Procesamiento del Lenguaje Natural, Revista nº 47 septiembre de 2011, pp 87-95 recibido 03-05-2011 aceptado 25-05-2011

Spanish Text Simplification: An Exploratory Study

Simplificación de textos en Español: Un estudio explorativo

Stefan Bott
Universitat Pompeu Fabra
Barcelona
stefan.bott@upf.edu

Horacio Saggion
Universitat Pompeu Fabra
Barcelona
horacio.saggion@upf.edu

Resumen: La simplificación de textos tiene como objetivo la transformación de un texto en una versión equivalente que es mas fácil de leer. En este artículo presentamos un estudio de operaciones de simplificación de textos para el español. Este estudio tiene consecuencias importantes para el desarrollo de un sistema automático.
Palabras clave: Simplificación de textos; operaciones de simplificación; alineación de oraciones

Abstract: Text simplification is the process of transforming a text into an equivalent which is more understandable for a target user. We focus on text simplification in the Spanish language and present a corpus-based study of simplification operations. The study has implications for the development of an automatic simplification system.
Keywords: Text Simplification; Simplification Operations; Sentence Alignment

1 *Introduction*

The question of whether a text is easy to read and understand depends very much on the abilities and experience of the reader. Some people can read official documents while others will find it difficult to understand short texts in popular newspapers or magazines. Even if the concept of ”easy-to-read” is not universal it's possible, taking the profile of the readers into account, to write a text that will suit the abilities of people with literacy and comprehension problems. Universal access to textual contents constitutes a fundamental human right; however it is far from being a reality, specially if we consider people with a cognitive disability or a linguistic handicap. It is estimated that at least 5% of the world population is functional illiterate due to such reasons. We therefore believe it is of paramount importance to develop natural language processing tools to allow easy access to textual information. Various organizations employ well-known techniques to prepare documentation or transform already available documentation in such a way that it becomes more readable by a target user group. For example the Easy-to-Read (Petz and Tronbacke, 2008) method is being used in several countries to produce texts which are adapted or simplified to particular user groups. Conventional text simplification requires a heavy load in human resources, a fact that limits the number of simplified digital content accessible today making it practically impossible to easily access already available legacy material for specific target groups, such as people with comprehension handicaps. Text simplification is also important in the context of Second Language Acquisition, in order to provide language learners with material that corresponds to their mastery of the target language.

Our research is concerned with the development of a Spanish text simplification system. To the best of our knowledge this is the first study in the area for the Spanish language. In this paper we report our research so far with special emphasis on the types of operations that make up text simplification. We present our preliminar results on the analysis of a corpus of simplified news which will constitute the basis for our simplification algorithm.

The rest of this paper is organized in the following manner: In Section 2 we report past work on text simplification and in Section 3 we describe the corpus of Spanish news and their simplifications we are using in the present work. In Section 4 we introduce an alignment algorithm to automatically align the corpus at sentence level and in Section 5 we present the result of a corpus study showing the types of operation involved in simplification and suggesting ways toward automa-

ISSN 1135-5948

tion. Finally, we close the paper in Section 6 with conclusions and future work.

2 Text Simplification

The simplification of written documents by humans has the objective of making texts more accessible to people with a handicap that makes linguistic comprehension more difficult for them. Manual simplification of written documents is, however, very expensive. If one considers people who can not read documents with heavy information load or documents from authorities or governmental sources the percent of need for simplification is estimated at around 25% of the population. It is therefore of great importance to develop methods and tools to tackle this problem. Automatic text simplification, the task of transforming a given text into an "equivalent" which is less complex in vocabulary and form, aims at reducing the efforts and costs associated with human simplification. In addition to transforming texts into their simplification for human consumsion, text simplification has other advantages since simpler texts can be processed more efficiently by different natural language processing processors such as parsers and used in applications such as machine translation, information extraction, question answering, and text summarization.

Early attempts to text simplification were based on rule-based methods where rules were designed following linguistic intuitions (Chandrasekar, Doran, and Srinivas, 1996). Steps in the process included linguistic text analysis (including parsing) and pattern matching and transformation steps. Other computational models of text simplification included processes of analysis, transformation, and phrase re-generation (Siddharthan, 2002) also using rule-based techniques. The PSET project (Carroll et al., 1998) proposes a news simplification system for aphasic readers and particular attention is paid to linguistic phenomena such as passive constructions and coreference which are difficult to deal with by people with disabilities. The PorSimples project (Aluísio et al., 2008) has looked into simplification of the Portuguese language. The methodology consisted on the creation of a corpus of simplification at two different levels and on the use of the corpus to train a decision procedure for simplification based on linguistic features. Simplification decisions about whether to simplify a text or sentence have been studied following rule-based paradigms (Chandrasekar, Doran, and Srinivas, 1996) or trainable systems (Petersen and Ostendorf, 2007) where a corpus of texts and their simplifications becomes necessary. (Zhu, Bernhard, and Gurevych, 2010) present an approach to text simplification which relies purely on machine learning techniques. Some resources are available for the English language such as parallel corpora created for or studied in various projects (Barzilay and Elhadad, 2003; Feng, Elhadad, and Huenerfauth, 2009; Petersen and Ostendorf, 2007; Quirk, Brockett, and Dolan, 2004); however there is no parallel Spanish corpus available for research into text simplification. The algorithm presented here will be used to create such a resource.

3 Dataset

We are working with a corpus of 200 news in Spanish covering the following topics: National News, Society, International News and Culture. Each of the texts is being adapted for people with learning disabilities following the Easy-to-Read methodology (Petz and Tronbacke, 2008). Simplification was done by trained experts. We had, however, no direct insight in the editing process in the form of editing histories or keystroke recordings. Original and adapted examples of texts in Spanish can be seen in Figure 1. The texts are being processed using parts-of-speech tagging, named entity recognition and parsing in order to create an automatically annotated corpus. The bi-texts are first aligned using the tool described in the next section and then post-edited with the help of a bi-text editor provided in the GATE framework (Cunningham et al., 2002). Figure 2 shows the texts in the alignment editor. This tool was, however, insufficient for our purposes since it does not provide mechanisms for uploading the alignments produced outside the GATE framework and for producing stand-alone versions of the bi-texts; we have therefore extended the functionalities of the tool for the purpose of corpus creation.

The size if the corpus is not big enough to make pure machine learning techniques a promising option. We are, however, confident, that it allows a hybrid approach, in which some of the most frequent simplification rules can be learned automatically and these can be manually polished and complemented by hand-crafted rules.

4 Alignment of source and simplified texts

In order to be able to study text simplification and to develop a text-simplification system the availability of parallel corpora is of crucial importance. Since there are no aligned corpora for Spanish and the corpus described in section 3 is not aligned, we had to develop an alignment tool.

The algorithm we use is based on two intuitions about simplified texts: As repeatedly observed, sentences in simplified texts use similar words than the original sentences that they stem from. The second observation is very specific to our data: the order in which information is presented in simplified texts roughly corresponds to the order of the information in the source text. So sentences which are close to each other in simplified texts correspond to original sentences which are also close to each other in the source text. In many cases two adjacent simplified sentences even correspond to one single sentence in the source text, which has been split in the simplification process. This leads us to apply a simple Hidden Markov Model which allows for a sequential classification.

We model sentence alignment as a chain of sentence positions in the source document and seek the most probable alignment sequence, given a simplified text and a set of source sentences. We can calculate the optimal alignment sequence $\widehat{align}$ as in (1):

$$\begin{aligned}\widehat{align} = arg\ max \textstyle\prod_{i=1}^{n} P(align_{i,j}) \times \\ P(source\_sent_j | align_{i,j})\end{aligned} \qquad (1)$$

Here $align_{i,j}$ is an alignment between a source sentence j and a (simplified) target sentence i, and $\widehat{align}$ is the most probable sequence of positions in the source text, such that every target sentence is associated to a position (i.e. a sentence identifier) in the source text. The given equation leaves us with two measures which determine the probability of each individual alignment: a measure of sentence similarity ($P(source\_sent_j|align_{i,j}$, the probability of alignment between two sentences) and a measure of consistency ($P(align_{i,j})$), under the assumption that a consistent simplified text presents the information in aproximately the same order as it is presented in the source text. In order to determine $\widehat{align}$, we apply the $Viterbi$ algorithm (Viterbi, 1967).

We calculate the probability of alignment proper as the probability of each word of a sentence in the simplified sentences as stemming from any of the sentences in the source text, calculated as the maximum likelihood estimate (MLE) based on indivual word frequencies. Since the MLE may equal zero in cases where the the is no word overlap, we have to recalculate the alignment probability, taking into accout a distortion probability (where distortion is a cover term for the insertion of new words or the substitution of a word by a different word).

For the consistency measure we made the Markov assumption that each alignment $align_{i,j}$ only depends on the proceeding alignment $align_{i-1,j'}$. We assume that this is the probability of a distance k between the corresponding source sentences of $targe\_sent_{i-1}$ and $target\_sent_i$ for each possible jump distance k, such that P(k) = $P(align_{i,j}|align_{i-1,j-k})$. Since long jumps are relatively rare, we used a normalized even probability distribution for all jump lengths greater than 2 and smaller than -1.

Since we have no training data, we have to initially set these probabilities by hand. The probability $P(source\_sent_j|align_{i,j})$ (the measure for lexical similarity) for each source sentence, however, is able to remedy bad values for the probabilities of the distance between words in the source text, which are only roughly estimated.

In order to derive more reliable values for each jump distance $P(k)$, we apply the Viterbi classifier iteratively: First we initiate the values for $P(k)$ by hand. Then we run the classifier and determine the values for $P(k)$ on its output. Then we run the classifier again, with the new model and so on. In practice stable values for $P(k)$ emerge after as little as two iterations.

As we have shown in (Bott and Saggión, 2011), we achieve very good results with this simple technique. We can compare our results to a baseline system, where the optimal sentence alignment is calculated on an TF*IDF score, either calculated for the frequency of words in the whole corpus (baseline 1) or the words in each source document (baseline 2). This baseline is inspired by,

Original Text	Adapted Text
Un Plan Global desde tu hogar El Programa GAP (Global Action Plan) es una iniciativa que se desarrolla en distintos países y que pretende disminuir las emisiones de CO2, principales causantes del cambio climático y avanzar hacia hábitos más sostenibles en aspectos como el consumo de agua y energía, la movilidad o la gestión de los residuos domésticos. San Sebastián de los Reyes se ha adherido a este Programa. Toda la información disponible para actuar desde el hogar en la construcción de un mundo más sostenible se puede encontrar en ssreyes.org o programagap.es.	Un Plan Global desde tu hogar San Sebastián de los Reyes se ha unido al Plan de Acción Global (GAP). El Plan es una iniciativa para luchar contra el cambio climático desde tu casa. Los objetivos del Plan son: Disminuir nuestros gastos domésticos de agua y energía. Reducir los efectos dañinos que producimos en el planeta con nuestros residuos. Mejorar la calidad de vida de nuestra ciudad. Tienes más información en ssreyes.org y en programagap.es. Apúntate al programa GAP y descárgate los manuales con las propuestas para conservar el planeta.

Figure 1: Original Full Document and Easy-to-Read Version

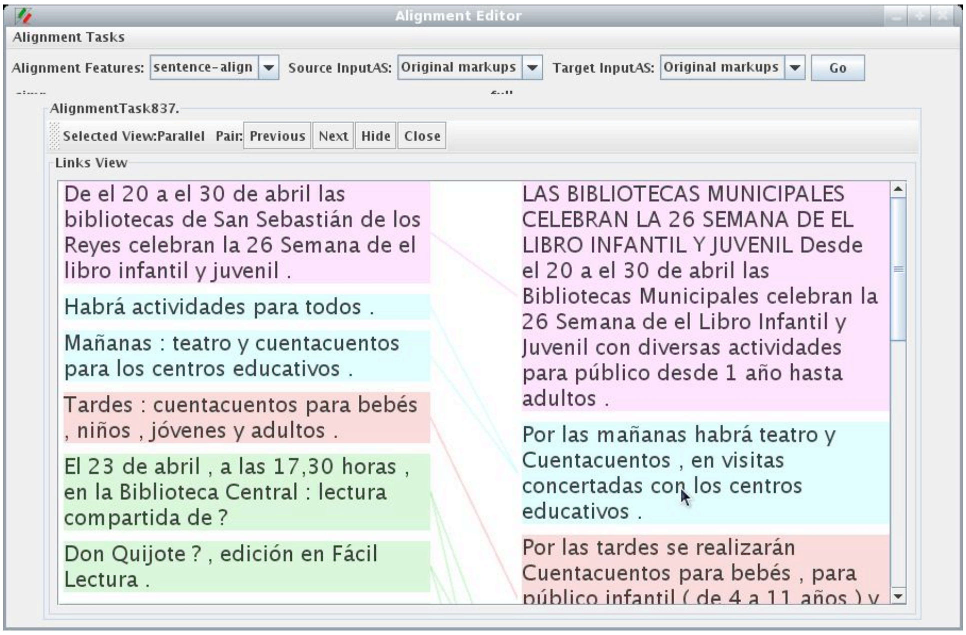

Figure 2: The Alignment Editor with Text and Adaptation

but not directly comparable to (Nelken and Shieber, 2006). One of the reasons why we cannot compare our results directly is that Nelken and Shieber use supervised learning in order to optimize the transformation of TF*IDF scores into probabilities, while we had no training data available. We can, however, compare our probability based sentence similarity measure, given in the second line of table 1 to TF*IDF scores. Also in other respects our alignment method is not directly comparable to any of the alignment methods we are aware of. Nelken and Shieber aligned a totally different type of bi-text (two different versions of the *Encyclopedia Britannica*, a dataset already used for the same task by

algorithm	precision	recall
HMM	82.4%	80.9%
Sentence similarity	81.13%	79.63%
TF*IDF (baseline 2)	75.47%	74.07%
TF*IDF (baseline 1)	62.2%	61.1%

Table 1: Results of the noisy channel based alignment module

(Barzilay and Elhadad, 2003)), and in this cases they could not make the assumption that the order of information presented is similar in the two texts to be aligned.

Table 1 summarizes our results. The scores our alignment module acieves lie clarly above the two baselines. The line *sentence similarity* represents our system when we only choose the best alignment for each target sentence with on our probability based similartity measure alone, diregarding the probability of the alignment sequence. It is interesting to note that our sentence similarity measure performs better than TF*IDF. A second interesting observation is that the sequential classification carried out by the Viterbi decoder, only pushes the performance slightly. The measure of sentence similarity seems to be the decisive factor. This was already found by Nelken and Shieber, when they performed micro-alignment on the paragraph level using dynamic programming.

5 Editing operations

When a human editor simplifies a text she performs a series of editing operations which can range from simple lexical substitutions to the complete rewriting of the original sentence. In principle, these editors follow a series of guidlines for the simplification process, such as the Easy-to-Read methodology. In practice, however, we could observe a high degree of variation in the way simplified texts in our corpus have been re-written. We are also interested in developing a reliable annotation-sceme for such editing operations, since we have to either learn such operations from the corpus automatically, create handcreafted rules for them or apply a combination of both. For all three cases we have to be able to classify the relevant transformations.

5.1 Types of editing operations

In our corpus we could distinguish eight different kinds of editing operations:

- *change*
- *delete*
- *insert*
- *split*
- *proximization*
- *reorder*
- *select*
- *join*

This taxonomy is meant to reflect the editing steps which are necessary in order to derive a simplified text, rather than describe the exact linguistic level at which they apply: *insert* operations, for example, can insert a whole phrase or only a single word. The same is true for *delete* and *change* operations.

The most frequent editing operation is *change*, which is a cover term for several types of replacement operations. The most basic subtype of these operations is a simple change of a lexical element. Sometimes infrequent or otherwise difficult words are replaced by simpler, shorter or more frequent words. The word *reconocimiento*, for example, may be substituted by the word *exámen* in a medical context. Other *change* operations include changing pronouns into full nouns, changing the voice of a verb or even rearanging whole syntactic constructions. The example pair(1)/(2), shows a lexical simplification, which replaces “poner en marcha” with the much shorter “activar”. In addition the title of the project (“Dinamización en parques para familias”) is simplified.

(1) El Ayuntamiento [...] ha puesto en marcha la 'Dinamización en parques para familias', con el objetivo [...]

(2) El Ayuntamiento [...] ha activado un Plan para que las Familias [...]

Delete operations act on single words (adjectives, adverbs), phrases or whole clauses. In (3)/(4) the adjective “diversas” is deleted, since it has no high informaional impact. *Insert* operations may recover the missing subject of a clause or the main verb, in case a full clause is derived from a nomilization. Somtimes these insertions are very hard to predict and involve a mayor reconstruction of the sentence.

(3) Sanse coopera con diversas comunidades de Bolivia y Guatemala

(4) Sanse coopera con comunidades de Bolivia y Guatemala

The *split* operation is very imortant for text simplification. Often long and complex sentences are split up into a series of smaller sentences. The points where such splits take place are often relative clauses, coordination (as the coordination “y” in (5)/(6)) and participle constructions, but sometimes also deverbal noun phrases may be turned into separate units and are then converted in full clauses, replacing the noun by a corresponding verb. The overall efect of these splits is that the number of sentences in a simplified text becomes larger, but the individual sentences become shorter and, hence, easier to read. Sometimes the *split* is combined with a *delete* operation and a part of the source sentence is omited in the simplified text.

(5) La muestra ofrece al público la oportunidad de acercarse a la fauna, la botánica y la cultura de esta inmensa región selvática americana, [...]

(6) La exposición nos muestra la cultura de esta gran selva americana. También nos muestra sus animales y plantas [...]

We found that there is a special operation, which we dubed *proximization*, which is largely orthogonal to the other editing operations. This operation type serves to make sentences psychologically closer to the reader. When the text is about an event in a certain city and this event is anounced in the local newspaper, sometimes a loctive phrase like "*in our city*" ("*en nuestra ciudad*") may be incerted, or a third person verb form ("*the interested person can*") is turned into second person ("*you can*"). These operations are often hard to predict when text simplification is taken to be a chain of editing operations.

(7) Se elabora el contrato de arrendamiento, [...]

(8) Elaboramos el contrato de alquiler.

Reorder operations change the order in which information is presented to the reader. A very typical case occurs with direct speech. Here the person speaking is commonly named before the quote in simplified texts, while in the original text often the person speaking is expressed after the quoted speech, in a clause separated by a comma. The example pair (9)/(10) shows this kind of operation, in addition to the expansion of a pronoun to a full NP and a lexical change.

(9) "Con ellos ofrecemos una nueva posibilidad para [...] propiciar un envejecimiento activo y saludable así como una mejor calidad de vida", afirma Dolores de Diego, concejala Personas Mayores.

operation	**percentage**
change	39.02 %
delete	24.80 %
insert	12.60 %
split	12.20 %
proximization	6.91 %
reorder	2.85 %
select	0.81 %
join	0.81 %

Table 2: Frequencies of different editing operations

(10) Dolores Diego, concejala de Personas Mayores, afirma: "Los aparatos propician un envejecimiento saludable y mejoran la calidad de vida de las personas mayores."

The last two edit operations are less common. *Select* operations may pick an NP out of a source sentence and use this NP as a headline and the relatively rare *join* operation combines two separate pieces of information into a single sentence.

5.2 Frequencies of edit operations

We annotated a part of our parallel corpus for these editing operations and counted the frequencies of these different operations. Table 2 shows the observed frequencies. We observed an average of slightly more than 3 edit operations per source sentence.

Change operations are by far the most frequent ones, followed by *delete* and *insert*. *Split* operations are not as frequent, but their effect is important: They increase the number of sentences by 13.6%. More interesting than the count of basic operations is a closer look at the detailed sub-types, which is given in table 3 and evaluate how easy these operations can be carried out automatically. This table lists the most frequent edit sub-types, an those which we consider an interesting challenge for an automatic text simplification tool.

Some of these editing operations can be modelled as well-defined computational tasks. The globally most frequent operation is lexical substitution. This kind of operation is relatively easy to carry out computationally, provided that we can find a good lexical equivalent for the target words. Also

operation	percentage
change: lexical	17.48 %
insert: unrestricted	3.66 %
change: syntax	2.85 %
change: voice	2.44 %
split: coordination	2.44 %
split: relative clause	2.03 %
insert: missing main verb	1.63 %
delete: adverbial	1.22 %
split: participle construction	1.22 %
change: pronoun → full noun	0.81 %
change: full clause → NP	0.81 %
insert: missing inflected verb	0.41 %
change: numbers	0.41 %
split: subordinate clause	0.41 %

Table 3: Frequencies of specific editing operations

the various kinds of split operations can be automated if we model them as operations on syntactic trees. The problem here is that often the two resulting parts of the split operation share the same subjet in the target sentence and for one of the resulting simplified sentence must be recovered. A coordination stucture of the type "X is Y and Z", for example, will result in "X is Y" and "X is Z" and a relative clause structure "Subj Verb X that Y" results in "Subj Verb X" and "X Y". It is also probable that a system can reliably learn from the training data at which points such split operations must be carried out. Other syntactic *change* operations can be treated in a similar fashion. Sometimes a noun phrase (e.g. "el riego con agua reciclada") is converted in a verb phrase ("se riega con agua reciclada"), in other ocasions (voice changes) impersonal constructions are transformed into second person plural ("se elabora" → "elaboramos"), which is also an instance of *proximization*. As with all proximization operations, the problem here is that the wider, even nonlinguistic, context has to be taken into account.

There is also a group of operations which treats specific word forms and converts them into clearer ones. Abreviations are turned into their full form equivalents and orthografic versions of numbers are replaced by digits, which can also be combinded with a rounding operation ("4.377" → "más de 4000").

A very interesting group of operations involves anaphoric resolution. In some cases pronouns are replaced with the full nominal form of its antecedent. While this is a relatively simple substitution, anaphoric resolution can be done computationally only with a limited degree of reliability.

On the other end of the spectrum, there are operations which are downright impossible to automate. We found 3.66% of the annotated operations in our corpus involved the unresticted insertion of material from the wider context, some of which even involve a considerable amount of inference on real world facts: for example, we found one instance of the adjective "*sostenible*" (*substainable*) in a context where recycled water was used for watering a park and the adjective itself does not apear in the source text. Many of the *proximization* operations fall under this category. Although proximization improves the readability of a text by estabishing closer links to the reader, it is very doubtfull that a computational system can ever perform such text transformations.

6 Conclusion

In this paper we have investigated in how far we can hope to automatize text simplifications with the use of a text-simplification tool, which is under development. We have automatically aligned a corpus of news texts with their manually simplified counterparts. This aligned corpus allowed us to gain important insights in the way human editors create simplified texts. We could identify a series of editing operations which, applied to a source text under the right circumstances, produce a simplified target text.

This study is only the first step towards the development of a text-simplification tool. We plan to combine machine learning techniques with hand-crafted rules and lexical resources for this purpose. We consider that the simplification operation of lexical change can be best attacked with a lexicon resource for "difficult word" which lists simpler expressions for given word and phrases. Some of the syntactic operations, such as splitting sentences, can probably be learned reliably with machine learning techniques, but the limited size of the corpus we have at our disposal makes us also think that such machine learning techniques will most probably have to be complemented by manually written simplification rules, which operate on syntactic trees.

Finally we will devote effort to the evaluation of the simplification system. For the system development automatic evaluation metrics will be needed. Several such metrics, such as Bleu, Rouge or Nist, are used in Machine Translation and Text Summarization and may be used for our purpose, although they will have to be adapted to the text simplification paradigm. Other aspects, such as difficulty of the used vocabulary or the sentence complexity, will require specific new metrics. We also plan to carry out a global human evaluation when the system development has reached maturity, which will be carried out with participants from one of the target user groups.

Acknowledgments

The research described in this paper arises from a Spanish research project called Simplext: An automatic system for text simplification (`http://www.simplext.es`). Simplext is led by Technosite and partially funded by the Ministry of Industry, Tourism and Trade of the Government of Spain, by means of the National Plan of Scientific Research, Development and Technological Innovation (I+D+i), within strategic Action of Telecommunications and Information Society (Avanza Competitiveness, with file number TSI-020302-2010-84). We are grateful to Programa Ramón y Cajal, Ministerio de Ciencia e Innovación, Spain.

References

Aluísio, Sandra M., Lucia Specia, Thiago Alexandre Salgueiro Pardo, Erick Galani Maziero, and Renata Pontin de Mattos Fortes. 2008. Towards brazilian portuguese automatic text simplification systems. In *ACM Symposium on Document Engineering*, pages 240–248.

Barzilay, Regina and Noemi Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *In Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 25–32.

Bott, Stefan and Horacio Saggión. 2011. An unsupervised alignment algorithm for text simplification corpus construction. In *Workshop on Monolingual Text-To-Text Generation, co-located with ACL 2011*, Porland, Oregon.

Carroll, John, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. In *In Proc. of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10.

Chandrasekar, Raman, Christine Doran, and Bangalore Srinivas. 1996. Motivations and methods for text simplification. In *COLING*, pages 1041–1044.

Cunningham, H., D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics.*

Feng, Lijun, Noemie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *EACL*, pages 229–237.

Nelken, Rani and Stuart M. Shieber. 2006. Towards robust context-sensitive sentence alignment for monolingual corpora. In *In 11th Conference of the European Chapter of the Association for Computational Linguistics.*

Petersen, Sarah E. and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *In Proc. of Workshop on Speech and Language Technology for Education.*

Petz, Andrea and Bror Tronbacke. 2008. People with specific learning difficulties: Easy to read and hci. In *ICCHP*, pages 690–692.

Quirk, Chris, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In *In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 142–149.

Siddharthan, Advaith. 2002. An architecture for a text simplification system. In *Proceedings of LREC 2002*, pages 64–71.

Viterbi, A. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13:260–269.

Zhu, Zhemin, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of The 23rd International Conference on Computational Linguistics*, COLING '10, pages 1353–1361, Beijing, China, Aug.