Text Simplification in Simplext: Making Texts more Accessible

Simplificación de textos en Simplext: haciendo textos más accesibles

Horacio Saggion
Universitat Pompeu Fabra, Grupo TALN, C/Tanger 122 - Barcelona - Spain
horacio.saggion@upf.edu

Elena Gómez-Martínez
Technosite-ONCE Foundation, R&D Department, C/ Albasanz, 16 28037 Madrid, Spain
megomez@technosite.es

Alberto Anula
Universidad Autónoma de Madrid, Grupo DILES, c/ Fco. Tomás y Valiente 2, 28049 Madrid
alberto.anula@uam.es

Lorena Bourg
Ariadna Servicios Informáticos, C/ Sanchez Barcaiztegui 33, 28007 Madrid
lbourg@asi-soft.com

Esteban Etayo
Technosite-ONCE Foundation, R&D Department, C/ Albasanz, 16 28037 Madrid, Spain
eetayo@technosite.es

Resumen: La simplificación de textos tiene como objetivo la transformación de un texto en uno equivalente que es de fácil lectura para un colectivo de personas determinado. El proyecto Simplext propone el desarrollo de un sistema ubicuo para la simplificación automática de textos. Simplext se basa en principios de Fácil Lectura y en la aplicación de técnicas robustas de procesamiento de lenguaje natural.
Palabras clave: Simplificación de textos; Fácil lectura; Texto informativo; Alineación de oraciones; Generación de textos

Abstract: Text simplification is the process of transforming a text into an equivalent which is more understandable for a target user group. The Simplext project aims at producing an ubiquitous text simplification system for Spanish. The automatic simplification system is being developed following the easy-to-read principles and applying robust Natural Language Processing techiques.
Keywords: Text Simplification; Easy-to-Read; Informative Text; Sentence Alignment; Text-to-Text Generation

1 Introduction

Text simplification is the process of transforming a text into an equivalent which is more understandable. This simplification is beneficial for many groups of readers, such as language learners, elderly persons and people with other special reading and comprehension necessities. Simplified texts are characterized by a simple and direct style and a smaller vocabulary as well as by simpler sentences. This means that in the simplification process complex sentences are often split into several smaller sentences.

Text simplification has many important applications such as:

- Facilitating access to information to

people with a low literacy level;

- Making news articles accessible to people with an intellectual disability or to people who need assisted reading;
- Making it easy to access content by language learners and help prepare adapted material for second language acquisition;
- Transforming text of high technical complexity (e.g. patents, technical manuals) for people unfamiliar with the intricancies of these types of texts.

Simplext ("Un sistema automático de simplificación de textos") is a 28-months research project funded by the Spanish Ministry of Industry, Tourism, and Trade under the Subprograma Avanza Competitividad to develop an ubiquitous text simplification solution for Spanish. In addition to its scientific interest and the fact of being the first application of text simplification to Spanish, Simplext has also an important social function since it aims at developing a solution to make text accessible to people with a cognitive impairment. The Simplext consortium has different expertises from different partners:

- Technological/Software companies in charge of software development and deployment of the ubiquitous solution: Abada, Ariadna, S.A.D.E. Consultoría Técnica, Technosite, Tilo Systems;
- Legal consultancy: CE Consulting Empresarial;
- Target users: Fundación PRODIS;
- Text information providers: Servimedia;
- Experts in the production of easy-to-read material: Universidad Autónoma de Madrid; and
- Natural language processing know-how: Universitat Pompeu Fabra.

This paper will give an overview of the work involved in the Simplext project. The rest of the paper is organized in the following way: in Section 2 we introduce text simplification initiatives and in Section 3 we give an overview of the manual simplification process adopted in Simplext. Then, in Section 4 we describe the corpus of simplifications we are creating as part of the project and in Section 5 we describe the role of natural language processing in Simplext. Section 6 outlines the adopted system architecture and Section 7 briefily explains the adopted evaluation. The paper ends with concluions and outlook in Section 8.

2 Background

From a foundational and methodological point of view there are various initiatives that promote accessible texts including the "Plain English" initiative for English in the United States and in the United Kingdom and the "Rational French", a controlled version of French used in the air industry to facilitate and standardize the technical jargon of that domain. In Europe there are associations dedicated to the adaptation of text materials (books, leaflets, laws, official documents, etc.) for people with disabilities or low literacy levels, examples of these associations are the Easy-to-Read Network in the Scandinavian countries, the Asociación Lectura Fácil in Spain, and the Centrum för Lättläst in Sweden.
The "easy-to-read" newspaper "8 Sidor" (http://8sidor.lattlast.se) is a publication by the Centrum för Lättläst to allow people access to "easy news". For Spanish, the Noticias Fácil (http://www.noticiasfacil.es) is another example of easy-to-read news for people with a disability. At the European level, the Inclusion Europe portal (http://www.inclusion-europe.org) provides good examples of how full text simplifications and simplified summaries in various European languages can provide improved access to relevant information. Finally, the Simple English Wikipedia (http://simple.wikipedia.org) provides encyclopaedic content which is more accessible than plain Wikipedia articles because of the use of simple language and simple grammatical structures.
Manual simplification of written documents is very expensive limiting the amount of simplified material available today. There is however an increased need for methods and techniques to make texts more accessible. For example, people with learning disabilities who need simplified text constitute a 5% of the population, however and according to data from the Easy-to-Read Network, if we

Original Text	Adapted Text
La Universidad Popular celebra este curso su 30º aniversario La Universidad Popular 'José Hierro´ de San Sebastián de los Reyes, la primera que se fundó en España, en el año 1980, celebra este curso su 30 Aniversario. El alcalde ha destacado "la importancia del trabajo que ha desarrollado y continúa llevando adelante la Universidad Popular, así como su dinamismo y cercanía a las necesidades de los vecinos".	La Universidad Popular celebra su 30º aniversario La Universidad Popular José Hierro de San Sebastián de los Reyes se fundó en 1980. Fue la primera Universidad Popular de España. En este año 2010 se celebra su 30º aniversario. El alcalde ha dicho que "la Universidad Popular desarrolla un trabajo muy importante". También ha dicho que "está muy cerca de las necesidades de los vecinos".

Figure 1: Original News in Spanish and its Easy-to-Read Adaptation. This text is an adaptation carried out by DILES for Revista "La Plaza".

consider people who can not read documents with heavy information load or documents from authorities or governmental sources the percent of need for simplification jumps to 25% of the population. The need for simplified texts is becoming more important as the incidence of disability increase as the population ages.

3 Simplification Methodology

Manual simplification methodology in Simplext follows work by Anula (2007; 2008) proven to contribute to the reduction of complexity in written language. Two types of simplifications are considered here:

- vocabulary simplification; and
- simplification of syntactic structures

3.1 Simplification of Vocabulary

The words that form a message belong to several types and fulfill different functions in the message. The so-called lexical or content words provide the semantic-denotative weight of the message, whereas functional words articulate the accurate grammatical relationships needed to appropriately assemble the lexical content. Together with these two types of words, we have referential and deictic ones. Each of these types of words has its specific features and, in the context of reading facilitation, its treatment differs depending on their nature. *Frequency of use* is an appropriate measure regarding lexical words, so a possible procedure for text simplification consists of replacing little-used (low frequency) words by others whose linguistic use is widespread. However, the control of low frequency does not assure, by itself, the improvement of comprehensibility, since for instance most lexical words are polysemic, have figurative uses and, furthermore, the meaning they denote in a sentence is strongly related to the syntactic context in which they are used. This complex reality will be taken into account in the basic linguistic research undertaken in Simplext.

Another instrument for the control of lexical words has to do with *lexical density*, see (Anula, 2008). This variable considers that the processes of linguistic decoding that take place during verbal comprehension real-time are facilitated if a reduced group of different words is used in texts, since this reduces the cognitive effort of pattern recovering and words meaning. Therefore, restricting the use of synonyms, in spite of impoverishing the text's verbal style, is a procedure that must be openly taken into account by automatic text simplification systems.

3.2 Simplification of Syntactic Structures

Regarding phrases and sentences, there are many elements than can and must be simplified to obtain simpler and more comprehensible discourses for people with reading difficulties. Among many others, we could mention the length of discursive segments (measured as the number of words per segment), the simple (a phrase or sentence per segment) or complex (more than one) nature of them, the use of periphrastic predicates, the abundance of subordinate structures, the use of impersonal or passive sentences, the insertion of subsections or circumlocutions, the alteration (without an informative purpose) of the constituents order, the profusion of non-argumental complements, the presence

of secondary predicates, etc. Here we will only present two of the different phenomena that we manipulate in order to reduce the structural complexity of discourse: *inlay* and *recursion*. One complex sentence pattern that needs to be dealt with is the following:

$$[\ X\ [\ X_1\ [\ X_2\ ]\ ]\ ]$$

which is a discursive segment such as a sentence or phrase or group of sentences or phrases syntactically locked bewteen two orthographic pauses that give the segment syntactic independence. This pattern entails inlaying the set of contracted units and generates three simple and independent discourse segments.

It is not infrequent, in written discourse, that in addition to the above scheme, we can have juxtaposed and/or coordinated relationships that make the segment more complex. In these cases, we use the term recursion to name the inlay of coordinated structures within subordinate structures or vice versa.

From the linguistic point of view, the main aim of the project is to identify those phenomena and categories which hinder the cognitive processing that takes place while and after reading, and which could undergo a linguistic simplification that allows us to create a formal variable accessible to the people with difficulties in reading comprehension, thus facilitating the cognitive processes to the reader.

4 Simplext Corpus

Simplext aims at the development of a corpus of original texts and their adaptations to carry out research and development into text simplification in Spanish. One example of Spanish news and its simplification is shown in Figure 1. Note that the adapted (simpler) text contains only simple sentences composed of subject-verb-object structures contrary to the more complext structures used in the original text where appositions, conjunctions, relatives phrases, etc. are used. In the show example, the sentence (O1) in the non-simplified text:

> (O1) La Universidad Popular 'José Hierro´ de San Sebastián de los Reyes, la primera que se fundó en España, en el año 1980, celebra este curso su 30 Aniversario.

has been simplified into three sentences (S1), (S2), and (S3) in the simplified text:

> (S1) La Universidad Popular José Hierro de San Sebastián de los Reyes se fundó en 1980.
>
> (S2) Fue la primera Universidad Popular de España.
>
> (S3) En este año 2010 se celebra su 30º aniversario.

where each of them conveys a single fact: (S1) the fundation date of "Universidad Popular José Hierro"; (S2) the fact that this university was the first "university of the people" in Spain; and (S3) the fact that it was 30 years old in 2010. All three facts were conveyed in a single but complex sentence (O1) in the non-simplified text. Manual simplification is a very elaborated process that requires lot of experience and human resources. The Simplext corpus will consist of a set of around 200 short news articles provided by one of the partners in the consortium (Servimedia) and adaptations of these articles to our target user groups. Adaptations are being produced by the DILES Research Group from Universidad Autónoma de Madrid with the methodology outlined in Section 3.

5 Automatic Text Simplification

Text simplification is not unknown to natural language processing, however, the first automatic simplification attempts were not based on human studies or on corpora to ground hypothesis, but on a set of intuitions about what linguistic phenomena should be the focus of simplification. These approaches generally adopted rule-based methods where the rules were manually designed to match complext sentences and produce simpler ones (Chandrasekar, Doran, and Srinivas, 1996). (Siddharthan, 2002) proposed an architecture of analysis, transformation, and phrase re-generation, the system uses parsing and pattern matching to detect and transform complext constructions. Where the target user of simplification is concerned, the PSET project (Carroll et al., 1998) developed a simplification system for English for aphasic readers and focused on the simplification of linguistic constructions such as passive voice (sentence level) and coreference phenomena (discourse level) which are both difficult to deal with by aphasic people. The PorSimples project (Aluísio et al., 2008) produced

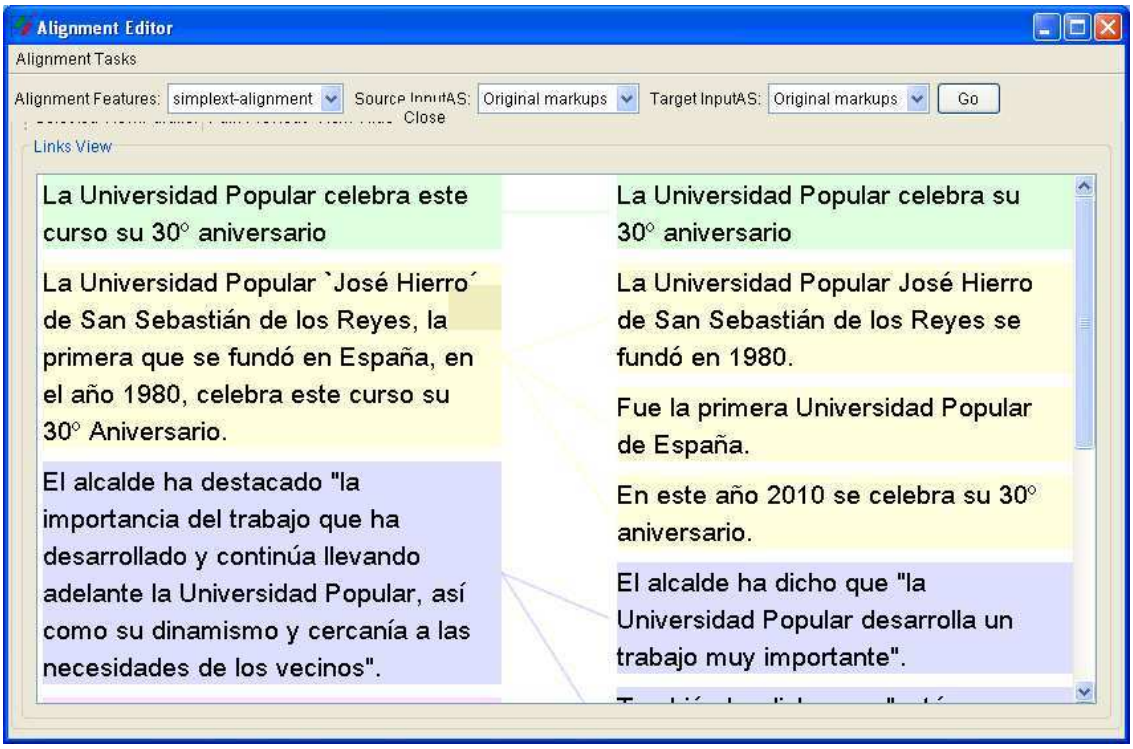

Figure 2: The Alignment Editor with Text and Adaptation

simplification technology for Portuguese aiming at adapting texts for people with a low literacy level. The project created a corpus of simplified texts at two different levels and used it to developed a trainable decision module to decide whether a sentence needs to be simplified. Simplification language resources are available for English: (Nelken and Shieber, 2006) have produced aligments between the simple English Wikipedia articles and the English Wikipedia counterparts while (Petersen and Ostendorf, 2007) studies simplification for second language acquisition and have enriched an available corpus with manual alignments indicating sentence split points. There are however no resources for the study of simplification in Spanish.

5.1 Natural Language Processing in Simplext

There are various problems we are addressing in the project from the NLP viewpoint:

- Corpus alignment: the Simplext corpus (Section 4) will be aligned at the sentence level in two steps: first, an automatic aligment algorithm will be applied to identify relations between original and simplified sentences, and then a bi-text editor will be used to correct the alignments proposed automatically to make sure that in the final version of the corpus these are correct. The alignment algorithm has already been proposed and results are reported in (Bott and Saggion, 2011b). As for the bi-text editor, we are relying on a plug-in from the GATE system (Maynard et al., 2002) which we use to post-edit the automatic alignments, see Figure 2. The tool has been extended with functionalities for document upload and document saving for Simplext.

- Development of text analysis pipelines to carry out linguistic processing of the corpus and the new documents to be simplied in application time. We are basing this work on various free-tools such as Freeling (Atserias et al., 2006) and GATE to implement syntactic and semantic analysis.

- Semi-automatic analysis of the corpus in order to identify potential simplification operations which could be implemented (e.g. systematic operations instead of

idiosyncratic ones). We have already carried out an analysis of a set of original and simplified texts and have detected a set of simplification operations including: change, delete, insert, split, etc. Results of this analysis are reported in (Bott and Saggion, 2011a).

- Development of a decision support module to take simplification decisions and implementation of metrics for text readability in the Spanish language.
- Development of a text-to-text generation component to produce simple sentences from extracted information chunks.

We have already developed varios linguistic processors, adapted a document alignment editor to the requirements of corpus construction in Simplext, and developed a robust sentence aligner. Our focus now is on the specification and implementation of the text simplification process.

6 Simplext Architecture

The Simplext architecture is an event-driven and service-oriented architecture that further develops the idea of delivering the simplification service. The rationale of such approach is that a person with an adapted device (e.g., mobile phone) should be able to receive easy-to-read digital contents, such as RSS or digital press. The Simplext architecture is inspired and reuses work from Limbourg et al. (Limbourg et al., 2004) and Kobsa's ideas about user modelling (Kobsa, 1990) and privacy (Kobsa, Chellappa, and Spiekermann, 2006) among others.

As outlined in Figure 3, the involved technologies are summarized in the following:

- The simplification web service is the software component that will allow to publish the simplification engine using SaaS (*Software as a Service*) as software delivery model. This web service will provide digital contents, such as RSS feeds and HTML documents, to translate into easy-to-read text applications deployed in different user devices, for instance, smartphones or personal computers.
- To develop this services layer, which will permit deliver the above mentioned functionalities, an architecture based on SOA and REST (*Representational State Transfer*) is planned. The main advantage of using REST instead of SOAP is that REST is a lightweight protocol, which exploits all the HTTP characteristics and methods, such as GET, POST, PUT and DELETE.
- To implement the functionality of translating news from third-party websites, simplification service must be able to handle RSS and ATOM responses, processing the content, making the translated response and returning the translated RSS/ATOM to the client, who has made the request by interacting API Web service.
- The simplification engine capabilities will be accessed using standard identifiers (URI), where each request may be made independently, atomic, stateless (following the principles laid down by HTTP protocol itself.)
- For the development of mobile applications, the technologies will be used to provide each of the platform manufacturers (SDK's, development tools), among those are Android SDK, iPhone SDK, Windows SDK for Visual Studio Phone and Symbian SDK.
- To develop the portal of easy-to-read news, standards and technologies are used and tested with content management and Drupal.
- For the native desktop application development, JavaSE technologies are used to ensure the portability of the product to different operating systems (MacOS, Windows, Linux).

7 Overview of the Evaluation

Simplext will carry out evaluation at three different levels. In an intrinsic evaluation, we will compare the automatic simplifications produced by Simplext with correct human simplifications. The comparison would be done with available metrics from text summarization or machine translation literature (e.g., string comparison metrics). This evaluation will help us test different system configurations and adjust the parameters of the final solution.

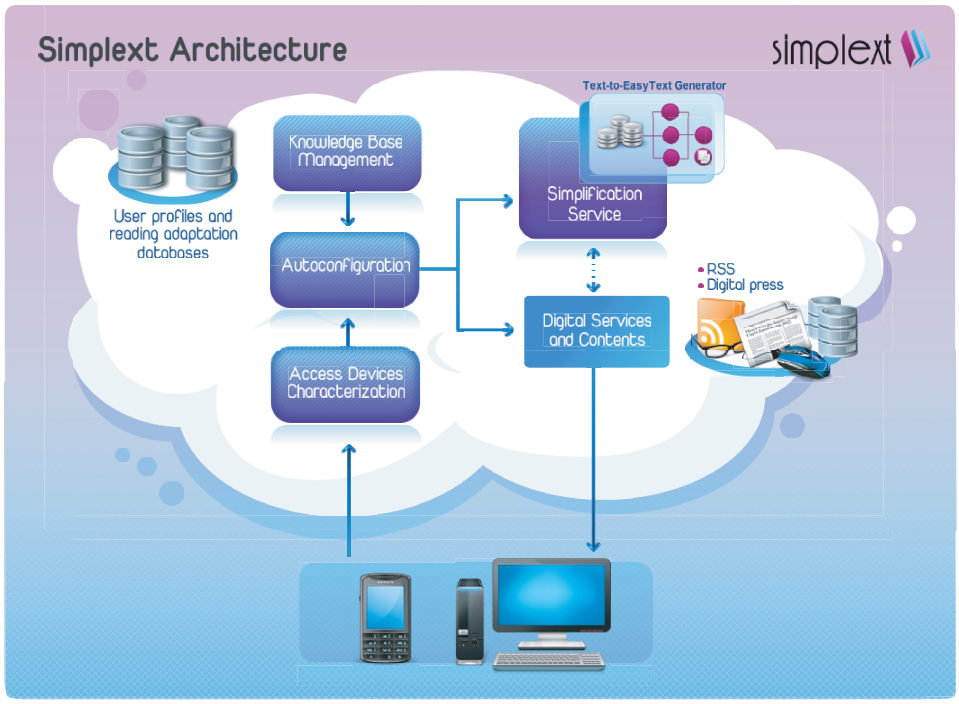

Figure 3: Simplext architecture simplified diagram.

In an extrinsic evaluation, we will evaluate the simplification solution with real users. We are planing reading and comprehension tests where the condition to be tested will be the text used: either original or simplified.

Finally, a technological evaluation will take place where the usability and accessibility of the Simplext solution will be tested. Here we will assess the solution with users accessing the system.

8 *Conclusion and Outlook*

In this paper we have presented an overview of the Simplext project to develop an ubiquitous text simplification solution for Spanish. Although the project is in its first year, we have already made the following contributions:

- First, we have established the principles for manual simplification for Simplext;
- Second, we have collected a set of over 200 informative texts in four different domains to create the Simplext corpus and have started the process of manual simplification;
- Third, we have developed and evaluated a robust sentence aligment algorithm for text simplification and we are applying it to the corpus;
- Fourth, we have carried out a study of text simplification operations that will inform the development of the simplification solution; and
- Finally, we have created a number of text analysis tools to enrich the corpus and use them for the final simplification application.

We are currently working on the development of the simplification solution which, given the reduced size of the corpus, will be a hybrid system combining machine learning with hand-crafted rules.

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). Horacio Saggion is grateful to a fellowship from 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.

Anula, A. 2007. Tipos de textos, complejidad lingüística y facilicitación lectora. In *Actas del Sexto Congreso de Hispanistas de Asia, Man-Ki, Jy-Eun, y Macías (eds.)*, pages 45–61, Seúl, República de Corea.

Anula, A. 2008. Lecturas adaptadas a la enseñanza del español como L2: variables lingüísticas para la determinación del nivel de legibilidad. In *La evaluación en el aprendizaje y la enseñanza del español como LE/L2, Pastor y Roca (eds.),*, pages 162–170, Alicante.

Atserias, Jordi, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluís Padró, and Muntsa Padró. 2006. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library . In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy. ELRA.

Bott, S. and H. Saggion. 2011a. Spanish text simplification: An exploratory study. *Revista de Procesamiento de Lenguaje Natural*.

Bott, S. and H. Saggion. 2011b. An unsupervised alignment algorithm for text simplification corpus construction. In *Proceedings of the ACL 2011 Workshop on Monolingual Text-to-Text Generation*, Portland, Oregon, USA, June. ACL, ACL.

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.

Kobsa, A. 1990. Modeling the user's conceptual knowledge in BGP-MS, a user modeling shell system. *Comput. Intell.*, **6**(3):193–208.

Kobsa, A., R.K. Chellappa, and S. Spiekermann. 2006. Privacy-enhanced personalization. In *CHI '06 extended abstracts on Human factors in computing systems*, CHI '06, pages 1631–1634. ACM.

Limbourg, Q., J. Vanderdonckt, B. Michotte, L. Bouillon, and V. López-Jaquero. 2004. USIXML: A Language Supporting Multi-path Development of User Interfaces. In *Proc. of Conf. Engineering Human Computer Interaction and Interactive Systems*, volume 3425 of *Lecture Notes in Computer Science*, pages 200–220. Springer.

Maynard, D., V. Tablan, H. Cunningham, C. Ursu, H. Saggion, K. Bontcheva, and Y. Wilks. 2002. Architectural Elements of Language Engineering Robustness. *Journal of Natural Language Engineering – Special Issue on Robust Methods in Analysis of Natural Language Data*, 8(2/3):257–274.

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*.

Siddharthan, Advaith. 2002. An architecture for a text simplification system. In *In LREC'02: Proceedings of the Language Engineering Conference*, pages 64–71.