Language Technologies applied to Document Simplification for Helping Autistic People

Eduard Barbu[a], M. Teresa Martín-Valdivia[a], Eugenio Martínez-Cámara[a,*], L. Alfonso Ureña-López[a]

[a]SINAI Research Group,
University of Jaén
Campus Las Lagunillas E-23071, Jaén (Spain)

Abstract

People affected by Autism Spectrum Disorders (ASD) have impairments in social interaction because they lack an adequate theory of mind. A significant percentile has inadequate reading comprehension skills. We present a multilingual tool called Open Book (OB) that applies Human Language Technologies (HLT) in order to identify reading comprehension obstacles in text documents and propose more simple alternatives with the aim of assisting the reading comprehension of users. OB involves several text transformations at lexical, syntactic and semantic level. In this paper we focus on three challenging components of the OB tool: the image retrieval component, the idiom detection component and the summarization module. There are very few studies that involve simplification by showing images associated to difficult concepts. In addition, the treatment of figurative language such as idioms or metaphors is one of the most challenging areas in Natural Language Processing (NLP). Finally, although text summarization is a more widely studied field in NLP, its application to text simplification remains as an open research issue. Thus, we focus on the integration of these three modules in our OB tool. We present the motivation for building these components and we describe how they are integrated in the whole system. Moreover, the usability and the usefulness of OB have been evaluated and analysed showing that the tool helps to produce texts easier to understand for autistic people.

*Corresponding author

Email addresses: eduard_barbu@yahoo.com (Eduard Barbu), maite@ujaen.es (M. Teresa Martín-Valdivia), emcamara@ujaen.es (Eugenio Martínez-Cámara), laurena@ujaen.es (L. Alfonso Ureña-López)

Keywords: Natural Language Processing, Text simplification, ASD, Image retrieval, Text Summarization, Topic Models, Idiom detection

1. Introduction

Autism Spectrum Disorder (ASD) is a condition that impairs the proper development of cognitive functions, social skills, and communicative abilities (Mesibov et al., 1997). It was shown in various studies (Nation et al., 2006) that individuals affected by ASD have deficits in reading comprehension.

Other psychological studies have showed that ASD children's lexical and syntactic knowledge remains delayed with respect to the other cognitive functions (Lord and Paul, 1997) and that pictures can improve the reading comprehension skills of children with a wide developmental disabilities spectrum (Fossett and Mirenda, 2006).

It is already proven by the significant amount of dedicated computer applications that information technology can enhance the communicative abilities of the people with ASD (Ploog et al., 2013; Mintz, 2013). Computer applications can also help people with ASD in other facet of their lifes. Chu et al. (2011) expose a framework to support problem-oriented e-learning, in plain English, a methodology to develop e-learning platforms focus on adaptive case-based learning, which is highly important to cover the specific learning needs of each person with ASD. Moreover, technology can assist people with ASD in other of their common symptoms: repetitive and stereotyped behaviors. In (Coronato et al., 2014) is described a method and infrastructure for the detection of the stereotyped motion disorders of people with ASD. The main goal of the system is the study of the stereotyped movements with the aim of reducing them.

This paper presents several components of a new software tool, called Open Book, developed to assist ASD carers to transform written documents into a format that is easier to read and understand using HLT.

The Open Book tool is being developed under the on-going European Project FIRST (Flexible Interactive Reading Support Tool). The project involves nine institutions from four countries (Spain, Bulgaria, UK and Belgium) and includes experts in language technologies, software development and autism. Currently, the Open Book tool processes documents written in English, Spanish and Bulgarian but can be extended to other languages provided that the required resources are available. To accommodate the variability of ASD users the tool functionality can be customized to fit their needs.

Based on literature research and on a series of studies performed in the United Kingdom, Spain and Bulgaria with a variety of autistic patients ranging from children to adults, a series of obstacles in reading comprehensions have been identified. From a linguistic point of view, they can be classified in lexical obstacles (difficulty in processing relative clauses, for example) and semantic obstacles (difficulty in understanding rare or specialized terms or in comprehension of idioms, for example). The tool applies a series of automatic transformations to user documents to identify and remove the reading obstacles to comprehension. These transformations include: the replacement of long and complex sentences with several simpler ones, identification of difficult terms and image retrieval for difficult terms, generation of concise summaries and automatic generation of tables of contents to help the navigation through the text.

In this paper, we focus on three specific challenging issues related to text simplification using Natural Language Processing (NLP). The first one is the retrieval of images associated with difficult or complex concepts detected in a document. Secondly, a module of idiom detection is presented. This feature is an open research issue in NLP and, to the best of our knowledge, there are no existing studies related to integrating idiom detection into a simplification system. Finally, we study two different approaches to text summarization based on topic model and the PageRank algorithm. The final evaluation demonstrates the usefulness of our tool and points out the advantages obtained by integrating the different modules. Of course we have also found some weaknesses in our systems that should be improved, most of them related to the imprecision of the integrated modules (images are in some situations inappropriate, the idiom detection is in some cases inaccurate, etc.). Thus, more research must be carried out in order to overcome the different issues. For example, improvement of the disambiguation module could imply greater accuracy in other related components such as the image retrieval or idiom detection modules.

Before going further we make some clarifications about the terminology. As we said before, a multidisciplinary team of experts is developing the Open Book tool. Throughout this paper we will refer to the clinicians and the specialists in psychology that contributed to the tool as *the psychologists*. Moreover, the partners from academy and industry that programmed the tool will be referred as *the technicians*. The people with ASD that benefit from using the tool will be denoted whenever possible as *the users*.

The rest of the paper is organized as follows: the next section includes a review of the state of the art for NLP and Automatic Text Simplification. In addition, we introduce some important software tools designed to help people with ASD and

we show how Open Book is different. Next, we indicate how the obstacles in user comprehension have been identified. Then the general Open Book architecture is presented and the different components of the system are shown. Finally an evaluation of the accuracy is presented. We end the paper with the conclusions.

2. Related Work

In this section, we first present some interesting works related to NLP and Text Simplification (TS). On the other side, we will comment the main current software tools available to help people with ASD.

2.1. NLP and Text Simplification

Text simplification is a NLP task that tries to reduce the linguistic complexity of a text while still retaining the original meaning. The process can deal with many different linguistic issues depending on whether the simplification is at lexical, syntactic, or discourse level. There is a large body of literature investigating the role of simplified text in relation to different kinds of users like, for example, second language learners (Tweissi, 1998) or people with some language deficits such as deafness (Luckner and Handley, 2008) or dyslexia (Rello et al., 2014).

It is clear that TS can be very useful for many automatic NLP applications (i.e., machine translation or information retrieval) and also for human readers, especially for people with language disabilities (i.e., autism or dislexia). However, the process of manual TS is highly costly and time consuming, and Automatic Text Simplification (ATS) has not received much attention until the last decade. Perhaps the work of (Chandrasekar et al., 1996) can be considered the first serious approach to the ATS problem. They provide a formalism of syntactic simplification based on dependency-trees, motivated by the need to improve the performance of a parser. From this research, the number of papers and projects related to TS has grown significantly. In fact, there are several good survey papers that summarize the state of the art and introduce the field (Feng, 2008; Siddharthan, 2014; Shardlow, 2014). Currently, one of the main focuses of interest is the generation of resources like the Simple English Wikipedia project (an alternative to English Wikipedia), which provides simplified versions of Wikipedia articles. There are over 88,000 articles which have been hand written in Simple English for this project. However, this project is only oriented to English although some researchers wish to extend it to other languages.

Regarding the Open Book tool, it addresses several simplification features related to NLP oriented towards helping people with autism, including features at

lexical, syntactic and semantic level. Other previous studies have treated these issues with more or less success. Thus, lexical simplification usually implies replacing difficult words with simpler synonyms (Devlin and Tait, 1998), images (Devlin and Unthank, 2006), or definitions from a dictionary (Kaji et al., 2002). For syntactic simplification several NLP technologies are applied such as parsing (Chandrasekar et al., 1996), converting passive voice to active (Canning, 2002), Part-of-Speech (PoS) and chucking (Siddharthan, 2003)... Finally, semantic simplification is considered a more complex task that involves anaphora resolution (Siddharthan and Copestake, 2002) or pronouns replacement (Canning, 2002).

In this paper we will focus on three specific simplification features: retrieval of image associated to complex concepts, idiom detection and summarization. Although we can find some papers where difficult concepts are replaced by images (Devlin and Unthank, 2006) and summarization (Margarido et al., 2008; Bouayad-Agha et al., 2009), we have not found any reference that identifies and simplifies figurative language. One of our main contributions is nevertheless to consider a complete architecture designed to carry out the simplification of a document in a global way and not by considering each feature individually.

2.2. Software tools for ASD

A number of software tools have been developed to support the learning and to enhance the communicative skills of the people with ASD. We think there are three classes of tools that can help users. The applications in the first category are not specifically designed for our users but have some functions that could be useful to them. Sound-Beginnings[1] is a tool designed to properly teach vocalization techniques to young children and people with special needs. Idiom Track[2] teaches a limited number of idiomatic expressions graphically illustrating their literal and figurative meaning. People having semantic or pragmatic disorders[3] as well as people that are learning English as second language use this tool. The second category of tools helps the users to understand the basic emotions. The paradigmatic tool for this class is Mind Reading[4] developed by a team at the University of

[1] `http://www.toolfactory.com/products/page?id=20104`

[2] `http://www.toolfactory.com/products/page?id=20103`

[3] The people with semantic disorders have difficulties understanding the meaning of particular language expressions. The people with pragmatic disorders have difficulties using appropriately the language in social contexts. The semantic and pragmatic disorders affect a larger category of individuals and it is not restricted to people with ASD.

[4] `http://www.jkp.com/mindreading/`

Cambridge. It teaches human emotions using a library set of 412 basic emotions illustrated by images and video. Its library of emotions is the fruit of the developmental studies of emotion recognition (Baron-Cohen, 2001). From the same class of applications it is worth mentioning Autism 5[5], a software application developed for iPhone and iPad that helps the students to interact with their teachers.

The third class of applications is the one that helps to understand texts or to compose textual stories. For example, VAST-Autism[6] is a tool that supports the understanding of linguistic units: words, phrase and sentences by combining spoken language and images. Likewise, "Stories about me" is an iPad application that allows early learners to compose stories about themselves thus helping the user's social integration. Boardmaker[7] helps the users to understand and compose stories using symbols. Boardmaker makes use of a database of Picture Communication symbols available in 44 languages.

All these tools and others from the same categories are complementary to Open Book. Actually, these tools are restricted to pre-stored images and texts and are not able to accommodate new pieces of information. The main characteristic that sets aside Open Book is its scalability. It works with texts of any genres in multiple languages. From a technical point of view Open Book is the first developed tool that uses Natural Language Processing (NLP) to aid carers and ASD users.

3. Obstacles in reading comprehension

To identify the obstacles in text comprehension we rely on four lines of evidence. The first line of evidence comes from the studies performed with autistic population and reported in the literature. For example, Frith and Snowling (1983) show that ASD children can understand the meaning of single words but they have difficulty using semantic context to disambiguate the ambiguous words. O'Connor and Klein (2004) point out that replacing difficult pronouns with their referents improves the reading comprehension of the people with ASD. Oakhill and Yuill (1986) found that students with autism make errors in identifying pronoun referents, and that these errors increased with the complexity of sentences.

[5] https://itunes.apple.com/us/app/autism-5-point-scale-ep/id467303313?mt=8

[6] https://itunes.apple.com/us/app/vast-autism-1-core/id426041133?mt=8

[7] http://www.mayer-johnson.com/boardmaker-software/

In her landmark book, "Thinking in Pictures: My Life with Autism", Temple Grandin, a scientist affected by ASD, gives an inside testimony for the importance of pictures in the life of the people with ASD (Grandin, 1996):

> "Growing up, I learned to convert abstract ideas into pictures as a way to understand them. I visualized concepts such as peace or honesty with symbolic images. I thought of peace as a dove, an Indian peace pipe, or TV or newsreel footage of the signing of a peace agreement. Honesty was represented by an image of placing one's hand on the Bible in court. A news report describing a person returning a wallet with all the money in it provided a picture of honest behavior."

Grandin suggests that not only the people with ASD need images to understand abstract concepts but that most of their thought process is visual. In an autobiographic study Grandin narrates that she uses language to retrieve pictures from the memory in a way similar to an image retrieval system. Other studies document the importance of images in ASD: Kana et al. (2006) show that the people with ASD use mental imagery even for comprehension of low-level imagery sentences.

The second line of evidence comes from the experience of carers with the autistic people. The psychologists involved in the FIRST project have worked as carers or have conducted interviews with the carers. It goes without saying that the insights provided by professionals are at least as valuable as the results of the strictly conducted experiments.

The third line of evidence comes from the controlled experiments the psychologists conducted with the users through questionnaires. The tests have been simultaneously administered in an online modality as well as in the traditional paper and pencil in United Kingdom, Spain and Bulgaria.

The fourth line of evidence comes from an experiment performed for English and Spanish[8]. The carers in UK and Spain simplified 25 documents for each language. The genres of the documents range from rent contracts to newspaper articles and from children literature to health care advices. Then the technicians compared the original and simplified texts and registered the main operations the carers performed when they have re-written the texts. The main operations for English and Spanish are the following:

[8]Unfortunately the same experiment could not be carried on for Bulgarian language texts because the professional carers in Bulgaria are scarce. Moreover they are very hard to contact and many are not very familiar with the new technologies.

1. Synonymous. A noun or an adjective is replaced by its less complex synonym. For example in the input sentence "The proletarians protested for a wage increase." the word *proletarian* is replaced by its simple synonymous *worker*.
2. Sentence Splitting. A long sentence is split in shorter sentences or in a bullet point list to make its content visually easier to process. For example, the relatively long sentence: "When walking along the beautiful beaches of Rhodes, Papadopoulos gathers colored shells, baths in the warm water, admires the sunset and stops to drink in all bars along the coast." is transformed using bullet points in: "When walking along beautiful beaches of Rhodes Papadopoulos:
   a Baths in the warm water.
   b Admires the sunset.
   c Drinks in all bars along the coast."
3. Definition. A difficult term is explained using an appropriate dictionary or Wikipedia definition. The sentence "The key enlightenment thinkers were: Locke, Voltaire Kant and Rousseau." contains the potentially difficult term *enlightenment*. The carer explains the term with a definition inserted immediately after the sentence containing the difficult word.
4. Near Synonymous. A noun or an adjective is replaced by a near synonym. For example, the sentence: "The subpoena sentenced Mr. Jack Bush to a fine of 1,000,000 dollars." is transformed to "The court sentenced Mr. Jack Bush to a fine of 1,000,000 dollars." The near synonymous *court* replaces the difficult word *subpoena*. The word *court* is a good approximate meaning for the word *subpoena*: their meaning overlap but do not coincide.
5. Image. A concept in the text or a concept related to a paragraph is illustrated with an image. For example, the sentence "The ambulance arrived and helped the injured people." is illustrated with an image of an ambulance.
6. Explanation. A sentence is rewritten using different words. The sentence: "The committee has spent several years working on a voluminous report about the detention and interrogation program, and according to one official interviewed in recent days, C.I.A. officers went as far as gaining access to computer networks used by the committee to carry out its investigation." is rewritten as "The C.I.A officers improperly monitored the work of the committee gaining access to their computer networks.".
7. Deletion. Parts of the sentence are removed thus making the transformed sentence easier to read. The sentence "Immanuel Kant, 22 April 1724 –

12 February 1804, was a Prussian philosopher who is widely considered to be a central figure of modern philosophy." is transformed into "Kant is considered to be a central figure of philosophy."

8. Coreference. A coreference resolution operation is performed. In general a pronoun or a definite description is replaced by the entity it refers to. For example, the sentence "It is located on the Iberian Peninsula in southwestern Europe." is replaced with the sentence: "Spain is located on the Iberian Peninsula in southwestern Europe." It was clear from the context that the pronoun it was referring back to the concept *Spain*.
9. Syntactic Operation. A transformation on the syntactic parse trees is performed. As an example consider the transformation from passive to active sentences.
10. Figurative Language. An idiom or other figurative language expression is explained. In the sentence: "Microsoft will cease to be a 800 pound gorilla", the idiom *800 pound gorilla* is being explained as "the dominant force in industry".
11. Summarization. The content of a sentence, a paragraph or of the whole text is condensed.
12. Paragraph Splitting. A paragraph is split into several smaller size paragraphs.
13. Paragraph Joining. Two or more paragraphs are joined.

The only obstacle that cannot be tackled automatically is Explanation, because it entails the interpretation of the sentence or paragraph and cannot be reduced to simpler operations. The number of operations for the selected texts for English and Spanish are registered in Table 1

The operations are similar across languages except for three operations: the carers in UK did not use images to illustrate concepts while the carers in Spain did not split or joined paragraphs. The reticence of the UK carers to make use of images is due to a regulation that restricts or prohibits the use of images for certain kind of users.

Based on the four lines of evidence discussed above the psychologists compiled the initial user requirements. The technicians made then comments about the implementation feasibility of these requirements and composed the final list of requirements to be implemented in the Open Book tool. These are the following:

1. Simplification of the sentences from a syntactic point of view. The technicians identified those syntactic phenomena that can be accurately simpli-

Operation Name	Number of English Operations	Number of Spanish Operations
Synonymous	40	64
Sentence Splitting	70	40
Definition	7	34
Near Synonymous	21	33
Image	0	27
Explanation	43	24
Deletion	27	17
Coreference	3	17
Syntactic Operations	42	9
Figurative Language	12	9
Summarization	10	3
Paragraph splitting	7	0
Paragraph joining	4	0

Table 1: Simplification Operations for English and Spanish

fied. The sentence simplification usually means the re-writing of sentence into two or more simpler sentences.

2. Coreference resolution. This user requirement is identical with the point 8 above. It means the resolution of a pronoun or a definite description.
3. Identification of difficult terms. The difficult terms include: rare, polysemous, specialized terms, etc.
4. Illustration by images of identified terms.
5. Presenting to the user the most important information of the text.
6. Identification of the figurative language, specifically of the idioms.

4. Open Book Architecture

The components listed above are either syntactical (the first two) or semantic (the last four). Before presenting the semantic components we make some brief considerations about the architecture of Open Book.

The Open Book System is a web application that has a three-tier architecture (Figure 1). The first layer is composed by Natural Language Processing Components implemented as SOAP web services. These web services add information in a GATE[9] document (Cunningham et al., 2011) as separate annotation sets. The

[9] https://gate.ac.uk/

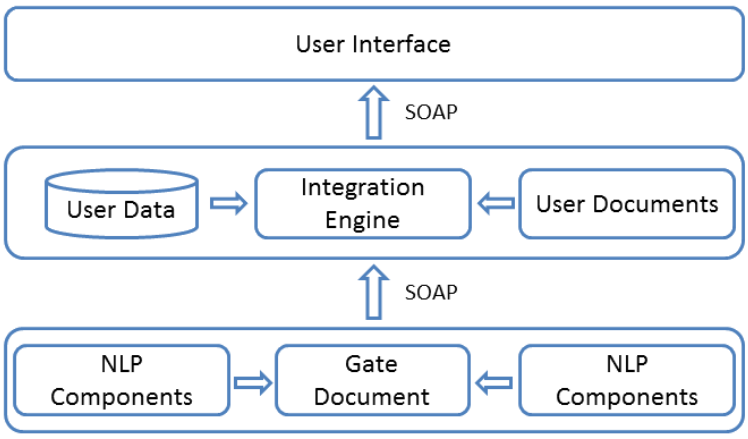

Figure 1: The architecture of the Open Book tool

second layer, called Integration Engine, combines the output of the NLP web services and manages the databases containing a representation of the users and their documents. The third layer is the User Interface (UI) that presents the information and interacts with the users.

The working flux starts with a text document loaded into the UI. The document is processed by the NLP web services and it is presented back to the carer or user who can accept or reject the suggestions of the automatic processing. The document is then saved and stored in the user library. In Figure 2 all NLP Processing components of the Open Book tool are illustrated. All the NLP components are executed in parallel with the exception of three components that should be executed serially: Disambiguation, Wikipedia Disambiguation and Offline Image Retrieval.

In what follows we present all semantic processing modules except for the module that identifies the difficult terms[10].

4.1. Image Retrieval

The Image Retrieval module is composed of two modules: the Offline Image Retrieval module that retrieves images for the automatically identified concepts and the Online Image Retrieval module that retrieves images for the concepts highlighted by the users. The architecture of the Image Retrieval Module is presented in Figure 3.

[10]The authors of this paper did not develop the module that identifies the difficult terms.

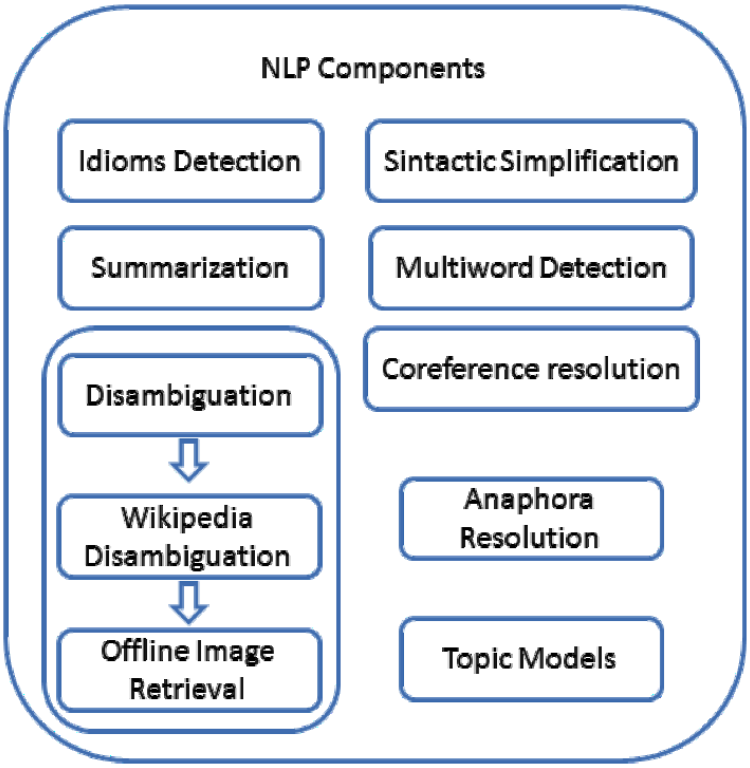

Figure 2: The NLP components of the Open Book tool

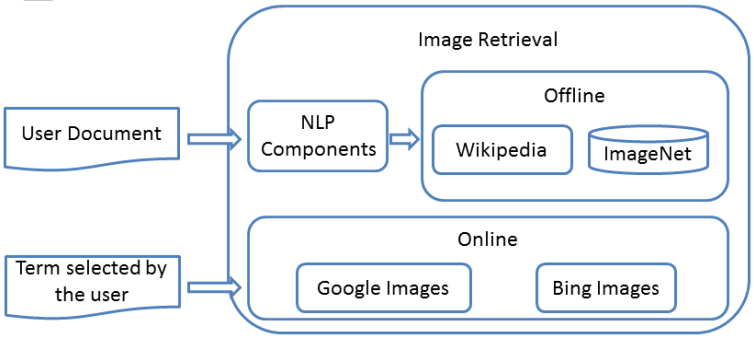

Figure 3: The architecture of the Image Retrieval Module

The difficult concepts that can be a problem for the user understanding are identified and disambiguated against Wordnet (Miller et al., 1990) by the Disambiguation component. The Offline Image Retrieval component extracts the corresponding images from two sources: ImageNet and Wikipedia.

ImageNet (Deng et al., 2009) is an image database that stores web images annotated with Wordnet noun synsets. At the time of writing of this article the ImageNet database links around 22000 synsets with more than 14 million images. The ImageNet is an accurate resource constructed in two steps. Exploiting the main web searching engines (Google, Yahoo) the images corresponding to synset words were automatically collected from the Web. Then the retrieved images were cleaned using the Amazon's Mechanical Turk service. The ImageNet database is wieldy used in various artificial intelligence tasks. For example the first large-scale image classification Deng et al. (2010) study that used more than 10000 Wordnet categories to classify around 9 million images made use of ImageNet. Open Book uses an internal version of ImageNet database that have been processed and indexed for fast access. For example all the URLs pointing to images[11] that no longer exist have been discarded. Moreover the synset identifiers are normalized to match the identifiers used by the Disambiguation module.

Wikipedia is another resource used to retrieve images. The difficult terms are disambiguated against Wikipedia by the Wikipedia Disambiguator module. To link the Wikipedia page we are applying an algorithm akin to the most frequent sense in Word Sense Disambiguation. When a concept has more than one corresponding page in Wikipedia the number of incoming links is computed[12]. The hypothesis we pursue is that the most likely page assignment is that with the major number of incoming links. In practice (Mihalcea and Csomai, 2007) this simple heuristics turned to be very efficient.

In the Online mode the user highlights the difficult concepts and the Online Image Retrieval component retrieves images through Google and Bing searching engines. The user or carer then adds the images to the appropriate place in the text.

[11]The image URL existed when the resource was first created but meanwhile they are relocated or deleted.

[12]The incoming links are links from other wikipedia pages to the page corresponding to the concept.

4.2. Idiom detection

The figurative language in general and the idioms in particular present specific problems for our users as they are not able to grasp the meaning of these expressions. When reading a text they tend to construct the literal meaning of figurative language expressions and therefore misconstrue the meaning of the linguistic unit of which the figurative language expression is part. Even if notable progress has been made in the last years, the Natural Language Processing algorithms are far from achieving enough precision to deal with the diversity of figurative language. The best we can do is to identify the idiomatic expressions and provide good definitions for them. Please note that giving definitions of identified idioms or words is one of the strategies recommended by the National Reading Panel as part of a strategy called explicit instruction.

In the linguistic discourse and lexicographical practice the term "idiom" is applied to a fuzzy category defined by prototypical examples: "kick the bucket", "keep tabs on", etc. Because it is hard to provide a definition for the whole class of idioms we give some characteristics of the whole class of idiomatic expressions. According to (Nunberg et al., 1994) there are three properties that the idioms exhibit:

- Conventionality. The meaning of idioms is not compositional, therefore it cannot be constructed from knowing the meaning of its components and the grammatical rules of the language.
- Inflexibility. Idioms appear in a limited range of syntactic constructions.
- Figuration. The line between idioms and other figurative language is somewhat blurred because other figurative constructions like literal metaphors (take the bull by the horns) or hyperboles (not worth the paper it's printed on) are also considered idioms.

The Open Book software incorporates three idiom dictionaries: one for each language in the project. The idiom dictionaries have been compiled from multiple web sources and specialized dictionaries in the public domain. The idiom dictionaries have been verified by native language speakers and matched for productivity (the number of occurrences) in big corpora. The size of the compiled dictionaries is given in Table 2.

Internally the idiom dictionaries are stored in XML format. An example of an entry in the English idiom dictionary is in Figure 4:

Dictionary of idioms	Number of idioms
Bulgarian	771
English	2401
Spanish	4519
TOTAL	7691

Table 2: The size of idiom dictionaries for each language of the project.

```
<IdiomUnit>
        <idiom>shake the world</idiom>
        <lemma>shake the world</lemma>
        <jape>
                {Token.root=="shake"}{Token.category==
   "DT"}({Token.category=="JJ"})*2{Token.root=="world"
   }
        </jape>
        <definition>
                To shake the world is to cause a large
    amount of shock or surprise.
        </definition>
</IdiomUnit>
```

Figure 4: English idiom dictionary example.

The idioms are grouped in idiom units. Each unit is composed of three elements but only the first and the last one are mandatory. The idiom element holds the dictionary entry of the idiom. The lemma element holds the lemmatized entry of the idiom. The jape element holds a JAPE (regular expressions over annotations) (Cunningham et al., 2011) expression version of the idiom. The last element is the dictionary definition of the idiom.

The architecture of the idiom web service is presented in Figure 5. The input document is tokenized and lemmatized using the appropriate language specific tokenizers and lemmatizers. Then, the Lucene [13] framework indexes the document. The matching of idioms against the text is done in three modalities corresponding to the entries in the idiom dictionary:

[13] http://lucene.apache.org/

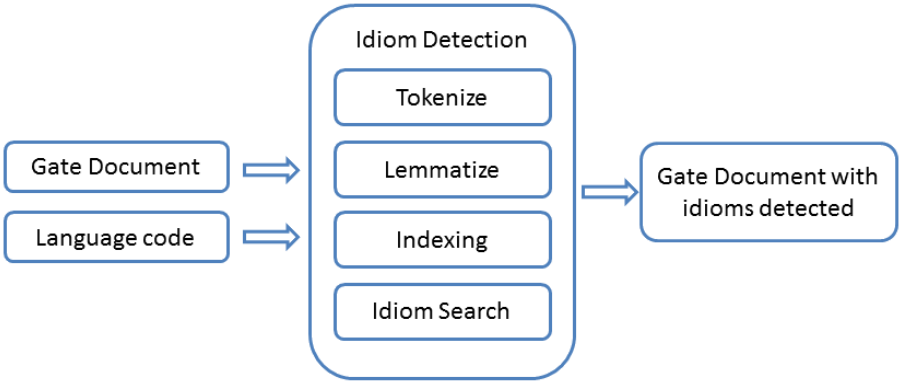

Figure 5: The architecture of the idiom detection module

1. *String matching*. The idioms are searched in text by simple string marching. This is the more precise and efficient way of searching the idioms in text. The disadvantage is that in this modality many idioms are missed. Considering the idiom entry above we will miss all idioms that do not have the exact string form (e.g. "shook the world", "shaken the world" etc.).
2. *Lemma matching*. The lemma form of the idiom is matched against the lemmatized version of the text. The lemma matching will match all the forms that the string matching misses. The precision only drops slightly for "poor" morphological languages like English.
3. *JAPE matching*. The matching with the regular expressions over annotations is the most general matching that could be performed but it is also the less precise. Some idioms might be realized in text with intervening words. The JAPE expression bellow matches the form "shook the whole world" where the adjective "whole" can be inserted inside the dictionary form of the idiom. The default level of matching is lemma but for the entries containing JAPE expressions tested on representative corpora the default matching level can be changed.

4.3. Summarization

The summarization is the task of presenting the information in a condensed form. Depending on the information type there are multiple forms of summaries like: a preview or a trailer of a movie, abstracts of scientific articles, statistic tables showing the performance of football players. Based on the user requirements and on the experiments performed with the texts simplified by the carers we implemented two forms of summarization of textual information. The first one is

a summarization technique widely used in Natural Language Processing called summarization by sentence extraction. The second one, called Topic Model, is a summarization technique for browsing large document collections.

4.3.1. Summarization by Sentence extraction

In Natural Language Processing there are two well-known summarization techniques called extraction and abstraction. An extract is a summary that contains only material from input text document. The abstract instead is a summary which uses some material that is not present in the input document. In general the abstracts are computed after the extraction phase by applying "glue" operations for the extracted sentences and possibly natural language inference to add information not present in the text. Because are easier to implement and can be used with a variety of texts with little or no change, the majority of the summarization system nowadays are sentence extractors. Moreover in terms of performance the extracts are not worst that the abstracts (Mani, 2001). Unlike the extraction systems, the abstraction systems are tailored to specific domains, require deep linguistic analysis like discourse understanding for example and make use of knowledge resources (e.g. ontologies). Given these considerations and the FIRST project objective constraints the sentence extraction system should have the following characteristics:

1. Language Independence. The core part of the sentence extraction system should be language independent. A totally independent summarization system might be impossible but it is desirable that the implemented system relies on low-level linguistic knowledge. If we consider the languages covered by FIRST project only English has good NLP processing resources. Bulgarian has few language resources: a part of speech tagger and a lemmatizer are the only resources we can use.
2. Performance. Ideally the implemented extraction system should be based on algorithms of proven efficiency.
3. Responsiveness. Like all the software components of the Open Book system the sentence extraction should respond quickly to the user request. Preferably the sentence extraction should not take more than 10 seconds to complete.

After a thorough literature review we choose to implement a sentence extraction system based on an algorithm known in literature as either TextRank (Mihalcea and Tarau, 2004) or LexRank (Erkan and Radev, 2004). This algorithm

is inspired by the Google PageRank (Page et al., 1998), the algorithm used by Google to rank web pages. In what follows we will only briefly describe the TextRank algorithm. The description we give should nevertheless enough for the reader to understand the sentence extraction system implemented in Open Book. If the reader needs further details is invited to read the articles cited before.

Google PageRank is an application of the concept of prestige in social networks to the realms of web. In social networks the reputation of a social actor depends of the number links it has with other social actors as well as on the prestige of the social actors is connected to. Similarly PageRank computes the importance of a web page using the votes (in this case the votes are the web links) it receives from other web pages as well as the importance of web pages that link to it. Text Rank is the PageRank algorithm tailored to summarization by sentence extraction. Text Rank represents the input document as a weighted undirected graph that has as vertexes the sentences of the text. An edge is drawn between two vertexes if the sentences corresponding to the vertexes are similar and the weight of the edge is the similarity score between the sentences at the vertexes. The similarity score can be computed using measures like word overlap or cosine similarity between sentences. The intuition behind Text Rank algorithm is that the sentences that have more prestige should be extracted in the summary. If we pause a little to think we understand why this is the case: the most prestigious sentences are those sentences that are voted by many other prestigious sentences. Because the links between sentences express sentence similarity this means the most prestigious sentences better compress the text information than the sentences that have less reputation.

The algorithm implemented in Open Book combines two strategies for computing the most salient sentences. The first one assigns Text Rank scores to the sentences and the second strategy assigns a score according to the sentence position in the text. The second strategy goes back to the seminal work of Edmunson (Edmundson, 1969) who noticed that the best predictor for certain kind of articles (e.g. scientific articles) for the sentences to be extracted is the sentence location in the text. The sentences near the title are more likely to be relevant for inclusion in the summary than the more distant ones.

The combined score of each strategy reflects the final relevance of the sentence. The architecture of the Summarization module is presented in Figure 6.

The final score reflecting the final relevance of the sentence combines the scores for each strategy.

The length of the summary is controlled by a parameter. The length can be expressed either as a percentage of the original text (e.g. 20%) or as the number

Figure 6: The architecture of the Summarization System

of sentences (e.g. 5 sentences, 10 sentences).

4.3.2. Topic Model

Topic Models are methods for organizing large document repositories based on the themes or discourses that permeate the collection. The computation of document topics is related to the user requirement: present the most important information in a manner that serves the user information needs. This requirement can be implemented either by computing the document summary or presenting before reading a set of topics that summarize the document's main ideas.

The mathematical details of the topics models are somewhat harder to grasp but the main intuition behind it is easily understood. For example if we consider an astrobiology document it will most likely talk about at least three topics: biology, computer models of life and astronomy.

It will contain words like: *cell*, *molecule*, *life* related to the **biology topic**; *model*, *computer*, *data*, *number* related to **computer models of life topic** and *star*, *galaxy*, *universe*, *cluster* related to **astronomy topic**.

From a practical point of view the topics can be viewed as clusters of words that frequently co-occur in the document collection. The Latent Dirichlet Allocation (LDA) is the simplest and the most efficient topic model technique Blei et al. (2003). The main assumption behind LDA is that the documents in the collection were generated by a random process in which the topics are drawn from a given distribution of topics and words are drawn from the topics themselves. The task of LDA is to construct the topic distribution and the topics (which are basically probability distributions over words) starting from the documents in the collection.

The Topic Model Web Service in Open Book is based on the implementation of LDA. This web service assigns topics to the submitted documents, thus informing the users about the themes traversing the documents and facilitating the browsing of the document repository. The architecture of the Topic Model Web Service is presented in Figure 7.

Once a document is received by the Topic Model Web Service it is first dis-

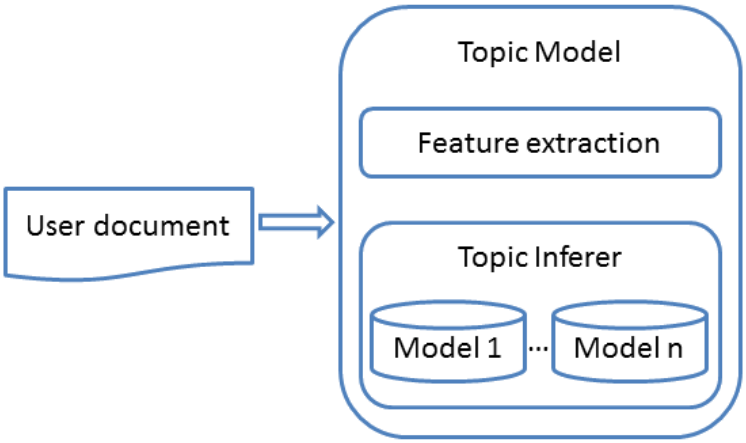

Figure 7: The architecture of the Topic Model module

patched to the Feature Extraction Module where it is POS tagged and lemmatized and the relevant features are extracted. The extracted features are all nouns and name entities in the document. Then the Topic Inferencer Module loads the appropriate domain model, performs the inference and assigns the new topic to the document. There are three domains/genders that the users of our system are mainly interested in: News, Literature and Health domain. For each of these domains we train topic models in each of the three languages of the project.

The Topic Model Web service is easily extensible to other domains. To add the new model to the web service is simply a matter of loading it in the system and modifying a configuration file.

The output of the Topic Model Web Service is a document containing the most important topics and the most significant words in the topics. The last two parameters can be configured.

5. Evaluation

In this section we evaluate the performance of the Image Retrieval System and Topic Model detection. The performance of the summarizer system has been thoroughly evaluated elsewhere (Erkan and Radev, 2004). The idiom detection system has a very high precision (close to 100%) because the idioms in our list are frozen phrases, multiple meanings being unlikely.

5.1. Evaluation of Image Retrieval

The evaluation of image retrieval has been performed for 12 documents selected by the psychologists: half English and half Spanish. The evaluation of the image retrieval process is made with respect to a set of query terms. The terms (Table 3) were randomly chosen from the Specialized Term Set, constructed by the Disambiguation module.

Each annotator evaluates the suitability of the retrieved image with regard to the terms considered intrinsically and in context. Both types of evaluation are necessary because the retrieved image may be appropriate for illustrating the intrinsic concept, but it might not be appropriate for illustrating the concept in particular linguistic contexts. Then the Kappa agreement (Siegel and Castellan, 1988) between annotators is computed and reported. The precision of image retrieval is computed for both types of evaluation (terms considered intrinsically and in context) and reported.

5.1.1. Evaluation of Online Image Retrieval Module

The Online Image Retrieval module was used to retrieve image for the concepts listed in Table 3. For each concept, the first two images retrieved by each engine were evaluated. Two annotators read the texts, identified the specialized concepts and annotated the appropriateness of the retrieved image for concepts considered both intrinsically (CI) and in context (CC).

The quality of the image retrieval modules was measured using K-score(Kappa Score) to assess the degree to which two human annotators considered the images returned by the systems to be appropriate. These results are displayed in Table 4.

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

The precision of the retrieval procedure (Table 5) is computed in the customary way as the proportion of correctly identified images divided by the total number of images. The computation is done both when concepts are considered intrinsically

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

Table 3: Selected Terms for Image Evaluation for each Document.

and in context, and for each annotator. The reported precision is the average precision score for the two annotators.

As expected, the precision with which the module retrieves suitable images for concepts considered in context is lower than that achieved when concepts are considered intrinsically. This indicates it is necessary, in future development, to

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

Table 4: The inter-annotator agreement for Online Image Retrieval Module.

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

Table 5: The overall precision for Online Image Retrieval Module.

improve the Online Image Retrieval web service by incorporating contextual information.

5.1.2. Evaluation of Offline Image Retrieval Module

The Offline Image Retrieval Module is based on the Wikipedia Disambiguation module mentioned before. Unfortunately, only 77% of the concepts listed in Table 1 can be mapped onto Wikipedia. The average accuracy with which concepts can be mapped is relatively high: 95% for English and 88% for Spanish.

The extent to which suitable images could be retrieved for successfully mapped concepts, both when considered intrinsically and in context, was measured. As in the previous case the evaluation task was performed by two human annotators and the inter-annotator agreement (Kappa) between the human annotators was obtained. The results are displayed in Table 6:

Language	CI	CC
English	0.77	0.70
Spanish	0.79	0.71

Table 6: The inter-annotator agreement for Offline Image Retrieval Module.

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

The overall precision for image retrieval, both intrinsically and in context, is displayed in Table 7.

Language	CI	CC
English	0.88	0.85
Spanish	0.68	0.61

Table 7: The overall precision for Offline Image Retrieval Module.

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

After computing the agreement and precision for Image Retrieval web Services in two modalities we conclude that the annotator agreement is moderate to substantial for Online Image Retrieval Web Service and substantial for Offline Image Retrieval web service. The precision of the Offline Image Retrieval web service is better than the precision of Online Image Retrieval Web Service but the recall is much lower. The drop of precision between the two modalities invites incorporation of contextual factors in the future versions of the two web services.

5.2. Evaluation of the Topic Model Web Service

To evaluate the topic model web service we trained two models using two corpora of news articles .The English corpus is the well-known Reuters corpus consisting of 21,578 news documents[14]. The Spanish corpus consists of 32189 news documents crawled from the sites of the main newspapers in Spain. Two topic models with 100 topics were trained for both languages. The topic model inferencer was used to infer the topics for 10 test documents. For evaluation we chose the three most salient topics and the three most salient concepts in the topics according to the assigned scores. Two annotators have annotated each concept in the selected topic with the label "yes", if the concept can be used to describe the content of the document, or with the label "no", if the concept cannot be used to describe the content of the document. The Kappa agreement scores are presented in Table 8.

The agreement for English is only moderate if we refer to the scale of Landis and Koch cited before. The explanation for the low agreement is that for the concepts that do not belong to the text, the significance is assessed based on the

[14] `http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html`

Language	Kappa
English	0.52
Spanish	0.79

Table 8: The inter-annotator agreement for Offline Image Retrieval Module.

text interpretation. The interpretation of the text is subjective and grounded in background knowledge. Two non-native speakers have performed the annotation for English therefore we expect that the text interpretation varies to a considerable extent. This effect is lower for Spanish because the annotation has been performed by two native language speakers. The agreement for Spanish is substantial.

5.3. Evaluation by the users

The Open Book tool has been evaluated by high-functioning autistic individuals [15]. 243 persons (193 males and 50 females) have been recruited in UK, Spain and Bulgaria to participate in the evaluation of the tool. All recruited users had basic level literacy skills and have been accompanied by carers . The users and their carers have been given access to a preliminary version of the tool and received training from the technical partners for two months. Reading comprehension tests under strict time conditions have been administered to all participants. Half of the texts used in the reading comprehension tasks have been simplified by the carers with the help of the Open Book tool. The texts presented to the users have been randomly selected. Thus the status of the texts to be evaluated (simplified or not) was transparent to the study participants. Following the reading of the texts the users filled up reading comprehension questionnaires. Based on the questionnaires a reading comprehension score was assigned to each text. A comparison of the reading comprehension scores for the simplified and normal texts showed a statistical significant difference favoring the simplified texts. Therefore the Open Book tool helps producing texts that are easier to understand for a wide range of users speaking different languages.

In addition to the reading comprehension tests the users gave feedback about the usability and the usefulness of the Open Book tool. On the positive side they stressed the fact the interface is easy to use (”is not over-engineered” as one user

[15]This section only gives a general overview of the evaluation process. Most of the evaluation work including the recruitment of the users , the sampling and administration of the tests has been performed by the clinical partners. For further information the reader is invited to consult the deliverable D7.8 that is in the public domain `http://first-asd.eu/?q=D7.8`

pointed out) and it is easy to set up requiring only a web browser. Different users appreciated different modules integrated in the system, the most appreciated feature being the ability to insert images. The most frequent complaint of the users was about the imprecision of the integrated modules: the images are in some situations inappropriate, the synonym suggestion is in some cases inaccurate, etc. Clearly, better algorithms can be devised. However there is a tradeoff between precision and speed. Better algorithms mean more computations and therefore delays between the time the user submitted his/her request and the times s/he receives the answer. A second complaint was about the coverage of the linguistic resources integrated into the tool. As stressed before, the resources for Bulgarian are scarce and the resources for English and Spanish have an uneven coverage (e.g. some domains are better covered than others). Nevertheless the tool architecture allows for easy integration of additional resources when these will be available.

6. Conclusions

In this paper we have presented the tool Open Book that applies HLT to identifying reading comprehension obstacles in text documents in three languages: English, Spanish and Bulgarian. We have implemented several NLP features in the whole architecture of OB oriented to simplifying texts for ASD sufferers. Although the tool integrates several modules into a unique system, our main contribution presented in this paper focuses on the implementation of three specific challenging tasks in NLP. Firstly, we develop an algorithm to simplify difficult and complex concepts by retrieving associated images. Secondly, a module for idiom detection is implemented. Although the method proposed is very simple, this is the first time that a module of idiom detection has been integrated into a simplification system. Finally, we have dealt with the text summarization task by studying two novel techniques based on the topic models and the PageRank algorithm. The results obtained as direct as indirect evaluation show that our approach is quite valid to be taken into account for real applications.

On the other hand, the Open Book tool also presents some limitations. The first one is related to the image retrieval and idiom detection module. The two modules separately have a good performance, but the image retrieval can be improved through resolving the ambiguity of idioms with images. For example, the idiom “it’s raining cats and dogs” is not considered by the image retrieval module as an idiomatic expression, so the system returns an image of cats and dogs falling from the sky. Thus, our on-going work is to improve the image retrieval module with the incorporation of the idiom detection module. Regarding the idiom detec-

tion module, we are now working on the definition of JAPE expressions (regular expressions over annotations) (Cunningham et al., 2011) for all the idioms in the three languages, because up until now there have been several idioms without their corresponding JAPE expression. It is known in the NLP research community that the solutions based on language or domain independent algorithms are good approaches, but if it is possible to incorporate into the algorithm specific knowledge related to the language or the domain this is likely to improve the performance of the algorithm. Thus, we are working on including knowledge of each of the languages that are supported by the tool with the aim of improving the performance of the text summarization module.

Finally, we are considering other lines of research for the future, most of them related to summarization because there are several points that can be improved and integrated into our system. The first one could be the generation of content tables. In addition, it would be an interesting approach to combine several algorithms in order to obtain better results for text summarization. On the other hand, regarding figurative language, there are several open issues not only related to idioms but also considering metaphors. That is one of the most complicated and challenging problems facing the NLP community. Thus, an in-depth study involving the detection of metaphors in text would be our most difficult milestone.

Acknowledgements

This work has been partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER), ATTOS project (TIN2012-38536-C03-0) from the Spanish Government. The project AORESCU (P11-TIC-7684 MO) from the regional government of Junta de Andalucía and the project CEATIC-2013-001 from the University of Jaén partially supports this manuscript. The work in this paper is partially funded by the European Commission under the Seventh (FP7 - 2007-2013) Framework Program for Research and Technological Development through the FIRST project (FP7-287607). This publication reflects only the views of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

References

G. B. Mesibov, L. W. Adams, L. G. Klinger, Autism: Understanding the disorder., Plenum Press, 1997.

K. Nation, P. Clarke, B. Wright, C. Williams, Patterns of reading ability in children with autism spectrum disorder, Journal of Autism and Developmental Disorders 36 (2006) 911–919.
URL http://link.springer.com/article/10.1007%2Fs10803-006-0130-1

C. Lord, R. Paul, Language and communication in autism., in: Handbook of autism and pervasive developmental disorders (2nd Edition), John Wiley & Sons, 1997.

B. Fossett, P. Mirenda, Sight word reading in children with developmental disabilities: A comparison of paired associate and picture-to-text matching instruction, Research in Developmental Disabilities 27 (4) (2006) 411–429.

B. Ploog, A. Scharf, D. Nelson, P. Brooks, Use of computer-assisted technologies (cat) to enhance social, communicative, and language development in children with autism spectrum disorders, Journal of Autism and Developmental Disorders 43 (2) (2013) 301–322.

J. Mintz, Additional key factors mediating the use of a mobile technology tool designed to develop social and life skills in children with autism spectrum disorders: Evaluation of the 2nd {HANDS} prototype, Computers & Education 63 (0) (2013) 17 – 27.

H.-C. Chu, M.-J. Liao, T.-Y. Chen, C.-J. Lin, Y.-M. Chen, Learning case adaptation for problem-oriented e-learning on mathematics teaching for students with mild disabilities, Expert Systems with Applications 38 (3) (2011) 1269 – 1281.
URL http://www.sciencedirect.com/science/article/pii/S0957417410005415

A. Coronato, G. D. Pietro, G. Paragliola, A situation-aware system for the detection of motion disorders of patients with autism spectrum disorders, Expert Systems with Applications 41 (17) (2014) 7868 – 7877.
URL http://www.sciencedirect.com/science/article/pii/S0957417414002917

A. I. Tweissi, The effects of the amount and type of simplification on foreign language reading comprehension., Reading in a foreign language 11 (2) (1998) 191–204.

J. L. Luckner, C. M. Handley, A summary of the reading comprehension research undertaken with students who are deaf or hard of hearing, American Annals of the Deaf 153 (1) (2008) 6–36.

L. Rello, H. Saggion, R. Baeza-Yates, Keyword highlighting improves comprehension for people with dyslexia, in: Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR), Association for Computational Linguistics, 2014, pp. 30–37.
URL http://aclweb.org/anthology/W14-1204

R. Chandrasekar, C. Doran, B. Srinivas, Motivations and methods for text simplification, in: Proceedings of the 16th Conference on Computational Linguistics - Volume 2, COLING '96, Association for Computational Linguistics, Stroudsburg, PA, USA, 1996, pp. 1041–1044.
URL http://dx.doi.org/10.3115/993268.993361

L. Feng, Text simplification: A survey, Tech. rep., The City University of New York (March 2008).

A. Siddharthan, A survey of research on text simplification, the International Journal of Applied Linguistics (2014) 259–98.

M. Shardlow, A survey of automated text simplification, International Journal of Advanced Computer Science and Applications (Special Issue on Natural Language Processing) (2014) 58–70.
URL http://dx.doi.org/10.14569/SpecialIssue.2014.040109

S. Devlin, J. Tait, The use of a psycholinguistic database in the simplification of text for aphasic readers, Linguistic databases (1998) 161–173.

S. Devlin, G. Unthank, Helping aphasic people process online information, in: Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility, Assets '06, ACM, New York, NY, USA, 2006, pp. 225–226.
URL http://doi.acm.org/10.1145/1168987.1169027

N. Kaji, D. Kawahara, S. Kurohash, S. Sato, Verb paraphrase based on case frame alignment, in: Proceedings of the 40th Annual Meeting on Association for

Computational Linguistics, ACL '02, Association for Computational Linguistics, Stroudsburg, PA, USA, 2002, pp. 215–222.
URL http://dx.doi.org/10.3115/1073083.1073120

Y. Canning, Syntactic simplification of text, Ph.D. thesis, University of Sunderland, United Kingdon (2002).

A. Siddharthan, Syntactic simplification and text cohesion, Ph.D. thesis, University of Cambridge, United Kingdon (2003).

A. Siddharthan, A. Copestake, Generating anaphora for simplifying text, in: Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2002), Lisbon, Portugal, Citeseer, 2002, pp. 199–204.

P. R. A. Margarido, T. A. S. Pardo, G. M. Antonio, V. B. Fuentes, R. Aires, S. M. Aluísio, R. P. M. Fortes, Automatic summarization for text simplification: Evaluating text understanding by poor readers, in: Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web, WebMedia '08, ACM, New York, NY, USA, 2008, pp. 310–315.
URL http://doi.acm.org/10.1145/1809980.1810057

N. Bouayad-Agha, G. Casamayor, G. Ferraro, L. Wanner, Simplification of patent claim sentences for their paraphrasing and summarization, in: FLAIRS Conference, 2009.

S. Baron-Cohen, Theory of mind and autism: a review, Int Rev Ment Retard 23 (2001) 169–184.

U. Frith, M. Snowling, Reading for meaning and reading for sound in autistic and dyslexic children, British Journal of Developmental Psychology 1 (4) (1983) 329–342.

I. M. O'Connor, P. D. Klein, Exploration of strategies for facilitating the reading comprehension of high-functioning students with autism spectrum disorders, Journal of Autism and Developmental Disorders 34 (2) (2004) 115–127.
URL http://link.springer.com/article/10.1023/B:JADD.0000022603.44077.6b

J. Oakhill, N. Yuill, Pronoun resolution in skilled and less-skilled comprehenders: Effects of memory load and inferential complexity, Language and Speech 29 (1) (1986) 25–37.

T. Grandin, Thinking In Pictures: and Other Reports from My Life with Autism, Vintage, 1996.

R. K. Kana, T. A. Keller, V. L. Cherkassky, N. J. Minshew, M. A. Just, Sentence comprehension in autism: Thinking in pictures with decreased functional connectivity (2006).

H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, W. Peters, Text Processing with GATE (Version 6), 2011.

G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller, WordNet: An on-line lexical database, International Journal of Lexicography 3 (1990) 235–244.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, 2009, pp. 248–255.

J. Deng, A. C. Berg, K. Li, L. Fei-Fei, What does classifying more than 10,000 image categories tell us?, in: Proceedings of the 11th European Conference on Computer Vision: Part V, ECCV'10, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 71–84.

R. Mihalcea, A. Csomai, Wikify!: Linking documents to encyclopedic knowledge, in: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM '07, ACM, New York, NY, USA, 2007, pp. 233–242.

G. Nunberg, I. Sag, T. Wasow, Idioms, Language.

I. Mani, Automatic Summarization, John Benjamins Publishing Co, 2001.

R. Mihalcea, P. Tarau, TextRank: Bringing order into texts, in: D. Lin, D. Wu (Eds.), Proceedings of EMNLP 2004, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 404–411.

G. Erkan, D. R. Radev, LexRank: graph-based lexical centrality as salience in text summarization, J. Artif. Int. Res. 22 (1) (2004) 457–479.

L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank citation ranking: Bringing order to the web, Tech. rep., Stanford Digital Library Technologies Project (1998).

H. P. Edmundson, New methods in automatic extracting, J. ACM 16 (2) (1969) 264–285.

D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

S. Siegel, N. J. Castellan, Nonparametric statistics for the behavioral sciences, 2nd Edition, McGraw-Hill, New-York, 1988.

J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data., Biometrics 33 (1) (1977) 159–174.