved in manual conversion of documents into a more accessible form and effort involved in curating the output of the prototype. It will also seek to quantify the appeal of the prototype to end users and intermediaries.

6.1 Processing structural complexity

Error analysis revealed that fully automatic conversion of compound and complex sentences into a more accessible form is quite unreliable, particularly for text of the literature category. It was noted that conversion of complex sentences into a more accessible form is more difficult than conversion of compound sentences. However, as noted in D7.3 subordinate clauses are significantly more prevalent than conjoint clauses in the training and testing data collected so far. While the system achieves improvements in readability for some categories of text, there is limited reduction in the average length of sentences in texts generated by the system. The evaluation of the rule sets used in the conversion of compound and complex sentences into a more accessible form motivates further specific development of the rule sets. This process includes deletion of rules that do not meet particular thresholds for accuracy and the development of new rules to address cases where input sentences fail to trigger any conversion rules (signalled by activation of redundant rules CCV-25a and SSCCV-78a). Although these findings are negative, it should be noted that in some use cases, use of the system will be semi-automatic. Provision is made for errors caused by the structural complexity processor to be manually corrected by carers/ intermediaries. The reduction in effort afforded by use of Open Book in this context is yet to be assessed. This will be investigated by comparing the time taken by intermediaries to convert a document into a more accessible form entirely by hand with the time taken to make the conversion by post-editing or otherwise exploiting the output of the services incorporated in the FIRST prototype.

With regard to the evaluation of the structural complexity processor, investigations will be made in the next six months into the reliability of the two assumptions that all structural obstacles to reading comprehension are indicated by signs of syntactic complexity (recall of the method, with regard to user

requirements) and that all signs of syntactic complexity indicate an obstacle to reading comprehension (precision of the method with regard to user requirements).

.

Meeting user requirements requires both detection and removal of structural obstacles to reading comprehension. While twelve of nineteen user requirements are met in a fully automatic way with regard to detection, only eight of nineteen are met in this fashion with regard to removal. Requirements that are not currently met by this service can be categorised as follows. They are ranked according to expected difficulty of automatic removal:

a) Requirements involving processing of double negation
b) Requirements involving movement of specific types of clause (UR307, UR308)
c) Requirements involving additional types of sentence rewriting (UR313)
d) Requirements involving identification and deletion of specific types of clause (UR303)
e) Requirements involving substitution of individual words or punctuation (UR304, UR305, UR306, UR310, UR311)

It is expected that detection of linguistic phenomena related to these types of user requirement will be developed in the next six months. Further, it is expected that user requirements of type (e) can be addressed significantly more easily than those of other types. An attempt to automatically address requirements of type (e) will be made in the next three months. Resources are currently being developed to detect and substitute items of this type. The ambiguity of certain words that may be adversative or final/illative conjunctions in some contexts and adverbs in others (e.g. *rather*, *so*) and the complexity of the patterns in which comparative conjunctions appear means that the process is not completely trivial. During the next six months, prototype services will be developed in order to assess the feasibility of automatically addressing user requirements of types (c) and (d). It is considered that there are insufficient resources in a project of this size to develop services fully addressing user requirements of types (a) and (b) in an automatic way.

6.2 Processing ambiguity in meaning

The methods developed for the purpose of removing obstacles to reading comprehension caused by difficult words are based on gazetteer look up. So far, neither the quality of the gazetteers or of the matching algorithm has been assessed. In the next six months, linguistic analysis will be performed in order to quantify the number of cases in which:

1. The occurrence of an item listed in some gazetteer is not a reference to the concept type indicated by that gazetteer.
2. References to concepts of a type indicated by their presence in some gazetteer are not detected by the system due to linguistic variation in the input text.

Before delivery of the software prototype (D6.4), improved semantic resources for Bulgarian and Spanish will be acquired. Acquisition of these resources will enable improved recall of the word sense disambiguation service over a larger number of concepts than is currently available by means of BalkaNet, MultiWordNet, or Freeling.

Another significant cause of error in all services based on consultation of semantic resources is word variation. Resources that do not provide distinct entries for morphologically inflected forms of content

words are incapable of providing a satisfactory response to user queries. This problem will be averted by integration in the next six months of a lemmatiser for Bulgarian, English, and Spanish. This component will enable the system to access semantic information about the morphological variants of a particular word via a single lemma common to all of its variants. In the next six months, the current preliminary implementation of a coreference resolution system for Bulgarian will also be updated and improved by exploitation of more sophisticated pre-processing tools.

6.3 Generation of personalised documents

Having computed agreement and precision and annotating concepts according to their degree of abstractness for the Image Retrieval web Services in two modalities, we reached the following conclusions:

1. The annotator agreement is moderate to substantial for Online Image Retrieval Web Service and substantial for Offline Image Retrieval web service.
2. The precision of the Offline Image Retrieval web service is better than the precision of Online image retrieval web service but the recall is much lower. We conclude that the two web services are complementary.
3. The drop in precision between the two modalities invites incorporation of contextual factors in the future versions of the two web services.
4. The annotators failed to reach a significant level of agreement for the abstract /concrete annotation of concepts. In subsequent month, ontologies may be exploited in order to verify the impact of this aspect in image retrieval.

In addition to improvements pertaining to the modules for image retrieval, the current personalised document generator currently lacks functions to generate additional assistive content in order to meet additional user requirements. In the next six months, the development of modules to meet user requirements UR501, UR503, and UR504 will be finalised. Methods exploiting large data sets will be applied to derive the main topics or themes of input documents. This information will then be displayed at the beginning of the personalised document. Existing approaches to automatic multiple choice question generation will be applied in order to generate reading comprehension questions from input documents. Finally, information will be provided on key concepts identified in the document to produce glossaries that end users may access.

References

Asterias J., Climent S., Farreres X., Rigau G. And Rodriguez H. (1997) Combining multiple methods for the automatic construction of multilingual Wordnets. In *Proceedings of the International Conference "Recent Advances on Natural Language Processing" RANLP'97*, Tzigov Chark, Bulgaria.

Attardi, G., Dei Rossi, S., and Simi, M. (2010) The Tanl Pipeline. In *Proceedings of Workshop on Web Services and Processing Pipelines in HLT, co-located LREC 2010*, Malta.

Fellbaum, C. (1998) *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Gaizauskas, R.,Foster, J., Wilks, Y., Arundel, J., Clough, P. and Piao, S. 2001. The METER corpus: a corpus for analysing journalistic text reuse. In *Proceedings of the Corpus Linguistics 2001 Conference*. Lancaster. p. 214-223.

Landis, R., & Koch, G. J. (1977). The measurement of observer agreement for categorical data. *Biometrics* (33), 159-174.

Padró, L. (2011) Analizadores Multilingües en FreeLing. *Linguamatica*, 3:2, 13-20. December, 2011.

Recasens, M., Marquez, L., Sapena, E., Antonia, M., Taule, M., Hoste, V., Poesio, M., and Versley, Y. (2010) SemEval-2010 Task 1: Coreference Resolution in Multiple Languages, *Proceedings of the 5th International Workshop on Semantic Evaluation*, July 2010, Uppsala, Sweden, Association for Computational Linguistics, pp. 1-8, http://www.aclweb.org/anthology/ S10-1001.

Siegel, S., & Castellan, N. (1988). *Nonparametric statistics for the Behavioral Sciences.* McGraw-Hill.

Stamou S., Oflazer K., Pala K., Christoudoulakis D., Cristea D., Tufiş D., Koeva S., Totkov G., Dutoit D., Grigoriadou M. (2002). BALKANET: A Multilingual Semantic Network for the Balkan Languages. In *Proceedings of the International Wordnet Conference*, January 21-25, Mysore, India, 12-14.

Tanev, H. and Mitkov, R. (2002): Shallow language processing architecture for Bulgarian In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-2002)*, 995-1001. Tapei, Taiwan.