Simplifying Drug Package Leaflets

Isabel Segura-Bedmar, Luis Núnez-Gómez, Paloma Martínez, Maribel Quiroz
Computer Science Department, University Carlos III of Madrid
Avd. Universidad, 30,Leganés, Madrid, Spain
isegura@inf.uc3m.es, lununezg@pa.uc3m.es, pmf@inf.uc3m.es

Abstract

Drug Package Leaflets provide information for patients on how to safely use medicines. European Commission and recent studies stress that further efforts must be made to improve the readability and understandability of package leaflets in order to ensure the proper use of medicines and to increase patient safety. To the best of our knowledge, this is the first work that directly deals with the automatic simplification of drug package leaflets. Our approach to lexical simplification combines the use of domain terminological resources to give a set of synonym candidates for a given target term, and the use of their frequencies in a large collection of documents in order to select the simplest synonym.

1 Introduction

Since 2001, according to a directive of the European Parliament (Directive 2001/83/EC) (EU, 2001), every drug product must be accompanied by a package leaflet before being placed on the market. This document provides informative details about a medicine, including its appearance, actions, side effects and drug interactions, contraindications, special warnings, etc. This directive also required that Drug Package Leaflets (DPL) must be written in order to provide clear and comprehensible information for patients since their misunderstanding could be a potential source of drug related problems, such as medication errors and adverse drug reactions. In 2009, the European Commission published a guideline (EC, 2009) with recommendations and advices in order to issue package leaflets with accessible and understandable information for patients. However, recent studies (Pires et al., 2015; Piñero-López et al., 2016) show that the readability and understandability of these documents have not been improved during the last seven years. Therefore, further efforts must be made to improve the understandability of package leaflets in order to ensure the proper use of medicines and to increase patient safety.

One of the main reasons why the understandability has not been improved is that these documents still contain a considerable number of technical terms describing adverse drug reactions, diseases and other medical concepts. Posology (dosage quantity and prescription), contraindications and adverse drug reactions seem to be the sections most difficult to understand (March et al., 2010). To help solving this problem, we propose an automatic system to simplify drug package leaflets.

Text simplification is a Natural Language Processing (NLP) task that aims to rewrite text into an equivalent with less complexity for readers. There are two main approaches to this task: lexical and syntactic simplification. Lexical simplification basically consists of replacing complex concepts with simpler synonyms, while syntactic simplification aims to reduce the grammatical complexity of a text while preserving its meaning.

Text simplification techniques have been applied to simplify texts from different domains such as crisis management (Temnikova, 2012), health information (Jonnalagadda et al., 2009; Kandula et al., 2010; Jonnalagadda and Gonzalez, 2011), aphasic readers (Devlin, 1999), language learners (Petersen and Ostendorf, 2007). Comprehensive surveys of the text simplification field can be found in (Shardlow, 2014; Siddharthan, 2014).

To the best of our knowledge, this is the first work that directly deals with the automatic simplification of drug package leaflets. In particular, we focus on the lexical simplification of adverse drug reactions that are described in these documents. Moreover, our work is one of the few studies that address the simplification of texts written in Spanish. Our approach for lexical simplification combines the use of terminological resources that provide a set of synonym candidates for a given target term, and the use of their frequencies in a large collection of documents in order to select the most common synonym.

The paper is organized as follows. Section 2 presents related work. Section 3 describes our approach. Experiments, results, and discussion are given in Section 4. Finally, the paper is concluded and future work is proposed in Section 5.

2 Related Work

First works in text simplification started 20 years ago (Chandrasekar et al., 1996). It is based on transforming a text in an equivalent text that is easier to read and probably easier to understand by a target audience.

There is a need to adapt contents for some groups of people because information is not equally accessible to everyone. It is unlikely that professional editors will adapt text for all literacy levels, and NLP techniques could help simplify texts by automating some tasks. In this way, it is possible to help content editors to generate adapted contents. On the other hand, text simplification is essential in several types of texts: News, Government and administrative information, laws and rights, etc. As it was mentioned before, there are two subtasks of text simplification (Saggion et al., 2011): (1) syntactic simplification that divides complex sentences in simplest sentences, (2) lexical simplification whose objective is to substitute complex vocabulary by common vocabulary (looking for synonyms that are simpler than the original word considering the context in the sentence). Moreover, a clarification step could be included to provide definitions and explanations for acronyms, abbreviations and unusual words. These tasks are not completely automatic, they have to be manually reviewed in some cases.

Firstly, we have to distinguish between readability and understandability because these concepts capture different aspects of the complexity of the text. Readability is about the structure of sentences (it concerns syntax and consequently requires syntactic simplification approaches). On the other side, understandability is about the difficulty to interpret a word (Barbieri et al., 2005) and lexical simplification approaches are required.

Concerning syntactic simplification it consists on transforming complex and long sentences into simplest and independent sentences eliminating coordination (of clauses, verbs, etc.), dropping subordination utterances (relative clauses, gerundive and participle utterances), resolving anaphora and transforming passive into active voice. First a parser is used to obtain a dependency tree that represents the syntactic structure of the sentence (noun, prepositional and verbal phrases and how they are related to) (Dorr et al., 2003). Then, rule-based approaches are used in syntactic simplification. Rules can be automatically learned from annotated corpora of text (syntactic trees of sentences where original sentences are related to their simplified sentences) (Zhu et al., 2010), or handcrafted rules (Chandrasekar et al., 1996; Siddharthan, 2002). The rules include split, drop, copying and reordering operations over syntactic trees.

Related to lexical simplification, this task consists on replacing words (taking into account the context) and complex utterances by easier words or phrases. A heuristic used is that complex words have a low frequency. Moreover, lexical resources, as Wordnet (Miller, 1995), are used to extract synonyms as candidates to replace a complex or difficult word. Combining a lexical resource and a probabilistic model is an approach that has been tried (De Belder et al., 2010). Probabilistic models are obtained from lexical simplifications, which have previously done applying E2R guidelines, as in the Simple Wikipedia. McCarthy and Navigli (McCarthy and Navigli, 2007) introduce work to propose candidates to replace a word using contexts. In Semeval 2012, English Lexical Simplification challenge (Specia et al., 2012) with ten participant systems, the evaluation results showed that proposals based on frequency give good results comparing to other sophisticated systems.

Focusing on research devoted to synonym substitution in Spanish texts, lack of semantic resources is a handicap. A recent work is described in (Bott et al., 2012), LexSiS system that uses Spanish OpenThesaurus to build a vector space model according to the distributional hypothesis that establishes that different uses of a word tend to appear in different lexical contexts. A vector is built in a window of nine words around each word-sense in a corpus extracted from the OpenThesarus and compared using the cosine similarity combined with word frequency and word length. This approach can be enhanced including rule-based lexical simplification, see (Drndarevic et al., 2012), where some patterns that avoid incorrect substitutions are defined, for instance, to replace reporting verbs (confirm, suggest, explain, etc.) that leaves correct syntactic structures as well as other editing transformations (numerical expressions or periphrasis). Following the same approach, CASSA method is reported in (Baeza-Yates et al., 2015) where the Spanish corpus used to extract word occurrences is the Google Books Ngram corpus that contains real web frequencies. This work also obtains word senses from OpenThesaurus.

But before simplifying we have to know the level of readability and understandability of a text by using complexity measures. There are simple measures based on frequency of words in texts as well as length of phrases, FOX index (Gunning, 1986), Flesch-Kinaid (Kincaid et al., 1975) measures are used in English. In Spanish texts, several indexes have been proposed to measure the structural complexity of a text (Anula, 2007): the number or verbal predicates in subordinate clauses, and the index of sentence recursion (a measure that counts the number of nested clauses in the text). To measure the lexical complexity two indexes are proposed: an index of low frequency words (the number of content words[1] with low frequency divided by the total number of lexical words) and an index of lexical density (number of distinct content words /total of discourse segments[2]). Finally, other indexes such as the average length of sentences and average length of words (syllables) although they are criticized. These indexes have to be validated by the end users. Knowing the readability level of a document, users have the opportunity to choose the most suitable text, from a collection of documents delivering the same information (Sbattella and Tedesco, 2012).

With respect to Spanish corpora for extraction of frequencies and word contexts, the CREA[3] corpus available online is not a useful resource when domain specific texts are required (for instance, biology or chemical texts). The latest version of June 2008 contains one hundred and sixty million of documents (from journals, books and newspapers covering more than one hundred subjects). In 2018 The Royal Spanish Academy (RAE) will deliver the CORPES XXI, a higher Spanish corpus with four hundred million of forms.

Finally, there are specific works to simplify numerical expressions. Bautista and Saggion (2014) (Bautista and Saggion, 2014) propose a rule-based lexical component that simplifies numerical expressions in Spanish texts. This work makes news articles more accessible to certain readers by rewriting difficult numerical expressions in a simpler way.

3 The EasyLecto system

The EasyLecto system aims to simplify the drug package leaflets, in particular, replacing the terms describing adverse drug reactions with synonyms that are easier to understand for the patients.

Figure 1 illustrates the EasyLecto system architecture. The first module of the EasyLecto system aims to automatically annotate adverse drug reactions in texts. This module uses a dictionary-based approach that combines terminological resources, such as MedDRA, the ATC system (a drug classification system developed by the World Health Organization) or CIMA (a database of medicines approved in Spain), or dictionaries gathered from websites about health and medicines such as MedLinePlus[4], vademecum.es[5] or prospectos.net[6]. The reader can find a detailed description of the NER module in (Segura-Bedmar et al., 2015).

Once adverse drug reactions are automatically identified in texts, a set of synonyms is proposed for each one of them. MedDRA[7] is a medical terminology dictionary about events associated with drugs. It is a multilingual dictionary (11 languages) and its main goal is to provide a classification system for efficient communication of adverse drug reactions data between countries. MedDRA is composed of a five-level hierarchy. The most specific level, ”Lowest Level Terms” (LLTs), contains a total of 72,072 terms that express how information is communicated in practice. The main

[1] A content word is a word with meaning (nouns, verbs, adjectives and adverbs)

[2] sentences or phrases

[3] http://corpus.rae.es/creanet.html

[4] https://www.nlm.nih.gov/medlineplus/spanish/

[5] http://www.vademecum.es

[6] https://www.prospectos.net

[7] http://www.meddra.org/

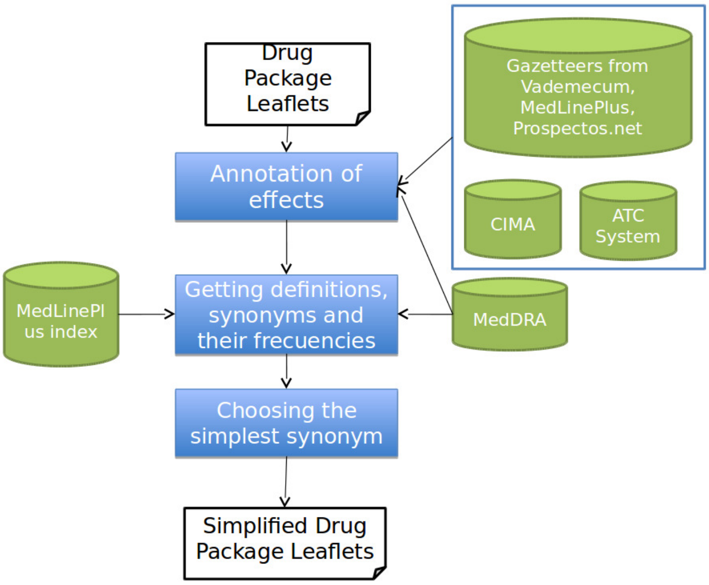

Drug Package Leaflets
Gazetteers from Vademecum, MedLinePlus, Prospectos.net
Annotation of effects
CIMA
ATC System
MedLinePlus index
Getting definitions, synonyms and their frecuencies
MedDRA
Choosing the simplest synonym
Simplified Drug Package Leaflets

Figure 1: The EasyLecto system arquitecture.

advantage of MedDRA is that its structured format allows easily obtaining a list of possible adverse drug reactions and their synonyms. Thus, we decided to use MedDRA as a source of synonyms for adverse drug reactions. Moreover, for a given effect in MedDRA, we used its longest synonym as definition for the effect.

The following step is to select the appropriate synonym, that is, the simplest synonym. The more common a term is in a collection of texts, the more familiar the term is likely to be to the reader (Elhadad, 2006). Thus, our system proposes those synonyms with higher frequency. In order to know how common a word is, we gathered a large collection of texts such as the MedLinePlus articles [8], and indexed it in order to obtain the frequency of each drug effect.

MedLinePlus is an online resource with health information for patients, which contains more than 1,000 articles about diseases and 6,000 articles about medicines. The Spanish version is one of the most comprehensive and trusted Spanish language health websites at the moment. We developed a web crawler to browse and download pages related to drugs and diseases from the MedLinePlus website. Each MedLinePlus article provides exhaustive information about a given medical concept, and also proposes a list of related health topics, which can be considered as synonyms of this concept. Moreover, an article related to a given medical concept can also be used to obtain the definition of this concept by getting its first sentence. Finally, all downloaded articles, the definitions (first sentence of each article) and their related health topics were translated into JSON objects in order to create an index (see Figure 2) using ElasticSearch[9], an open source search engine.

All told, the EasyLecto system proposes a definition and a set of synonyms from MedDRA, as well as a definition and a set of synonyms from MedLinePlus, for each drug effect. Then, the frequency of each synonym is calculated using the index built from MedLinePlus, and finally the synonym with the highest frequency is selected as the simplest synonym.

Due to the horizontal scalability provided by ElasticSearch, it is possible to index large collections of documents, as is the case of the MedlinePlus. The main advantage of ElasticSearch is its capacity to create distributed systems by specifying only the configuration of the hierarchy of nodes. Then, ElasticSearch is self-managed to maintain better fault tolerance and load distribution. Another important advantage of ElasticSearch is that it does not require very high computing power and a high storage capacity to index large collections. In this study, ElasticSearch (version 2.2) was installed on a Ubuntu Server 14.04

[8] https://www.nlm.nih.gov/medlineplus/spanish/

[9] http://elasticsearch.org

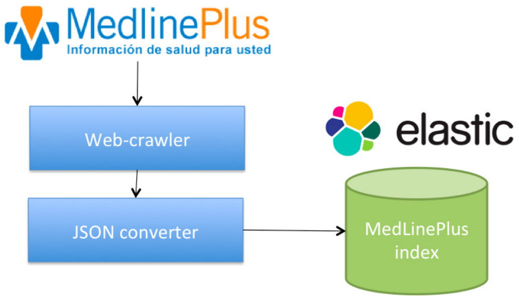


Figure 2: An index was generated from the MedLinePlus articles using ElasticSearch.

with 8GB of RAM and 500GB of disk space.

A demo of the EasyLecto system is available at: http://jacky.uc3m.es/EasyLecto/. This tool allows to load a document highlighting the adverse drug reactions (in blue) (see Figure 3). If the user selects any of these adverse drug reactions, the tool displays a popup window with information about the definitions and synonyms proposed by the system. Figure 4 shows the synonyms and definitions proposed for the effect 'dispepsia' (dyspepsia). While the most frequent MedDRA synonym was 'indigestión' (indigestion), the most common synonym from MedLinePlus was 'enfermedades del estómago' (stomach diseases).

4 Evaluation

The dataset used for the evaluation is the EasyDPL (easy drug package leaflets) corpus[10], which contains 306 package leaflets annotated with 1,400 adverse drug reactions and their simplest synonyms. The corpus was manually annotated by three trained annotators. The quality and consistency of the corpus were evaluated by measuring inter-annotator agreement (IAA). IAA also determines the complexity of the task and provides an upper bound on the performance of the automatic systems for the simplification of adverse drug reactions in drug package leaflets. In particular, the Fleiss' kappa (Fleiss, 1971) was calculated, which is an extension of Cohen's kappa (Cohen, 1960) that measures the degree of consistency for two or more annotators. The assessment showed a kappa of 0.709, which is considered substantial on the Landis and Koch scale (Landis and Koch, 1977).

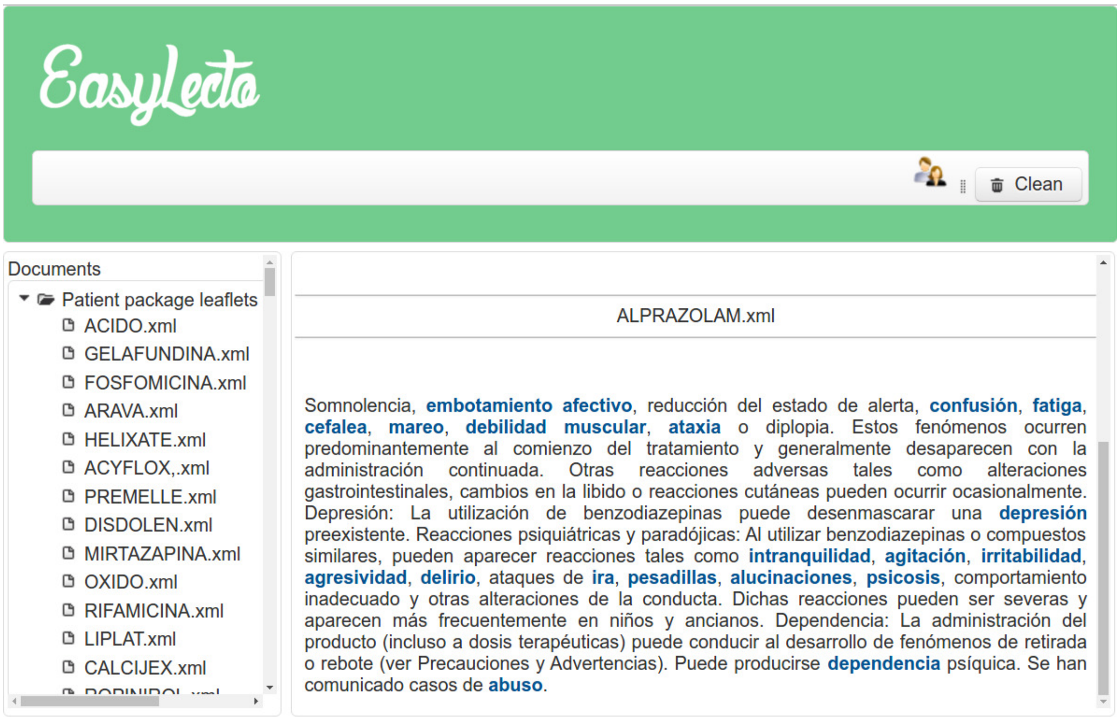
For each drug effect annotated in the EasyDDI corpus, the evaluation consisted in comparing the gold-standard synonym, that is, the synonym proposed by the human annotators, to the simplest synonym, that is, the synonym with the highest frequency in the index built from the MedLinePlus articles. Since we used two different resources, MedDRa and MedLinePlus, in order to achieve the set of synonym candidates, we evaluated the simplest synonym from each of the resources. Thus, for the synonym obtained from MedLinePlus, EasyLecto achieves an accuracy of 68.7%, while for the MedDRA synonym, the accuracy is much lower (around 37.2%). This is mainly due to MedDRA being a highly specific standardized medical terminology, which implies its terms are not familiar to most people. MedLinePlus on the other hand is a health information website for patients, which uses a more readable language and a lay vocabulary.

We conducted an error analysis in order to obtain the main causes of false positives and false negatives in our system. In particular, we studied in detail a random sample of 30 documents. Table 1 presents some errors that our system makes on the EasyDPL corpus. Most errors are due to the absence of a simpler synonym for a term; some terms could only be explained by a small sentence or phrase (for example, terms such as akathisia or eosinophilia). Another cause of error was that some terms were replaced by their hypernyms in the gold-standard corpus (for example, allergic alveolitis was substituted by allergy), whereas the system failed because it does not exploit the hierarchical relationships between terms and is not able to propose more general terms as synonyms for a specific term. Some errors, such as dysphoria-hoarseness or diaphoresis - sweating, may occur due to the lack of synonyms in the resources. An approach based on a word vector model able to compute the similarity between words based on their contexts, could reduce such errors.

In addition to the quantitative evaluation, we also used SurveyMonkey to collect some quick user feedback on the EasyLecto system [11]. We defined a survey with 10 closed-ended questions, in which users should pick just one answer from a list of given options. We asked users about the usefulness and the performance of the EasyLecto, as well as about its usability, design and visual appeal. A total of 26 users completed the survey,

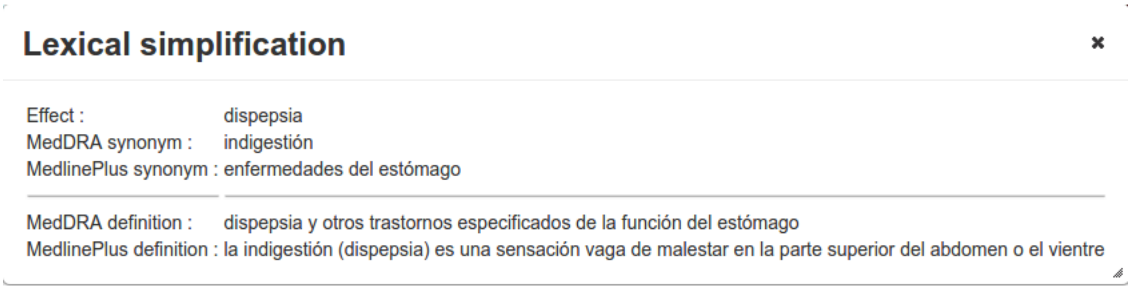
[10] http://labda.inf.uc3m.es/doku.php?id=en:labda_recursosPLN

[11] https://es.surveymonkey.com/r/8HMVJKV



Somnolencia, **embotamiento afectivo**, reducción del estado de alerta, **confusión**, **fatiga**, **cefalea**, **mareo**, **debilidad muscular**, **ataxia** o diplopia. Estos fenómenos ocurren predominantemente al comienzo del tratamiento y generalmente desaparecen con la administración continuada. Otras reacciones adversas tales como alteraciones gastrointestinales, cambios en la libido o reacciones cutáneas pueden ocurrir ocasionalmente. Depresión: La utilización de benzodiazepinas puede desenmascarar una **depresión** preexistente. Reacciones psiquiátricas y paradójicas: Al utilizar benzodiazepinas o compuestos similares, pueden aparecer reacciones tales como **intranquilidad**, **agitación**, **irritabilidad**, **agresividad**, **delirio**, ataques de **ira**, **pesadillas**, **alucinaciones**, **psicosis**, comportamiento inadecuado y otras alteraciones de la conducta. Dichas reacciones pueden ser severas y aparecen más frecuentemente en niños y ancianos. Dependencia: La administración del producto (incluso a dosis terapéuticas) puede conducir al desarrollo de fenómenos de retirada o rebote (ver Precauciones y Advertencias). Puede producirse **dependencia** psíquica. Se han comunicado casos de **abuso**.

Figure 3: A drug package leaflet annotated with the EasyLecto system. Adverse drug reactions are highlighted in blue



Effect : dispepsia
MedDRA synonym : indigestión
MedlinePlus synonym : enfermedades del estómago

MedDRA definition : dispepsia y otros trastornos especificados de la función del estómago
MedlinePlus definition : la indigestión (dispepsia) es una sensación vaga de malestar en la parte superior del abdomen o el vientre

Figure 4: Simplification (synonyms and definitions) for the effect 'dispepsia'.

Drug Effect	Gold-standard synonym	EasyLecto synonym
acatisia (akathisia)	incapacidad de quedarse quieto (inability to sit still)	acatisia
bursitis	hinchazón alrededor de los músculos (swelling around the muscles)	bursitis
eosinofilia (eosinophilia)	problemas en la sangre (blood problems)	eosinofilia
cloasma (chloasma)	manchas durante el embarazo (spots during pregnancy)	cloasma
miositis (myositis)	inflamación en la piel (skin inflammation)	miositis
alveolitis alérgica (allergic alveolitis)	alergía (allergy)	alveolitis alérgica
diaforesis (diaphoresis)	sudoración (sweating)	diaforesis
disforia (dysphoria)	ronquera (hoarseness)	disforia

Table 1: Some errors of the EasyLecto system.

most of them being software engineers or PhD students in computer science. The analysis of the survey shows that most users have positive opinions about the EasyLecto system. Almost 97% of users think that the EasyLecto system helps to simplify drug package leaflets. Regarding the definitions proposed by the system, 75% of users believe that the definitions help to understand the text. Almost 30% of them would like to obtain three or more synonyms from the system. Around 81% of users think that the EasyLecto has a friendly interface.

5 Conclusions and future work

Although drug package leaflets should be designed and written ensuring complete understanding of their contents, several factors can have an influence on patient understanding of drug package leaflets. Low literacy is directly associated with limited understanding and misinterpretation of these documents (Davis et al., 2006b; Davis et al., 2006a). Older people are more likely to have lower literacy skills, as well as decreased memory and poorer reading comprehension (Kutner et al., 2006). Therefore, low literacy along with older age may lead to an unintentional non-compliance or inappropriate use of drugs, leading to dangerous consequences for patients, such as therapeutic failure or adverse drug reactions. Several studies (March et al., 2010; Pires et al., 2015; Piñero-López et al., 2016) have shown that there is an urgent need to improve the quality of drug package leaflets because they are usually too difficult to understand for patients, and this could be a potential source of drug related problems, such as medication errors and adverse drug reactions. In particular, patients have problems to understand those sections describing dosages and adverse drug reactions.

The EasyLecto system aims the simplification of drug package leaflets, in particular, the simplification of terms describing adverse drug reactions by synonyms that are easier to understand by patients. The system uses a dictionary-based approach in order to automatically identify adverse drug reactions in drug package leaflets. MedDRA and MedLinePlus are used as sources of synonyms and definitions for these effects. Our main hypothesis is that a simple word will likely be more common in a collection of texts than their more difficult synonyms. We built an index from a large collection of texts such as MedLinePlus. This index provides us information about how common a word is. EasyLecto was evaluated on a gold-standard corpus with 306 texts manually annotated by three trained experts. Experiments show an accuracy of 68.7% for the MedLinePlus synonym and 37.1% for the MedDRA synonym. Therefore, resources that have been specially written for patients are a better source of simpler synonyms that the specialized terminological resources (such as MedDRA). On the other hand, the error analysis shows that some of the system answers might as well be valid and simple synonyms, even though they are not the same as proposed by the gold-standard corpus. In order to obtain a more realistic evaluation, we plan to extend the EasyDPL corpus by adding several simpler synonyms for each term.

In addition to the quantitative evaluation, the subjective impression of 26 users was documented by a simple questionnaire published in SurveyMonkey. In general, users have positive perceptions of the EasyLecto system. We are aware that our evaluation system based on user experience has a lot of shortcomings (e.g., the number of users is very small and they are not representative of the general public). Therefore, we plan to extend and improve the evaluation with a large set of users that includes elderly users, people with disabilities or with low literacy levels.

In this work, we only focus on the simplification of adverse drug reactions, however we plan to extend our approach in order to simplify not only other medical concepts (such as diseases, medical procedures, medical tests, etc), but also complex words from open-domain texts. As future work, we also plan to integrate additional resources such as BabelNet (Navigli and Ponzetto, 2012) or the UMLS Metathesaurus (Lindberg et al., 1993). In addition to providing broader coverage for terms and more synonyms, these resources will allow to develop a multilingual simplification system.

To the best of our knowledge, while word vector models based on n-grams have already been used (Bott et al., 2012), word vector models trained using deep learning techniques have not been explored for the task of simplification yet. We also plan to study the use of word embeddings learned by Word2Vec (Mikolov et al., 2013) or Glove (Pennington et al., 2014). One important advantage of these models is that they allow to compute the similarity between terms without the need of using synonym dictionaries that are generally domain-dependent.

Acknowledgments

This work was supported by eGovernAbility-Access project (TIN2014-52665-C2-2-R).

References

Alberto Anula. 2007. Tipos de textos, complejidad lingüística y facilitación de la lectura. In *Actas del IV Congreso de la Asociación Asiática de Hispanistas*.

Ricardo Baeza-Yates, Luz Rello, and Julia Dembowski. 2015. Cassa: A context-aware synonym simplification algorithm. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, page 13801385.

Thimoty Barbieri, Antonio BIANCHI, Licia SBATTELLA, Ferdinando CARELLA, and Marco FERRA. 2005. Multiabile: A multimodal learning environment for the inclusion of impaired e-learners using tactile feedbacks, voice, gesturing, and text simplification. *Assistive Technology: From Virtuality to Reality*, 16(1):406–410.

Susana Bautista and Horacio Saggion. 2014. Can numerical expressions be simpler? implementation and demostration of a numerical simplification system for spanish. In *LREC*, pages 956–962.

Stefan Bott, Luz Rello, Biljana Drndarevic, and Horacio Saggion. 2012. Can spanish be simpler? lexsis: Lexical simplification for spanish. In *Proceedings of COLING 2012*, pages 357–374.

Raman Chandrasekar, Christine Doran, and Bangalore Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 1041–1044. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scale. *Educ Psychol Meas*, 20:37–46.

Terry C Davis, Michael S Wolf, Pat F Bass, Mark Middlebrooks, Estela Kennen, David W Baker, Charles L Bennett, Ramon Durazo-Arvizu, Anna Bocchini, Stephanie Savory, et al. 2006a. Low literacy impairs comprehension of prescription drug warning labels. *Journal of general internal medicine*, 21(8):847–851.

Terry C Davis, Michael S Wolf, Pat F Bass, Jason A Thompson, Hugh H Tilson, Marolee Neuberger, and Ruth M Parker. 2006b. Literacy and misunderstanding prescription drug labels. *Annals of Internal Medicine*, 145(12):887–894.

Jan De Belder, Koen Deschacht, and Marie-Francine Moens. 2010. Lexical simplification. In *Proceedings of ITEC2010: 1st international conference on interdisciplinary research on technology, education and communication*.

Siobhan Lucy Devlin. 1999. *Simplifying natural language for aphasic readers*. Ph.D. thesis, University of Sunderland.

Bonnie Dorr, David Zajic, and Richard Schwartz. 2003. Hedge trimmer: A parse-and-trim approach to headline generation. In *Proceedings of the HLT-NAACL 03 on Text summarization workshop-Volume 5*, pages 1–8. Association for Computational Linguistics.

Biljana Drndarevic, Sanja Štajner, and Horacio Saggion. 2012. Reporting simply: A lexical simplification strategy for enhancing text accessibility. In *Proceedings of Easy-to-Read on the Web Symposium*.

EC. 2009. Guideline on the readability of the labelling and package leaflet of medicinal products for human use.

Noémie Elhadad. 2006. Comprehending technical texts: predicting and defining unfamiliar terms. In *AMIA*.

Council EU. 2001. Directive 2001/83/ec of the european parliament and of the council of 6 november 2001 on the community code relating to medicinal products for human use. *Official Journal L*, 311(28):11.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Robert Gunning. 1986. The technique of clear writing.

Siddhartha Jonnalagadda and Graciela Gonzalez. 2011. Biosimplify: an open source sentence simplification engine to improve recall in automatic biomedical information extraction. *arXiv preprint arXiv:1107.5744*.

Siddhartha Jonnalagadda, Luis Tari, Jörg Hakenberg, Chitta Baral, and Graciela Gonzalez. 2009. Towards effective sentence simplification for automatic processing of biomedical text. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 177–180. Association for Computational Linguistics.

Sasikiran Kandula, Dorothy Curtis, and Qing Zeng-Treitler. 2010. A semantic and syntactic text simplification tool for health content. In *AMIA Annu Symp Proc*, volume 2010, pages 366–70.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document.

Mark Kutner, Elizabeth Greenberg, and Justin Baer. 2006. A first look at the literacy of america's adults in the 21st century. nces 2006-470. *National Center for Education Statistics*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Donald A Lindberg, Betsy L Humphreys, and Alexa T McCray. 1993. The unified medical language system. *Methods of information in medicine*, 32(4):281–291.

Cerdá JC March, Rodríguez MA Prieto, Azarola A Ruiz, Lorda P Simón, Cantalejo I Barrio, and Alina Danet. 2010. [quality improvement of health information included in drug information leaflets. patient and health professional expectations]. *Atención primaria/Sociedad Española de Medicina de Familia y Comunitaria*, 42(1):22–27.

Diana McCarthy and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 48–53. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Sarah E Petersen and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *Proceedings of Workshop on Speech and Language Technology for Education*, pages 69–72.

Ángeles María Piñero-López, Pilar Modamio, F. Cecilia Lastra, and L. Eduardo Mariño. 2016. Readability analysis of the package leaflets for biological medicines available on the internet between 2007 and 2013: An analytical longitudinal study. *J Med Internet Res*, 18(5):e100.

Carla Pires, Marina Vigário, and Afonso Cavaco. 2015. Readability of medicinal package leaflets: a systematic review. *Revista de saude publica*, 49:1–13.

Horacio Saggion, Elena Gómez Martínez, Esteban Etayo, Alberto Anula, and Lorena Bourg. 2011. Text simplification in simplext. making text more accessible. *Procesamiento del lenguaje natural*, 47:341–342.

Licia Sbattella and Roberto Tedesco. 2012. Calculating text complexity during the authoring phase. In *Proceedings of Easy-to-Read on the Web Symposium*.

Isabel Segura-Bedmar, Paloma Martínez, Ricardo Revert, and Julián Moreno-Schneider. 2015. Exploring spanish health social media for detecting drug effects. *BMC medical informatics and decision making*, 15(2):1.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1).

Advaith Siddharthan. 2002. Resolving attachment and clause boundary ambiguities for simplifying relative clause constructs. In *Proceedings of the Student Workshop, 40th Meeting of the Association for Computational Linguistics (ACL02)*, pages 60–65.

Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):259–298.

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. Semeval-2012 task 1: English lexical simplification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 347–355. Association for Computational Linguistics.

Irina Temnikova. 2012. *Text Complexity and Text Simplification in the Crisis Management domain*. Ph.D. thesis, University of Wolverhampton.

Zhemin Zhu, 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*, pages 1353–1361. Association for Computational Linguistics.