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.

D7.8: Final evaluation report

Author: Vesna Jordanova, Richard Evans, and Arlinda Cerga Pashoja

Affiliation: LNFT, WLV

Date: 19th November 2014

Document Number: FIRST_D7.8_20141119

Status/Version: Approved v 1.1

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	30th September 2014
Actual Date of Delivery	19th November 2014
Document Number	FIRST_D7.8_20141119
Status & Version	Approved v1

Number of Pages	100
WP Contributing to the Deliverable	WP7
WP Task responsible	Vesna Jordanova
Authors	Richard Evans and Arlinda Cerga Pashoja
Other Contributors	The whole consortium
Reviewer	Constantin Orasan
EC Project Officer	Krister Olson
Keywords:	evaluation, intrinsic evaluation, extrinsic evaluation, qualitative methodology, quantitative methodology, text accessibility, reading comprehension, text editing, language technology, readability

Abstract:

In the FIRST project, language technologies (LT) were acquired and developed to enable deployment of two services intended to convert texts into a more accessible form for people with autism. In one, the service is fully automatic, with end users directly consuming the output of the LT. In the other, the LT is exploited by intermediaries (carers) to assist in the manual conversion of texts into a more accessible form. In this context, conversion is a semi-automatic process. The set of language technologies integrated to support these text conversion services is called *OpenBook*. The quality of the text conversion services was assessed using two broad categories of evaluation: extrinsic and intrinsic. *Extrinsic evaluation* assesses the text conversion services from the perspective of users, intermediaries exploiting the LT to convert texts, and end users (people with autism), who are the final consumers of the text conversion services. Extrinsic evaluation includes an investigation of the impact of the service on the reading comprehension of end users and of the perceived usefulness of the LT for intermediaries. These aspects of extrinsic evaluation are conducted using quantitative methods (reading comprehension tests). Another important aspect of extrinsic

evaluation involves the use of qualitative methods to investigate the impact of the text conversion services on quality of life, taking into account perceived changes in independence, education, and stigmatisation. Interviews with end users and intermediaries were conducted and the transcriptions analysed to support this aspect of extrinsic evaluation. *Intrinsic evaluation* involves quantitative assessment of the LT components supporting text conversion. This evaluation is made via standard methods used in the field of natural language processing (NLP) together with assessments of the readability of texts in their original forms and in converted forms generated by intermediaries exploiting OpenBook. Comparisons are drawn between the readability of texts converted by carers using OpenBook and a benchmark comprising the readability of texts produced by unaided conversion of texts (D7.2).

Contents

1. Executive Summary

In the FIRST project, language technologies (LT) were acquired and developed to enable deployment of two services intended to convert texts into a more accessible form for people with autism. In one, the service is fully automatic, with end users directly consuming the output of the LT. In the other, the LT is exploited by intermediaries (carers) to assist in the manual conversion of texts into a more accessible form. In this context, conversion is a semi-automatic process. The set of language technologies integrated to support these text conversion services is called *OpenBook*. The quality of the text conversion services was assessed using two broad categories of evaluation: extrinsic and intrinsic. *Extrinsic evaluation* assesses the text conversion services from the perspective of users who include intermediaries exploiting the LT to convert texts and end users (people with autism), who are the final consumers of the text conversion services.[1] Extrinsic evaluation includes an investigation of the impact of the service on the reading comprehension of end users and of the perceived usefulness of the LT for intermediaries. This aspect of extrinsic evaluation was conducted using quantitative methods (reading comprehension tests). Another important aspect of extrinsic evaluation involves the use of qualitative methods to investigate the impact of the text conversion services on quality of life, taking into account perceived changes in independence, education, and stigmatisation. Interviews with end users and intermediaries were conducted and the transcripts analysed to support this aspect of extrinsic evaluation. *Intrinsic evaluation* involves quantitative assessment of the LT components supporting text conversion. This evaluation is made via standard methods used in the field of natural language processing (NLP) together with assessments of the readability of texts in their original forms and in converted forms generated by carers exploiting OpenBook. Comparisons are drawn between the readability of texts converted by carers using OpenBook and the readability of texts produced by unaided conversion, which constitutes a benchmark (D7.2).

[1] Throughout this report, the term *user* will be use to denote carers and end users. The term *carer* will be used to denote people using the OpenBook tool to convert texts into a more accessible form for end users (third parties). The term *end user* will denote people with autism consuming the automatic and semi-automatic text conversion services supported by the OpenBook tool.

2. Extrinsic Evaluation

In the FIRST project, software was developed to support the provision of two text conversion services. In the first, a limited number of LT services are integrated in an interface providing end users with the ability to improve the accessibility of input texts. In the second, automatic LT services are integrated in an editing tool to enable carers such as teachers, healthcare professionals, family members, and friends to convert texts into a more accessible form for end users. The software for carers and end users is referred to as OpenBook. Extrinsic evaluation of Open Book consists of two parts. The first is a study of the impact on end users of accessing the software directly for the purpose of automatic text conversion. The second is a study of the impact on end users of accessing texts converted by carers using OpenBook to facilitate the conversion process. In the project, extrinsic evaluation comprises quantitative methods such as reading comprehension testing together with qualitative methods used in the conduct and analysis of interviews with end users and carers.

2.1 Reading Comprehension

Reading comprehension testing was used to test the effectiveness of Open Book as a tool used to convert texts into a more accessible form for end users (people with Autistic Spectrum Disorders (ASD)). 243 patients (193 males and 50 females) with high-functioning autism were recruited in the UK, Spain and Bulgaria, with 50 controls from Bulgaria. Comprehension tests were conducted in a controlled environment under time-limited conditions. Each participant was presented with a battery of texts followed by multiple choice questions and a subjective text rating. Half of the texts were presented in their original form while the remainder were presented in a more accessible form generated by carers using OpenBook. The order of text presentation was random and both researchers and participants were blind to this ordering. Participants were also blind to the status of each text (original or converted). More details about how the reading comprehension tests were carried out are presented in deliverable D7.6. In this Section, analysis of participants' responses to the tests is presented. In the tests, items testing reading comprehension of a text in its original form are denoted $Item_{ORIG}$. Items testing reading comprehension of a text in a form converted using OpenBook are denoted $Item_{SIMP}$. References to more than one item will be denoted $Items_{ORIG}$ and $Items_{SIMP}$, respectively.

2.1.1 Outcomes

2.1.1.1 Primary outcomes

The primary outcome of testing is the comprehension score calculated by adding the item scores for each question. Scores obtained from *Items*$_{SIMP}$ were compared with scores obtained from *Items*$_{ORIG}$. Items presented to adults consisted of a text followed by 6 questions. Items presented to children consisted of a text followed by 4 questions. Every correct answer was scored 1, and each incorrect answer was scored 0. As a result, each item score could range from 0 (no correct answers) to 6 (all correct answers) for adults, and 0-4 for children. Each test contained 3 items. The overall score for a test was calculated by aggregating the score obtained for each of the 3 items presented. The overall range of scoring values for adults are 0-18 (3 texts × 6 questions per test) and for children 0-12 (3 texts × 4 questions). Please see Appendix 1 for an example of the scoring method.

The main study hypothesis was that text conversion improves reading comprehension for participants with ASD therefore the overall score for *Items*$_{SIMP}$ will be higher than the overall score for *Items*$_{ORIG}$. This hypothesis was tested by repeated measures t-tests considering the scores satisfying the parametric data conditions.

2.1.1.2 Secondary outcomes

There are three secondary outcomes of the testing process:

1) *To explore the 'subjective rating' of texts by participants between conditions (original vs converted texts).*

   Subjective rating is measured in a Likert type scale ranging from 1 (very easy) to 5 (very difficult) (Appendix 2). Therefore, the range of scores for each text is 1-5 and overall (three texts) 3-15 (detailed scoring example provided in Appendix 1). Higher scores indicate higher level of difficulty; while lower scores indicate the opposite. It is hypothesized that participants will find converted texts easier to understand and therefore rate them lower overall. T-tests will be used to investigate this hypothesis. Pearson's correlation coefficient will also be used to test the relationship between comprehensive scoring and subjective rating.

2) *To examine the relationship of 'time required to complete reading task' with text condition (original vs converted text).*

3) *To explore the effects of demographic measures on the primary outcome by means of logistic regression.*

A feature of the study was that each participant was assessed using both $Items_{ORIG}$ and $Items_{SIMP}$. As a result there was a ‘pair’ of measurements from each participant. An examination of the data suggested that the differences in text and subjective scores between the two methods were approximately normally distributed. As a result the paired t-test was used for the analyses.

Subjects originally had three timing variables, each of which was classified as 'could not finish, 'on time' or 'before time'. A combined measure of time was calculated which was the number of the three item that were completed either on time or before time. The differences in this measure between the scores obtained for $Items_{ORIG}$ and $Items_{SIMP}$ were found to be normally distributed, and so the paired t-test was also used for these analyses.

Analyses were conducted in series, initially for all ASD participants, and then for different subgroups.

A second set of analyses compared the difference in scores between Bulgarian participants with ASD and Bulgarian controls. Comparisons were made for both $Items_{ORIG}$ and $Items_{SIMP}$, and also for the difference between the two types of item. The unpaired t-test was used for the analyses.

An additional analysis examined the association between the text and subjective scores. As both variables were continuous in nature, Pearson correlation was used for the analyses. All participants were included in the analyses, but separate analyses were performed scores obtained for $Items_{ORIG}$ and $Items_{SIMP}$.

The objective of the analyses was to examine factors associated with the scores obtained for $Items_{SIMP}$ in the whole cohort of participants. The scores were found to be approximately normally distributed, and so linear regression was used for all analyses.

The separate association between each factor and the outcome was examined in a series of univariate analyses. Subsequently, the joint association between the factors and the scores was examined in multivariate analyses. This has the advantage that the association between each factor and the outcome is adjusted for the effects of the other variables. To simplify the final analysis, a backwards selection procedure was used to retain only the statistically significant variables. This involves omitting non-significant variables, one at a time, until all remaining variables are significant. Bulgarian controls were excluded from this analysis.

2.1.2 Results

2.1.2.1 Comparison between Items$_{ORIG}$ and Items$_{SIMP}$

The first analyses compared the scores obtained for the two sets of items, and a summary of the results is given in Table 1. The first figures reported are the number of participants in each analysis, and the mean and standard deviation for each set of scores. Also reported are the mean differences between the two types of item, along with a corresponding confidence interval. P-values indicating the significance of the results are also presented.

Participant group	N	*Items$_{ORIG}$* Mean (SD)	*Items$_{SIMP}$* Mean (SD)	Difference (*) Mean (95% CI)	P-value	Effect size d
Adults & children (†)	243	10.0 (4.1)	11.2 (4.1)	1.2 (0.9, 1.6)	**<0.001**	**0.3**
Adults	153	12.0 (3.5)	13.3 (3.3)	1.3 (0.8, 1.8)	**<0.001**	**0.4**
Children (†)	90	6.6 (2.6)	7.8 (2.8)	1.1 (0.7, 1.6)	**<0.001**	**0.4**
UK adults	99	12.3 (3.9)	13.8 (3.7)	1.5 (0.8, 2.2)	**<0.001**	**0.4**
Spain adults & children	95	9.3 (3.5)	10.6 (3.2)	1.3 (0.8, 1.7)	**<0.001**	**0.4**
Spain adults	54	11.5 (2.6)	12.4 (2.1)	1.0 (0.3, 1.7)	**0.009**	**0.4**
Spain children	41	6.5 (2.1)	8.1 (2.8)	1.7 (1.2, 2.2)	**<0.001**	**0.7**
Bulgaria children	99	8.7 (3.0)	9.2 (2.8)	0.5 (0.0, 0.9)	**0.03**	**0.2**
Bulgaria children (†)	49	6.8 (2.9)	7.4 (2.9)	0.7 (-0.1, 1.4)	0.08	
Bulgaria controls	50	10.7 (1.4)	11.0 (1.0)	0.3 (-0.2, 0.8)	0.22	

Table 1: Comprehension score analysis

(*) Difference calculated as converted score minus original score

(†) Omitting Bulgarian controls

The results of the majority of the analyses indicate higher scores for $Items_{SIMP}$ than for $Items_{ORIG}$. This was the case for all participants combined, and also for the different subgroups, with the exception of Bulgarian children, where there was only slight evidence of a difference between the two sets of scores.

The effect size (Cohen's d) measures the size of differences between the two means. The effect sizes were of medium magnitude overall, and for the children sample in Spain the effect size was large.

There was not found to be any difference between the two sets of scores for the Bulgarian controls.

In all subgroups where a difference was observed, scores obtained for $Items_{SIMP}$ were higher than scores obtained for $Items_{ORIG}$. When all participants (apart from controls) were combined together, the scores obtained from $Items_{SIMP}$ were 1.2 units higher than the scores obtained from $Items_{ORIG}$.

A similar set of analyses were performed for the subjective rating, and the results are summarised in Table 2. Higher scores represent more difficult text.

Participant group	N	$Items_{ORIG}$	$Items_{SIMP}$	Difference [(*)]	P-value	Cohen's
		Mean (SD)	Mean (SD)	Mean (95% CI)		d
Adults & children [(†)]	243	8.7 (2.6)	7.6 (2.4)	-1.0 (-1.3, -0.7)	**<0.001**	**0.4**
Adults	153	9.1 (2.3)	8.0 (2.2)	-1.2 (-1.6, -0.8)	**<0.001**	**0.5**
Children [(†)]	90	7.8 (2.9)	7.0 (2.7)	-0.8 (-1.3, -0.3)	**0.001**	**0.3**
UK adults	99	9.3 (2.3)	8.0 (2.1)	-1.3 (-1.8, -0.8)	**<0.001**	**0.6**
Spain adults & children	95	8.1 (2.4)	7.3 (2.4)	-0.8 (-1.2, -0.3)	**0.001**	**0.3**
Spain adults	54	8.7 (2.4)	7.8 (2.3)	-0.9 (-1.5, -0.3)	**0.006**	**0.4**
Spain children	41	7.3 (2.3)	6.7 (2.4)	-0.7 (-1.4, 0.1)	0.07	0.3
Bulgaria children overall	99	7.0 (3.0)	6.2 (2.6)	-0.9 (-1.3, -0.4)	**<0.001**	**0.3**
Bulgaria children with ASD[(†)]	49	8.3 (3.2)	7.3 (3.0)	-0.9 (-1.6, -0.3)	**0.008**	**0.3**
Bulgaria controls	50	5.8 (2.2)	5.0 (1.4)	-0.8 (1.4, -0.2)	**0.006**	**0.4**

Table 2: Analysis of subjective scoring

(*) Difference calculated as simplified scores obtained for $Items_{SIMP}$ minus scores obtained for $Items_{ORIG}$

(†) Omitting Bulgarian controls

The results for the subjective scores suggested that in all instances there was evidence of a difference between subjective scores for *Items*$_{ORIG}$ and subjective scores for *Items*$_{SIMP}$. However, the result was only of borderline statistical significance for Spanish children.

In all instances the scores for *Items*$_{SIMP}$ were significantly lower than the scores for *Items*$_{ORIG}$. When all participants (aside from controls) were included in the analysis, scores for *Items*$_{SIMP}$ were, on average, 1.0 units lower than the scores for *Items*$_{ORIG}$.

Comparisons between the two methods were also made for the timing measure, and the results are summarised in Table 3. The summary figures are the mean and standard deviation number of the three individual times where the task was completed on or before time.

Participant group	N	*Items*$_{ORIG}$	*Item*$_{SIMP}$	Difference [(*)]	P-value
		Mean (SD)	Mean (SD)	Mean (95% CI)	
Adults & children [(†)]	73	2.6 (0.8)	2.6 (0.8)	0.0 (-0.1, 0.2)	0.52
Adults	48	2.8 (0.5)	2.9 (0.5)	0.1 (-0.1, 0.2)	0.37
Children [(†)]	25	2.2 (1.1)	2.2 (1.1)	0.0 (-0.3, 0.3)	1.00
UK adults	48	2.8 (0.5)	2.9 (0.5)	0.1 (-0.1, 0.2)	0.37
Spain adults & children	0	-	-	-	-
Spain adults	0	-	-	-	-
Spain children	0	-	-	-	-
Bulgaria children	25	2.2 (1.1)	2.2 (1.1)	0.0 (-0.3, 0.3)	1.00
Bulgaria children [(†)]	25	2.2 (1.1)	2.2 (1.1)	0.0 (-0.3, 0.3)	1.00
Bulgaria controls	0	-	-	-	-

Table 3: Analysis of times needed to complete reading tests (original vs converted)

(*) Difference calculated as score for *Items*$_{SIMP}$ minus score for *Items*$_{ORIG}$

(†) Omitting Bulgarian controls

The results indicated no significant difference in the time taken between the two types of item for any of the groups of participants examined.

2.1.2.2 Comparison between Bulgarian ASD and control participants

The next analyses examined the difference in reading comprehension scores between ASD and control participants from Bulgaria. A summary of the analysis results is given in Table 4. The figures reported are the mean and standard deviation in each group, along with p-values indicating the significance of the results.

The results suggested significant differences between the two groups for both the comprehension and subjective scores. Scores for both types of item were significantly higher in the control group, whilst the subjective scores for both $Items_{ORIG}$ and $Items_{SIMP}$ were significantly lower in the control group.

The difference between scores for $Items_{ORIG}$ and $Items_{SIMP}$ was also examined. For both comprehension and subjective scores, the difference between the sets of scores did not differ significantly between groups.

Score	ASD Mean (SD)	Control Mean (SD)	P-value
Comprehension score - Original	6.8 (2.9)	10.7 (1.4)	**<0.001**
Comprehension score - Simplified	7.4 (2.9)	11.0 (1.0)	**<0.001**
Comprehension score - Difference (*)	0.7 (2.6)	0.3 (1.7)	0.40
Subjective score - Original	8.3 (3.2)	5.8 (2.2)	**<0.001**
Subjective score - Simplified	7.3 (3.0)	5.0 (1.4)	**<0.001**
Subjective score - Difference (*)	-0.9 (2.4)	-0.8 (2.0)	0.75

Table 4: Between-group analysis for item scores (Bulgarian children with ASD vs controls)

(*) Difference calculated as simplified score minus original score

2.1.2.3 Association between text and subjective scores

The next analyses examined the strength of association between the comprehension and subjective scores. A summary of the results is given in Table 5. The figures presented are the correlation coefficients and p-values indicating the significance of the results.

Scores	Correlation Coefficient	P-value
$Items_{ORIG}$	0.03	0.56
$Items_{SIMP}$	0.03	0.67

Table 5: Correlation between comprehension scores and subjective rating

The results suggested that there was no significant association between the comprehension scores and the subjective scores for either the scores for $Items_{ORIG}$ or the scores for $Items_{SIMP}$. This means that although participants gave more correct answers in response to questions about texts converted using OpenBook, they did not identify them as being easier to understand.

Initially the separate association between each factor and the scores for $Items_{SIMP}$ was examined separately. A summary of the analysis results is given in Table 6. Column *N* presents the number of participants in each category, and the mean and standard deviation score in each category. The regression coefficients are also presented, along with corresponding confidence intervals. This gives the mean difference in scores between each category and a baseline category. The exception is for IQ, where the scores represent the change in scores for a 10-unit increase in IQ score. P-values indicating the significance of the result are also presented.

Variable	Category	N	Mean (SD)	Coefficient (95% CI)	P-value
Age group	Adult	153	13.3 (3.3)	0	**<0.001**
	Child	90	7.8 (2.9)	-5.5 (-6.4, -4.7)	
Gender	Male	193	10.9 (4.1)	0	**0.02**
	Female	50	12.5 (4.2)	1.6 (0.3, 2.8)	
ADHD	No	204	11.4 (4.1)	0	0.26
	Yes	29	10.5 (3.7)	-0.9 (-2.5, 0.7)	
Psychiatric diagnosis	No	180	10.5 (3.8)	0	**<0.001**
	Yes	49	14.1 (3.6)	3.6 (2.4, 4.8)	
IQ $^{(*)}$	-	-	-	0.9 (0.7, 1.2)	**<0.001**
Education	None/elementary	25	7.4 (10.6)	0	**<0.001**
	Secondary	144	10.6 (3.7)	3.2 (1.7, 4.7)	
	University	65	14.0 (3.2)	6.6 (5.0, 8.2)	
	Married	59	11.1 (4.1)	0	**0.007**

Marital Status (†)	Divorced/widow	16	11.4 (4.5)	0.3 (-1.8, 2.4)	
	Single	115	12.9 (3.4)	1.8 (0.6, 3.0)	
Occupation (†)	Unemployed/retired	68	12.6 (3.7)	0	0.51
	Student	41	12.5 (2.9)	-0.1 (-1.6, 1.4)	
	Employed	76	11.9 (4.4)	-0.7 (-2.0, 0.6)	

Table 6: Regression analysis of simple main effects

(*) Regression coefficient given for a 10-unit increase in IQ

(†) Data unavailable for Spanish children

The results of the univariate analyses indicated that the majority of variables examined were associated with the scores for reading comprehension obtained for $Items_{SIMP}$. The exception was occupation and ADHD, which were not found to be significant.

Children were found to obtain significantly lower scores than adults, on average by 5.5 units. Females scored higher than males, with scores 1.6 units higher.

Participants with higher IQ values achieved higher scores for $Items_{SIMP}$. A 10-unit increase in IQ was associated with a 0.9 unit increase in reading comprehension score.

A higher level of education was also associated with higher outcome values. Those with university education achieved scores that were 6.6 units higher, on average, than those with no education or only elementary education.

There was little difference in scores achieved by married and divorced/widowed participants. However, single participants achieved the highest scores.

The second stage in the analysis process examined the joint association between the variables and the scores achieved for $Items_{SIMP}$ in a multivariate analysis. A backwards selection procedure was performed to retain only the statistically significant variables. The final model is summarised in Table 7.

Variable	Category	Coefficient (95% CI)	P-value
Age group			
	Adult	0	<0.001
	Child	-4.3 (-5.4, -3.2)	
Education			
	None/elementary	0	0.04
	Secondary	0.8 (-0.8, 2.3)	

	University	2.1 (0.2, 4.1)

Table 7: Regression analysis of combined effects

The results suggested that age and education were all significantly independently associated with the reading comprehension scores from the multivariate analysis. After adjusting for these two factors, there was no longer a significant effect of gender, psych diagnosis, IQ or marital status, all of which were significant in the univariate analyses.

As in the univariate analyses, children achieved lower scores than adults. Again, a higher level of education was associated with higher scores. The multivariate analysis indicated that the size of effects of all variables (especially autism and education) was reduced when compared to the results of the univariate analysis.

2.1.3 Conclusions of reading comprehension testing

The reading comprehension test indicated that participants performed better with MCQs based on versions of texts converted using OpenBook than on the original versions of the same texts (t=4.42, p<0.001, CI[2] [0.63, 0.79]). Participants also rated blindly converted texts as easier to understand. (t=6.96, p<0.001, CI [0.71, 1.26]).

Bulgarian children without ASD (controls) had higher comprehension scores than Bulgarian children with ASD for items based on texts in both their original and converted forms.

Univariate analysis indicated that reading comprehension scores were associated with age, gender, Psychiatric diagnosis, IQ score, and education. Regression analysis indicated that age and education were significantly associated with reading comprehension scores. Thus, older and better educated participants gave more correct answers and consequently achieved better reading comprehension.

Both adults and children participating in the tests gave more correct answers to questions posed about the converted forms of texts than to questions posed about the original forms of texts, although they took similar times to process both versions. This was consistent for UK, Spain and Bulgaria, for both adults and children. Participants also provided blind, subjective ratings, indicating that versions of texts converted using

[2] Confidence Interval.

OpenBook were easier to understand in comparison with texts in their original form. The evaluation indicates that the OpenBook software is useful in converting texts to a form that is easier to comprehend by people with ASD.

2.2 User Feedback

Individual face-to-face in-depth interviews were carried out in Bulgaria, Spain, and the UK to explore the experiences of people using OpenBook and to explore its impact on social inclusion. The Topic Guides presented in deliverable D7.4 were used to guide the interviews. Questioning was structured by the interviewer to ensure coverage of key themes but was also responsive to issues that emerged from respondents' accounts. Transcripts and interview summaries were subject to thematic analysis. Qualitative data was downloaded for analysis using the *Atlas.it* computer package.

2.2.1 Methodology

2.2.1.1 Sampling methods

Subjects formally diagnosed with high-functioning autism and basic literacy skills and/or their carers, who had already participated in the reading comprehension tests detailed above were included in the study. The clinicians in each collaborating centre in London, Madrid and Plovdiv identified suitable patients formally diagnosed with high-functioning autism from the participant's lists. Fully-informed written consent was obtained from those who were willing to participate in the study.

Purposeful sampling was used in order to collect information from a diverse population with a wide range of socio-demographic variables (Glaser & Strauss 1967). The fact that the accessibility and utility of the new technology may be influenced by socio-demographic factors such as age and sex was also taken into account. Clinicians selected diverse participants in terms of age, sex, educational attainment and employment history. Participants were required to be native Bulgarian/English/Spanish speakers and were required to be able to give written informed consent.

Sampling was continued until saturation was achieved - the point at which no new information or themes emerge from the data. Based on evidence from previous qualitative research applying purposeful sampling methods in relation to subjects with mental and behavioural disorders or their carers (Marsden et al. 2007; Lipman et al. 2010), 18 in-depth interviews were conducted: 6 interviews in the UK, 6 in Spain, and 6 in Bulgaria.

2.2.1.2 Procedures

Qualitative evaluation was led and coordinated by the UK centre (LNFT). Qualitative data was collected from each of the collaborating centres in London, Madrid and Plovdiv. Participants were given access to and started testing OpenBook 3 months before the qualitative interviews were conducted. Unfortunately, due to unforeseen delays in the development of OpenBook, participants had a much shorter period of time (down from the 6 months originally planned to 2-3 months) to test and be exposed to OpenBook.

Interviewees participated in workshops, in which they were trained by technical partners in the use of OpenBook. They were given access to the OpenBook interface and were asked to actively test the software and use it to improve the accessibility of documents.

Each participant was interviewed once. Initially participants were asked to sign a consent form, which had two main elements:

1. declaration of agreement to participate in the study and
2. declaration of consent for the interview to be audio-recorded.

Interviews were audio-recorded, with digital voice recorders, which were voice activated and dealt sensitively with ambient noise.

The questions were open and non-leading, neutral, sensitive and clear to the interviewee (Patton, 1987). Interviews lasted between 20 and 50 minutes. A naïve interviewer approach was adopted in order to encourage participants to provide comprehensive accounts. Field notes were written up after each interview to maintain a record of researcher’s impressions and to minimize researcher bias.

2.2.1.3 Ethical Issues

- Informed consent was obtained from participants, who signed a consent form, agreeing to participate in the study and to being audio taped. It was made clear that they were under no obligation to participate and that participants could withdraw from the study at any stage without any detriment to them.
- All tapes, transcripts and data were anonymised. Details that could identify participants were carefully considered and removed or altered.
- All data from the study complies with the requirements of the GCP and Data Protection Act (1998). Tapes and interview transcripts were anonymised and kept in separate locked filing cabinets. Only

the research team had access to the data. Identifiable personal data was not transferred between centres and countries.

2.2.1.4 Data collection methods

Individual in-depth interviews were conducted with patients and carers. Collectively these interviews were designed to generate information from all key perspectives. General themes were explored with specific reference to the impact of OpenBook to improved access to written information and promotion of social inclusion.

Topic guides

Interviews were based on a topic guide (D7.4) that was explored in depth. Questioning was structured by the interviewer to ensure coverage of key themes but was also responsive to issues which emerged from respondents' accounts. Initial drafting of topic guides was informed by findings of the cross-sectional survey and the suggestions of the project advisory group. They covered a range of topics including carers' perceptions of the impact of improved reading abilities on social inclusion of people with high-functioning autism, factors that facilitate and hinder successful social inclusion and impact of the new reading support tool on users' degree of social inclusion. Topic guides were refined through progressive focusing during the course of fieldwork.

Interviews were audio taped with the permission of respondents and transcribed verbatim into English by each clinical team in Spain and Bulgaria. The researchers listened to the recordings and verified the accuracy of transcriptions.

Place and Time

Geographic location has had a considerable impact on the project. Being located in three different sites with a considerable geographic distance between them meant that Bulgarian, Spanish, and UK teams had few opportunities for informal and formal face-to-face communication. Team-wide communications between meetings were mainly through skype, email and, to a lesser extent, by telephone. Most documentation was sent as attachments via email, and some effort was required to standardize to compatible software.

Transcripts of interviews were translated by each clinical team from the source language into English. They were then anonymised and sent to LNFT team. The LNFT researchers downloaded copies of transcripts (English versions) into *Atlas.it* software and performed thematic analysis.

Training

Researchers involved in qualitative research received formal training provided by Prof Mike Crawford (Department of Psychological Medicine, Imperial College London) in Alicante, 20-21st September 2012. The purpose of the workshop was to ensure that all researchers involved in WP7 took a consistent approach to qualitative interviewing across study centres. A second training session was organised and led by LNFT and delivered by Skype. Standard Operating Procedure for conduct of the interviews was produced by LNFT to aid reliability, and disseminated to all interviewing partners.

Analysis

Thematic analysis was the framework used to analyse interview transcripts. Thematic analysis is a principal technique that is used by qualitative researchers to analyse data. It is a method for identifying, analysing, and reporting patterns (themes) within data. This process may be based on prior categories, or on categories that become clear to the researcher only as the analysis proceeds. Thematic analysis differs from other analytic methods that seek to describe patterns across qualitative data – such as thematic discourse analysis, thematic decomposition analysis, IPA and grounded theory. Both IPA and grounded theory seek patterns in the data, but are theoretically bounded. As thematic analysis does not require the detailed theoretical and technological knowledge of approaches such as grounded theory, it can offer a more accessible form of analysis. Themes or patterns within data can be identified in one of two primary ways in thematic analysis: inductive or ‘bottom up’ (Frith & Gleeson, 2004), or theoretical, deductive, or ‘top down’ (Boyatzis, 1998; Hayes, 1997). A theoretical thematic analysis would tend to be driven by the researcher’s theoretical or analytic interest in the area, and is thus more explicitly analyst-driven. The choice between inductive and deductive approaches maps onto how and why the data is coded. In the case of FIRST an effort was made to explore a very specific research question (the effect of OpenBook on people with ASD and their social inclusion). Although there was a primary theoretical interest in the area which drove a deductive approach, an effort was made to use elements of inductive analysis by not trying to fit coded data into a pre-existing coding frame.

Thematic analysis is characterized by 6 stages:

1. Becoming familiar with the data and transcribing
2. Generating initial codes and data reduction
3. Searching for Themes
4. Reviewing Themes
5. Defining and naming themes
6. Reporting

These stages were followed strictly by the researchers who analysed the data. Three LNFT researchers (1 of whom is bilingual in English and Spanish) carried out independent coding. All researchers in the UK analysed the data (18 cases) separately, each developing separate coding frameworks. All transcripts were accessed in Word document files. Transcripts were transformed into PDF files which were later on uploaded into *Atlas.ti*. Constant comparative method was used: "The constant comparative method involves going through your data again and again, comparing each element – phrase, sentence or paragraph – with all of the other elements." (Thomas, 2013). Researchers made two copies of the data: The raw data and working data files.

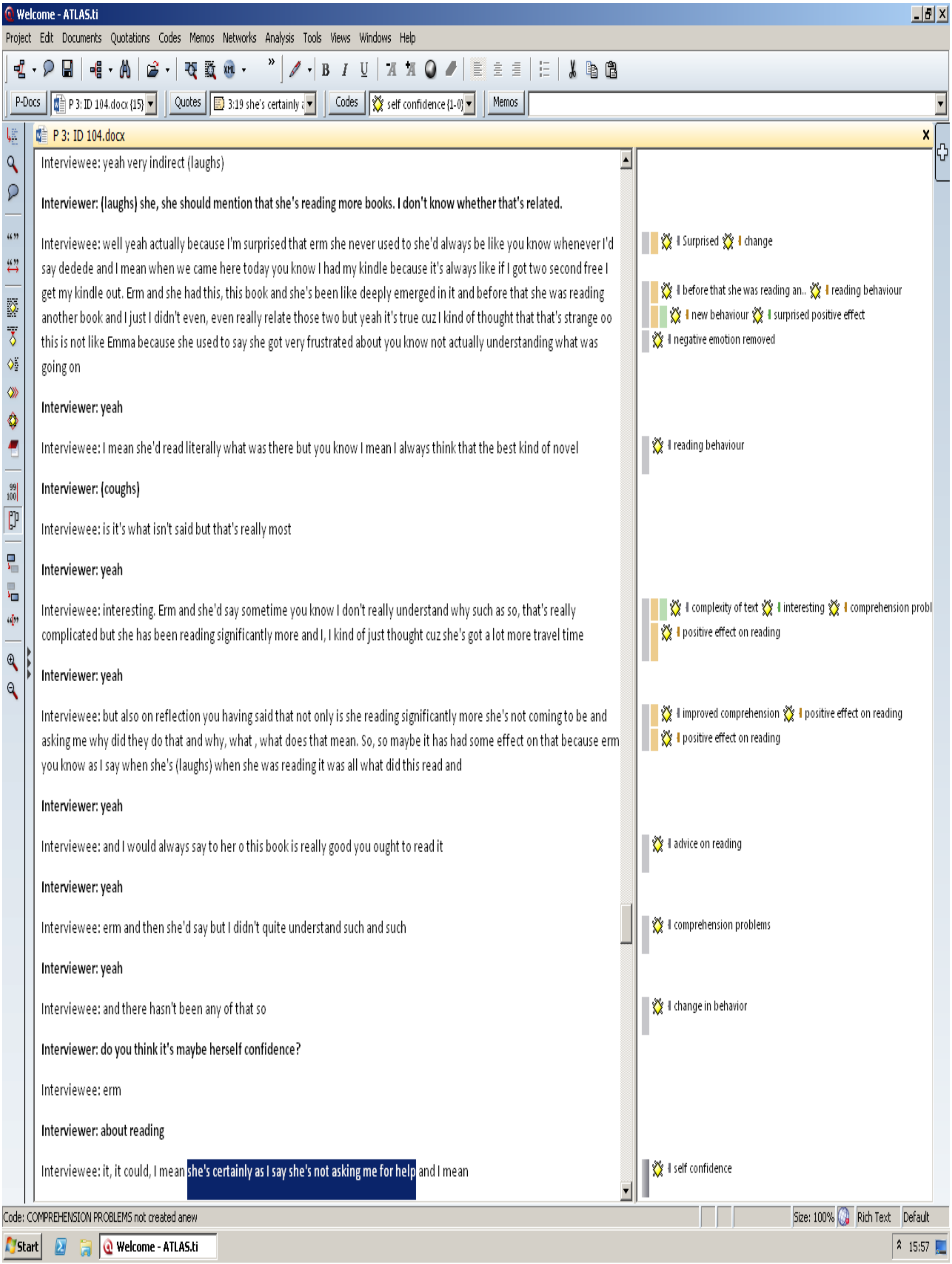
Initially the researchers achieved immersion in the data by reading transcripts of all interviews several times as they were generated. Each researcher read all of the data (transcripts) several times (at least twice) to get an impression of important ideas or subjects that recur. These were considered temporary constructs (categories). Researchers made a list of these categories.

Then the process of coding was initialised by considering one transcript at a time. Line-by-line initial descriptive open coding was carried out and each researcher developed separate coding frameworks. As the researchers were reading they underlined, circled or highlighted parts that were considered important. At the margins developing codes were added as annotations. Specific words were picked out and possible meanings listed. These meanings were validated against the rest of the text and against other researchers' codes. Codes were reviewed throughout the process. As the data was coded researchers created new codes, and went back to check the units of data previously coded to update. This was done to check whether there were any more data that could have been coded at the newly created node. Repetition of identical codes was eliminated and similar codes were combined together.

The codes were then compared between researchers and were integrated into a common coding framework, drawing together the overlap and the diverse concepts that arose during this process. Analytic induction was employed in a primary analysis whereby emerging themes were identified and incorporated into subsequent interviews. A reflexive approach was taken, continuously reviewing and refining the topic guide and coding framework to ensure that all areas that respondents had spoken about were covered. Data was summarised in relation to these emergent themes with an emphasis upon description of the utility of the new tool, its impact on improved access to written information and personal experience.

A thematic framework emerged during the course of the study. A two level coding framework was used at this stage to code all transcripts. The first level incorporated thematic descriptive codes and the second, subcategories of the first level codes and conceptual codes, which were identified through the analytical process naturally occurring in the formal coding of transcripts. In the full final analysis this thematic framework was used to code transcripts, which were re-read and indexed to indicate the presence of key themes.

The main aims of analysis interrogated this thematic framework to identify different users' experiences of their social inclusion. Under each theme we retained verbatim quotes for use in the reporting of findings.



Interviewee: yeah very indirect (laughs)

Interviewer: (laughs) she, she should mention that she's reading more books. I don't know whether that's related.

Interviewee: well yeah actually because I'm surprised that erm she never used to she'd always be like you know whenever I'd say dedede and I mean when we came here today you know I had my kindle because it's always like if I got two second free I get my kindle out. Erm and she had this, this book and she's been like deeply emerged in it and before that she was reading another book and I just I didn't even, even really relate those two but yeah it's true cuz I kind of thought that that's strange oo this is not like Emma because she used to say she got very frustrated about you know not actually understanding what was going on

Interviewer: yeah

Interviewee: I mean she'd read literally what was there but you know I mean I always think that the best kind of novel

Interviewer: (coughs)

Interviewee: is it's what isn't said but that's really most

Interviewer: yeah

Interviewee: interesting. Erm and she'd say sometime you know I don't really understand why such as so, that's really complicated but she has been reading significantly more and I, I kind of just thought cuz she's got a lot more travel time

Interviewer: yeah

Interviewee: but also on reflection you having said that not only is she reading significantly more she's not coming to be and asking me why did they do that and why, what , what does that mean. So, so maybe it has had some effect on that because erm you know as I say when she's (laughs) when she was reading it was all what did this read and

Interviewer: yeah

Interviewee: and I would always say to her o this book is really good you ought to read it

Interviewer: yeah

Interviewee: erm and then she'd say but I didn't quite understand such and such

Interviewer: yeah

Interviewee: and there hasn't been any of that so

Interviewer: do you think it's maybe herself confidence?

Interviewee: erm

Interviewer: about reading

Interviewee: it, it could, I mean she's certainly as I say she's not asking me for help and I mean

Code: COMPREHENSION PROBLEMS not created anew

Figure 1: Sample from coding process in Atlas.ti

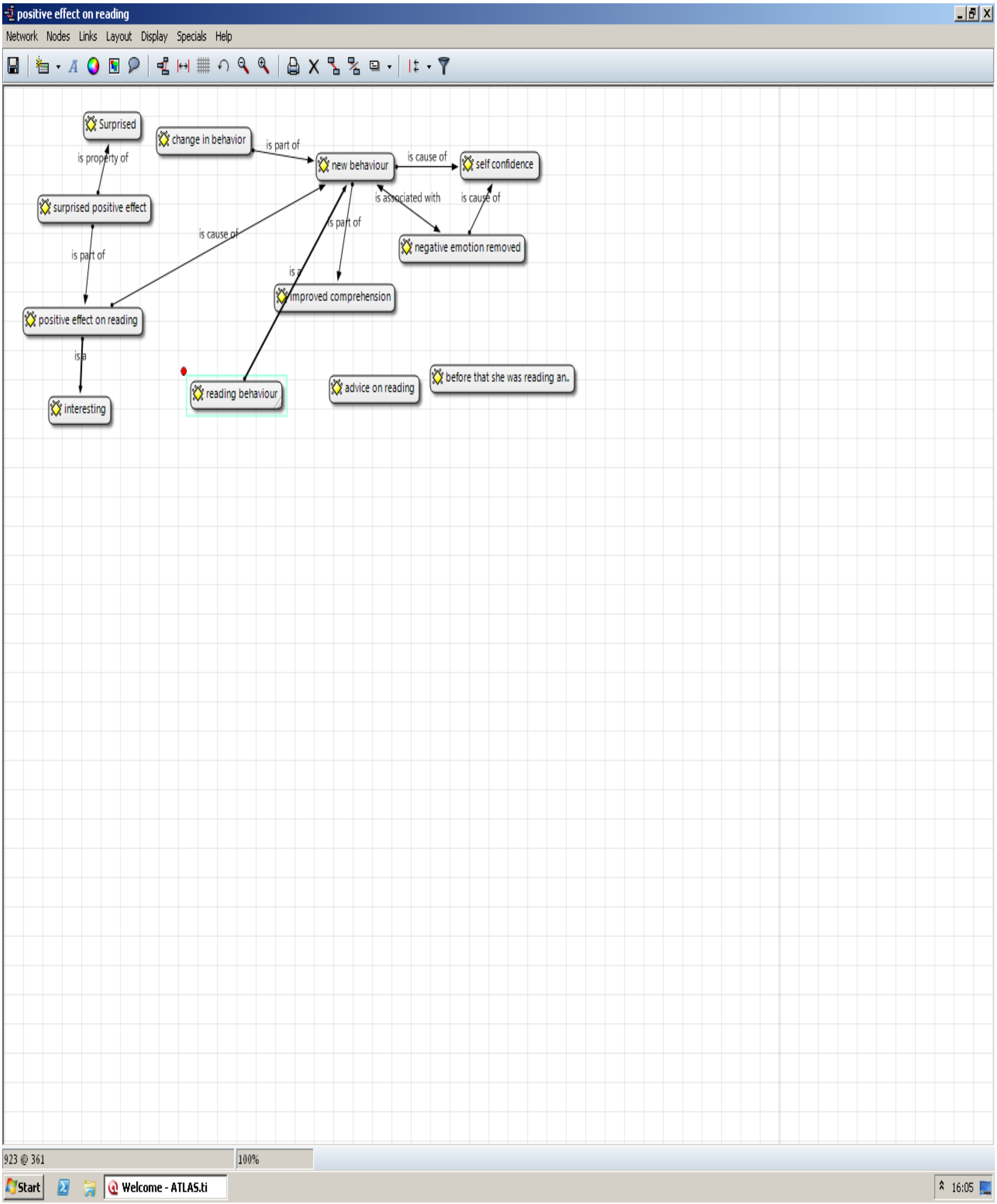


Figure 2 Example of initial hierarchical networks in Atlas.ti

'Negative cases' such as 'challenges of using OpenBook' were identified and reported in order to capture the complexity of the phenomena and add depth to the data.

Triangulation of data sources was achieved through communication by skype and emails with colleagues in Spain and Bulgaria. Efforts were made to ensure that researchers were fully aware of developing codes; that ideas were not lost in translation and that valid inferences were made during data analysis. Researchers in Bulgaria and Spain agreed with all the codes and themes and there were no contentions, mistranslations or misunderstandings to report.

2.2.1.5 Participants

Six adults with ASD were interviewed in the UK and Spain, 1 (17 year old) child was interviewed in Spain and 11 carers were interviewed in all three centres in total (Table 8).

Characteristics		**Children with ASD**	**Adults with ASD**	**Carers**
N	UK		4 (1 female, 3 male)	2 (female)
	Spain	1(male)	2 (male)	3 (female)
	Bulgaria			6(5 female, 1 male)
Age		14 (12-18)[3]	33 (25-43)	42 (37-50)
Relationship				5 mothers (1UK, 1 Spain, 3 Bulgaria) 1 wife (UK) 2 teachers (Bulgaria) 2 SLT[4] (1 Spain, 1Bulgaria) 1 psychologist (Spain)

Table 8: Interviewee Characteristics

Testers discussed extensively about the usability of OpenBook including its pros and cons. OpenBook was not finalised at the stage of participants' testing, which happened during the development phase. Therefore several issues reported during the interviews have been resolved post-interview. Nevertheless, all issues

[3] This includes the age of the children that were not interviewed directly, but, whom were cared by the interviewees. The child that was interviewed in Spain was 17 years old.

[4] Speech and Language Therapist

raised are reported in Section 2.2.2 and Section 2.2.3 to accurately reflect tester's perceptions at the time of interview.

2.2.2 User Feedback on OpenBook (End Users and Carers)

2.2.2.1 Positive aspects of OpenBook

Almost everyone agreed that the concept of OpenBook is strong and the tool has great potential (Table 9). Neli in Bulgaria said ***"It (OpenBook) has been quite easy to access and the things follow naturally, I mean, it's simple to work with. Even people who haven't got good computer skills and knowledge can use it on their own, without any need of assistance. It has got enough explanations; the buttons are very large and people can find them easily. It's just the beginning, but I think it's very important because one can easily find what they look for in a program..."*** Testers liked the way that OpenBook is set up and said that it allows fast access to information. Thus, John (adult with ASD) stated ***"Sometimes these things get over-engineered...it does look quite simple; and it's this one page as well, most of it is done on this one page where you don't have to keep going into sub-menus and stuff like that, and that's its strong point, I think."*** Testers liked different features of OpenBook such as insert image, definitions, add note etc. They reported that personalised settings for fonts and themes were very useful.

A user in Spain stated that summaries can also be quite useful "***...for example for making summaries easier in tough school subjects, in socials (Note: Social Science), biology and let's see... in grammar and, in the case of literature and... then also, let's see... and in other subjects, except for English and maths.***"

Many users have to wait for family members to simplify information for them, especially when they receive letters from government departments. It was also reported that the waiting time until they see a family member to explain such letters can be exasperating. Use of Open book in such situations can lower anxiety and increase independence. John stated "***No one wants to dread the postman coming, because you don't know what's coming through the door. Whereas if you got something coming through the door...which is council tax, housing benefits all these complicated forms, and so we can just slap them onto the scanner, find out exactly what it's really about***."

2.2.2.2 Challenges of using OpenBook

Given that testing happened alongside the development phase, a few issues were reported when using OpenBook (Table 10: OpenBook challenges

Positive aspects of OpenBook	Frequency of statements
helpful and easy to use	8
insert image and definition feature	7
the concept of open book is very good	6
the way open book is set up	3
it's good for self-studies, help children with difficulties	3
the personalized options such as colour font etc.	3
the idea of 'explain' words function	2
the separate carer and user aspect	2
OpenBook allows fast access to info	2
the help of word definitions when carer doesn't know	2
it saves time and effort	2
it is accurate (with summary for example)	1
the idea of 'add note' function	1

Table 9: Positive aspects of OpenBook

Several testers explained that images were not always accurate, thus Jim stated "***So ...one of the pictures, erm... that was not helpful - I looked up snorkelling, just the word snorkelling. And some of the pictures showed a snorkeller, but some of the other pictures, showed somebody just, like, with the goggles: they had a snorkel, but they were underwater so if it was, like, you know, if I didn't really know what snorkelling was...***" A carer discussed inappropriate images that could cause further confusion in people with ASD ***"… she (a person with ASD) looked up the phrase 'It's raining cats and dogs'.......and one of the pictures was literally, er… cats and dogs falling from the sky with umbrellas or something***".

Challenges when using OpenBook	Frequency of statements
problems with definitions: explanation not right	17
pictures not being accurate	9
slow processing of text	8
incorrect synonyms	5
simplified text not being changed significantly from the original	4
having to re-highlight to be able to use different functions	4
label button not working	2
phrases not being recognised	2
the 'explain’ function not working	2
summaries not accurate	2
highlighting making it hard to read	2
text disappearing	2
messages not being sent from career to user	2
difficulties with font size	2
colours not working	1
function of 'add note' not working properly	1
URL doesn’t work in loading files	1
‘help’ tab not working	1
difficulty logging in open book	1

Table 10: OpenBook challenges

2.2.2.3 Progress of OpenBook

Ivan in Bulgaria explained that “***Some words have many meanings but the program gives only one definition. Sometimes this definition doesn’t correspond with the meaning of the word in a certain text.***” Ivan went on to suggest “***I would add more synonyms and definitions. I would change some pictures with more appropriate ones. I would delete the marking of some common words as hard.***” Several users suggested that making OpenBook available as a mobile application would be very useful for them. They recommended improvement of image selection and dictionaries (to enable provision of more comprehensible synonyms), the introduction of tutorials and voice recognition functions, as well as functions for the optical scanning of printed text (further details in Table 11).

Suggested improvements for OpenBook	Frequency of statements
make it available as a phone app (an app in general)	4
improve images	5
integrate OpenBook in school and social services	3
have dictionary for foreign languages	3
use children dictionary	2
tutorial should be made on how to use OpenBook	2
user page should be more like career one	2
OpenBook should include notes	2
voice recognition	1
photographing text and getting instant translation	1
make dictionary more context-specific	1
have a audio facility (read the text to you)	1
improve highlighting function	1
careers should be able to input pictures or definitions	1
labels should show notes	1
changes should be saved automatically	1

Table 11: Suggested improvements for OpenBook

Users reported that they had seen changes and improvements to OpenBook since the start of testing. Thus, John (adult with ASD) explained ***“so I tried it again and it seems a little better ...we can do more at home.”*** Mark also stated ***“the copy- paste one - that works faster...cos initially there was a problem with the servers. Erm... so then I got more adventurous.... the copy and paste thing seems to work quite well.”*** Maria in Bulgaria stated that OpenBook reduces the time and effort required to simplify information and therefore reduces the burden on carers.

Overall, all interviewees said that they will continue to use OpenBook and would suggest its use to others as well. Interviewees said that they would suggest OpenBook to adults and children with ASD as well as teachers and other government departments that work with people with ASD. For example Galina in Bulgaria stressed ***"... I would recommend it (OpenBook) because it will help them (teachers) and their children. The children will become more independent, will understand things which are difficult to explain, and if there is a way to implement it in schools it will be great. That way we won't have to tell the teachers where the obstacles can be and the teachers will do just fine."*** EC in Spain also agreed that: ***"I would recommend it to some of my mates that are stressed out with studies."***

Neli in Bulgaria stated "***I'm telling lots of people about the program. I think this program is very suitable for children with autism. Although I don't have much experience, in my opinion it should be integrated in schools and other social services where disabled children are taken in, I mean places which obtain documents for disabled people- it' suitable for these places***."

2.2.2.4 Technical analysis

Technical teams analysed the interview transcripts presented in deliverable D7.7 with regard to their relevance to LT services developed in the FIRST project. The expressed views relevant to each one are summarised in the remainder of this Section. In each case, the points are listed in order of the frequency with which carers and end users drew attention to them in the interviews.

OpenBook overall (144 comments overall)

1. End users and carers would recommend the use of OpenBook to others (6% of comments)
2. OpenBook should be improved to reduce the amount of effort required of carers converting texts (5% of comments)
3. Further improvements are needed to address problems with OpenBook (4% of comments)
4. OpenBook is being used by end users independently of their carers (3% of comments)
5. Carers and end users would like to continue using OpenBook in the future (3% of comments) and since starting to use OpenBook, end users have become more independent and keen to solve problems for themselves (3% of comments).

Examination of the set of comments made about the system reveals that there is a demand to apply OpenBook over a wide range of texts, including official documents such as medical correspondence, applications for benefits, council tax and utilities bills, work contracts, as well as news, science, and academic texts.

WP3 - Syntactic processor (6 comments overall)

1. The syntactic processor only works well on restricted types of texts (BBC news reports) (17% of comments on this function)
2. Sentence rewriting is inaccurate (17% of comments on this function)
3. Rewritten sentences should be formatted as bullet points (17% of comments)

4. Repetition in syntactically processed texts helps reading comprehension (17% of comments)
5. Rewritten sentences should be highlighted in the interface as it is currently hard for users to detect them (17%). Additionally, the syntactic processor fails to process types of complexity involving unusual word order (such as passive sentences or fronted sentences)

WP4 - Definitions/explanations of complex terms (51 comments overall)

1. Explanations of words are frequently incorrect or inappropriate, given the context of use (16% of comments about this function)
2. Definitions should be available for a greater proportion of words in input texts (14% of comments about this function)
3. End users successfully used this function, independent of their carers (14% of comments about this function)
4. The ability to retrieve definitions of complex terms reduces the burden on the carer (14% of comments about this function)
5. The function is very useful (6%) and retrieving definitions of terms constitutes the main way in which the system is used (6%)

WP4: Synonym retrieval (13 comments overall)

1. The ability to retrieve definitions of complex terms reduces the burden on the carer to define words for the end user (23% of comments about this function)
2. Retrieved synonyms are sometimes incorrect/inappropriate, given the context of use (23% of comments about this function)
3. Synonyms should be available for a greater proportion of words in input texts (15% of comments about this function)
4. Provision of synonyms is useful for both end users and carers (15% of comments about this function)
5. Retrieval of synonyms of words to improve understanding of word meanings is the main function of OpenBook used by carers and end users (8% of comments about this function). Users noted that OpenBook works well for making complex words more accessible (8%) but not complex phrases (8%).

WP5: Image retrieval (30 comments overall)

1. Retrieved images are sometimes incorrect/inappropriate, given the context of use (23% of comments about this function)
2. The ability to retrieve images to explain complex terms reduces the burden on the carer to define words for the end user (17% of comments on this function)
3. End users successfully use this function of OpenBook independently on their carers to assist reading comprehension (13% of comments on this function)
4. Images should be available for a greater proportion of words in input texts (10% of comments about this function)
5. Carers found the quality of image retrieval to be good (7% of comments about this function) and the function was seen to improve the attention of end users when reading (3% of comments about this function)

WP5: Summarisation (12 comments overall)

1. Users and carers consider the summarisation function to be useful (33% of comments about this function)
2. Summaries should be formatted as sequences of bullet points (8% of comments about this function)

3. OpenBook should be changed so that it is possible to save summaries (8%)
4. End users should have access to the function to generate and insert summaries into their documents (8%)
5. It should be possible for carers and end users to edit summaries, insert them into the document, and save them. (8%) It was also felt that summaries should be derived from the converted forms of texts, not their original form and that summaries may have a role to play in helping end users to write without including repetition and redundancy.

WP6: Messaging between users and carers (3 comments overall)

1. The messaging function seems not to work very well (100% of comments on this function)

WP6: Interface (50 comments overall)

1. The interface is easy to use for carers and end users (16% of comments)
2. End users would like a version of the interface that works well on tablets and mobile phones (8% of comments)
3. End users should have access to several of the functions currently reserved for carers (8% of comments)
4. Obstacle detection functions help both end users and carers (4% of comments)
5. The ability to build up a library of converted documents is useful (4% of comments) as is the ability to adjust the font and background colours of the texts (4%)

2.2.2.5 Conclusions

Clinicians in Bulgaria, Spain, and the UK interviewed 18 users of OpenBook, both people with ASD and carers. The interviewees had used OpenBook for a period of 2-3 months. Interviews were driven by topic guides developed early in the study and lasted between 20-50 minutes. Purposeful sampling was used in order to ensure a representation of people from diverse backgrounds. LNFT led the qualitative analysis and qualified staff in Bulgaria and Spain were trained accordingly.

Interviews were tape-recorded in all centres in the source language (English, Spanish, and Bulgarian). All teams transcribed verbatim all interviews and anonymised them. The Spanish and Bulgarian teams translated transcripts into English and sent anonymised transcripts to the UK team. All transcripts were uploaded into *Atlas.ti*, which was used to organise codes and themes. Thematic analysis was the framework used to analyse interview transcripts, whereby researchers simultaneously read and re-read interviews, generated initial codes and themes, checked those codes and themes between each other and reviewed them. Codes and themes were shared with the Spanish and Bulgarian colleagues in order to achieve triangulation and verify findings. Final themes were then agreed upon and reported.

Even though the users involved in the interviews experienced problems, the overall experience of using OpenBook was reported to be a positive one. Testers agreed that the concept of OpenBook is excellent and most of its functions are very useful. Some of the challenges such as long processing times and idiom identification have now been resolved. Other issues such as image and synonym selections have been improved, but still perform worse than users would like due to the limitations of LT methods.

Generally, users said that the system is simple to use and not only improves the studies of children with ASD and relieves anxiety of adults with ASD regarding text comprehension but it also relieves the burden of teachers and carers. Users said that they will continue to use OpenBook in the future and will recommend it to friends and colleagues. Many of the findings made by the FIRST technical team when analysing the interviews will serve as promising directions for future work in the use of LT as a tool for converting texts into a more accessible form.

2.2.3 OpenBook and Inclusion

As detailed in Section 2.2.1.5, 18 users of OpenBook were interviewed face to face and in-depth by clinical teams in the UK, Spain and Bulgaria. 11 carers of people with autism were interviewed in all three centres in total; six adults with ASD were interviewed in UK and Spain and 1 adolescent (17-year old) was interviewed in Spain. Carers were predominantly female (10/11) and participants with ASD were predominantly male (6/7). Carers were of a relatively older age (mean: 42 years old) compared to adults with ASD (mean: 33 years old). Carers were a mix of family members (mothers, spouses) and professional caregivers (teachers and therapists).

Four researchers in the UK carried out independent coding by analysing the 18 interviews separately, each developing separate coding frameworks. The codes were then compared and integrated into a common coding framework, drawing together the overlap and the diverse concepts that arose out of this process. Data was summarised in relation to emergent themes with an emphasis upon description of the utility of the new tool, its impact on improved access to written information, and personal experience.

Adult participants stated in the interviews that they used OpenBook to simplify Wikipedia articles, work contracts, and news items. Children used it mainly to carry out homework and Google different learning topics.

Self-reported frequency of use of OpenBook varied from *rarely* to *more than once per day*.

Interviewees explained that OpenBook is particularly useful for reading books and manuals, to make summaries for homework and to help with filling out important documents (such as council tax documents). Another positive function of OpenBook was explained to be that it gives carers hints on what users need help with, by highlighting possible obstacles to reading comprehension.

2.2.3.1 Findings of the analysis

Eight themes emerged from data analysis relating to the impact of OpenBook on participants' reading, writing, comprehension, communication, emotions, relationships, self-confidence and independence.

Improved reading comprehension as a result of using OpenBook was widely reported (Table 12, C). The positive impact in comprehending written texts was described as "***obvious***" and "***encouraging***" by one user: "***For me it was kind of like black and white, completely clear ....it was very encouraging***". Other users reported improved comprehension when accessing complex information such as reading about formulas and mathematical curves (C2) but also to understand subtext (C5), which is very important especially when reading fiction literature.

Improved comprehension seemed to have an impact on participants' reading skills. Thus, improved reading abilities were reported by both adults with ASD and their carers (Table 12, A). Adults with ASD reported that they were reading more as a result of using OpenBook (A1, A3) and focusing better on reading (A2). A mother in London also reported that her daughter's vocabulary increased considerably as a direct result of using OpenBook (A1).

Surprisingly, both adults and carers of children reported a positive impact of OpenBook on their writing skills. They seemed to be more confident in writing emails and notes (Table 12, B). Some participants started to write notes for the first time (B2 & B4), while others felt more confident (B1) and consequently became more active writers (B3).

Improved communication was another theme that emerged from the interview analysis, and was consistent for both adults and children (Table 12, D). A speech therapist in Bulgaria stated that the student who used OpenBook started to initiate contact with other children, which he never did before (D1). Adults with ASD, on the other hand, reported enhancement of spoken language particularly using richer vocabulary and complex phrases and sentences (D3 and D4).

Testing the tool (OpenBook) during the beginning of the production phase was described as frustrating for all users (Table 12, D3, D4, D6). At initial phases, text took a considerably long time to upload and convert into a simplified form. Other issues were: inappropriate image suggestions, the suggestion of synonyms that were more complex than original word, inappropriate synonym suggestions (e.g. ‘Cinderella went to the ball’: the synonyms and images for the word ‘ball’ were bouncy balls). However, most of these issues were resolved through the development phase, and users reported improvement in processing times and synonym suggestions. Improved comprehension and communication were evident, factors which seemed to have a positive impact on the emotions of both children and adults with ASD (Table 12, E). A mother stated that her son’s behaviour had improved and her son was not getting angry when he did not understand what his mother was saying (E1). A teacher also explained that a student (Alex) seemed to overcome his shyness and become more sociable.

Carers talked appreciatively about the way OpenBook made users more self-sufficient and consequently self-confident (Table 12, E4, and F) in the way that they became able to look independently for information and communicate with others (F1, F4, F6 and F8). Both adult users and children became tenacious in using OpenBook although they faced challenges (F2). Children were also reported to study longer and more effectively (F5). However, OpenBook did not work well for everyone. Martin, for example, an adult with autism found it difficult to deal with the obstacles highlighted by the software. His mother explained that he was made uncomfortable by the red font and felt he had done something wrong (F7). John, on the other hand, an adult with ASD talked about his experience with OpenBook in a very positive light “*Because you're self-reliant... your confidence grows. No one wants to dread the postman coming, because you don't know what's coming through the door...*’’

Relationships between people with ASD and carers were also a focal point of interview statements. Children were reported to engage more frequently with their peers and carers and the quality of these interactions was consistently reported to have been enhanced (Table 12, G, B2, B3, D1 and E6). A teacher in Bulgaria stated that she expected her workload to be alleviated as a result of the student engaging independently with OpenBook (G5). Adults also reported changes in relationships. Lisa, for example, explained that she started leaving notes for her work colleagues to facilitate her workload. Jim (an adult with ASD) talked about being less reliant on others: "*...But it's kind of you're not so reliant on other people so it makes more of a difference to me than other people, because they don't mind doing it.*’’ Lisa also spoke about being less reliant on others (G4), and her mother confirmed this: ‘*’Not only is she reading significantly more, she's not coming to me asking me, 'why did they do that’ and 'what does that mean*'” (C1 and C6). Concerns were raised, however, about the accessibility of OpenBook especially from paid carers such as social workers: *''My social worker.... he doesn't really use computers. I don't think he'd really be that into it... He doesn't use email. Not sure if some social worker will use the carer part of OpenBook.* "

The increased independence of users of OpenBook was a recurring theme throughout the interviews. Although some concern was expressed about the possibility of adults with ASD becoming more dependent on their carers (H6), overall, statements described increased independence of both adults and children with ASD when:

- accessing OpenBook and reading texts (H1, H3, H4)
- accessing technology (H2)
- reading paper books (H9)
- writing notes and emails (B3, B4)

A. Reading	1. ‘’Lisa’s increase in vocabulary and reading more... probably are directly as a result of this (OpenBook)....’ *(mother of adult with ASD)* 2. ‘’I find it easier to focus on reading.’’ *(adult with ASD)* 3. ‘’ I've been reading more since I started using it (OpenBook).’’ *(adult with ASD)*
B. Writing	1. "I think that Jake is more confident in the written communication and expression" *(teacher of child with ASD)* 2. "Ben narrates better. He started to write notes to his teachers and classmates to express himself" *(teacher of child with ASD)* 3. "Jake is more active in written communication with his friends- through emails and notes.’’ *(teacher of child with ASD)* 4. ‘’I don't know whether this is (due to) OpenBook or just me but I, what I've recently started doing is that ...when I'm leaving work for the last day of the week I now write little notes to whoever is in that day or the day after. .... Just to tell them what needs stocking up... little notes of encouragement and stuff like that.” *(adult with ASD)*

C. Comprehension	1. ‘’Not only is she reading significantly more, she's not coming to me asking me, 'why did they do that’ and 'what does that mean'.’’ *(mother of adult with ASD)* 2. ‘’ There was a Google doodle the other day for ... Maria Gaetana .....She was a philosopher from Italy. ...she came up with all these formulas and mathematical curves. I found OpenBook made it easier to appreciate and understand what she has done.’’ *(adult with ASD)* 3. " For me it was, it was kind of like black and white, completely clear.... for me it was very encouraging" *(adult with ASD)* 4. " (OpenBook) is making me think a little and not get bogged down in having to read and get the jest and tend to select more key phrases from a paragraph. It's woke me up to that....’’ *(adult with ASD)* 5. ‘’.... she does know what it means now, because she's not asking me; and I think she still would have if she was having difficulty. So, erm... she may be more confident: it may be just that she’s...getting better since using this (OpenBook) in understanding subtext.” *(mother of adult with ASD)*
D. Communication	1. " I can see that Andrew starts to communicate with other children. Andrew goes to them and starts talking, and he never did that before." *(Speech Therapist of child with ASD)* 2. " Ben overcame his uneasiness in communicating with others" *(teacher of child with ASD*) 3. ‘’ (OpenBook) it has made my language become a little bit more formal.’’ *(adult with ASD)*

	4. "I think it (OpenBook) increased my vocabulary, I think it (OpenBook made me use more complex phrases and sentences. " *(adult with ASD)*
E. Emotions	1. " Yes, he doesn't get angry when I can't give him the answer or when he doesn't understand what I am saying" *(mother of child with ASD)* 2. " It's just very frustrating" *(wife of adult with ASD)* 3. "it makes me want to scream" *(mother of adult with ASD)* 4. ‘’It's great, ‘cuz, I mean, anything that makes her feel better about herself is fantastic.’’ *(mother of adult with ASD)* 5. 'it’s beyond frustrating that something isn't working" *(adult with ASD)* 6. " Alex’s communication has improved. He was shy before but now he is starting to overcome that.” *(teacher of child with ASD)*
F. Self-efficacy @ Confidence	1. " George looks for things which he's personally interested in, things he's curious about" *(mother of child with ASD)* 2. "Ben was trying to correct the text because he is seeing it harder. He wanted to do it alone. He didn't give up and that is the most important thing. *(teacher of child with ASD)* *3.* ‘’Because you're self-reliant... your confidence grows. No one wants to dread postman coming, because you don't know what's coming through the door....’’ *(adult with ASD)* 4. "Tim even wants to log in the program by himself" *(mother of child with ASD)*

	5. "Ben is more motivated. When using the program he agrees to study longer- reading asking and answering questions" *(teacher of child with ASD)* *6.* "When something actually gets in his way Alex doesn't ask me, he just looks for it in the program" *(teacher of child with ASD)* *7.* "Martin is kinda made uncomfortable. He found it disconcerting all that red (obstacles are written in red font in the text on OpenBook) - it was like... 'What have I done wrong?!" *(mother of adult with ASD)* *8.* ‘’If Lisa has got a document she doesn't understand and she puts it through OpenBook, and she understands it, then that, that's a big reward; and it's a big incentive to do it again and again.’’ *(mother of adult with ASD)*
G. Relationships	1. "The change I saw was that the involvement of John in a group task. Before he refused to work like that.*" (teacher of child with ASD)* 2. " My social worker.... he doesn't really use computers. I don't think he'd really be that into it... He doesn't use email. Not sure if social worker will use carer part of OpenBook. " *(adult with ASD)* 3. " But it's kind of you're not so reliant on other people so it makes more of a difference to me than other people, because they don't mind doing it.’’ *(adult with ASD)* 4. ‘’Erm... I don't bother so many people, I suppose, I will be less reliant on other people.’’ *(adult with ASD)*

	5. " I even believe that at a certain stage it will help me, it will reduce my engagement" (*teacher of child with ASD*)
H. Independence	1. " It's fun for him to look for the word and its meaning, on his own" *(mother of child with ASD)* 2. " Jake can now work with tablet and computer. He feels more independent" (*teacher of child with ASD*) 3. " It's fun for Alex to look for the word and its meaning, on his own" (*teacher of child with ASD*) 4. "He is more independent.... the program (OpenBook) itself has changed him....to do everything by himself without my help at all." (*teacher of child with ASD*) 5. " Yes, yesterday was the biggest success because he showed independence with his behaviour. Without wanting any help, although he had difficulties with the text, he didn't ask for my help, he wanted to do it by himself." (*teacher of child with ASD)* 6. "..that's actually reducing their individual independence rather than increasing it, if ... you know mum has to sort everything out" (*mother of adult with ASD*) 7. " I think I have needed a little less help" (*adult with ASD)* 8. ‘’Main autonomy bit is that, you know Lisa seems to be reading fiction now without actually asking for help. So, I mean, that is an increase in independence and I've never thought about it,

	but actually is." (*mother of adult with ASD*)

Table 12: Themes and quotations from the one to one interviews

Every participant stated that they will recommend OpenBook strongly to others and that they would like to continue using OpenBook in the future.

2.2.3.2 Conclusions on the analysis of OpenBook and Inclusion

Clinical teams in the UK, Spain and Bulgaria conducted face to face interviews with 18 adults and children with ASD. The interviews were tape recorded and transcribed in the original language. Transcripts were translated into English and forwarded to the team in UK. Analysis was carried out in the UK, but codes were triangulated through all clinical teams.

The main themes that emerged from data analysis were: reading, writing, comprehension, communication, emotions, relationships, self-confidence and independence. Users reported challenges when using OpenBook during the development phase. They often experienced frustration when working with texts that took a very long time to process, inaccurate retrieval of images and retrieval of incorrect synonyms. However, most of these issues have now been resolved: input from testers during the critical development phase has been essential.

OpenBook was reported to have had a positive impact on reading comprehension of both adults and children with ASD. Improved comprehension seemed to improve not just the reading skills of people with ASD but also their writing and communication abilities. Although some users found it disconcerting to use OpenBook, most users stated that it had a positive effect on their relationships, self-confidence and ultimately their independence.

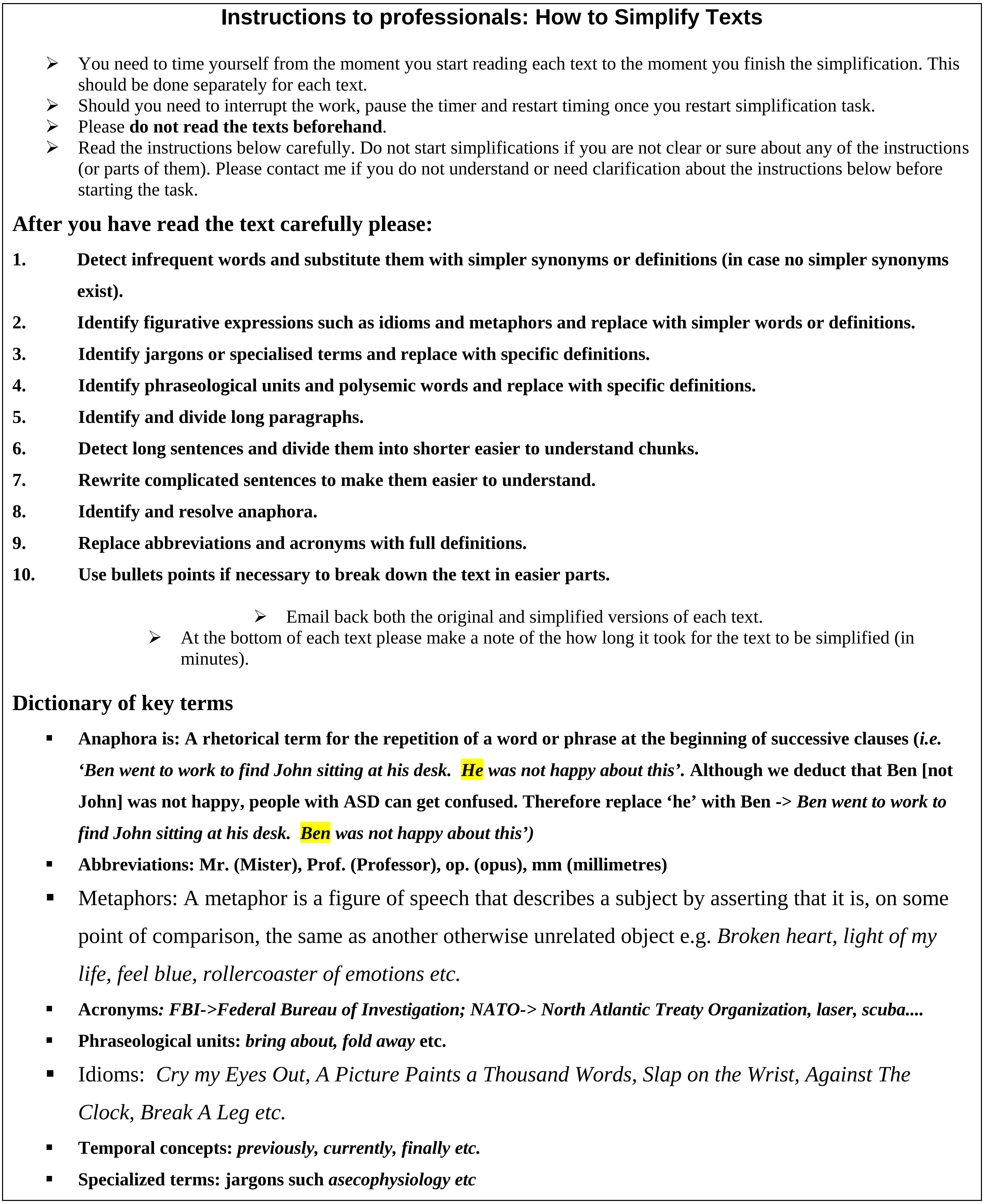
3. Intrinsic Evaluation

In the FIRST project, intrinsic evaluation is a study of the efficacy of the text conversion services supported by the OpenBook software. It assesses the quality of the semi-automatic conversion process in terms of its ability to improve the readability of converted texts and the extent to which OpenBook reduces the burden on carers converting texts into a more accessible form for end users. Intrinsic evaluation includes an empirical assessment of the accuracy of the LT supporting the text conversion services.

3.1 Evaluation of Text Conversion using OpenBook: Readability Assessment

Deliverable D7.2 presents a benchmark for unaided text conversion, detailing the time taken for professional carers to convert 25 heterogeneous texts in Bulgarian, English, and Spanish into a more accessible form and detailing differences in the readability of those texts before and after conversion. The task was repeated for a set of 25 science texts in each of the three languages. The second task was semi-automatic in nature: carers used the interface to OpenBook when converting the texts, and were able to exploit LT functions to assist in the process.

To ensure validity, the same professional carers involved in the benchmarking process detailed in D7.2 were recruited for the second task. They were given the same instructions as for the unaided conversion phase (Figure 3). All recruited professionals were sent five different science texts, identified locally by random allocation. They were asked to complete the work and return the original texts, the converted versions and their timing figures to the UK centre where analysis was conducted. UK professionals worked with an adult audience in mind, while the Bulgarian colleagues converted texts for the needs of children. Professionals in Spain converted texts for both children and adults.



Instructions to professionals: How to Simplify Texts

- You need to time yourself from the moment you start reading each text to the moment you finish the simplification. This should be done separately for each text.
- Should you need to interrupt the work, pause the timer and restart timing once you restart simplification task.
- Please **do not read the texts beforehand**.
- Read the instructions below carefully. Do not start simplifications if you are not clear or sure about any of the instructions (or parts of them). Please contact me if you do not understand or need clarification about the instructions below before starting the task.

After you have read the text carefully please:

1. **Detect infrequent words and substitute them with simpler synonyms or definitions (in case no simpler synonyms exist).**
2. **Identify figurative expressions such as idioms and metaphors and replace with simpler words or definitions.**
3. **Identify jargons or specialised terms and replace with specific definitions.**
4. **Identify phraseological units and polysemic words and replace with specific definitions.**
5. **Identify and divide long paragraphs.**
6. **Detect long sentences and divide them into shorter easier to understand chunks.**
7. **Rewrite complicated sentences to make them easier to understand.**
8. **Identify and resolve anaphora.**
9. **Replace abbreviations and acronyms with full definitions.**
10. **Use bullets points if necessary to break down the text in easier parts.**

- Email back both the original and simplified versions of each text.
- At the bottom of each text please make a note of the how long it took for the text to be simplified (in minutes).

Dictionary of key terms

- **Anaphora is: A rhetorical term for the repetition of a word or phrase at the beginning of successive clauses (*i.e.* *'Ben went to work to find John sitting at his desk. He was not happy about this'*. Although we deduct that Ben [not John] was not happy, people with ASD can get confused. Therefore replace 'he' with Ben -> *Ben went to work to find John sitting at his desk. Ben was not happy about this'*)**
- **Abbreviations: Mr. (Mister), Prof. (Professor), op. (opus), mm (millimetres)**
- Metaphors: A metaphor is a figure of speech that describes a subject by asserting that it is, on some point of comparison, the same as another otherwise unrelated object e.g. *Broken heart, light of my life, feel blue, rollercoaster of emotions etc.*
- **Acronyms*: FBI->Federal Bureau of Investigation; NATO-> North Atlantic Treaty Organization, laser, scuba....***
- **Phraseological units: *bring about, fold away* etc.**
- Idioms: *Cry my Eyes Out, A Picture Paints a Thousand Words, Slap on the Wrist, Against The Clock, Break A Leg etc.*
- **Temporal concepts: *previously, currently, finally etc.***
- **Specialized terms: jargons such *asecophysiology etc***

Figure 3: Instructions given to professionals regarding text conversions

Since the focus of OpenBook changed from general texts to science texts, only science texts were included in the second conversion task. Science texts used in the manual benchmarking task were exploited in the second task, but care was taken to ensure that they were converted by a different professional than was the case for benchmarking. Clinical teams identified new science texts with a word length of between 250 and 350 words.

Researchers were mindful to identify texts containing examples of obstacles to reading comprehension such as: phraseological units, polysemic words, anaphors, abbreviations, acronyms, temporal concepts, and specialized terms. A library of selected science texts was sent to technical partners. Technical teams analysed obstacles in the texts and matched new texts with non-scientific texts that had been used in the manual benchmarking task. These non-scientific texts were replaced with scientific texts matched for word length and prevalence of obstacles to reading comprehension. The similarity between pairs of scientific and non-scientific texts in terms of readability was computed by measuring cosine similarity between vectors consisting of readability index scores. Pairs of texts with greatest similarity were substituted. Table 13 and **Error! Reference source not found.** display the set of substitutions of Bulgarian and English texts for which the global level of similarity was greatest, together with the cosine similarity of each pair.

Substitution pair		**Cosine similarity**
Anna Karenina	Longitude	0.956
Tom Sawyer	MusicalGlass	0.956
Cannabis	GiantTubeWorms	0.93
Movie paradoxes	ClimateChange	0.989
James Joyce	Superbugs	0.988
Fall North	Voltammeter	0.967
Forrest wedding	Cancer	0.977
Rock Sound Mag	SpaceCentre	0.991
Melt in the body electronics	Mars	0.983
Lady Gaga	Skeleton	0.984
To Kill A Mockingbird	CharlesDarwin	0.973
Syrian News	NorthernLights	0.981
Pulp Fiction	MysteriousIllness	0.994
Leeds vs. Wigan	VirungaPark	0.988
Casual Vacancy	AlzheimersResearch	0.936
µ		0.973

Table 13: Text substitutions (EN)

Substitution pair		Cosine similarity
bg.Text9Original.txt	TheBGPolicyMidXX.txt	0.847076998
bg.Text1Original.txt	DevelopmentGeography.txt	0.872263165
bg.Text16Original.txt	WatersAfrica.txt	0.924704566
bg.Text24Original.txt	RespSystem.txt	0.903697441
bg.Text17Original.txt	MyForefathHome.txt	0.905683575
bg.Text2Original.txt	Patriotism_nat.nihilism.txt	0.957435408
bg.Text4Original.txt	NationalLeaders.txt	0.921681007
bg.Text10Original.txt	AfricasClimate.txt	0.964868919
bg.Text3Original.txt	Minerals.txt	0.900724175
bg.Text23Original.txt	Nat_Identity.txt	0.921827496
bg.Text22Original.txt	TheAtlanticOcean.txt	0.900396656
bg.Text19Original.txt	Taiga.txt	0.892393621
bg.Text6Original.txt	BulgariaWWII.txt	0.859927749
bg.Text7Original.txt	Blood_vessels.txt	0.952297375
bg.Text25Original.txt	SavannasSemidesertsDeserts.txt	0.898380735
bg.Text8Original.txt	Climate_ZonesAfrica.txt	0.900918485
bg.Text20Original.txt	LaysAndMinerals.txt	0.969090364
bg.Text5Original.txt	Mount_nature_area.txt	0.975552242
bg.Text21Original.txt	DevelopmentAgriculture.txt	0.929061845
bg.Text18Original.txt	TheBulgarianBeginXX.txt	0.919845579
μ		0.91589137

Table 14: Text substitutions (BG)

3.1.1 Analysis of times taken to convert texts

The mean time taken to convert texts containing 250-350 words decreased significantly from 54 minutes for unaided conversion to 29 minutes for conversion using OpenBook (Table 16). The length of time required increased only for conversion of English texts, but this increase is not statistically significant. Discussion with the carers involved in the benchmarking experiment revealed that the scientific articles were found by some carers to be more difficult to understand than the texts used in the benchmarking task described in Deliverable D7.2. In addition, when using OpenBook, they spent time inserting images to further improve the accessibility of the text. This operation was not performed in the original benchmarking task because participants considered it too tedious to perform without access to a tool such as OpenBook. During unaided conversion, huge disparities in conversion rates between centres were noted. These disparities diminished when Open book was used and the speed of processing was similar for all professionals based at the same centre (Figure 4). Speed of processing increased consistently in all centres; however, the improvement was most significant (more than threefold) in Bulgaria (Figure 5).

Participating Centres	Mean times taken to simplify (in minutes)	
	Unaided conversion	Conversion using OpenBook
UK (adults)	23	*27*
Spain (adults & children)	43	*27*
Bulgaria (children)	97	*35*
Overall means	**54.33**	**29.48**

Table 15: Outcomes of the text conversion processes

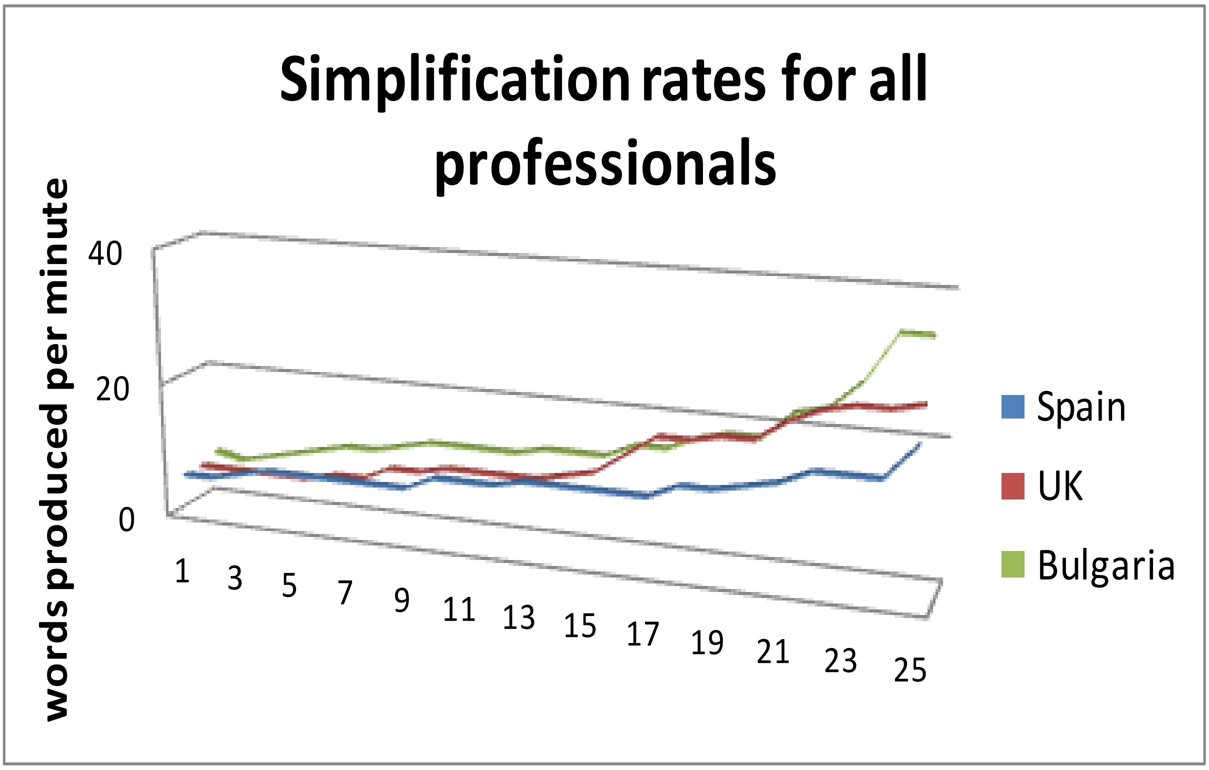


Figure 4: Speed of conversion using OpenBook per text for all centres

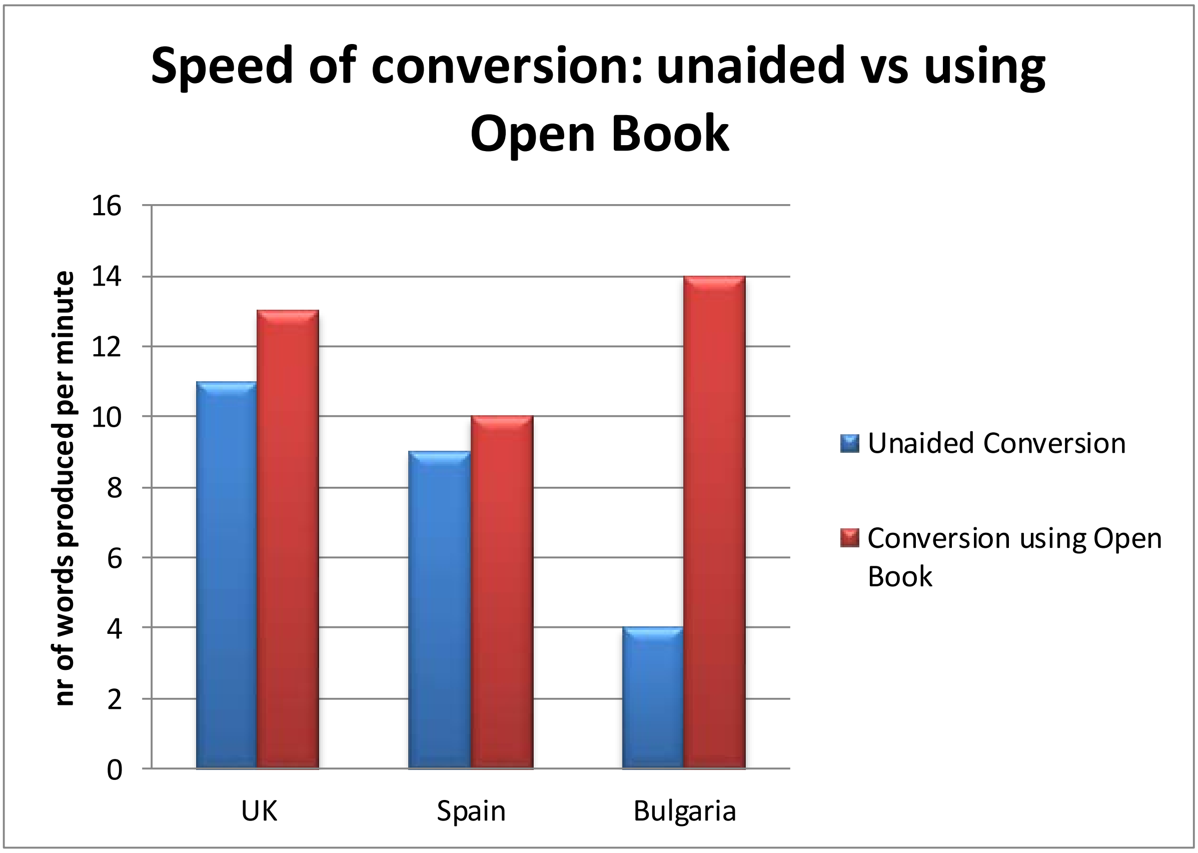


Figure 5: Speed of unaided conversion vs. speed of conversion exploiting OpenBook

3.1.2 Readability

Table 16 and Table 17 (Section 3.1.2.1), Table 18 and Table 19 (Section 3.1.2.2), and Table 21 and Table 22 (Section 3.1.2.3) display readability statistics for BG, EN, and ES texts in their original form and in a form produced by the semi-automatic conversion process supported by the OpenBook system. The choice and method of derivation of these statistics was reported in Section 5.2 of D7.2. In the six tables, underlining of column headings indicates that there is a statistically significant difference in the values of the metric for the original and converted forms of the texts (α=0.01).

3.1.2.1 Bulgarian

Metric	Comma index		Index of words with > 3 syllables		Index of words in sentences		Index of word diversity		Pronoun index	
File	**ORIG**	**SIMP**	**ORIG**	**SIMP**	**ORIG**	**SIMP**	**ORIG**	**SIMP**	**ORIG**	**SIMP**
AfricasClimate.txt	0.36	0.30	3.68	3.65	15.71	15.46	0.52	0.53	4.32	5.31
Blood_vessels.txt	0.51	0.54	3.05	2.97	17.65	17.74	0.53	0.59	7.68	8.82
BulgariaWWII.txt	0.43	0.42	3.53	3.32	12.82	14.84	0.53	0.50	3.72	5.09
Climate_ZonesAfrica.txt	0.25	0.14	2.75	3.17	13.44	11.95	0.52	0.53	5.15	5.20
CommunSkills.txt	0.59	0.14	3.26	3.17	17.89	11.95	0.55	0.53	14.91	5.20
DevelopmentAgriculture.txt	0.43	0.51	3.51	3.81	16.06	15.14	0.53	0.55	6.05	6.04
DevelopmentGeography.txt	0.66	0.53	3.36	3.39	16.06	15.00	0.54	0.49	11.74	10.06
Habitat.txt	0.66	0.71	3.36	3.50	16.06	21.73	0.54	0.56	11.74	11.04
LaysAndMinerals.txt	0.50	0.37	3.37	3.63	15.51	13.21	0.48	0.52	7.37	7.38
Minerals.txt	0.61	0.62	2.98	2.90	19.13	18.06	0.45	0.48	4.77	4.89
Mount_nature_area.txt	0.63	0.61	3.52	3.42	22.24	25.33	0.56	0.57	5.03	6.58
MyForefathHome.txt	0.62	0.45	2.89	2.61	24.96	17.00	0.54	0.56	8.68	9.88
Nat_Identity.txt	0.59	0.60	3.35	3.00	17.47	17.42	0.54	0.54	10.33	13.25
NationalLeaders.txt	0.39	0.42	3.62	3.43	15.89	17.07	0.52	0.57	8.16	9.00
Patriotism_nat.nihilism.txt	0.58	0.47	3.86	3.93	22.81	13.64	0.58	0.59	10.14	10.67
RespSystem.txt	0.50	0.48	3.06	3.28	22.04	20.08	0.51	0.52	7.62	7.47
SameSimTriangles.txt	0.60	0.48	3.23	3.28	16.73	20.08	0.35	0.52	2.79	7.47
SavannasSemidesertsDeserts.txt	0.90	0.99	3.13	2.93	12.76	18.15	0.49	0.47	4.94	3.90
SelfImage.txt	0.90	0.44	3.13	3.09	12.76	14.41	0.49	0.56	4.94	16.40
Taiga.txt	0.88	0.85	2.91	3.01	16.35	13.64	0.50	0.50	4.96	4.32
TheAtlanticOcean.txt	0.41	0.55	3.37	3.31	15.78	22.06	0.51	0.49	4.45	6.23
TheBGPolicyMidXX.txt	0.39	0.29	3.90	3.78	14.09	13.58	0.54	0.52	4.78	6.06
TheBulgarianBeginXX.txt	0.58	0.55	3.61	3.74	14.60	13.63	0.54	0.55	4.42	5.32
TheText_PublicCommun.txt	0.49	0.55	4.02	3.74	13.30	13.63	0.53	0.55	7.84	5.32
WatersAfrica.txt	0.34	0.48	2.97	2.83	14.09	16.34	0.49	0.49	6.67	6.50
µ	0.55	0.50	3.34	3.32	16.65	16.45	0.52	0.53	6.93	7.50

Table 16: Readability of original (ORIG) and semi-automatically converted (SIMP) forms of BG science texts

Inspection of Table 16 reveals that, while statistically insignificant, differences in values of the *index of word diversity* are more frequently smaller for the original versions of Bulgarian texts than those of their converted versions. This finding is unexpected as it implies that the converted texts contain a larger number of different words than their converted counterparts. This is explained by the fact that carers of children with ASD (target group in Bulgaria, age 12-18) decided, in many cases, to retain the original term and to add definitions, explanations, images, and their own additions as a way of teaching students new words. Parents and professionals are seen to use language that they find suitable for their particular students, weighing their individual levels of knowledge and needs. The finding that the converted versions of texts contain a wider range of vocabulary than their original versions is a result of the addition of new explanatory text for specialised and ambiguous terms. It was noted that, rather than substituting inaccessible terms by more accessible ones, these terms were retained and the text was supplemented with explanatory text.[5] Though not statistically significant, the original versions of the text also contain smaller values of the *pronoun index* than is the case for the converted versions. This implies that the converted texts contain a higher number of pronouns per word, which can be a source of ambiguity in meaning. As noted in Section 5.4 of D7.2, the number of anaphoric pronouns used in a converted text can increase as editors seek to avoid verbatim repetition when providing explanatory text about specialised or ambiguous terms and when rewriting syntactically complex sentences. Differences in the values of the remaining indices were not statistically significant, though in those cases the values indicate that the converted versions of the texts are more accessible than the originals.

One observation that may provide some explanation for statistically insignificant differences between the two versions of each text is that the converted versions tend to be shorter than the original versions (by an average of 20 words). Many of the readability indices are computed as ratios with the number of word tokens in the text as the denominator. When applied to shorter texts, they are therefore evaluated to higher scores.

Inspection of Table 17 reveals that the manually converted versions of Bulgarian texts contain statistically significant differences in the values of the *metaphor index* and the *polysemic word index*. With respect to the former, the differences are in line with expectation. However, values of the *polysemic word index* are unexpectedly larger in converted versions of the texts than in their original versions. As noted in Section 5.4 of D7.2, statistics on the frequency of occurrence of semantically ambiguous words may make little contribution to assessing the readability of a text. Unfortunately, resources enabling measurement of more relevant phenomena are currently insufficient in the state of the art. There is no statistically significant difference in the values of other indices derived from the different versions of the texts. In this experiment, it was observed that values of the *passive verb index* are unexpectedly larger for the converted forms of the majority of the BG texts than for the original versions.

[5] In English, images were also inserted into the converted forms of the texts to assist comprehension.

Metric	**Metaphor Index**		**Passive verb index**		**Polysemic word index**[6]		**Verb temperature**	
File	**ORIG**	**SIMP**	**ORIG**	**SIMP**	**ORIG**	**SIMP**	**ORIG**	**SIMP**
AfricasClimate.txt	0.00	0.00	0.00	0.00	59.43	60.58	**0.10**	**0.10**
Blood_vessels.txt	0.00	0.00	**0.03**	**0.04**	**78.45**	**79.25**	**0.11**	**0.12**
BulgariaWWII.txt	0.00	0.00	**0.02**	**0.04**	**69.12**	**68.45**	**0.12**	**0.12**
Climate_ZonesAfrica.txt	0.00	0.00	0.10	0.03	**55.05**	**54.55**	**0.09**	**0.09**
CommunSkills.txt	1.00	0.00	0.06	0.03	**74.47**	**78.82**	0.14	0.09
DevelopmentAgriculture.txt	0.00	0.00	0.00	0.00	62.05	67.09	0.09	0.09
DevelopmentGeography.txt	2.00	0.00	**0.03**	**0.03**	**72.38**	**72.56**	0.14	0.12
Habitat.txt	0.00	0.00	**0.03**	**0.07**	**60.71**	**67.86**	0.14	0.13
LaysAndMinerals.txt	0.00	0.00	0.14	0.13	56.64	58.12	**0.13**	**0.13**
Minerals.txt	0.00	0.00	0.09	0.06	**56.38**	**61.19**	**0.09**	**0.10**
Mount_nature_area.txt	0.00	0.00	**0.06**	**0.07**	**54.55**	**56.36**	**0.08**	**0.10**
MyForefathHome.txt	1.00	0.00	**0.04**	**0.08**	66.67	67.26	**0.12**	**0.14**
Nat_Identity.txt	3.00	0.00	**0.03**	**0.04**	**71.93**	**72.09**	**0.12**	**0.14**
NationalLeaders.txt	1.00	0.00	**0.00**	**0.04**	**65.28**	**78.00**	**0.11**	**0.13**
Patriotism_nat.nihilism.txt	1.00	0.00	0.00	0.00	**64.22**	**78.00**	**0.10**	**0.12**
RespSystem.txt	0.00	0.00	0.00	0.00	**68.46**	**70.34**	0.10	0.10
SameSimTriangles.txt	0.00	0.00	0.00	0.00	**72.73**	**73.33**	0.10	0.10
SavannasSemidesertsDeserts.txt	0.00	0.00	**0.07**	**0.20**	**44.51**	**44.37**	0.11	0.10
SelfImage.txt	2.00	1.00	0.07	0.09	79.00	74.16	**0.11**	**0.14**
Taiga.txt	0.00	0.00	0.11	0.05	48.30	48.44	0.10	0.10
TheAtlanticOcean.txt	0.00	0.00	0.00	0.00	**60.00**	**66.29**	0.08	0.08
TheBGPolicyMidXX.txt	0.00	0.00	**0.04**	**0.04**	67.46	68.83	0.11	0.11
TheBulgarianBeginXX.txt	0.00	0.00	0.04	0.03	55.90	55.70	0.09	0.09
TheText_PublicCommun.txt	1.00	0.00	**0.00**	**0.03**	**70.64**	**77.11**	0.13	0.09
WatersAfrica.txt	0.00	0.00	**0.03**	**0.03**	**55.36**	**60.00**	0.10	0.09
μ	0.48	0.04	0.04	0.04	63.59	66.35	0.11	0.11

Table 17: Readability of original (ORIG) and semi-automatically converted (SIMP) forms of BG science texts

[6] The converted versions of the 25 texts have significantly larger values of this index than the original versions. This is unexpected as scores derived from more accessible texts were assumed to have smaller values, indicative of smaller proportions of ambiguous words. It may be that ambiguous words were preserved and unambiguous words omitted during the reduction in text length that occurred in simplification.

3.1.2.2 English

METRIC	Comma index		Index of words with > 3 syllables		Index of words in sentences		Index of word diversity		Pronoun index	
File	ORIG	SIMP	ORIG	SIMP	ORIG	SIMP	ORIG	SIMP	ORIG	SIMP
AlzheimersResearch.txt	0.50	0.55	1.68	1.75	16.47	21.47	0.63	0.61	3.93	1.92
Blood booster	0.44	0.10	0.55	0.54	24.91	11.25	0.26	0.24	3.28	3.81
Cancer.txt	0.38	0.41	1.29	0.60	11.00	16.07	0.54	0.45	5.02	4.15
CharlesDarwin.txt	0.50	0.31	1.83	1.29	18.65	19.00	0.63	0.56	5.05	3.35
Climate change (Conversion 1: D7.2)	0.43	0.12	2.62	1.37	15.50	12.05	0.27	0.24	1.08	3.32
ClimateChange.txt (Conversion 2)	0.47	0.19	1.33	1.34	19.93	20.79	0.53	0.51	0.72	0.51
DNA profiling	0.63	0.21	1.54	1.35	17.81	10.67	0.30	0.28	3.86	4.17
GiantTubeWorms.txt	0.75	0.33	1.58	1.26	25.77	15.15	0.61	0.53	3.28	3.88
IPhone 5 (Conversion 2)	0.39	0.14	1.02	0.41	16.00	12.82	0.29	0.20	3.62	5.05
Longitude.txt	0.37	0.26	1.77	1.07	24.92	21.28	0.60	0.56	2.01	2.35
Mars.txt	0.18	0.16	0.76	0.77	25.23	20.18	0.51	0.48	3.35	2.70
Melt in the Body Electronics	0.18	0.00	1.35	1.06	17.00	9.40	0.31	0.27	2.94	3.19
MusicalGlass.txt	0.06	0.06	0.53	0.61	16.95	19.05	0.48	0.48	7.76	5.80
MysteriousIllness.txt	0.50	0.32	1.04	1.13	21.20	15.86	0.61	0.53	2.20	1.35
NorthernLights.txt	0.33	0.22	0.91	0.72	19.41	15.14	0.56	0.57	5.15	4.09
PCs Features	0.35	0.30	1.09	1.42	11.18	9.91	0.26	0.25	5.43	4.90
PeriodicTable.txt	0.61	0.37	2.07	2.12	22.69	39.23	0.51	0.48	1.36	0.59
Photosynthesis.txt	1.00	0.54	2.00	1.85	22.31	50.58	0.53	0.44	0.34	0.82
Skeleton.txt	0.69	0.59	1.94	1.90	25.33	21.06	0.62	0.63	1.97	1.68
SpaceCentre.txt	0.80	0.41	1.93	1.66	25.85	36.88	0.60	0.48	2.68	1.69
StemCells.txt	0.74	0.33	1.86	1.17	17.00	18.88	0.51	0.45	0.93	1.32
StringTheory.txt	0.47	0.33	2.89	2.39	27.70	23.46	0.54	0.52	1.44	1.97
Superbugs.txt	0.19	0.09	2.12	1.88	15.55	19.59	0.63	0.53	4.50	3.94
VirungaPark.txt	0.47	0.33	1.15	1.18	14.10	19.11	0.60	0.58	3.04	1.93
Voltammeter.txt	0.63	0.32	2.14	2.18	19.38	20.89	0.63	0.50	7.94	3.72
µ	0.48	0.28	1.56	1.32	19.67	19.99	0.50	0.45	3.32	2.89

Table 18: Readability of original (ORIG) and semi-automatically converted (SIMP) forms of EN science texts

Inspection of Table 18 reveals statistically significant differences between the original and converted versions of the English science texts in the values of the *Comma index*, the *index of words with > 3 syllables*, and the *index of word diversity*. This indicates that the converted forms of the texts contain less complex sentences (fewer commas per sentence), shorter words, and more restricted vocabulary than the original versions. Although the converted forms of the texts do not always contain a smaller number of pronouns (*pronoun index*) or shorter sentences (*index of words per sentence*) than the originals, the findings are broadly in line with expectation.

METRIC	Metaphor index		Passive verb index		Polysemic word index		Index of productive syntax	
File	**ORIG**	**SIMP**	**ORIG**	**SIMP**	**ORIG**	**SIMP**	**ORIG**	**SIMP**
AlzheimersResearch.txt	82.35	47.06	5.88	0.00	49.64	48.22	74.00	32.00
Blood booster	50.00	25.00	54.55	28.57	53.28	49.52	74.00	76.00
Cancer.txt	3.45	3.33	27.59	30.00	52.66	46.89	77.00	74.00
CharlesDarwin.txt	29.41	13.64	23.53	0.00	45.43	47.85	73.00	47.00
Climate change (Conversion 1: D7.2)	16.67	0.00	16.67	15.00	57.35	55.19	70.00	66.00
ClimateChange.txt (Conversion 2)	14.29	10.53	7.14	10.53	53.41	51.90	68.00	67.00
DNA profiling	21.88	14.81	43.75	51.85	44.56	48.26	83.00	82.00
GiantTubeWorms.txt	7.69	2.94	23.08	2.94	42.69	45.83	74.00	56.00
IPhone 5 (Conversion 2)	41.67	21.43	27.78	25.00	49.83	50.49	80.00	79.00
Longitude.txt	50.00	5.56	25.00	0.00	46.15	43.08	75.00	31.00
Mars.txt	7.69	4.55	53.85	31.82	46.04	45.72	71.00	75.00
Melt in the Body Electronics	2.78	3.57	44.44	17.86	49.05	47.59	77.00	73.00
MusicalGlass.txt	36.84	26.32	5.26	0.00	45.96	45.86	82.00	26.00
MysteriousIllness.txt	13.33	10.71	33.33	14.29	46.54	42.57	79.00	73.00
NorthernLights.txt	29.41	19.05	23.53	19.05	46.97	49.06	74.00	74.00
PCs Features	10.71	8.11	10.71	8.11	50.80	51.77	74.00	74.00
PeriodicTable.txt	7.69	0.00	76.92	76.92	44.41	45.29	71.00	74.00
Photosynthesis.txt	15.38	16.67	100.00	0.00	46.55	43.99	68.00	35.00
Skeleton.txt	58.33	29.41	25.00	0.00	47.70	41.90	78.00	61.00
SpaceCentre.txt	69.23	62.50	15.38	12.50	52.38	44.75	71.00	63.00
StemCells.txt	10.53	4.17	42.11	12.50	49.85	52.32	64.00	41.00
StringTheory.txt	10.00	0.00	60.00	7.69	55.60	55.74	72.00	50.00
Superbugs.txt	15.00	13.64	20.00	9.09	45.98	42.23	84.00	52.00
VirungaPark.txt	28.57	15.79	23.81	0.00	39.86	40.22	75.00	16.00
Voltammeter.txt	23.08	11.11	15.38	22.22	45.63	44.15	76.00	73.00
M	26.24	14.80	32.19	15.84	48.33	47.21	74.56	58.80

Table 19: Readability of original (ORIG) and semi-automatically converted (SIMP) forms of EN texts used in benchmarking experiments

Inspection of Table 19 reveals that there is a statistically significant difference in the values of the *metaphor index*, the *passive verb index*, and the *index of productive syntax* (IPS) for the two versions of the English texts. Converted versions contain fewer examples of lexical and non-lexical metaphors (indicators of ambiguity in meaning) and passive verbs per sentence (an indicator of structural complexity). These versions also contain a more restricted range of the language features exploited when computing the index of productive syntax (i.e. they are characteristic of less highly developed/mature language).

3.1.2.3 Spanish

	Sents	Words/Text	Char/Text	Words/Sent	Char/Word	SC Sign Usage
Original (μ)	8.52	267.52	1437.04	34.89	5.39	19.13
Simplified (μ)	13.2	306	1638	31.18	5.36	18.77
t-test	0.04	0.169	0.153	0.471	0.868	0.387

Table 20: Readability statistics of ES science texts

Table 20 presents statistics for readability criteria of the original and semi-automatically converted versions of Spanish science texts. The first row contains the mean values of each characteristic calculated for the original forms of the documents. The second row contains this information for the documents converted semi-automatically using OpenBook. The final row displays 2-tailed statistical significance of the paired t-test. The columns present different characteristics of the documents. These are the number of sentences in the document (Sents), the number of words in the document (Words), the number of characters (letters) in the document (Chars), the average sentence length (Words/Sent), the average word length (Chars/Word), and the percentage of tokens in the document that are signs of syntactic complexity[7] (SC Sign Usage). For α=0.001 no significant differences were found between the original texts and those converted by carers using OpenBook.

Table 21 and Table 22 display the values of readability indices calculated for ES texts.[8] It can be observed that values of the *metaphor index* and the *index of words in sentences* for semi-automatically converted (SIMP) versions of the texts are significantly smaller than those of the original (ORIG) versions.

[7] Punctuation, conjunctions, complementisers and wh-words.
[8] In Table 21 and Table 22, the statistics are presented as decimals rather than percentages.

METRIC	Comma index		Index of words with > 3 syllables		Index of words in sentences		Index of word diversity		Pronoun index	
File	ORIG	SIMP	ORIG	SIMP	ORIG	SIMP	ORIG	SIMP	ORIG	SIMP
Descubren como convertir la luz en materia	0.47	0.06	3.09	3.21	53.33	11.41	0.45	0.23	2.81	3.63
Indicios de un possible entorno habitable en Marte	0.37	0.38	2.85	2.72	30.00	30.43	0.46	0.48	3.33	3.76
Los Dinosaurios	0.73	0.69	2.77	2.81	41.29	76.71	0.55	0.30	3.81	3.35
Diez alimentos que combaten el cáncer	0.55	0.05	4.00	2.83	31.43	11.81	0.54	0.20	5.00	3.44
El oro negro	1.00	0.89	3.55	3.41	23.25	20.77	0.52	0.52	5.38	4.81
Enigmas de la ciencia	0.79	0.78	2.82	3.10	26.33	25.50	0.41	0.42	3.80	5.10
Mosca Asesina	0.51	0.38	3.36	3.24	28.44	21.64	0.54	0.53	4.30	4.20
El Universo	0.76	0.64	3.47	3.42	13.94	13.69	0.57	0.55	4.78	3.65
Propiedades de los cítricos	0.38	0.40	3.42	3.47	33.25	37.75	0.50	0.27	5.26	3.09
El Uso de plantas medicinales	0.66	0.68	3.25	3.27	41.88	24.18	0.49	0.48	8.66	9.02
Las Aves	0.54	0.42	3.07	2.76	25.25	24.00	0.60	0.55	100.00	7.29
Física cuántica	1.14	0.57	3.58	3.63	33.88	26.17	0.52	0.59	5.54	5.10
La Antártida	0.76	0.64	3.00	3.30	50.00	21.80	0.42	0.48	4.80	9.17
Gorro pilotar aviones	0.45	0.49	3.33	3.47	30.38	31.67	0.53	0.53	6.17	5.26
Aloe Vera	0.78	0.60	3.42	3.33	33.30	28.64	0.56	0.57	4.20	2.86
Lluvias Torrenciales	0.53	0.38	3.33	3.18	27.00	16.23	0.55	0.58	4.53	2.84
Cancez y Alzheimer	0.69	0.16	3.57	3.08	58.20	13.74	0.50	0.19	3.78	3.99
Fibrosis Quística	0.54	0.56	3.91	3.72	29.70	39.00	0.52	0.49	4.71	4.36
Colapso Urgencias	0.75	0.13	3.13	3.69	26.67	11.85	0.55	0.24	3.75	1.25
Trabajar con disolventes puede provocar deterioro cognitivo	0.41	0.28	3.52	3.42	24.40	40.57	0.47	0.34	2.46	3.87
Las TIC	0.78	0.81	3.80	3.57	61.25	113.83	0.60	0.23	6.53	6.88
Hallan un fósil	0.57	0.37	3.08	3.58	32.43	14.89	0.53	0.54	3.96	4.48
La Mosca de laFruta	0.63	0.49	3.25	3.16	28.00	53.20	0.53	0.49	7.14	6.02
La luna y el sol	1.20	1.15	3.10	3.31	15.78	26.09	0.04	0.51	1.06	5.92
El Fantasma de la obesidad acosa a la población infantil	0.47	0.51	3.69	3.30	23.18	19.70	0.56	0.55	5.88	6.09
µ	0.66	0.50	3.33	3.28	32.90	30.21	0.50	0.43	8.23	4.78

Table 21: Readability of original (ORIG) and semi-automatically converted (SIMP) forms of ES science texts

Examination of Table 21 reveals that versions of the ES science texts converted using OpenBook have, in 68% of cases, more restricted vocabulary than their original versions, implying that they are easier to read. This is one indication of the contribution made by OpenBook in the conversion process.

Examination of Table 22 indicates that the conversion process exploiting OpenBook is effective in reducing the number of phraseological units and non-lexicalised metaphors in the text (indicated by the *metaphor index*), which is in accordance with the user requirements reported on in deliverable D2.2. The results reported there also indicate that the texts converted by carers using OpenBook contain fewer passive verbs, which are noted to be potential obstacles to reading comprehension.

The *pronoun index* reveals that the 32% of the semi-automatically converted Spanish science texts contain more pronouns per word than the originals. In the benchmarking experiment reported on in deliverable D7.2, it was noted that 56% of texts produced as a result of unaided conversion contained a larger number of pronouns per word than the originals. The use of OpenBook leads to a reduction in the number of pronouns introduced during the conversion process. As noted earlier, carers tend to introduce pronouns in order to avoid verbatim (redundant) repetition of phrases during the conversion process.

Values of the *polysemic word index* also show that the use of OpenBook in the text conversion process improves the readability of the converted text. In the original benchmark, exploiting unaided conversion, 68% of the converted versions of Spanish texts had larger values for the *polysemic word index* than the original versions. In the context of text conversion using OpenBook, the percentage of converted texts with a larger value for the *polysemic word index* is smaller (48% of the documents).This is further evidence of the contribution made by OpenBook to the conversion process.

3.1.2.4 Comparing changes in readability due to unaided conversion and conversion using OpenBook

The current deliverable and deliverable D7.2 include tables presenting readability scores for Bulgarian, English, and Spanish texts. These results are derived from the original and converted forms of 25 texts in each language. In D7.2, conversion was unaided: performed by intermediaries with no access to OpenBook. In this deliverable, conversion was performed by intermediaries using OpenBook. A small proportion of the texts were converted using both methods and were analysed in both D7.2 and in the current deliverable (5 Bulgarian texts, 6 English texts, and 3 Spanish texts).

METRIC	Metaphor Index		Passive verb index		Polysemic word index	
File	ORIG	SIMP	ORIG	SIMP	ORIG	SIMP
Descubren como convertir la luz en materia	0.00	0.00	0.17	0.29	0.37	0.34
Indicios de un possible entorno habitable en Marte	0.00	0.00	0.22	0.14	0.34	0.34
Los Dinosaurios	0.14	0.00	0.86	0.71	0.30	0.32
Diez alimentos que combaten el cáncer	0.14	0.00	0.43	0.13	0.37	0.41
El oro negro	0.25	0.00	0.50	0.46	0.42	0.43
Enigmas de la ciencia	0.00	0.00	0.25	0.30	0.39	0.45
Mosca Asesina	0.00	0.00	0.22	0.27	0.36	0.36
El Universo	0.00	0.00	0.11	0.06	0.44	0.44
Propiedades de los cítricos	0.13	0.00	0.25	0.50	0.35	0.39
El Uso de plantas medicinales	0.38	0.00	0.38	0.27	0.41	0.42
Las Aves	0.00	0.00	1.75	0.38	0.01	0.40
Física cuántica	0.13	0.00	0.25	0.17	0.38	0.41
La Antártida	0.00	0.00	0.80	1.00	0.32	0.28
Gorro pilotar aviones	0.00	0.00	0.38	0.33	0.42	0.44
Aloe Vera	0.10	0.00	0.20	0.18	0.38	0.40
Lluvias Torrenciales	0.22	0.00	0.67	0.23	0.30	0.27
Cancez y Alzheimer	0.00	0.00	0.80	0.10	0.33	0.35
Fibrosis Quística	0.00	0.00	0.20	0.30	0.37	0.35
Colapso Urgencias	0.00	0.00	0.33	0.00	0.30	0.42
Trabajar con disolventes puede provocar deterioro cognitivo	0.00	0.00	0.10	0.71	0.37	0.35
Las TIC	1.25	0.00	0.75	2.17	0.44	0.44
Hallan un fósil	0.00	0.00	0.14	0.22	0.31	0.38
La Mosca de laFruta	0.00	0.00	0.78	1.00	0.33	0.35
La luna y el sol	0.00	0.00	5.61	0.27	0.00	0.37
El Fantasma de la obesidad acosa a la población infantil	0.00	0.00	0.55	0.30	0.37	0.40
µ	0.11	0.00	0.67	0.42	0.34	0.38

Table 22: Readability of original and manually converted forms of ES texts used in benchmarking experiments

Changes in the readability of texts produced as a result of the two types of conversion were compared. The comparison makes reference to statistics presented in Tables 5a-b, 7a-b, and 9a-b in D7.2 (Section 5.2) and Table 16-Table 17, Table 18-Table 19, and Table 21-Table 22 in Section 3.1.2 of the current deliverable. The fact that several texts were converted using both methods enables direct comparison of the changes in readability brought about by unaided conversion and by conversion exploiting OpenBook.

Converting the 25 Bulgarian texts, it was observed that, when compared with the originals, unaided carers produced texts containing:

- fewer commas per word (an average of 3 fewer commas per 100 words),[9]
- shorter sentences (an average of 8.48 fewer words per sentence),[10]
- fewer phraseological units and non-lexicalised metaphors (an average of 4 fewer metaphors per 100 sentences),[11] but
- more polysemic words (an average of 3 more polysemic words per 100 words).[12]

The changes, in each of these features, between the original and converted versions of texts produced in the experiment were statistically significant.

Carers using OpenBook to make the conversion produced 25 BG texts containing, when compared with the originals:

- fewer phraseological units and non-lexicalised metaphors (an average of 4 fewer metaphors per 1000 sentences),[13] but
- more polysemic words (an average of 3 more polysemic words per 100 words)[14]

Again, the difference in these features between original and converted forms of the texts was statistically significant. It can be observed that, for Bulgarian, unaided conversion tends to lead to the production of more readable texts than conversion using OpenBook, at least as measured by the readability indices introduced in Section 5.2 of D7.2.

Five of the BG texts used in the benchmarking experiment were converted both by carers using OpenBook and by unaided carers (*CommunSkills* (*Text11*), *Habitat* (*Text12*), *SelfImage* (*Text13*), *SameSimTriangles*

[9] A change in *comma index* of -0.29
[10] A change in *index of words in sentences* of -8.48
[11] A change in *metaphor index* of -4.24
[12] A change in *polysemic word index* of +3.75
[13] A change in *metaphor index* of -0.44
[14] A change in *polysemic word index* of +2.76

(*Text14*), and *TextPublicCommun* (*Text15*)). Analysis of these ten conversions revealed that use of OpenBook led to smaller reductions in the number of metaphors mentioned in the converted texts (an average of 1 fewer per hundred sentences when using OpenBook vs. 3 fewer per hundred sentences when unaided), but smaller undesirable increases in the number of polysemic words introduced by the conversion process (an average of 3 more per hundred words using OpenBook vs. 6 more per hundred words when making unaided conversions).[15]

Converting the 25 English texts, it was observed that unaided carers produced texts containing:

- fewer commas per word (an average of 3 fewer commas per 100 words)[16]
- shorter sentences (an average of 6 fewer words per sentence)[17]
- restricted vocabulary (an average of 4 fewer vocabulary items per 100 tokens).[18]
- fewer phraseological units and non-lexicalised metaphors (an average of 1 fewer metaphor every 5 sentences)[19]

The changes, in each of these characteristics, between the original and converted versions of texts produced in the experiment were statistically significant.

Carers using OpenBook for conversion of English produced 25 texts containing:

- fewer commas per word (an average of 2 fewer commas every 100 words),[20]
- shorter words (an average of 2 fewer words with three or more syllables every 100 words),[21]
- more restricted vocabulary (an average of 4 fewer vocabulary items every 100 tokens),
- fewer phraseological units and non-lexicalised metaphors (an average of 1 fewer metaphors every 5 sentences),[22]
- reduced syntactic complexity (measured using the *index of productive syntax*).[23]

[15] Three ambiguous words added in every hundred when using OpenBook vs. six ambiguous words added in every hundred during unaided conversion.
[16] A change in *comma index* of -0.31
[17] A change in *index of words in sentences* of -6.43
[18] A change in *index of word diversity* of -0.04
[19] A change in *metaphor index* of -20.99
[20] A change in *comma index* of -0.2
[21] A change in *index of words with more than 3 syllables* of -0.23
[22] A change in *metaphor index* of -20.99
[23] A change in the *index of productive syntax* of -15.76

Conversion of EN texts exploiting OpenBook thus differs from unaided conversion. It leads to the production of texts containing shorter words and with reduced syntactic complexity. By contrast, unaided conversion leads to the production of texts containing shorter sentences.

Six of the EN texts used in the benchmarking experiment were converted both by carers using OpenBook and by unaided carers (*Melt in the body electronics*, *Climate change*, *Blood booster*, *DNA profiling*, *iPhone5*, and *PCs Features*). Analysis of these 12 conversions revealed that use of OpenBook led to smaller reductions in the number of metaphors mentioned in the converted texts (an average of 11 fewer per hundred sentences when using OpenBook vs. 21 fewer per hundred sentences when unaided). Changes in *comma index* were of similar magnitude under both conversion methods, with both leading to the production of texts with an average of 3 fewer commas per 100 words.[24] Changes in the *index of word diversity* were also of similar magnitude under both methods, with conversion exploiting OpenBook being slightly more effective in reducing the range of vocabulary used. Unaided conversion led to the loss of 2 vocabulary items (word types) per 100 tokens while conversion exploiting OpenBook led to the loss of 3 vocabulary items (word types) per 100 tokens.[25] By contrast, unaided conversion led to a much greater reduction in the number of phraseological units and non-lexicalised metaphors in the six converted texts. One reason for this is that while unaided conversion often involves replacement of metaphorical language with more literal expressions, conversion exploiting OpenBook frequently involves the addition of explanatory definitions of metaphors. Metaphors are not replaced – instead, explanations are provided. In English, unaided conversion led to a reduction of 127 metaphors per 1000 sentences whereas conversion exploiting OpenBook led to a reduction of just 2 metaphors per 1000 sentences.

Although the differences were not statistically significant, it was observed that, on average, unaided conversion led to the undesirable addition of 3 polysemic words per 1000 words[26] while conversion using OpenBook led to the *removal* of 11 polysemic words per 1000 words. [27]

Despite the integration of LT to convert long and complex sentences into sequences of shorter sentences, it was found that conversion exploiting OpenBook did not lead to significant differences in sentence length between the original and converted forms of texts. One explanation for this is that OpenBook provides users with the option to insert parenthetical definitions of words and metaphors into the sentences in which those elements appear. As a result, use of this function increases the length of sentences being converted.

[24] Changes in *comma index* of -0.26 in manual conversion and -0.25 in conversion using OpenBook.
[25] Changes in *index of word diversity* of -0.02 in manual conversion and -0.03 in conversion using OpenBook.
[26] A change in *polysemic word index* of +0.28.
[27] A change in *polysemic word index* of -1.12

When considering the changes in readability induced by unaided text conversion and changes induced by text conversion using OpenBook, no statistically significant differences were observed. A t-test of changes of the values of *comma index*, *metaphor index*, and the *index of word diversity* derived from the two different types of conversion gave p > 0.2. Use of OpenBook reduces the time required to convert texts to a more accessible form without affecting the quality of conversion. It should be noted that while differences between these changes were insignificant, the indices used to estimate the readability of text are insensitive to the effects on readability of introducing new elements such as summaries, explanatory definitions, and illustrative images into the converted texts.

Converting the 25 Spanish texts, it was observed that unaided carers produced texts containing:

- fewer commas per word (an average of 2 fewer commas every 100 words)[28]
- shorter sentences (an average of 5 fewer words per sentence)[29]
- fewer phraseological units and non-lexicalised metaphors (an average of 9 fewer metaphors per 100 sentences)[30]

The changes, in each of these characteristics, between the original and converted versions of texts produced in the experiment were statistically significant.

Converting the 25 Spanish texts, it was observed that carers exploiting OpenBook produced texts containing:

- fewer commas per word (an average of 2 fewer commas per 100 words)[31]
- shorter sentences (an average of 3 fewer words per sentence)[32]
- fewer phraseological units and non-lexicalised metaphors (an average of 11 fewer metaphors per 100 sentences)[33]

Three of the ES texts used in the benchmarking experiment were converted both by carers using OpenBook and by unaided carers (*El oro negro*, *Fisica Cuantica*, and *Lluvias Torrenciales*). Analysis of these six conversions revealed that use of OpenBook led to similar reductions in the number of commas used when compared with unaided conversion (3 fewer per hundred words in each case). Use of OpenBook led to larger reductions in sentence length than unaided conversion did (an average of 7 fewer words per sentence vs. 3

[28] A change in *comma index* of -0.23
[29] A change in *index of words in sentences* of -5.12
[30] A change in *metaphor index* of -0.09 (for ES in D7.2, this index is not reported as a percentage)
[31] A change in *comma index* of -0.16
[32] A change in *index of words in sentences* of -2.69
[33] A change in *metaphor index* of -0.11 (for ES in D7.2, this index is not reported as a percentage)

fewer words) and larger reductions in the number of metaphors used in the converted texts (20 fewer per 1000 sentences vs. 3 fewer per 1000 sentences).

3.1.2.5 Text Conversion Operations

Texts used in the semi-automatic conversion task were inspected manually as well as electronically. Manual inspections were focused on features that could not be identified electronically such as metaphors, idioms, specialised terms, etc. (Table 23).

Use of OpenBook was found to improve overall recall in the detection and removal of obstacles to reading comprehension. The mean number of obstacles removed overall in unaided conversion and for all centres was 35.1. The number of obstacles removed in conversion exploiting OpenBook was 61.3. Almost twice as many obstacles were resolved when using OpenBook than when conversions were made in an unaided fashion. This was consistent for all centres (Figure 6). Professionals detected and removed more obstacles when using OpenBook than when they carried out the same work independently (manually and/or exploiting functions built-in to their PC operating systems).

Evaluation of the contribution made by OpenBook to the semi-automatic text conversion process is complex. In this report, the ability of a text conversion process exploiting OpenBook to improve the readability of the texts being processed was compared with that of an unaided manual text conversion process. The impact on carers performing conversion operations was investigated, and information on the changes in time and effort required for unaided text conversion and conversion exploiting OpenBook was reported.

Unaided conversion of texts in the three languages is more time-consuming than conversion exploiting OpenBook (Table 15). However for Bulgarian, comparison of the results presented in Table 16-Table 17 in this deliverable with Tables 5a-b in deliverable D7.2 indicates that unaided conversion leads to the production of texts in which sentences are shortened to a greater extent and in which the number of commas (an indication of syntactic complexity) has been more greatly reduced than those produced when OpenBook is exploited. Both unaided conversion and conversion using OpenBook lead to the production of texts that are significantly more readable than the originals when assessed using the *metaphor index*. However, unaided conversion leads to significant improvements in readability measured in terms of sentence length and the number of commas occurring in sentences.

Conversion Operations	Spanish Texts (frequency)		English Texts (frequency)		Bulgarian Texts (frequency)	
	Manual	OpenBook	Manual	OpenBook	Manual	OpenBook
Synonyms **An entity usually a noun or an adjective is replaced with its less complex synonym**	64	142	83	120	67	136
Sentence Splitting **A long sentence is split in shorter sentences or in bullet list.**	40	38	70	97	70	78
Definition **A difficult term is explained using an explanation extracted from Wikipedia, a dictionary or Web**	34	26	65	94	--	97
Near - Synonymous	33	18	21	84	36	17
Image **A concept is illustrated by an image**	27	52	--	61	29	--
Explanation **The sentence is rewritten with different words**	24	42	43	55	116	152
Anaphora **Resolving the anaphor to its antecedent**	17	29	--	47	1	28
Deletion **Parts of the sentence are removed**	17	66	58	44	69	203
Coreference resolution	9	--	3	--	--	--
Syntactic Operation **A transformation on the syntactic parse trees**	9	20	42	31	73	17
Metaphor **A metaphor is explained**	8	20	30	--	19	--
Summarization **Providing the gist of the sentence or the paragraph**	3	2	10	--	1	12
Overall Mean	**24**	**41**	**42.5**	**70.3**	**43**	**73.5**

Table 23: Frequency of text conversion operations for each centre: unaided conversion vs. conversion exploiting OpenBook

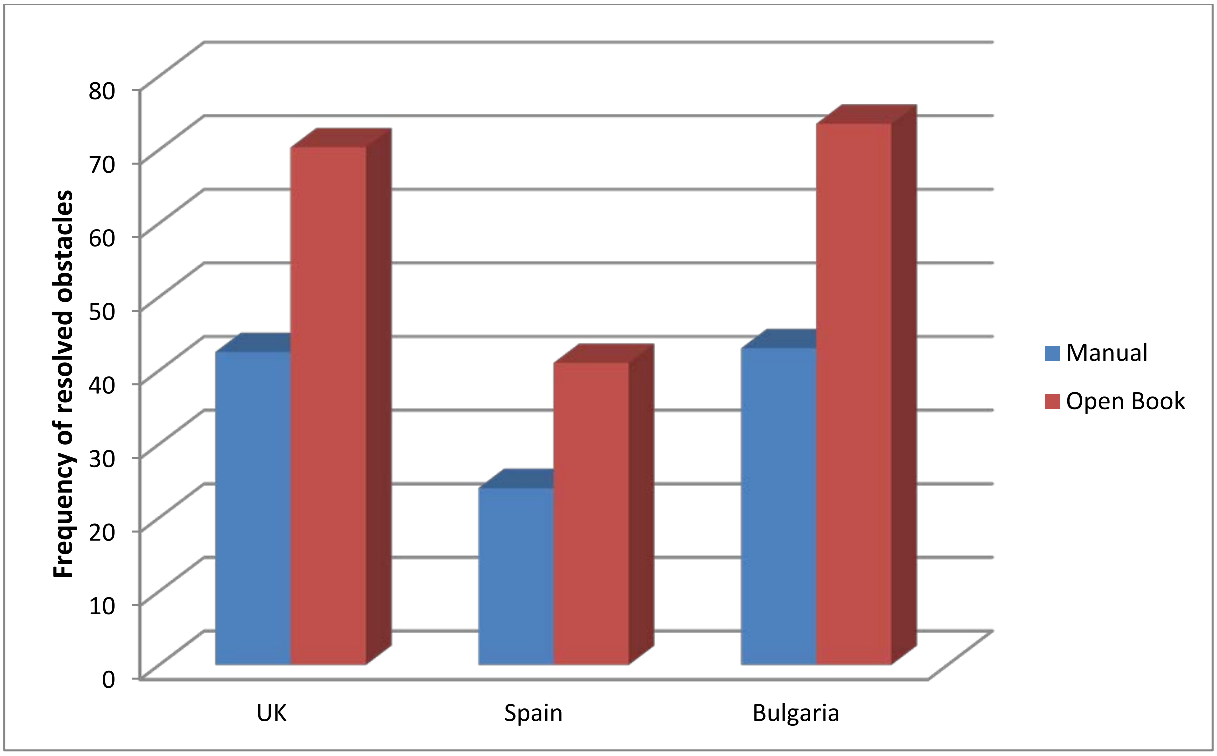


Figure 6 Frequency of removed obstacles to reading comprehension: Manual vs. OpenBook

For English, unaided conversion leads to the production of texts in which both the number of commas and the length of sentences has been more greatly reduced than those produced when OpenBook is exploited. However, values of the *index of productive syntax*, which assesses syntactic complexity via a wider range of morpho-syntactic indicators, show that the syntactic complexity of texts produced with the aid of OpenBook has been reduced to a greater extent than that of texts produced by unaided conversion. In texts produced by conversion using OpenBook, there is also a greater reduction in the number of long words than in texts produced by unaided conversion.

For Spanish, unaided conversion leads to the production of texts in which sentence length has been reduced to a greater extent than those produced by conversion using OpenBook. However, texts produced using OpenBook show a greater reduction in the number of phraseological units and non-lexicalised metaphors than those produced by unaided conversion.

OpenBook has reportedly had a positive effect on professionals' task of converting complex texts into a more accessible form for people with autism. Significantly more obstacles to reading comprehension were removed overall for all centres (**61.3 Vs 35.1**) when Open book was utilised and this was achieved in almost half the time (**29.48 mins Vs 54.33 mins**) taken for unaided conversions. Therefore, we can conclude that OpenBook will be a useful tool that will not only be effective in supporting professionals in converting text appropriately and correctly but this will significantly reduce the amount of time and effort needed for conversion, improving their productivity.

3.2 Evaluation of LT Integrated in OpenBook: Updated Error Analysis [D7.vii]

Detailed evaluation and error analysis of the LT services developed in the project were presented in Deliverable D7.5 and in Deliverables D3.1, D4.1, and D5.1. The current report outlines three recent studies evaluating the *structural complexity processor* developed for Bulgarian, evaluating new functions based on machine learning to detect the occurrence of relative clauses in English, and evaluating the *meaning disambiguator* developed for all three languages.

3.2.1 Evaluation of the Structural Complexity Processor (BG)

The Bulgarian structural complexity processor consists of a rule based approach to identify and extract subordinate clauses from complex sentences. Typically these clauses convey additional information which may be less important and can be safely deleted as a way to improve the readability of sentences. In cases where removing the information leads to loss of essential information, the editor (the intermediary) has to intervene and either restore the original sentence or manually rephrase the deleted clause.

This LT service detects relative clauses and other types of embedded clause. To do so, the system examines the syntactic dependency tree generated by a syntactic parser for Bulgarian. The left boundaries of embedded clauses are signaled by the occurrence of punctuation marks, conjunctions, and relative pronouns. Several rules are then used to identify dependency tree sub-structures which signal the right boundary of the embedded clause. There are three main cases addressed by the current LT service:

- Type 1: subordinate clauses which extend to the end of the sentence
- Type 2: embedded clauses which occur in the middle of a complex sentence
- Type 3: appositives and parenthetical expressions,

When dealing with embedded clauses of Type 1, it may be possible to simply split the sentence in two, and manually correct the new sentence if it contains essential information.

1. Юнкер, който най-вероятно ще поеме поста през ноември, казва, че не вижда в следващите 5 години да има разширяване на ЕС, но не уточнява дали това включва и казуса с Шотландия.
   Juncker, who will most probably take the post in November, says that he does not see the EU expanding

in the next 5 years but he does not clarify whether this includes the case of Scotland.

2. Англичаните бомбардират пристанищата на континента, включително Антверпен, Остенде, Кале, Дюнкерк.
   The English are bombing the harbours on the continent, including Antwerpen, Ostend, Calais, Dunkirk.

3. Рано сутринта на 9 май немските ръководители в Берлин подписват подобна капитулация спрямо съветската армия, и генералисимус Сталин обявява края на войната
   In the early morning of 9 May the German leaders in Berlin sign a similar Surrender to the Soviet army and Generalissimus Stalin proclaims the end of the war.

4. Целта на това германско нашествие е да помогне на съюзниците си и да изгони британските войски от Гърция, предотвратявайки по този начин евентуална заплаха за находищата на нефт в Румъния от страна на вражеските войски.
   The purpose of this German invasion is to help their allies and to push the British troops out of Greece, thus preventing a possible threat to the oil fields in Romania from the enemy troops.

Type 2 clauses occur in the middle of long and complex sentences. In these cases, detecting the right boundary of the clause correctly can be very challenging.

1. Вотът може да попречи на малко над 5 милиона граждани на ЕС да останат в съюза и това да се случи против волята им, но пък ще успокои всички държави, които се страхуват, че успешен референдум в Шотландия ще отвори Кутията на Пандора за претенции на Каталуня, баските, фламандците, бретанците, ломбардците и т.н. "Каквото и да казват адвокатите, всичко в крайна сметка опира до политиците", казва служител на ЕС в Брюксел, пожелал – подобно на всички дипломати и чиновници – анонимност и говорещ само на теория.
   The vote may prevent a bit more than 5 million EU citizens to stay in the Union and this will happen against their will, but it will comfort all countries, which are afraid that a successful referendum in Scotland is going to open the Pandora's box for demands from Catalonia, the Basques, Flamandians, Bretons, Lombardians, etc. "Whatever lawyers may say, it is all, in the end of the day, up to the politicians" - an EU officer in Brussels said, who - similar to all diplomats and officers - has requested anonymity and is speaking only in theory.

When detecting appositions and parenthetical constructions (Type 3), these elements are removed from the sentence as they are not part of the main sentence and their deletion does not affect its meaning.

1. Немските военни операции в Гърция и Северна Африка отлагат с няколко седмици запланувaното германско нападение, започнато на 22 юни 1941 г. Прахосана е голяма част от хубавото лятно време (за подробности относно военните операции вижте Операция Барбароса).
   The German military operations in Greece and North Afrika postpone the planned German attack, which started on 22 June, 1944, with a few weeks. A big part of the nice summer weather is wasted (for more details on military operations see Operation Barbarossa).

2. Върху японския град Хирошима на 6 август 1945 е хвърлена втората в света атомна бомба, разработена тайно от американското правителство (първата е използвана за тест два месеца преди това.
Upon the Japanese town Hiroshima on the 6th of August 1945, is thrown the second in the world nuclear bomb, secretly developed by the American government (the first one was used as a test two months prior to this).

As is the case for English, automatic processing of Bulgarian syntax has limitations. Relatively small errors can lead to essential change in the meaning of the resulting text. For this reason, the service should be used to provide suggestions for editing by intermediaries (carers, teachers, or parents) rather than being provided for direct consumption by end users.

The LT service for BG was evaluated by linguistic experts in its processing of 82 sentences. Table 24 displays statistics derived during this process.

Rule Type	#Correct	#Incorrect	Minor grammatical mistakes	Major grammatical mistakes	Unimportant information missing	Important information missing	Accuracy
Type 1	32	7	0	7	0	21	0.82
Type 2	0	1	0	1	0	1	0
Type 3	28	14	3	12	0	12	0.66
Totals	**60**	**22**	**3**	**20**	**0**	**34**	**0.73**

Table 24: Evaluation of the rules exploited by the structural complexity processor for Bulgarian

The figures in each row relate to each of the three types of rule exploited by the LT service detailed in Section 3 of deliverable D3.1. The columns summarise information on the number of sentences converted correctly by the rules (*#Correct*), the number converted incorrectly (*#Incorrect*), the number of conversions that led to the generation of sentences containing minor and major grammatical errors (*Minor grammatical mistakes* and *Major grammatical mistakes*, respectively), the number of sentences in which unimportant or important information had been omitted (*Unimportant information missing* and *Important information missing*, respectively), and the accuracy of rules of the three different types (*Accuracy*). In this study, *accuracy* is computed as the ratio of *#Correct* to the sum of *#Correct* and *#Incorrect*.

The *accuracy* of rules of Type 1 and Type 3 was found to be acceptable for exploitation by carers for the semi-automatic conversion of text to a more accessible form for end users in this small experiment. It was observed that application of rules of Type 1 (with an accuracy of 0.82) led to the generation of converted sentences containing major grammatical errors in 17.95% of cases and in which important information had been omitted in 53.85% of cases. Rules of Type 3 had an accuracy of 0.67. Application of rules of Type 3 led to the generation of converted sentences containing minor grammatical errors in 7.14% of cases, sentences containing *major* grammatical errors in 28.6% of cases, and sentences in which important information had been omitted in 28.6% of cases. Rules of Type 2 applied just once, incorrectly, when processing the small set of test data involved in this experiment. This single error exemplified both major grammatical errors and the omission of important information.

3.2.2 Evaluation of the Structural Complexity Processor (EN)

Evaluation and error analysis of the *structural complexity processor* (v2) for EN was presented in Deliverable D7.5 (Section 3). In D7.5, evaluation was focused on LT services for sign tagging and on rule-based approaches to the rewriting of compound sentences and complex sentences. The current report details a recent study evaluating alternate approaches to the detection of relative clauses for the purpose of rewriting sentences which contain complex noun phrases (user requirements UR302 and UR309, described in deliverable D2.2).

3.2.2.1 Relative clause extraction for syntactic simplification

This study investigates non-destructive text conversion, a type of syntactic text processing which focuses on extracting embedded clauses from structurally complex sentences and rephrasing them without affecting their original meaning. This process reduces the average sentence length and complexity to make text more accessible.

To evaluate the performance of the relevant syntactic processing components, a dataset covering three genres was manually annotated and used to develop and compare several approaches for automatically detecting appositions and non-restrictive relative clauses.

In this experiment, the rule-based method for detecting relative clauses (presented in more detail in Section 4.2.2 of D3.1 and evaluated in D7.5) was compared with a machine learning (ML)-based tagging model developed using conditional random fields (CRF).

This LT component performs a form of syntactic processing in which text is rephrased in such a way that the meaning of the original text is preserved as much as possible. This is specifically linked to certain types of clausal structures which can be deleted from the original sentence without affecting meaning. These types include appositions and non-restrictive relative clauses (Siddharthan, 2002). This research investigates a method specifically developed for identifying appositions and non-restrictive relative clauses which can be removed from a text without losing essential information.

3.2.2.2 Dataset

To carry out this evaluation, it was necessary to annotate text from the FIRST corpus covering three genres: news, literature and health. A set of annotation guidelines were created, given to the annotators and then explained in a group discussion where several examples were also analysed. Subsequently, annotators were given a small set of sentences to trial individually. Their subsequent questions and feedback led to development of a revised set of guidelines.

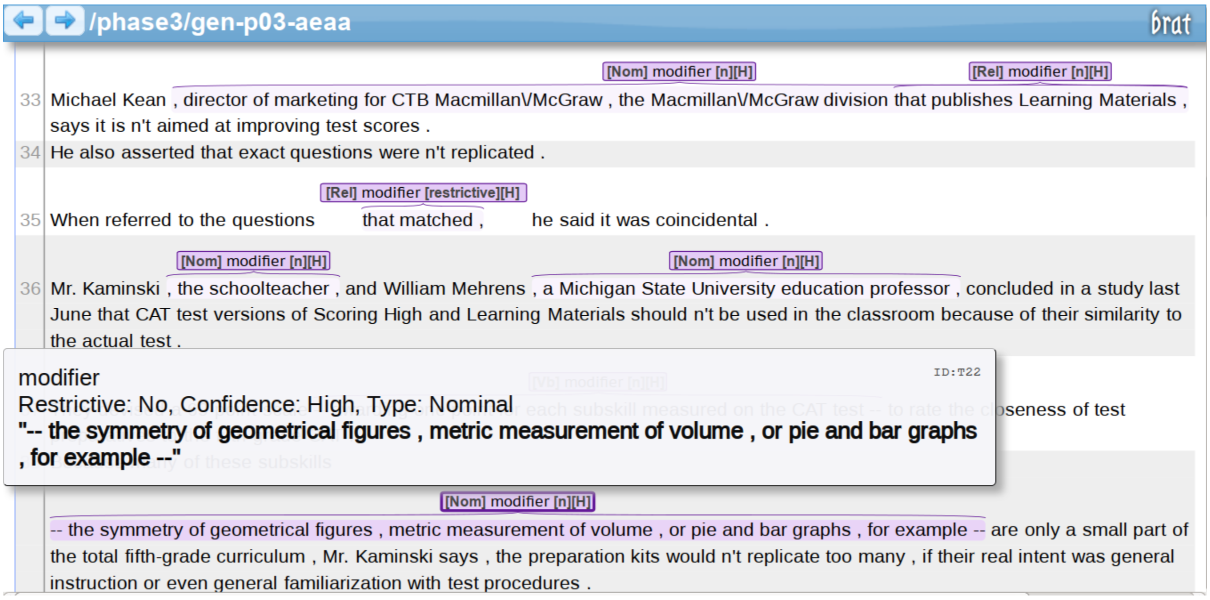
After this training phase was complete, the actual annotation was carried out. The corpus was split randomly with each part being annotated by at least two annotators. The annotation was performed using the BRAT tool (Stenetorp et al., 2012).

Figure 7 shows the interface of the annotation tool. For each annotated span, annotators were required to evaluate three attributes:

a) **type** (*relative, nominal, adjectival, verbal, prepositional*),
b) whether it is **restrictive** (*no, yes, unknown*), and
c) the annotator’s **confidence** (*low, medium, high*).

The amount of data in the corpus is listed for each genre in Table 25, i.e. number of sentences and tokens. On average, around half of the sentences contain an annotated span, but they occur more frequently in newswire and healthcare texts than in literature.

The annotated data used in this evaluation was made in accordance with the guidelines presented in Dornescu et al. (2014).



33 Michael Kean , director of marketing for CTB MacmillanVMcGraw , the MacmillanVMcGraw division that publishes Learning Materials , says it is n't aimed at improving test scores .

34 He also asserted that exact questions were n't replicated .

35 When referred to the questions that matched , he said it was coincidental .

36 Mr. Kaminski , the schoolteacher , and William Mehrens , a Michigan State University education professor , concluded in a study last June that CAT test versions of Scoring High and Learning Materials should n't be used in the classroom because of their similarity to the actual test .

modifier
Restrictive: No, Confidence: High, Type: Nominal
"-- the symmetry of geometrical figures , metric measurement of volume , or pie and bar graphs , for example --"
ID:T22

-- the symmetry of geometrical figures , metric measurement of volume , or pie and bar graphs , for example -- are only a small part of the total fifth-grade curriculum , Mr. Kaminski says , the preparation kits would n't replicate too many , if their real intent was general instruction or even general familiarization with test procedures .

Figure 7: Annotation tool and example of nominal apposition

Genre	(Corpus)	Sentences	Tokens	Spans	Span tokens	Sent. len.	Span len.
healthcare		1214	27379	958	6094	22.55	6.36
news	METER1	1038	28367	732	5592	27.33	7.64
news	METER2	1377	37515	1165	9203	27.24	7.9
literature		1946	48620	431	3834	24.98	8.9
news	Penn T.B.	1733	39740	625	5652	22.93	9.04
	Overall	**7308**	**181621**	**3911**	**30375**	**24.85**	**7.77**

Table 25: Annotated dataset

3.2.2.3 Annotation insights

The corpus was split into chunks of roughly 100-150 sentences and each was annotated by 2 to 5 annotators. Sentences were randomly selected from the corpus based on length and the presence of signs of syntactic complexity (conjunctions, commas, parentheses). Where annotated spans did not match, a reviewer made the final adjudication.

The agreement on detecting the span of post-modifiers was relatively low, on average, with pairwise F1 score of 54.90%. This is mainly because of the way annotators interpreted the instructions. For example, some annotators consistently marked all parenthetical expressions whereas others consistently failed to do so.

All annotators typically reached higher precision than recall. This suggests a systematic source of disagreement which affects files annotated by few annotators. One way to address this problem would be by aggregating all annotated spans for each document and asking annotators to confirm which of them are post-modifiers. This process will be carried out in a future study, using a voting scheme to mitigate the recall problem.

With respect to only those cases for which two annotators marked the same span as a post-modifier, it is possible to investigate the level of agreement reached on evaluation of the individual attributes: type and restrictiveness. Annotators reached high pairwise agreement (kappa=0.78) when marking the **types** of post-modifiers. This suggests that the left boundaries of post-modifiers have reliable markers which could be used to automatically predict the type of the bounded post-modifier. The pairwise agreement for **restrictiveness** is lower (kappa=0.51), but still at an acceptable level, given that the two values have an unbalanced distribution (70% of post-modifiers are non-restrictive). Possible causes of this are: lack of context (sentences were considered in isolation from their source documents), lack of domain knowledge (especially where post-modifiers do not provide information about named entities, but about specialised terminology, such as symptoms, procedures, strategies).

Although agreement on the two attributes can be improved, the biggest challenge is to ensure that all post-modifiers are annotated, i.e. to address situations in which only one annotator marks a span.

3.2.2.4 Automatic detection of relative clauses

In this study, the scheme for annotation of signs of syntactic complexity introduced by Evans and Orasan (2013) is followed. In this scheme, punctuation marks and functional words are considered explicit markers of coordination and subordination, the two syntactic phenomena underlying a large proportion of structurally complex sentences.

The signs of syntactic complexity comprise conjunctions ([*and*], [*but*], [*or*]), a complementiser ([*that*]), wh-words ([*what*], [*when*], [*where*], [*which*], [*while*], [*who*]), punctuation marks ([,], [;], [:]), and 30 compound signs consisting of one of these lexical items immediately preceded by a punctuation mark (e.g. [, *and*]). These signs are automatically tagged with a label indicating the type of constituent they bound, such as finite clauses (EV) or strict appositives (MN), and the position of the sign, such as start/left boundary (SS*) or end/right boundary (ES*). For example, the label ESMA indicates the right boundary (end) of an adjectival phrase. An automatic tagger of signs of syntactic complexity was developed using a sequence tagging approach (Dornescu et al., 2013) and is used in this work to select complex sentences from the corpus and to provide linguistic information exploited by the proposed approach.

3.2.2.5 Baseline system: RC1

System RC1 uses a set of rules to detect appositives which are delimited by punctuation marks and do not contain any verbs. Such expressions are typically nominal appositives or parenthetical expressions e.g.:

a. *The chief financial officer*, GREGORY BARNUM, *announced the merger in an interview.*

b. *Oxygen can be given with a face mask or through little tubes* (NASAL CANNULAE OR ’NASAL SPECS’) *that sit just under your nostrils.*

c. *The business depends heavily on the creativity of its chief designer*, SEYMOUR CRAY.

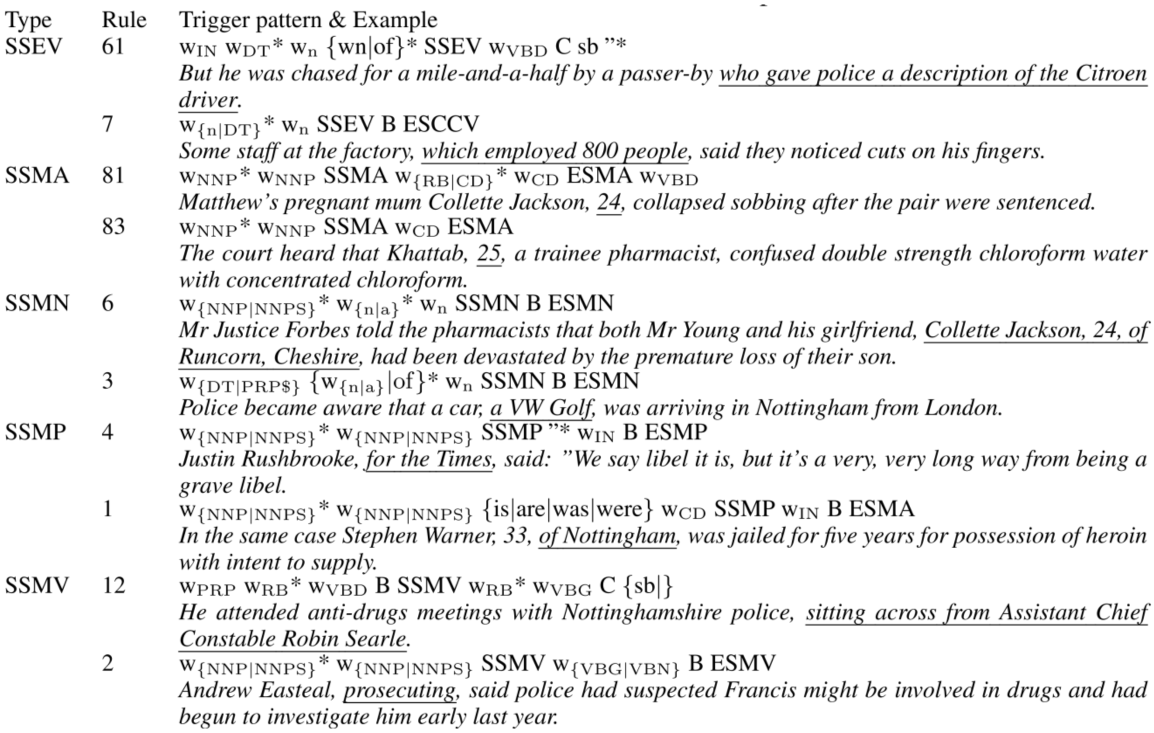
This baseline is intended to capture a set of frequently occurring noun post-modifiers, even though this set is relatively restricted.

3.2.2.6 Rule based system: DAPR

The second system, DAPR (Detection of Adnominal Post-modifiers by Rules), is a LT component developed for people with autistic spectrum disorders as part of the FIRST project (Evans et al., 2014). Although OpenBook can also rephrase complex sentences, in this report, it is only the appositive constituents detected in a sentence by DAPR that are considered.

DAPR employs several hand-crafted linguistic rules which detect the extent of appositions based on the presence of signs of syntactic complexity, in this case punctuation marks, relative pronouns, etc. DAPR exploits rules and patterns to convert sentences containing noun post-modifiers such as finite clauses (EV), strict appositives (MN), adjective phrases (MA), prepositional phrases (MP), and non-restrictive non-finite clauses (MV) into a more accessible form.

The conversion procedure is implemented as an iterative process. When a pattern matches the input sentence, the detected post-modifier is deleted and the resulting sentence is then processed. The priority of each pattern determines the order in which they are matched when processing sentences which contain multiple left boundaries of relevant constituents (i.e. signs of syntactic complexity tagged with certain labels). The patterns are implemented to match the first (leftmost) sign of syntactic complexity in the sentence. More details about the rules have been reported in D7.5. Figure 8 lists some examples.



Type	Rule	Trigger pattern & Example
SSEV	61	$w_{IN}\ w_{DT}^{*}\ w_n\ \{wn \mid of\}^{*}$ SSEV w_{VBD} C sb ”* *But he was chased for a mile-and-a-half by a passer-by who gave police a description of the Citroen driver.*
	7	$w_{\{n \mid DT\}}^{*}\ w_n$ SSEV B ESCCV *Some staff at the factory, which employed 800 people, said they noticed cuts on his fingers.*
SSMA	81	$w_{NNP}^{*}\ w_{NNP}$ SSMA $w_{\{RB \mid CD\}}^{*}\ w_{CD}$ ESMA w_{VBD} *Matthew’s pregnant mum Collette Jackson, 24, collapsed sobbing after the pair were sentenced.*
	83	$w_{NNP}^{*}\ w_{NNP}$ SSMA w_{CD} ESMA *The court heard that Khattab, 25, a trainee pharmacist, confused double strength chloroform water with concentrated chloroform.*
SSMN	6	$w_{\{NNP \mid NNPS\}}^{*}\ w_{\{n \mid a\}}^{*}\ w_n$ SSMN B ESMN *Mr Justice Forbes told the pharmacists that both Mr Young and his girlfriend, Collette Jackson, 24, of Runcorn, Cheshire, had been devastated by the premature loss of their son.*
	3	$w_{\{DT \mid PRP\$\}}\ \{w_{\{n \mid a\}} \mid of\}^{*}\ w_n$ SSMN B ESMN *Police became aware that a car, a VW Golf, was arriving in Nottingham from London.*
SSMP	4	$w_{\{NNP \mid NNPS\}}^{*}\ w_{\{NNP \mid NNPS\}}$ SSMP ”* w_{IN} B ESMP *Justin Rushbrooke, for the Times, said: ”We say libel it is, but it’s a very, very long way from being a grave libel.*
	1	$w_{\{NNP \mid NNPS\}}^{*}\ w_{\{NNP \mid NNPS\}}\ \{is \mid are \mid was \mid were\}\ w_{CD}$ SSMP w_{IN} B ESMA *In the same case Stephen Warner, 33, of Nottingham, was jailed for five years for possession of heroin with intent to supply.*
SSMV	12	$w_{PRP}\ w_{RB}^{*}\ w_{VBD}$ B SSMV $w_{RB}^{*}\ w_{VBG}$ C $\{sb \mid \}$ *He attended anti-drugs meetings with Nottinghamshire police, sitting across from Assistant Chief Constable Robin Searle.*
	2	$w_{\{NNP \mid NNPS\}}^{*}\ w_{\{NNP \mid NNPS\}}$ SSMV $w_{\{VBG \mid VBN\}}$ B ESMV *Andrew Easteal, prosecuting, said police had suspected Francis might be involved in drugs and had begun to investigate him early last year.*

Figure 8: Examples of rules for detecting noun post-modifiers used by DAPR

3.2.2.7 Tagging-based system

According to the reports on user requirements produced in WP2 (deliverables D2.1 and D2.2), the only requirement of OpenBook that specifically concerns subordinate constituents is the requirement to minimise the occurrence of subordinate clauses occurring in sentences. However, the occurrence of other types of subordinate phrases can also increase average sentence length. In order to address the requirement that long sentences be avoided (UR309), the method described in this section was developed. It is used to convert sentences containing other types of subordinate constituents into a more accessible form, detecting and removing noun phrase post modifiers from long sentences.

As many types of appositive modifiers are simple in structure, a tagging approach was adopted for the task of detecting noun post-modifiers. The commonly used IOB2 format was employed, in which the beginning of each noun post-modifier is tagged as B-PM and tokens inside it are tagged as I-PM. All other tokens are tagged as other: O. The task can be considered similar to a named entity recognition or to a chunking task where only one type of entity/chunk is detected.

The annotated corpus was used to build two supervised tagging models based on Conditional Random Fields (Lafferty et al., 2001): CRF++ and crfsuite. Four feature sets were used. Model A contains standard features used in chunking, such as word form, lemma and part of speech (POS) tag. Model B includes predictions made by the baseline system RC1 as an additional feature: using the IOB2 models, tokens have one of three values: B-RC1, I-RC1 or O-RC1. Similarly, model C adds the predictions of the DAPR system also using the IOB2 approach. This allows us to test whether the baseline systems are robust enough to be employed as input to the sequence tagging models. Model D adds information about the tokens of the sentence which are signs of syntactic complexity. These are produced automatically using the method described in Dornescu et al. (2013).

3.2.2.8 Results and analysis

Results reported by conlleval, the standard tool for evaluating tagging, are presented in Table 26. Although the two baselines, RC1 and DAPR, out-perform the CRF++ models, the best overall performance is achieved by the crfsuite models.

The rules employed by the RC1 baseline can be misled by sentences containing enumerations, numerical expressions and direct speech due to false positive matches. The rules perform well, despite being few in number and addressing the simplest kinds of post-modifiers.

The more complex baseline, DAPR, appears to be more conservative (it makes the fewest predictions overall), which suggests it covers fewer types of appositions than covered by the test dataset. Compared to the previous baseline, DAPR detects more complex appositions and relative clauses with better precision, but with reduced recall.

		#predicted modifiers	#correct modifiers	Accuracy	Precision	Recall	F1
RC1	baseline	1287	371	81.01	28.83	17.68	21.92
DAPR	baseline	535	163	81.25	30.47	7.77	12.38
CRF++	A:word & POS	3372	289	85.48	8.57	13.78	10.57
	+B:RC1	3381	315	85.66	9.32	15.01	11.50
	+C:DAPR	3586	319	85.63	8.90	15.20	11.22
	+D:tagged signs	3680	319	85.60	8.67	15.20	11.04
crfsuite	A:word & POS	1391	790	87.54	56.79	37.65	45.29
	+B:RC1	1437	825	87.55	57.41	39.32	46.68
	+C:DAPR	1470	838	87.56	57.01	39.94	46.97
	+D:tagged signs	1481	838	87.56	56.58	39.94	46.83

Table 26: Results reported by conlleval on the test set (90076 tokens, 2098 annotated post-modifiers)

Although the CRF models also use as features the predictions made by the two baseline models, due to the level of noise, the improvement is small, between 1 and 2 points. Adding information about the tagged signs of syntactic complexity actually has a negative impact on both models, suggesting that the signs are less relevant for this type of syntactic constituent. A large difference in performance is noted between the two CRF tools: whereas CRF++ is outperformed by both baselines, crfsuite achieves much better performance despite using the same input features.

To gain better insights into the performance of the best model, Table 27 presents label-wise results. Given that the average length of a post-modifier is 7 tokens, inside tokens (I-PM) are 7 times more prevalent than beginning tokens (B-PM). Despite this, the model achieves similar performance for both (F1 score just below 0.60). The two tables also bring evidence suggesting that detecting the end token of a post-modifier is challenging: although the start is correctly detected for 48.89% of appositives, only 39.94% are a perfect match. This suggests that more work is necessary to improve the ability to detect post-modifiers but also to better determine their correct extent. The second part is critical to the perceived performance of the text converion system, as incorrect detection usually leads to incorrect text being generated for users, whereas a loss in recall may be invisible.

Label	#match	#predicted	#reference	precision	recall	F1
O	70452	77884	73955	90.46	95.26	92.8
B-PM	1014	1469	2074	69.03	48.89	57.24
I-PM	7406	10723	14047	69.07	52.72	59.8
			Macro-average	**76.18**	**65.63**	**69.95**

Table 27: Label-wise performance for the best model (crfsuite C)

3.2.2.9 Error analysis

The most frequent error was omission of relevant post-modifiers; most of the automated systems typically reached higher precision than recall. This is similar to the phenomenon occurring during annotation. A way to mitigate this in the future is to consider predictions made by a committee of automatic models, which complement each other in terms of recall.

One common error concerns noun modifiers within the same NP, such as prepositional phrases. For example: *The $2.5 billion Byron 1 plant near Rockford was completed in 1985.*

While this span modifies a noun, it is part of the NP itself, and it is arguably too short to be relevant for rephrasing the sentence in an automatic text simplification system; it is more likely a candidate for deletion in the context of destructive sentence compression systems.

Another frequent issue concerns nested modifiers, where systems usually fail to distinguish both constituents. A related issue is how to deal with nested and overlapping spans, not only from the point of view of data representation, but also in the way the detected modifiers can be transformed.
The new plant, located in Chinchon, ***about 60 miles from Seoul,*** *will help meet increased demand.*

An interesting debate concerns ambiguous constituents which can have several interpretations. In the previous example, the second constituent about 60 miles from Seoul can be considered an apposition modifying the proper noun Chinchon, or a prepositional phrase modifying the verb located; both entail a similar meaning to a human reader. This example illustrates a situation which frequently occurs in natural language text: for stylistic or editorial reasons writers omit words which are implied by the context. The effect is that the syntactic structure becomes ambiguous, but the information communicated to the reader is nevertheless unaffected.

A consequence of this phenomenon is that distinguishing the type of a post-modifier (i.e. relative, nominal, adjectival, verbal, prepositional) only reflects its form and less so its role. For example, it is easy to rephrase most post-modifiers as relative clauses, e.g.

a) Nominal-appositives: *My wife*, **[who is]** a nurse by training, *has helped the accident victim.*
b) Verbal-appositives: *Lord Melchett led a dawn raid on a farm in Norfolk*, **[which caused]** ~~**causing**~~ 17,400 of damage... , *a court was told yesterday*.
c) Prepositional-appositives: *Boe*, **[who lives in]** ~~**of**~~ Chelmsford, Essex, *admitted six fraud charges and asked for 35 similar offences to be taken into consideration*.
d) Adjectival-appositives: *Student Richard*, **[who is]** 5ft 1Oins tall, *has now left home*.

The type of post modifier is a piece of information necessary in the text generation phase, but, as the above examples show, it is less critical to automatically identify type at the time of detection. A similar conclusion concerns the other attribute: restrictiveness. Based on initial experiments, type can be deduced based on the initial tokens in the noun phrase post-modifier, but restrictiveness is more challenging: in many cases pragmatic information, context and world knowledge is required to correctly determine whether a post-modifier is restrictive or not. Depending on restrictiveness, different strategies will be employed for rephrasing the content in order to best preserve the meaning of the sentence.

3.2.2.10 Conclusions on updated evaluation and error analysis of the structural complexity processor

As described in Deliverable D7.5, OpenBook originally integrated a rule-based method to identify relative clauses in complex sentences for the rewriting those sentences into a more readable form (DAPR). Evaluation of DAPR revealed that only a small set of noun post-modifiers are detected, resulting in poor recall. As a result, a new data-driven approach to the identification of relative clauses and other types of noun post-modifier was developed with the aim of obtaining better coverage than the original system.

Implementation of the data-driven approach was supported by the development of a set of guidelines and human annotation of a corpus with information about the noun post-modifiers occurring in it. These processes provided insights into the challenges and into the frequent errors made by both human annotators and the automatic system when identifying the spans of noun post-modifiers. Under the data-driven approach, a model was developed to complement the detection capabilities of DAPR: together, the joint system reaches superior levels of precision and recall than DAPR (above 50%).

It was found that a method converting complex sentences into a more accessible form by deleting detected noun post-modifiers generates incorrect output in almost half of the cases. For this reason, the developed component is not enabled for direct consumption by end users, though it could prove useful for intermediaries converting documents into a more accessible form. To preserve the original meaning, post-modifiers should not simply be deleted from complex noun phrases. Instead, new stand-along sentences should be generated which explicitly assert the links between noun post-modifiers and the modified nouns. This process involves detecting the correct subject and the predicate, as these two elements are typically omitted for brevity. Text generation is recognised to be an open problem in NLP and finding solutions to it was not one of the stated objectives of the FIRST project. However, it is an area where OpenBook can be used to collect valuable data: original and transformed sentence pairs. At present, OpenBook continues to deploy DAPR for the purpose of rewriting sentences which contain complex noun phrases.

3.2.3 Evaluation of the Meaning Disambiguator

Different methods were used to evaluate the components integrated in the Meaning Disambiguator v2.1, as appropriate for the component and the metrics and resources available. Different methods were used for:

- Evaluation of the coreference module
- Evaluation of the detection, disambiguation and resolution of polysemy, mental verbs, less common words and specialised slang
- Evaluation of the detection and resolution of infrequent slang
- Evaluation of the detection and resolution of acronyms/abbreviations
- Evaluation of the Meaning Disambiguator v2.1 with Science documents

More detail on the evaluation is reported in deliverable D4.1, where initial evaluations of some of the components have already been presented.

3.2.4.1 General evaluation of the coreference module

The different approaches for coreference resolution in Bulgarian, English, and Spanish were evaluated in a quantitative manner, due to the existence of available corpora (see D7.5 and D4.1 for more information).

Broadly, it is concluded that after evaluating in detail different coreference approaches for English, the most appropriate for being included in the Meaning Disambiguation framework, and therefore in OpenBook, was the Stanford Deterministic coreference resolution system. For Spanish, the coreference module integrated in OpenBook relied on Freeling for the detection of pronominal anaphora and definite descriptions; Naïve Bayes machine learning algorithm for detecting ellipsis; Voted Feature Interval machine learning algorithm for resolving pronominal anaphora and ellipsis, and PART algorithm for resolving definite descriptions. Comparing the results for Spanish coreference with respect to the state of the art, we find that the best performing system in the SemEval-2010 task[34] (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 Spanish 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. These results from a research point of view are a good contribution to the state of the art in this task.

3.2.4.2 General evaluation of the detection, disambiguation and resolution for polysemy, mental verbs, less common words and specialised slang

The Most Frequent Sense (MFS) approach for word sense disambiguation was evaluated using gold-standard corpora developed by the research community. The availability of such corpora allowed us to compare the output of our disambiguation approach with the correct one. Table 28 summarises the results for the MFS approach for English and Spanish (more detailed experiments and results for this method were provided in D7.5 and D4.1).

Language	Corpora	Precision
English	Semeval 2013, Multilingual Word Sense Disambiguation task	52.10%
Spanish		57.45%

Table 28: Results for the Most Frequent Sense approach for the Disambiguation stage

The corpus used was the one provided in Semeval 2013 competition[35], and the results obtained are in line with the state of the art for this task (McCarthy, 2011).

Furthermore, thanks to the efforts made in the development of an annotated corpus for Spanish, we analysed in more detail the disambiguation task, testing other approaches apart from the MFS approach. In this respect, we studied and experimented with alternative approaches that use Wikipedia and Wiktionary as resources, instead of WordNet and its related-resources for Bulgarian and Spanish. The reasons for also analysing these resources were due to the fact that:

i) Wikipedia/wiktionary have become very popular among all users, as a means of searching for information

This is shown by the statistical analysis provided in Wikipedia[36] and Wiktionary[37] projects.

[34] http://stel.ub.edu/semeval2010-coref/
[35] https://www.cs.york.ac.uk/semeval-2013/

Focusing on the usage of Wikipedia, only in March 2014, for English, 9,557 million Wikipedia pages were visited, with a rate of 319 million pages per day. In the case of the other languages for the same month, 1,406 million pages were visited in Spanish (46.9 million page per day), and 36.4 million pages for Bulgarian (1.2 million a day). From the same statistics, it is shown that English and Spanish are the languages in which Wikipedia is most looked up.

Regarding the usage of Wiktionary, in March 2014, the statistics showed that 77.2 million pages (2.6 million per day); 7.9 million pages (263 thousand a day); and, 830 thousand pages (28 thousand a day) were visited for English, Spanish and Bulgarian, respectively.

Although Wiktionary is not as widely used as Wikipedia, both resources are becoming more and more popular, and also provide a valuable LT resource that can be further exploited due their wide acceptance as main references on the Internet, as well as their multilingual nature.

ii) Wikipedia and Wiktionary include more data than WordNet and its extensions

The main difference between WordNet related resources and Wikipedia/Wiktionary resources is that the former were specifically developed by the LT research community and its purpose was to have a resource capable of supporting semantic analysis of text, particularly in the word sense disambiguation task. By contrary, Wikipedia and Wiktionary emerged as a collaborative online encyclopedia and dictionary, respectively. Given the rise of the Internet, and the popularity that these resources gained, they rapidly were also exploited by the LT community for analysing and experimenting with multilinguality. Table 29 shows a comparison between Wikipedia, Wiktionary, and WordNet and its related resources for Bulgarian (Bulgarian WordNet) and Spanish (MultiWordNet) with respect to their size.

[36] http://stats.wikimedia.org/EN/Sitemap.htm
[37] http://stats.wikimedia.org/wiktionary/EN/Sitemap.htm

Language	Semantic Resource	Number of articles/synsets/words	Number of definitions	Number of synonyms
Bulgarian	Bulgarian WordNet	23,502	23,502	23,502
	Wikipedia	152,794	> 152,794	-
	Wiktionary	821,547	> 821,547	Unknown
English	WordNet	117,000	117,000	117,000
	Wikipedia	4,334,000	> 4,334,000	-
	Wiktionary	3,164,724	> 3,164,724	Unknown
Spanish	MultiWordNet	118,712	20,094	118,712
	Wikipedia	1,052,162	> 1,052,162	-
	Wiktionary	77,362	> 77,362	Unknown

Table 29: Comparative between WordNet, Wikipedia, and Wiktionary

As can be seen from the table, Wikipedia and Wiktionary have much more coverage for detecting obstacles than WordNet-based resources. The main limitation of WordNet is related to its coverage. Although it is a very fine-grained semantic resource for LT, there is a considerable difference between the English version and the Bulgarian and Spanish ones. On the one hand, for the Bulgarian version, the number of synsets is reduced almost fivefold, and therefore, definitions and synonyms are only available for a limited number of synsets. On the other hand, the Spanish MultiWordNet contains more synsets than the English WordNet, and synonyms are provided for all of them. However, the definition is only provided for 17% of these concepts. This is a problem for the Meaning Disambiguator, since the definition of a word constitutes one of the most important assistive elements defined in the user requirements (D2.2). In light of this comparison, it was determined to experiment with Spanish Wikipedia and Wiktionary as alternative resources in order to study other disambiguation methods, and analyse whether the approach could be improved.

The disambiguation approaches proposed and compared were:

- ***MFS-Freeling:*** in this approach, we rely on the results provided by Freeling, which can be configured to obtain the MFS for a word, providing the corresponding MultiWordNet synset.
- ***Wikipedia-baseline*** consists of the first paragraph of the article corresponding to the word searched for in Wikipedia. In case the word refers to a Wikipedia Disambiguation Page (e.g. the Disambiguation page for the word "plant"[38]), the first paragraph of the first sense returned in this page is returned instead.

[38] http://en.wikipedia.org/wiki/Plant_%28disambiguation%29

- ***Wikipedia-MFS:*** if Wikipedia returns a Disambiguation page for the query word, the number of external links for the first paragraph of each of the possible senses is computed, assuming that the one containing the largest number of links will be the correct sense for that word. In case the word is directly associated to a Wikipedia article, the first paragraph of the article is associated as the definition of this word (in this latter case, the behaviour is the same as for Wikipedia-baseline).
- ***Wiktionary-baseline:*** the first definition for an occurrence in Wiktionary, regardless of the senses and definitions it has.
- ***Wiktionary-MFS:*** if the occurrence has more than one definition, the definition that contains more external links to other terms in Wiktionary is returned. If several definitions contain the same number of links, the first definition listed will be returned.

To evaluate these approaches, a portion of the corpus annotated for the FIRST project (please see D4.1, section "Available and Annotated Corpora for Evaluation") was selected. First, the number of obstacles that had to be selected in order to obtain a representative sample was computed. For this, the formula described in (Fernández, 1996) was used, since it returns the minimum value of obstacle to build representative sample:

$$M = \frac{N * K^2 * P * Q}{E^2 * (N - 1) + K^2 * P * Q}$$

In this formula, N is the population, K in the confidence interval, E is the error, P is the success probability and Q is the failure probability. The parameters were set in the same way as in (Gutiérrez Vázquez et al., 2010), except the population that was set to the total number of annotated obstacles (N=2949). The obstacles are considered to include polysemous words, mental verbs, less common words, and specialised slang. It was found to be necessary to evaluate our approaches taking into account at least 88 obstacles in order for the sample to be representative, if we wanted our sample to be representative. With these findings, a 20% sample of the corpus (635 obstacles) was randomly selected for the purpose of evaluation.

Having determined the section of the corpus to be tested, the extent to which the proposed approaches returned definitions for those obstacles in MultiWordNet, Wikipedia and Wiktionary was assessed. This corresponds to the evaluation of the **detection stage**. Table 30 shows the results.

Approach	Number of obstacles to be evaluated	Number of obstacles with definition	Coverage (%)
MFS-Freeling	635	222	35%
Wikipedia-baseline		323	51%
Wikipedia-MFS		343	54%
Wiktionary-baseline		484	76%
Wiktionary-MFS		504	79%

Table 30. Definition coverage for the proposed approaches.

The comparison previously shown provides data on the number of obstacles that can be detected and resolved (by providing its definition) under each approach. As expected, the use of MultiWordNet leads to the lowest coverage, with the ability to detect only 35% of the obstacles. By contrast, the differences between Wiktionary and Wikipedia are considerable. Wiktionary is, in this case, the resource that provides the best coverage with the MFS approach (i.e., returning the definition with more external links), despite the fact that 21% of the words could not be detected using this approach. The superior performance of Wiktionary over Wikipedia in detection, despite containing less entries, is logical, since Wiktionary is conceived as an online dictionary, whereas Wikipedia contains encyclopaedic knowledge, where named entities and events may be predominant over words (i.e., Battle of Trafalgar, London Bridge, Museo del Prado, etc.). With these findings, the next stage is to evaluate the proposed approaches on the disambiguation task.

For evaluating the **disambiguation stage**, several similarity metrics are computed. Here, the concern is with how correct the definitions obtained for the words are, and for this, experiments were made with lexical and semantic measures. Lexical similarity metrics were computed using the SimMetrics library (SimMetrics Library, 2014). Using this library, Cosine Similarity and QGramDistance were applied. Cosine similarity was computed because it is a popular and widely-used metric in the LT research area, and QGramDistance was computed as there is research showing that it is an effective similarity metric (Fernández et al., 2012). The limitation when computing only lexical similarity between a candidate definition and a gold-standard one is that if they do not contain the same words, similarity values will be very low, even though they may be expressing the same idea. In light of this, the lexical-semantic similarity metric proposed in (Chávez et al., 2014)[39] was used to test the approaches.

The average similarity results obtained for each approach are shown in Table 31 (the higher the values, the better as it indicates greater similarity).

Approach	Lexical similarity		Lexical-semantic similarity
	Cosine similarity	QGramDistance	(Chávez et al., 2014)
MFS-Freeling	25%	26%	51%
Wikipedia-baseline	14%	12%	36%
Wikipedia-MFS	13.5%	11%	35%
Wiktionary-baseline	20%	20%	51%
Wiktionary-MFS	17%	18%	48%

Table 31. Comparison similarity results for the disambiguation approaches

As can be seen from the results in Table 31, the lexical semantic similarity methods obtain better results than the lexical ones. This is as expected, since these methods are not only taking into account the words appearing in the definitions, but also semantic relationships, such as synonymy, that links multiple definitions.

It was concluded, on the basis of the experiments, that both the MFS-Freeling and the Wiktionary-baseline are the two disambiguation methods that perform best, taking into account both the lexical and the lexical-semantic similarity metrics.

iii) Wikipedia and Wiktionary are not as specific as WordNet, and therefore, the explanations/definitions/ synonyms provided may be easier to understand, and thus more useful for the ASD population

Another problem encountered when using WordNet and its related resources to provide assistive elements was the complexity associated with the definitions. In the particular context of ASD, the definition provided should be easy to understand; otherwise it will not be useful for end users. In light of the feedback obtained from the initial testing of the Meaning Disambiguator integrated in the Open Book tool (i.e., the outcome of

[39] This metric was ranked as the best similarity metric for Spanish in *SemEval 2014, task 10: Multilingual Semantic Textual Similarity*.

the project), apart from analysing a suitable approach for disambiguating a word in a specific context using Wikipedia and Wiktionary, a comparative analysis of the readability results for the different definitions based on the experiments performed in *ii)* was conducted. In this case, the readability of the definitions provided by each of the methods was assessed, thus evaluating the **resolution stage**. The readability metric employed to assess the quality of the definitions was the Fernández-Huerta Formula (Fernández Huerta, 1959). This is a Spanish readability metric based on Flesh Reading Ease. This index is computed as:

Readability Index = 206.84 -0.60 * (syllables per 100 words) - 1.02 * (number of sentences per 100 words)

For the interpretation of the results given, higher values of the index indicate easier texts, whereas lower values indicate more complex texts. Specifically, the scale shown in Table 32 was designed for interpreting this index:

Readability index	Difficulty
90-100	Very Easy
80-90	Easy
70-80	Fairly Easy
60-70	Normal
50-60	Fairly Difficult
30-50	Difficulty
0-30	Very Difficult

Table 32. Interpretation of the readability index (Fernández-Huerta, 1959)

In this experiment, for each of the obstacles analysed, the average readability value for the definitions provided by each of the disambiguation approaches and the definitions provided in the gold-standard corpus was computed. Table 33 shows the results obtained.

Approach	Our approach
Gold-standard	68.03
MFS-Freeling	56.31
Wikipedia-baseline	73.24
Wikipedia-MFS	68.81
Wiktionary-baseline	69.75
Wiktionary-MFS	65.97

Table 33. Readability analysis over the definitions provided in each approach.

According to the Fernández-Huerta readability index, the definitions provided in the gold-standard are *Normal*. As can be seen in Table 33, the definitions provided by MFS-Freeling, which implies the use of WordNet related resources, are the most difficult ones from the readability point of view (56.31, meaning *"fairly difficult"*). This coincides with intuition, as WordNet and its related resources for other languages were specifically developed for LT research, and not for education or text accessibility. Here, the best readability results is obtained for Wikipedia-MFS definitions, achieving 73.24 (*"fairly easy"*). It is also worth mentioning that the Wiktionary definitions are in line with the ones in the gold-standard corpus (*"normal"*). Another interesting observation is that for none of the resources, the definitions obtained are *"easy"* or *"very easy"*. This is explained by the fact that none of the analysed resources was developed taking into account reading comprehension problems and/or difficulties.

From all the experiments performed, and having analysed different resources, it can be concluded that Wiktionary is the most appropriate resource to be used to detect and resolve potential semantic obstacles in input text, that include polysemy, mental verbs, less common words, and specialised slang. Finally, to conclude this analysis, it is important to recap the main reasons that this resource can be more useful than others:

i) it is multi-lingual, available for a large number of languages (including Bulgarian, English, and Spanish);

ii) it has good coverage in terms of the words contained. From our analysis, this resource contained almost 80% of the obstacles detected in a document, compared to only 35% that could be resolved using MultiWordNet;

iii) it performs well with respect to the disambiguation of words, achieving the same results as using MultiWordNet, and being in line with the state of the art for this task in the LT research community (around 50%); and

iv) the definitions provided as assistive elements are easier to understand, obtaining a readability score around 70, which classifies the definition complexity between *"normal"* and *"fairly easy"*. This is a very good readability score, considering that the best definitions in this aspect were Wikipedia definitions with a readability score of 73.

Wiktionary is the most appropriate resource for use in assistive LT for end users of OpenBook. However, it was not integrated into the current system because it lacks a standardised and fixed structure for the internal format of definitions. This is a problem for developing the parsing process to extract the information in the definitions, since it cannot be generalised and there are a wide range of different cases. Moreover, the number of possible formats for the definitions is different and varied for other languages. These limitations are not visible through the online Website that shows the definitions of the words correctly; however from the LT perspective, a manual revision for all the words contained in Wiktionary should be done for the three languages in order to take into account all the possible cases, and this is a very time-consuming and costly task.

Given this limitation and the problems encountered when dealing with Wiktionary definitions, the most appropriate disambiguation method to be integrated in Open Book is based on WordNet, and its related resources for Bulgarian, English, and Spanish, (for users and carers). However, as an extra functionality, a dictionary module based on Wiktionary that implements the Wiktionary-baseline by extracting the first definition of the query word is also integrated in the interface for users, with the caveat that the information shown to the user may need post-editing.

3.2.4.3 General evaluation of the detection and resolution of infrequent slang

Quantitative evaluation of TENOR (Mosquera and Moreda, 2012; Mosquera et al., 2012a) was conducted for English and Spanish over an informal corpus of tweets. Table 34 reports the results obtained for detection and resolution.

Language	Corpora	PROCESS	RESULTS %		
			Recall	Precision	F1
English	Han’s Twitter dataset	Detection	95.2	91.7	93.4
		Resolution	91.2	74.5	82.1
Spanish	A hand-annotated corpus of 1000 Tweets texts	Detection	82.7	98	89.7
		Resolution	96.1	73	83

Table 34: Results for infrequent slang suing TENOR as LT tool

State-of-the-art performance is at the level of 71% (F1 score) for this task (Han and Baldwin, 2011). The results obtained using the approach integrated with OpenBook compare favorably with this.

3.2.4.4 General evaluation of the detection and resolution of acronyms/abbreviations

In the annotated corpora used for this evaluation (reported in Deliverable D4.1 in the Section on "Available and Annotated Corpora for Evaluation"), a total of 47 acronyms/abbreviations were detected and explained/expanded (e.g. *IVA*. (English: *VAT*)). Testing focused on detection and expansion of a subset of 18 of these acronyms/abbreviations. Recall, precision and F1 scores were calculated (Table 35).

Language	Corpora	PROCESS	RESULTS %		
			Recall	Precision	F1
Spanish	Subcorpus of newswire text	Detection	38.3	100	55.39
		Resolution	38.3	100	55.39

Table 35: Results for acronyms/abbreviations

The same figures are reported for detection and resolution of these obstacles because the lexicon of acronyms contains the expanded form of each acronym. The low level of recall indicates that the lexicon of acronyms/abbreviations employed needs to be expanded with a greater number of acronyms. Although it was not possible to perform the same evaluation for Bulgarian and English, the results could be extrapolated by consulting the English and Bulgarian dictionary of acronyms developed within the project.

3.2.4.5 General evaluation of the detection and resolution of the Meaning Disambiguator v2.1 with Science documents

The objective of this evaluation is to analyse whether adapting the Meaning Disambiguator to the Science domain, when we know that an input text belongs to that category, could improve the accuracy of the definitions returned for the semantic obstacles detected.

When testing the Meaning Disambiguation in the Science Domain, a set of experiments with Spanish texts were performed, using the documents in the annotated corpus developed for the project (see Deliverable D4.1: "Available and Annotated Corpora for Evaluation").

Specifically, 7 documents of the corpus belonged to the Science domain, and the processing of 114 obstacles was evaluated. The definitions obtained using the word sense disambiguation approaches Wiktionary-baseline and MFS-Freeling were evaluated against the gold-standard annotations in the corpus (D4.1: "Available and Annotated Corpora for Evaluation"). This evaluation enabled determination of whether focusing on a particular domain would lead to improved performance with respect to the MFS-Freeling method.

The results obtained are shown in Table 36. In this table, we distinguish three cases:

i) the case in which both methods have not disambiguated correctly (*wrong WSD*);

ii) the case in which the disambiguation using MFS-Freeling performs well, and therefore MultiWordNet resource is a good resource (*MFS-FreelingOK*); and

iii) the case in which the Wiktionary-baseline method has provided accurate disambiguation, and therefore, Wiktionary has proven to be a good resource for disambiguation (*Wiktionary-baseOK*).

Lang.	Corpus	Docs	Annotated obstacles	Analysed obstacles	RESULTS				
					Wrong WSD	MFS-FreelingOK		Wiktionary-baseOK	
ES	Science documents of the Subcorpus of newswire text	7	916	114	39.5%	Rec	25%	Rec	43.9%
						Prec	48.3%	Prec	43.9%
						F1	**33%**	F1	**44%**

Table 36: Results for the Science domain

Generally speaking, the results indicate that in 39.5% of the cases (aprox. 40%) none of the resources analysed (MultiWordNet and Wiktionary) disambiguates correctly. Considering each of these resources independently, recall, precision and F1 score were also computed. As can be seen from the results obtained, MultiWordNet disambiguates slightly better than Wiktionary; however, it offers very limited coverage, being able to resolve only the 25% of the obstacles. Wiktionary performs equally well as far as recall and precision are concerned. This is explained by the fact that Wiktionary provides much more coverage in words than MultiWordNet, and for most of the words, at least one definition is available. The improvement achieved using the Wiktionary-based method with regard to F1 is mainly due to the improvements in coverage provided by this resource over MultiWordNet.

Analysing in more detail the words and definitions included in the 40% that is wrongly disambiguated by both methods, it was found that most of these errors occur because, even though the document belongs to the Science domain, not all the words that it contains do necessarily belong to the Science domain, so extracting the sense belonging to the Science domain for all the words is not always appropriate. For instance, for the word "*análisis*" (English: *analysis*), in one of the testing documents, this term referred to *the process of investigation of the parts of a set* (Spanish: investigación de las partes componentes de un conjunto y de sus vínculos en la composición del todo). In this case, the MFS-Freeling method using the most frequent sense for the word "analysis" in MultWordNet worked well, obtaining a similar definition in the same line. However, when using Wiktionary for the disambiguation process, since it is assumed that the domain of the text is Science, the definition extracted was the analysis referring to text in the medical domain (Spanish: *5: (Medicina) Examen químico o bacteriológico de los humores o tejidos con un fin diagnóstico.*). The disambiguation task is still one of the most challenging tasks in the LT area, and it is very difficult for an automatic method to inspect word by word in a text to guess whether the word in this text is acting as a word belonging to the domain or with a general meaning. A possible approach to dealing with this limitation could be to establish a context window of *N* words before and after the obstacle, and analyse the domain of the words in this context. In this way, it would be possible to be more accurate when selecting the Wiktionary Science category from which the definition is retrieved. It should be noted that current word sense disambiguation systems have not yet resolved this challenge.

After all the experiments and tests performed, and given the impossibility of determining the specific cases, within the same text, in which a word may be used with the most frequent sense (most of the words in a text) or, less frequently, with a domain-specific sense (when the document belongs to a domain, and the word is used for that domain), **the main conclusion of this evaluation,** is that the MFS approach should be exploited in Open Book using WordNet and its extensions (MultiWordNet and Bulgarian WordNet). This is motivated by the following issues:

i) the noise (incorrect disambiguation and definitions) that would be introduced when using the Science-adapted approach, and especially if the domain of the text is not known in advance; and

ii) the lack of standards for encoding the Wiktionary definitions that will pose challenges for the resolution of the obstacles detected.

It is also important to note that it was not possible to find any other appropriate resources for the Science domain specifically addressed for users with ASD. Moreover, the results obtained for meaning disambiguation in Open Book could be improved upon if approaches different from the most frequent sense approach could be applied. Currently this is not possible due to the current computational "glass ceiling" for the word sense disambiguation task[40].

3.2.4 Conclusions of Evaluation of LT

In processing structural complexity in Bulgarian, a syntactic parser was used in combination with rules to identify less important constituents for deletion and positions at which long sentences could be split into shorter sentences. The rules were evaluated with respect to a modest set of test data (82 sentences) and were found to perform at a level that compared favourably with those exploited by the English structural complexity processor. At present the processing of Bulgarian sentences is limited to a relatively narrow range of syntactic constructions. In future work it would be useful to explore the possibility of supplementing this module with additional methods addressing a wider range of complexities identified via the Bulgarian syntactic parser.

For the processing of English, evaluation focused on a new method based on machine learning to identify relatively unimportant subordinate clauses and appositions in complex sentences for the purpose of sentence rewriting. The approach was found to be superior in terms of both precision and recall to the rule-based method for processing complex sentences presented in Deliverable D3.1. The updated method is limited to identification of subordinate clauses. Future work will focus on the development of conversion rules to exploiting this method for the purpose of rewriting complex sentences by both deleting embedded clauses and appositions and also generating new declarative sentences linking the subordinate elements to the phrases that they modify.

[40] http://alt.qcri.org/semeval2014/

Updated evaluation of the meaning disambiguator revealed that, in line with the findings reported in D7.5, the component of OpenBook performing coreference resolution performs well in comparison with state-of-the-art systems.

The most suitable approach to WSD was found to exploit the “most frequent sense” in Freeling and a Wiktionary baseline. The performance of these methods still obtained results only just over 51%. Attempts to adapt the *meaning disambiguator* to texts from the science domain revealed that the most effective approach to word sense disambiguation is based on the *most frequent sense* method, exploiting lexical resources such as WordNet, MultiWordNet, and the Bulgarian WordNet.

Definitions provided by Wikipedia are most readable (Flesch Reading Ease of ~73 “Fairly easy”). This resource also provides good coverage of concepts and would make Wiktionary the idea source of definitions in OpenBook. However, its lack of a standardised and fixed structure are limitations that prompted UA to use WordNet definitions instead, despite their poor coverage.

With regard to the detection and resolution of slang in EN and ES text, the LT service integrated in OpenBook performs at a level that compares favourably with the state of the art. The LT performing detection and resolution of acronyms and abbreviations in ES text has F1 score of 55.39, with excellent precision but recall at 38.3. This implies that additional lexical resources should be acquired or developed in order to improve on this level of performance.

4. Conclusions

Evaluation of the text conversion services supported by OpenBook was made using extrinsic and intrinsic approaches.

OpenBook was used by carers to provide a semi-automatic service aimed at converting texts into a more accessible form for end users. Extrinsic evaluation assesses the extent to which this service improves comprehension of converted texts by people with ASD. OpenBook can also be exploited directly by end users who can freely apply different LT services in order to improve their comprehension of texts. In this context, the service provided is automatic. Extrinsic evaluation includes an analysis of the transcripts of interviews with end users and carers who have exploited the automatic service.

Reading comprehension testing involving 293 participants showed that texts converted by carers using OpenBook were understood better than texts in their original form. Further, participants rated questions based on converted texts to be easier than questions based on texts in their original form. No statistically significant difference was observed in the time taken to answer the two different types of test item. There was also no correlation between participants' subjective opinions of the ease of particular test items and the reading comprehension score obtained. This serves to indicate that participants did not recognise that items based on converted texts were easier to answer than items based on the originals. It was also found that participants' test scores correlated with age, gender, psychiatric diagnosis of autism, IQ, education level, and marital status. No correlation was found between reading comprehension test scores and employment status.

Feedback from end users enabled evaluation of OpenBook as an automatic service for use by people with ASD to facilitate better understanding of texts that they are reading. End users appreciated the ease of use of the interface and there was enthusiasm on the part of all participants for the concept of software providing support for reading comprehension. Functions providing users with the ability to retrieve explanations of complex words and idioms, to retrieve images to explain those concepts, and to generate succinct summaries of input texts were all appreciated by end users and carers using OpenBook to convert texts into a more accessible form. There was criticism of the inaccuracy of these functions and the poor handling/lack of coverage of some domain specific terms. Users also criticised the slow speed of processing offered by the system. End users and carers made many suggestions for improvements of OpenBook, suggesting that the software be adapted for use with mobile devices, and changed to exploit different lexical resources such as children's dictionaries which may offer more accessible definitions of difficult terms and dictionaries for foreign loan words occurring in texts.

Overall, users stated that they would recommend the use of OpenBook to others. All users stated that they are enthusiastic about using the system independently, that they will continue to use OpenBook in the future, and that they have become more independent and keen to solve problems for themselves since starting to use OpenBook.

Users did observe that OpenBook would benefit from additional improvements (e.g. in accuracy and coverage of the LT services) to help reduce the amount of manual effort required to convert a text into a more accessible form. Carers said that they benefited from the ability to retrieve definitions of complex terms and that this function reduced the burden on them. Many users said that retrieving such definitions, constituted the main way in which they used OpenBook.

Interviews conducted with 18 participants, including carers and end users provided evidence of benefits to inclusion brought by the use of OpenBook to improve the accessibility of texts. It was reported that improvements in users' ability to comprehend texts included improvements in their understanding of subtext. Overall, there was greater motivation to read, greater engagement in reading, and improved attention in reading as well as improvements in vocabulary. Improvements were also noted in the writing skills of both young and mature end users.

With regard to communication, carers noted that children became more confident and willing to initiate conversations during their period of use of OpenBook. Adults with ASD reported enhancement of their spoken abilities, including the use of more extensive vocabulary and more complex phrases/sentences. When considering social interaction, children were reported to engage more frequently with their carers and with their peers. The quality of these interactions was consistently reported to have been enhanced. In the context of employment, one adult with ASD reported becoming less reliant on others, and better able to manage their workload through more frequent use of written communication with colleagues. The increased independence of OpenBook users was a recurring theme throughout the interviews.

In some cases, carers hypothesised that improvements in text comprehension resulting from the use of OpenBook had a positive impact on the behaviour of end users. For example, a child who formerly became angry at his inability to understand what his mother was saying became less likely to do so. End users of school age were also noted to become less shy and more sociable. Carers identified the use of OpenBook as a potential factor causing this change. They also noted that users became more self-sufficient and consequently self-confident during their period of use of OpenBook. Children were also reported to study for longer periods of time, and more effectively, with improvements noted in their school grades.

Intrinsic evaluation focused on the contribution brought by OpenBook as an editing tool for use by carers converting texts to a more accessible form and on the quality of the LT services integrated into the system.

As an editing tool, it was found that use OpenBook enabled more rapid conversion of text to a form that was not significantly different in terms of accessibility than those created by unaided carers. A study of the use of OpenBook by 16 carers (including teachers, psychologists, speech and language therapists, and psychiatrists) converting 75 texts (25 in each of BG, EN, and ES) demonstrated that the average time required to make the conversion was just 29 minutes per text, which compares favourably with the 54 minutes required per text for unaided conversion. Assessment of readability showed that converting texts using OpenBook generated texts that were not significantly less readable than those generated as a result of unaided conversion. Regardless of the method used, conversion of the texts generated versions that were, by most indices, more readable than the originals.

For detailed evaluation of the LT services developed in the FIRST project, the reader is directed to Deliverable D7.5 (Evans et al., 2013). The updated evaluation included in the current report focuses on the *structural complexity processor* developed for use in converting Bulgarian and English sentences into a more accessible form. The updated evaluation also provides new perspectives on the *meaning disambiguator* exploited by the system.

Overall, the findings of the current evaluation are in line with those detailed in D7.5. That is, many of the components integrated in OpenBook do not perform with sufficient accuracy to support fully automatic conversion of text into a more accessible form for people with autism. Instead, they are better exploited in a semi-automatic conversion process, in which an interface is provided to facilitate post-editing of system output by intermediaries (carers). In light of this, OpenBook provides two interfaces. The first is for end users, which provides access to a limited set of reliable LT components that can be used automatically. The second is for carers, which provides access to the full set of LT components together with post-editing functions.

Intrinsic evaluation of the LT services integrated in the OpenBook system and feedback from carers and end users indicate several directions for future work. These include:

- improvement of the accuracy of LT services
- development of methods for automatic domain adaptation and selection of appropriate semantic resources,
- use of speech synthesis and speech recognition technologies,
- development of an improved interface for mobile devices, and
- optimisation of OpenBook in domains of interest to users to enable better comprehension of correspondence with health services, government, utilities companies, banks, etc.

5. References

Iustin Dornescu, Richard Evans, and Constantin Orasan. 2013. *A Tagging Approach to Identify Complex Constituents for Text Simplification*. In Galia Angelova, Kalina Bontcheva, and Ruslan Mitkov, editors, Proceedings of Recent Advances in Natural Language Processing, RANLP'13, pages 221 – 229, Hissar, Bulgaria. RANLP 2011 Organising Committee / ACL.

Iustin Dornescu, Richard Evans, and Constantin Orasan. 2014. *Relative clause extraction for syntactic simplification*. In Proceedings of the Workshop on Automatic Text Simplification – Methods and Applications in the Multilingual Society (ATS-MA 2014), pages 1 – 10, Dublin, Ireland. ACL.

Richard Evans and Constantin Orasan. 2013. *Annotating signs of syntactic complexity to support sentence simplification*. In Ivan Habernal and Vclav Matouˇsek, editors, Text, Speech, and Dialogue, volume 8082 of Lecture Notes in Computer Science, pages 92–104. Springer Berlin Heidelberg.

Richard Evans, Iustin Dornescu, and Mijail Kabadjov. 2013. *FIRST Project - Evaluation of Language Technology: Error Analysis*. University of Wolverhampton, UK. FIRST_D7.5_20130930.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. In Proceedings of the 18th International Conference on Machine Learning, pages 282–289.

Advaith Siddharthan. 2002. *Resolving Attachment and Clause Boundary Ambiguities for Simplifying Relative Clause Constructs*. Association for Computational Linguistics Student Research Workshop, pages 60–65.

Pontus Stenetorp, Sampo Pyysalo, Goran Topi´c, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. *Brat: A web-based tool for nlp-assisted text annotation*. In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, EACL '12, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

6. Appendices

Appendix 1

Scoring sheet

R Code

Test order												
Text Nr	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**

Text 1 ***Participant Rating***

Question 1	D x	C
Question 2	A x	B
Question 3	B x	A
Question 4	A x	B
Question 5	B √	B
Question 6	B √	B
Total T1	2	

Participant Rating: 4

Text 2 ***Participant Rating***

Question 1		B
Question 2		C
Question 3		C
Question 4		D
Question 5		F
Question 6		B
Total T2		

Text 3 ***Participant Rating***

Question 1	B √	B
Question 2	C x	B
Question 3	A √	A
Question 4	C √	C
Question 5	D x	A
Question 6	B x	A
Total T3	3	

5

Text 4			***Participant Rating***
Question 1		A (True)	
Question 2		B (False)	
Question 3		B (False)	
Question 4		D (Tadpole to frog)	
Question 5		B (Aqua-Light)	
Question 6		B (Swim-a-long)	
Total T4			

Text 5			***Participant Rating***
Question 1		C (to prepare students for University)	
Question 2		B (computers)	
Question 3		A (writing summaries)	
Question 4		B completion of assignments,...)	
Question 5		C (manage their time well)	
Question 6		D (Upper-intermediate)	
Total T5			

Text 6			***Participant Rating***
Question 1	B x	A (Ptolemy)	3
Question 2	F √	F (Newton)	
Question 3	C √	C (William Shakespeare)	
Question 4	E √	E (Bruno)	
Question 5	D x	B (George Rheticus)	
Question 6	B x	D (Galileo)	
Total T6	3		

Total Original Score	**8**	***Total Participant Rating***	**12**

Text 7 (1 Simplified)			***Participant Rating***
Question 1		C	
Question 2		B	
Question 3		A	
Question 4		B	
Question 5		B	
Question 6		B	

Total T7

Text 8 (2 Simplified)

Question 1	B √	B
Question 2	C √	C
Question 3	C √	C
Question 4	D √	D
Question 5	A x	F
Question 6	A x	B
Total T8	4	

Participant Rating

1

Text 9 (3 Simplified)

Question 1		B
Question 2		B
Question 3		A
Question 4		C
Question 5		A
Question 6		A
Total T9		

Participant Rating

Text 10 (4 Simplified)

Question 1	B x	A (True)
Question 2	B √	B (False)
Question 3	B √	B (False)
Question 4	D √	D (Tadpole to frog)
Question 5	B √	B (Aqua-Light)
Question 6	B √	B (Swim-a-long)
Total T10	6	

Participant Rating

1

Text 11 (5 Simplified)

Question 1	A x	C (to prepare students for University)
Question 2	B √	B (computers)
Question 3	A √	A (writing summaries)
Question 4	B √	B completion of assignments,...)
	C √	

Participant Rating

2

Question 5		C (manage their time well)
Question 6	A x	D (Upper-intermediate)
Total T11	4	

Text 12 (6 Simplified)			***Participant Rating***
Question 1		A (Ptolemy)	
Question 2		F (Newton)	
Question 3		C (William Shakespeare)	
Question 4		E (Bruno)	
Question 5		B (George Rheticus)	
Question 6		D (Galileo)	
Total T12			

Total Simplified Score 14 ***Total Participant Rating*** 4

Appendix 2

Text Rating

How did you find the first text?

Please cross one box only

Very easy	Easy	OK	Difficult	Very difficult

How did you find the second text?

Please cross one box only

Very easy	Easy	OK	Difficult	Very difficult

How did you find the third text?

Please cross one box only

Very easy	Easy	OK	Difficult	Very difficult

How did you find the fourth text?

Please cross one box only

□ Very easy □ Easy □ OK □ Difficult □ Very difficult

How did you find the fifth text?

Please cross one box only

□ Very easy □ Easy □ OK □ Difficult □ Very difficult

How did you find the sixth text?

Please cross one box only

□ Very easy □ Easy □ OK □ Difficult □ Very difficult