A Survey of Automated Text Simplification

Matthew Shardlow
Text Mining Group, School of Computer Science
University of Manchester, Manchester, United Kingdom
Email: mshardlow@cs.man.ac.uk

***Abstract*—Text simplification modifies syntax and lexicon to improve the understandability of language for an end user. This survey identifies and classifies simplification research within the period 1998-2013. Simplification can be used for many applications, including: Second language learners, preprocessing in pipelines and assistive technology. There are many approaches to the simplification task, including: lexical, syntactic, statistical machine translation and hybrid techniques. This survey also explores the current challenges which this field faces. Text simplification is a non-trivial task which is rapidly growing into its own field. This survey gives an overview of contemporary research whilst taking into account the history that has brought text simplification to its current state.**

***Keywords*—Text Simplification, Lexical Simplification, Syntactic Simplification**

I. Introduction

Text Simplification (TS) is the process of modifying natural language to reduce its complexity and improve both readability and understandability. It may involve modifications to the syntax, the lexicon or both. The automation of this process is a difficult problem which has been explored from many angles since its conception in the nineties [1]–[7]. This survey paper is intended to give an overview of the field of TS in its current state. To the author's knowledge, there is no similar publicly available survey since 2008 [8]. Whereas the previous survey identified eight separate systems, this work has exposed closer to fifty. The recent growth in TS research can be seen in Figure 1 where it is clear that TS is steadily increasing in size as a field. The last few years have seen a growing maturity in the field, marked by an increased use of both external resources [9]–[11] and methods [12]–[14].

Whereas there has been much work in manual TS over the years, especially with a focus on second language learners [15], there is less work in automated simplification. The first effort towards automated simplification is a grammar and style checker developed for writers of simplified English [16]. This was developed at Boeing for the writers of their commercial aircraft manuals, to help them keep in accordance with the ASD-STE100 standard for simplified English[1]. Further work to automate simplification for controlled language was undertaken [17]. This was later extended for the case of general language in the areas of syntactic simplification [3] and lexical simplification [4]. These methods have heavily influenced future efforts to date. Work to improve the preservation of discourse in syntactic simplification [18] and to improve the context–awareness of lexical simplification [12]–[14] has been carried out. Other work has involved applying phrase based statistical machine translation techniques to produce simple English [10], [19], [20].

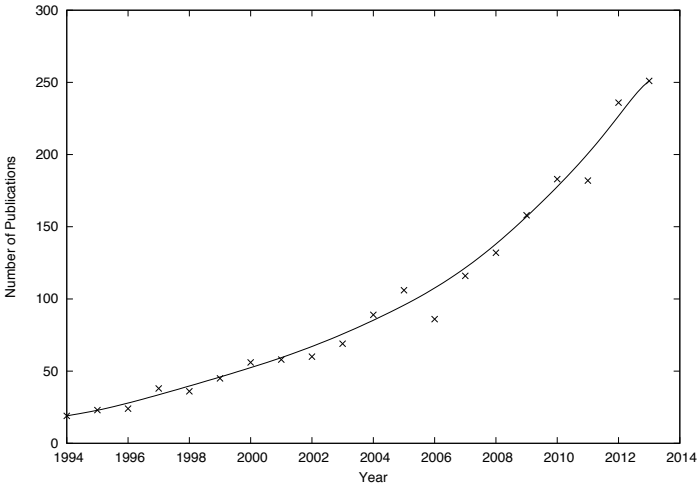

Fig. 1. This graph was produced by polling Google Scholar with the search query: 'Text Simplification' OR 'Lexical Simplification' OR 'Syntactic Simplification'. It shows the sustained growth in TS and associated sub-fields between 1994 and 2013.

TS is within the field of natural language processing. Within this field it is very similar to other techniques such as machine translation, monolingual text-to-text generation, text summarisation and paraphrase generation. These fields all draw on each other for techniques and resources and many techniques within TS come from these other fields [19], [21]. TS is different to text summarisation as the focus of text summarisation is to reduce the length and content of input. Whilst simplified texts are typically shorter [22], this is not necessarily the case and simplification may result in longer output — especially when generating explanations [23]. Summarisation also aims at reducing content — removing that which may be less important or redundant. This is typically not explored within simplification, where all the content is usually kept. Some efforts have explored the use of simplification alongside summarisation systems [24]–[28]. Here, TS is used to improve the readability of the final summary.

When talking about TS the words *simple* and *complex* are often used in relation to each other as shown in Table I. For example, in a parallel corpus of simplified English and regular English, the former will be called *simple* and the latter *complex*. In a corpus of technical English and regular English, the former will be called *complex* and the latter *simple*. This shows that simplicity and complexity are relative to each other and should be used with care. When creating *simple* text, we actually intend to create text which is more *simple* (and so

[1]http://www.asd-ste100.org/

TABLE I. SIMPLICITY IS RELATIVE TO COMPARISON

	Simple Text	Lay Text	Technical Text
Simple vs. Lay	Broken Arm. > Fractured Arm. (is simpler than)		
Lay vs. Technical		Fractured Arm. > Fractured Tibia. (is simpler than)	

less *complex*) than it originally was. Two other important terms to define are *readability* and *understandability*. At first, these may seem like the same things, however they may be measured independently depending on the context of an application. *Readability* defines how easy to read a text may be. This is typically governed by factors such as the complexity of grammar, length of sentences and familiarity with the vocabulary. *Understandability* is the amount of information a user may gain from a piece of text. This can be affected by factors such as the user's familiarity with the source's vocabulary, their understanding of key concepts or the time and care taken to read the text. It may be the case that a text has high *readability*, but low *understandability*. For example: trying to read a well written scientific article with no scientific training. It may also be possible that a text has low *readability*, but is still *understandable*. For example: an author who communicates a simple point uses misleading grammatical structures. *Readability* and *understandability* are related and a text which is easier to read is likely to be more understandable, as the reader will find it easier to take the time to look over the difficult concepts. Similarly, a text which is easily understandable will encourage the reader to keep reading, even through difficult readability.

Simplicity is intuitively obvious, yet hard to define. Typical measures take into account factors such as sentence length [29], syllable count [30] and other surface text factors. Whilst these give a good estimate, they are not always accurate. For example, take the case of sentence length. One long sentence may use many complex terms and anaphora (words which refer to a previous entity: he, she, it, etc.). A simplified version of this would be lexically longer, but may be more explicative. In the case of explanation generation, complex words are appended with short definitions, increasing sentence length. Automatic heuristic measures will judge these sentences which have been simplified to be more complex. The final text may be longer, however it is also easier to understand and therefore simpler.

Different forms of simplification will address different needs. No two users are exactly the same and what one finds easy, another may find difficult. This is true both at the level of different user groups (the low literacy user requires different simplifications to the second language learner), but also within user groups. Factors such as dialect, colloquialisms and familiarity with vocabulary and syntax can all affect the user's understanding of a text. This means that simplification is best done at a general level. Text which is made very simple for one user may be more complex for another. However text which is made slightly simpler for one user will generally be easier for most other users.

This still leaves the question of how to measure simplicity. Automatic measures are ineffective [31]. In fact, they may even be detrimental to the simplification process. For example, if a measure favours short sentences and the aim of simplification is to get the best score with that measure, we could easily succeed by reducing our text to a series of two or three word stubs. This would be much more difficult to read and understand, yet score highly.

Although many efforts have been made towards TS techniques over the past two decades, few have been used in production. Those used in production are generally developed with a focus as an aid to the user in translating their text to simplified language [16], [32]. A governing factor in the low take-up of TS systems is inaccuracy. In some natural language processing applications, a low accuracy may be acceptable as the application is still usable. For example, in information retrieval, even if a system has a moderate accuracy, it will still enable the user to find some portion of the documents they were looking for. Without the information retrieval system, the user would not have been able to find the documents as easily. However, this does not transfer in the same way to TS. If a system is not accurate, then the resultant text will not make sense. If the text is not understandable, then it will definitely not be more simple than the original. If a user is routinely presented with inaccurate simplifications, then they will not find it helpful. Assistive technology must be accurately assistive, otherwise the user will be more confused and less able to interact with the text than in its original "more complex" form.

TS is a largely unsolved task. Whereas many areas of natural language processing and computer science have a flagship system, method or technique, TS has many varied approaches (as outlined in Section II). Whilst these techniques employ differing methodologies and may have differing outputs, their purpose is always to simplify text. The field is fast moving and research into new areas is regularly produced. Further, more and more people are becoming interested in TS with an increased number of projects and publications year on year, as shown in Figure 1. Whilst many techniques have been implemented, there is still much work to be done in comparing, evaluating and refining these.

Text is a fundamental part of our daily interaction with the information world. If text is simplified for an end user, then this may improve their experience and quality of life. Whether we are reading the newspaper, checking emails or following instructions, it is highly important to be able to understand the text used to convey this information. TS can be applied to reduce the complexity of information and increase a user's understanding of the text they encounter in their day to day lives. This has great advantages for both readers and authors. The reader gains a better understanding of the world around them and authors can ensure their written material will be understandable by those recipients with a low reading level.

The related fields of machine translation and text summarisation allow for the crossover and sharing of techniques. For example, corpus alignment techniques have been borrowed from summarisation [33], [34] and statistical machine translation techniques have been used [19], [35], along with their evaluation methods [36]. This crossover means that, as these fields progress, there will be new techniques available for the task of simplification. As techniques are developed in the context of TS, they will also be useful in the context of other related domains.

The need for simplified English in particular is evidenced by the popularity of 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. Many groups with low levels of English benefit. The size of Simple Wikipedia indicates the need for simple English, however the process of hand crafting these articles is time consuming. Improvements in automating simplification would help to address this need.

II. Approaches

TS has been carried out in a number of different ways. Many systems use a combination of approaches to simplify text in different manners. These different methods of TS are largely independent and methodologically distinct of each other. In this section, we observe the development of methods from: lexical and syntactic simplification, explanation generation, statistical machine translation and TS techniques in languages other than English.

A. Lexical Approaches

Lexical simplification is the task of identifying and replacing complex words with simpler substitutes. This involves no attempt to simplify the grammar of a text but instead focusses on simplifying complex aspects of vocabulary. An overview of research papers in lexical simplification is presented in Table II. Lexical simplification may be formulated as a phrase based substitution system, which takes limited syntactic information into account. There are typically 4 steps to lexical simplification as shown in Figure 2. Firstly, the complex terms in a document must be identified. Secondly, a list of substitutions must be generated for each one. Thirdly, those substitutions should be refined to retain those which make sense in the given context. Finally, the remaining substitutions must be ranked in their order of simplicity. The most simple synonym is used as a replacement for the original word. Systems have made differing variations on this theme with many approaches missing out the word sense disambiguation step.

In the first notable work in automated lexical simplification [4], the authors rank synonyms from the semantic thesaurus WordNet [49] using Kučera-Francis frequency [50] to identify the most common synonym. This work has heavily influenced lexical simplification systems since [12]–[14], [34], [37], [51], providing a framework with many avenues to explore and build upon. Recently, work has also focussed on the simplification of numerical expressions for improved reader comprehension [52], [53].

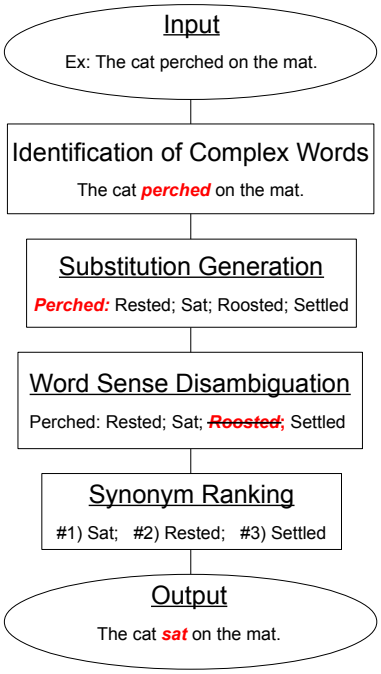

Fig. 2. The lexical simplification pipeline. Many simplifications will be made in a document concurrently. In the worked example the word 'perched' is transformed to sat. 'Roosted' is eliminated during the word sense disambiguation step as this does not fit in the context of 'cat'.

One area for improvement is the method of substitution ranking. Kučera-Francis frequency is the counts of words from the Brown corpus, which consists of just over one million words from 50 sources which are intended to be representative of the English language. Modern technology allows for frequency counts of much larger corpora to be carried out [54], [55]. Larger corpora are naturally better estimators of the true frequency counts of a language.

One of the major stumbling blocks with primitive lexical substitution systems is a loss of meaning due to word sense ambiguity. This occurs when a word has multiple meanings and it is difficult to distinguish which is correct. Different meanings will have different relevant substitutions and so replacing a word with a candidate substitution from the wrong word sense can have disastrous results for the cohesion of the resultant sentence. Early systems [4] did not take this into account, at the expense of their accuracy. Word sense disambiguation may be used to determine the most likely word sense and limit the potential synonyms to those which will maintain coherence.

Word sense disambiguation has been applied to lexical simplification in a number of different ways. These usually involve taking a standard lexical substitution system and applying a word sense disambiguation algorithm at some point. One such system is the latent words language model (LWLM) [56], which is applied to lexical simplification during the substitution generation step. The LWLM is used to generate a set of words which are semantically related to the original word. These are then compared against the substitutions returned by WordNet to remove any antonyms found by the LWLM. WordNet is useful for word sense disambiguation as it gathers words according to their semantic similarities into a group

TABLE II. RESEARCH EFFORTS IN LEXICAL SIMPLIFICATION ORDERED BY YEAR. LATER SYSTEMS ARE TYPICALLY MORE SOPHISTICATED. RECENT YEARS HAVE SEEN THIS AREA GATHERING MOMENTUM.

Year	Title	Notes
1998	The Use of a Psycholinguistic Database in the Simplification of Text for Aphasic Readers (PSET) [4]	Seminal work on lexical simplification.
2003	Text Simplification for Reading Assistance: A Project Note (KURA) [37]	Paraphrasing for deaf Japanese students in an educational setting.
2006	Helping Aphasic People Process Online Information (HAPPI) [38]	An update of PSET project for Web deployment.
2007	Mining a Lexicon of Technical Terms and Lay Equivalents [39]	Corpus alignment for paraphrasing.
2009	FACILITA: Reading Assistance for Low-literacy Readers (PorSimples) [40]	Designed for Brazilian Portuguese readers.
2009	Extracting Lay Paraphrases of Specialized Expressions from Monolingual Comparable Medical Corpora [41]	Paraphrasing medical corpora.
2010	Lexical Simplification [13]	Applying word sense disambiguation during the synonym generation phase.
2010	For the Sake of Simplicity: Unsupervised Extraction of Lexical Simplifications from Wikipedia [9]	Paraphrasing.
2011	Putting It Simply: a Context-aware Approach to Lexical Simplification [12] (SIMPLEXT)	A word sense disambiguation approach to lexical simplification.
2012	Can Spanish Be Simpler? LexSiS: Lexical Simplification for Spanish [42]	Spanish lexical simplification. c.f. [43], [44]
2012	English Lexical Simplification (SemEval Task 1) [45]	The project description for the SemEval 2012 task on lexical simplification.
2012	WordNet-based Lexical Simplification of a Document [46]	using WordNet hypernymy to perform substitutions.
2012	Automatic Text Simplification via Synonym Replacement [47]	Masters thesis focussing on the challenges of lexical simplification in Swedish.
2013	User Evaluation of the Effects of a Text Simplification Algorithm Using Term Familiarity on Perception, Understanding, Learning, and Information Retention [48]	Semi-automated lexical simplification for medical literature.

called a "synset". One particular use of WordNet [46] develops a tree of simplification relationships based on WordNet hypernym relations. This tree is used to reduce the size of the vocabulary in a document. Word sense disambiguation is carried out to place content words into their correct WordNet synset. Simplification may then be carried out by looking at the relevant node in the tree. Word sense disambiguation is also carried out by the use of context vectors [12], [42]. In this method, a large amount of information is collected on the surrounding context of each word and is used to build a vector of the likely co-occurring words. Vector similarity measures are then used to decide which word is the most likely candidate for substitution in any given context. These methods show the diversity of word sense disambiguation as applied to lexical simplification.

Other work has attempted to improve lexical simplification by improving the frequency metrics which are used. Frequent words have been shown to increase a text's readability [57]. Simple Wikipedia has been shown to be more useful than English Wikipedia as a method for frequency counting [58]. N-Grams have shown some use in providing more context to the frequency counts, with higher order n-grams giving improved counts [59]. However, the most effective method has so far proven to be the usage of a very large initial data-set [58], [59]. Namely, the Google Web 1T [55].

As well as performing substitutions at the single word level, lexical substitution may also be carried out at the phrase level, which requires some knowledge of how words cluster into individual phrases and how these can be recognised and substituted. A phrase may be replaced by a single word which conveys the same sentiment or by another phrase which uses simpler language. This may be done by comparing revisions in the edit histories of Simple Wikipedia [9] or by comparing technical documents with simplified counterparts [39]. A corpus which can be used to identify simplifications made by a human editor is required. Phrase based simplification is

similar to the task of paraphrasing [41], [60], where phrases with high semantic similarity are aligned for use in tasks such as question answering [61] or the automatic evaluation of machine translation [62]. Techniques could be drawn from this area to improve the work in lexical simplification. Two advantages are as follows: Firstly, it allows some rudimentary syntactic simplification to be carried out, altering the structure within a phrase to make it more readable. Secondly, it allows more diversity in the range of simplifications which can be made. It may be the case that simplifying a single word which is part of a complex phrase is actually detrimental to the understanding of that phrase, whereas simplifying the whole phrase itself is helpful.

A recent important development in the field of lexical simplification is the lexical substitution task from SemEval 2012 [45]. Participants designed a system to rank words in terms of their simplicity. The words were given as valid replacements for a single annotated word in a sentence. Many such sentences were provided and systems were able to train and test on sample data before being deployed for the final testing data. The corpus was developed by crowd sourcing through Amazon's Mechanical Turk[2]. Annotators were asked to rank the substitutions in order of their simplicity. These rankings were then combined to form one final ranking.

The SemEval task isolates the synonym ranking problem within lexical simplification where the aim is to find the easiest synonym. Systems do not have to focus on other distractions, such as identifying complex words or synonym generation, but can focus solely on ranking. Several systems were developed to produce these rankings and the techniques used considered a variety of methods such as: language models for word context [63]–[65], decompositional semantics [66] and machine learning techniques [65], [67]. A comprehensive overview and comparison of these is given in the task description [45]. The SemEval task benefits TS in two separate ways: firstly, it has promoted the field and specifically the area of lexical simplification. Hopefully, interest will be generated and more time and resources will be channeled into TS. Secondly, it has provided an evaluation of different methods for synonym ranking. This should drive research forward as new systems will have both a reasonable baseline and evaluation method to compare against.

B. Syntactic Approaches

Syntactic simplification is the technique of identifying grammatical complexities in a text and rewriting these into simpler structures. There are many types of syntactic complexity which this may apply to: Long sentences may be split into their component clauses; Sentences which use the passive voice may be rewritten and anaphora may be resolved. Poorly written texts are very difficult to engage with. Readers may struggle to follow the text, lose interest at some point in a sentence and eventually give up trying. In the case of people with cognitive impairments such as aphasia, some grammar structures may even cause a loss of meaning. Patients may not be able to distinguish between subject and object when the passive voice is used. For example, the passive voice sentence: "the boy was kicked by the girl" may appear to read as: "the boy kicked the girl" for someone with aphasia. A list of research in syntactic simplification is presented in Table III.

[2]www.mturk.com

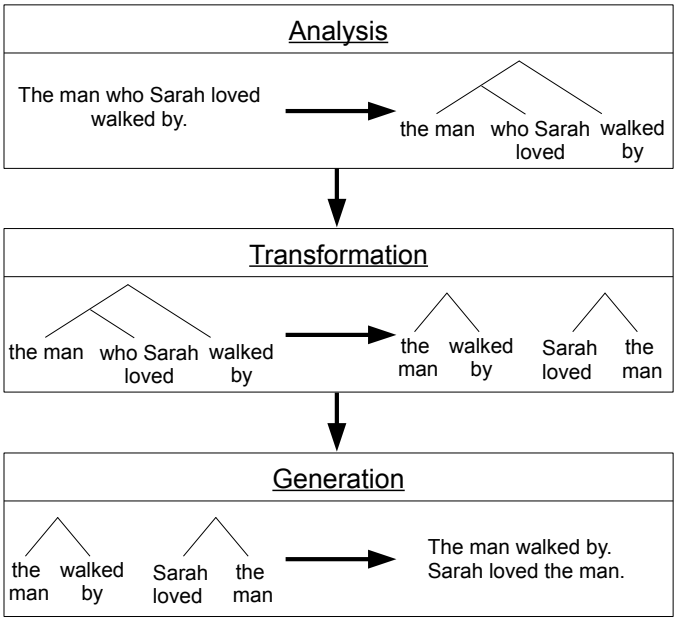

Fig. 3. The syntactic simplification pipeline, with worked example. Predetermined rewrite rules govern the simplifications that occur during the transformation step. The generation step is important to ensure the cohesion of the resultant text.

Work on syntactic simplification began with a system for the automatic creation of rewrite rules for simplifying text [3]. This system takes annotated corpora and learns rules for domain specific sentence simplification. The main purpose is as a preprocessing step to improve other natural language applications. Later work [68], [75], [77], [79] focussed on applying this syntactic simplification as an assistive technology. Improvements to the discourse structure were made to ensure that clauses of sentences appeared in the correct order [18]. More recent work has focussed on applying syntactic simplification as a preprocessing tool for named entity recognition in the biomedical domain [5], [26]. There have also been efforts to apply this technique for languages other than English [51], [69], [71]–[73].

Syntactic simplification is typically done in three phases as shown in Figure 3. Firstly, the text is analysed to identify its structure and parse tree. This may be done at varying granularity, but has been shown to work at a rather coarse level. At this level, words and phrases are grouped together into 'super-tags' which represent a chunk of the underlying sentence. These super-tags can be joined together with conventional grammar rules to provide a structured version of the text. During the analysis phase, the complexity of a sentence is determined to decide whether it will require simplification. This may be done by automatically matching rules, but has also been done using a support vector machine binary classifier [80]. The second phase is transformation, in which modifications are made to the parse tree according to a set of rewrite rules. These rewrite rules perform the simplification operations such as sentence splitting [68], clause

TABLE III. THE STATE OF SYNTACTIC SIMPLIFICATION RESEARCH ORDERED BY YEAR. RECENT EFFORTS HAVE PARTICULARLY SEEN THIS AS APPLIED TO LANGUAGES OTHER THAN ENGLISH.

Year	Title	Notes
1997	Automatic Induction of Rules for Text Simplification [3]	Seminal work in field.
1998	Practical Simplification of English Newspaper Text to Assist Aphasic Readers (PSET) [68]	Shortened sentences for aphasic users.
2004	Automatic Sentence Simplification for Subtitling in Dutch and English [69]	Dutch language simplification
2004	Text Simplification for Information-seeking Applications [6]	Introduce the notion of Easy Access Sentences.
2006	Syntactic Simplification and Text Cohesion [18]	Maintaining discourse when performing syntactic simplification
2009	Sentence Simplification Aids Protein-protein Interaction Extraction [70]	Preprocessing for biomedical interaction recognition.
2010	A Semantic and Syntactic Text Simplification Tool for Health Content [23]	Long sentences split after explanation generation.
2010	Simplifica: a Tool for Authoring Simplified Texts in Brazilian Portuguese Guided by Readability Assessments (PorSimples) [32]	An authoring tool which provides text simplification techniques whilst writing a document.
2012	Acquisition of Syntactic Simplification Rules for French [71]	A comprehensive list of rules for simplifying the French language.
2012	Sentence Splitting for Vietnamese-English Machine Translation [72]	Vietnamese language splitting to improve machine translation.
2012	Transforming Complex Sentences using Dependency Trees for Automatic Text Simplification in Basque [73]	Basque language syntactic simplification.
2012	Enhancing Multi-document Summaries with Sentence Simplification [26]	Syntactic simplification as a preprocessing aid.
2013	ERNESTA: A Sentence Simplification Tool for Children's Stories in Italian [74]	Italian Syntactic Simplification
2013	Enhancing Readability of Web Documents by Text Augmentation for Deaf People [75]	Simplification of Korean for deaf readers.
2013	Sentence Simplification as Tree Transduction [76]	Direct manipulation of parse trees.
2013	Simple, Readable Sub-sentences [77]	Removing unnecessary parts of sentence
2013	Corpus-based Sentence Deletion and Split Decisions for Spanish Text Simplification [78]	Spanish syntactic simplification.

rearrangement [18] and clause dropping [74], [78]. Although techniques for automatically inducing these rules exist [3], most other systems implementing syntactic simplification use hand written rewrite rules. Two reasons for this are the removal of the need for annotated corpora and the improved accuracy of the final rules. After transformation, a regeneration phase may also be carried out, during which further modifications are made to the text to improve cohesion, relevance and readability.

Syntactic simplification is an essential component to any working TS system and has been implemented in both PSET [4] and PorSimples [51] which both seek to provide ubiquitous TS as an assistive technology. It has been particularly useful outside this application and has been implemented for improving the accuracy of other natural language techniques with significant success. Syntactic simplification will be incorporated into future TS systems, as it has the ability to reduce grammatical complexities in a way which is not possible with other techniques. Creation and validation of the rewrite rules is a difficult process and one aspect of further work may concentrate on new techniques to automatically discover these.

C. Explanation Generation

Explanation generation is the technique of taking a difficult concept in a text and augmenting it with extra information, which puts it into context and improves user understanding. Table IV lists the research in this area. It has been shown that in some cases, this is more appropriate than lexical simplification [83]. A specific example can be taken from

TABLE IV. RESEARCH INTO EXPLANATION GENERATION, ORDERED BY YEAR.

Year	Title	Notes
2006	SIMTEXT Text Simplification of Medical Literature [81]	Dictionary definitions appended for some terms.
2009	FACILITA: Reading Assistance for Low-literacy Readers (PorSimples) [40]	References to Wikipedia articles for difficult terms.
2010	A Semantic and Syntactic Text Simplification Tool for Health Content [23]	Long sentences split after explanation generation.
2012	Sense-specific Lexical Information for Reading Assistance [82]	Lexical elaboration for the education of second language learners.

Health Informatics, where explanations are generated for terms in health literature [23], [81]. These are categorised by their semantic type (disease name, anatomical structure, device, etc.) Explanations are then generated by finding an easier term and adding in a short connecting phrase to explain the more complex term:

"Pulmonary atresia *(a type of birth defect)*"[3]

'Pulmonary atresia' is found to be semantically related to 'birth defect' and the connecting phrase 'a type of' is added to maintain cohesion. Five semantic types are identified and a medical thesaurus is employed for identifying valid substitutions. This is highly specific to the medical terminology in question. The semantic types were discovered by manually analysing simplified literature and so applications to the general case would require much analytical work and many more categories to be discovered. Whilst analysis could be automated, the accuracy would then suffer. This technique could also be used in another equally specific technical domain.

A more general form of simplification is carried out as part of the PorSimples project [51]. The application 'Educational FACILITA' [84], provides a browser plug-in which can simplify Web content for users. Named entities are recognised and annotated with short explanatory extracts from Wikipedia articles. These allow the user to learn more about the difficult concepts in a text. This information is presented to the user in a separate text box at their request. Lexical simplification is also carried out to improve the text's overall readability. The largest challenge lies in the named entity labelling task. Here, words must be matched to their semantic concepts. This is a difficult task which requires word sense disambiguation and some mapping between the concepts and their explanations.

More recently, this has been applied to the case of second language learners [82]. Here, the learner has the opportunity to highlight words which they find difficult and see a dictionary entry for that word. Word sense disambiguation (as discussed above in Section II-A) is carried out to ensure that only the correct sense of the word is presented to the user. This is shown to increase the language learner's reading comprehension for the explained words.

In its present form, explanation generation has particular potential for users with some understanding who wish to learn more about a text. By providing explanations alongside difficult terms, the user is able to better understand the concept and will hopefully not require the explanation next time they encounter the complexity. Explanation generation is not confined to presentation alongside the complex terms however and may also be done to replace the original word. A semantically simple phrase which explains the original term could be used as its replacement. Due to the potentially complex levels of processing involved, this is a technique which is prone to error. If errors occur and are left undetected and unresolved, then they may result in the final text becoming misleading and unhelpful to an end user, which should naturally be avoided wherever possible. This technique may also be useful when deployed alongside lexical [84] or syntactic simplification [23]. The explanations which are generated may add to the structural complexity of the text, resulting in diminished readability. Any steps to increase the readability will help the reader to interact with the text.

[3]From [23]. Generated explanation in italics

D. Statistical Machine Translation

Automated machine translation is an established technique in natural language processing, for a comprehensive review see [88]. It involves automatic techniques to convert the lexicon and syntax of one language to that of another, resulting in translated text. Its application to TS involves casting our problem as a case of monolingual text-to-text generation. Table V gives the research in this field to date. We consider our translation task as that of converting from the source language of complex English to the target of simple English. Each has its own unique syntax and lexicon and is sufficiently distinct to permit the use of machine translation techniques. Recent research (as described below) in machine translation has focussed on phrase based statistical techniques. These learn valid translations from large aligned bilingual corpora and are then able to apply these to novel texts. This task is made easier as the source and target languages are very similar, and so few changes are necessary. It is this type of machine translation that has been applied to TS.

Work to perform TS by statistical machine translation has been performed for English [10], [19], [20], Brazilian Portuguese [35] and has been proposed for German [87]. Practically, systems often use and modify a standard statistical machine translation tool such as Moses [89]. A difficult task can be finding aligned sentences in complex and simple language. This has been done by manual creation [35] and by mining English and Simple Wikipedia [19] using techniques from monolingual corpus alignment [90].

Using this corpus, Moses has been applied to the TS task for English [10]. Moses was augmented with a phrase deletion module which removed unnecessary parts of the complex

TABLE V. PAPERS PRESENTING TS BY STATISTICAL MACHINE TRANSLATION ORDERED BY YEAR.

Year	Title	Notes
2010	Translating from Complex to Simplified Sentences (PorSimples) [35]	Part of the PorSimples project for TS in Brazilian Portuguese
2010	A Monolingual Tree-based Translation Model for Sentence Simplification [19]	Tree based model, produced the PWKP dataset of aligned complex-simple sentences from Wikipedia.
2011	Learning to Simplify Sentences using Wikipedia [10]	Improves previous work.
2012	Sentence Simplification by Monolingual Machine Translation [20]	Further improves on previous work.
2012	A Simplification Translation Restoration Framework for Cross-domain SMT Applications [85]	Chinese – English. Simplification as processing aid.
2013	Statistical Machine Translation with Readability Constraints [86]	English – Swedish. Simplification for improved readability.
2013	Building a German/Simple German Parallel Corpus for Automatic Text Simplification [87]	A corpus for the production of German monolingual statistical machine translation simplification.

source text. The evaluation used BLEU [91], a standard measure in machine translation. Recent research [20] has used human judges to evaluate the quality of the simplified text against a lexical substitution baseline, something which has not been done before. The use of human judges is a valuable method for evaluation in TS.

A recent development has been simplification in more traditional bilingual statistical machine translation, which has occurred for translating English to Chinese [85] and Swedish to English [86]. Simplification aids readability in the target language, making it useful for language learners. It is also useful to transform source and target texts to a common format to improve their alignment, thus improving translation accuracy. In the example of English to Chinese translation, the final text is restored to it's original level of complexity, the simplification is only required for improving the quality of the translation.

As new techniques and evaluation methods are developed for machine translation, they will be directly applicable to this task. Simplification through monolingual machine translation gives a form of simplified text which appears to reflect human simplified text. This may be useful when simplifying for different domains as the types of simplification are automatically learnt by the algorithm. Statistical machine translation is a technique with applications in real world systems. However, the nature of a statistical technique is that it will not work perfectly in every single case. Every statistical technique has a number of false positives (simplification operations made in error) and false negatives (simplifications which should have been made). Whilst the aim is to reduce these at the same time as improving the levels of true positives and negatives, there will always be some errors that creep in during the learning process. As discussed previously (see Section I), the introduction of errors results in a diminished understandability and increased text complexity — the opposite to the desired outcome. This highlights the importance of accuracy and output validation in TS.

E. Non-English Approaches

As with many natural language processing applications, the majority of TS research is conducted solely for the English language. However, TS is also applied across many different languages as shown in Table VI. The KURA project [37] worked on Japanese language simplification for deaf students and introduced the concept of phrase based simplification identifying and simplifying complex terms. Similarly, the PorSimples project has contributed much to the wider field of TS. This is undoubtedly the largest TS project to date with 3 main systems and many types of simplification investigated. The Simplext project is an ongoing project, currently in the process of developing simplification tools and resources for Spanish. It has particularly focussed on the application of simplification for dyslexic readers.

Most projects do not focus on introducing new techniques for TS, but instead focus on implementing existing techniques in their own language. This is an interesting challenge, as language specific characteristics make it non-trivial to re-implement existing techniques. The main barrier is usually in discovering appropriate resources for the language. For lexical simplification, an extensive word frequency list and some electronic thesaurus is usually employed. If no such word frequency list exists, this may be easily calculated from a count of a sufficiently large corpus (such as pages from Wikipedia). Syntactic simplification typically requires more work to be done. The differences between simplified text and complex text in the language must be analysed to discover language specific simplification rules. These will typically not be transferable between languages due to differing grammar structures. Some constructs such as passive voice and WH-phrases may be common points of confusion across languages and so research may be aided by identifying these known complexities. Techniques to learn these automatically [3] and statistical machine translation may be of use here.

It can be seen from Table VI that recent times have seen a proliferation in TS techniques in languages that are not English. Of the fourteen systems presented, eight have publications in 2012-13. This may be in part due to projects

TABLE VI. A TABLE OF TS IN DIFFERENT LANGUAGES. SS = SYNTACTIC SIMPLIFICATION. LS = LEXICAL SIMPLIFICATION. EG = EXPLANATION GENERATION. PBMT = PHRASE BASED MACHINE TRANSLATION. INFORMATION IS OMITTED WHERE UNAVAILABLE

Year	Language	Methodology	Notes
2003	Japanese (KURA) [37]	LS	No continuing work evident
2004	Dutch [69]	SS	Completed study
2007–10	Portuguese (PorSimples) [51]	SS, LS, EG	Completed study
2010–2013	Spanish (Simplext) [42]	LS	Ongoing work
2011	Italian (Read-it) [92]	LS, SS	No continuing work evident
2012	French [71]	SS	Present a set of syntactic rules
2012	Bulgarian (FIRST) [93]		Preliminary study
2012	Danish (DSIM) [94]	PBMT	Aligned corpus for training
2012	Swedish [47]	LS	Masters Thesis
2012	Vietnamese [72]	SS	Preprocessing for Machine Translation
2013	Basque [73]	SS	Preliminary study
2013	Italian [74] (ERNESTA)	SS	Simplification of children's stories.
2013	Korean [75]	SS	Simplification for sign language users
2013	German [87]	PBMT	Aligned corpus for training

such as PorSimples and Simplext publicising TS as a research field, especially for non-English natural language processing research.

III. RESEARCH CHALLENGES

Throughout this survey, many open areas have been identified and this Section will gather these together and suggest future directions for research. These directions have been grouped into three categories: resources, systems and techniques. Section III-A describes the need for novel evaluation methods and corpora for the further development of existing TS systems. Section III-B outlines the need for TS systems and some methods for the deployment of these. Section III-C explains the need for the development of new algorithms in the field.

A. Resources

Resources are the foundation upon which a system is built. TS has seen many different approaches to the task of providing resources such as evaluation methods and corpora. These have often been done with little consideration to prior techniques and so one general aspect of future work is the comparison and evaluation of potential resources.

Current techniques for automatically evaluating readability are of limited use in TS research. A strong contribution to the field would be an automatic evaluation measure which reliably reported the effects of TS. Some progress has been made [95]. However, this is only useful for the highly specific task of ordering synonyms in terms of their complexity. This is very useful when evaluating a system designed for the specific task, but not as useful for the general task of TS. An evaluation method is needed which has the generality of a readability formula [29], [30], [96] but with the specificity and speed of an automated measure [95].

Manual techniques for the evaluation of automatic TS may also be investigated and developed. Whilst automated techniques give some impression as to the efficacy of a system, they are a step removed from the actual intended audience and so will never be as accurate as direct user evaluation. Many authors have used some manual evaluation for their results [4], [23], [31], [37] and research should aim towards this, especially when deploying a TS system for a specific user group. Experiments to determine the best manual methods of evaluation may also take place.

In addition to research on the evaluation methods, the development of new corpora is equally paramount to the progression of the field. As there are different approaches to TS (see Section II), different types of corpora are necessary. Simplification is inherently difficult to evaluate as there is no obvious correct answer. This means that a corpus cannot be in the standard format of a set of problems labelled with their solutions. Instead, more abstract corpora must be developed to address specific evaluation needs within the TS domain. As these are developed, evaluation methods will be developed alongside them. Corpora which draw on human annotation and where possible the input of the eventual users of a TS system will be more effective than those that do not.

One promising method for corpus development and evaluation comes from the field of statistical machine translation. Some authors have formulated TS as a monolingual translation problem [10], [19], [20]. This creates the possibility of using machine translation evaluation methods such as BLEU [91] and NIST [36]. These techniques compare a given translation

with a reference translation and report an accuracy measure based on the co-occurring words across the two translations. These may also be applicable to the wider scope of TS where sample simplifications could be compared to one or many reference texts. As machine translation evaluation techniques are advanced, the benefits may also be reaped by the text simplification community.

B. Systems

Another research challenge is the development of TS applications. These will exist as a layer of assistive technology upon information gathering systems. There are two clear options for the development of publicly available TS systems, as outlined below.

Firstly, TS can be applied at the user's level. In this model, the user receives some complex text which they automatically simplify by some means. This could take on the form of a Web browser plug-in which allows the user to select and simplify text (similar to the the FACILITA project for Brazilian Portuguese [40]). This could also take on the form of an application which allows the user to identify text and simplify. Some users may not even require the choice to simplify text. For example, in the context of browsing the Internet, some users may find the complex text which is presented to them at first distracting, demoralising or off-putting. Here, it may be helpful to automatically reduce the complexity of any text on a webpage before presenting it to a user.

Secondly, TS may be applied by the author to a text he is creating [97]. In this model, the author may write a document and then use automatic techniques to identify any complexities and to automatically simplify or receive suggestions as to simplifications he may apply. The main advantage is that the author can check the quality of simplifications before the text is presented to a user. Grammaticality, cohesion and intended meaning are definitely preserved, whilst understandability and readability are increased. This is useful in many different applications where text is being written for audiences who may not necessarily understand the final product. The research challenge here is to develop helpful ways of doing this which allow an author to target his text to many levels of understanding.

TS is currently not a commercialised application. This may be in part due to low accuracy in test systems and the youth of the field. As work is done to increase the accuracy of TS systems, they will become more commercially viable. TS is a useful product which can be packaged and sold in the form of software and Web services. As an industry develops around TS, this will create interest in the area which will drive the field to further developments.

C. Techniques

The identification and evaluation of new techniques is paramount to the progression of the field, as the potential solution space for TS is currently sparsely explored. This is mainly because previous projects have been limited by the resources available. As more TS research applications are developed, a few underexplored areas for focus are as follows. These are not intended as an exhaustive list of all the potential future work in TS, but instead to highlight some areas which may be of future interest. These have all been explored initially and references are provided as appropriate.

Firstly, word sense disambiguation is highly important for lexical simplification. Initial work ignored ambiguity in the hope that complex words would belong to only one potential sense. This has not been the case and word sense errors (where a synonym with a drastically different meaning is selected) are a common problem among lexical substitution systems. Some work has previously addressed this [12]–[14], however future efforts must focus on incorporating state of the art word sense disambiguation techniques and adapting these for best use within the TS context. This may be implemented at the synonym ranking step of lexical substitution to combine the simplicity score with a 'relevance' score produced by a disambiguation system. Words which are of low relevance in a context will make the text less understandable.

Secondly, work should be undertaken to improve techniques for identifying candidates for simplification within a text. Whilst there has been plenty of work into readability measures, little has been transferred to a TS setting, although exceptions do exist [21], [80]. Machine learning techniques hold some promise and should be investigated further. The existing techniques are for sentence level simplification and further work could focus on candidate identification at the lexical level. This would involve looking at features of given words and developing some classification system to identify those of sufficient complexity to require simplification.

IV. Conclusion

TS is a domain which has emerged as a reaction to difficult texts. This has occurred for different applications such as preprocessing for machine translation [72] and assistive technology for people with Aphasia [4]. These applications promise to reduce the complexity of text whilst improving readability and understandability. This is a highly useful task and is highly applicable in many settings such as second language learners and lay readers of technical documents. TS is not solely confined to the reader, it may also be applied by the author to a text in order to ensure his point is clearly communicated, or even in a natural language processing pipeline to improve the performance of later components.

There are also many approaches to the task. Some focus on the lexical level, replacing complex words with simpler synonyms. Some modify the syntax of a text to remove complex grammatical structures. Yet others perform phrase based machine translation in an effort to automatically learn valid methods of simplification. The field is currently seeing a wave of growth with many new research projects and new approaches being developed. As the field progresses, more techniques will become available and TS will be widely distributed.

TS is on its way to becoming a household application. As it does so, it is likely that people will often not even know they are benefitting from it. Campaigns for simplified English have existed for many years. TS offers an answer.

References

[1] S. Crossley, D. Allen, and D. McNamara, "Text simplification and comprehensible input: A case for an intuitive approach," *Language Teaching Research*, vol. 16, no. 1, pp. 89–108, 2012.

[2] D. J. Young, "Linguistic simplification of SL reading materials: Effective instructional practice?" *Modern Language Journal*, vol. 83, no. 3, pp. pp. 350–366, 1999.

[3] R. Chandrasekar and B. Srinivas, "Automatic induction of rules for text simplification," *Knowledge-Based Systems*, vol. 10, no. 3, pp. 183 – 190, 1997.

[4] S. Devlin and J. Tait, "The use of a psycholinguistic database in the simplification of text for aphasic readers." *Linguistic Databases*, pp. 161–173, 1998.

[5] S. Jonnalagadda and G. Gonzalez, "BioSimplify: an open source sentence simplification engine to improve recall in automatic biomedical information extraction," in *Annual Proceedings of AMIA 2010*, November 2010, pp. 13–17.

[6] B. B. Klebanov, K. Knight, and D. Marcu, "Text simplification for information-seeking applications," in *On the Move to Meaningful Internet Systems, Lecture Notes in Computer Science*. Springer Verlag, 2004, pp. 735–747.

[7] D. Vickrey and D. Koller, "Applying sentence simplification to the conll-2008 shared task," in *Proceedings of the Twelfth Conference on Computational Natural Language Learning*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 268–272.

[8] L. Feng, "Text simplication: A survey," CUNY, Tech. Rep., March 2008.

[9] M. Yatskar, B. Pang, C. Danescu-Niculescu-Mizil, and L. Lee, "For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 365–368.

[10] W. Coster and D. Kauchak, "Learning to simplify sentences using Wikipedia," in *Proceedings of the Workshop on Monolingual Text-To-Text Generation*. Portland, Oregon: Association for Computational Linguistics, June 2011, pp. 1–9.

[11] C. Napoles and M. Dredze, "Learning Simple Wikipedia: A cogitation in ascertaining abecedarian language," in *NAACL HLT 2010 Workshop on Computational Linguistics and Writing*, 2010, pp. 42–50.

[12] O. Biran, S. Brody, and N. Elhadad, "Putting it simply: a context-aware approach to lexical simplification," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 496–501.

[13] J. De Belder, K. Deschacht, and M.-F. Moens, "Lexical simplification," in *1st International Conference on Interdisciplinary Research on Technology, Education and Communication*, 2010.

[14] J. De Belder and M. Moens, "Text simplification for children," in *Prroceedings of the SIGIR workshop on accessible search systems*, 2010, pp. 19–26.

[15] S. Blum and E. A. Levenston, "Universals of lexical simplification," *Language Learning*, vol. 28, no. 2, pp. 399–415, 1978.

[16] J. E. Hoard, R. Wojcik, and K. Holzhauser, "An automated grammar and style checker for writers of simplified English," in *Computers and Writing*. Springer Netherlands, 1992, pp. 278–296.

[17] G. Adriaens, "Simplified English grammar and style correction in an MT framework: the LRE SECC project," in *Aslib proceedings*, vol. 47. MCB UP Ltd, 1995, pp. 73–82.

[18] A. Siddharthan, "Syntactic simplification and text cohesion," *Research on Language & Computation*, vol. 4, pp. 77–109, 2006.

[19] Z. Zhu, D. Bernhard, and I. Gurevych, "A monolingual tree-based translation model for sentence simplification," in *Proceedings of the 23rd International Conference on Computational Linguistics*, 2010, pp. 1353–1361.

[20] S. Wubben, A. van den Bosch, and E. Krahmer, "Sentence simplification by monolingual machine translation," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Jeju Island, Korea: Association for Computational Linguistics, July 2012, pp. 1015–1024.

[21] K. Woodsend and M. Lapata, "Learning to simplify sentences with quasi-synchronous grammar and integer programming," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 409–420.

[22] S. E. Petersen and M. Ostendorf, "Text simplification for language learners: A corpus analysis," in *Speech and Language Technology for Education workshop*, 2007.

[23] S. Kandula, D. Curtis, and Q. Zeng-Treitler, "A semantic and syntactic text simplification tool for health content," in *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, 2010, pp. 366–370.

[24] H. Jing, "Sentence reduction for automatic text summarization," in *Proceedings of the sixth conference on Applied natural language processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2000, pp. 310–315.

[25] C. Blake, J. Kampov, A. K. Orphanides, D. West, and C. Lown, "Query expansion, lexical simplification and sentence selection strategies for multi-document summarization," in *Document understanding conference (DUC-2007)*, 2007.

[26] S. Silveira and A. Branco, "Combining a double clustering approach with sentence simplification to produce highly informative multi-document summaries," in *Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference*, Aug 2012, pp. 482–489.

[27] J. M. Conroy, J. G. Stewart, and J. D. Schlesinger, "Classy query-based multi-document summarization," in *Proceedings of the Document Understanding Conf. Wksp. 2005 (DUC 2005) at the Human Language Technology Conf./Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, 2005.

[28] A. Siddharthan, A. Nenkova, and K. Mckeown, "Syntactic Simplification for Improving Content Selection in Multi-Document Summarization," in *In Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004*, 2004, pp. 896–902.

[29] J. Kincaid, R. Fishburne, R. Rogers, and B. Chissom, "Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for navy enlisted personnel," *Research Branch report*, 1975.

[30] G. H. McLaughlin, "Smog grading: A new readability formula," *Journal of reading*, vol. 12, no. 8, pp. 639–646, 1969.

[31] C. Napoles, B. Van Durme, and C. Callison-Burch, "Evaluating sentence compression: Pitfalls and suggested remedies," in *Proceedings of the Workshop on Monolingual Text-To-Text Generation*. Portland, Oregon: Association for Computational Linguistics, June 2011, pp. 91–97.

[32] C. Scarton, M. de Oliveira, A. Candido, Jr., C. Gasperin, and S. M. Aluísio, "Simplifica: a tool for authoring simplified texts in brazilian portuguese guided by readability assessments," in *Proceedings of the NAACL HLT 2010 Demonstration Session*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 41–44.

[33] R. Barzilay and N. Elhadad, "Sentence alignment for monolingual comparable corpora," in *Proceedings of the 2003 conference on Empirical methods in natural language processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 25–32.

[34] S. Bott and H. Saggion, "An unsupervised alignment algorithm for text simplification corpus construction," in *Proceedings of the Workshop on Monolingual Text-To-Text Generation*. Portland, Oregon: Association for Computational Linguistics, June 2011, pp. 20–26.

[35] L. Specia, "Translating from complex to simplified sentences," in *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language*. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 30–39.

[36] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in *Proceedings of the second international conference on Human Language Technology Research*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002, pp. 138–145.

[37] K. Inui, A. Fujita, T. Takahashi, R. Iida, and T. Iwakura, "Text simplification for reading assistance: A project note," in *Proceedings of the Second International Workshop on Paraphrasing*. Sapporo, Japan: Association for Computational Linguistics, July 2003, pp. 9–16.

[38] S. Devlin and G. Unthank, "Helping aphasic people process online information," in *Proceedings of the 8th international ACM SIGACCESS conference on computers and accessibility*. New York, NY, USA: ACM, 2006, pp. 225–226.

[39] N. Elhadad and K. Sutaria, "Mining a lexicon of technical terms and lay equivalents," in *Proceedings of the Workshop on BioNLP*

2007: Biological, Translational, and Clinical Language Processing. Association for Computational Linguistics, 2007, pp. 49–56.

[40] W. M. Watanabe, A. C. Junior, V. R. Uzêda, R. P. d. M. Fortes, T. A. S. Pardo, and S. M. Aluísio, "Facilita: reading assistance for low-literacy readers," in *Proceedings of the 27th ACM international conference on Design of communication*. New York, NY, USA: ACM, 2009, pp. 29–36.

[41] L. Deléger and P. Zweigenbaum, "Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora," in *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*. Association for Computational Linguistics, 2009, pp. 2–10.

[42] S. Bott, L. Rello, B. Drndarević, and H. Saggion, "Can Spanish be simpler? LexSiS: Lexical simplification for Spanish," in *Proceedings of COLING 2012*. Mumbai, India: The COLING 2012 Organizing Committee, December 2012, pp. 357–374.

[43] B. Drndarević and H. Saggion, "Towards automatic lexical simplification in Spanish: An empirical study," in *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*. Montréal, Canada: Association for Computational Linguistics, June 2012, pp. 8–16.

[44] H. Saggion, S. Bott, and L. Rello, "Comparing resources for Spanish lexical simplification," in *Statistical Language and Speech Processing*, ser. Lecture Notes in Computer Science, A.-H. Dediu, C. Martín-Vide, R. Mitkov, and B. Truthe, Eds. Springer, Berlin Heidelberg, 2013, vol. 7978, pp. 236–247.

[45] L. Specia, S. K. Jauhar, and R. Mihalcea, "SemEval-2012 task 1: English lexical simplification," in **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*. Montréal, Canada: Association for Computational Linguistics, 7-8 June 2012, pp. 347–355.

[46] S. R. Thomas and S. Anderson, "WordNet-based lexical simplification of a document," in *Proceedings of KONVENS 2012*, J. Jancsary, Ed. ÖGAI, September 2012, pp. 80–88.

[47] R. Keskisärkkä, "Automatic text simplification via synonym replacement," Ph.D. dissertation, Linköping, 2012.

[48] G. Leroy, J. E. Endicott, D. Kauchak, O. Mouradi, and M. Just, "User evaluation of the effects of a text simplification algorithm using term familiarity on perception, understanding, learning, and information retention," *Journal of Medical Internet Research*, vol. 15, no. 7, p. e144, 2013.

[49] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[50] H. Kučera and W. N. Francis, *Computational analysis of present-day American English*. Providence, RI: Brown University Press, 1967.

[51] S. M. Aluísio and C. Gasperin, "Fostering digital inclusion and accessibility: the PorSimples project for simplification of Portuguese texts," in *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 46–53.

[52] S. Bautista, R. Hervás, P. Gervás, R. Power, and S. Williams, "A system for the simplification of numerical expressions at different levels of understandability." in *Worshop Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, Atlanta, USA, 06/2013 2013.

[53] L. Rello, S. Bautista, R. Baeza-Yates, P. Gervás, R. Hervás, and H. Saggion, "One half or 50%? an eye-tracking study of number representation readability," in *Human-Computer Interaction INTERACT 2013*, ser. Lecture Notes in Computer Science, P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, and M. Winckler, Eds. Springer Berlin Heidelberg, 2013, vol. 8120, pp. 229–245.

[54] M. Brysbaert and B. New, "Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English," *Behavior Research Methods*, vol. 41, no. 4, pp. 977–990, 2009.

[55] T. Brants and A. Franz, "Web 1T 5-gram corpus version 1.1," *Linguistic Data Consortium*, 2006.

[56] K. Deschacht and M. Moens, "The latent words language model," in *Proceedings of the 18th Annual Belgian-Dutch Conference on Machine Learning*, 2009.

[57] L. Rello, R. Baeza-Yates, L. Dempere-Marco, and H. Saggion, "Frequent words improve readability and short words improve understandability for people with dyslexia," in *Human-Computer Interaction INTERACT 2013*, ser. Lecture Notes in Computer Science, P. Kotz, G. Marsden, G. Lindgaard, J. Wesson, and M. Winckler, Eds. Springer Berlin Heidelberg, 2013, vol. 8120, pp. 203–219.

[58] D. Kauchak, "Improving text simplification language modeling using unsimplified text data," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 1537–1546.

[59] A. Ligozat, C. Grouin, A. Garcia-Fernandez, and D. Bernhard, "Approches à base de fréquences pour la simplification lexicale," in *Actes de TALN'2013 : 20e conférence sur le Traitement Automatique des Langues Naturelles*, vol. 1, Les Sables d'Olonne, France, 2013, pp. 493–506.

[60] C. Quirk, C. Brockett, and W. Dolan, "Monolingual machine translation for paraphrase generation," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 2004, pp. 142–149.

[61] F. Rinaldi, J. Dowdall, K. Kaljurand, M. Hess, and D. Mollá, "Exploiting paraphrases in a question answering system," in *Proceedings of the second international workshop on Paraphrasing - Volume 16*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 25–32.

[62] D. Kauchak and R. Barzilay, "Paraphrasing for automatic evaluation," in *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 455–462.

[63] A. Ligozat, C. Grouin, A. Garcia-Fernandez, and D. Bernhard, "Annlor: A naïve notation-system for lexical outputs ranking," in **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*. Montréal, Canada: Association for Computational Linguistics, 7-8 June 2012, pp. 487–492.

[64] R. Sinha, "Unt-simprank: Systems for lexical simplification ranking," in **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*. Montréal, Canada: Association for Computational Linguistics, 7-8 June 2012, pp. 493–496.

[65] S. K. Jauhar and L. Specia, "UOW-SHEF: SimpLex – lexical simplicity ranking based on contextual and psycholinguistic features," in **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*. Montréal, Canada: Association for Computational Linguistics, 7-8 June 2012, pp. 477–481.

[66] M. Amoia and M. Romanelli, "Sb: mmsystem - using decompositional semantics for lexical simplification," in **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*. Montréal, Canada: Association for Computational Linguistics, 7-8 June 2012, pp. 482–486.

[67] A. Johannsen, H. Martínez, S. Klerke, and A. Søgaard, "Emnlp@cph: Is frequency all there is to simplicity?" in **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*. Montréal, Canada: Association for Computational Linguistics, 7-8 June 2012, pp. 408–412.

[68] J. Carroll, G. Minnen, Y. Canning, S. Devlin, and J. Tait, "Practical simplification of English newspaper text to assist aphasic readers," in *Proceedings of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, 1998, pp. 7–10.

[69] W. Daelemans, A. Höthker, and E. Sang, "Automatic sentence simplification for subtitling in Dutch and English," in *Proceedings of the 4th International Conference on Language Resources and Evaluation*, 2004, pp. 1045–1048.

[70] S. Jonnalagadda and G. Gonzalez, "Sentence simplification aids protein-

protein interaction extraction," in *The 3rd International Symposium on Languages in Biology and Medicine*, November 2009, pp. 8–10.

[71] V. Seretan, "Acquisition of syntactic simplification rules for french," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012.

[72] B. T. Hung, N. L. Minh, and A. Shimazu, "Sentence splitting for Vietnamese-English machine translation," in *Knowledge and Systems Engineering (KSE), 2012 Fourth International Conference*, August 2012, pp. 156 –160.

[73] M. J. Aranzabe, A. D. de Ilarraza, and I. Gonzalez-Dios, "Transforming complex sentences using dependency trees for automatic text simplification in Basque," *Procesamiento del Lenguaje Natural*, vol. 50, pp. 61–68, 2012.

[74] G. Barlacchi and S. Tonelli, "Ernesta: A sentence simplification tool for childrens stories in Italian," in *Computational Linguistics and Intelligent Text Processing*, ser. Lecture Notes in Computer Science, A. Gelbukh, Ed. Springer Berlin Heidelberg, 2013, vol. 7817, pp. 476–487.

[75] J.-W. Chung, H.-J. Min, J. Kim, and J. C. Park, "Enhancing readability of web documents by text augmentation for deaf people," in *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*. New York, NY, USA: ACM, 2013, pp. 30:1–30:10.

[76] D. Feblowitz and D. Kauchak, "Sentence simplification as tree transduction," in *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 1–10.

[77] S. Klerke and A. Søgaard, "Simple, readable sub-sentences," in *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 142–149.

[78] S. Štajner, B. Drndarević, and H. Saggion, "Corpus-based sentence deletion and split decisions for Spanish text simplification," *Revista Computación y Sistemas; Vol. 17 No. 2*, 2013.

[79] B. Drndarević, S. Štajner, S. Bott, S. Bautista, and H. Saggion, "Automatic text simplification in spanish: A comparative evaluation of complementing modules," in *Computational Linguistics and Intelligent Text Processing*, ser. Lecture Notes in Computer Science, A. Gelbukh, Ed. Springer Berlin Heidelberg, 2013, vol. 7817, pp. 488–500.

[80] C. Gasperin, L. Specia, T. Pereira, and S. M. Aluísio, "Learning when to simplify sentences for natural text simplification." in *Encontro Nacional de Inteligência Artificial*, 2009, pp. 809–818.

[81] J. Jan, S. Damay, G. Jaime, D. Lojico, D. B. Tarantan, and E. C. Ong, "Simtext text simplification of medical literature," in *3rd National Natural Language Processing Symposium - Building Language Tools and Resources*, 2006, pp. 34–38.

[82] S. Eom, M. Dickinson, and R. Sachs, "Sense-specific lexical information for reading assistance," in *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. Montréal, Canada: Association for Computational Linguistics, June 2012, pp. 316–325.

[83] L. Rello, R. Baeza-Yates, S. Bott, and H. Saggion, "Simplify or help?: text simplification strategies for people with dyslexia," in *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*. New York, NY, USA: ACM, 2013, pp. 15:1–15:10.

[84] W. M. Watanabe, A. Candido, Jr., M. A. Amâncio, M. de Oliveira, T. A. S. Pardo, R. P. M. Fortes, and S. M. Aluísio, "Adapting web content for low-literacy readers by using lexical elaboration and named entities labeling," in *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)*. New York, NY, USA: ACM, 2010, pp. 8:1–8:9.

[85] H.-B. Chen, H.-H. Huang, H.-H. Chen, and C.-T. Tan, "A simplification-translation-restoration framework for cross-domain SMT applications," in *Proceedings of COLING 2012*. Mumbai, India: The COLING 2012 Organizing Committee, December 2012, pp. 545–560.

[86] S. Stymne, J. Tiedemann, C. Hardmeier, and J. Nivre, "Statistical machine translation with readability constraints," in *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); Linkoping Electronic Conference Proceedings*, 2013, pp. 375–386.

[87] D. Klaper, S. Ebling, and M. Volk, "Building a German/simple German parallel corpus for automatic text simplification," in *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 11–19.

[88] A. Lopez, "Statistical machine translation," *ACM Comput. Surv.*, vol. 40, no. 3, pp. 8:1–8:49, Aug. 2008.

[89] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2007, pp. 177–180.

[90] R. Nelken and S. Shieber, "Towards robust context-sensitive sentence alignment for monolingual corpora," in *Proceedings of EACL 2006, the 11th Conference of the European Chapter of the ACL*, Trento, Italy, April 2006, pp. 3–7.

[91] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 311–318.

[92] F. Dell'Orletta, S. Montemagni, and G. Venturi, "Read-it: assessing readability of Italian texts with a view to text simplification," in *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 73–83.

[93] S. Štajner, R. Evans, C. Orasan, and R. Mitkov, "What can readability measures really tell us about text complexity?" in *Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, 2012, pp. 14–21.

[94] S. Klerke and A. Søgaard, "Dsim, a Danish parallel corpus for text simplification," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 4015–4018.

[95] J. De Belder and M. Moens, "A dataset for the evaluation of lexical simplification," in *Computational Linguistics and Intelligent Text Processing*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2012, vol. 7182, pp. 426–437.

[96] E. Dale and J. Chall, "A formula for predicting readability: Instructions," *Educational research bulletin*, pp. 37–54, 1948.

[97] A. Max, "Writing for language-impaired readers," *Computational Linguistics and Intelligent Text Processing*, pp. 567–570, 2006.