UNIVERSITY OF THE BASQUE COUNTRY (UPV/EHU)
Basque Language and Communication Departament

PhD dissertation

Readability Assessment and Automatic Text Simplification. The Analysis of Basque Complex Structures

Itziar Gonzalez Dios

2016

UNIVERSITY OF THE BASQUE COUNTRY (UPV/EHU)
Basque Language and Communication Departament

Readability Assessment and Automatic Text Simplification. The Analysis of Basque Complex Structures

This summary is a shortened and translated version of the dissertation entitled "Euskarazko egitura sintaktiko konplexuen analisirako eta testuen sinplifikazio automatikorako proposamena", written by Itziar Gonzalez Dios under the supervision of Dr. Arantza Díaz de Ilarraza and Dr. María Jesús Aranzabe. It also includes the papers which the candidate has published in English on the research presented here.

Donostia (2016).

Caminante,
no hay camino,
se hace camino al andar.

Antonio Machado

Acknowledgements

Lehenik eta behin eskerrak eman nahi nizkieke tesi-lan honen zuzendariak izan diren Arantzari eta Maxuxi, zuen denboragatik, arretagatik eta konfiantzagatik, bidea irekitzeagatik eta bidean gidatzeagatik. Eskerrak eman nahi dizkiet Xabier Artolari eta Izaskun Aldezabali azken orrazketak eta zuzenketak egiteagatik.

Modu batera edo bestela, lan honetan kolaboratu duzuenoi: Itziar Aduriz, Rodrigo Agerri, Manex Agirrezabal, Edurne Aldasoro, Itziar Aldabe, Izaskun Aldezabal, Begoña Altuna, Nora Aranberri, Maxux Aranzabe, Zuhaitz Beloki, Klara Ceberio, Ainara Estarrona, Uxoa Iñurrieta, Mikel Iruskieta, Gorka Labaka, Mikel Lersundi, Oier Lopez de Lacalle, Inigo Lopez-Gazpio, Unai Lopez-Novoa, Vanessa Martin, Itziar Otaduy, Arantxa Otegi, Iñaki San Vicente, Aitor Soroa eta Larraitz Uria (esango nuke denak zaudetela... norbait ahaztu bazait, barkatu!). Zuen " marroitxoak" nire fruitu izan direlako... Amaiari, Estherri eta Kikeri beti laguntzeko prest egon zaretelako!

Nola ez, Ixa taldeko kideei kafe orduak, bazkaltzeko tarteak eta "bestelako ekintzak" ezinbestekoak izan dira tesia aurrera eramateko eta arinagoa izateko! Eta batez ere, bulegokidei eta bulegokide izandakoei. Gora 314 bulegoa eta krisi komiteak! Bestelako "lantxoetan" izan ditudan taldekideei, zuen esperientzia irakasle onena izan baita!

3. pisuan pasillotik bueltaka aurkitu ditudanei, eta, bereziki, Mendiri, txiste txarrekin irribarrea ateratzeagatik... babesagatik eta laguntzagatik!

Il gruppo ItaliaNLP Lab, per la vostra hospitalità, e farmi sentire come a casa! E Chiara, la megliore coinquilina! Non ho parole, grazie mille a tutti!!!

Koadrilakoei eta uniko lagunei, batzuk urrun besteak gertu baina hor egon zeatelako!

Y, por último, a los de casa, por ser vosotros, por estar siempre a mi lado y dispuestos a todo por ayudarme. Eskerrik asko, aita eta ama! Esti, eredua izategatik, eta Izar, egunak alaitzeagatik! A las abuelas, porque una sonrisa vuestra vale más que nada, y a los abuelos, donde quiera que esteis, porque espero que estéis orgullosos de mi!

Bidean topatutako edonori, bidea erraztu edo zaildu duen orori, indartsuago eta ulerkorragoa bihurtzeagatik!

Mila esker!!!

Official acknowledgments

The Department of Education, Universities and Research of the Basque Government who awarded a pre-doctoral fellowship (BFI-2011-392) to the author of this PhD dissertation to conduct this research.

Abstract

In this thesis we have paved the way for the automatic readability assessment and text simplification in the automatic processing of Basque. In order to analyse the complexity of the texts, we have considered the works for other languages targeted to automatic text simplification and we have performed linguistic analyses in Basque corpora. Based on these analysis we have set the linguistic foundations to simplify texts automatically. To assess the readability of the texts automatically, we have implemented *ErreXail*, a system based on linguistic features that uses machine learning techniques. To simplify texts automatically, we have defined the operations that the text simplification system *EuTS* should perform and we have connected them with the modules of the architecture. We have also provided the linguistic information these modules need. As case study, we have implemented a multilingual tool that simplifies parenthetical structures containing biographical information, and we have shown that the results of the linguistic analysis for Basque are also useful for other languages. To contrast our corpus-study-based approach, we have created the ETSC-CBST corpus that contains original and simplified texts. To make the comparison among different approaches, we have defined an annotation scheme.

Contents

List of Figures

List of Tables

INTRODUCTION

1 Presentation of the Ph.D. Project

1.1 Introduction

Millions of texts are produced every day in our society but these texts are not accessible to everybody due to their complexity. Not only people have troubles processing texts[1], Natural Language Processing (NLP) advanced applications get also stuck with long and complex sentences. Let us give an example a machine translation[2] of an original sentence.

Original sentence	**Translation of the original sentence**
1962an Charles De Gaulle eta Konrad Adenauer Bonnen elkartu zirenean 55 miloi lagun bizi ziren herrialde horretan, eta 47 milioi Frantzian.	Charles De Gaulle and Konrad Adenauer in Bonn, when 55 million people were living together in this country, and 47 million in France.

Table 1.1 – Machine translations of an original sentence

In the translation of the original sentence, the verb *elkartu* (meet) has been translated with the word "together", which adds nothing new to the meaning. So, we do not really know what Charles De Gaulle and Konrad Adenauer did in Bonn. Moreover, the main clause has been translated as

[1] When we use the term text we refer always to written text.

[2] This machine translation was done in February, 2013 with the web service of Google Translate https://translate.google.es/.

a subordinate clause headed by "when", that is, the elements of the clauses have been mixed due to its length.

To overcome the problems long and complex sentences cause is the main goal of this work. To that end, we have studied in this thesis two research lines of Natural Language Processing (NLP): Automatic Text Simplification (ATS) and Readability Assessment (RA). ATS seeks to get simpler texts out of the complex ones, keeping the original meaning of the original complex text, and RA analyses the complexity of the texts. In this thesis, we have analysed other works in other languages and we have made the effort to bring them to Basque.

ATS and RA are really important in the digital age since the manual readability assessment and simplification are an expensive and time-consuming task. By means of NLP technology, however, this task can be easier and faster.

ATS has been also considered as part of Natural Language Generation (NLG), since when texts are simplified, language is generated. Two main types of simplification have been carried out in ATS: syntactic simplification and lexical simplification. In the syntactic simplification, complex syntactic structures are rewritten in order to give a simpler one, while in lexical simplification difficult or low frequency words are substituted with more frequent or known words. The systems that perform the simplifications are rule-based, data-driven or hybrid. In general, rule-based systems are based on linguistic knowledge while data-driven systems are based on corpora and statistical methods. Hybrid systems combine both techniques.

When texts are simplified, the target audience of the text is usually taken into account. There are two main target audiences: people and machines. Below, we explain how a simplified text can help to each target audience.

- People:
  - Foreign language learners: as structures and vocabulary are learned step by step, they do not know all of them until the learning process is fulfilled or almost completed.
  - Illiterate or low-literate: as their study or reading level is low refined or developed texts seem difficult for them.
  - Child: as they are learning, they are not able to understand all the concepts.

  - Aphasic: as they have lost a part of the language ability, they have troubles to understand certain structures.
  - Deaf: as they have a different conceptualisation of the world, it is difficult for them to understand common language.
  - Cognitively disabled: due to Alzheimer and other disabilities, they lose the capability to understand.

- Machines:

  - NLP advanced applications[3] (Parsers, machine translation, Q&A, summarization systems...): simplification can be used as preprocess, since texts with short sentences are easily and effectively processed.
  - Devices with small screens (smartphones, tablets...): shorter sentences are displayed in a comfortable way.

ATS is a research line that has gained popularity in the recent years. This is borne by the workshops that have been organised in some of the NLP main conferences (LREC, EACL, Coling): PITR (*Predicting and Improving Text Readability for target reader populations*) was organised in 2012, 2013 and 2014; NLP4ITA (*Natural Language Processing for Improving Textual Accessibility*) in 2012 and 2013, and ATS-MA (*Automatic Text Simplification - Methods and Applications in the Multilingual Society*) in 2014. Two other workshops have been organised in 2016: ISI-NLP (*Improving Social Inclusion using NLP: Tools and resources*) eta QATS (*Quality Assessment for Text Simplification*). There is also shared-task in the latter.

In the works presented in these workshops, conferences and journals, TSA has been diversely considered. The first difference among approaches is the treatment of the information. In some works all the information in the original is retained or tried to retain (Siddharthan, 2006; Gasperin et al., 2009; Aranzabe et al., 2012a), while in others non required information is deleted (Bott et al., 2012b; Barlacchi and Tonelli, 2013). In our opinion, the latter is more related to automatic summarization, because in addition to simplifying texts are also shortened. In fact, TSA has been confused due to their similarities with other NLP research lines and tasks such as automatic

[3]The term sentence simplification has also been used in the works that target NLP advanced applications. Nowadays, however, it is not so used.

summarization, sentence reduction, sentence compression or sentence fusion. Their aim is to give also shorter sentences, but they delete information while it is kept in most of the simplification approaches.

Apart from the treatment of the information, there are other approaches in ATS. For instance, tools to adapt the texts to a certain person have been created in the FIRST project[4]. The coordinator of the project called this kind of simplification "text personalisation" in the ATS-MA workshop. They perform, indeed, text adaptation for each individual.

Two other works also presented in the ATS-MA workshop brought new insights to ATS; for instance, the adaptation of historical texts to the current writing (Vertan and von Hahn, 2014) or the representation as graph of the relations found in the patents (Sheremetyeva, 2014). Bringing text to the current writing is not, in our opinion, simplification but text normalisation. Texts are more accessible in current writing and putting the information as a graph can be easily interpreted but, these insights do not follow the definition so far given in the community.

Another field that has a direct interaction with simplification is in the case of the controlled language. For instance, to simplify texts of crisis management domain controlled languages are used (Temnikova, 2012). Plain language has also to do with simplification. Indeed, plain language guidelines have been used in various TSA works (Bott et al., 2012b; Mitkov and Štajner, 2014). In Table 1.2 we present some approaches in ATS.

Work	**Keeping information**	**Deleting information**	**Plain language**	**Controlled language**	**Others**
Siddharthan (2006)	✔	-	-	-	-
Gasperin et al. (2009)	✔	-	-	-	-
Temnikova (2012)	-	-	-	✔	-
Bott et al. (2012b)	-	✔	✔	-	-
Barlacchi and Tonelli (2013)	-	✔	-	-	-

(Continued on the next page)

[4] http://www.first-asd.eu/ (last retrieved January, 2016)

Work	Keeping information	Deleting information	Plain language	Controlled language	Others
Vertan and von Hahn (2014)	-	-	-	-	✔
Sheremetyeva (2014)	-	-	-	-	✔

Table 1.2 – TSA according to various approaches

Both controlled languages and plain language are methods to get simpler texts. As far as we know, there is no controlled language for Basque and the presence of plain language movements for simplification is quite recent. In any case, we want to mention that the member of the Royal Academy of Basque Language *Euskaltzaindia* Imanol Berriatua published in 1978 a method to learn "basic Basque" based on his experience in Israel. He learned indeed the methods that were been used to recover the Hebraic.

On the other hand, RA analyses the complexity level of the text, for example whether texts are simple or complex. To that end, the linguistic and/or statistic features of the texts are taken into account. In RA the content of the text is taken into account (*readability*), and it is important not to make confusion with legibility, where the impact of the shape or the form of the (the fonts, the justification, the spaces and so on) are considered.

RA has been used in ATS as preprocess or as evaluation. As preprocess, RA is used to know which are the complex text, and therefore, to know which texts should be simplified. As evaluation, readability formulae have been used to test if the simplified text is simpler than the original one. In this thesis we have above all concentrated on the RA that is intended to ATS. Readability has often been confued with legibility but in the former the content is taken into account while in the latter the form of the texts (spacing, font type and size...) is analysed.

1.2 Motivation and scope

This thesis project has been carried out in the Ixa research group[5] of the University of the Basque Country (UPV/EHU). The Ixa group has been working

[5] http://ixa.eus (last retrieved January, 2016)

in NLP for Basque for 27 years creating basic resources and advanced applications. The works of this thesis belong to the part of advanced applications. But it is not only restricted to that area, it also includes a linguistic analysis that is necessary for the computational formalisation of language.

The main motivation of this thesis has been to solve the problem the long and/or complex sentences cause in NLP advanced applications and to offer to people learning Basque easier texts. In addition to that, we want to examine reusability of the tools and resources developed in the Ixa research group.

Following that motivation we have set two scopes: 1) from a linguistic point of view, to make an analysis of the complexity of the sentences and do simplification proposals and 2) from a computational point of view, to provide the readability assessment and text simplification systems the required linguistic information. To accomplish these scopes we raise research questions in four groups:

- **Analysis of text complexity:** Which are the Basque complex structures? What is complexity? How can it be measured?
- **Treatment of text complexity (simplification):** How can be these structures simplified? Which is the process to simplify them? Which operations should be performed?
- **Resources:** Which resources (corpora, tools...) are needed to carry out this process?
- **Comparison to other languages:** Does Basque have special needs, if we compare it with other languages?

As starting point for our thesis, we looked into the resources we needed to perform our study. These resources are the system that performs the auto-evaluation of essays *Idazlanen Autoebaloaziorako Sistema* (IAS) (Castro-Castro et al., 2008), the analysis chain of the Ixa research group (Aduriz et al., 2004), the Reference Corpus for the Processing of Basque EPEC (*Euskararen Prozesamendurako Erreferentzia Corpusa*) (Aduriz et al., 2006a) and the Consumer corpus (Alcázar, 2005). These resources are displayed in Figure 1.1.

We analysed the re-usability of these resources and tools and we consider that IAS is useful for the readability assessment, EPEC and Consumer corpus

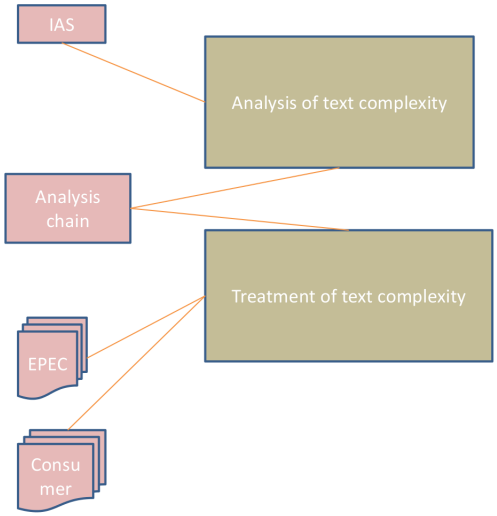

Figure 1.1 – Resources and tools we had at the beginning of the thesis

to perform a linguistic analysis of text complexity and the analysis chain of the Ixa research group to process the texts automatically.

1.3 Outline of this report

This report is the summary of the thesis report in Basque entitled "*Euskarazko egitura sintaktiko konplexuen analisirako eta testuen sinplifikazio automatikorako proposamena*". Here, we present the translations or summaries of chapter of the previously mentioned report and the publications related to this thesis in English.

The report is organised as follows. After this general presentation of the PhD project, we introduce the work done in the readability assessment part, in Chapter 2 we present the system that performs readability assessment called *ErreXail*. In the automatic text simplification part, we summarise the works done in other languages in Chapter 3; in Chapter 4 we present the linguistic analysis of complex structures; in Chapter 5 we explain our approach and the preprocessing tools and in Chapter 6 the proposal of the system that will carry out the automatic simplification called *EuTS*. In the part of the analysis of manually simplified texts, we introduce the Corpus

of Basque Simplified Texts in Chapter 7. We will finish this report the conclusions and the the future work in Chapter 8. There are also three appendixes in this report: the structures of adverbial clauses (Appendix A), the syntactic simplification rules (Appendix B) and the list of the common operations to enlarge the ETSC-CBST corpus (Appendix C).

The publications that have been included are the following:

- Readability Assessment part:

  1. Gonzalez-Dios, I., Aranzabe, M.J., Díaz de Ilarraza, A. Salaberri, H. (2014) Simple or Complex? Assessing the Readability of Basque Texts. In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. pp. 334-344, Dublin City University and Association for Computational Linguistics, Dublin (Ireland). ISBN: 978-1-941643-26-6.

- Automatic Text Simplification part:

  1. Aranzabe, M.J., Díaz de Ilarraza, A., Gonzalez-Dios, I. (2012) First Approach to Automatic Text Simplification in Basque. In Rello, L., Saggion, H., eds.: *Proceedings of the Natural Language Processing for Improving Textual Accessibility (NLP4ITA) workshop (LREC 2012)*, pp. 1–8, Istanbul, Turkey.
  2. Gonzalez-Dios, I., Aranzabe, M.J., Díaz de Ilarraza, A. (2015) Simplifying Basque Texts: the Shallow Syntactic Substitution Simplification. In: *Proceedings the 7th Language & Technology Conference*. pp. 450-454. ISBN: 978-83-932640-7-0.
  3. Aranzabe, M.J., Díaz de Ilarraza, A., Gonzalez-Dios, I. (2013) Transforming Complex Sentences using Dependency Trees for Automatic Text Simplification in Basque. *Procesamiento del Lenguaje Natural*, 50, pp. 61–68.
  4. Gonzalez-Dios, I., Aranzabe, M.J., Díaz de Ilarraza, A., Soraluze, A. (2013) Detecting Apposition for Text Simplification in Basque. *Lecture Notes in Computer Science (LNCS)* 7817, Alexander Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing. Springer. 13th International Conference, CICLing 2013. Part II. pp. 513–524.

5. Gonzalez-Dios, I., Aranzabe, M.J., Díaz de Ilarraza, A. (2014) Making Biographical Data in Wikipedia Readable: A Pattern-based Multilingual Approach. In: *Proceedings of the Workshop on Automatic Text Simplification- Methods and Applications in the Multilingual Society (ATS-MA 2014)*. Workshop at Coling 2014. pp. 11–20.

1.4 Publications and awards

To conclude the presentation chapter, we will show the publications by the candidate. The first publications related to the thesis are signed in alphabetical order. Below, we present the publications closely connected to the thesis:

- Gonzalez-Dios, I., Aranzabe, M.J., Díaz de Ilarraza, A. (2015) Simplifying Basque Texts: the Shallow Syntactic Substitution Simplification. In: *Proceedings the 7th Language & Technology Conference*. pp. 450-454. ISBN: 978-83-932640-7-0.

- Gonzalez-Dios, I., Aranzabe, M.J., Díaz de Ilarraza, A. (2015) *Perpaus adberbialen agerpena, maiztasuna eta kokapena EPEC-DEP corpusean* [Presence, frequency and Position of Basque Adberbial Clauses in The BDT corpus]. UPVEHU LSI TR 02-2015

- Gonzalez-Dios, I., Aranzabe, M.J., Díaz de Ilarraza, A. Salaberri, H. (2014) Simple or Complex? Assessing the Readability of Basque Texts. In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. pp. 334-344, Dublin City University and Association for Computational Linguistics, Dublin (Ireland). ISBN: 978-1-941643-26-6.

- Gonzalez-Dios, I. (2014) Euskarazko testuak errazten: euskal testuen sinplifikazio automatikoa [Making Basque Texts Easier: Automatic Simplification of Basque Texts]. In Aduriz, I. and Urizar, R., eds.: *Euskal hizkuntzalaritzaren egungo zenbait ikerlerro. Hizkuntzalari euskaldunen I. topaketa*. UEU. pp. 135–149.

- Gonzalez-Dios, I., Aranzabe, M.J., Díaz de Ilarraza, A. (2014) Making Biographical Data in Wikipedia Readable: A Pattern-based Multilin-

gual Approach. In: *Proceedings of the Workshop on Automatic Text Simplification- Methods and Applications in the Multilingual Society (ATS-MA 2014)*. Workshop at Coling 2014. pp. 11–20.

- Gonzalez-Dios, I. (2014) Simplificación automática de textos en Euskera [Automatic Simplification of Basque Texts]. In: L. Alfonso Ureña López, Jose Antonio Troyano Jiménez, Francisco Javier Ortega Rodríguez, Eugenio Martínez Cámara (eds.): *Actas de las V Jornadas TIMM*, Cazalla de la Sierra, España, http://ceur-ws.org. pp.45–50

- Gonzalez-Dios, I., Aranzabe, M.J., Díaz de Ilarraza, A. (2013). Testuen sinplifikazio automatikoa: arloaren egungo egoera [Automatic Text Simplification: State of Art]. *Linguamática*. 5(2), 43–63.

- Gonzalez-Dios, I. (2013) Euskarazko testuen sinplifikazio automatikoa [Automatic Simplification of Basque Texts]. *Hizkuntzalari Euskaldunen I. Topaketak. Egungo ikerlerroak.* UEU.

- Gonzalez-Dios, I., Aranzabe, M.J., Díaz de Ilarraza, A., Soraluze, A. (2013) Detecting Apposition for Text Simplification in Basque. *Lecture Notes in Computer Science (LNCS)* 7817, Alexander Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing. Springer. 13th International Conference, CICLing 2013. Part II. pp. 513–524.

- Aranzabe, M.J., Díaz de Ilarraza, A., Gonzalez-Dios, I. (2013) Transforming Complex Sentences using Dependency Trees for Automatic Text Simplification in Basque. *Procesamiento del Lenguaje Natural*, 50, pp. 61–68.

- Aranzabe, M.J., Díaz de Ilarraza, A., Gonzalez-Dios, I. (2012) First Approach to Automatic Text Simplification in Basque. In Rello, L., Saggion, H., eds.: *Proceedings of the Natural Language Processing for Improving Textual Accessibility (NLP4ITA) workshop (LREC 2012)*, pp. 1–8, Istanbul, Turkey.

- Gonzalez-Dios, I. (2011) Euskarazko egitura sintaktikoen azterketa testuen sinplifikazio automatikorako: aposizioak, erlatibozko perpausak eta denborazko perpausak [Analysis of Basque Syntactic Structures for Automatic Text Simplification]. Master's Thesis, University of the Basque Country (UPV-EHU)

These publications have been related to the chapter of in report in Table 1.3.

Chapters	Publications
Chapter 2	Gonzalez-Dios et al. (2014b)
Chapter 3	Gonzalez-Dios et al. (2013b)
Chapters 2, 4, 5 and 6	Gonzalez-Dios (2011), Aranzabe et al. (2012a), Gonzalez-Dios (2013), Gonzalez-Dios (2014b), Gonzalez-Dios (2014a)
Chapters 4, 5, and 6	Aranzabe et al. (2013), Gonzalez-Dios et al. (2013a), Gonzalez-Dios et al. (2014a), Gonzalez-Dios et al. (2015b), Gonzalez-Dios et al. (2015a)

Table 1.3 – Publications connected to the chapters

The following publications are, however, connected to other NLP research lines:

- Agirrezabal, M., Gonzalez-Dios, I., Lopez-Gazpio, I. (2015). Euskararen sorkuntza automatikoa: lehen urratsak [Automatic Generation of Basque: First Steps]. In: *I. Ikergazte Nazioarteko ikerketa euskaraz Kongresuko artikulu-bilduma*. pp. 15–23.

- Aduriz I., Arriola J., Gonzalez-Dios I., Urizar R. (2015) *Funtzio Sintaktikoen Gold Estandarra eskuz etiketatzeko gidalerroak* [Guidelines to Annotate the Gold-standard of Syntactic Functions]. UPVEHU LSI TR 01-2015

- Iruskieta M., Aranzabe M., Diaz de Ilarraza A., Gonzalez-Dios I., Lersundi M., Lopez de Lacalle O. (2013) The RST Basque TreeBank: an online search interface to check rhetorical relations. In: *Proceedings of the 4th Workshop RST and Discourse Studies*, pp. 40-49, Sociedad Brasileira de Computaçao, Fortaleza, CE, Brasil.

- Aldabe I., Gonzalez-Dios I., Lopez-Gazpio I., Madrazo J., Maritxalar M. (2013) Two Approaches to Generate Questions in Basque. *Procesamiento del Lenguaje Natural*, 51 pp. 101-108.

The summary of this thesis won the "most comprehensible twitter user" in the "*Txiokatu zure tesia 6 mezutan #Txiotesia2*" (Tweet your Ph.D. thesis in 6 tweets) contest in 2014. This contest is organised by the *Udako Euskal Unibertsitatea* (Basque Summer University) (UEU).

READABILITY ASSESSMENT

2

Readability Assesment: the *ErreXail* System

In this chapter we present the readability system for Basque *ErreXail*. This system will tell us if a text is simple or complex. If the text is complex, it should be simplified by the system *EuTS* presented in Chapter 6.

Although *ErreXail* was born a preprocessor of *EuTS*, it has already been used to analyse the complexity of the texts in a postedition experiment in machine translation (Aranberri et al., 2014). *ErreXail* has also been used as a basis (and baseline) to distinguish the B1, B2, C1 and C2 levels (Madrazo, 2014).

The details of *ErreXail* are presented in the paper *Simple or Complex? Assessing the Readability of Basque Texts* (Gonzalez-Dios et al., 2014b).

In Figure 2.1 we have added the contributions of this chapter, namely the readability assessment *ErreXail* and the Elhuyar and Zernola corpora, the corpora that have been created to train the system.

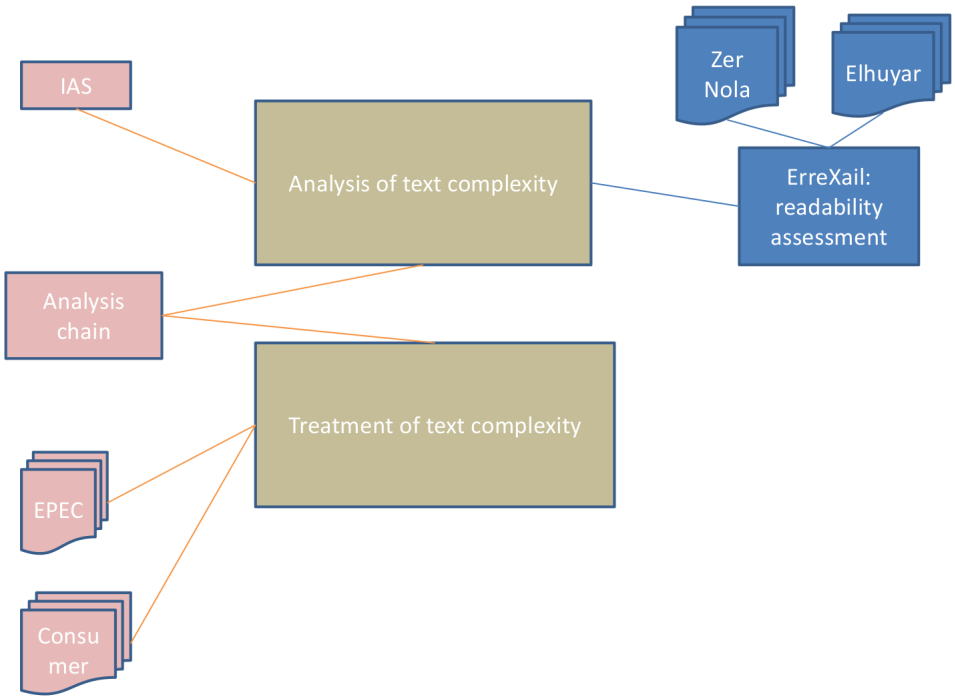

Figure 2.1 – Resources and tools used during thesis, and the contributions

Simple or Complex? Assessing the readability of Basque Texts

Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza, Haritz Salaberri
IXA NLP Group
University of the Basque Country (UPV/EHU)
itziar.gonzalezd@ehu.es

Abstract

In this paper we present a readability assessment system for Basque, *ErreXail*, which is going to be the preprocessing module of a Text Simplification system. To that end we compile two corpora, one of simple texts and another one of complex texts. To analyse those texts, we implement global, lexical, morphological, morpho-syntactic, syntactic and pragmatic features based on other languages and specially considered for Basque. We combine these feature types and we train our classifiers. After testing the classifiers, we detect the features that perform best and the most predictive ones.

1 Introduction

Readability assessment is a research line that aims to grade the difficulty or the ease of the texts. It has been a remarkable question in the educational domain during the last century and is of great importance in Natural Language Processing (NLP) during the last decade. Classical readability formulae like Flesh formula (Flesch, 1948), Dale-Chall formula (Chall and Dale, 1995) and The Gunning FOG index (Gunning, 1968) take into account raw and lexical features and frequency counts. NLP techniques, on the other hand, make possible the consideration of more complex features.

Recent research in NLP (Si and Callan, 2001; Petersen and Ostendorf, 2009; Feng, 2009) has demonstrated that classical readability formulae are unreliable. Moreover, those metrics are language specific.

Readability assessment is also used as a preprocess or evaluation in Text Simplification (TS) systems e.g. for English (Feng et al., 2010), Portuguese (Aluísio et al., 2010), Italian (Dell'Orletta et al., 2011), German (Hancke et al., 2012) and Spanish (Štajner and Saggion, 2013). Given a text the aim of these systems is to decide whether a text is complex or not. So, in case of being difficult, the given text should be simplified.

As far as we know no specific metric has been used to calculate the complexity of Basque texts. The only exception we find is a system for the auto-evaluation of essays *Idazlanen Autoebaluaziorako Sistema* (IAS) (Aldabe et al., 2012) which includes metrics similar to those used in readability assessment. IAS analyses Basque texts after several criteria focused on educational correction such as the clause number in a sentence, types of sentences, word types and lemma number among others. It was foreseen to use this tool in the Basque TS system (Aranzabe et al., 2012). The present work means to add to IAS the capacity of evaluating the complexity of texts by means of new linguistic features and criteria.

In this paper we present *ErreXail*, a readability assessment system for Basque, a Pre-Indo-European agglutinative head-final pro-drop language, which displays a rich inflectional morphology and whose orthography is phonemic. *ErreXail* classifies the texts and decides if they should be simplified or not. This work has two objectives: to build a classifier which will be the preprocess of the TS system and to know which are the most predictive features that differ in complex and simple texts. The study of the most predictive features will help in the linguistic analysis of the complex structures of Basque as well.

This paper is organised as follows: In section 2 we offer an overview about this topic. We present the corpora we gathered and its processing in section 3. In section 4 we summarise the linguistic features we

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 334–344, Dublin, Ireland, August 23-29 2014.

implemented and we present the experiments and their results in section 5. The present system, *ErreXail*, is described in section 6 and in section 7 we compare our work with other studies. Finally, we conclude and outline the future work (section 8).

2 Related work

In the last years new methods have been proposed to assess the readability in NLP. For English, Si and Callan (2001) use statistical models, exactly unigram language models, combined with traditional readability features like sentence length and number of syllables per word. Coh-Metrix (Graesser et al., 2004) is a tool that analyses multiple characteristics and levels of language-discourse such us narrativity, word concreteness or noun overlap. In the 3.0 version[1] 108 indices are available. Pitler and Nenkova (2008) use lexical, syntactic, and discourse features emphasising the importance of discourse features as well. Schwarm and Ostendorf (2005) combine features from statistical language models, parse features, and other traditional features using support vector machines.

It is very interesting to take a look at readability systems for other languages as well. Some readability metrics take them into account special characteristics linked to languages. For example, in Chinese the number of strokes is considered (Pang, 2006), in Japanese the different characters (Sato et al., 2008), in German the word formation (vor der Brück et al., 2008), in French the *passé simple* (François and Fairon, 2012) and the orthographic neighbourhood (Gala et al., 2013) and in Swedish vocabulary resources (Sjöholm, 2012; Falkenjack et al., 2013) among many other features. For Portuguese, Coh-metrix has been adapted (Scarton and Aluísio, 2010) and in Arabic language-specific formulae have been used (Al-Ajlan et al., 2008; Daud et al., 2013). Looking at free word order, head final and rich morphology languages, Sinha et al. (2012) propose two new measures for Hindi and for Bangla based on English formulae. Other systems use only machine learning techniques, e.g. for Chinese (Chen et al., 2011).

The systems whose motivation is Text Simplification analyse linguistic features of the text and then they use machine learning techniques to build the classifiers. These systems have been created for English (Feng et al., 2010), Portuguese (Aluísio et al., 2010), Italian (Dell'Orletta et al., 2011) and German (Hancke et al., 2012). We follow the similar methodology for Basque since we share the same aim.

Readability assessment can be focused on different domains such as legal, medical, education and so on. Interesting points about readability are presented in DuBay (2004) and an analysis of the methods and a review of the systems is presented in Benjamin (2012) and Zamanian and Heydari (2012).

3 Corpora

Being our aim to build a model to distinguish simple and complex texts and to know which are the most predictive features based on NLP techniques, we needed to collect the corpora. We gathered texts from the web and compiled two corpora. The first corpus, henceforth *T-comp*, is composed by 200 texts (100 articles and 100 analysis) from the *Elhuyar aldizkaria*[2], a monthly journal about science and technology in Basque. *T-comp* is meant to be the complex corpus. The second corpus, henceforth *T-simp*, is composed by 200 texts from *ZerNola*[3], a website to popularise science among children up to 12 years and the texts we collected are articles. To find texts specially written for children was really challenging. Main statistics about both corpora are presented in Table 1.

Corpus	Docs.	Sentences	Tokens	Verbs	Nouns
T-comp	200	8593	161161	52229	59510
T-simp	200	2363	39565	12203	13447

Table 1: Corpora statistics

Both corpora were analysed at various levels:

1. Morpho-syntactic analysis by *Morpheus* (Alegria et al., 2002)

[1] `http://cohmetrix.memphis.edu/cohmetrixpr/cohmetrix3.html` (accessed January, 2014)
[2] `http://aldizkaria.elhuyar.org/` (accessed January, 2014)
[3] `http://www.zernola.net/` (accessed January, 2014)

2. Lemmatisation and syntactic function identification by *Eustagger* (Aduriz et al., 2003)

3. Multi-words item identification (Alegria et al., 2004a)

4. Named entities recognition and classification by *Eihera* (Alegria et al., 2004b)

5. Shallow parsing by *Ixati* (Aduriz et al., 2004)

6. Sentence and clause boundaries determination by *MuGak* (Aranzabe et al., 2013)

7. Apposition identification (Gonzalez-Dios et al., 2013)

This preprocess is necessary to perform the analysis of the features presented in section 4.

4 Linguistic features

In this section we summarise the linguistic features implemented to analyse the complexity of the texts. We distinguish different groups of features: global, lexical, morphological, morpho-syntactic, syntactic and pragmatic features. There are in total 94 features. Most of the features we present have already been included in systems for other languages but others have been specially considered for Basque.

4.1 Global features

Global features take into account the document as whole and serve to give an overview of the texts. They are presented in Table 2.

Averages
Average of words per sentence
Average of clauses per sentence
Average of letters per word

Table 2: Global features

These features are based on classical readability formulae and in the criteria taken on the simplification study (Gonzalez-Dios, 2011), namely the sentence length and the clause number per sentence. They are also included in IAS (Aldabe et al., 2012).

4.2 Lexical features

Lexical features are based on lemmas. We calculate the ratios of all the POS tags and different kinds of abbreviations and symbols. We concentrate on particular types of substantives and verbs as well. Part of theses ratios are shown in Table 3. In total there are 39 ratios in this group.

Ratios
Unique lemmas / all the lemmas
Each POS / all the words
Proper Nouns / all the nouns
Named entities / all the nouns
Verbal nouns / all the verbs
Modal verbs / all the verbs
Causative verbs / all the verbs
Intransitive verbs with one arg. (*Nor* verbs) / all the verbs
Intransitive verbs with two arg. (*Nor-Nori* verbs) / all the verbs
Transitive verbs with two arg. (*Nor-Nork* verbs) / all the verbs
Transitive verbs with three arg. (*Nor-Nori-Nork*) verbs / all the verbs
Acronyms / all the words
Abbreviations / all the words
Symbols / all the words

Table 3: Lexical features

Among those features, we want to point out the causative verbs and the intransitive or transitive verbs with one, two or three arguments (arg.) as features related to Basque. Causative verbs are verbs with the

suffix *-arazi* and they are usually translated as "to make someone + *verb*", e.g. *edanarazi*, that stands for "to make someone drink". Other factitive verbs are translated without using that paraphrase like *jakinarazi* that means "to notify", lit. "to make know". The transitivity classification is due to the fact that Basque verb agrees with three grammatical cases (ergative *Nork*, absolutive *Nor* and dative *Nori*) and therefore verbs are grouped according to the arguments they take in Basque grammars.

4.3 Morphological features

Morphological features analyse the different ways lemmas can be realised. These features are summarised in Table 4 and there are 24 ratios in total.

Ratios
Each case ending / all the case endings
Each verb aspect / all the verbs
Each verb tense / all the verbs
Each verb mood / all the verbs
Words with ellipsis / all the words
Each type of words with ellipsis / all the words with ellipsis

Table 4: Morphological features

Basque has 18 case endings (absolutive, ergative, inessive, allative, genitive...), that is, 18 different endings can be attached to the end of the noun phrases. For example, if we attach the inessive *-n* to the noun phrase *etxea* "the house", we get *etxean* "at home". The verb features considered the forms obtained with the inflection.

Verb morphology is very rich in Basque as well. The aspect is attached to the part of the verb which contains the lexical information. There are 4 aspects: puntual (aoristic), perfective, imperfective and future aspect. Verb tenses are usually marked in the auxiliary verb and there are four tenses: present, past, irreal and archaic future[4]. The verbal moods are indicative, subjunctive, imperative and potential. The latter is used to express permissibility or possible circumstances.

Due to the typology of Basque, ellipsis[5] is a normal phenomenon and ellipsis can be even found within a word (verbs, nouns, adjective...); for instance, *dioguna* which means "what we say". This kind of ellipsis occurs e.g. in English, Spanish, French and German as well but in these languages it is realised as a sentence; but it is expressed only by a word in Basque.

4.4 Morpho-syntactic features

Morpho-syntactic features are based on the shallow parsing (chunks[6]) and in the apposition detection (appositions). These features are presented in Table 5.

Ratios
Noun phrases (chunks) / all the phrases
Noun phrases (chunks) / all the sentences
Verb phrases / all the phrases
Appositions / all the phrases
Appositions / all the noun phrases (chunks)

Table 5: Morpho-syntactic features

Contrary to the features so far presented, the morpho-syntactic features take into account mainly more than a word. About apposition, there are 2 types in Basque (Gonzalez-Dios et al., 2013) but we consider all the instances together in this work.

[4]The archaic future we also take into account is not used anymore, but it can be found in old texts. Nowadays, the aspect is used to express actions in the future.

[5]Basque is a pro-drop language and it is very normal to omit the subject, the object and the indirect object because they are marked in the verb. We do not treat this kind of ellipsis in the present work.

[6]Chunks are a continuum of elements with a head and syntactic sense that do not overlap (Abney, 1991).

4.5 Syntactic features

Syntactic features consider average of the subordinate clauses and types of subordinate clauses. They are outlined in Table 6 and there are 10 ratios in total. The types of adverbial clauses are temporal, causal, conditional, modal, concessive, consecutive and modal-temporal. The latter is a clause type which expresses manner and simultaneity of the action in reference to the main clause.

Ratios
Subordinate clauses / all the clauses
Relative clauses / subordinate clauses
Completive clauses / subordinate clauses
Adverbial clauses / subordinate clauses
Each type of adverbial clause / subordinate clauses

Table 6: Syntactic features

In this first approach we decided not to use dependency based features like dependency depth or distance from dependent to head because dependency parsing is time consuming and slows down the preprocessing. Moreover, the importance of syntax is under discussion: Petersen and Ostendorf (2009) find that syntax does not have too much influence while Sjöholm (2012) shows that dependencies are not necessary. Pitler and Nenkova (2008) pointed out the importance of syntax. but Dell'Orletta et al. (2011) demonstrate that for document classification reliable results can be found without syntax. Anyway, syntax is necessary for sentence classification.

4.6 Pragmatic features

In our cases, the pragmatic features we examine are the cohesive devices. These features are summed up in Table 7. There are 12 ratios in total.

Ratios
Each type of conjunction / all the conjunctions
Each type of sentence connector / all the sentence connectors

Table 7: Pragmatic features

Conjunction types are additive, adversative and disjuntive. Sentence connector types are additive, adversative, disjuntive, clarificative, causal, consecutive, concessive and modal.

5 Experiments

We performed two experiments, the first one to build a classifier and the second one to know which are the most predictive features. For both tasks we used the WEKA tool (Hall et al., 2009).

In the first experiment we ran 5 classifiers and evaluated their performance. Those classifiers were Random Forest (Breiman, 2001), the J48 decision tree (Quinlan, 1993), K-Nearest Neighbour, IBk (Aha et al., 1991), Naïve Bayes (John and Langley, 1995) and Support Vector Machine with SMO algorithm (Platt, 1998). We used 10 fold cross-validation, similar to what has been done in other studies.

Taking into account all the features presented in section 4, the best results were obtained using SMO. This way, 89.50 % of the instances were correctly classified. The F-measure for complex text was 0.899 %, for simple texts was 0.891 % and the MAE was 0.105 %. The results using all the features are shown in Table 8.

Random Forest	**J48**	**IBk**	**Naïve Bayes**	**SMO**
88.50	84.75	72.00	84.50	**89.50**

Table 8: Classification results using all the features

We classified each feature type on their own as well and the best results were obtained using only lexical features, 90.75 %. The classification results according to their feature group are presented in Table 9. We only present the classifiers with the best results and these are remarked in bold.

Classifier	Random Forest	J48	SMO
Global	74.25	73.50	**74.75**
Lex.	88.00	85.00	**90.75**
Morph.	**82.00**	71.75	75.00
Morpho-synt.	**78.25**	76.25	72.75
Synt.	71.25	**73.75**	67.75
Prag.	67.50	**70.50**	65.75

Table 9: Classification results of each feature type

We also made different combinations of feature types and the accuracy was improved. The best combination group was the one formed by lexical, morphological, morpho-syntactic and syntactic features and they obtain 93.50 % with SMO. Best results are show in Table 10.

Feature Group	Random Forest	SMO
Global+Lex	87.50	**89.50**
Global+Lex+Morph	87.75	**89.00**
Global+Lex+Morph+Morf-sint	89.25	**89.50**
Global+Lex+Morph+Morph-sint+Sintax	87.25	**90.25**
Morph+Morph-sint	**84.25**	82.25
Morph+Morph-sint+Sintax	**83.25**	80.75
Morph+Morof-sint+Sintax+Prag	**83.75**	82.00
Lex+Morph	88.75	**92.75**
Lex+Morph+Morph-sint	**89.25**	**89.25**
Lex+Morph+Morph-sint+Sintax	89.75	**93.50**
Lex+Morph+Morph-sint+Sintax+Prag	88.50	**90.25**
Sintax+Prag	**78.25**	73.50

Table 10: Classification results using different feature combinations

Combining the feature types, SMO is the best classifier in most of the cases but Random Forest outperforms the results when there are no lexical features.

In the second experiment, we analysed which were the most predictive linguistic features in each group. We used Weka's Information Gain (InfoGain AttributeEval) to create the ranking and we ran it for each feature group. In Table 11 we present the 10 most predictive features taking all the features groups into account.

The results of this experiment are interesting for the linguistic studies on Text Simplification. It shows us indeed which phenomena we should work on next. In these experiment we notice as well the relevance of the lexical features and that syntactic features are not so decisive in document classification.

The features with relevance 0 have been analysed as well. Some of them are e.g. the ratio of the inessive among all the case endings, the ratio of the indicative mood among all the verbal moods, the ratio of the adjectives among all the words and the ratio of the ratio of the present tense among all the verbal tenses.

We also performed a classification experiment with the top 10 features and J48 is the best classifier (its best performance as well). These results are presented in Table 12.

To sum up, our best results are obtained using a combination of features (Lex+Morph+Morph-sint+Sintax). We want to remark the importance of lexical features as well, since they alone outperform all the features and 5 of them are among the top ten features.

6 System overview

The readability system for Basque *ErreXail* has a three-stage architecture (Figure 1).

So, given a Basque written text, we follow next steps:

1. The linguistic analysis will be carried out, that is, morpho-syntactic tagging, lemmatisation, syntactic function identification, named entity recognition, shallow parsing, sentence and clause boundaries determination and apposition identification will be performed. We will use the tools presented in section 3.

Feature and group	Relevance
Proper nouns / common nouns ratio (Lex.)	0.2744
Appositions / noun phrases ratio (Morpho-synt.)	0.2529
Appositions / all phrases ratio (Morpho-synt.)	0.2529
Named entities / common nouns ratio (Lex.)	0.2436
Unique lemmas / all the lemmas ratio (Lex.)	0.2394
Acronyms / all the words ratio (Lex.)	0.2376
Causative verbs / all the verbs ratio (Lex.)	0.2099
Modal-temporal clauses / subordinate clauses ratio (Synt.)	0.2056
Destinative case endings / all the case endings ratio (Morph.)	0.1968
Connectors of clarification / all the connectors ratio (Prag.)	0.1957

Table 11: Most predictive features

Random Forest	J48	IBk	Naïve Bayes	SMO
87.75	**88.25**	72.00	83.25	87.00

Table 12: Classification results using the top 10 features

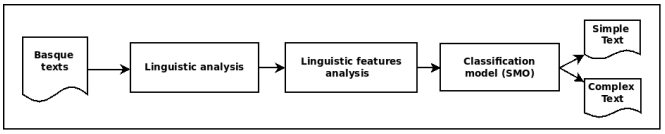

Figure 1: The architecture of system

2. Texts will be analysed according to the features and measures presented in section 4.

3. We will use the SMO Support Vector Machine as classification model, since that was the best classifier in the experiments exposed in section 5. To speed up the process for Text Simplification, we will analyse only the combination of lexical, morphological, morpho-syntactic and syntactic (Lex+Morph+Morph-sint+Sintax) features.

Although the first application of this system will be the preprocessing of texts for the Basque TS system, the system we present in this paper is independent and can be used for any other application. We want to remark that this study, as it is based on other languages, could be applied to any other language as well provided that the text could be analysed similar to us.

7 Discussion

The task of text classification has been carried out by several studies before. Due to our small corpus we were only able to discriminate between complex and simple texts like Dell'Orletta et al. (2011) and Hancke et al. (2012), other studies have classified more complexity levels (Schwarm and Ostendorf, 2005; Aluísio et al., 2010; François and Fairon, 2012). In this section we are going to compare our system with other systems that share our same goal, namely to know which texts should be simplified.

Comparing our experiment with studies that classify two grades and use SMO, Hancke et al. (2012) obtain an accuracy of 89.7 % with a 10 fold cross-validation. These results are very close to ours, although their data compiles 4603 documents and ours 400. According to the feature type, their best type is the morphological, obtaining 85.4 % of accuracy. Combining lexical, language model and morphological features they obtain 89.4 % of accuracy. To analyse their 10 most predictive features, they use Information Gain as well but we do not share any feature in common.

Dell'Orletta et al. (2011) perform three different experiments but only their first experiment is similar to our work. For that classification experiment they use 638 documents and follow a 5 fold cross-validation process of the Euclidian distance between vectors. Taking into account all the features the accuracy of their system is 97.02 %. However, their best performance is 98.12 % when they only use the combination of raw, lexical and morpho-syntactic features.

Aluísio et al. (2010) assess the readability of the texts according to three levels: rudimentary, basic and advanced. In total they compile 592 texts. Using SMO, 10 fold cross-validation and standard classification, they obtain 0.276 MAE taking into account all the features. The F-measure for original texts is 0.913, for natural simplification 0.483 and for strong simplification 0.732. They experiment with feature types as well but they obtain their best results using all the features. Among their highly correlated features they present the incidence of apposition in second place as we do here. We do not have any other feature in common.

Among other readability assessment whose motivation is TS, Feng et al. (2010) use LIBSVM (Chang and Lin, 2001) and Logistic Regression from WEKA and 10 fold cross-validation. They assess the readability of grade texts and obtain as best results 59.63 % with LIBSVM and 57.59 % with Logistic Regression. Since they assess different grades and use other classifiers it is impossible to compare with our results but we find that we share predictive features. They found out that named entity density and and nouns have predictive power as well.

8 Conclusion and perspectives

In this paper we have presented the first readability assessment system for the Basque language. We have implemented 94 ratios based on linguistic features similar to those used in other languages and specially defined for Basque and we have built a classifier which is able to discriminate between difficult and easy texts. We have also determined which are the most predictive features. From our experiments we conclude that using only lexical features or a combination of features types we obtain better results than using all the features. Moreover, we deduce that we do not need to use time consuming resources like dependency parsing or big corpora to obtain good results.

For the future, we could implement new features like word formation or word ordering both based in other languages and in neurolinguistic studies that are being carried out for Basque. Other machine learning techniques can be used, e.g. language models and in the case of getting a bigger corpora or a graded one, we could even try to differentiate more reading levels. We also envisage readability assessment at sentence level in near future.

Acknowledgements

Itziar Gonzalez-Dios's work is funded by a PhD grant from the Basque Government. We thank Lorea Arakistain and Iñaki San Vicente from *Elhuyar Fundazioa* for providing the corpora. We also want to thank Olatz Arregi for her comments. This research was supported by the the Basque Government (IT344-10), and the Spanish Ministry of Science and Innovation, Hibrido Sint project (MICINN, TIN2010-202181).

References

Steven P. Abney. 1991. Parsing by Chunks. In Robert C. Berwick, Steven P. Abney, and Carol Tenny, editors, *Principle-Based Parsing: Computation and Psycholinguistics*. Kluwer Academic.

Itziar Aduriz, Izaskun Aldezabal, Iñaki Alegria, Jose Mari Arriola, Arantza Díaz de Ilarraza, Nerea Ezeiza, and Koldo Gojenola. 2003. Finite State Applications for Basque. In *EACL'2003 Workshop on Finite-State Methods in Natural Language Processing.*, pages 3–11.

Itziar Aduriz, María Jesús Aranzabe, Jose Mari Arriola, Arantza Díaz de Ilarraza, Koldo Gojenola, Maite Oronoz, and Larraitz Uria. 2004. A cascaded syntactic analyser for Basque. *Computational Linguistics and Intelligent Text Processing*, pages 124–134.

David W. Aha, Dennis Kibler, and Marc C. Albert. 1991. Instance-based learning algorithms. *Machine Learning*, 6:37–66.

Amani A Al-Ajlan, Hend S Al-Khalifa, and A Al-Salman. 2008. Towards the development of an automatic readability measurements for Arabic language. In *Digital Information Management, 2008. ICDIM 2008. Third International Conference on*, pages 506–511. IEEE.

Itziar Aldabe, Montse Maritxalar, Olatz Perez de Viaspre, and Uria Larraitz. 2012. Automatic Exercise Generation in an Essay Scoring System. In *Proceedings of the 20th International Conference on Computers in Education*, pages 671–673.

Iñaki Alegria, María Jesús Aranzabe, Aitzol Ezeiza, Nerea Ezeiza, and Ruben Urizar. 2002. Robustness and customisation in an analyser/lemmatiser for Basque. In *LREC-2002 Customizing knowledge in NLP applications workshop*, pages 1–6, Las Palmas de Gran Canaria, May.

Iñaki Alegria, Olatz Ansa, Xabier Artola, Nerea Ezeiza, Koldo Gojenola, and Ruben Urizar. 2004a. Representation and treatment of multiword expressions in Basque. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, pages 48–55. Association for Computational Linguistics.

Iñaki Alegria, Olatz Arregi, Irene Balza, Nerea Ezeiza, Izaskun Fernandez, and Ruben Urizar. 2004b. Design and development of a named entity recognizer for an agglutinative language. In *First International Joint Conference on NLP (IJCNLP-04). Workshop on Named Entity Recognition.*

Sandra Aluísio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9. Association for Computational Linguistics.

María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Itziar Gonzalez-Dios. 2012. First Approach to Automatic Text Simplification in Basque. In Luz Rello and Horacio Saggion, editors, *Proceedings of the Natural Language Processing for Improving Textual Accessibility (NLP4ITA) workshop (LREC 2012)*, pages 1–8.

María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Itziar Gonzalez-Dios. 2013. Transforming Complex Sentences using Dependency Trees for Automatic Text Simplification in Basque. *Procesamiento de Lenguaje Natural*, 50:61–68.

Rebekah George Benjamin. 2012. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24(1):63–88.

Leo Breiman. 2001. Random Forests. *Machine Learning*, 45(1):5–32.

Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability Revisited: The New DaleChall Readability Formula*. Brookline Books, Cambridge, MA.

Chih-Chung Chang and Chih-Jen Lin. 2001. Libsvm - a library for support vector machines. The Weka classifier works with version 2.82 of LIBSVM.

Yaw-Huei Chen, Yi-Han Tsai, and Yu-Ta Chen. 2011. Chinese readability assessment using TF-IDF and SVM. In *Machine Learning and Cybernetics (ICMLC), 2011 International Conference on*, volume 2, pages 705–710. IEEE.

Nuraihan Mat Daud, Haslina Hassan, and Normaziah Abdul Aziz. 2013. A Corpus-Based Readability Formula for Estimate of Arabic Texts Reading Difficulty. *World Applied Sciences Journal*, 21:168–173.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. READ-IT: assessing readability of Italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, SLPAT '11, pages 73–83, Stroudsburg, PA, USA. Association for Computational Linguistics.

William H. DuBay. 2004. The Principles of Readability. *Impact Information*, pages 1–76.

Johan Falkenjack, Katarina Heimann Mühlenbock, and Arne Jönsson. 2013. Features indicating readability in Swedish text. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 27–40.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 276–284. Association for Computational Linguistics.

Lijun Feng. 2009. Automatic Readability Assessment for People with Intellectual Disabilities. *SIGACCESS Access. Comput.*, (93):84–91, January.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Thomas François and Cédrick Fairon. 2012. An AI readability formula for French as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477. Association for Computational Linguistics.

Núria Gala, Thomas François, and Cédrick Fairon. 2013. Towards a French lexicon with difficulty measures: NLP helpnig to bridge the gap between traditional dictionaries and specialized lexicons. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, and M. Tuulik, editors, *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia.*, pages 132–151, Ljubljana/Tallinn. Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.

Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Ander Soraluze. 2013. Detecting Apposition for Text Simplification in Basque. In *Computational Linguistics and Intelligent Text Processing*, pages 513–524. Springer.

Itziar Gonzalez-Dios. 2011. Euskarazko egitura sintaktikoen azterketa testuen sinplifikazio automatikorako: Aposizioak, erlatibozko perpausak eta denborazko perpausak. Master's thesis, University of the Basque Country (UPV/EHU).

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods*, 36(2):193–202.

Robert Gunning. 1968. *The technique of clear writing*. McGraw-Hill New York.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability Classification for German using lexical, syntactic, and morphological features. In *COLING 2012: Technical Papers*, page 10631080.

George H. John and Pat Langley. 1995. Estimating Continuous Distributions in Bayesian Classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, San Mateo. Morgan Kaufmann.

Lau Tak Pang. 2006. *Chinese Readability Analysis and its Applications on the Internet*. Ph.D. thesis, The Chinese University of Hong Kong.

Sarah E. Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer Speech & Language*, 23(1):89–106.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 186–195. Association for Computational Linguistics.

John C. Platt. 1998. Fast Training of Support Vector Machines using Sequential Minimal Optimization. In Bernhard Schlkopf, Christopher J. C Burges, and Alexander J. Smola, editors, *Advances in Kernel Methods-Support Vector Learning*. MIT Press.

Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

Satoshi Sato, Suguru Matsuyoshi, and Yohsuke Kondoh. 2008. Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Joseph Maegaard, Benteand Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA).

Carolina Evaristo Scarton and Sandra Maria Aluísio. 2010. Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português. *Linguamática*, 2(1):45–61.

Sarah E Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 523–530. Association for Computational Linguistics.

Luo Si and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 574–576. ACM.

Manjira Sinha, Sakshi Sharma, Tirthankar Dasgupta, and Anupam Basu. 2012. New Readability Measures for Bangla and Hindi Texts. In *Proceedings of COLING 2012: Posters*, pages 1141–1150, Mumbai, India, December. The COLING 2012 Organizing Committee.

Johan Sjöholm. 2012. Probability as readability: A new machine learning approach to readability assessment for written Swedish. Master's thesis, Linköping.

Tim vor der Brück, Sven Hartrumpf, and Hermann Helbig. 2008. A readability checker with supervised learning using deep indicators. *Informatica*, 32(4):429–435.

Sanja Štajner and Horacio Saggion. 2013. Readability Indices for Automatic Evaluation of Text Simplification Systems: A Feasibility Study for Spanish. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 374–382, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Mostafa Zamanian and Pooneh Heydari. 2012. Readability of texts: State of the art. *Theory and Practice in Language Studies*, 2(1):43–53.

AUTOMATIC TEXT SIMPLIFICATION

3 State of the Art

In this chapter we present the related works in Automatic Text Simplification (ATS) in a schematic way. In the following chapters, we also present the related works to compare them with ours. The review of the State of Art until 2013 has also been published in the paper *Testuen sinplifikazio automatikoa: arloaren egungo egoera* [Automatic Text Simplification: State of Art] (Gonzalez-Dios et al., 2013b).

ATS is a research line in Natural Language Processing (NLP) that aims to reduce the complexity of the text to a certain target audience. As can be seen in Figure 3.1, ATS has flourished the last years.

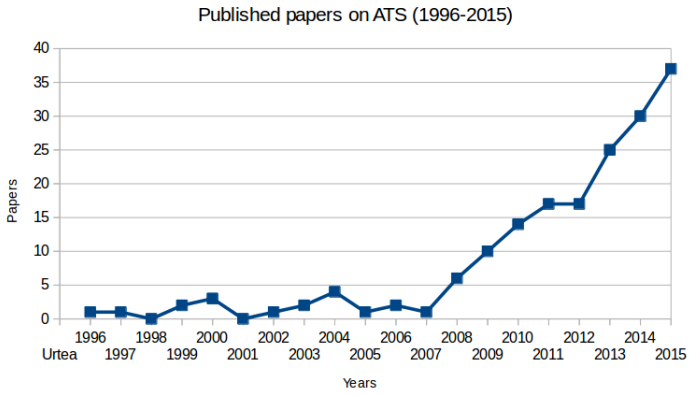

Figure 3.1 – Published paper in ATS during 1996-2015

Below we summarise the most important points for our project. Two main

types of simplification have been performed in ATS: syntactic simplification and lexical simplification. In Table 3.1 we show some examples of syntactic simplification in various languages and in Table 3.2 we give examples of lexical simplification in English and Spanish.

Language	Original sentence	Simplified sentences
English Siddharthan (2002)	*Needing* money to pay my rent, I forced myself to beg my parents.	I needed money to pay my rent. I forced myself to beg my parents.
Brazilian Portuguese Specia et al. (2008)	Vários produtos brasileiros enfrentam barreiras para entrarem nos EUA, *ao passo que* o mercado brasileiro está basicamente aberto.	Vários produtos brasileiros enfrentam barreiras para entrarem nos EUA. Mas o mercado brasileiro está basicamente aberto.
Spanish Bott and Saggion (2012)	Los participantes (...) recibirán como obsequio un libro *editado* por el Ayuntamiento (...).	Los participantes (...) recibirán como obsequio un libro. Este libro está editado por el Ayuntamiento (...).
French Seretan (2012)	Il faut favoriser l'éducation des enfants et des adultes pour une prise de conscience des risques, *mais* aussi développer la sécurisation des réseaux routiers (...).	Il faut favoriser l'éducation des enfants et des adultes pour une prise de conscience des risques. Mais il faut aussi développer la sécurisation des réseaux routiers (...).
Italian Barlacchi and Tonelli (2013)	Ernesta *stava mangiando* la torta *con i suoi amici*.	Ernesta mangia la torta.

Table 3.1 – Examples of syntactic simplification in various languages

Language	Original sentence	Simplified sentences
English Specia et al. (2012)	Hitler committed terrible *atrocities* during the second World War.	Hitler committed terrible *cruelties* during the second World War.
Spanish Bott et al. (2012a)	El visitante puede contemplar los óleos y esculturas que se exponen en la *pinacoteca*.	El visitante puede contemplar los óleos y esculturas que se exponen en el *museo*.

Table 3.2 – Examples of lexical simplification in English and Spanish

Concentrating on syntactic simplification, in Table 3.3 we present the phenomena that have been treated and in Table 3.4 the simplification operations

that have been performed. We give the term used by each author.

Work	Treated phenomena
English	
Siddharthan (2002)	Adjectival or relative clauses, adverbial clauses, coordinated clauses, subordinated clauses, correlated clauses, participial clauses, appositive clauses and voice
De Belder and Moens (2010)	Appositions, relative clauses, prefix subordination, infix coordination and subordination
Siddharthan (2010)	Lexico-syntactic discourse markers
Evans (2011)	Coordinate structures
Poornima et al. (2011)	Relative pronouns, coordinating and subordinating conjunctions
Siddharthan (2011)	Coordination (verb phrases and full clauses), subordination, apposition, relative clauses, passive voice
Peng et al. (2012)	Coordination, relative clauses, and appositions
Brazilian Portuguese	
Aluísio et al. (2008a)	Appositions, relative clauses, subordinate clauses, coordinate clauses, sentences with non-finite verbs, and passive voice
Spanish	
Bott et al. (2012b)	Relative clauses, gerundive constructions and participle constructions, coordinations of clauses, coordinations of objects clauses
French	
Seretan (2012)	Hedging phrases, appositive phrases and clauses, reporting phrases and clauses, coordinated clauses, subordinated clauses, object relative clauses, gerundial clauses, long adjuncts, long post-nominal modifiers, long participial modifiers, small clauses and de-cleft

Table 3.3 – Treated syntactic phenomena in various languages

Work	Operations
Japonese	
Inui et al. (2003)	Syntactic/structural paraphrase remove a cleft construction from a sentence and divide a sentence
English	
Zhu et al. (2010)	Splitting, dropping, reordering and substitution
Vu et al. (2014)	Splitting, dropping, reordering, and substitution

(Continued on next page)

Work	Operations
Brazilian Portuguese Aluísio et al. (2008a)	Splitting sentences, changing discourse markers, changing passive to active voice, inverting clause order and non-simplification
Spanish Saggion et al. (2011)	Change, delete, insert, split
Swedish Rybing et al. (2010)	Remove or replace sub phrases and add new syntactical information to the text
Rennes and Jönsson (2015)	Substitute phrase, delete phrases, rearrangement of words and split
French Seretan (2012)	Delete, extract, split, promote and de-cleft
Brouwers et al.; Brouwers et al. (2012; 2014)	Deletion, modification and splitting
Bulgarian Lozanova et al. (2013)	Clause splitting, simplification of syntactic structure of complex sentences, anaphora resolution, subject recovery, clause reordering and insertion of additional phrases
Korean Chung et al. (2013)	Split sentences, relocate arguments

Table 3.4 – Syntactic simplification operations

In Tables 3.5 and 3.6 we have classified the ATS systems and works according to their simplification type (syntactic, lexical or other) and the technique they used (rule-based, statistical or data-driven, using a machine translation (MT)[1] approach or hybrid).

Systems	Syntactic simpl.				Lexical simpl.	Other
	Rule	Stat.	MT	Hyb.		
English Chandrasekar et al. (1996), Chandrasekar and Srinivas (1997)	✔	-	-	-	-	-

(Continued on next page)

[1]Although MT techniques are a subgroup of data-driven techniques, we give them a special place because they have been productive in the last years.

Systems	**Syntactic simpl.**				**Lexical simpl.**	**Other**
	Rule	Stat.	MT	Hyb.		
PSET project, Systar Carroll et al. (1998), Canning and Tait (1999)	✔	-	-	-	✔	-
Siddharthan (2002), Siddharthan (2006), Siddharthan (2010), Siddharthan (2011), Angrosh and Siddharthan (2014)	✔	-	-	-	✔	✔
Beigman Klebanov et al. (2004)	✔	-	-	-	-	-
Daelemans et al. (2004)	✔	✔	-	-	-	-
Doi and Sumita (2004)	-	✔	-	-	-	-
Max (2006)	✔	-	-	-	-	
SIMTEXT Damay et al. (2006), Ong et al. (2007)	✔	-	-	-	✔	-
Vickrey and Koller (2008)	-	-	-	✔	-	-
BioSimplify Jonnalagadda et al. (2009), Jonnalagadda and Gonzalez (2010a), Jonnalagadda and Gonzalez (2010b)	✔	✔	-	-	-	-
De Belder and Moens (2010), De Belder et al. (2010)	-	-	-	✔	✔	-
Kandula et al. (2010)	-	-	-	-	-	✔
Yatskar et al. (2010)	-	-	-	-	✔	-
Zhu et al. (2010)	-	-	✔	-	✔	-
Bach et al. (2011)	-	✔	-	-	-	-
Bawakid and Oussalah (2011)	✔	-	-	-	-	-
Biran et al. (2011)	-	-	-	-	✔	-
Coster and Kauchak (2011)	-	-	✔	-	✔	-
Evans (2011)	-	-	-	✔	-	-

(Continued on next page)

Systems	**Syntactic simpl.**				**Lexical simpl.**	**Other**
	Rule	Stat.	MT	Hyb.		
Medero and Ostendorf (2011)	-	✔	-	-	-	-
Poornima et al. (2011)	✔	-	-	-	-	-
Tur et al. (2011)	✔	-	-	-	-	-
Woodsend and Lapata (2011)	-	✔	-	-	✔	-
Amoia and Romanelli (2012)	-	-	-	-	✔	-
Chen et al. (2012)	✔	-	-	-	✔	-
Jauhar and Specia (2012)	-	-	-	-	✔	-
Johannsen et al. (2012)	-	-	-	-	✔	-
Ligozat et al. (2012), Ligozat et al. (2013)	-	-	-	-	✔	-
Minard et al. (2012)	-	-	-	-	✔	-
iSimp Peng et al. (2012), Peng et al. (2014)	✔	-	-	-	-	-
Shardlow (2012), Shardlow (2013)	-	-	-	-	✔	-
Silveira Botelho and Branco (2012)	✔	-	-	-	-	-
Sinha (2012)	-	-	-	-	✔	-
Specia et al. (2012)	-	-	-	-	✔	-
Srivastava and Sanyal (2012)	✔	-	-	-	-	-
Temnikova et al. (2012)	✔	-	-	-	✔	-
Thomas and Anderson (2012)	-	-	-	-	✔	-
Wubben et al. (2012)	-	-	✔	-	✔	-
Bautista et al. (2013)	-	-	-	-	-	✔
Feblowitz and Kauchak (2013), Kauchak (2013)	-	✔	-	-	✔	-
Leroy et al. (2013)	-	-	-	-	✔	✔
Nunes et al. (2013)	-	-	-	-	✔	-

(Continued on next page)

Systems	Syntactic simpl.				Lexical simpl.	Other
	Rule	Stat.	MT	Hyb.		
Paetzold and Specia (2013), Paetzold (2015), Paetzold and Specia (2015)	✔	-	-	-	✔	-
Vu et al. (2014)	-	✔	-	-	-	-
Narayan and Gardent (2014), Narayan and Gardent (2015)	-	✔	✔	-	✔	-
Štajner and Saggion (2015)	-	-	✔	-	✔	-

Table 3.5 – ATS systems for English according to their simplification type and technique

Systems and languages	Syntactic simpl.				Lexical simpl.	Other
	Rule	Stat.	MT	Hyb.		
Japonese						
Inui et al. (2003)	✔	-	-	-	✔	-
Kajiwara and Yamamoto (2015)	-	-	-	-	✔	-
Portuguese						
PorSimples proiektua Aluísio et al. (2008a), Candido et al. (2009), Scarton et al. (2010)	✔	-	-	-	✔	-
Specia (2010)	-	-	✔	-	✔	-
Silveira Botelho and Branco (2012)	✔	-	-	-	-	-
Štajner and Saggion (2015)	-	-	✔	-	✔	-
Swedish						
CogFLUX Rybing et al. (2010), Rennes and Jönsson (2015)	✔	-	-	-	-	-
Keskisärkkä (2012)	-	-	-	-	✔	-
Arabic						
Al-Subaihin and Al-Khalifa (2011)	✔	-	-	-	✔	-
Spanish						

(Continued on next page)

Systems and languages	**Syntactic simpl.**				**Lexical simpl.**	**Other**
	Rule	Stat.	MT	Hyb.		
Simplex project Saggion et al. (2011) Bott et al. (2012a), Bott et al. (2012b), Saggion et al. (2013), Drndarević et al. (2013), Štajner (2014), Štajner et al., 2015, Štajner and Saggion (2015), Saggion et al. (2015), Baeza-Yates et al. (2015), Saggion et al. (2016)	✔	-	✔	-	✔	-
Bautista et al. (2012), Bautista and Saggion (2014)	-	-	-	-	-	✔
Fajardo et al. (2013)	-	-	-	-	-	✔
French						
Brouwers et al. (2012) Brouwers et al. (2014)	-	-	-	✔	-	-
Seretan (2012)	-	-	-	✔	-	-
Danish						
Klerke and Søgaard (2013)	-	✔	-	-	-	-
Italian						
ERNESTA Barlacchi and Tonelli (2013)	✔	-	-	-	-	-
Bulgarian						
Lozanova et al. (2013)	✔	-	-	-	-	-
Korean						
Chung et al. (2013)	✔	-	-	-	-	-

Table 3.6 – ATS systems for several languages according to their simplification type and technique

To evaluate these systems several methods have been proposed: readability assessment techniques, questions to users, MT metrics, evaluations against gold standards and so on. Most of them systems tend to be evaluated intrinsically and extrinsically.

4 Linguistic Analysis of Complex Syntactic Structures

In this chapter we present the linguistic analysis of the complex phenomena and we have shown their simplification proposals.

4.1 Target phenomena, resources and methodology

For us, complex phenomena are the sentences that have more that one verb (coordinate and subordinate clauses), apposition and parenthetical structures. Those phenomena have also been considered as complex in other studies, e.g. for English (Siddharthan, 2002) and Brazilian Portuguese (Specia et al., 2008; Aluísio et al., 2008a, b) (See Table 3.3).

The corpora that we have used in this study are the *Consumer corpusa* (Alcázar, 2005), *Euskararen Prozesamendurako Erreferentzia Corpusa* (the Reference Corpus for the Processing of Basque) (EPEC) (Aduriz et al., 2006a), the Wikipedia and the *Elhuyar corpusa*. These corpora have been used to study the complex phenomena in different domains. For the study of adverbial clauses, we also have used the corpus *Lexikoaren Behatokia*[1] to make a list of lemma frequency.

[1] http://lexikoarenbehatokia.euskaltzaindia.net/aurkezpena.htm (last retrieved: April 2015)

The grammars we have mainly used are *Euskal Gramatika Lehen Urratsak* (EGLU) ('Basque Grammar First Steps') (Euskaltzaindia, 1999, 2005, 2011), *Sareko Euskal Gramatika*[2] (SEG) ('Online Basque Grammar'). These grammars have been used to help and support our linguistic analysis. We also have look up the online version of 'A Brief Grammar of Euskara, the Basque Language'[3] (Laka, 1996), *A Grammar of Basque* (Hualde and Ortiz de Urbina, 2003) and *Euskal Gramatika Laburra* (EGLA) ('Short Basque Grammar') (Euskaltzaindia, 2002).

The methodology to study the complex phenomena has been the following:

1. Select a phenomena and extract sentences with it in corpora
2. Perform the analysis of the sentences and contrast the data we find with the information in the descriptive grammar
3. If necessary, sub-classify the phenomena
4. Make simplification proposals
5. Check if the simplification proposal is also valid in other domain or helps in a NLP tool
6. Document the simplification proposal

We have made a simplification proposal for each phenomena considered as complex. In what follows we give examples of each one. The original sentences are given in the (a) part of the example with glosses and the targeted phenomena are marked in bold both in Basque and English. The simplified sentences are given in the (b) part of the example.

4.2 Simplification proposals

In the section we present the simplification proposals for coordinate clauses, noun clauses, relative clauses, adverbial clauses, apposition and parenthetical structures. The general proposal is i) to split the sentences, ii) reconstruct

[2] http://www.ehu.eus/seg/aurkezpena (last retrieved April, 2015)

[3] http://www.ehu.eus/eu/web/eins/a-brief-grammar-of-euskara (last retrieved April, 2015)

new simple sentences by removing relational marks and adding elements that will keep the meaning of the original, iii) reordering the created sentences in text and iv) correcting possible mistakes and punctuation marks.

4.2.1 Coordinate clauses

The simplification proposal for coordinate clauses is i) to split both coordinate clauses. ii) The coordinating conjunction will be kept in the second clause in all the cases, except for *and*, that will be removed. In (1) we present a sentence where this proposal has been applied.

(1) a. *Irlandako Poliziak RIRAko 14 ustezko kide*
Ireland-ADN police-ERG RIRA-ADN 14 alleged-Ø member-ABS
atxilotu ditu azken astean,
arrest-PRF aux-3SGERG.3PLABS.PRS.IND last-Ø week-INE,
baina *horietako zazpi jada aske utzi*
but them-ADN seven-ABS already free-ABS leave-PRF
ditu, kargurik gabe.
aux-3SGERG.3PLABS.PRS.IND, charges-PART without.
'The police of Ireland has arrested 14 alleged RIRA members in the last week, **but** seven of them are already free, without charges.'

b. i. *Irlandako Poliziak RIRAko 14 ustezko kide atxilotu ditu azken astean.*
'The police of Ireland has arrested 14 alleged RIRA members in the last week.'
ii. *Baina horietako zazpi jada aske utzi ditu, kargurik gabe.*
'But seven of them are already free, without charges.'

The ordering of the new sentences will be the same as the clauses had in the original sentence.

4.2.2 Noun clauses

Noun clauses can be sub-classified as completive clauses and indirect questions. Our simplification proposal, in both cases, is the same: the indirect speech should be changed to direct speech. To that end, i) we will split the

clauses, ii) remove the complementisers and iii) we will add the phrase *honako hau* (this) in the main clause, to work as cataphora for the next sentence. iv) The sentence ordering will be main$_{orig}$-subordinate$_{orig}$ and v) the new sentences will be correctly punctuated. In some cases vi) pronouns should also be adapted. This proposal is illustrated in (2) and (3).

(2) a. *Eri, gaixorik naiz***ela** *esan genezake...*
ill-ABS, sick-ABS be-1SGABS.PRS.IND.COMP say-PRF aux-3SGABS.1PLERG.PST.POT...
'We could say **that** I am sick, ill.'

b. i. *Honako hau esan genezake:*
'We could say this:'

ii. *"Eri, gaixorik naiz."*
"'I am sick, ill."'

(3) a. *Fiskalak galdetu zidan* **ea** *desobedientzia zibila eraikuntza nazionalerako egiten genu***en**.
prosecutor-ERG ask-PRF aux-3SGERG.3SGABS.1SGDAT.PST.IND whether disobedience-Ø civil-ABS construction-Ø national-MOT do-IMPF aux-1PLERG.3SGABS.PST.IND.COMPL
'The prosecutor asked me **whether** we made civil disobedience for the national building.'

b. i. *Fiskalak honako hau galdetu zidan:*
'The prosecutor asked me this:'

ii. *"Desobedientzia zibila eraikuntza nazionalerako egiten zenuten?"*
"'Did you make civil disobedience for the national building?"'

The modal clauses with *-enez* + reporting verb and postpositional phrases that express thoughts or statements such as *-en arabera* (according to), *-en hitzetan* (in words of) and *-en adierazpenetan* (in declarations of) will be also simplified as noun clauses.

4.2.3 Relative clauses

In our corpora we have mainly found two types of relative clauses (common relatives and relatives with pronoun *zein*) and we have made a simplification

proposal for each one. In both cases we will split the clauses, the complementisers or the pronoun will be removed and the antecedent will be added in the new sentences. If the antecedent is a common noun, a demonstrative pronoun will be also added in the second sentence. If the antecedent is a named-entity no demonstrative pronoun will be added. We also must add the required case marker (the one that the antecedent had in the original sentence).

The new sentence ordering will be different according to the relative type. So, in common relatives (4) the ordering will be first subordinate and then main (subordinate$_{orig}$-main$_{orig}$) and in the relatives with pronoun *zein* (5) will be main$_{orig}$-subordinate$_{orig}$. These proposals are summed up in Table 4.1.

(4) a. *Konstituzioari eta Estatutuari eskaini*
constitution-DAT and by-law-DAT offer-PRF
*zaizkie***n** *ihardunaldietan*
aux-3SGABS.3PLDAT.PRS.IND.COMPL meetings-INE
Zuzenbide Zibila bazterrean geratzen
law-Ø civil-ABS corner-INE stay-IMPF
da.
aux-3SGABS.PRS.IND.
'The meetings **which** have been offered to the constitution and to the by-law are in the limit of the civil law.'

b. i. *Konstituzioari eta Estatutuari jardunaldiak eskaini zaizkie.*
'The meetings that have been offered to the constitution and to the by-law.'

ii. *Jardunaldi horietan Zuzenbide Zibila bazterrean geratzen da.*
'Those meetings are in the limit of the civil law.'

(5) a. *1873ko urriaren 6ko dekretu gehigarri batek,*
1873-ADN october-GEN 6-ADN decret-Ø additional-Ø one-ERG,
zeina *Errepublikako Gobernu-Presidente Emilio*
which-ABS republic-ADN government-president-Ø Emilio
Castelar-ek eta Estatuko Ministro, Jose de Carvajal-ek
Castelar-ERG and estate-ADN minister-Ø, Jose de Carvajal-ERG
izenpetu **bait***zuten,* *bi*
sign-PRF aux-COMPL.3SGABS.3PLERG.PST.IND, two

pentsio gehitzen zituen
pension-ABS add-IMPF aux-3PLABS.3PLERG.PST.IND
merituzko plaza banarentzat, arkitektura eta
merit-INS.ADN post-Ø one-JARRI, architecture-Ø and
sakongrabatuko alorretan.
rotogravure-ADN area-INE.
'An additional decret law of the 6th october, 1873, **which** was signed by the president of the government Emilio Castelar and the minister of the state Jose de Carvajal, added two pensions for a merit post, in the areas of architecture and rotogravure.'

b. i. *1873ko urriaren 6ko dekretu gehigarri batek bi pentsio gehitzen zituen merituzko plaza banarentzat, arkitekturako eta sakongrabatuko alorretan.*
'An additional decret law of the 6th october, 1873 added two pensions for a merit post, in the areas of architecture and rotogravure.'

ii. *Dekretu hori Errepublikako Gobernu-Presidente Emilio Castelarek eta Estatuko Ministro Jose de Carvajal-ek izenpetu zuten.*
'That decret was signed by the president of the government Emilio Castelar and the minister of the state Jose de Carvajal.'

If we find a word that is not written in standard Basque or if we find errors in the sentences, they will be corrected in the simplified sentences. That is why the lemma *ihardunaldi* has been corrected to *jardunaldi*.

Relative type	**Treatment of the antecedent**	**Sentence ordering**
Common relative	Antecedent + demonstrative	$\text{subordinate}_{orig}$-$\text{main}_{orig}$
Common relative (with named entity)	Antecedent	$\text{subordinate}_{orig}$-$\text{main}_{orig}$
Zein **relative**	Antecedent + demonstrative	main_{orig}-$\text{subordinate}_{orig}$
Zein **relative (with named entity)**	Antecedent	main_{orig}-$\text{subordinate}_{orig}$

Table 4.1 – Summary of simplification proposals of relative clauses

The simplification of relative clauses have also been presented in the Sec-

tions 6 and 7 of the paper *First Approach to Automatic Text Simplification in Basque* (Aranzabe et al., 2012a) and the Section 2 of the paper *Transforming Complex Sentences using Dependency Trees for Automatic Text Simplification in Basque* (Aranzabe et al., 2013).

4.2.4 Adverbial clauses

As the adverbial clauses are a diverse group and point out different relations we have decided to make a deeper analysis of them. So, we present a corpus study on the presence, frequency and position of the adverbial clauses together with the summary of the corpus analysis that leads us to make their simplification proposals.

For the adverbial clauses we have made two types of simplification proposals. In the first proposal we make structural changes in the syntax (syntactic simplification). In the second proposal, we do not perform any structural change on the syntax but substitute the structure with a most frequent one (syntactic substitution simplification).

Quantitative corpus analysis

The quantitative corpus analysis uses the information of list of structures[4] we have created from the descriptive grammar *Euskal Gramatika: Lehen Urratsak* (Euskaltzaindia, 1999, 2005, 2011) and the Basque Dependency Treebank (BDT) (Aranzabe, 2008). The list of structures can be consulted in Appendix A.

To analyse the presence of the structures, we have looked up the list of the structures in the BDT and in Table 4.2 we present the percentage of structures listed by the Academia that is found in the corpus. We are interested in finding out how many are present in the corpus, since some structures in the list are dialectal variants and diachronic. The percentage shows the number of the structures in the academic list we found in the BDT corpus, that is, e.g. 40.00 % of the finite temporal structures in the academic list were found in the corpus while all of them were found in the case of finite

[4]Clause type denotes for us the semantic classification of clauses, that is, temporal clauses, purpose clauses etc. Structures are the different realisations of those clause types. For example, clauses built with 'when', 'after' and 'before' are structures that belong to the temporal clause type.

purpose, concessive and consecutive clauses[5].

Clause type	Finiteness	Percentage
Temporal	**finite**	40.00
	non-finite	60.00
Causal	**finite**	80.00
	non-finite	50.00
Concessive	**finite**	100.00
	non-finite	66.67
Modal	**finite**	63.64
	non-finite	63.33
Purpose	**finite**	100.00
	non-finite	33.33
Consecutive	**finite**	100.00
Conditional	**finite**	50.00
	non-finite	55.56

Table 4.2 – Percentage of the presence in corpus

As we can see, finite temporal clauses (40.00 %) and non-finite purpose clauses (33.33 %) are the clause types that show less variety in the corpus, that is, round the 35 % of the structures listed by the Academia are found in the corpus. The presence of rest of the types ranges between 50 % and 80 %. With this analysis we got an overview of the corpus. We have seen which the percentage of the structures that are found in the corpus is, and therefore, used nowadays and also its limits.

Once we know which structures are found in the corpus, we want to know their frequency of use. We will see in Table 4.3 which is the frequency of each clause type. In the third and the fourth columns we present the percentage of each type in relation to all the types and in brackets we present the distribution of each type.

As we see in the second column of Table 4.3, the most used clause type is modal (29.95 %) followed by purpose clauses (22.37 %). Frequencies for temporal (17.55 %) and causal clauses (17.10 %) are quite similar. Conditional (6.99 %) and concessive (5.76 %) are less frequent but similar too.

Taking into account the frequency of the finite and non-finite clauses (third and forth column), the most used clause type is the non-finite modal (24.09 %) followed by non-finite purpose clauses (21.05 %). On the other

[5]We have to point out that there is only one structure for finite purpose clauses and that is why they get 100 %.

Clause type	All	Finite	Non-finite
Temporal	17.55	9.01 (51.34)	8.54 (48.66)
Causal	17.10	16.61 (97.11)	0.49 (2.89)
Concessive	5.76	3.01 (52.24)	2.75 (47.76)
Modal	29.95	5.86 (19.56)	24.09 (80.44)
Purpose	22.37	1.32 (5.89)	21.05 (94.11)
Consecutive	0.28	0.28 (100.00)	-
Conditional	6.99	5.86 (83.84)	1.13 (16.16)

Table 4.3 – Frequency of use the clause types

hand, finite purpose (1.32 %), non-finite causal (0.49 %) and finite consecutive clauses (0.28 %) are the less used. Except for the finite causal clauses (16.61 %) the rest do not achieve 10 % of frequency.

Looking at the distribution of finite and non-finite clauses (third and forth column, data in brackets), the results show that temporal (finite 51.34 % and non-finite 48.66 %) and concessive clauses (finite 52.24 % and non-finite 47.76 %) have similar distribution. Causal (97.11 %) and conditional (83.84 %) clauses tend to be used as finite clauses while modal (80.44 %) and purpose (94.11 %) have a tendency to non-finite. In Basque there are not non-finite consecutive clauses. This analysis has also been made at structure level and it is presented in the technical report *Perpaus adberbialen agerpena, maiztasuna eta kokapena EPEC-DEP corpusean [Presence, frequency and position of adverbial clauses in the BDT]*.

Word ordering in Basque has been studied over all at phrase level from both corpus (Hidalgo, 1999; Aldezabal et al., 2003) and neurolinguistic perspective (Erdozia, 2006). Here we will present the position adverbial clauses take in the sentence in reference to the main verb.

In Table 4.4 we show the results of the finite and non-finite adverbial clause types. Except for the modal clauses where finite tend to be preposed and non-finite postposed, all the clause types keep the same tendency in both finite and non-finite clauses. The clause types that tend to be preposed are temporal, concessive and conditionals while causal, purpose and consecutive clauses tend to be postposed.

Some tendencies we found here quantitatively were already pointed out by Zabala (2000). She states that we usually give concession and condition before the consequence and that temporal clauses tend to precede the main. She adds as well that explanations tend to appear after the main clause.

It is remarkable that causal clauses do not follow the logical and chrono-

Type	Finiteness	Preposed	Postposed
Temporal	**finite**	67.27	32.73
	non-finite	74,19	25,81
Causal	**finite**	29.49	70.51
	non-finite	37,88	62,12
Concessive	**finite**	75.00	25.00
	non-finite	64,75	35,25
Modal	**finite**	63.29	36.71
	non-finite	33,63	66,37
Purpose	**finite**	14.29	85.71
	non-finite	41,74	58,26
Consecutive	**finite**	14.29	85.71
Conditional	**finite**	81.12	18.88
	non-finite	84,91	15,09

Table 4.4 – Position of finite and non-finite adverbial clauses

logical order of cause (subordinate) and effect (main). The logical order, however, is used in consecutive clauses and conditional clauses. In order to study if the chronological order is fulfilled, we need to analyse the temporal clauses according to their subgroup (anteriority, posteriority, simultaneity, delimitation, impendency and duration). That is why we have also analysed the structures in each clause type (results only in the Basque report).

For the adverbial clauses we have made two types of simplification proposals. In the first proposal we make structural changes in the syntax (syntactic simplification). In the second proposal, we do not perform any structural change on the syntax but substitute the structure with a most frequent one (syntactic substitution simplification).

Syntactic simplification

The simplification proposals that involve structural changes are different according to the adverbial type and sub-type but in general we perform the following steps: i) split the sentences into clauses; ii) remove case markers and relational suffixes; iii) add an adverb or a phrase that is going to keep the relation (added element and alternative added elements); iv) order the sentences in text and v) correct errors and standardise, if needed. In (6) we show a sentence that contains a temporal clause and its simplified sentences.

(6) a. *Edurnezuri printzearekin ezkondu*
Snow_White prince-COM marry-PRF
*zen***ean**, *zazpi ipotxek*
aux-3SGABS.PST.IND.COMP.INE, seven-Ø dwarfs-ERG
edateari eman zioten.
drink-VEN.DAT give-PRF aux-3PLERG.3SGDAT.3SGABS.PST.IND

'**When** Snow White married the prince, the seven dwarfs started to drink.'

b. i. *Edurnezuri printzearekin ezkondu zen.*
'Snow White married the prince.'

ii. *Orduan zazpi ipotxek edateari eman zioten.*
'Then, the seven dwarfs started to drink.'

As we mentioned the added elements are the adverbs or noun phrases that stick to the meaning. The alternative added elements are the elements that will be added in the text if the standard added element has already been frequently used in the text. To decide which added elements should be used we have looked them up in the lemma frequency list we have compiled from the corpus *Lexikoaren Behatokia* and we have chosen the most frequents.

To reorder the sentences in the text we have used the information of the quantitative corpus study. In most of the cases we have respected the clause ordering we have seen them but in other cases we have decided to use the chronological or logical ordering as pointed in other simplification studies (Specia et al., 2008; Klerke and Søgaard, 2012).

In Table 4.5 we present the added elements and alternative added elements and the reordering of temporal clauses.

Sub-type	**Structure**	**Added element**	**Alternative added element**	**Sentence ordering**
Simultaneity in general	*-enean; -ela(rik); noiz eta ... bait- /-en* *-tzean; -tzerakoan; -tzekoan; -tzearekin; -tzeari/-tzerat; -tu(k)eran*	Orduan	Une hartan; Aldi berean	subordinate$_{orig}$-main$_{orig}$

(Continued on next page)

Sub-type	**Structure**	**Added element**	**Alternative added element**	**Sentence ordering**
Repeated simultaneity	*-enetan; -en bakoitzean; -en guztietan; -en aldikal/aldiro; zenbat aldiz -en ... hainbat aldiz, -tu aldiro; -tu bakoitzean; -tu guztian; -tu ahala/arau*	Une horietan guztietan	Aldiro	$\text{subordinate}_{orig}$-$\text{main}_{orig}$
Narrow simultaneity	*-eneko; -en orduko* *-tzerako; -tu orduko* *-en bezain laster/sarri/agudo/fite; -en ber; -enaz batera* *-tu bezain laster/pronto; -tu eta berehala; -tu eta laster; -tuaz/-tzearekin bat(era); -tu berri(t)an; -tu ahala/arau*	Orduko Une horretan bertan	Segidan Orduko; Segidan	$\text{subordinate}_{orig}$-$\text{main}_{orig}$
Anteriority	*-en baino lehen; -en aurrean/aitzinean* *-tu baino lehen; -tu aurretik/ -tu aitzinean; -tu gabe; -tu orduko*; -tzerako**	Gero	Ondoren; Ostean	main_{orig}-$\text{subordinate}_{orig}$
Posteriority	*-en ondoan; -en ondoren; -en ostean* *-tu eta; -tu eta gero; -tuta; -tu ondoan; -tu ostean; -tu(a)z; -tuz gero; -turik*	Ondoren	Ostean	$\text{subordinate}_{orig}$-$\text{main}_{orig}$
Front bound of the duration	*-enetik; -enez gero; -enik ...-ra* *-tuz gero**	Ordutik	Une hartatik; Harrezkero	$\text{subordinate}_{orig}$-$\text{main}_{orig}$
Last bound of the duration	*-en arte* *-tu arte; -tu artean; -tu bitartean; -tzeraino*	Ordura arte	Orduraino	$\text{subordinate}_{orig}$-$\text{main}_{orig}$
The whole duration	*-eno/-eino; -en bitartean; -en artean; -en arteko*	Bitartean	Artean	$\text{subordinate}_{orig}$-$\text{main}_{orig}$

(Continued on next page)

Sub-type	Structure	Added element	Alternative added element	Sentence ordering
	-tu bitarte(an); -tu artean**			

Table 4.5 – Added elements and ordering of the temporal clauses

The added elements of the causal clauses according to their sub-types and sentence reordering of the simplified clauses is presented in Table 4.6. In this case, contrary to the data found in the corpus analysis, we have decided to follow the chronological and logical ordering in the pure causal clauses. On the other hand, we keep the ordering found in the corpora for the explicative, as explanations tend to be after the main fact (Zabala, 2000).

Sub-type	Structure	Added element	Alternative added elements	Sentence ordering
Pure causal	*-elako/elakoz/-lakotz; -elakoan; bait-; zeren eta ... bait-/-(e)n -tzeagatik; -tzearren*	Horregatik	Hori dela eta	$\text{subordinate}_{orig}$-$\text{main}_{orig}$
Explicative	*Bait-; ... eta; zeren eta ... [(bait-/-(e)n)]*	Izan ere	-	main_{orig}-$\text{subordinate}_{orig}$

Table 4.6 – Added elements and sentence reordering of the causal clauses

The added elements of the conditional clauses are presented in Table 4.7. We have also made a sub-classification according to their type.

Sub-type	Added element in the subordinate	Alternative added element in the subordinate	Added element in the main clause
Real (present)	Demagun	Eman dezagun	Kasu horretan
Real (past)	Demagun	Eman dezagun	
Irreal	(Polarity change)	-	Bestela

Table 4.7 – Added elements of the conditional clauses

To simplify the modal clauses, we propose also to add other elements like modal verbs, *-ten* aspect or the negative adverbs. Those special added

elements are presented in Table 4.8.

Structure	Special added elements
-tu nahirik; -tu ezinik; -tu beharrean/beharrez egon	nahi izan; ezin izan; behar izan
-tu gabe/barik/ezta;-tzeke; -tu ordez/ordean;-tu beharrean	ez (negate the clause)
-tu aginean/aginik, -tu hurran; -tzeko zorian	ia
-tu ahala/arau	*-ten* aspect

Table 4.8 – Special added elements of the modal clauses

When simplifying the consecutive clauses, the quantifier of the original sentence should be changed. In Table 4.9 we present their substitutions for the simplified sentences.

Quantifier in the original sentence	Quantifier in the simplified sentence
hain	oso
hainbeste	asko
hainbat	hainbat
hala	hala, honela
halako maneran	hala, honela
halako modez	hala, honela
halako x	halako x

Table 4.9 – Quantifiers in the original and simplified consecutive clauses

To sum up, we present the added elements and the alternative added elements for the adverbial clauses and their reordering in Table 4.10. In the cases where sub-types had been made, the reference to that table is made.

Clause type	Added element	Alternative added elements	Sentence ordering
Temporal	Table 4.5	Table 4.5	Table 4.5
Causal	Table 4.6	Table 4.6	Table 4.6
Concessive	Hala ere	Nolanahi ere; Edonola ere; Hala eta guztiz ere	$\text{subordinate}_{orig}$-$\text{main}_{orig}$
Modal	Hala	Honela; Horrela; Modu horretan; Era berean; Era horretan	$\text{subordinate}_{orig}$-$\text{main}_{orig}$
Consecutive	Ondorioz	Beraz; Hortaz; Honenbestez	main_{orig}-$\text{subordinate}_{orig}$
Purpose	nahi izan	gura izan	main_{orig}-$\text{subordinate}_{orig}$

(Continued on next page)

Clause type	Added element	Alternative added elements	Sentence ordering
Conditional	Table 4.7	Table 4.7	subordinate$_{orig}$-main$_{orig}$

Table 4.10 – Added elements, alternative added elements and sentences reordering of the adverbial clauses

More information about the simplification proposals for the Basque adverbial clauses can be found in the Sections 6 and 7 of the paper *First Approach to Automatic Text Simplification in Basque* (Aranzabe et al., 2012a) and the Section 2 of the paper *Transforming Complex Sentences using Dependency Trees for Automatic Text Simplification in Basque* (Aranzabe et al., 2013).

Syntactic substitution simplification

The following proposals are based on the frequencies got in the quantitative analysis (Section 4.2.4). Taking into account the meaning equivalences, we propose to substitute the the less frequent structures with the most frequent one, without altering the syntax of the sentence. In (7) we show an example of this simplification type.

(7) a. *Abuztuaren amaieran beste goi bilera bat*
August-GEN end-INE other-Ø high-Ø meeting-Ø one-ABS
egitea aztertzen ari dira Israel
do-VEN.ABS study-IMPF be-doing-3SGBS-PRS-IND Israel-ABS
eta PAN Palestinako Aginte Nazionala, Ekialde
and PNA-Ø Palestine-ADN Authority-Ø National-ABS, East-Ø
Erdiko bake prozesua suspertzearren.
Middle-ADN peace-Ø proccess-ABS promote-VEN.RM
'Israel and the PNA, the Palestinian National Authority, are studying to organise another summit at the end of August **in order to** promote the peace process in the Middle East.'

b. i. *Abuztuaren amaieran beste goi bilera bat egitea aztertzen ari dira Israel eta PAN Palestinako Aginte Nazionala, Ekialde Erdiko bake prozesua suspertzeko*.
'Israel and the PNA, the Palestinian National Authority, are studying to organise another summit at the end of August **in order to** promote the peace process in the Middle East.'

In Table 4.11 we present the substitution candidates for the less frequent structures. This simplification proposal is detailed in the paper *Simplifying Basque Texts: the Shallow Syntactic Substitution Simplification* (Gonzalez-Dios et al., 2015a).

Adverbial type	**Less frequent structures**	**Substitution candidate**
Temporal	*-tzeari; -tzera(t); -tu(k)eran*	-tzean
	-tu arau	-tu ahala
	-tu baino lehen; -tu gabe; -tu aintzinean	-tu aurretik
	-tu eta; -tu eta gero; -tu ondoan; -tuz gero; -tuaz gero; -tu ostean; -tu ondotik	-tu ondoren
	-tu artean	-tu arte
Causal	*-tzearren*	-tzegatik
Concessive	*-tuagatik; -tuz gero ere, -tuta (gabe/ezta) ere; -tzearren; -ik ere*	-tu arren
Modal	*-tu arau*	-tu ahala
	-tu gurarik	-tu nahian
	-tu barik; -tu ezta	-tu gabe
	-tu ordean	-tu ordez
	-tzeko gisan; -tzeko maneran	-tzeko moduan
	-tu hurran; -tu aginean; -tu aginik;	-tzeko zorian
Purpose	*-tzekotzat; -tzeagatik; -tzearren; -tze alde(ra)*	-tzeko
	-tzekotan	-tzeko asmotan
Conditional	*-tuenenean; -tzera(t)*	-tuz gero
	-tu ezin	-tu ezean
	-tzez gero; -tzekoz	-tzekotan

Table 4.11 – Substitution options for the less frequent structures

4.2.5 Apposition and parenthetical structures

Apposition is a phenomenon that increases the length of the sentences and it has been reported in the context of ATS as a complex phenomenon. There are two types of apposition in Basque: apposition that occurs inside a noun phrase (8a) and a noun phrase as appositive (9a).

To simplify the apposition, i) we will split the appositives, ii) we will remove the appositive that expresses the explanation from the main clause, and iii) we will create new copulative sentences with the appositives. iv) The sentence ordering will be main and the the sentences created out of the appositives. Examples of this proposal can be seen in (8) (apposition that

occurs inside a noun phrase) and in (9) (a noun phrase as appositive).

(8) a. **Jasser Arafat buru palestinarra Egiptoko**
Jasser-Ø Arafat-Ø head-Ø palestinian-ABS Egypt-ADN
presidente Hosni Mubarak-ekin *bildu*
president-Ø Hosni-Ø Mubarak-COM meet-PRF
zen *atzo* *Kairon.*
aux-3SGABS.PST.IND yesterday Cairo-INE.
'**Palestinian Chairman Jasser Arafat** met **President of Egypt Hosni Mubarak** yesterday in Cairo.'

b. i. *Jasser Arafat Hosni Mubarak-ekin bildu zen atzo Kairon.*
'Jasser Arafat met Hosni Mubarak yesterday in Cairo.'

ii. *Jasser Arafat buru palestinarra da.*
'Jasser Arafat is a Palestinian chairman.'

iii. *Hosni Mubarak Egiptoko presidentea da.*
'The president of Egypt is Hosni Mubarak.'

(9) a. *Aitor Mendiluzek,* **hogei urteko andoaindar**
Aitor Mendiluze-ERG, twenty-Ø year-ADN andoaindar-Ø
bertsolariak, *irabazi* *zuen*
bertsolari-ERG win-PRF aux-3SGERG.3SGABS.PST.IND
Gipuzkoako *txapelketa.*
Gipuzkoa-ADN championship-ABS.
'Aitor Mendiluze, **a twenty-year-old** *bertsolari*[6] **from Andoain,** won the championship of Gipuzkoa.'

b. i. *Aitor Mendiluzek irabazi zuen Gipuzkoako txapelketa.*
'Aitor Mendiluze won the championship of Gipuzkoa.'

ii. *Aitor Mendiluze hogei urteko andoaindar bertsolaria da.*
'Aitor Mendiluze is a twenty-year-old *bertsolari* from Andoain.'

Parentheticals are expressions, somehow structurally independent, that integrated in a text function as modifiers of phrases, sentences..., and add information or comments to the text. We have mainly analysed parenthetical structures that contain biographical information (10).

[6] A *bertsolari* is a singer that improvises musical verses.

To simplify the parentheticals containing biographical information, i) we will split the parenthetical and ii) split the elements inside. Then, iii) we will write first the main sentence (10b-i), and iv) we will build the following sentences with the information found in the parenthetical. If the person is dead (10), we will find birth data (town, state and date) and death data (town, state and date). If the person is alive, we will only find birth data. In the new sentences we will make explicit the kind of information, so we will add the verbs *jaio zen* (was born) or *hil zen* (died). v) The sentence ordering will be first the main sentence followed by a new sentence with the information about the birth. If the birthplace is composed by more than a place entity, we will add the place specifications. After the birth information, a sentence will contain the information about the death, and, if appear, the place specifications.

(10) a. *Ernest Rutherford, Nelsongo lehenengo baroia,*
Ernest Rutherford-ABS, Nelson-ADN first Baron-ABS,
(Brightwater, Zeelanda Berria, 1871ko abuztuaren
(Brightwater, Zeeland New, 1871-ADN August-GEN
30a - Cambridge, Ingalaterra, 1937ko urriaren
30-ABS - Cambridge, England, 1937-ADN October-GEN
19a) *fisika nuklearraren aita izan*
19)-ABS Physics nuclear-GEN father-ABS be-PRF
zen.
aux-3SGABS.PST.IND

'Ernest Rutherford, 1st Baron of Nelson, **(Brightwater, New Zeeland, 30th August, 1871 - Cambridge, England, 19th October, 1937)** was the father of the nuclear Physics.'

b. i. *Ernest Rutherford, Nelsongo lehenengo baroia, fisika nuklearraren aita izan zen.*
'Ernest Rutherford, 1st Baron of Nelson, was the father of the nuclear Physics.'

ii. *Ernest Rutherford 1871ko abuztuaren 30ean Brightwateren jaio zen.*
'Ernest Rutherford was born on the 30th of August, 1871 in Brightwater.'

iii. *Brightwater Zeelanda Berrian dago.*
'Brightwater is in New Zeeland.'

iv. *Ernest Rutherford 1937ko urriaren 19an Cambridgen hil zen.*
'Ernest Rutherford died on the 19th of October, 1937 in Cambridge.'

v. *Cambridge Ingalaterran dago.*
'Cambridge is in England.'

The proposals of the appositions and parenthetical structures are also presented in the Section 6 of the paper *First Approach to Automatic Text Simplification in Basque* (Aranzabe et al., 2012a), in the Section 6 of the paper *Detecting Apposition for Text Simplification in Basque* (Gonzalez-Dios et al., 2013a) and in the Sections 2 and 3 of the paper *Making Biographical Data in Wikipedia Readable: A Pattern-based Multilingual Approach* (Gonzalez-Dios et al., 2014a).

4.3 Summary

In this chapter we have presented the summary of the simplification proposals for the phenomena we considered to be complex. These phenomena are coordinate clauses, subordinate clauses, apposition parenthetical structures and postpositional phrases that express thoughts or statements. We have also proposed two simplification types for adverbial clauses.

Glossary

1 first person
3 third person
ABS absolutive case marker
ADN adnominaliser (relational suffix *-ko*) (local genitive) case marker
AUX auxiliary verb
COM comitative case marker
COMPL complementiser
DAT dative case marker
ERG ergative case marker
GEN genitive case marker
IND indicative mood
INE inessive (locative) case marker
INS instrumental case marker
IMPF imperfective aspect
MOT motivative case marker
PRF perfective aspect (perfect participle)
PL plural
PART partitive case marker
POT potential mood
PRS present tense
PST past tense
RM relation mark
SG singular
VEN verbal noun

5 Framework for the ATS in Basque

In this chapter we present the framework and our approach to Automatic Text Simplification in Basque. The system we are proposing, *EuTS*, is a linguistic knowledge, rule based system, that performs simplification at syntactic level. We have decided to build a rule based system since there are no data to train a data-based system for Basque and we have also decided to concentrate on syntactic simplification since the fundational systems in ATS used to treat that kind of simplification mainly. Exactly, *EuTS* will perform two types of simplification: syntactic substitution simplification and syntactic simplification. In the following sections we explain the main points of the system. We also describe the tools that perform the automatic analysis of the texts.

5.1 Simplification decisions

In order to know which texts should be simplified and at which level should be simplified, we apply the simplification decision algorithm (Figure 5.1). This algorithm takes into account the complexity of the texts and the level of the target audience to decide which kind of simplification should be performed and at which level the texts should be simplified.

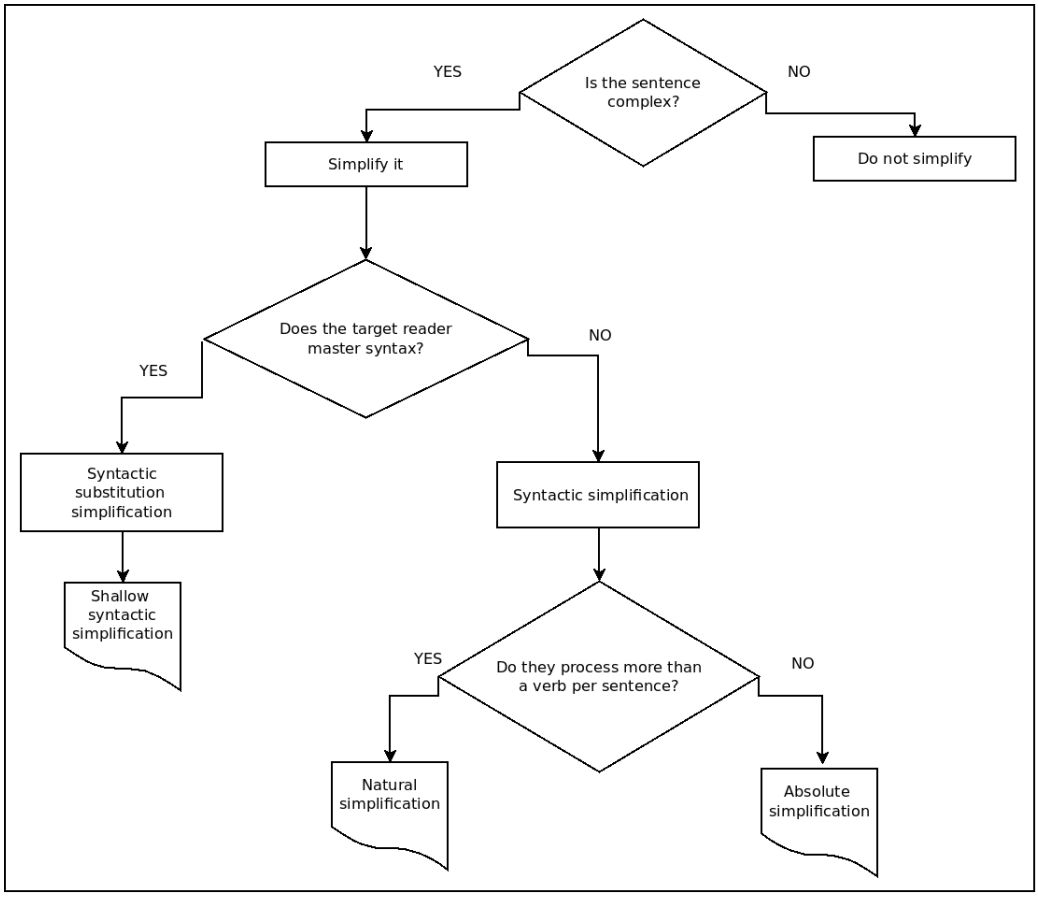

Figure 5.1 – Algorithm of simplification decisions

5.2 Simplification levels

To adapt the simplification of the texts we have defined three simplification levels:

1. **Shallow Syntactic Simplification (SSS):** the frequency based simplification of adverbial non-finite structures is performed. This level is intended for people who have a good level of Basque and master Basque syntax but do not know unusual, dialectal and synchronic variations. That is, at this level the depth of the syntactical structure is kept but the structure that is used is more frequent. These people are usually

advanced learners or non-fluent speakers.

2. **Natural Simplification (NS):** compound and complex sentences with finite verb simplification will follow the syntactic simplification process for Basque (Section 5.3.1) together with the syntactic substitution simplification. In this level the depth of the syntactic tree of the sentences is altered. That is, structural changes are performed in the sentence. The target of this level is people who are learning Basque but get stuck with long sentences and do not master syntax. Advanced NLP applications can benefit from this level and get better results with shorter sentences.

3. **Absolute Simplification (AS):** everything is simplified. Both sentences with finite and non-finite verbs will follow the syntactic simplification process. Syntactic substitution simplification will be also applied. The structure of the sentences is also altered as in the previous level. This level will be useful for people with low knowledge of Basque syntax or advanced NLP applications that get better results by processing only one verb per sentence.

The simplification levels of Basque are presented in the Section 2 of the paper *Simplifying Basque Texts: the Shallow Syntactic Substitution Simplification* (Gonzalez-Dios et al., 2015a). In (11) we present an example of a sentence simplified at different levels.

(11) a. *1991 eta 1993an errepideko Munduko Txapelketak eta bi Tour irabaz***i ondoren**, *mendian gora aise igo***tzearren** *pisua galtzen hasi zen,* **eta** *1994. urtean 48 kiloko infernura jaitsi zen, anorexiara.*
After having won in 1991 and 1993 the UCI Road World Championships and two Tours, **in order to** comfortably **climb up** the mountains, he began to lose weight, **and** in 1994 he descended to the hell of 48 kg, to the anorexia.

b. i. **Shallow syntactic simplification:** *1991 eta 1993an errepideko Munduko Txapelketak eta bi Tour irabaz***i ondoren**[1], *mendian gora aise igo***tzeko** *pisua galtzen hasi zen, eta 1994. urtean 48 kiloko infernura jaitsi zen, anorexiara.*

[1]This structure is not changed because *-tu ondoren* is the most frequent.

After having won in 1991 and 1993 the UCI Road World Championships and two Tours, **in order to**[2] comfortably **climb up** the mountains, he began to lose weight, **and** in 1994 he descended to the hell of 48 kg, to the anorexia.

ii. **Natural simplification:** *1991 eta 1993an errepideko Munduko Txapelketak eta bi Tour irabaz***i ondoren**, *mendian gora aise igo***tzeko** *pisua galtzen hasi zen. 1994. urtean 48 kiloko infernura jaitsi zen, anorexiara.*
After having won in 1991 and 1993 the UCI Road World Championships and two Tours, **in order to** comfortably **climb up** the mountains, he began to lose weight. In 1994 he descended to the hell of 48 kg, to the anorexia.

iii. **Absolute simplification:** *1991 eta 1993an errepideko Munduko Txapelketak eta bi Tour irabazi zituen.* **Ondoren**, *pisua galtzen hasi zen. Mendian gora aise igo* **nahi zuen**. *1994. urtean 48 kiloko infernura jaitsi zen, anorexiara.*
He won in 1991 and 1993 the UCI Road World Championships and two Tours. After, he began to lose weight. He wanted to climb up the mountains comfortably. In 1994 he descended to the hell of 48 kg, to the anorexia.

5.3 Syntactic simplification

In this section we present the decisions taken to carry out the syntactic simplification.

5.3.1 Syntactic simplification process

The syntactic simplification process is the we have defined to carry out the syntactic simplification in Basque. This process based on our corpus analysis and work to automate the simplification proposals where structural changes occur.

1. **Splitting:** make as many new sentences as clauses out of the original.
2. **Reconstruction:** two operations take place in the split sentences:

[2]Notice that in Basque the structure has been changed.

(a) Removing no longer needed morphological features such as complementisers and suffixes (Relation_Marks_List). Being Basque an agglutinative language we have to remove parts of words and not a whole word.

(b) Adding new elements like adverbs or paraphrases (Added_Elements_List). The goal is to maintain the meaning. In other words, the features that have been deleted should be replaced by new words. In that list of lists the added elements are compiled. The alternative added elements appear from the second position of each list on.

3. **Reordering:** reorder the elements in the new sentences, and ordering the sentences in the text (Reordering_List).

4. **Adequation and Correction:** correct the possible grammar and spelling mistakes, and fix punctuation and capitalisation.

This process is explained in the Section 4 of the paper *First Approach to Automatic Text Simplification in Basque* (Aranzabe et al., 2012a) and in the Section 3 of the paper *Transforming Complex Sentences using Dependency Trees for Automatic Text Simplification in Basque* (Aranzabe et al., 2013).

5.3.2 Other decisions

We also have taken three other decisions:

- Minimum clause length: the minimum clause length of a candidate clause should be verb plus two arguments or adjunts. With this constraint we want to avoid too short sentences that can sound unnatural or even make more difficult the continuity of the text.

- Alternative added elements: the alternative added elements will be used we the default added element has been used in the sentence we are working on, or in the two previous or next ones. With this decision we want to avoid the monotony of the text.

- Rule application ordering: rules will be applied top-down in the dependency tree. If two phenomena are at the same level, they can be simplified at the same time. The rules for noun-clauses can be applied

only once per sentence or coordinative clause and the rules for adverbial clauses cannot be applied more than twice. In (12) we show the example of a sentence simplified top-down clause by clause.

(12) a. *Haize handia ibili* **eta** *euria egiten zue***n arren**, *hegazkinari bidaia hasteko baimena* **eman zitzaionean** *eguraldia hegan egiteko modukoa ze***la** *azpimarratu zuten Taiwango Babes Zibileko agintariek,* **eta** *Boeing 747a* **bota zuen** *haize bolada bereziki indartsu hura* **ezin zela aurreikusi** *(azpimarratu zuten).*
Although the wind was strong and it rained, when the plane was given permission to start the journey, the authorities of civil Protection of Taiwan emphasised that the weather was adequate to fly and (they emphasised) that it was not possible to foresee that that strong squall that knocked over the plane.

b. i. *Honako hau azpimarratu zuten Taiwango Babes Zibileko agintariek:*
ii. *"Haize handia ibili zen.*
iii. *Euria egiten zuen.*
iv. *Hala ere, hegazkinari bidaia hasteko baimena eman zitzaion.*
v. *Orduan eguraldia hegan egiteko modukoa zen."*
vi. *Honako hau ere azpimarratu zuten:*
vii. *"Boeing 747a bota zuen haize boladak.*
viii. *Haize bolada bereziki indartsu hura ezin zen aurreikusi."*
The authorities of civil Protection of Taiwan emphasised this: "The wind was strong. It rained. However, the plane was given permission to start the journey. Then, the weather was adequate to fly." They also emphasised this: "The strong squall that knocked over the plane. It was not possible to foresee that strong squall."

In the future, we plan add to the system the coreference resolution, the treatment of the ellipsis and links to Wikipedia. We also are working on

lexical simplification.

5.4 Automatic text analysis

The automatic analysis is the preprocess required before we analyse the complexity of the texts and before we simplify them. Based on the lexical database EDBL-LBDBL (Aldezabal et al., 2001), the tools we have used to analyse the texts automatically are:

- Analysis chain[3]:
  - Morpho-syntactic analysis by *Morfeus* (Aduriz et al., 1998): tokenisation, segmentation or morphological analysis, morpho-syntax (Aduriz et al., 2000; Gojenola, 2000), treatment of multiword units (Alegria et al., 2004; Urizar, 2012).
  - Lemmatisation/annotation by *Eustagger* (Ezeiza, 2002; Aduriz, 2000; Aduriz and Díaz de Ilarraza, 2003; Aduriz et al., 1997)
  - Chunking by *Ixati*: NERC by *EIHERA* (Alegria et al., 2003; Fernandez Gonzalez, 2012), postposition recognition, syntactic function disambiguation (Aduriz, 2000), phrasal and verbal chunks (Aranzabe, 2008; Arrieta, 2010)
  - Dependency parsing by *EDGK* (Aranzabe, 2008), *Maltparser-Maltixa* (Bengoetxea and Gojenola, 2007; Bengoetxea, 2014) and the hybrid analyser (Aranzabe et al., 2012b)
- Tools improved or developed during this project:
  - Clause boundary detection by *Mugak* (Ondarra, 2003; Aduriz et al., 2006b; Arrieta, 2010; Aranzabe et al., 2013).
  - Aposition detection by *Aposizioak* (Gonzalez-Dios et al., 2013a).

All these tools are also useful for the implementation of the syntactic simplification rules. To remove the morphemes, we will use the analysis provided by *Morfeus*. *Eustagger* will tell us which the sentence and clause type is. Moreover, it will give us the disambiguated verb information (finite or

[3]The analysis chain is the pipeline of tools and resources developed in the Ixa group.

non-finite verb, tense, aspect and person). *Ixati* will give us the information about the named entities necessary for the apposition detection, the postposition boundaries to split the treated pospositional structures and the chunks to calculate the minimum length. The dependency parsing will give us the information about needed to apply the rules. *Mugak* will tell us where clauses are should be split and *Aposizioak* will give us the information about the appositives also to determine where they should be split. In Figure 5.2 we present the output of these tools.

```
"<Asperren>"<HAS_MAI>"   S:137/0
        "asper" IZE ARR GEN NUMP MUGM ZERO HAS_MAI w1,L-A-IZE-ARR-10,lsfi2 @IZLG> %SIH S:137 %ESALDI_HAS_1_BEREZIA &NCMOD>
"<kasua>"
        "kasu" IZE ARR BIZ- ABS NUMS MUGM w2,L-A-IZE-ARR-14,lsfi3 @SUBJ %SIB &NCSUBJ>
"<emeki-emeki>"
        EZEZAG "emeki-emeki" ADB ARR ZERO w3,L-G-ADB-ARR-2,lsfi6 @ADLG %SINT &NCMOD>
"<aitzinatu>"
        "aitzinatu" ADI SIN PART BURU NOTDEK w4,L-A-ADI-SIN-8,lsfi7 @-JADNAG %ADIKATHAS &MENOS>
"<bada>"
        "izan" ADL BALD A1 NOR NR_HURA w5,L-A-ADL-3,lsfi8 @+JADLAG_MP_ADLG %ADIKATBU &MENOS>
"<ere>"
        "ere" LOT LOK EMEN w6,L-A-LOT-LOK-6,lsfi11 @LOK &CMOD>
"<,>"<PUNT_KOMA>"  S:608/0
        PUNT_KOMA S:608 }MUGA
"<Sa>"<HAS_MAI>"
        EZEZAG "Sa" IZE IZB PLU- ENTI_HAS_PER AORG HAS_MAI w8,L-G-IZE-IZB-3,lsfi12 @KM> %SIH &NCMOD>
"<Pintoren>"<HAS_MAI>"
        "pinto" ADJ ARR IZAUR- GEN MG ENTI_BUK_PER HAS_MAI w9,L-A-ADJ-ARR-10,lsfi14 @IZLG> &NCMOD>
"<etorkizuna>"
        "etorkizun" IZE ARR BIZ- ABS NUMS MUGM w10,L-A-IZE-ARR-18,lsfi15 @SUBJ> %SIB &NCSUBJ>
"<fite>"
        "fite" ADB ARR ZERO w11,L-A-ADB-ARR-3,lsfi18 @ADLG %SINT &NCMOD>
"<argituko>"
        "argitu" ADI SIN PART GERO NOTDEK w12,L-A-ADI-SIN-10,lsfi19 @-JADNAG %ADIKATHAS
"<da>"
        "izan" ADL A1 NOR NR_HURA w13,L-A-ADL-5,lsfi20 @+JADLAG %ADIKATBU &<AUXMOD
"<$.>"<PUNT_PUNT>"  S:123/0  S:148/0
       PUNT_PUNT S:123 %ESALDI_BUK_1 S:148 }MUGA
```

Figure 5.2 – An automatic analysis of a sentence

Mugak and *aposizioak* have been specially improved and developed during this thesis since we think they are necessary for simplification purposes. *Mugak* performs the clause splitting and it is presented in the Sections 4 and 5 of the paper *Transforming Complex Sentences using Dependency Trees for Automatic Text Simplification in Basque* (Aranzabe et al., 2013). The apposition detector detects and splits the appositives and it is presented in the paper *Detecting Apposition for Text Simplification in Basque* (Gonzalez-Dios et al., 2013a).

5.5 Summary

In this chapter we have presented our approach and the framework for the ATS of Basque texts. We also have mentioned the tools which will perform

the automatic text analysis. In Figure 5.3 we have added to the contributions the basic tools (*Mugak* and *Aposizioak*) we have introduced in this chapter.

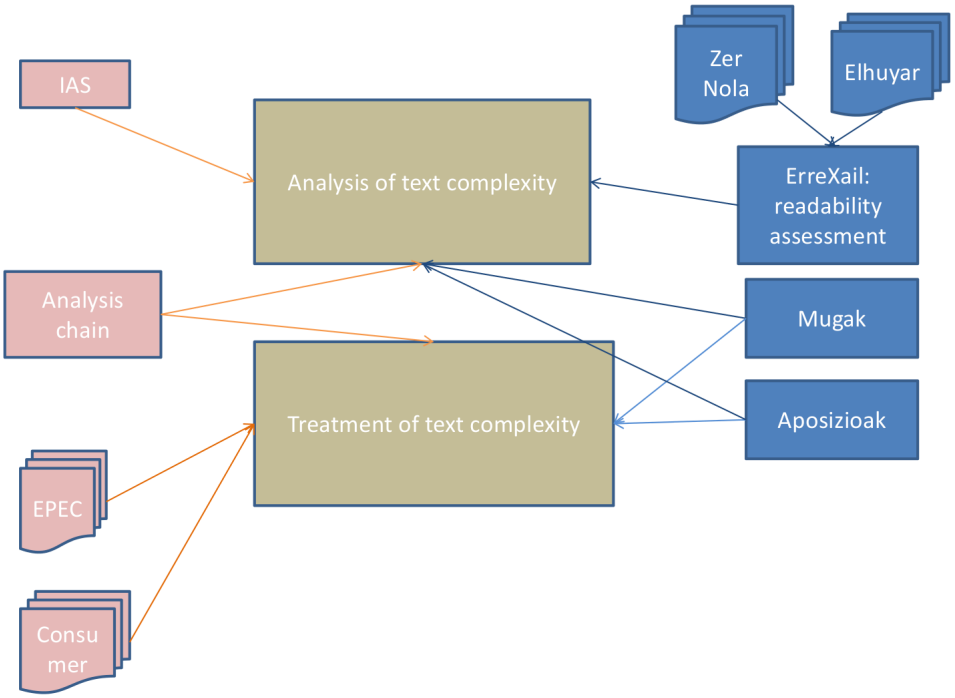

Figure 5.3 – Resources and tools used during thesis, and the contributions

First Approach to Automatic Text Simplification in Basque

María Jesús Aranzabe*, Arantza Díaz de Ilarraza, Itziar Gonzalez-Dios****

IXA NLP Group, Basque Philology Department*, Languages and Information Systems** University of the Basque Country
Sarriena auzoa zg 48940 Leioa*, Manuel Lardizabal 1 48014 Donostia***
maxux.aranzabe@ehu.es, a.diazdeilarraza@ehu.es, igonzalez010@ikasle.ehu.es

Abstract

Analysis of long sentences are source of problems in advanced applications such as machine translation. With the aim of solving these problems in advanced applications, we have analysed long sentences of two corpora written in Standard Basque in order to make syntactic simplification. The result of this analysis leads us to design a proposal to produce shorter sentences out of long ones. In order to perform this task we present an architecture for a text simplification system based on previously developed general coverage tools (giving them a new utility) and on hand written rules specific for syntactic simplification. Being Basque an agglutinative language this rules are based on morphological features. In this work we focused on specific phenomena like appositions, finite relative clauses and finite temporal clauses. The simplification proposed does not exclude any target audience, and the simplification could be used for both humans and machines. This is the first proposal for Automatic Text simplification and opens a research line for the Basque language in NLP.

1. Introduction

Automatic Text Simplification (TS) is a NLP task which aims to simplify texts so that they are more accessible, on one hand, among others to people who learn foreign languages (Petersen and Ostendorf, 2007); (Burstein, 2009) or people with disabilities (Carroll et al., 1999); (Max, 2005). And, on the other hand, it is useful for advanced NLP applications such us machine translation, Q&A systems or dependency parsers (Chandrasekar et al., 1996). In either cases, it is of prime importance to keep the meaning of original text, or at least trying not to lose information.

TS systems and architectures have been proposed for languages like English (Siddharthan, 2006), Portuguese (Candido et al., 2009), Swedish (Rybing et al., 2010), and there is ongoing work for Arabic (Al-Subaihin and Al-Khalifa, 2011) and Spanish (Saggion et al., 2011). Considering the advantages that these systems offer, we will explain here the architecture for a TS system based on the linguistic approach done so far for the Basque language, an agglutinative free-order language, in which grammatical relations between components within a clause are represented by suffixes.

This paper is structured as follows: in section 2 we explain briefly the linguistic typology of Basque associated to our problem. After that, in section 3 we present the corpora we have used. In section 4 we explain the process to simplify we propose and after it our architecture in section 5. The syntactic simplification proposals of the phenomena we have treated will be explained in section 6 and in section 7 we will expose this process by means of an example. We will finish the paper with the conclusion in section 8.

2. Typology of Basque

Basque is not an Indo-European language and differs considerably in grammar from the languages spoken in surrounding regions. It is, indeed, an agglutinative head-final pro-drop isolated language. The case system is ergative-absolutive. Due to its rich morphology, we have to take into account the structure of words (morphological analysis) to achieve this simplification task.

Basque displays a rich inflectional morphology. Indeed, it provides information about the case (Absolutive, Ergative or Dative) on either synthetic or auxiliary verbs. Basque declarative sentences are composed of a verb and its arguments and they can contain postpositional phrases too. The inflected verb is either synthetic or periphrastic. The synthetic (*noa*) in (1) is only composed by a word and it contains all the lexical and inflective information. The periphrastic (*joan nintzen*) in (2) is composed, however, of two (or three) words: main verb with lexical and aspect information and auxiliary verb containing agreement morphemes, tense and modality (Laka, 1996).

(1) *Etxera* *noa*
House-ALL go-1SG.PUNTUAL
'I go home'

(2) *Etxera* *joan* *nintzen*
House-ALL go-PRF AUX-1SG
'I went home'

In order to build subordinating clauses we attach complementisers[1] (comp) to the part of the verb containing inflection information. After the complementiser *-(e)n* in (3)(it is both past and comp) suffixes can be attached *-(e)an*-INE)[2]

(3) *Etxera* *joan* *nintzenean*
House-ALL go-PRF aux-1SG.COMP.INE
'When I went home'

The canonical element order is Sub Dat Obj Verb, but it can be easily changed according to the focus. Adjuncts can be placed everywhere in the sentence and arguments are often elided (pro-drop). The order changes in negative sentences as well. Let us see the first sentence in negative in (4).

[1] In sense of a morpheme which introduces all types of subordinating clauses

[2] INE=inessive(locative), ALL=allative, PRF=perfective

(4) *Ez noa etxera*
not go-1SG.PUNTUAL House-ALL
'I'm not going home'

3. Corpora analysis

We have used two corpora for this task: EPEC: *Euskararen Prozesamendurako Errenferentzia Corpusa*-Reference Corpus for the Processing of Basque (Aduriz et al., 2006a) and *Consumer* corpus (Alcázar, 2005).

EPEC corpus contains 300 000 words written in Standard Basque and it is tagged at morphological, syntactical levels (dependency-trees) (Aranzabe, 2008), and semantic level: word senses according to Basque WordNet and Basque Semcor (Agirre et al., 2006) and thematic roles in (Aldezabal et al., 2010). It is being tagged too at the pragmatic level: discourse markers (Iruskieta et al., 2011) and anaphora (Aduriz et al., 2006b).

Consumer corpus[3] is used in machine translation since the texts it contains are written in four languages (Spanish, Basque, Catalonian and Galician). It is a specialised corpus, compliling texts published the *consumer* magazine: critics, product comparison and so on.

The main characteristic of those corpora is that they contain authentic text.

In order to study the structures that should be simplified in Basque, to get better results in advanced application such as machine translation, we have taken the longest sentences from both corpora. We based our hypothesis on the results obtained by the machine translation system developed in our group when translating sentences of different length (Labaka, 2010). The results show that, the longer sentence longer, the higher error rate in Basque Spanish translation (table 1). The error rate used for scoring the results is HTER (Human-targeted Translation Error Rate) (Snover et al., 2006).

Words per sentence	**0-5**	**0-10**	**10-20**	**$>$ 20**
Sentences in corpora	5	41	100	59
HTER	17,65	28,57	32,54	49,16

Table 1: Sentence length and error rate in MT

Taking into account the results of the analysis of both corpora, we show in table 2 the sentence number we have treated in the corpora analysis and number that should be simplified, since they are complex sentences (with one or more complementisers). The third and fourth lines show the number of words that the longest and the shortest sentences we have in both corpora.

4. Simplification Process

The simplification process illustrates the operations that should be done and the steps we follow in order to produce simple sentences out of long sentences. Some of the operations we make have already been proposed in other TS works for other languages (Siddharthan, 2006) and (Aluísio et al., 2008).

In what follows we explain the operations considered:

[3]http://corpus.consumer.es/corpus/

	EPEC	**Consumer**
Long sentences	595	196
Complex sentences	488	173
Words/longest sentence	138	63
Words/shortest sentence	14	22

Table 2: Number of sentences and sentence length in Corpora

1. **Spliting:** Make as many new sentences as clauses out of the original.
2. **Reconstruction:** Two operations take place:
   (a) Removing no longer needed morphological features like complementisers (comp). Being Basque an agglutinative language we have to remove parts of words and not a whole word in case of finite verbs.
   (b) Adding new elements like adverbs or paraphrases. The main goal is to maintain the meaning.
3. **Reordering:** Reorder the elements in the new sentences, and ordering the sentences in the text.
4. **Adequation and Correction:** Correct the possible grammar and spelling mistakes, and fix punctuation and capitalisation.

This process will be ilustrated in section 7 by means of an example.

5. System Architecture

In this section we will present the architecture of the system we propose (see figure 1) to perform the steps mentioned in section 4. Having as input the text to be simplified, we distinguish different steps in our process:

1. The first step will be to evaluate the complexity of the text by means of a system already developed by our group for the auto-evaluation of essays *Idazlanen Autoebaloaziorako Sistema (IAS)* module (Castro-Castro et al., 2008). This module examines the text in order to determine its complexity based on several criteria such as the clause number in a sentence, types of sentences, word types and lemma number among others.
2. Once a sentence has been categorised as complex in the previous step, *Mugak* module (a system created in our group for detecting chunks and clauses) (Arrieta, 2010) will help us in the task of splitting long sentences into simple ones. *Mugak* is a general purpose clause identifier that combines rule-based and statistical-based clause identifiers previously developed for Basque. It works on the basis of the output produced by several tools implemented in our group[4]:

[4]http://ixa.si.ehu.es/Ixa

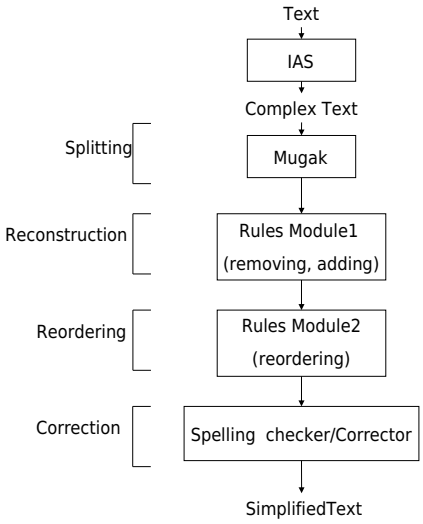

Figure 1: The architecture of system

- **Morpho-syntactic analysis:** *Morpheus* (Aduriz et al., 1998) makes word segmentation and PoS tagging. Syntactic function identification is made by *Constraint Grammar* formalism (Karlsson et al., 1995).
- **Lemmatisation and syntactic function identification:** *Eustagger* (Aduriz et al., 2003) resolves the ambiguity caused at the previous phase.
- **Multi-words items identification:** The aim is to determine which items of two or more words are always next to each other (Ezeiza, 2002).
- **Named entity recognition:** *Eihera* (Alegria et al., 2003) identifies and classifies named-entities in the text (person, organisation, location).

3. DAR (Deletion and Addition Rules) module includes a set of rules to perform the necessary deletions of morphological features and additions of grammatical elements in the split sentences. For example figure 2, shows the rule that would be applied to an auxiliary verb (aux) with a suffix in inessive, we remove the complementiser and the suffix (ine) and we add the adverb *ordu*-INE:

   We are defining the basic rules for the treatment of the phenomena explained in this paper. We are testing 15 rules and this process will be enriched while we go forward in our linguistic research.

4. ReordR (Reordering Rules) module includes a set of rules to perform the reordering needed in the created new sentences.

```
 if aux comp +ine {
remove comp and ine;
add ordu+ine in main clause;
}
```

Figure 2: A rule for an adverbial temporal sentences

5. Finally, the spell checker for Basque Xuxen (Agirre et al., 1992) will be applied in order to correct the created sentences.

6. Treated Phenomena

In the following subsections we give examples of the structures we have analysed and after them we give their simplifications. We follow the order that this structures have been explained in (Specia et al., 2008), i.e. apposition, relative clauses, adverbial subordinated clauses, coordinated clauses, non-inflected verb clauses and passive voice. In this paper we explain the simplification procedure for three structures: i) apposition and parenthetical structures, ii) finite relative clauses and iii) finite adverbial temporal clauses.
These structures are analysed in more details in (Gonzalez-Dios and Aranzabe, 2011).

6.1. Apposition and parenthetical structures

These structures give additional information about something that has been previously mentioned. Following we explain in (5) and (6) the process proposed for these structures. Sentences correspond to real text but have been shortened for clarity.
The steps for the treatment of (5) are:

1. When splitting we take the nominal group (NG) and the apposition to make several clauses out of the original one. In (5) NG are *Jose Maria Aznar* and *Javier Arenas* and their corresponding appositions are *Espainiako presidenteak* and *PPko idazkari nagusia*.

2. (a) We remove the apposition out of the original sentence.
   (b) Then, we add the copula verb to nominal group and the apposition, and so a new sentence is built (as we have here two apposition, two sentences will be built).

3. To reorder the elements in the sentence that has been built we follow this pattern:

```
NG(subj) apposition(pred) copula
```

   The ordering of the new sentences will be according to the order the appositions had in the original sentence (b) and (c) but the main clause in the original sentence will be the first one (a).

4. To check that the new sentences are grammatically correct and fix the punctuation by means of *XUXEN*.

(5) *Pankarta eraman zuten, besteak beste, Jose Maria Aznar Espainiako presidenteak eta Javier Arenas PPko idazkari nagusiak.*

The President of Spain Jose Maria Aznar and the Secretary-general of PP Javier Arenas carried the placard among others.

And those are the simplified sentences (a), (b) and (c):

a. *Pankarta eraman zuten, besteak beste, Jose Maria Aznarrek, eta Javier Arenasek.*

Jose Maria Aznar and Javier Arenas, carried the placard among others.

b. *Jose Maria Aznar Espainiako presidentea da.*

Jose Maria Aznar is President of Spain.

c. *Javier Arenas PPko idazkari nagusia da.*

Javier Arenas is Secretary-general of PP.

For parenthetical structures (6), we should repeat the process explained before. Sometimes we should retrieve the previously mentioned information as well to replace an elided element.

(6) *Hala ere, badirudi Sabino (Badajozetik fitxatuta), Moha (Barcelona B-tik) eta Aitor Ocio (Athleticek utzita) ez direla aurtengo fitxaketa bakarrak izango.*

However, it seems that Sabino (signed up from Badajoz), Moha (from Barcelona B) and Aitor Ocio (transferred from Athletic Bilbao) are not going to be the only signings.

And those are the simplified sentences (a), (b), (c) and (d):

a. *Hala ere, badirudi Sabino, Moha, eta Aitor Ocio ez direla aurteko fitxaketa bakarrak izango.*

However, it seems that Sabino, Moha and Aitor Ocio are not going to be the only signings.

b. *Sabino Badajozetik fitxatua da.*

Sabino is signed up from Badajoz.

c. *Moha Barcelona Btik fitxatua da.*

Moha is signed up from Barcelona B.

d. *Aitor Ocio Athleticek utzita da.*

Aitor Ocio is transfered from Athletic.

By simplifying the appositions this way the meaning of several entities will be *ipso facto* explained. Anyway, it would be necessary to explain the other entities in sentences, which are not appositions, if our target audience were humans (foreigners, second language learners, people lacking general knowledge). Sentences similar to the one presented here (with named-entities, references to persons, places etc.) could be enriched by facilitating access to Wikipedia[5]. This could be useful in a future proposal.

6.2. Relative clauses

Contrary to other subordinated clauses, relative clauses modify a noun and not a verb. There are different relativisation strategies in Basque: ordinary embedded relative clauses and appositive and extraposed relatives with relative pronouns (Oiarzabal, 2003). We consider that both can be simplified the same way. Sentence (7) is an example of the first strategy (ordinary embedded).

1. We split the sentence into relative clause and main clause. *Mugak* produces this output.
   (a) We will remove the complementiser.
   (b) We will copy the substantive they modified (the antecedent). In (7) the antecedent is *Ollanta Moises Humala teniente koronelak*. We will add the substantive to the previously removed relative clause, in the place of PRO[6], building a new simple sentence. We have to take into account the inflection case that the antecedent will have in the new sentence and give it the case that PRO has. If the clause is introduced by a relative pronoun, we use its inflection.
2. The subordinated clause will be left as it was, after having removed the complementiser.
3. To order the sentences we will keep the order they have in the original (relt (a) + main (b)).

This sentence (7) also presents an apposition linked to *Alberto Fujimori*, so in this case the treatment defined for appositions should be applied (here we just focused on finite relative clauses).

(7) *JOAN den igandeaz geroztik Alberto Fujimori Peruko presidentearen aurka altxamendu militar bat gidatzen ari den Ollanta Moises Humala teniente koronelak ez du uste bakarrik dagoenik (...)*

Since last Sunday Lt. Cr. Ollanta Moises Humala who is leading a military uprise against Peru president Alberto Fujimori does not think that he is alone.

And those are the simplified sentences (a) and (b):

a. *Joan den igandeaz geroztik Alberto Fujimori Peruko presidentearen aurka altxamendu militar bat gidatzen ari da Ollanta Moises Humala teniente koronela.*

Since last Sunday Lt. Cr. Ollanta Moises Humala is leading a military uprise against Peru president Alberto Fujimori.

[5] http://eu.wikipedia.org/wiki/Azala

[6] Phonetically null but syntactically active element

b. *Ollanta Moises Humala teniente koronelak ez du uste bakarrik dagoenik (...)*

Lt. Cr. Ollanta Moises Humala does not think that he is alone.

This will be the simplification of the most common finite relative clause type in Basque.

6.3. Adverbial temporal clauses

Adverbial clauses are adjuncts that specify relations like time, place, cause, consequence...with a reference to a main verb. As they constitute a heterogeneous group, we have decided to begin our experiment with the finite temporal adverbial clauses, and in the future we will expand our research.

1. So, we will split the original sentence (8).

2. The original main sentence will only be changed by adding an adverb (in (8) *orduan*) and by removing the subordinated clause. The subordinated will be left as the original, after having removed the complementiser and the suffix, which are attached to the auxiliary verb in case of periphrastic verbs, or to the main verb if the verb is synthetic.

   The element we add will be built this way: *ordu*-SUFFIX. The suffix is the one that is in the verb of the subordinated clause after the complementiser.

3. The problem with these clauses will be the ordering of new sentences and it will be more problematic if there are anaphoric elements. Meanwhile we have decided to keep the order the clauses in the original sentence, and if there is more than a subordinate clause, to put the former subordinated before the main clause, when they become simple sentences. In (8) both ordering have the same effect (a) and (b).

4. The new sentences will be corrected, if necessary, and punctuated.

(8) *erabakia hartu behar izan zuenean, ez zuen inolako zalantzarik izan don Polikarpo Gogorzak.*

'When he/she/it needed to decide, Sir Polikarpo Gogorza had no doubt.'

The simplified sentences are (a) and (b):

a. *Erabakia hartu behar izan zuen.*

   'He/she/it needed to decide.'

b. *Orduan, ez zuen inolako zalantzarik izan don Polikarpo Gogorzak.*

   'Then/in that time Sir Polikarpo Gogorza had no doubt.

We think that the procedure we have presented here will be useful for other adverbial clauses.

7. Example

We will explain here the process explained in section 4. Sentence (9) has the three phenomena we have presented in this paper. The changes we want to point out are underlined. We use the glosses in order to illustrate the morphological process properly, when needed.
Let us explain some morpho-syntactic aspects of the sentence (9) before showing the simplification steps:
There are 5 verbs in sentence (9), and each one builds a clause. The main verb is *da*, therefore it builds the main clause. The verb *dute* is main too, but in our analysis system it is dependent on the substantive it is referring to as apposition. The periphrastic verbs *igurtzitzen ditugunean* and *sortzen den* build subordinated clauses, and contrary to *gertatu* they are inflected. The non-inflected verb *gertatu* will be simplified although it is not treated in this approach. It will be treated when we treat non-inflected verbs.[7]

1. **Spliting:** Each verb forms a clause and they will be separated from the original one.
   Temporal adverbial clause: *(S Metalak igurtzitzen ditugunean S)*
   Non-finite verb concessive clause: *(S nahiz_eta kargen bereizketa berdin gertatu S)*
   Relative clause: *(S sortzen den S)*
   Main clause: *(S partikulen mugimendua oso erraza da material hauetan S)*
   Apposition: *(S eroankortasun elektriko haundia dute S)*

2. **Reconstruction:** Two steps are performed:

   (a) **Removing:** The complementisers (-(e)n) and suffixes in subordinated clauses (-(e)an).
   (S Partikulen mugimendua sortzen da s)
   (S Metalak igurtzitzen ditugu S)
   (S sortzen da S)

   (b) **Adding:** Adverbs and nominal groups in simple sentences.
   (S Orduan partikulen mugimendua oso erraza da material hauetan S)
   material hauek (S eroankortasun elektriko haundia dute S)

3. **Reordering:** This step is not needed in this sentence.
   (S Metalak igurtzitzen ditugu S)
   (S partikulen mugimendua sortzen da S)
   (S Orduan nahiz_eta kargen bereizketa berdin gertatu, partikulen mugimendua oso erraza da material hauetan S)
   (S material hauek eroankortasun elektriko haundia dute S)

[7] IMPF=imperfective, GEN=genitive, ERG=ergative ABS=Absolutive

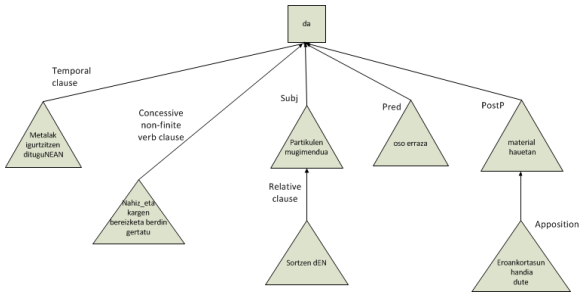

Figure 3: Tree of original sentence in example (9)

(9) *Metalak* *igurtzitzen* *dituguneanean,* *nahiz_eta* *kargen* *bereizketa* *berdin*
Metal-ABS.PL rub-IMPF aux-ABS3PL.ERG1PL.COMP.INE although charge-GEN separation-ABS equal
gertatu, *sortzen* *den* *partikulen* *mugimendua* *oso* *erraza* *da*
happen-PRF create-IMPF aux-ABS3SG.COMP(REL) particle-GEN movement-ABS grad easy-ABS is
material *hauetan* *(eroankortasun* *elektriko* *haundia* *dute).*
material det-INE conductivity-ABS electrical big have.

'When we rub metals, although charge separation happens equally, the particle movement that is generated is very easy in these materials (they have a high electrical conductivity).

4. **Correction and Adequation:**

Correct sentences can be seen glossed in (10) and the trees in figure 4. Sentences have been punctuated and a non standard verb *igurtzitzen* and a non standard adjective *haundia* have been corrected (standardised) in this step.

(10) a. *Metalak* *igurzten* *ditugu.*
Metal-ABS.PL rub-IMPF auxABS3PL.ERG1PL
'We rub metals.'

b. *Partikulen* *mugimendua* *sortzen*
Particle-GEN movement-ABS generate-IMPF
da.
aux-3SG
'The particle movement is generated'

c. *Orduan,* *nahiz_eta* *kargen*
Then(hour-INE) although charge-GEN
bereizketa *berdin* *gertatu,*
separation-ABS equal happen-PRF
partikulen *mugimendua* *oso*
particle-GEN movement-ABS grad
erraza *da* *material* *hauetan.*
easy-ABS is material det-INE
'Then although charge separation happens equally, the particle movement is very easy in these materials.'

d. *Material* *hauek* *eroankortasun*
conductivity-ABS electrical big
handia *dute.*
have
'These materials have a high electrical conductivity.'

At the end of the simplification process, the tree in figure 3 becomes 4 trees that we can see in figure 4. The inserted elements are ovals, main verbs are squares, and other constituents are triangles.

8. Conclusions

In this paper we have presented an approach for building a TS system for the Basque language, proposing an architecture and explaining simplification proposals for apposition and parenthetical structures, finite relative clauses and finite temporal clauses.
The approach is based on the linguistic study we have performed on long sentences taken from two corpora (EPEC and Consumer).
Similarly to other studies (Specia et al., 2008) our analysis leads us to detect the sentence structures susceptible of being simplified.
Although our first motivation was to produce simple sentences to help in advanced applications such as machine translation, we think that this study is valid for other purposes: education, foreign language learners and so on.
Most of the tools that are proposed in this work have been developed for general purposes and we are reusing them. Besides, we have evaluated them while we looked at the

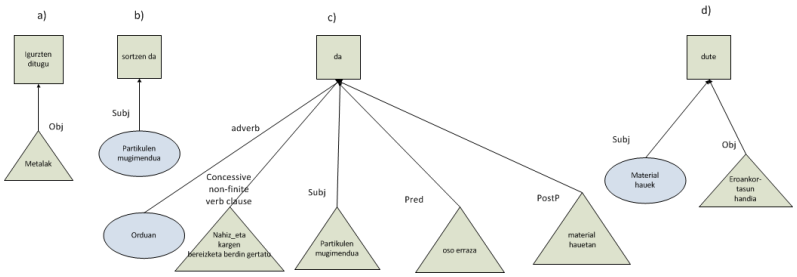

Figure 4: Tree of simplified sentences in example (10)

way to adapt them for our purpose. In this evaluation process we have concluded that *IAS* and *Mugak* are useful and that they can be a module of our architecture.

In any case, applying these rules we propose we get shorter sentences (Gonzalez-Dios and Aranzabe, 2011), which are translated automatically more easily, without losing the original meaning.

Although we have focused on syntactic simplification in this approach, it is important not to forget that in the future we should work on lexical simplification and text adaptation like proposed in (Siddharthan, 2006). We should remark as well that a part of this syntactic simplification approach is based on morphological constituents, which is necessary for high inflection languages like such a Basque.

It is important to mention too that the operations and the steps we make are similar to those which are made in other languages e.g. Portuguese (Specia et al., 2008), even though the tipology is different.

For the future, we should continue with this task by analysing other structures, improving the rules and their ordering, testing other methods (Woodsend and Lapata, 2011) (Siddharthan, 2011) using our dependency-based parsers (Aranzabe, 2008) (Bengoetxea et al., 2011), adapting the rules according to target audience etc.

9. Acknowledgements

Itziar Gonzalez-Dios's work is funded by a PhD grant from the Basque Government. This research was supported by the the Basque Government (IT344-10), and the Spanish Ministry of Science and Innovation (MICINN, TIN2010-202181).

10. References

I. Aduriz, E. Agirre, I. Aldezabal, I. Alegria, O. Ansa, X. Arregi, J.M. Arriola, X. Artola, A. Díaz de Ilarraza, N. Ezeiza, K. Gojenola, M. Maritxalar, M. Oronoz, K. Sarasola, A. Soroa, R. Urizar, and M. Urkia. 1998. A framework for the automatic processing of basque. In *Proceedings of Workshop on Lexical Resources for Minority Languages. First LREC Conference. Granada. 1998.”*.

I. Aduriz, Aldezabal. I., I. Alegria, J.M. Arriola, X. Artola, A. Díaz de Ilarraza, N. Ezeiza, and Gojenola. 2003. Finite State Applications for Basque. In *EACL'2003 Workshop on Finite-State Methods in Natural Language Processing. pp. 3- 11”*.

I. Aduriz, M. Aranzabe, J.M. Arriola, A. Atutxa, A. Díaz de Ilarraza, N. Ezeiza, K. Gojenola, M. Oronoz, A. Soroa, and R. Urizar, 2006a. *Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing*.

I. Aduriz, K. Ceberio, and A. Díaz de Ilarraza. 2006b. Pronominal anaphora in basque: annotation of a real corpus. In *XXII Congreso de la SEPLN (Sociedad Espanola para el Procesamiento del Lenguaje Natural), pp. 99-104, ISSN: 1135-5948”*.

E. Agirre, I. Alegria, X. Arregi, X. Artola, A. Díaz de Ilarraza, M. Maritxalar, K. Sarasola, and M. Urkia. 1992. Xuxen: A spelling checker/corrector for basque based in two-level morphology. In *Proceedings of NAACL-ANLP'92, 119-125. Povo Trento. 1992.*

E. Agirre, I. Aldezabal, J. Etxeberria, M. Iruskieta, E. Izagirre, K. Mendizabal, and E. Pociello. 2006. A methodology for the joint development of the basque wordnet and semcor. In *Proceedings of the 5th International Conference on Language Resources and Evaluations (LREC). ISBN 2-9517408-2-4. Genoa (Italy).”*.

A.A. Al-Subaihin and H.S. Al-Khalifa. 2011. Al-baseet: A proposed simplification authoring tool for the arabic language. In *Communications and Information Technology (ICCIT), 2011 International Conference on*.

A. Alcázar. 2005. Towards linguistically searchable text. In *Proceedings of BIDE Summer School of Linguistics*.

I. Aldezabal, M. Aranzabe, A. Díaz de Ilarraza, A. Estarrona, K. Fernandez, and L. Uria. 2010. EPEC-RS: EPEC (Euskararen Prozesamendurako Erreferentzia Corpusa) rol semantikoekin etiketatzeko eskuliburua. Technical report.

I. Alegria, N. Ezeiza, I. Fernandez, and R. Urizar. 2003. Named entity recognition and classification for texts in basque. In *II Jornadas de Tratamiento y Recuperacin de Informacin, JOTRI, Madrid. 2003. ISBN 84-89315-33-7”*.

S. M. Aluísio, L. Specia, T. A.S. Pardo, E. G. Maziero, and R. P.M. Fortes. 2008. Towards brazilian portuguese automatic text simplification systems. In *Proceedings of the eighth ACM symposium on Document engineer-*

ing, DocEng '08, pages 240–248, New York, NY, USA. ACM.

M. Aranzabe. 2008. Dependentzia-ereduan oinarritutako baliabide sintaktikoak: zuhaitz-bankua eta gramatika konputazionala. In *Euskal Filologia Saila (UPV/EHU). EHUko Gipuzkoako Campuseko Joxe Mari Korta I+G+b zentroan, 2008ko urriaren 30ean*.

N. Areta, A. Gurrutxaga, I. Leturia, I. Alegria, X. Artola, A. Díaz de Ilarraza, N. Ezeiza, and A. Sologaistoa. 2007. ZT Corpus: Annotation and tools for Basque corpora. In *Corpus Linguistics Conference. Birmingham.*

B. Arrieta. 2010. Azaleko sintaxiaren tratamendua ikasketa automatikoko tekniken bidez: euskarako kateen eta perpausen identifikazioa eta bere erabilera komazuzentzaile batean. In *Informatika Fakultatea (UPV-EHU)*.

K. Bengoetxea, A. Casillas, and K. Gojenola. 2011. Testing the Effect of Morphological Disambiguation in Dependency Parsing of Basque. In *International Conference on Parsing Technologies (IWPT). 2nd Workshop on Statistical Parsing Morphologically Rich Languages (SPMRL)*.

J. Burstein. 2009. Opportunities for Natural Language Processing Research in Education. In *Computational Linguistics and Intelligent Text Processing*. Springer Berlin / Heidelberg.

A. Candido, Jr., E. Maziero, C. Gasperin, T. A. S. Pardo, L. Specia, and S. M. Aluisio. 2009. Supporting the adaptation of texts for poor literacy readers: a text simplification editor for brazilian portuguese. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, EdAppsNLP '09, pages 34–42, Stroudsburg, PA, USA. Association for Computational Linguistics.

J. Carroll, G. Minnen, D. Pearce, Y. Canning, S. Devlin, and J. Tait. 1999. Simplifying Text for Language-Impaired Readers. volume 9th Conference of the European Chapter of the Association for Computational Linguistics.

D. Castro-Castro, R. Lannes-Losada, M. Maritxalar, I. Niebla, C. Pérez-Marqués, N.C. Alamo-Suarez, and A. Pons-Porrata. 2008. A multilingual application for automated essay scoring. In *Lecture Notes in Advances in Artificial Intelligence - LNAI 5290 - IBERAMIA ISBN 3-540-99308-8 Springer New York pp. 243-251"*.

R. Chandrasekar, Christine Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of the 16th conference on Computational linguistics - Volume 2*, COLING '96, pages 1041–1044, Stroudsburg, PA, USA. Association for Computational Linguistics.

N. Ezeiza. 2002. *CORPUSAK USTIATZEKO TRESNA LINGUISTIKOAK. Euskararen etiketatzaile morfosintaktiko sendo eta malgua.* Ph.D. thesis.

I. Gonzalez-Dios and M.J. Aranzabe. 2011. Euskarazko egitura sintaktikoen azterketa testuen sinplifikazio automatikorako: Aposizioak, erlatibozko perpausak eta denborazko perpausak. Master's thesis, University of Basque Country, September.

M. Iruskieta, A. Díaz de Ilarraza, and M. Lersundi. 2011. Unidad discursiva y relaciones retricas: un estudio acerca de las unidades de discurso en el etiquetado de un corpus en euskera. In *Procesamiento del Lenguaje Natural 47"*.

F. Karlsson, A. Voutilainen, J. Heikkila, and A. Anttila. 1995. *Constraint Grammar, A Language-independent System for Parsing Unrestricted Text*. Mouton de Gruyter.

G. Labaka. 2010. EUSMT: Incorporating Linguistic Information into SMT for a Morphologically Rich Language. Its use in SMT-RBMT-EBMT hybridation. In *Lengoaia eta Sistema Informatikoak Saila (UPV-EHU). Donostia. 2010ko martxoaren 29a*.

I. Laka. 1996. A brief grammar of euskara, the basque language.

A. Max. 2005. Simplification interactive pour la production de textes adapts aux personnes souffrant de troubles de la comprhension. volume Proceedings of Traitement Automatique des Langues Naturelles (TALN).

B. Oiarzabal, 2003. *A Grammar of Basque*, chapter Relatives. Mouton de Gruyter.

S. E. Petersen and M. Ostendorf. 2007. Text Simplification for Language Learners: A Corpus Analysis. *Electrical Engineering*, pages 69–72.

J. Rybing, C. Smith, and A. Sivervarg. 2010. Towards a Rule Based System for Automatic Simplification of texts. In *The Third Swedish Language Technology Conference (SLTC 2010)*.

H. Saggion, E. Gómez-Martínez, E. Etayo, A. Anula, and L. Bourg. 2011. Text Simplification in Simplext: Making Text More Accessible. *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*.

A. Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language & Computation*, 4(1):77–109.

A. Siddharthan. 2011. Text Simplification using Typed Dependencies: A Comparision of the Robustness of Different Generation Strategies. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 2–11, Nancy, France, September. Association for Computational Linguistics.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA'2006*, pages 223–231, Columbus, Ohio, June.

L. Specia, S.M. Aluisio, and T.A.S Pardo. 2008. Manual de Simplificao Sinttica para o Portugułs. Technical Report NILC-TR-08-06, So Carlos-SP.

K. Woodsend and M. Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Simplifying Basque Texts: the Shallow Syntactic Substitution Simplification

Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza

Ixa NLP Group, University of the Basque Country (UPV/EHU)
Manuel Lardizabal, 1 Donostia
{itziar.gonzalezd, maxux.aranzabe, a.diazdeilarraza}@ehu.eus

Abstract

In this paper we present the automatic simplification levels we have defined for Basque. These levels will be chosen according to the requirements and level of the target audience. Along with that, we go through the details of the first simplification level, namely the Shallow Syntactic Substitution Simplification (SSSS). We explain its motivation, our frequency based approach and evaluate the output taking into account the correction, grammaticality and simplicity. The latter is evaluated by linguists and the target audience. To carry out this experiment we have compiled a corpus of infrequent syntactic structures.

1. Introduction and related work

Automatic Text Simplification (ATS) is a research line in Natural Language Processing (NLP) that, given a source text, aims to create a simpler version of that text. The original texts can be simplified according to the required level and can be used with different target audiences: they can be oriented to people with impairment, languages learners and also to facilitate the processing of NLP advanced applications (Gonzalez-Dios et al., 2013; Shardlow, 2014; Siddharthan, 2014). For example, in the project PSET (Practical Simplification of English Text), they concentrated on the need of Aphasic readers (Carroll et al., 1998). Simplification strategies have also been proposed for people with dyslexia (Rello et al., 2013), or autism (Evans et al., 2014), children (De Belder and Moens, 2010; Barlacchi and Tonelli, 2013), language learners (Petersen and Ostendorf, 2007) and poor literacy readers (Gasperin et al., 2009). NLP advanced applications which have been target audience for ATS systems are e.g. parsers (Chandrasekar et al., 1996), information retrieval systems (Beigman Klebanov et al., 2004) or machine translation (Doi and Sumita, 2004; Poornima et al., 2011).

There are two main simplification types in ATS: syntactic simplification and lexical simplification. Syntactic simplification aims to rewrite sentences to get a more simple equivalent of them that will be accessible to a target audience. Lexical simplification seeks to rewrite complex or low frequency words by substituting them with synonyms or paraphrases. So far, syntactic simplification has concentrated mainly on sentence splitting and sentence transformation and generation while lexical simplification has principally treated word and phrase substitutions (Siddharthan, 2002). The work done so far in ATS for Basque has focused mainly on syntactic simplification (Aranzabe et al., 2012) with the aim of getting shorter sentences that preserve the meaning of the original one.

In this paper we present the three simplification levels we define for Basque (shallow syntactic substitution, natural and strong or absolute simplifications) and go through the first level. Shallow Syntactic Substitution Simplification (SSSS) is a substitution operation similar to those that are applied at lexical simplification but at syntactic level, which is the domain of syntactic simplification. So, SSSS can be understood as a mixture of both or as a continuum operation between both and it is intended for advanced language learners and non-fluent speakers. Apart from the simplification, which is our main motivation, we think the approach we present here can be used for other applications, such as standardisation or normalisation of historical texts.

Although nowadays ATS for English is getting more attention from the data driven methods, lesser resourced languages still concentrate on knowledge-based or semi-data driven methods. Due to the fact that Basque is a language with a data scarce problem, we based our study and approach on corpus analysis and linguistic knowledge.

This paper is structured as follows: we define the simplification levels for Basque in section 2. We go through the Shallow Syntactic Substitution Simplification, explaining our approach and evaluation in section 3. Finally, we conclude and outline future work in section 4.

2. Simplification framework: levels and operations

Texts can be simplified according to the needs of the target group. In The PorSimples project, targeting poor literacy readers, two simplification levels are defined: natural simplification and strong simplification (Gasperin et al., 2009). The former is intended for people with a basic literacy level and the latter to people with a rudimentary level. In natural simplification certain operations such as splitting and inversion of clause ordering are dealt with, while in strong simplification a set of pre-defined simplification operations is applied with the aim of making the sentence as simple as possible.

In our study, based on those two syntactic simplification levels, we add a third one. In what follows, we define our three levels of simplification targeting Basque language learners and/or non-fluent speakers.

1. **Shallow Syntactic Substitution Simplification (SSSS):** Frequency based simplification of syntactical structures. This level is intended for people

who have a good level of Basque and master Basque syntax but do not know unusual, dialectal and synchronic variations. That is, at this level the depth of the syntactical structure is kept but the structure that is used is more frequent. These people are usually advanced learners or non-fluent speakers.

2. **Natural Simplification (NS):** Compound and complex sentences with finite verb simplification will follow the simplification process for Basque (Aranzabe et al., 2012) together with the SSSS. That is, the following operations will take place: 1) splitting: sentences will be split into clauses; 2) reconstruction: morphological features such as complementisers (comp) and case markers will be removed and new elements, such as adverbs, connectors, verbs or phrases that will keep the meaning of the original sentence, will be added (added elements); 3) reordering: sentences will be ordered in the text; and 4) correction: possible mistakes (grammatical errors and standardisation) will be corrected. In this level the syntactic depth of the sentences is altered. The target of this level is people who are learning Basque but get stuck with long sentences and do not master syntax. Advanced NLP applications can benefit from this level and get better results with shorter sentences.

3. **Strong or Absolute Simplification (AS):** Everything is simplified. Both sentences with finite and non-finite verbs will follow the simplification process. SSSS will also be applied. The syntactic depth of the sentences is also altered as in the previous level. This level will be useful for people with low knowledge of Basque syntax or advanced NLP applications that get better results by processing only one verb per sentence.

However, our system can apply only needed or required phenomena, depending on the needs of a special target audience (customised simplification (CS)). However, the sentence that undergoes the simplification process should have more than one complement or adjunct. With this premise we want to avoid sentences that are too short and could sound unnatural. The operations performed in the SSSS will be explained in section 3.

3. Shallow Syntactic Substitution Simplification

The SSSS is a frequency based simplification that aims at providing a simpler option but which keeps the subordinate clause. That is, we use lexical simplification techniques applied to syntax. This approach is useful above all with the adverbial clauses, where we have found a high diversity of structures. Although we focus and perform our experiments in Basque, we think that this approach is also viable in other languages.

3.1. Motivation

The main motivation for SSSS is that some target audiences such as advanced learners or non-fluent speakers do not need big structural changes in syntax processing but some structures are unknown to them because they are dialectical or synchronic variations. Other structures are also ambiguous at pointing out different relations. Our aim with this approach is to give the text a simple equivalent without making structural changes using the clearest and most frequent option.

In example (1) we see a sentence simplified at absolute level which has undergone the simplification process defined for Basque. In that sentence we find a non-finite purpose structure *-tzearren* (in order to) and to simplify it we follow the defined simplification operations: the sentence has been split, the relation marker has been removed, the verb of the subordinate clause has been put in the participial form (*suspertu*) and the verb *nahi izan* (to want) has been included, according to its rule. Then, following the rule of the purpose clauses, the sentences have been checked to see if they follow the main-subordinate order. Otherwise, they would have been reordered. Finally, the correction of the simplified sentences has been checked.

(1) a. *Abuztuaren amaieran beste goi bilera bat egitea aztertzen ari dira Israel eta PAN Palestinako Aginte Nazionala, Ekialde Erdiko bake prozesua susper***tzearren**. ('Israel and the PNA, the Palestinian National Authority, are studying the organization of another summit at the end of August to promote the peace process in the Middle East.')

b. i. *Abuztuaren amaieran beste goi bilera bat egitea aztertzen ari dira Israel eta PAN Palestinako Aginte Nazionala.* ('Israel and the PNA, the Palestinian National Authority, are studying the organization of another summit at the end of August.')

ii. *Ekialde Erdiko bake prozesua susper***tu nahi dute**. ('They want to promote the peace process in the Middle East.')

Using this 'classical' syntactic simplification, the subordinate clause has disappeared from the main clause and has become an independent clause. But advanced learners or low literacy speakers may understand that there is a subordinate clause in the original sentence but do not understand the relation it points out. So, we consult in the structure frequency list (Table 1) and we see that *-tzearren* is used as a non-finite purpose structure 1.68 % while *-tzeko* is used the 88.38 %. Then, to simplify the sentence, we substitute that syntactic structure with its most frequent equivalent (in this case *-tzeko*) as lexical simplification does with words. This way, a simpler option has been given but the subordinate clause is kept.

Structure	Quantity	Percentage
-tzeko (in order to)	791	88.38
-tzekotzat (in order to)	0	0.00
-tzearren (in order to)	15	1.68
-tzeagatik (in order to)	0	0.00
-tze alde(ra) (in order to)	0	0.00
-tzekotan (in order to)	0	0.00

Table 1: Frequency list of non-finite purpose clauses

In (2) we have performed a SSSS of (1a) by substituting *-tzearren* with *-tzeko*, *suspertzearren -> suspertzeko* only

being changed in the sentence. The meaning (and therefore the translation) and the syntactic tree do not change at all.

(2) a. *Abuztuaren amaieran beste goi bilera bat egitea aztertzen ari dira Israel eta PAN Palestinako Aginte Nazionala, Ekialde Erdiko bake prozesua sus-per***tzearren**.

b. i. *Abuztuaren amaieran beste goi bilera bat egitea aztertzen ari dira Israel eta PAN Palestinako Aginte Nazionala, Ekialde Erdiko bake prozesua suspertzeko*.

SSSS will be used above all with non-finite clauses but it can be also used with finite clauses.

3.2. Methodology

In order to perform the SSSS, we have carried out the following steps:

1. We have made a list of the structures presented by *Euskaltzaindia*, the Royal Academy of the Basque Language, in its descriptive grammar *Euskal Gramatika: Lehen Urratsak* (Euskaltzaindia, 1999; Euskaltzaindia, 2005; Euskaltzaindia, 2011). This grammar collects the synchronic and dialectical structures that have been used in written Basque.

2. The list of structures has been consulted in the Basque Dependency Treebank (BDT)[1] (Aranzabe, 2008) and their presence, frequency and position has been examined. To formalise our approach, we have taken the information about the frequencies of that corpus analysis (Gonzalez-Dios et al., 2015).

3. We have checked the meaning equivalences of the structures manually. To that end, we have also used the information of the grammar. That is, we have assembled the structures that have the same meaning.

4. Based on the frequencies, we have substituted the uncommon syntactic structures with a more frequent equivalent syntactic structure.

The list of structures and the frequency information are language dependent resources. These should be changed to apply this method to other languages.

3.3. Corpus and approach

To carry out the SSSS, we have compiled a corpus with the examples given in the descriptive grammar (Euskaltzaindia, 1999; Euskaltzaindia, 2005; Euskaltzaindia, 2011). One half of the examples were used for the training part and the other half for the test. Each part had 54 instances. More details about the corpus can be found in Table 2.

In table 3 we detail the number of target structures, the substitution options and the implemented rules. 17 options have been defined to substitute 39 structures. That is, there are 17 frequent structures that are going to substitute 39 non-frequent or less frequent structures. To perform these substitutions, 42 rules based on regular expressions have been implemented. Those rules are applied at text level. The substitution option is the most frequent structure of each type or subtype. For example, there are 5 options for temporal relations because there is one option for each subtype of temporal clause (anteriority, posteriority, impendency, simultaneity and repeated simultaneity). No target structure is a substitution option. This way, we eliminate the possible relation ambiguity (structures that point out more than one relation are never used as a substitution option).

[1]BDT is the version of the Reference Corpus for the Processing of Basque (EPEC) (Aduriz et al., 2006) and compiles 200 000 words written in standard Basque.

Part	Sentences	Clauses	Words
Training	54	583	155
Test	54	588	138

Table 2: Sentence, clause and word number found in the corpus

	Target structures	Substitution option	Rules
Total	39	17	42
Temporal	15	5	16
Causal	1	1	1
Purpose	5	1	5
Conditional	6	3	9
Concessive	2	1	3
Modal	10	6	8

Table 3: Summary of the structures and rules

3.4. Evaluation and error analysis

To evaluate our approach we have taken into account two parameters: correct substitution and grammatically correct output (correct sentences). The correct substitutions column show the percentage of the sentences correctly performed and the correct sentences column shows the percentage of grammatically correct sentences for the cases where the substitution was correct. These results can be seen in table 4.

	Correct substitutions	Correct sentences
Total	79.63	88.64
Temporal	62.50	93.34
Causal	100.00	100.00
Purpose	100.00	100.00
Conditional	88.89	62.50
Concessive	100.00	100.00
Modal	90.00	100.00

Table 4: Results of the performance in total and by clause type

As we can see, the results with the most adverbial clause types are satisfactory. When we deal with the types with more changes and more structures the results, are, however worse. We have performed an error analysis and we discovered that most of the errors happen a) when changing the form of the verb (participial <-> verb noun)

and b) when the participles are marked with ø[2]. The former involves incorrect substitutions and the latter ungrammatical sentences. To overcome these problems, the substitution should be made at analysis level with tools to work with the morphology of Basque (Alegria, 1995) and using advanced Natural Language Generation (NLG) techniques (Agirrezabal et al., 2015). In fact, we should work with the form found in the two-level morphology.

We also evaluated the simplicity of the generated sentences to see if our frequency based approach is valid to get simpler sentences. To that end, two linguists from different parts of the Basque Country with expertise in language learning and teaching evaluated the corrected substituted sentences. They were given both the original and the simplified sentences and we asked them to evaluate whether the generated sentences were simpler, equal or more difficult than the original taking into account Basque learners of their surroundings.

	Simpler	Equal	More difficult
Western linguist	76.74	23.26	0.00
East-central linguist	30.23	48.84	20.93

Table 5: Simplicity of the sentences evaluated by linguists

As we can see in the results of Table 5, the western linguist considered that all the sentences were mainly simpler (76.74 %) or equal (23.26 %). On the other hand, the east-central linguist considered that the most of the simplified sentences were equal (48.84 %) and only (30.23 %) of the them were simpler. She also judged that some sentences were more difficult (20.93 %).

We also tested the simplicity of the generated sentences with our target audience. We asked 2 advanced learners and 2 non-fluent speakers to take the test. All of them had at least the B2 level in Basque, universitary studies and they all came from different parts of the Basque Country. They were asked if the simplified sentences were simpler, equal or more difficult than the original for them.

	Simpler	Equal	More difficult
Total	75.00	25.00	0.00
Temporal	33.33	58.33	8.33
Causal	87.50	0.00	12.50
Purpose	75.00	25.00	0.00
Conditional	25.00	0.00	75.00
Concessive	37.50	37.50	25.00
Modal	75.00	25.00	0.00

Table 6: Simplicity judgements of the advanced learners and non-fluent speakers

The results of the advanced learners and non-fluent speakers is presented in Table 6. For brevity, we show the percentages of the number of testers that mainly gave that evaluation in total taking into account the sentence type. 75.00 % of the testers mainly found that the sentences were in total simpler and 25.00 % found them mainly equal. No one found that they were more difficult in general. That is, taking into account all the sentences, 3 out of the 4 testers considered that they were in general simpler. Looking at the origin of the tester, only the Est-central speaker considered that sentences were mainly of equal complexity as the Est-central linguist did.

If we see the results by clause type, we can see that conditional sentences were more difficult in general after the simplification and that temporal sentences were equal to the originals. But, looking at these results and taking into account the origin, we can see that both conditional and temporal simplified sentences are considered simpler by the western speaker (the western linguist also considered this). The interpretation of the concessive sentences shows no pattern and other types show good results.

Based on the outcome of this subjective experiment, we conclude that the frequency based approach is valid but it can be more helpful according to the origin. That is, the origin and the surrounding dialect of the target should be taken into account when simplifying the texts. This dialectical parametrisation can easily be included in the system.

4. Conclusion and future work

In this paper we have presented the simplification levels for the automatic text simplification of Basque written texts: the shallow syntactic substitution simplification, natural simplification, and strong or absolute simplification. We have detailed the approach of SSSS presenting its motivation, our approach and the evaluation. The performance results we obtain are satisfactory. We also evaluate the simplicity of the generated sentences with linguists and advanced learners and non-fluent speakers. We find that the results vary depending on the origin of the speaker. We conclude that this simplification level is suitable for people who do not know all the dialectical and synchronic adverbial structures of Basque but we have also seen that the effectiveness of the simplification depends on the origin of that target. That is, we think that the origin is important when simplifying texts.

In the future, we plan to correct the errors found in our analysis using morphological tools and NLG advanced techniques. Moreover, we are working on the implementation of the rest of the simplification levels. Further testing with other kinds of learners (with other levels) will also be interesting to perform. It will also be interesting to see how this approach can be used in other languages.

Acknowledgments.

Itziar Gonzalez-Dios' work is funded by a Ph.D. grant from the Basque Government. This research is also supported by the the Basque Government (IT344-10). We are also very grateful to Rodrigo Agerri, Begoña Altuna, Unai Lopez-Novoa, Vanessa Martin, Itziar Otaduy and Larraitz Uria that took part in our experiments.

5. References

Aduriz, Itziar, María Jesús Aranzabe, Jose Mari Arriola, Aitziber Atutxa, Arantza Díaz de Ilarraza, Nerea Ezeiza, Koldo Gojenola, Maite Oronoz, Aitor Soroa, and Ruben

[2]In Basque the participle is formed with *ø, -tu, -du, -i.*

Urizar, 2006. *Methodology and Steps Towards the Construction of EPEC, a Corpus of Written Basque Tagged at Morphological and Syntactic levels for Automatic Processing*, volume 56. Rodopi, pages 1–15.

Agirrezabal, Manex, Itziar Gonzalez-Dios, and Iñigo Lopez-Gazpio, 2015. Euskararen Sorkuntza Automatikoa: lehen urratsak [Automatic Generation of Basque: First Steps]. In *Proceedings of Ikergazte*.

Alegria, Iñaki, 1995. *Euskal morfologiaren tratamendu automatikorako tresnak [Tools for the Treatment of Basque Morphology]*. Ph.D. thesis, University of the Basque Country (UPV/EHU).

Aranzabe, María Jesús, 2008. *Dependentzia-ereduan oinarritutako baliabide sintaktikoak: zuhaitz-bankua eta gramatika konputazionala [Syntactic Resources based on the Dependency Model: the Treebank and the Computational Grammar]*. Ph.D. thesis, University of the Basque Country (UPV/EHU).

Aranzabe, María Jesús, Arantza Díaz de Ilarraza, and Itziar Gonzalez-Dios, 2012. First Approach to Automatic Text Simplification in Basque. In Luz Rello and Horacio Saggion (eds.), *Proceedings of the Natural Language Processing for Improving Textual Accessibility (NLP4ITA) workshop (LREC 2012)*.

Barlacchi, Gianni and Sara Tonelli, 2013. ERNESTA: A Sentence Simplification Tool for Childrens Stories in Italian. In *Computational Linguistics and Intelligent Text Processing*. Springer, pages 476–487.

Beigman Klebanov, Beata, Kevin Knight, and Daniel Marcu, 2004. Text Simplification for Information-Seeking Applications. *On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE*:735–747.

Carroll, John, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait, 1998. Practical Simplification of English Newspaper Text to Assist Aphasic Readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*. Citeseer.

Chandrasekar, Raman, Christine Doran, and Bangalore Srinivas, 1996. Motivations and Methods for Text Simplification. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96. Stroudsburg, PA, USA: Association for Computational Linguistics.

De Belder, Jan and Marie-Francine Moens, 2010. Text Simplification for Children. In *Proceedings of the SIGIR workshop on accessible search systems*.

Doi, Takao and Eiichiro Sumita, 2004. Splitting Input Sentence for Machine Translation Using Language Model with Sentence Similarity. In *Proc. of the 20th international conference on Computational Linguistics*.

Euskaltzaindia, 1999. V, (Mendeko perpausak-1, osagarriak, erlatiboak, konparaziozkoak, ondoriozkoak) [V (Subordinate Clauses-1, Completive, Relative, Comparative, Consecutive]. In *Euskal Gramatika Lehen Urratsak [Basque Grammar First Steps]*. Bilbo: Euskaltzaindia.

Euskaltzaindia, 2005. VI, (Mendeko perpausak-2, baldintzazkoak, denborazkoak, helburuzkoak, kausazkoak, kontseziozkoak eta moduzkoak) [VI (Subordinate Clauses-2, Conditional, temporal, Purpose, Causal, Concessive and Modal]. In *Euskal Gramatika Lehen Urratsak [Basque Grammar First Steps]*. Bilbo: Euskaltzaindia.

Euskaltzaindia, 2011. VII, (Perpaus jokatugabeak: denborazkoak, kausazkoak eta helburuzkoak, baldintzazkoak, kontzesiozkoak, moduzkoak, erlatiboak eta osagarriak) [VII (Subordinate Clauses-2, temporal, Causal and Purpose, Conditional, Concessive, Modal, Relative and Completive]. In *Euskal Gramatika Lehen Urratsak [Basque Grammar First Steps]*. Bilbo: Euskaltzaindia.

Evans, Richard, Constantin Orasan, and Iustin Dornescu, 2014. An evaluation of syntactic simplification rules for people with autism. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*. Gothenburg, Sweden: Association for Computational Linguistics.

Gasperin, Caroline, Erick Maziero, Lucia Specia, Thiago A.S. Pardo, and Sandra M. Aluisio, 2009. Natural Language Processing for Social inclusion: a Text Simplification Architecture for Different Literacy Levels. *the Proceedings of SEMISH-XXXVI Seminário Integrado de Software e Hardware*:387–401.

Gonzalez-Dios, Itziar, María Jesús Aranzabe, and Arantza Díaz de Ilarraza, 2013. Testuen sinplifikazio automatikoa: arloaren egungo egoera [Automatic Text Simplification: State of Art]. *Linguamática*, 5(2):43–63.

Gonzalez-Dios, Itziar, María Jesús Aranzabe, and Arantza Díaz de Ilarraza, 2015. Perpaus adberbialen agerpena, maiztasuna eta kokapena EPEC-DEP corpusean [Presence, frequency and Position of Basque Adberbial Clauses in The BDP corpus]. Technical report, University of the Basque Country (UPV/EHU) UPV/EHU/LSI/TR 02-2015.

Petersen, Sarah E. and Mari Ostendorf, 2007. Text Simplification for Language Learners: A Corpus Analysis. In *In Proceedings of Workshop on Speech and Language Technology for Education. SLaTE*. Citeseer.

Poornima, C., V. Dhanalakshmi, K.M. Anand, and KP Soman, 2011. Rule based Sentence Simplification for English to Tamil Machine Translation System. *International Journal of Computer Applications*, 25(8):38–42.

Rello, Luz, Ricardo Baeza-Yates, Stefan Bott, and Horacio Saggion, 2013. Simplify or Help? Text Simplification Strategies for People with Dyslexia. *Proc. W4A*, 13.

Shardlow, Matthew, 2014. A Survey of Automated Text Simplification. *International Journal of Advanced Computer Science and Applications (IJACSA), Special Issue on Natural Language Processing*:58–70.

Siddharthan, Advaith, 2002. An Architecture for a Text Simplification System. In *Proceedings of the Language Engineering Conference (LEC'02)*. Washington, DC, USA: IEEE Computer Society.

Siddharthan, Advaith, 2014. A Survey of Research on Text Simplification. *The International Journal of Applied Linguistics*:259–98.

Procesamiento del Lenguaje Natural, Revista nº 50 marzo de 2013, pp 61-68 recibido 30-11-12 revisado 07-02-13 aceptado 19-02-13

Transforming Complex Sentences using Dependency Trees for Automatic Text Simplification in Basque

Transformación de las oraciones compuestas utilizando árboles de dependencias para la Simplificación Automática de Textos en Euskera

María Jesús Aranzabe, Arantza Díaz de Ilarraza, Itziar Gonzalez-Dios
IXA NLP Group, University of the Basque Country (UPV/EHU)
Manuel Lardizabal 1 48014 Donostia
maxux.aranzabe@ehu.es, a.diazdeilarraza@ehu.es, igonzalez010@ikasle.ehu.es

Resumen: En este artículo se presenta uno de los módulos que forma parte del sistema de simplificación automática de textos escritos en euskera que se está implementando. Concretamente, se describe el módulo donde se lleva a cabo la transformación de las oraciones compuestas en oraciones simples. Esta transformación se realiza mediante las herramientas de alta precisión y cobertura general desarrolladas para el tratamiento automático del euskera. Además de adaptar y enriquecer el identificador de oraciones se ha implementado un algoritmo basado en árboles de dependencias sintácticas cuyo objetivo es dividir las oraciones complejas en oraciones más simples.
Palabras clave: Simplificación automática de textos, división de oraciones, euskera, identificación de las oraciones compuestas y simples

Abstract: In this paper we present a module of the Text Simplification architecture that we are implementing. Exactly, we describe the module that carries out the task of splitting sentences into clauses. This module is based on general-coverage tools. We have adapted the clause identifier in this module and we have added a algorithm based on dependency-trees to split the sentences. This way, we get simple sentences.
Keywords: Text Simplification, sentence splitting, Basque, clause boundary identification

1 Introduction

Automatic Text Simplification (TS) is a Natural Language Processing (NLP) task that aims the transformation of difficult texts to get a equivalent simple text. This may involve simplifyng syntactic phenomena, performing operations like sentence spliting, changing passive to active voice, inverting the order of the clauses, changing discourse marker by a simpler and/or more frequent one. As a result, this new text should easier to understand for humans and/or easier to process by NLP advanced applications and it should keep the meaning of original text, or at least information loss should be avoid.

TS systems and architectures have been proposed for languages like English (Siddharthan, 2006), Brazilian Portuguese (Candido et al., 2009), Swedish (Rybing, Smith, and Silvervarg, 2010), Japanese (Inui et al., 2003), Arabic (Al-Subaihin and Al-Khalifa, 2011), Spanish (Saggion et al., 2011), and French (Seretan, 2012). As method, dependency trees have been used in TS systems like (Zhu, Bernhard, and Gurevych, 2010) and (Siddharthan, 2011) among others.

The target audiences of the TS systems have been people with disabilities (Carroll et al., 1999), illiterate (Candido et al., 2009) or people who learn foreign languages (Petersen and Ostendorf, 2007) (Burstein, 2009) among others. There are TS system for NLP advanced applications such us machine translation (Poornima et al., 2011), Q&A systems (Bernhard et al., 2012), information extraction system (Jonnalagadda and Gonzalez, 2010), and so on.

One of the operations in TS is sentence splitting. In fact, it is a compulsory need to find precise splitting points in order to continue the next operations in the TS task. In this study we analyse two linguistic diverse structures in Basque like relative clauses and adverbial temporal clauses in other to evaluate how accurate our tools are. Besides, we implement and algorithm to create simple

ISSN 1135-5948

sentences out of a complex one. Although we get simple sentences, the simplification process is not achieved: complementisers and suffixes should be removed in other to get grammatically correct sentences.

This paper is strucutured as follows: In section 2 we describe the phenomena we have treated in this paper, namely relative clauses (subsection 2.1) and temporal adverbial clauses (subsection 2.2). In section 3 we describe the simplification process we follow together with our system architecture. In section 4 we explain how we transform the trees. After that in section 5 we present the evaluation. The conclusion and future work are presented in section 6.

2 Treated Phenomena

In order to make a deep analysis of the clause boundary identifier implemented in the splitting module we explain the two phenomena we have focused on: relative clauses and adverbial temporal clauses. We selected relative clauses since they are attached to a noun and on the other hand, adverbial temporal clauses have been chosen because they show varied structures.

The corpus that has been used for this task has been EPEC (*Euskararen Prozesamendurako Erreferentzia Corpusa*-Reference Corpus for the Processing of Basque). EPEC corpus contains 300,000 words written in Standard Basque and it is tagged at morphological and syntactical levels (dependency-trees) (Aduriz et al., 2006a). At semantic level the most frequent nouns have been tagged with their corresponding synset in EusWordNet and EusSemcor (Agirre et al., 2006). Besides, the instances of the most frequent verbs have been tagged with their thematic roles in (Aldezabal et al., 2010). At the pragmatic level, discourse markers (Iruskieta, Díaz de Ilarraza, and Lersundi, 2011) and coreference (Soraluze et al., 2012) are also tagged.

We will see in next sections examples illustrating the treated phenomena. We will only show the relevant morphological information in the glosses.

2.1 Relative clauses

Basque uses gapping as strategy for relativisation, which is marked as PRO[1]. Basque relative clause can be built with finite verbs (1) using the complementiser (comp) *-(e)n* and with non finite verbs (2), attaching to the participle the suffixes *(-ta/da, -ik, -i)* + *-ko* (rel). Let us see some examples where the relative clause is marked between brackets in the examples.

(1) *Horixe zen (magoak eta nik*
That was magician and I
genuen) sekretua.
had-COMP secret.

'That was the secret the magician and me shared.'

(2) *(Bildutako diruarekin,)*
CollectREL money-SOZ,
Afganistanerako hegazkin-txartela
Afghanistan-ALL plane-ticket
erosi zitzaion Pepitari.
buy aux Pepita-DAT

'With the collected money, a plane-ticket to Afghanistan was bought to Pepita.'

The location of finite relative clauses and non finite verb relative clauses within the sentence is at the left side of the antecedent. The subordinate verb is at the end of the relative sentence.

2.2 Adverbial temporal clauses

Adverbial temporal clauses are adjuncts that specify chronological ordering (anteriority, posteriority, simultaneity, delimitation, impendency and duration) having the reference of a main verb/clause. Temporal clauses constitute a heterogeneous group, not only semantically but syntactically too. They can be built with finite verbs and non finite verbs. In both cases free elements can be added.

Finite verb temporal clauses are headed by complementisers and suffixes are attached to verb (V) like *zu#-(e)n*V.COMP *#an*-INE in example (3). In some cases like (4) a free element (*bitartean*) is added after the verb with the complementiser. Let us see these examples, where the temporal clause is marked between brackets.

[1]Phonetically null but syntactically active element

(3) *(Jontxu ikusi zuenean,)* *laster*
Jontxu see aux-COMP.INE, soon
ezagutu zuen.
recognise aux
'When s/he saw Jontxu, s/he recognised him soon.'

(4) *(Indarrean egon den*
force-INE be aux-COMP
bitartean) ez du mugapenik
meanwhile not aux delimitation
izan
be
'While it has been in force, it had no delimitation.'

Non finite verb temporal clauses are formed on the basis of the verbal noun (VN) or participle. After that suffixes are added like the inessive (INE) in *itzultze*-VN#*an*-INE from (5) example. Free elements like *ostean* in (6) can be added after the verb.

(5) *(Etxera itzultzean,)*
Home-ALL come_back-INE,
Annikak makinaz pasatzen
Annika-ERG machine-INS pass
zuen testua.
aux text-ABS
'At coming back home, Annika used to type the text'.

(6) *(Maistrak agindutakoa egin*
Teacher-ERG order-REL.ABS do
ostean,) arratsalde osoa zeukaten
after, afternoon whole had
jolasteko (...)
play-FINAL CLAUSE
'After having done what the teacher ordered, they had all the afternoon to play.'

Contrary to relative clauses, the subordinate verb does not need to be always in the last position, so we can find arguments or adjuncts after it. This canonical word order alteration is difficult too for a rule based chunker, above all if there are more than one element after the verb and no punctuation marks, that could help us by giving a clue.

3 Simplification process and system architecture

In this section we present the simplification process we follow and the architecture of the system (see figure 1) we are implementing to perform the simplification process.

The simplification process illustrates the operations that should be done and the steps we follow in order to produce simple sentences out of long sentences. Before this process is initiated, the readability of the text is analysed. This task is performed by *Idazlanen Autoebaloaziorako Sistema (IAS)*[2] module (Castro-Castro et al., 2008), a system already developed by our group for the auto-evaluation of essays, which discriminates the texts that should continue the process.

Having as input a complex text, following operations are performed:

1. **Splitting:** Make as many new sentences as clauses out of the original. This operation is performed by *Mugak* (Arrieta, 2010).
2. **Reconstruction:** Two operations take place in the split sentences:
   (a) Removing no longer needed morphological features like complementisers and suffixes. Being Basque an agglutinative language we have to remove parts of words and not a whole word.
   (b) Adding new elements like adverbs or paraphrases. The goal is to maintain the meaning. In other words, the features that have been deleted should be replaced by new words. This is included in DAR (Deletion and Addition Rules) module.
3. **Reordering:** Reorder the elements in the new sentences, and ordering the sentences in the text. The set of these rules is included in ReordR (Reordering Rules) module.
4. **Adequation and Correction:** Correct the possible grammar and spelling mistakes, and fix punctuation and capitalisation. The spell checker for Basque Xuxen (Agirre et al., 1992) will carry put this operation.

[2]System for auto-evaluation of essays

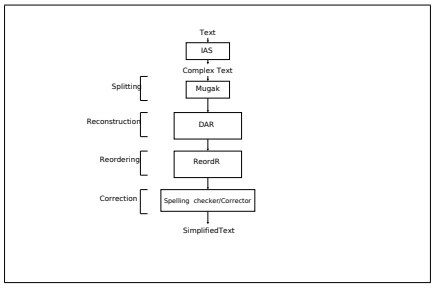

Figure 1: The architecture of system

The work presented in this paper is framed in the splitting operation and at the same time it guides the sentences to the reconstruction operation.

4 Transformation of complex sentences

Our splitting module is based on two stages: first, we apply a grammar that tags the splitting point, that is, the clause boundary is marked, and secondly, we apply an algorithm to make dependency-trees of the clauses out of the original sentence.

4.1 Splitting Point Tagging

The task of splitting point tagging is made by *Mugak* following the Constraint Grammar (CG) (Karlsson et al., 1995) formalism.

Mugak works on the basis of the output produced by several tools implemented in our group: Morpho-syntactic analysis by *Morpheus* (Alegria et al., 2002), lemmatisation and syntactic function identification by *Eustagger* (Aduriz et al., 2003), multi-words items identification (Ezeiza, 2002) (Urizar, 2012) and named entity recognition by *Eihera* (Alegria et al., 2003).

Our work consists on improving the grammar in *Mugak* (Ondarra, 2003) (Aduriz et al., 2006b) by means of adding new rules and adapting older rules based on linguistic knowledge, that lead us to get better results.

In this moment there are 78 rules and 22 of them are especially written for the phenomena we are presenting in this paper. Major improvements have been made this time in the detection of clauses headed by compound verbs and the comma. We have to remark that this is an ongoing work, that is optimised by using new corpora to find new structures and above all to determine the precision in case of non canonical order sentences.

4.2 Splitting algorithm

We have implemented an algorithm to apply several heuristics defined to transform a complex sentence into simple sentences, once the splitting point has been tagged. The usage of this algorithm is to create the dependency-trees of the new sentences. To create this algorithm and to help the following reconstruction step, we have carried out an experiment with sentences in EPEC-DEP (Basque Dependency Treebank) (Aranzabe, 2008) that were sintactically deep tagged, that is PRO[3] and pro[4] elements had a tag.

Let us explain this process by means of an example. Figure 2 shows the tree of the original sentence *Bere zeregina zatituta dagoen alderdia batzea izango dela esan zuen* (S/he said that her/his mission is to unify the political party that is divided).

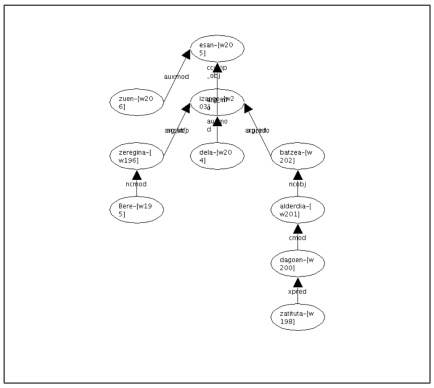

Figure 2: Original sentence: *Bere zeregina zatituta dagoen alderdia batzea izango dela esan zuen*

Having this input our algorithm works as follows:

1. The relative clause *zatituta dagoen* (that is divided) is removed out of the original sentence. This way we get two trees: the main clause *Bere zeregina alderdia batzea izango dela esan zuen.* (S/he said that her/his mission is to unify the political party.) (figure 3) and the relative

[3] see footnote 1

[4] elided arguments (pro-drop)

clause *zatituta dagoen* (that is divided) (figure 4).

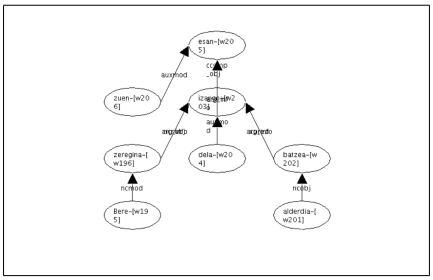

Figure 3: The main clause: *Bere zeregina alderdia batzea izango dela esan zuen*

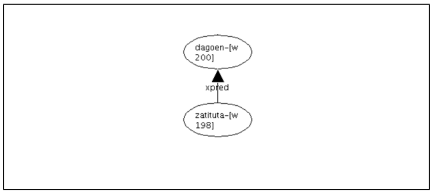

Figure 4: The relative clause: *zatituta dagoen*

2. The PRO antecedent of relative clause *alderdia* (The political party) is included in the new sentence. This way, the sentence *alderdia zatituta dagoen* is formed as shown in the tree of figure 5.

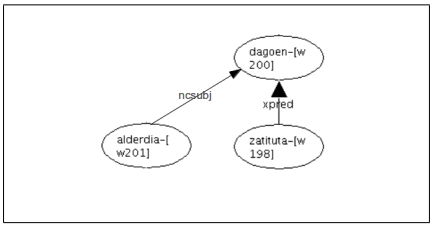

Figure 5: The new simple sentence (relative + antecedent): *alderdia zatituta dagoen*

In the case of adverbial temporal clauses, the adverbial clause is removed in the first step and an adverb will be added in the second step.

This way the reconstruction operation is over in both cases at tree level. That is simple sentences are formed, but they are not grammatically correct. The reconstruction will be over, continuing with this example, by removing the *-(e)n* complementiser of the verb.

5 Evaluation

In this section we evaluate the correctness assessing the splitting point tag and splitting the sentences.

The corpus that has been used to develop and to evaluate the grammar has been EPEC. We divided the corpus in two sets: devel and eval. We used devel for designing the rules of the grammar and eval for automatic evaluation. The latter was previously manually tagged. In table 1 we see the word and sentence number we have used for this task in the development part and the evaluation part of the corpus.

	Devel	**Eval**
Word number	61121	63766
Sentence number	5068	5211
Clause number	18301	18356

Table 1: Word, sentence and clause number in corpus

In table 2 we show the results we obtained by relative clauses and adverbial temporal clauses. The measures that we have used are precision (correctly detected clauses/detected clauses), recall (correctly detected clauses/all clauses) and F-measure (2 * precision * recall / (precision + recall)). Forth column shows the clause number of each structure.

For relative clauses, the results are high. The F-measure for the finite verbs is 0,988 and for the non finite verbs it is 0,992. By analysing the errors the chunker made we concluded that:

- We have a problem with a rule that aims a finite verb temporal clause with free elements structure that can be mixed with relative sentences.
- Another kind of error was due to errors in the PoS tagging.
- Non finite modal verbs structures were not found in the development part.

For temporal clauses, we have to divide the results in two groups: clauses without free elements and clauses with free ele-

	Precision	Recall	F-measure	Clause number
Relative finite verb clauses	0,998	0,978	0,988	547
Relative non finite verb clauses	1	0,985	0,992	335
Temporal finite verb clauses	0,955	0,964	0,960	111
Temporal non finite verb clauses	0,966	0,966	0,966	29
Temporal finite verb clauses + free element(s)	1	0,556	0,714	18
Temporal non finite verb clauses + free element(s)	0,970	0,372	0,538	86

Table 2: Evaluation results of the treated phenomena

ments. The results for the first group are quite high and similar for finite and non finite verbs. The F-measure for temporal finite verb clauses is 0,960 and for the non temporal finite verb clauses is 0,966. We analysed the errors and they are due to canonical word order alteration.

The results for the second group are, however, lower. The F-measure for the temporal finite verb clauses + free element(s) is 0,714 and for the temporal non finite verb clauses + free element(s) is 0,538. The main problem here is that the recall is very low (finite verbs 0,556 and non finite verbs 0,372). Those results are due to:

- The ambiguity of the free elements
- The richness of those structures (all of them were not found in the development part)

Anyway, apart from the problem of the ambiguity the precision we get is high (finite verbs 1 and non finite verbs 0,970).

Since our aim consists on getting accuracy (precision) it is widely achieved, so we consider that we have a basis to continue with the simplification process. This basis is extremely remarkable for relative clauses. The results of the temporal clauses are good. Nevertheless, we should keep on improving the rules, and if possible, getting more structures. It is remarkable too that recall goes down resounding when the clause has free elements, since it is difficult to cover all the possible structures with a corpus. So, defining the clause boundaries is a continuous task we have to keep on working on in order to improve our clause boundary identifier.

6 Conclusion and Future work

In this paper we have focused on the splitting module in our text simplification architecture, since we think that it is important to have a good basis to continue with the simplification process. As we have explained, this module works on two phases: clause boundary detection and splitting point tagging and building simple sentence dependency-trees out of original sentence. The first phase tagging is made by means of *Mugak* a linguistic knowledge based grammar written in the Constraint Grammar formalism and the second phase is carried out by an algorithm based on dependencies-trees as well to create so many sentences out of the clauses in the original sentences. Furthermore, this algorithm introduces the clause in the reconstruction operation.

For this task, we have deeply analysed two diverse structures, namely relative clauses and adverbial temporal clauses. We have explained their different formation and the challenge they suppose.

We have made an evaluation and concluded that we have great basis to continue with the simplification process. Moreover, the algorithm we have implemented introduces the clauses in the reconstruction step fulfilling almost the simplification process in the case of relative sentences. But, on the other hand, the improvements made here to the clause boundary identifier will serve to improve the performance of other tools which use older versions of this identifier, for example, the statistical clause boundary identifier (Arrieta, 2010).

Our next step is actually to keep on working with the syntactic simplification process. For the verb state changing, that is becoming a subordinate verb into a main verb, we plan to use finite state technology tools like FOMA (Hulden, 2009). This tool will be useful as well to implement deletion and addition rules so far defined in (Gonzalez-Dios, 2011).

Acknowledgements

Itziar Gonzalez-Dios's work is funded by a PhD grant from the Basque Government. This research was supported by the Basque Government (IT344-10), and the

Spanish Ministry of Science and Innovation (MICINN, TIN2010-20218). We thank Iñigo Lopez-Gazpio for the support in programming task.

References

Aduriz, Itziar, Izaskun Aldezabal, Iñaki Alegria, Jose Mari Arriola, Arantza Díaz de Ilarraza, Nerea Ezeiza, and Koldo Gojenola. 2003. Finite State Applications for Basque. In *EACL'2003 Workshop on Finite-State Methods in Natural Language Processing.*, pages 3–11.

Aduriz, Itziar, María Jesús Aranzabe, Jose Mari Arriola, Aitziber Atutxa, Arantza Díaz de Ilarraza, Nerea Ezeiza, Koldo Gojenola, Maite Oronoz, Aitor Soroa, and Ruben Urizar. 2006a. Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for automatic processing. *Language and Computers*, 56(1):1–15.

Aduriz, Itziar, Bertol Arrieta, Jose Mari Arriola, Arantza Díaz de Ilarraza, Elixabete Izagirre, and Ainara Ondarra. 2006b. Muga Gramatikaren Optimizazioa. Technical report, UPV/EHU/LSI/TR 9-2006.

Agirre, Eneko, Izaskun Aldezabal, Jone Etxeberria, Mikel Iruskieta, Elixabete Izagirre, Karmele Mendizabal, and Eli Pociello. 2006. A methodology for the joint development of the Basque WordNet and Semcor. In *Proceedings of the 5th International Conference on Language Resources and Evaluations (LREC)*.

Agirre, Eneko, Iñaki Alegria, Xabier Arregi, Xabier Artola, Arantza Díaz de Ilarraza, Montse Maritxalar, Kepa Sarasola, and Miriam Urkia. 1992. Xuxen: A Spelling Checker/Corrector for Basque based in Two-Level Morphology. In *Proceedings of NAACL-ANLP'92*, pages 119–125.

Al-Subaihin, Afnan A. and Hend S. Al-Khalifa. 2011. Al-Baseet: A proposed simplification authoring tool for the Arabic language. In *International Conference on Communications and Information Technology (ICCIT)*, pages 121–125.

Aldezabal, Izaskun, María Jesús Aranzabe, Arantza Díaz de Ilarraza, Ainara Estarrona, Kike Fernandez, and Larraitz Uria. 2010. EPEC-RS: EPEC (Euskararen Prozesamendurako Erreferentzia Corpusa) rol semantikoekin etiketatzeko eskuliburua. Technical report, UPV/EHU/LSI/TR 02-2010.

Alegria, Iñaki, María Jesús Aranzabe, Aitzol Ezeiza, Nerea Ezeiza, and Ruben Urizar. 2002. Robustness and customisation in an analyser/lemmatiser for Basque. In *LREC-2002 Customizing knowledge in NLP applications workshop*, pages 1–6.

Alegria, Iñaki, Nerea Ezeiza, Izaskun Fernandez, and Ruben Urizar. 2003. Named Entity Recognition and Classification for texts in Basque. In *II Jornadas de Tratamiento y Recuperación de Información, JOTRI*.

Aranzabe, María Jesús. 2008. *Dependentzia-ereduan oinarritutako baliabide sintaktikoak: zuhaitz-bankua eta gramatika konputazionala*. Ph.D. thesis, Euskal Filologia Saila (UPV/EHU).

Arrieta, Bertol. 2010. *Azaleko sintaxiaren tratamendua ikasketa automatikoko tekniken bidez: euskarako kateen eta perpausen identifikazioa eta bere erabilera koma-zuzentzaile batean*. Ph.D. thesis, Informatika Fakultatea (UPV-EHU).

Bernhard, Delphine, Louis De Viron, Véronique Moriceau, and Xavier Tannier. 2012. Question Generation for French: Collating Parsers and Paraphrasing Questions. *Dialogue and Discourse*, 3(2):43–74.

Burstein, Jill. 2009. Opportunities for Natural Language Processing Research in Education. In *Computational Linguistics and Intelligent Text Processing*. Springer Berlin / Heidelberg.

Candido, Jr., Arnaldo, Erick Maziero, Caroline Gasperin, Thiago A. S. Pardo, Lucia Specia, and Sandra M. Aluisio. 2009. Supporting the adaptation of texts for poor literacy readers: a text simplification editor for Brazilian Portuguese. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, EdAppsNLP '09, pages 34–42. ACL.

Carroll, John, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying Text for

Language-Impaired Readers. In *9th Conference of the European Chapter of the Association for Computational Linguistics*.

Castro-Castro, Daniel, Rocío Lannes-Losada, Montse Maritxalar, Ianire Niebla, Celia Pérez-Marqués, Nancy C. Alamo-Suarez, and Aurora Pons-Porrata. 2008. A Multilingual Application for Automated Essay Scoring. In *Lecture Notes in Advances in Artificial Intelligence - LNAI 5290 - IBERAMIA*, pages 243–251. Springer New York.

Ezeiza, Nerea. 2002. *Corpusak ustiatzeko tresna linguistikoak. Euskararen etiketatzaile morfosintaktiko sendo eta malgua.* Ph.D. thesis, Informatika Fakultatea, UPV-EHU.

Gonzalez-Dios, Itziar. 2011. Euskarazko egitura sintaktikoen azterketa testuen sinplifikazio automatikorako: Aposizioak, erlatibozko perpausak eta denborazko perpausak. Master's thesis, University of the Basque Country.

Hulden, Mans. 2009. Foma: a Finite-State Compiler and Library. In *EACL (Demos)'09*, pages 29–32.

Inui, Kentaro, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, and Tomoya Iwakura. 2003. Text simplification for reading assistance: a project note. In *Proceedings of the second international workshop on Paraphrasing-Volume 16*, pages 9–16. ACL.

Iruskieta, Mikel, Arantza Díaz de Ilarraza, and Mikel Lersundi. 2011. Unidad discursiva y relaciones retóricas: un estudio acerca de las unidades de discurso en el etiquetado de un corpus en euskera. *Procesamiento del Lenguaje Natural*, (47).

Jonnalagadda, Siddhartha and Graciela Gonzalez. 2010. Sentence simplification aids protein-protein interaction extraction. *Arxiv preprint arXiv:1001.4273*.

Karlsson, Fred, Atro Voutilainen, Juha Heikkila, and Atro Anttila. 1995. *Constraint Grammar, A Language-independent System for Parsing Unrestricted Text*. Mouton de Gruyter.

Ondarra, Ainara. 2003. Murriztapen Gramatikaren sintaxia. EUSMG optimizatzen. Esaldi-mugak. Master's thesis, Euskal Herriko Unibertsitatea.

Petersen, Sarah E. and Mari Ostendorf. 2007. Text Simplification for Language Learners: A Corpus Analysis. *Electrical Engineering*, (SLaTE):69–72.

Poornima, C., V. Dhanalakshmi, K.M. Anand, and KP Soman. 2011. Rule based Sentence Simplification for English to Tamil Machine Translation System. *International Journal of Computer Applications*, 25(8):38–42.

Rybing, Jonas, Christian Smith, and Annika Silvervarg. 2010. Towards a Rule Based System for Automatic Simplification of texts. In *The Third Swedish Language Technology Conference (SLTC 2010)*.

Saggion, Horacio, Elena Gómez-Martínez, Esteban Etayo, Alberto Anula, and Lorena Bourg. 2011. Text Simplification in Simplext: Making Text More Accessible. *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*.

Seretan, Violeta. 2012. Acquisition of syntactic simplification rules for french. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA).

Siddharthan, Advaith. 2006. Syntactic simplification and text cohesion. *Research on Language & Computation*, 4(1):77–109.

Siddharthan, Advaith. 2011. Text Simplification using Typed Dependencies: A Comparison of the Robustness of Different Generation Strategies. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 2–11. ACL.

Soraluze, Ander, Olatz Arregi, Xabier Arregi, Klara Ceberio, and Arantza Díaz de Ilarraza. 2012. Mention Detection: First Steps in the Development of a Basque Coreference Resolution System. In *Proceedings of KONVENS 2012 (Main track: oral presentations)*, pages 128–163.

Urizar, Ruben. 2012. *Euskal lokuzioen tratamendu konputazionala*. Ph.D. thesis, UPV-EHU.

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*, pages 1353–1361.

Detecting Apposition for Text Simplification in Basque

Itziar Gonzalez-Dios*, María Jesús Aranzabe,
Arantza Díaz de Ilarraza, and Ander Soraluze**

IXA NLP Group, University of the Basque Country (UPV/EHU),
Manuel Lardizabal 1 48014 Donostia
{itziar.gonzalezd,maxux.aranzabe,a.diazdeilarraza,ander.soraluze}@ehu.es
http://ixa.si.ehu.es/Ixa

Abstract. In this paper we have performed a study on Apposition in Basque and we have developed a tool to identify and to detect automatically these structures. In fact, it is necessary to detect and to code this structures for advanced NLP applications. In our case, we plan to use the Apposition Detector in our Automatic Text Simplification system. This Detector applies a grammar that has been created using the Constraint Grammar formalism. The grammar is based, among others, on morphological features and linguistic information obtained by a named entity recogniser. We present the evaluation of that grammar and moreover, based on a study on errors, we propose a method to improve the results. We also use a Mention Detection System and we combine our results with those obtained by the Mention Detector to improve the performance.

Keywords: Apposition Detector, Basque, Automatic Text Simplification, Mention Detection.

1 Introduction

Automatic Text Simplification (TS) is a Natural Language Processing (NLP) task whose aim is to simplify texts automatically, keeping the meaning of original text, or at least avoiding information loss. TS is a necessary research line in NLP since the texts which are simplified are easier to process both for people and advanced NLP applications.

TS systems have already been proposed for people with disabilities [1], illiterate [2] or people who learn foreign languages [3] [4] among others. There are TS systems for advanced applications such us machine translation [5], Q&A systems [6], information extraction systems [7], and so on.

Our main motivation for TS is that long sentences cause problems in advanced applications like machine translation [8]. Apposition is a phenomenon that increases the length of the sentences and it has been reported in the context of TS as

* Itziar Gonzalez-Dios's work is funded by a PhD grant from the Basque Government.
** Ander Soraluze's work is funded by PhD grant from Euskara Errektoreordetza.

A. Gelbukh (Ed.): CICLing 2013, Part II, LNCS 7817, pp. 513–524, 2013.

a complex phenomenon and rules to simplify these structures have been studied e.g. in [9] and [10] and for Basque in [11]. The information that an appositional phrase contains is not syntactically necessary and therefore it can be taken out of the sentence. This will mean the loss of some information, unless we create a new sentence out of the apposition. So if we remove apposition out of the sentence and create shorter sentences, for example, the task of machine translation will be more affordable.

In NLP, apposition detection has been mainly studied in the context of its integration in other general tools. However, there are tools that identify apposition explicitly [12] by means of machine learning techniques. Other techniques that have been used to detect apposition are heuristics [13] or full parse information [14]. In [15] appositive detection is applied as preprocess of a mention detection system and they use patterns to identify these structures. In [16] they use sequence mining to detect linguistic patterns in French like appositive qualifying phrases.

There are two tools in Basque that can be useful to detect Apposition. The first is a named entity recogniser and classifier, *Eihera* [17] and the second is the combination of the rule based (*IXAti* [18]) and the statistical-based (*ML-IXAti* [19]) shallow syntactic parsers for Basque. These tools consider apposition inside a noun phrase (restrictive) as a chunk, and apposition, that is expressed by noun phrase as appositive (non-restrictive), as more than an independent chunk. Since there is no explicit way to mark the apposition, we need a special tool to detect them.

So, in this paper we present a rule based Apposition Detector, based on linguistic knowledge, that is able to identify these structures and classify them according to their type. The output of this tool is human friendly, but it can be easily coded for machines as well. Although the first use of this Detector is TS, the Apposition Detector can be useful for other NLP advanced applications like mention detection, coreference resolution, parsing, textual entailment, text summarisation, Q&S systems, information extraction, event extraction, opinion mining etc. In the evaluation, we obtain 0.80 in F-measure. However, we analyse the errors and to improve the results, we use a Mention Detection System [20].

This paper is structured as follows: in section 2 we present the apposition types in Basque language. In section 3 the framework and the formalism of the Apposition Detector is explained. In section 4 we show the evaluation results. To improve this result we show in section 5 the experiments we carried out using the Mention Detector. In section 6 we describe how this tool will be used for Automatic Text Simplification and finally, in section 7 we expose the conclusion and the future work.

2 Apposition in Basque

Basque is Pre-Indo-European language and differs considerably in grammar from the languages spoken in surrounding regions. It is, indeed, an agglutinative head-final pro-drop isolated language whose case system is ergative-absolutive. Basque displays a rich inflectional morphology. Basque is still undergoing the normalisation process, and in charge of that, among others, there is *Euskaltzaindia* (Royal Academy of the Basque Language).

Apposition detection grammar has been built according to *Euskaltzaindia* [21]. As regulated, there are two types of apposition in Basque:

- **First type (restrictive):** Apposition that occurs inside a noun phrase. There are two ways to realise this type: a) example (1), the named entity *Luis Uranga* precedes the common name *presidenteak* (henceforth, type 1A):

(1) *Luis Uranga presidenteak (...)*
Luis Uranga president_the
'The president Luis Uranga (...)'

or b) example (2), the common name *presidente* precedes the named entity *Luis Uranga* (henceforth, type 1B):

(2) *Errealeko presidente Luis Uranga (...)*
Real_Sociedad_of president Luis Uranga
'The president of Real Sociedad Luis Uranga (...)'

- **Second type (non-restrictive):** A noun phrase as appositive like (3)[1]:

(3) *Jakinduria hori, guretzat harrapezina dena, (...)*
Wisdom that, us_for unattainable is_which_the,
'That wisdom, that is unattainable for us, (...) '

It is possible as well to combine both types (4):

(4) *Simon Peres laborista, Israelgo lehen ministro izana,*
Shimon Peres Labour_the, Israel_of Prime Minister have_been_the
'Labour Shimon Peres, the former Prime Minister of Israel, (...)'

and to merge the both structures (1A and 1B), example (5):

(5) *Vatikanoko Estatuekiko Harremanetarako idazkari Jean Louis Tauran*
Vatican_of states_with relations_for secretary Jean Louis Tauran
artzapezpikuak (...)
archbishop_the
'The archbishop Jean Louis Tauran, Secretary for Relations with States of The Vatican, (...)'

Parenthetical structures are not considered as apposition by *Euskaltzaindia*, since there is no agreement. However, some kind of parenthetical structures follow the same pattern as apposition in the simplification rules [11], so we have included rules to treat them in this grammar. For non simplification uses, these rules can be omitted. In (6) we see an example of a parenthetical structure the grammar covers.

[1] Notice that the equivalent translation is a relative clause.

(6) *Durangon* *(Bizkaia)*
Durango_in (Biscay)
'in Durango (Biscay)'

These are the target structures for our Apposition Detector. Each structure is given a tag, so they are classified.

If we applied only our shallow syntactic parser *IXAti* [18], type one apposition (both 1A and 1B) will be considered as a chunk, which is correct and valid for shallow parsing. But for some tasks like Automatic Text Simplification they should be distinguished. Apposition type two is considered by *IXAti* as more than one chunk. In both cases there is no explicit tag to express the appositional relation. This way Apposition Detector accomplishes this tagging task before the chunker *IXAti* is applied.

3 Architecture of the Apposition Detector

In this section we explain how our Apposition Detector works. Having as input a text, we perform the following analysis before we apply the Apposition Detector:

- **Morpho-syntactic analysis:** *Morpheus* [22] makes word segmentation and part of speech tagging. Syntactic function identification is made by *Constraint Grammar* formalism [23].
- **Lemmatisation and syntactic function identification:** *Eustagger* [24] resolves the ambiguity caused at the previous phase.
- **Multi-words items identification:** The aim is to determine which items of two or more words are always next to each other [25] [26].
- **Named entity recognition:** *Eihera* [17] identifies and classifies named-entities in the text (person, organisation, location).

To detect the apposition we have written a grammar following Constraint Grammar formalism [23]. The linguistic features we have used to write the rules in grammar are category, subcategory, and named entity tags.

Our detection system works in two phases: first, a grammar tags the named entities that are candidates to be a part of an apposition and secondly, based on the previous tags another grammar tags the second part of the apposition, if it fulfils the conditions of being a real apposition. The phrase with both tags is an apposition. There are 37 rules for the first phase, and 21 rules for the second phase. The rules are classified according to the entity type as well.

Each structure presented in section 2 has a tag (Table 1). This is the way apposition classification is made. This classification is valid, for example to know what kind of structures are used frequently or which rule should be applied for Text Simplification.

Once the apposition has been tagged we apply the rule based chunker *IXAti* [18] and *ML-IXAti* [19], which identifies chunks and clauses by combining rule-based grammars and machine learning techniques, exactly the version implemented in

Table 1. Tags applied by the grammar

Type	1 appositional phrase	2 appositional phrases
1A	]APOS1	[APOS2
1B	]APOS1_KONTRA	[APOS2_KONTRA
2	]APOS1SINT	[APOS2SINT
Parenthetical structures	]APOS1_EGON	[APOS2_EGON

[20] to get the both appositional phrases. The algorithm is the following: the first appositional phrase begins where the chunker has tagged the phrase begin and it finishes with the word that has the first tag by our grammar. The second appositional phrase is formed by the word(s) between the first tag and second tag.

Let see this process with example (1), *Luis Uranga presidenteak*. The first rule (Figure 1) tags *]APOS1* and targets the end boundary of a named entity classified as person *Luis Uranga*, that is composed only by two words[2] and that is in the context of an apposition, in this example *Uranga*.

```
MAP (]APOS1) TARGET (ENTI_BUK_PER) IF (-1 ENTI_HAS_PER)  (0 NEXT_KM)
                                        (1  IZE + ARR) (NOT 1 NEXT_KM);
```

Fig. 1. CG rule to tag a candidate appositional phrase

The second rule (Figure 2) tags *[APOS2* and targets a common name, if previously an apposition candidate has been tagged (i.d. there is previously *]APOS1* tag), that is not followed by a adjective, in this example *presidenteak*.

```
MAP ([APOS2) TARGET  (IZE) IF (0C ARR) (NOT 0 PUNT_KARDI OR LEKU )
                                        (NOT 0 HM OR DM) (-1 APO) (NOT 1 ADJ) ;
```

Fig. 2. CG rule to tag second appositional phrase and confirm the apposition

Taking into account the information of *IXAti* and *ML-IXAti* and the previously mentioned tags, the whole appositional phrases are *Luis Uranga* and *presidenteak*. In figure 3 we see the output of example (1) in text version (human-friendly).

4 Evaluation and Error Analysis

The corpus that has been used to develop and to evaluate the grammar has been EPEC (*Euskararen Prozesamendurako Errreferentzia Corpusa*-Reference Corpus for the Processing of Basque) [27]. EPEC Corpus is interesting for this task since

[2] *ENTI_HAS_PER* and *ENTI_HAS_PER* tag the beginning and the ending of a named entity, and the other tags express morphological features.

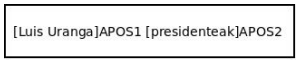

Fig. 3. Output of Apposition Detector in Text Version

it compiles text from newspapers, where apposition is a normal feature. In the first column of table 2 we see the quantities of the apposition found in the evaluation part of the corpus, in general and classified according to their type. To evaluate this grammar we have created a gold standard, where the apposition has been manually tagged.

In table 2 we also show the results[3] obtained by Apposition Detection and the quantities that are in the corpus. We show the results according to the apposition type as well.

Table 2. Evaluation results of the Apposition Detection

	Quantities	**Precision**	**Recall**	**F measure**
All types	336	0.87	0.74	0,80
1A type	286	0.90	0.62	0.73
1B type	30	0.85	0.73	0.79
2 type	9	1	0.44	0.62
Parenthetical structures	11	1	0.64	0.78

Except for a case, appositions were classified correctly. It was the case of a parenthetical structure that was considered as 1A type.

These results have been analysed qualitatively and we found out following errors and missing structures:

- Due to errors in named entity detection, rules were not applied or misapplied
- Apposition was detected, but a tag was not in the correct place. For example, the tag was in the substantive, when it should be in the adjective
- Complex appositional phrases that were already dismissed in development phase because they made a lot of errors for a correct one, like coordination in appositional phrases.

5 Improving Apposition Detection Using a Mention Detector

By analysing the results (section 4) we noticed that in some cases Apposition Detector has tagged the candidate (first tag) but due to the complexity of the

[3] Precision = correctly detected apposition/detected apposition; Recall = correctly detected apposition/all apposition; F-measure = 2 * precision * recall / (precision + recall).

appositional phrases, the tag for the second appositional phrase has been omitted (rule failed or dismissed rule) and in other cases nothing was retrieved. Those were considered as errors. This is the case of example (7).

(7) *Manuel Contreras Inteligentzia Nazionaleko Zuzendaritzako (DINA)*
Manuel Contreras Intelligence national_of direction_of (DINA)
buruzagi ohiak
head former
'Manuel Contreras, former head of the National Intelligence Directorate (DINA), '

In order to get this complex structures (e.g, (7)), we have carried out an experiment with the Mention Detector [20]. This system identifies mentions that are potential candidates to be part of coreference chains in Basque written texts. The aim of this experiment is to see if the Mention Detector can help to improve the results, without making changes in the system. In other words, we want to combine the output of the grammar and the output of the Mention Detector to see if we can get the discarded instances. This process is illustrated in figure 4.

We have formed two hypotheses that we explain next and developed a technique for each one. To test these hypotheses we made a subcorpus with the errors the grammar made, that is, we used the phrases which the first candidate was tagged,

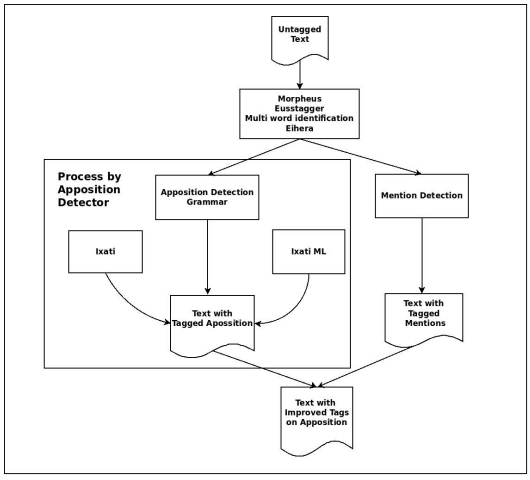

Fig. 4. Architecture of Apposition Detector and Improvement Process through Mention Detection

but the second one was omitted. There are 25 instances in this subcorpus. We only used 1A type, because other type quantities were insignificant.

Taking into account that this subcorpus was formed by the structures the grammar failed, we form **the first hypothesis**: If inside a mention is an appositional phrase candidate according to the grammar, it may be an apposition. So, the algorithm we implemented is next: if a mention has first tag inside (candidate), the rest of the mention is given second tag, and therefore considered as an apposition. Out of 25 instances 5 were were retrieved correctly.

To continue improving the results and taking into account the results of the first hypothesis, we formed **the second hypothesis**: if a mention is an appositional phrase candidate, the following mention in text should be its appositive. The technique we use to track is the mention identification number. If the candidate mention has identification number 1 in text, mention with identification number 2 should be its appositive. Applying this method, the 13 instances of the 20 left were correctly retrieved. Three instances more were retrieved, but as the whole appositional phrase was not correct, they were consider as errors.

So, we concluded that Mention Detection, without having been tuned, can improve the detection of apposition, retrieving 18 instances out of 25 and obtaining following results in the error subcorpus (Table 3). This approach using the Mention Detector is above all helpful to retrieve the cases which grammarians had discarded the rule due to error increasing.

Table 3. Evaluation Results of Error Detection through Mention Detection

	Quantities	**Precision**	**Recall**	**F measure**
A1 type	25	0.86	0.72	0.78

It is important to mention that these algorithms have been tested with errors. To prove both hypotheses in a normal corpora, we think that the Mention Detector should be tuned. That is, instead of applying the second grammar, if we want to use only the Mention Detector, we should make severe changes. These algorithms should be more accurate, since not all the candidates form apposition. That is, we should incorporate the information of the second grammar adequated to the rules of the mention detection system, so that instances like named entities referring to a place followed by cardinal directions like *Londres mendebalean* (in West London) or followed by complex postpositions like *Erroma inguruan* (in the surroundings of Rome) are not retrieved. Anyway, we could not get rid of the grammar, since there are instances that Mention Detector would not retrieve.

6 Use of Apposition Detector in Text Simplification

The Apposition Detector presented here will be a part of the framework in our TS system, together with *Mugak* [28], the clause identifier. Based on its output apposition follows the simplification process [29], that will be explained by means of

example (8): *Jasser Arafat buru palestinarra Egiptoko presidente Hosni Mubarak-ekin bildu zen atzo Kairon* (Palestinian Chairman Jasser Arafat met President of Egypt Hosni Mubarak yesterday in Cairo).

1. **Splitting:** First apposition is detected: there are two in sentence (9): [*Jasser Arafat buru palestinarra*] (Palestinian Chairman Jasser Arafat)and [*Egiptoko presidente Hosni Mubarak-ekin*] (President of Egypt Hosni Mubarak). Secondly, a chunk is created for each appositional phrase in each apposition (figure 5). This is the task that the Apposition Detector presented in section 3 carries out.

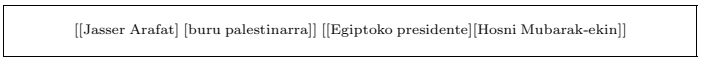

Fig. 5. Appositional phrases in sentence (8)

2. **Reconstruction:**
   (a) Removing: The chunks with the second tag (second appositional phrase) will be removed from the original sentence, obtaining following output: [*Jasser Arafat Hosni Mubarak-ekin bildu zen atzo Kairon*] (Jasser Arafat met Hosni Mubarak yesterday in Cairo). If a chunk has a suffix like *-ekin* (with) in *Hosni Mubarak-ekin* it should be removed.
   (b) Adding: Chunks with both tags will be added together with the copula, in these examples *da* (is), to form simple sentences: The absolutive suffix *-a* should be added in the phrase *Egiptoko presidentea* (the President of Egypt).

   This is the output of this operation: [*Jasser Arafat buru palestinarra da*] (Jasser Arafat is a Palestinian chairman) and [*Egiptoko presidentea Hosni Mubarak da*] (The president of Egypt is Hosni Mubarak). In this operation sentences have been created, but the simplification process is not yet fulfilled.
3. **Reordering:**
   (a) Internal word reordering in sentence: First the internal order will be checked: the order of former original sentence is kept untouched, the new sentences follow this rule pattern: Chunk with first tag (SUBJ), chunk with second tag (PRED), copula in present tense, 3 person, singular or plural depending on the subject. The first apposition follows the pattern of the rule, so it is left untouched but the second should be reordered to follow that pattern[4]: [*Hosni Mubarak Egiptoko presidentea da*] (Hosni Mubarak is the president of Egypt).
   (b) Sentence reordering in text: First, the former original sentence; then, new simple sentences following the order they appear in the original sentence.

[4] Before reordering this sentence was already grammatically correct, since Basque is a free word order language. But according to the simplification rule, the order should change.

4. **Correction:** There is no grammatical error to correct but sentences should be punctuated. This will be the final output: [*Jasser Arafat Hosni Mubarak-ekin bildu zen atzo Kairon. Jasser Arafat buru palestinarra da. Hosni Mubarak Egiptoko presidentea da.*] (Jasser Arafat met Hosni Mubarak yesterday in Cairo. Jasser Arafat is a Palestinian chairman. Hosni Mubarak is the President of Egypt.).

Following this process we have got shorter sentences which are useful for advanced applications like machine translation. Anyway, as simplification rules can be tuned according to the target audience, another option is to make a coordinate sentence with *eta* (and) to unify the new simple sentences. This will be the final output: *Jasser Arafat Hosni Mubarak-ekin bildu zen atzo Kairon. Jasser Arafat buru palestinarra da eta Hosni Mubarak Egiptoko presidentea da.* (Jasser Arafat met Hosni Mubarak yesterday in Cairo. Jasser Arafat is a Palestinian chairman and Hosni Mubarak is the President of Egypt.).

7 Conclusion and Future Work

In this paper we have presented an Apposition Detector based on linguistic knowledge. Moreover, it is able to classify the apposition in corpora according to their type and structure, which is helpful for linguistic analysis and research on apposition.

We have evaluated this tool and looking at the results (F-measure 0.80), we realised that they could be improved. So we have made an experiment on errors with another tool, the Mention Detector. We have formed two hypotheses and created to techniques to combine the output of the grammar and the output of the Mention Detector. This way, the instances that were not covered by the grammar were retrieved (F-Measure 0.78), without having changed the Mention Detection system.

We have explained as well how we are going to use the output of the Mention Detector in Automatic Text Simplification by means of an example. Performing the syntactic simplification process, we get shorter sentences that are easier to process for NLP advanced applications such us machine translation.

Although the first use of the Apposition Detector is Automatic Text Simplification, it can be used for other tasks like coreference resolution, information extraction, lexicon elaboration or text summarisation. Indeed, we plan to implement this Detector to improve the mention detection system and in the coreference resolution system.

Acknowledgments. Itziar Gonzalez-Dios's work is funded by a PhD grant from the Basque Government and Ander Soraluze's work is funded by PhD grant from Euskara Errektoreordetza, the University of the Basque Country (UPV/EHU). This research was also supported by the the Basque Government (IT344-10).

References

1. Carroll, J., Minnen, G., Pearce, D., Canning, Y., Devlin, S., Tait, J.: Simplifying Text for Language-Impaired Readers. In: 9th Conference of the European Chapter of the Association for Computational Linguistics (1999)
2. Candido Jr, A., Maziero, E., Gasperin, C., Pardo, T.A.S., Specia, L., Aluisio, S.M.: Supporting the adaptation of texts for poor literacy readers: a text simplification editor for Brazilian Portuguese. In: Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications. EdAppsNLP 2009, pp. 34–42. Association for Computational Linguistics, Stroudsburg (2009)
3. Petersen, S.E., Ostendorf, M.: Text Simplification for Language Learners: A Corpus Analysis. In: Electrical Engineering (SLaTE), pp. 69–72 (2007)
4. Burstein, J.: Opportunities for Natural Language Processing Research in Education. In: Gelbukh, A. (ed.) CICLing 2009. LNCS, vol. 5449, pp. 6–27. Springer, Heidelberg (2009)
5. Poornima, C., Dhanalakshmi, V., Anand, K., Soman, K.: Rule based Sentence Simplification for English to Tamil Machine Translation System. International Journal of Computer Applications 25(8), 38–42 (2011)
6. Bernhard, D., De Viron, L., Moriceau, V., Tannier, X.: Question Generation for French: Collating Parsers and Paraphrasing Questions. Dialogue and Discourse 3(2), 43–74 (2012)
7. Jonnalagadda, S., Gonzalez, G.: Sentence simplification aids protein-protein interaction extraction. Arxiv preprint arXiv:1001.4273 (2010)
8. Labaka, G.: EUSMT: Incorporating Linguistic Information into SMT for a Morphologically Rich Language. Its use in SMT-RBMT-EBMT hybridation. PhD thesis, UPV-EHU (2010)
9. Siddharthan, A.: Syntactic simplification and text cohesion. Research on Language & Computation 4(1), 77–109 (2006)
10. Specia, L., Aluisio, S.M., Pardo, T.A.: Manual de Simplificação Sintática para o Português. Technical Report NILC-TR-08-06, So Carlos-SP (2008)
11. Gonzalez-Dios, I.: Euskarazko egitura sintaktikoen azterketa testuen sinplifikazio automatikorako: Aposizioak, erlatibozko perpausak eta denborazko perpausak. Master's thesis, University of the Basque Country (September 2011)
12. Freitas, M.C., Duarte, J.C., Santos, C.N., Milidiú, R.L., Rentería, R.P., Quental, V.: A machine learning approach to the identification of appositives. In: Sichman, J.S., Coelho, H., Rezende, S.O. (eds.) IBERAMIA 2006 and SBIA 2006. LNCS (LNAI), vol. 4140, pp. 309–318. Springer, Heidelberg (2006)
13. Phillips, W., Riloff, E.: Exploiting strong syntactic heuristics and co-training to learn semantic lexicons. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 125–132. Association for Computational Linguistics (2002)
14. Roth, D., Sammons, M.: Semantic and logical inference model for textual entailment. In: Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, pp. 107–112. Association for Computational Linguistics (2007)
15. Kummerfeld, J.K., Bansal, M., Burkett, D., Klein, D.: Mention detection: heuristics for the OntoNotes annotations. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. CONLL Shared Task 2011, pp. 102–106. ACL, Stroudsburg (2011)
16. Béchet, N., Cellier, P., Charnois, T., Crémilleux, B.: Discovering linguistic patterns using sequence mining. In: Gelbukh, A. (ed.) CICLing 2012, Part I. LNCS, vol. 7181, pp. 154–165. Springer, Heidelberg (2012)

17. Fernandez Gonzalez, I.: Euskarazko Entitate-Izenak: identifikazioa, sailkapena, itzulpena eta desanbiguazioa. PhD thesis, UPV-EHU (2012)
18. Aduriz, I., Aranzabe, M.J., Arriola, J.M., de Ilarraza, A.D., Gojenola, K., Oronoz, M., Uria, L.: A cascaded syntactic analyser for basque. In: Gelbukh, A. (ed.) CICLing 2004. LNCS, vol. 2945, pp. 124–134. Springer, Heidelberg (2004)
19. Arrieta, B.: Azaleko sintaxiaren tratamendua ikasketa automatikoko tekniken bidez: euskarako kateen eta perpausen identifikazioa eta bere erabilera komazuzentzaile batean. PhD thesis, UPV-EHU (2010)
20. Soraluze, A., Arregi, O., Arregi, X., Ceberio, K., Díaz de Ilarraza, A.: Mention Detection: First Steps in the Development of a Basque Coreference Resolution System. In: Proceedings of KONVENS 2012, pp. 128–163 (2012)
21. Euskaltzaindia: Euskal gramatika laburra: perpaus bakuna. Euskaltzaindia (2002)
22. Alegria, I., Aranzabe, M.J., Ezeiza, A., Ezeiza, N., Urizar, R.: Robustness and customisation in an analyser/lemmatiser for Basque. In: LREC-2002 Customizing Knowledge in NLP Applications Workshop, pp. 1–6 (2002)
23. Karlsson, F., Voutilainen, A., Heikkila, J., Anttila, A.: Constraint Grammar, A Language-independent System for Parsing Unrestricted Text. Mouton de Gruyter (1995)
24. Aduriz, I., Aldezabal, I., Naki Alegria, I., Arriola, J.M., de Ilarraza, A.D., Ezeiza, N., Gojenola, K.: Finite State Applications for Basque. In: EACL 2003 Workshop on Finite-State Methods in Natural Language Processing, pp. 3–11 (2003)
25. Ezeiza, N.: Corpusak ustiatzeko tresna linguistikoak. Euskararen etiketatzaile morfosintaktiko sendo eta malgua. PhD thesis, UPV-EHU (2002)
26. Urizar, R.: Euskal lokuzioen tratamendu konputazionala. PhD thesis, UPV-EHU (2012)
27. Aduriz, I., Aranzabe, M.J., Arriola, J.M., Atutxa, A., Díaz de Ilarraza, A., Ezeiza, N., Gojenola, K., Oronoz, M., Soroa, A., Urizar, R.: A corpus of written Basque tagged at morphological and syntactic levels for automatic processing. In: Methodology and Steps Towards the Construction of EPEC, vol. 56, pp. 1–15. Rodopi (2006)
28. Aranzabe, M.J., Díaz de Ilarraza, A., Gonzalez-Dios, I.: Transforming Complex Sentences using Dependency Trees for Automatic Text Simplification in Basque (manuscript)
29. Aranzabe, M.J., Díaz de Ilarraza, A., Gonzalez-Dios, I.: First Approach to Automatic Text Simplification in Basque. In: Rello, L., Saggion, H. (eds.) Proceedings of the Natural Language Processing for Improving Textual Accessibility (NLP4ITA) Workshop (LREC 2012), Istanbul, Turkey, pp. 1–8 (2012)

6

Automatic Text Simplification: the Proposal of the *EuTS* System

In this chapter we present the proposal of the *EuTS* system, (*Euskarazko Testuen Sinplifikatzailea*) [Simplifier of Basque Texts]. *EuTS* performs two types of simplification: the syntactic substitution simplification and syntactic simplification. The architecture of the system is presented in Figure 6.1 and each module of them performs an operation presented in Chapter 5. In addition to the architecture of the system, we have also presented as case study the simplification of the parenthetical structures that contain biographical information, performed by the tool *Biografix*.

- **Syntactic substitution simplification:**
  - **Shallow syntactic substitutions -> SintSubs module:** this module has also been presented in the paper *Simplifying Basque Texts: the Shallow Syntactic Substitution Simplification* (Gonzalez-Dios et al., 2015a).
- **Syntactic simplification:** in what follows we present the operations of the simplification process together with the modules that perform those operations. These modules have also been presented in the Section 5 of the paper *First Approach to Automatic Text Simplification in Basque* (Aranzabe et al., 2012a).
  1. **Splitting -> Mugak module:** in this module the sentences are split into clauses; appositions into appositives and parenthet-

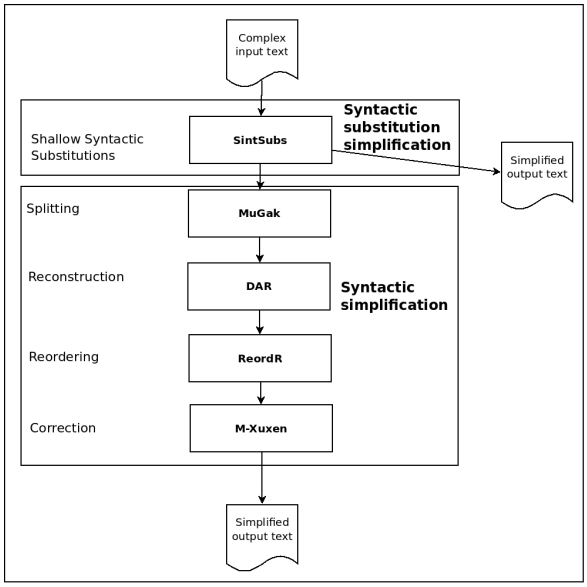

Figure 6.1 – The architecture of *EuTS* system

ical structures into clauses and parentheticals and postpositional structures into postpositional phrases and clauses. Before the splitting, the minimum length will be checked based on the output of *Ixati* (counting the chunks).

To perform the splitting this module will base on the output of *Mugak* (clause splitting), apposition detector *Aposizioak* (splitting of the appositions), the postposition recognition of *Ixati* (splitting of the postpositional phrases) and *Biografix* (splitting of the parenthetical structures).

2. **Reconstruction -> DAR (deletion and addition rules) module:** in this module the syntactic simplification rules obtained in the corpus analysis will be implemented. These rules are based on morphological features and are presented in Appendix B. Exactly, the relation marks that should be removed are compiled in the Relation_Marks_List and the added elements in the Added_Elements_List[1]. This module will based on the output of *Morfeus*

[1]The first element in the list is the default added element and the following are the alternative added elements.

and *Eustagger* to perform the removing. For example, in Table 6.1 we present the relation marks that should be removed in the case of the finite subordinate clauses.

Phenomena	**Clause type**	**Morphological features**	**Others to remove**	**Elements to remove in the syntactic functions**
Relative	Common	ERLT		@+JADNAG_MP_IZLG>
	Zein	ZHG	*zein, non*	@+JADNAG_MP_IZLG>
	Zein	KAUS	*zein, non*	@+JADNAG_MP_IZLG>
Noun clause	Completive	KONPL		@+JADNAG_MP_SUBJ
	Completive	KONPL		@+JADNAG_MP_PRED
	Completive	KONPL		@+JADNAG_MP_OBJ
	Indirect question	ZHG	*ea*, wh-words	@+JADNAG_MP_OBJ
	Indirect question	ZHG	*ea*, wh-words	@+JADNAG_MP_SUBJ
Adverbial	Temporal	DENB		@+JADNAG_MP_ADLG
	Temporal	MOD/DENB		@+JADNAG_MP_ADLG
	Temporal	ZHG	*gehienetan, aldiro, bezain laster, bezain ber...*	@+JADNAG_MP_OBJ
	Temporal	ERLT	*guztietan..*	@+JADNAG_MP_IZLG>
	Causal	KAUS		@+JADNAG_MP_ADLG
	Concessive	MOS	*nahiz eta, a-rren*	@+JADNAG_MP_ADLG
	Concessive	BALD	*ere*	@+JADNAG_MP_ADLG
	Modal	MOD/DENB		@+JADNAG_MP_ADLG
	Modal	MOS	*bezala, moduan...*	@+JADNAG_MP_ADLG
	Purpose	HELB		@+JADNAG_MP_ADLG
	Conditional	BALD	*baldin*	@+JADNAG_MP_ADLG

Table 6.1 – Features to remove from the finite subordinate clauses

3. **Reordering -> ReordR module:** two kinds of reordering will take place: reordering new simplified sentences in text and reordering the elements inside the sentence. The former is based in our corpus analysis and the latter will keep so far (until we get more information from neurolinguistic studies) as in the original sentence. The defined orderings are compiled in the Reorder-

ing_List.

4. **Correction and adequation -> M-Xuxen module:** the Basque spell and grammar checker will be applied. Punctuation will also be adapted. In the future the coreference resolution system will be also integrated.

As case study, we have concentrated on parenthetical structures that contain biographical information. The simplification of these structures is performed by a tool called *Biografix* which is presented in the paper *Making Biographical Data in Wikipedia Readable: A Pattern-based Multilingual Approach* (Gonzalez-Dios et al., 2014a).

6.1 Summary

In this chapter we have presented the modules of the architecture of the system and the tools that are involved in them. As a case study, we have also performed the simplification of parenthetical biographical information.

In Figure 6.2 we have added the system *EuTS*, the tool *Biografix* and the datasets EGLU DS and Wikipedia DS to the contributions.

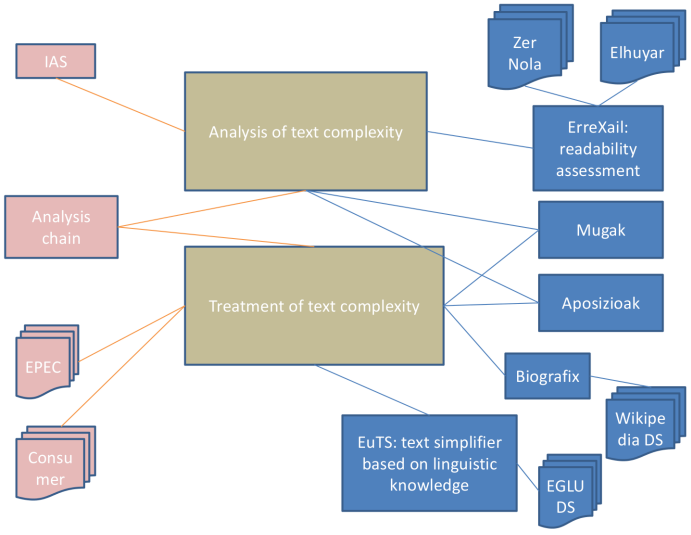

Figure 6.2 – Resources and tools used during thesis, and the contributions

Making Biographical Data in Wikipedia Readable: A Pattern-based Multilingual Approach

Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza
IXA NLP Group
University of the Basque Country (UPV/EHU)
itziar.gonzalezd@ehu.es

Abstract

In this paper we present *Biografix*, a pattern based tool that simplifies parenthetical structures with biographical information, whose aim is to create simple, readable and accessible sentences. To that end, we analysed the parenthetical structures that appear in the first paragraph of the Basque Wikipedia, and concentrated on biographies. Although it has been designed and developed for Basque we adapted it and evaluated with other five languages. We also perform an extrinsic evaluation with a question generation system to see if *Biografix* improve its results.

1 Introduction and motivation

Parentheticals are expressions, somehow structurally independent, that integrated in a text function as modifiers of phrases, sentences..., and add information or comments to the text. Therefore, it has been argued that they interrupt the prosodic flow, breaking the intonation. According to Dehé and Kavalova (2007), parentheticals can be realised in different ways: one-word parentheticals, sentence adverbials, comment clauses and reporting verbs, nominal apposition and non-restrictive relative clauses, question tags, clauses and backtracking. Besides, the authors argue that sometimes the parentheticals are not related to the host sentence neither semantically nor pragmatically, but they are understood in the text due to the situational context.

Some parentheticals can be the result of a stylistic choice (Blakemore, 2006) and that is the case of parenthetical information found in the first paragraph of some Wikipedia articles. As stated in the Wikipedia guidelines[1] the first paragraph of the articles should contain resuming and important information. That is why the information is there so condensed. Apart from condensing the information parentheticals cause long sentences, which are more difficult to process both for humans and for advanced applications. Moreover, web writting style books (Amatria et al., 2013) suggest not to use parenthetical constructs because they make more difficult the access to the information. Simple wikipedia guidelines[2] recommend also not to use two sets of brackets next to each other.

NLP applications such as question generation systems (QG) for educational domain[3] may fail when finding important information in brackets. For example, if we want to create questions, systems such as the presented in Aldabe et al. (2013) will look for a verb[4]. In the case of parenthetical biographical information there is no verb which makes explicit when the person is born or when she or he died. So, no question will be created based on that information.

The study of parentheticals in Basque has been limited to the analysis of the irony in the narrativity of Koldo Mitxelena (Azpeitia, 2011). In the present study we analyse the parentheticals that are used in the first paragraph of the Basque Wikipedia and developed a rule-based tool *Biografix* to detect these

[1] http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style (last accessed: March, 2014)

[2] http://simple.wikipedia.org/wiki/Wikipedia:Manual_of_Style (last accessed: March, 2014)

[3] Question generation is important in learning technologies (intelligent tutoring systems, inquiry-based environments, and game-based learning environments), virtual environments and dialogue systems among others. http://www.questiongeneration.org/ (last accessed: April, 2014)

[4] Both systems (one chunk-based and another dependency-based) presented in Aldabe et al. (2013) follow the guidelines presented in Rus and Graesser (2009).

Proceedings of the Workshop on Automatic Text Simplification: Methods and Applications in the Multilingual Society, pages 11–20,
Dublin, Ireland, August 24th 2014.

structures and to create new sentences out of them. To be more concrete, we concentrate on biographical information since there are not explicit words in text that give a clue about what type of information it is. Our aim is to make more readable sentences and, consequently, to eliminate the interruption they cause. About the domain of biographies, their automatic generation has been studied (Duboue et al., 2003) in Natural Language Generation (NLG). In this research line, referring expressions to people have been studied for automatic summarisation of news (Siddharthan et al., 2011). The quality of the biographies (linguistic and content) has been recently analysed in the English Wikipedia (Flekova et al., 2014).

We want also to make a first step towards the simplification of Basque Wikipedia, since English simple wikipedia has been a great resource for Text Simplification (TS) and Readability Assessement (RA). Efforts for simple wikipedia have also been made for Portuguese (Junior et al., 2011) using TS techniques. Although *Biografix* has been specially developed for Basque, being pattern-based, we have also evaluated its adaptation to other languages. This work is not limited to wikipedia, *Biografix* can be used on other types of text as well, since these structures can be found in educational texts, newspapers and so on.

This paper is structured as follows: after this introduction we report in section 2 the treatment of parentheticals in TS and in Wikipedia. In section 3 we describe *Biografix* and in section 4 we report its evaluation. Finally, we conclude and outline the future work in section 5.

2 Parenthetical Structures

In this section we report the treatment that parenthentical structures have undergone in TS and other NLP applications. We also describe the parentheticals found in Basque Wikipedia.

Parentheticals have been object of study in TS and three main operations have been proposed: a) parentheticals have been removed out of the texts (Drndarević et al., 2013), b) parentheticals have been removed but they have been kept in another form (Aranzabe et al., 2012; Seretan, 2012) or c) parentheticals have been added to explain the meaning by short paraphrases (Hallett and Hardcastle, 2008) or hyperonyms (Leroy et al., 2013). In any case, it is usually recommended to avoid them (Aluísio et al., 2008). In other NLP applications such as summarisation they are usually removed and even some QG works follow the same strategy, in case they are not relevant (Heilman and Smith, 2010).

2.1 Parenthetical Structures in Basque Wikipedia

Wikipedia guidelines emphasise the importance of the first paragraph. It should indeed contain a summary of the most significant information. To concentrate all the information, stylistic resources such as parenthentical structures are used. The information that is written in brackets in the Basque Wikipedia can be classified in two groups: a) information about people and b) information about concepts. About people biographical data and mandates are usually found and about concepts the etymology of words is frequent. Translations or transliterations of the named entity or the concept is found for both groups.

On the other hand, there are other frequent parenthetical structures that are found in the first paragraph, but they are not written in brackets. This is the case of the nicknames, which are written in commas. This kind of information is also found in other languages. After this analysis, we decided to concentrate on biographical data to create new sentences out of that information.

Biographical data Contrary to English Wikipedia, in Basque Wikipedia the information contained in bracket is, if known, birthplace (town, province, state), date of birth, and if the person is dead, date of death and place of dead. This is the case as well of the Catalan, Spanish, Italian, Portuguese, German and French Wikipedia among others, although sometimes paraphrases are found in brackets. For French there is, for example, more than a way to write the biographical data[5].

In Basque Wikipedia guidelines[6] it is stated that biographical data should be written as in examples 1 and 2. If the person is dead, we see in example 1 that the birth data (town, state and date) and the death data (town, state and date) are linked by a dash.

[5] `http://fr.wikipedia.org/wiki/Wikipedia:Conventions_de_style` (last accessed: March, 2014)
[6] `http://eu.wikipedia.org/wiki/Wikipedia:Artikuluen_formatua` (last accessed: March, 2014)

(1) *Ernest Rutherford, Nelsongo lehenengo baroia, (Brightwater, Zeelanda Berria, 1871ko abuztuaren 30a - Cambridge, Ingalaterra, 1937ko urriaren 19a) fisika nuklearraren aita izan zen.*
'Ernest Rutherford, 1st Baron Rutherford of Nelson, (Brightwater, New Zeeland, 30th August, 1871 - Cambridge, England, 19th October, 1937) was the father of the nuclear Physics.'

And if the person is alive, only birth data (town, province, date) is provided as in example 2.

(2) *Karlos Argiñano Urkiola, nazioartean Karlos Arguiñano grafiaz ezagunagoa, (Beasain, Gipuzkoa, 1948ko irailaren 6a) sukaldari, aktore eta enpresaburu euskalduna da.*
'Karlos Argiñano Urkiola, internationally known with the Karlos Arguiñano spelling, (Beasain, Gipuzkoa, 6th September, 1948) is a basque chef, actor and businessman.'

In both cases, the places (if known) should precede the date and these should be separated by commas. However, biographical data is not frequently written uniformly. Places do not precede the date, the date is incomplete (only year) and sometimes other characters like the question mark appear to denote that the place or the date is known.

Taking into account this guidelines and the articles we have analysed, we have developed *Biografix*, a pattern based tool that detects biographical data and creates new sentences with this information. This tool was originally developed for Basque but it has been adapted to other languages. An adaptation of this tool, moreover, could be used as a first step into Text Summarisation, if we only remove the parentheticals and do not create new sentences.

Biographical information is contained in brackets in other Wikipedias as well but formats may be different. The way of writing, for example, in Catalan, German and Portuguese is similar to Basque. In Spanish, French, and Italian that format is also used but, as mentioned beforehand, other formats are also accepted.

3 Inside *Biografix*

Biografix is a pattern-based tool that simplifies the biographical data and creates new sentences out of that information. Having as an input the example 1 in subsection 2.1, *Biografix* will produce the sentences 3, 4, 5, 6 and 7.

(3) *Ernest Rutherford, Nelsongo lehenengo baroia, fisika nuklearraren aita izan zen.*
'Ernest Rutherford, 1st Baron Rutherford of Nelson, was the father of the nuclear Physics.'

(4) *Ernest Rutherford 1871ko abuztuaren 30ean Brightwateren jaio zen.*
'Ernest Rutherford was born on the 30th of August, 1871 in Brightwater.'

(5) *Brightwater Zeelanda Berrian dago.*
'Brightwater is in New Zeeland.'

(6) *Ernest Rutherford 1937ko urriaren 19an Cambridgen hil zen.*
'Ernest Rutherford died on the 19th of October, 1937 in Cambridge.'

(7) *Cambridge Ingalaterran dago.*
'Cambridge is in England.'

So, if the person is dead, *Biografix* will write first the main sentence (3) followed by a new sentence (4) with the information about the birth. If the birthplace is composed by more than a place entity, sentences like (5) will be written. After the birth information, a sentence will contain the information about the death (6). For the cases that more than a place appear, those will be rewritten (7).

If the person is alive like in example 2 in subsection 2.1, the same process will take place, but no death information will appear by creating the new sentences 8, 9 and 10.

(8) *Karlos Argiñano Urkiola, nazioartean Karlos Arguiñano grafiaz ezagunagoa, sukaldari, aktore eta enpresaburu euskalduna da.*
'Karlos Argiñano Urkiola, internationally known with the Karlos Arguiñano spelling, is a basque chef, actor and businessman.'

(9) *Karlos Argiñano 1948ko irailaren 6an Beasainen jaio zen.*
'Karlos Argiñano was born on the 6th of September, 1948 in Beasain.'

(10) *Beasain Gipuzkoan dago.*
'Beasain is in Gipuzkoa.'

So, first, main information will be kept (8) and then the information about the birth will appear (9). As a second place information (the province) original sentence (2), it will be rewritten as well (10).

We have to mention that we use the title of the article as the subject of the sentences containing the biographical information. That is way we see that in sentences 9 and 10 the subject is *Karlos Argiñano* and the subject in sentence 8 is *Karlos Argiñano Urkiola*. We took this decision for cases where the real name of person is not so known, e.g. Cherilyn Sarkisian. Had we used Cherilyn Sarkisian in all the sentences, would someone have known we are talking about Cher?

To carry out these simplifying transformations *Biografix* follows the simplification process explained in Aranzabe et al. (2012):

- **Splitting:** In this stage we get the parts of the sentences we are going to work with. To that end, three steps take place: a) the parenthetical structure is removed from the original sentence; b) the type of parenthetical expression is checked looking at whether there are birth and death data or only the former; c) dates and places are split. We use simple patterns to detect the dates and the places. As it is possible to find more than a place, they will be split by the commas. This stage is common for all the languages.

- **Reconstruction:** The new simplified sentences are created in this stage. This part is language-dependent, since we add the verbs, determinants, prepositions and case markers. In the case of Basque we also remove the absolutive case that is found in some articles[7]. Anyway, we create three kind of sentences that are common for all the languages with the constructs obtained in the splitting stage: a) sentences indicating birth data, b) sentences indicating death data and c) sentences indicating place specifications. The main sentence will be kept as in the original version (the parenthetical has been removed in the splitting stage).

- **Reordering:** The sentences will be ordered in text. First, the main sentence; second, the information about the birth; if there is more than a place, the following sentences will contain that information (place specifications); third, the information about the death (if dead) and finally, the death place specifications.

- **Correction:** The aim of this stage is to check if there are any mistake in the new sentences and to correct them. As one of our goals is to know the correctness of *Biografix*'s output this stage has not been implemented yet.

Biografix has been designed for Basque and then the reconstruction stage has been adapted to other 7 languages: French, German, Spanish, Catalan, Galician, Italian and Portuguese. To develop the Basque version we implemented the guidelines in Wikipedia (see subsection 2.1) and we used a small corpus of 50 sentences to find possible cases, where the guidelines are not fulfilled. These 50 sentences were randomly crawled.

For other languages, we did not make any change in the splitting stage but for German. According to German Wikipedia guidelines birth and death data are separated by a semicolon and not by a dash. Although French, Spanish and Italian have other options to express the biographical information between bracket we did not implement them. Our aim is not to create a tool specially for these languages, but to see if the design for Basque can be applicable to other languages. That is why, the adaptations to other languages are available at our website[8], if someone wants to improve them.

[7] The absolutive case is used according to the format of the date.

[8] `https://ixa.si.ehu.es/Ixa/Produktuak/1403535629`

Other improvements could be done in the reconstruction stage. To rewrite the sentences we have used the most familiar past tense in each language. The only exception was French. The most familiar past tense according to the context is the *passé composé* but this tense requires the agreement of the gender between subject and verb[9]. As the *passé simple* is not very familiar we decided to use the present tense to avoid the concordance problem. So, this could be one of the things to take into account for future developers.

No other changes should be done in the reordering stage but the correction has to be adapted to each language. No training was performed for the other languages. Only 3-5 sentences were used to check that there were no errors.

4 Evaluation

In order to evaluate *Biografix* we crawled the first sentence of 30 Wikipedia articles. The method to select these articles was the following: a) we used CatScan V2.0β[10] to get a list of the Biographies in Basque Wikipedia; b) we randomised that list and make another list to see which articles were written in 8 languages (Basque, Catalan, French, Galician, German, Italian, Portuguese and Spanish); c) we selected the first 32 articles. The first two articles were used to explain and train the annotators. The final test-sample had, therefore, 30 items.

Having that sample, we performed two evaluations: a manual evaluation (section 4.1) and a extrinsic evaluation with a question generation system (section 4.2).

4.1 Manual evaluation

The manual evaluation was carried out for 6 languages: Basque, Catalan, French, Galician, German and Spanish. 10 linguists took part in the evaluation process and they evaluated three aspects of the task: the original sentences (*JatTestua*), *Biografix* performance (*Prog*) and the grammaticality of the new simplified sentences (*Gram*). In total they answered nine yes/no questions. This evaluation method we are proposing is useful to perform an error analysis and find out which are the weak points of our tool.

To evaluate the performance and the adaptation of *Biografix* we chose six languages according to the format of the biographical data: i) Basque (the language *Biografix* has been designed for) ii) Catalan (same format as Basque), iii) German (same format but a slightly variation), iv) Spanish (same format as Basque but other options as well), v) French (same format as Basque in one of the parenthetical formats and other options), vi) Galician (without defined format). Portuguese and Italian were not evaluated because their case studies were already evaluated with Catalan and Spanish. All the sample were evaluated by two annotators except for Catalan and Galician, because Catalan has the same case study as Basque and Galician has not a predefined format that could cause confusion.

Questions concerning the original sentences (*JatTestua*) Three questions were presented in regards to the original sentence in Wikipedia. The aim is to know if the original sentences do have parenthetical structures and therefore, how many of them are candidates to simplification (coverage).

1. Are there parenthetical structures written between brackets?

2. Is the sentence grammatically correct and standard?

3. Is the punctuation correct?

We asked about the grammaticality and the punctuation of original sentences (correctness) because it was shown in Aldabe et al. (2013) that many source sentences were incorrect and that fact decreased the performance of the question generators and the correctness of the created questions.

[9] e.g. *Cher est né**e** en Californie.*, but *Ernest Rutherford est né**ø** en Angleterre*.

[10] `http://tools.wmflabs.org/catscan2/catscan2.php` (last accessed: March, 2014)

Questions concerning the performance of *Biografix* (*Prog*) Four questions were designed to check if *Biografix* carries out the process it has been implemented for (precision).

1. Have parenthetical structures been removed?
2. Is all the information kept?
3. Taking into account the original sentence, is all the information correct?
4. Is there new information?

Second and third questions are essential to know if at rewriting in the reconstruction stage no information has been omitted or changed. The aim of the forth question is to know, for example, if sentences with other kind of information like translations have been added and treated as biographical or if a sentence referring to the death of a living person has been created.

Questions concerning the grammaticality of the new simplified sentences (*Gram*) Two questions were prepared to check the correctness of the simplified questions, since to create correct sentences is very important to understand the text. These questions should be answered for each simplified sentence (grammatical precision).

1. Is the sentence grammatically correct and standard?
2. Is the punctuation correct?

If these questions get negative results, we cannot forget that in our simplification study we consider the correction as a last step. This way, the output of *Biografix* will be checked and, were there any mistakes, they would be corrected.

4.1.1 Results of the manual evaluation

In table 1 we present the results obtained in the manual evaluation and it shows the results considering the following measures:

1. The coverage is the percentages out of 30 (the size of the sample) *Biografix* processed, that is, the sentences that had parentheticals.
2. The correctness is the percentage of the source sentences whose grammar and punctuation is correct.
3. The recall is the division between the number of the created simple sentences and the number of the sentences it should have created taking into account all the information in the original sentences.
4. The precision is the division between the correct performed, that is, all the *Prog* questions have been correctly answered and the processed sentences. We call this precision at performance.
5. The grammatical precision is the correctly created sentences among the created sentences.

In the second-last column we show the κ agreement of the evaluators (Cohen, 1960). As we have few examples, the expected agreement is very high and it causes low scores. That is the reason why we also show the percentage agreement (observed agreement) in the last column.

Taking a look at the results for Basque, we see that Biografix is able to create almost all the sentences (recall: 0.94) and that they are correct (grammatical precision: 0.87), although there are little problems keeping all the information and keeping it right (precision: 0.79). Taking into account that the percentage of the correct source sentences is low (82.76), we follow Aldabe et al. (2013) and recalculate the results without the incorrect sentences. This way, recall is 0.93, precision is 0.80, grammatical precision is 0.88. As we see, results do not vary that much, since the grammaticality of the source sentence has only influence in the first of the created sentences. About the agreement between annotators, we see that κ is really low (0.37) due to the few disagreements that annotators had above all about the grammar. However, the observed agreement is high (90.63).

Language	Coverage	Correctness	Recall	Precision	Gram. Prec.	κ	%
Basque	97.00	82.76	0.94	0.79	0.87	0.37	90.63
Catalan	93.33	98.21	0.77	0.53	0.78	-	-
French	73.00	88.64	0.80	0.18	0.37	0.39	85.06
Galician	43.00	88.46	0.76	0.15	0.62	-	-
German	100	100.00	0.78	0.60	0.78	-	100
Spanish	100	85.00	0.71	0.33	0.67	0.52	88.76

Table 1: Results of *Biografix* language by language

In the case of Catalan, we see that *Biografix* is not able to create as many sentences as information in the original source (recall: 0.77) and this tendency occurs in the other languages as well. Precision at performance goes down (0.53) due to added and lost information but grammatical precision is acceptable (0.78). We think, that this is a quite satisfactory adaptation.

The results for French indicate that something went wrong. There is more than a way to express the biographical information and, as expected, the performance goes down. The precision is very low (0.18) due to the fact that a lot of information is lost and as sometime paraphrases do appear in the original sentence, this fact implies grammatical error. Anyhow, the recall is acceptable (0.80) and that is a good starting point for the further development of French version. The average of the obtained κ measures is really low (0.39) and that is why having few instances Cohen's kappa penalises the disagreement too much.

The case of Galician is quite different. It is not stated in the guidelines how biographical data should be written and the parentheticals we found are few (coverage: 43.00) and different from the Basque. However, we wanted to try *Biografix* and what we see is, that, although its precision at performance is really low (0.15), the created sentences are quite correct (0.62). We think the Galician Wikipedia should be analysed thoroughly and then *Biografix* should be adjusted.

The German version of *Biografix* was able to simplify all the sentences found in the test-sample and its recall is high (0.88). Its weak point is the precision at performance (0.60), as in other languages, due to the fact that the second question of *Prog* is not satisfied. The sentence it creates are quite acceptable (0.71) as well. Surprisingly, both linguists agreed in all the cases and questions. So, we conclude that the German adaptation was successful.

Finally, in the case of the Spanish adaptation, we see that the precision is very low (0.33) since there was an important information loss. However, the grammatical precision (0.67) is acceptable. Although κ is higher (0.52) than in other languages, observed agreement is not far from Basque (88.76). It is remarkable as well that being Spanish a long time normalised language only the 85.00 % of the source sentence are correct and that although there are other formats to express the biographical information the coverage is absolute (100.00).

The main disagreement was found when evaluating the grammar and the punctuation due to different criteria of the annotators. For some of them sentences without verb were correct because they considered that there was an elided verb. In our opinion, as we are trying to simplify, we think that all the sentences should have a finite verb. Annotators did not have to much trouble to answer the four *Prog* questions, so we think that this is a good methodology, and, moreover, it makes easy to perform error analysis. We want to point out that κ has not been the best measure but we have used it as we consider that it is a standard to measure data reliability.

To conclude, we find that there is room to improve the versions in other languages, above all trying not to lose information but the adaptation of *Biografix* has been a good starting point. In fact, the adaptation has been quite satisfactory for German and Catalan, because they share the format with Basque but they should be further analysed. As foreseen, the languages with different formats like Galician, Spanish and French require a bigger analysis.

4.2 Extrinsic evaluation

To evaluate the performance of *Biografix* throughout another NLP advanced application, we used the web application *Seneko* (Lopez-Gazpio and Maritxalar, 2013)[11], the application of the chunk-based question generation system for educational purposes presented in Aldabe et al. (2013). This kind of evaluation was only performed for Basque.

We ran *Seneko* with the original sentences and the simplified sentences. The number of the generated questions is presented in table 2. We break down the results on the basis of the case markers as well. In agglutinative languages like Basque case markers are the morphemes that express the grammatical functions.

Source file	Total	Absolutive	Inessive	Genitive	Other
Original sentences	34	23	7	2	2
Simplified sentences	142	65	66	8	3

Table 2: Questions generated by *Seneko* using the original and the simplified sentences

Using as input the original sentences *Seneko* is able to create 34 questions, more or less a question per sentence. 23 of them have been generated for the absolutive case, that is, for the subject and the predicative, and only 7 of them have been generated for the inessive. Taking into account that we are working with biographical information, this is a bad result because the inessive case in Basque is used to express time and place relations. That is, the inessive is used to create questions with the question words *When* and *Where*. On the other hand, using as source the simplified sentences, 65 questions have been generated for the absolutive and 66 for the inessive. This way, we see that using *Biografix*'s output *Seneko* has been able to generate questions about place and time expressions.

Next, in 11 and 12 we show an example of the questions generated by *Seneko*. In 11 we find that using the original input it was only able to create a question, and it makes no sense but using the simplified text (example 12) *Seneko* creates two correct questions.

(11) a. **Source text:** *Eduardo Hughes Galeano (Montevideo, 1940ko irailaren 3a -) Uruguaiko kazetari eta idazlea da.*
'Eduardo Hughes Galeano (Montevideo, 3rd of September, 1940 -) is an Uruguayan journalist and writer.'

b. **Generated question:** *Nor da Eduardo Hughes Galeano Montevideo 1940ko irailaren 3a?*
'Who is Eduardo Hughes Galeano Montevideo 3rd of September, 1940?'

(12) a. **Simplified text:** *Eduardo Hughes Galeano Uruguaiko kazetari eta idazlea da. Eduardo Galeano 1940ko irailaren 3an Montevideon jaio zen.*
'Eduardo Hughes Galeano is an Uruguayan journalist and writer. Eduardo Hughes Galeano was born the 3rd of September, 1940 in Montevideo.'

b. **Generated questions:** *Nor jaio zen 1940ko irailaren 3an Montevideon? Non jaio zen Eduardo Galeano 1940ko irailaren 3an?*
'Who was born on the 3rd of September, 1940 in Montevideo? Where was born Eduardo Hughes Galeano on the 3rd of September, 1940?'

This way, we conclude that *Biografix* is an useful tool to improve the performance of question generation systems like *Seneko*.

5 Conclusion and future work

In this paper we have presented *Biografix*, a tool that detects parenthetical structures and simplifies the biographical data in order to create new more readable sentences. Although *Biografix* has been

[11] `http://ixa2.si.ehu.es/seneko/` (last accessed: March, 2014)

designed and developed for Basque, we have applied it to the parenthetical biographical information written in other seven languages: French, German, Spanish, Catalan, Galician, Italian and Portuguese. The results of the evaluation show that the Basque version obtains very good results but the adaptations should be further developed. Anyway, good results have been obtained for Catalan and German and promising for Spanish and French. Besides, we have shown its validity through an extrinsic evaluation with *Seneko*, a question generation system. These systems are important for the educational domain, and the improvement *Biografix* offers is considerable. Although we have used Wikipedia to develop and evaluate *Biografix*, it can be used for other kind of text with parenthetical biographical information.

For the future, we plan to continue analysing and implementing rules for other kind of parenthetical structures like etymology, translations of named entities or mandates of relevant people. We also plan to link the entities to the their articles in Wikipedia to offer additional information. Patterns could also be improved using previously developed analysers or tools, but this way the splitting stage will become language-dependent. Moreover, we cannot forget that this work is included in the main framework of the TS system for Basque that we are developing and this is another step towards the main aim of getting easier and more readable Basque texts.

Acknowledgements

Itziar Gonzalez-Dios's work is funded by a PhD grant from the Basque Government. We thank Aitor Soroa for his help with the language links in Wikipedia. A great part of this work would have not been able without the collaboration of the linguists Itziar Aduriz, Izaskun Aldezabal, Begoña Altuna, Nora Aranberri, Klara Ceberio, Ainara Estarrona, Mikel Iruskieta, Mikel Lersundi and Uxoa Iñurrieta. We also do appreciate the help Ander Soraluze offered during the implementation of *Biografix* and Oier Lopez de Lacalle for his quick tutorial on R. This research was supported by the Basque Government (IT344-10), and the Spanish Ministry of Science and Innovation, Hibrido Sint project (MICINN, TIN2010-202181).

References

Itziar Aldabe, Itziar Gonzalez-Dios, Iñigo Lopez-Gazpio, Ion Madrazo, and Montse Maritxalar. 2013. Two Approaches to Generate Questions in Basque. *Procesamiento del Lenguaje Natural*, 51:101–108.

Sandra M. Aluísio, Lucia Specia, Thiago A. S. Pardo, Erick G. Maziero, Helena M. Caseli, and Renata P. M. Fortes. 2008. A corpus analysis of simple account texts and the proposal of simplification strategies: first steps towards text simplification systems. In *Proceedings of the 26th annual ACM international conference on Design of communication*, SIGDOC '08, pages 15–22, New York, NY, USA. ACM.

Xabier Amatria, Urtzi Barrenetxea, Irene Fernández, Rakel Olea, Joseba Uskola, and Izaskun Zuntzunegi. 2013. *Komunikazio elektronikoa. IVAPen gomendioak web-orriak idazteko*. IVAP.

María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Itziar Gonzalez-Dios. 2012. First Approach to Automatic Text Simplification in Basque. In Luz Rello and Horacio Saggion, editors, *Proceedings of the Natural Language Processing for Improving Textual Accessibility (NLP4ITA) workshop (LREC 2012)*, pages 1–8.

Agurtzane Azpeitia. 2011. Enuntziatu parentetikoak: Koldo Mitxelenaren intentzio ironikoaren ispilu. *Gogoa*, 10(1&2).

Diane Blakemore. 2006. Divisions of labour: The analysis of parentheticals. *Lingua*, 116(10):1670–1687. Language in Mind: A Tribute to Neil Smith on the Occasion of his Retirement.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20:37–46.

Nicole Dehé and Yordanka Kavalova. 2007. Parentheticals. An introduction. In Nicole Dehé and Yordanka Kavalova, editors, *Parentheticals*, pages 1–22. John Benjamins Publishing Company.

Biljana Drndarević, Sanja Štajner, Stefan Bott, Susana Bautista, and Horacio Saggion. 2013. Automatic Text Simplification in Spanish: A Comparative Evaluation of Complementing Modules. In *Computational Linguistics and Intelligent Text Processing*, pages 488–500. Springer.

Pablo A. Duboue, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. 2003. PROGENIE: Biographical Descriptions for Intelligence Analysis. In *Proceedings of the 1st NSF/NIJ Conference on Intelligence and Security Informatics*, ISI'03, pages 343–345, Berlin, Heidelberg. Springer-Verlag.

Lucie Flekova, Oliver Ferschke, and Iryna Gurevych. 2014. What Makes a Good Biography?: Multidimensional Quality Analysis Based on Wikipedia Article Feedback Data. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 855–866, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Catalina Hallett and David Hardcastle. 2008. Automatic Rewriting of Patient Record Narratives. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Joseph Maegaard, Benteand Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Michael Heilman and Noah A Smith. 2010. Extracting Simplified Statements for Factual Question Generation. In *Proceedings of QG2010: The Third Workshop on Ques-tion Generation*, page 11.

Arnaldo Candido Junior, Ann Copestake, Lucia Specia, and Sandra Maria Aluísio. 2011. Towards an on-demand simple portuguese wikipedia. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 137–147. Association for Computational Linguistics.

Gondy Leroy, David Kauchak, and Obay Mouradi. 2013. A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty. *International Journal of Medical Informatics*, 82(8):717–730.

Iñigo Lopez-Gazpio and Montse Maritxalar. 2013. Web application for Reading Practice. In IADAT, editor, *IADAT-e2013: Proceedings of the 6th IADAT International Conference on Education*, pages 74–78.

Vasile Rus and Arthur C. Graesser. 2009. The Question Generation Shared Task and Evaluation Challenge. In *The University of Memphis. National Science Foundation*.

Violeta Seretan. 2012. Acquisition of Syntactic Simplification Rules for French. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Advaith Siddharthan, Ani Nenkova, and Kathleen McKeown. 2011. Information Status Distinctions and Referring Expressions: An Empirical Study of References to People in News Summaries. *Comput. Linguist.*, 37(4):811–842, December.

ANALYSIS OF MANUALLY SIMPLIFIED TEXTS

7 Corpus of Basque Simplified Texts

In this chapter we present the Corpus of Basque Simplified Texts (CBST), or *Euskarazko Testu Sinplifikatuen Corpusa* (ETSC) in Basque, henceforth ETSC-CBST. ETSC-CBST will be used to evaluate our simplification framework and proposals. As this part of the work has not been previoulsy published, we have translated the most of the chapter.

7.1 Introduction

Corpora of simplified text are text collections where each original text has its simplified counterpart. That is, for each text in corpora there is at least a simplified version. We can consider that these texts form a parallel corpus, since most of the sentences in each version should be related. The goal of this kind of corpora, is, therefore, to compile different versions of the same text that vary according to their difficulty. Contrary to other corpora, there is neither gold standard nor perfect simplified text, since the simplified texts can also be very different in consonance with their target audience.

The simplified texts can be written either following structural approaches or intuitive approaches. The structural approaches are used to create graded readings. In this approach, predefined word and structure lists are used to adapt the texts to the required level. Readability formulae are also used in this approach to check their complexity. These formulae take into account features such as word and sentence length, lexical lists to mention a few. On the other hand, intuitive approaches rely on the experience and intuition of

the teacher or the expert who is simplifying the text (Crossley et al., 2012). So, texts can also be simplified in different levels and having different target audiences in mind.

The corpus we are presenting here is the Corpus of Basque Simplified Texts (CBST), or *Euskarazko Testu Sinplifikatuen Corpusa (ETSC)* in Basque. It is a compilation of three original texts and two different versions of that one (one simplified version for each simplification approach). The main aims of ETSC-CBST are to see which are the operations performed to simplify texts in Basque and to see which common operations in both approaches are. We also want to compare these operations with our proposal. As a result of this, we have also developed an annotation scheme.

7.2 Corpus building and annotation

The originals texts we have used to be simplified are part of the Elhuyar corpus that was used to train the readability assessment system for Basque *ErreXail* (Gonzalez-Dios et al., 2014b). We selected 227 sentences corresponding to three long texts from different topics: social sciences, medicine and technology. We decided to use long texts to see the continuity of the simplification operations on the same topic. We differentiate three phrases to create the corpus:

1. **Starting phase:** a text from each topic has been simplified to see if these texts are suitable for this task. A list of operations (changes carried out to create the simplified text) performed has been created based on that experience and other languages. This list is the CBTS-annotationScheme-v0 and operations such as split clauses, substitute synonyms, or reorder clauses are considered. In total, there are 16 operations.

2. **Comparison phase:** a text of each topic has been given to simplify to two different persons: a court translator who has never worked on simplification before and a languages teacher that used to simplify texts for learners of Basque as foreign language. The translator was given easy-to-read guidelines and the list of operations created in the starting phase to help her (structural approach) and the teacher followed her intuition (intuitive approach). This phase has different aims:

- Compare different approaches
- Compare those approaches with our simplification proposal (evaluation)
- Update our criteria and operations with new ones
- Look for common criteria when simplifying

To carry out those aims various analysis of the corpus have been done until the definitive annotation scheme has been created. The outcome of this phase is the corpus we are presenting in this chapter.

3. **Extension phase:** the corpus will be enlarged applying the common operations (Appendix C).

Now we present in detail the work done in the second phase. The ETSC-CBST corpus has been annotated and analysed in two different phases:

1. **Exploratory analysis of the tagging:** we tagged the texts at paragraph level based on the operations extracted from the starting phase. We identified and classified the new phenomena that were not covered (classified as others) in the CBTS-annotationScheme-v0. We created a new set of operations (CBTS-annotationScheme-v1). This improved set has 31 operations and it is divided in syntactic operations, lexical operations, discourse level operations, reordering operations, ellipsis treatment, information treatment, others and no operation. We confronted the CBTS-annotationScheme-v1 to the Italian operations and annotation scheme (Brunato et al., 2015) because it fit better to our study.

2. **Definitive analysis of the tagging:** we tagged and analysed the texts at sentence level followed the annotation scheme (Subsection 7.3). The tool we used to annotate the corpus is Brat (Stenetorp et al., 2012).

In Figure 7.1 we can see an example of an annotated text. Texts are presented divided in sentences. The annotators choose the operation they want to perform (among a list provided to them) and the point or element implied in the operation.

In the following section we present our annotation scheme. We also explain the different macro-operations and operations involved.

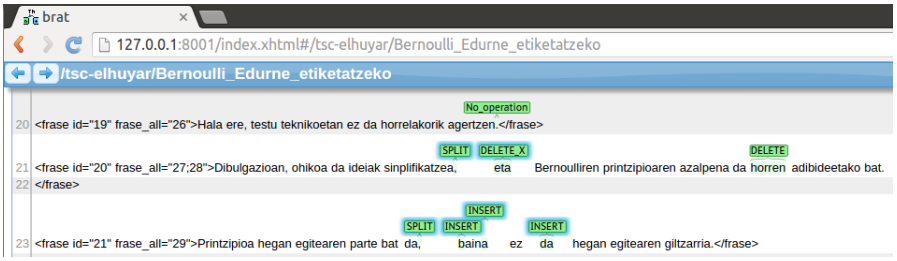

Figure 7.1 – A part of the text annotated with Brat

7.3 Annotation scheme: macro-operations and operations

The annotation scheme we present here is the result of two phases of annotation. It is organised eight in macro-operations. In the following points we briefly define the macro-operations for Basque:

- *Delete*: delete case markers, words, phrases, clauses or sentences
- *Merge*: create a clause or a sentence by joining other clauses or sentences
- *Split*: split or divide phrases, clause or sentences
- *Transformation*: alter the words, phrases, clauses, or sentences
- *Insert*: insert new elements (words, phrases, clauses or sentences).
- *Reordering*: change the order of words, phrases, clauses or sentences.
- *No_ operation*: do not perform any change
- *Other*: other kind of operations or operations difficult to classify

In Table 7.1 we sum up our annotation scheme and the operations included in each macro-operation. These will be explained by means of examples in the following subsections. The Basque sentences will be presented in italics and the operation or the cue words we are describing at each moment will be underlined. That operation will be also underlined in the English

translations. Sometimes, the English translations may sound unnatural or ungrammatical but we have taken that decision to be able to illustrate the Basque phenomena.

Macro-operation	Criteria	Sub-criteria
Delete **Merge**	Information	Information vs. functional words
Split	Strength	Hard vs. soft
	Phenomena	Coordination Adverbial clauses Relative clauses Apposition/ Parenthetical structures Noun clauses Postpositions Others
Transformation	Type	Lexical Morphological Syntactic Discourse Correction Other
Insert	Ellipsis	Marked morphologically vs. not marked (non required)
	Place	In new sentences vs. in the former original sentence
Reordering	Element	Phrases Clauses Auxiliary Verb
	Place	In new sentences vs. in the former original sentence
No_operation		
Other		

Table 7.1 – Annotation scheme

7.3.1 Delete

A delete operation is performed when some elements are eliminated from the original text. The elements that can be deleted are morphemes, words, phrases, clauses or sentences. We distinguish two types of deletions based on the criterion of the information deleted element contained:

- **Information deletion (delete-info)**: deletion of information is the case when the element that has been deleted added information to the whole sentence. In the example of Table 7.2, the relative clause *Sortzen*

den (that is created) containing a piece of information (maybe not relevant) has been deleted. The element deleted can be content/lexical words, phrases, clauses or even sentences.

- **Functional deletion (delete-functional)**: deletion of functional words such as conjunctions, discourse markers, morphemes (case markers and intensifiers) and punctuation marks. When a functional deletion is performed, there is no impact on the information of the text, although some nuances could disappear. In the example of Table 7.2, we consider that the deletion of the *eta* (and) conjunction does not delete information; so, we tagged it as delete-functional.

Operation	Original	Simplified
delete-info	*Sortzen den aldea oso handia da* The part that is created is very big *Azterketak erakutsi du askotariko aldaerak daudela gene horietan:* The study has shown there are many variances in these genes	*Aldea oso handia da* The part is very big *Askotariko aldaerak daude gene horietan:* There are many variances in these genes
delete-functional	*Eta beste edozein hegazkinekin ere gauza bera gertatzen da* And it also happens with any other kind of plane	*Beste edozein hegazkinekin ere gauza bera gertatzen da* It also happens with any other kind of plane

Table 7.2 – Examples of delete operations

7.3.2 Merge

When a merge operation is performed elements are fused. This macro-operation has not been found in the corpus frequently, so we have not been able to distinguish different operations or to sub-classify it. In the example we show in Table 7.3, two sentences have been merged to create one using as link the pronoun in genitive case *haien* (their). In this case, the merge has been performed by means of a coreference resolution, since the pronoun has been substituted with its referent to link the sentences.

Operation	Original	Simplified
Merge	*Adibide bat gaur egungo hegazkin komertzialen hegoak dira. Haien diseinua plano aerodinamiko superkritikoan oinarrituta dago.* The wings of the modern commercial planes are an example. Their design is based on the supercritical airfoil.	*Gaur egungo hegazkin komertzialen hegoen diseinua plano aerodinamiko superkritikoan oinarritzen da.* The design of the wings of the modern commercial planes is based on the supercritical airfoil.

Table 7.3 – Examples of merge operations

7.3.3 Split

The split is the operation where clauses, phrases or morphemes are divided with the aim of creating new sentences. We distinguish different types of splits based on two criteria::

- **Depending on the strength**: the punctuation mark of the resulting simplified sentence is taken into account to determine if the split is soft or hard. The soft split occurs when a new sentence has been delimited by a comma or a semicolon and the hard happens when the new simplified sentence has been delimited by a full stop.
- **Depending on the phenomena**: the phenomena we take into account are coordination, noun-clauses, relative clauses, adverbial clauses (and different adverbial types), appositions, postpositions and other.

In Table 7.4 we show two instances of split. These examples show the split depending on the strength and in both cases the phenomena that has been split is the coordination.

Operation	Original	Simplified
split-hard-coordination	*Dibulgazioan, ohikoa da ideiak sinplifikatzea, eta Bernoulliren printzipioaren azalpena da horren adibideetako bat.* It is normal to simplify the ideas in the science popularisation, and Bernoulli's principle is an example of that.	*Dibulgazioan ohikoa da ideiak sinplifikatzea. Bernoulliren printzipioaren azalpena da adibideetako bat.* It is normal to simplify the ideas in the science popularisation. Bernoulli's principle is an example of that.

(Continued on the next page)

Operation	Original	Simplified
split-soft-coordination	*Hortik aurrerako azalpena konplexua da, eta hegalari batetik bestera asko aldatzen da.* From that on, the explanation is complex, and it changes considerably from one flyer to another.	*Hortik aurrerako azalpena konplexua da; hegalari batetik bestera asko aldatzen da.* From that on, the explanation is complex; it changes considerably from one flyer to another.

Table 7.4 – Examples of split operations

7.3.4 Transformation

Transformations suppose the change of a word, a phrase or a structure. We distinguish transformations of different types: lexical, morphological, syntactic, discursive, and corrections. Combinations of the are also possible. These are there distinguished transformation operations:

- **Lexical:** Subst_Syn (synonym substitution) and Subst_MultiWord (substitution of phrases)
- **Morphological:** Pas2Act (passive -> active or impersonal-> personal), Fin2NonFin (finite -> non-finite), NonFin2Fin (finite -> non-finite), Subst_Per (change of the person) and Verb_Feats (changes in the verb)
- **Syntactic:** Clause2Phrase (clause -> phrase), Phrase2Clause (phrase-> clause), Ind2Dir_Speech (style change: indirect -> direct), Dir2Ind_Speech (style change: direct -> indirect), Sub2Main (subordinate clause -> main clause), Main2Sub (main clause -> subordinate clause), Connect_Syntax (change the syntactic connector) and Sub2Coor (subordinate clause -> coordinate clause)
- **Discourse:** Coref (coreference resolution) and Connect_Disc (change of discourse marker)
- **Correction :** Correction (correction of spelling, grammatical and punctuation mistakesk)
- **Combinations:** Reform (reformulations edo periphrasis) and Other_Subst (others)

Examples of the transformation operations are shown in Table 7.5. It is possible that some instances represent more than an operation. Indeed, it is difficult to find examples with one operation only.

Operation	Original	Simplified
Subst_Syn	*ahaleginetan* in the efforts	*lanetan* in the works
Subst_MultiWord	*urteetan zehar* through the years	*urtero* every year
Pas2Act	*ikusi da* it has been seen	*ikusi dute* they have seen
Fin2NonFin	*hegazkin horiei airean eusten dien printzipio fisikoa* the physical principle that keeps those planes in the air	*hegazkin horiei airean eusteko printzipio fisikoa* the physical principle to keep those planes in the air
NonFin2Fin	*Airea beherantz bultzatuta* pushing down the air	*Airea beherantz bultzatzen da* the air is pushed down
Subst_Per	*orduan odolean begiratzen dugu* so, we look in the blood	*orduan odolean begiratzen dute* so, they look in the blood
Verb_Feats	*gai izango litzateke* they might be able	*gai izango da* he will be able
Clause2Phrase	*Jatorri genetikoa duten minbizi gehienetan* in the most of the cancers that have genetic origin	*Jatorri genetikodun minbizi gehienetan* in the most of the cancers with genetic origin
Phrase2Clause	*bakoitzak oso diseinu ezberdinarekin* each one with its different design	*Bakoitzak bere diseinua du* each one has its own design
Ind2Dir_Speech	*familian zenbat kasu dauden galdetzen dugu* we ask how many cases there are in the family	*zenbat kasu daude familian?* how many cases are there in the family?
Dir2Ind_Speech	*horiekin "ez da eragozten" minbizia sortzea* the creation of 'is not impeded' with those	*horiekin ez dela galarazten minbizia sortzea* that the creation of is not hindered with those
Sub2Main	*fluxu horrek presio handiagoa egiten diola hegoari behetik goitik baino*	*fluxu horrek presio handiagoa egiten dio hegoari behetik goitik baino*

(Continued on the next page)

Operation	Original	Simplified
	that that flux makes more pressure to the wing downwards than upwards	the flux makes more pressure to the wing downwards than upwards
Main2Sub	*Familia barruan minbizi horietako kasu asko dituzten pertsonak iristen dira kontsultara* People that have those cancer cases in the family arrive at the consultation	*Mujikak esan du kontsultara etortzen direla familia bereko pertsonak.* Mujika has said that people that have those cancer cases in the family come to the consultation
Connect_Syntax	*angelu horren inguruan irauten duen bitartean* while it lasts around that angle	*angelu horren inguruan irauten badu* if it lasts around that angle
Sub2Coor	*Hartara, mutazioa identifikatuta,* Thus, identified the mutation,	*Hartara, mutazioa identifikatzen dugu;* Thus, we identify the mutation;
Coref	*Mende hartan* in that century	*XVIII. mendean* in 18th century
Connect_Disc	*beraz* thus/ therefore	*ondorioz* as a result of
Correction	*abiadura (...) izan beharko luke* the speed (abs) should have	*abiadurak (...) izan beharko luke* the speed (erg) should have
Reform	*Zama guztiarekin, 573 tonara irits daiteke.* With all the load, it can arrive to 573 tones	*Zama guztiarekin, 573 tona pisatzen du, gutxi gorabehera.* With all the load, it weights 573 tones, approximately
Subst_Other	*hegaldiaren azalpenetik* from the explanation of the flight	*hegaldiaren azalpenean* in the explanation of the flight

Table 7.5 – Examples of transformation operations

7.3.5 Insert

Insert operations occur when a new element is introduced in the text. This new element can be a word, a clause or a sentence and it is added to recover a functional relation or to treat the ellipsis. So, we have taken two criteria:

1. The place, where the insertion has been done: in a former original sentence or in a new simplified sentence.

2. The ellipsis type: if the ellipsis is marked morphologically (elided_morph) or not (not_required).

Based on the two criteria, those are the three types of insertions we have distinguished:

- **Funct_NS**: elements that have been included in the new simplified sentences. These insertions happen after a split operation and they are usually used to recover a deleted functional relation. This insertion cannot happen is a split has not been performed. In the example presented in Table 7.6 the coordinated apposition has been split and when creating the simplified sentences out of that apposition the verb *da* (is) has been added.

- **Ellided_morph**: verbs or nouns that are marked morphosyntactically (there is a morphological mark of the ellipsis, usually a determinant) but have been made explicit. This operation happens in the former original sentence. In the example of Table 7.6, there is marked ellipsis in the word *obulutegietakoa* (the ovarian); to recover this ellipsis, *minbiziaren pronostikoa* (prognosis of cancer) has been added in the simplified sentence.

- **Non required**: ellided arguments, adjective, adverbs, sentences or whatever that has been inserted to make the meaning clearer. This operation also happens in the former original sentence. In the example of Table 7.6, the subject *proteinak* (the proteins) has been added.

Operation	Original	Simplified
Funct_NS	*Hala esaten du La Pizarra de Yuri blogeko Antonio Cantó dibulgatzaile eta hegazkinetan adituak* So states the blogger of La Pizarra de Yuri, science populariser and expert on planes Antonio Cantó	*Antonio Cantó dibulgatzailea da; Antonio Cantó hegazkinetan aditua da.* Antonio Cantó is a science populariser; Antonio Cantó is an expert on planes

(Continued on the next page)

Operation	Original	Simplified
Ellided-_morph	*endometrioko minbiziaren pronostikoa obulutegietakoa baino askoz ere hobea izaten da* the prognosis of the endometrial cancer is so much better than the ovarian	*endometrioko minbiziaren pronostikoa obulutegietako minbiziaren pronostikoa baino askoz ere hobea izaten da* the prognosis of the endometrial cancer is so much better than prognosis of the ovarian cancer
Non required	*Eraldatuta badaude* If (they) are transformed	*Proteinak eraldatuta badaude* If the proteins are transformed

Table 7.6 – Examples of insert operations

7.3.6 Reordering

In the reordering operation the order of the elements is altered. We have found different types of reordering operations and the criteria have been element that has been moved (phrase, clause or auxiliary verb) and where it has been done (place). That is, the Reord$_P$hrase, Reord$_C$lause and Reord$_A$ux operations happen in the former original sentence, when no structural change has been performed. The Reord$_N$S operation happens, as in the case of the insertion in new sentences, when after a split, a phrase or a clause has been moved to a new simplified sentence. These are the reordering operations we have found:

- **Reord_Phrase**: the ordering of the phrases has been changed, but they still remain in the same sentence.
- **Reord_Clause**: clause ordering has been altered, but they are kept in the same sentence.
- **Reord_Aux**: the auxiliary verb has been moved.
- **Reord_NS_Phrases**: phrases that have been moved to new sentences. This reordering cannot be done unless a split has been performed and it happens in the simplified sentences. In the example presented in Table 6, a noun clause has been split and after that, the main clause of the former original sentence *adituek aurreikusten dute* (the experts foresee), which was preceding the subordinate clause, has been postponed in the simplified sentence. Note that there is also a Reord_Phrase in that example among other operations.

The instances of the reordering operations are shown in Table 7.7.

Operation	Original	Simplified
Reord_Phrase	*(...) argitu du Bachiller astronomoak* (...) has clarified the astronomer Bachiller	*Bachiller astronomoak argitu du (...)* The astronomer Bachiller has clarified (...)
Reord_Clause	*Aireak hegazkinaren inguruan duen jokabidea zoruak alda dezake, hegaldia oso baxua denean.* The soil can change the behaviour that the air has around the plane, when the flight is very low.	*Hegaldia oso baxua denean zoruak hegazkinaren inguruko airearen jokabidea alda dezake.* When the flight is very low, the soil can change the behaviour that the air has around the plane.
Reord_Aux	*Orain dela 25 urte, berriz, eguzki-sistemako planetak baino ez ziren ezagutzen.* 25 years ago, on the contrary, only planets in the solar system known were.	*Orain dela 25 urte, berriz, eguzki-sistemako planetak bakarrik ezagutzen ziren.* 25 years ago, on the contrary, only planets in the solar system were known.
Reord_NS-_Phrases	*Hala ere, adituek aurreikusten dute planetagaien % 90, gutxi gorabehera, benetako planetak izango direla.* However, the experts foresee that more or less the 90 % of the candidates to be planets is going to be real planets.	*Hala ere, planetagaien % 90, gutxi gorabehera, benetako planetak izango dira; hala aurreikusi dute adituek.* However, more or less the 90 % of the candidates to be planets is going to be real planets; so foresee the experts.

Table 7.7 – Examples of reordering operations

7.3.7 No_operation

The operation no_operation is used when no change or alteration has been produced, that is, when the simplified sentence remains like the original one. The sentences that have this tag are also interesting so we can explore why they have not been simplified. An example of this operation is shown in Table 7.8.

Operation	Original	Simplified
No_operation	*Azken batean, hori da hegan egitearen sekretua.* After all, that is the secret of flying.	*Azken batean, hori da hegan egitearen sekretua.* After all, that is the secret of flying.

Table 7.8 – Example of no_operation

7.3.8 Other

This macro-operation is used, on the one hand, to tag the operations that have not been covered by this annotation scheme and, on the other hand, to tag the cases that are tricky to classify. This macro-operation will be used as less as possible and the sentences with this tag will be further analysed.

7.3.9 Annotation schemes in other languages

In this subsection we are going to compare our annotation scheme to the Italian (Brunato et al., 2015) and the Spanish (Bott and Saggion, 2014) annotation schemes. To make the comparison clearer, in Table 7.9 we sum up the terms used in these works and our equivalents.

Basque	Italian	Spanish
Macro-operations	Classes	First dimension
Operations	Sub-classes	Second dimension

Table 7.9 – Terminology used in different annotation schemes

Let us begin explaining the similarities and the differences found in relation to the Italian annotation schema. At, macro-operation level, we have defined the same macro-operations being the only difference that we have grouped those cases that cannot be classified properly in the others (other and no_operation) with the aim of storing them to be deeply studied further on. At operation level (sub-classes in the Italian scheme), we found three main differences: a) in the deletion operation the sub-classes are defined according to the part of speech (PoS) of the element to be deleted, while we also consider whether the deleted element is a content word or not. b) In the insertion operation, they use again the PoS of the inserted elements to define the sub-classes while we distinguish the types of inserts. c) In the transformation operation, they also classify them according to their type, but

as expected, we find different operations since transformations form a wide range of operations.

The Spanish annotation scheme is a two-level dimensional taxonomy. Our main macro-operations (all but other and no_operation) have their equivalent in their first dimension (in some cases using different terminology). Moreover, they define what they call proximization (make the information closer to the reader) and select (emphasise information, or make it as title), two macro-operation we did not identify in our work. Referring to the categorisation of the second dimension, we cannot establish a comparison because it is not explicitly stated, but from the results table we can conclude that they are quite similar to our types and phenomena. Some of them are, for example, change:lexical, split:coordination and insert:missing main verb.

7.4 Annotation results and trends

In this section we present the results and the analysis of the operations performed to create the simplified texts. First, we will present the alignment results and then the incidence of the macro-operations and operations. When possible, we will relate our work to other languages.

With these results we want to know which are the operations performed to create a simplified text and also, we want to compare different approaches when simplifying texts. These results and comparison will help us to establish common criteria. Our final aim is also to compare these approaches with the simplification approach we have presented in this thesis.

Before we get into the results, we show the details of the ETSC-CBST corpus. ETSC-CBST is formed by the original text and two different simplifications of that text. Each simplified version of the text has been done following a different approach. The translator has followed easy to read guidelines and the teacher has followed her experience and intuition. The details of the corpus are to see in Table 7.10.

Looking at the sentence number, we find more sentences in the simplified texts than in the original texts. In the case of the word number, it is incremented in the texts simplified by the translator but that tendency occurs only in one of the texts simplified by the teacher. The other two simplified texts have, therefore, fewer words.

Let us give an overview of the corpora simplified manually in other lan-

Text	Version	Sentences	Words
Bernoulli (technology)	original	89	1446
	translator	123	1472
	teacher	105	1253
Etxeko (medicine)	original	70	1535
	translator	84	1611
	teacher	75	1608
Exoplanetak (social science)	original	68	1512
	translator	75	1608
	teacher	96	1258
Total	original	227	4493
	translator	282	4691
	teacher	276	4119
Total	corpus	785	13303

Table 7.10 – Sentence and word number in the original and simplified texts

guages. In English 104 articles have been used, each one with its abridged[1] version. There are 2,539 sentences in the original version and 2,459 in the abridged; word number is 41,982 in the original and 29,584 in the abridged (Petersen and Ostendorf, 2007).

In Brazilian Portuguese, two different levels of simplification are compiled. So, texts are found in the original, natural simplified and strong simplified versions. The sentence numbers are 2,116 in the original, 3,104 in the natural and 3,537 in the strong; word numbers are 41,897 in the original, 43,013 in the natural and 43,676 in the strong (Caseli et al., 2009).

In Spanish, in the sample of corpus that has been used to align the sentences automatically, 110 sentences are found in the original part and 145 sentences in the simple part. The word numbers are 2,456 and 1,840 respectively (Bott and Saggion, 2011). In another work 37 documents are used (Štajner et al., 2013). On the other hand, the FIRST corpus has 25 documents and 330 sentences.

In Danish, the corpus has been compiled with 3,701 parallel documents. In the aligned part of the corpus there are 48,186 original sentences and 62,365 simplified. There are more sentences in the corpus which have not been aligned (Klerke and Søgaard, 2012).

In Italian, the corpus has two parts or subcorpora where the structural

[1]This is the term the authors use to refer to the simplified texts.

approach (*Terence*) and the intuitive approach (*Teacher*) are compared. In the *Terence* part there are 1,036 original sentences and 1,060 simplified. In the *Teacher* part, there are 24 parallel documents (Brunato et al., 2015).

If we compare the size of ETSC-CBST is in general smaller. The only exception is the Spanish sample. About the trend of sentence and word number difference from original to simplified, we see that sentence number also increases in Portuguese, Spanish, Danish and the Italian *Terence*. Word number rises in Portuguese but decreases in Spanish. In English both sentence and word number decline. This comparison is shown in Table 7.11.

Language/Corpus	**Art.s Doc.s**	**Sentences**		**Words**	
		Original	**Simplified**	**Original**	**Simplified**
English Petersen and Ostendorf (2007)	104	2,539	2,459	41,982	29,584
Brazilian Portuguese Caseli et al. (2009)	-	2,116	3,104 (nat.) 3,537 (str.)	41,897	43,013 (nat.) 43,676 (str.)
Spanish Bott and Saggion (2011)	-	110	145	2,456	1,840
Danish Klerke and Søgaard (2012)	3,701	48,186	62,365	-	-
Italian *Terence* Brunato et al. (2015)	-	1,036	1,060	-	-
Italian *Teacher* Brunato et al. (2015)	24	-	-	-	-

Table 7.11 – Sentence and word number in the original and simplified texts in other corpora

7.4.1 Alignment

The alignment of the corpus is basically to know which sentences of the simplified texts have been created out of the original n sentence. This is an important step in order to know which operations have been performed when simplifying the texts. We have also explored in which scale this alignment happens. That is, we have analysed how many sentences are related to an original one. So, the scale 1:1 means that for an original sentence a simpli-

fied has been created and the scale 1:2 means that there are two simplified sentences for each original. The results in percentages can be seen in Table 7.12.

Scale	Translator	Teacher
1:1	76.21	73.25
1:2	18.50	19.74
1:3	3.52	4.39
2:1	0.88	0.44
Others	0.88	2.19

Table 7.12 – Alignment results

Most of the sentences have been aligned in 1:1 scale. The translator has performed that alignment the 76.21 % and the teacher the 73.25 %. The second most used scale has been 1:2; the translator has performed it in the 18.50 % of the sentences and the teacher in the 19.74 %. The 1:3 and 2:1 scales are less frequent in both approaches. Other scales cover the cases where a sentence has been aligned to more than three sentences or to half sentences. We want to mention that the percentages of the alignments are quite similar in both approaches.

We have also analysed the alignments in other languages. The scale 1:1 has also been the most used in English (Petersen and Ostendorf, 2007), Italian (Brunato et al., 2015) and Spanish (Štajner, 2015). The second most used scales are in English 1:0, in Italian 2:1 in the intuitive approach and 1:2 in the structural approach and in Spanish 1:n in both corpora.

7.4.2 Incidence of macro-operations and operations

In this subsection we are going to see the incidence of the operations performed to create the simplified texts. We will start analysing the results of the macro-operations in general (Table 7.13).

The macro-operation that has been performed most of the times is the transformation. It has been done 24.92 % by the translator and 33.62 % by the teacher. The second most used operation differs in the approaches: the translator has used the split (23.55 %) while the teacher has used the delete (20.78 %). The less frequent macro-operations are merge and other in both approaches.

The transformations have been used more frequently by the teacher (33.62

Macro-operations	Translator	Teacher
Transformation	24.92	33.62
Split	23.55	12.30
Insert	21.88	18.61
Delete	17.66	20.78
Reordering	7.95	8.27
No_operation	3.53	6.20
Merge	0.40	0.22
Other	0.10	0.00

Table 7.13 – Results of the macro-operations in both approaches

%) than the translator (24.92 %). The sentences which have not been simplified (no_operation) are also more frequent in the teacher's approach (6.20 %) than in the translator's approach (3.53 %). The percentages of reordering, insert and delete are quite similar. The split has been used more time by the translator (23.55 %) than the teacher (12.30 %).

It is not surprising to find that the transformation is the most used macro-operation. It is, indeed, a macro-operation that incorporates many different operations, and as simplification is also considered as rewriting, and many of the rewriting operations are usually transformations. Let us show now go through the different transformations.

The most frequent transformation found in the texts of the translator is Sub2Main (48.50 %) and the reformulation (19.09 %) has been the most used in the texts of the teacher. With these results, we see that the translator has a tendency to convert subordinate clauses into main clauses while the teacher has used a broader variety of operations.

Sorting the transformations according to their type (Table 7.14), we see that in both approaches the most used transformations are the syntactic transformations. The less used are corrections.

Transformation type	Translator	Teacher
Syntactic	41.34	33.01
Morphological	22.05	19.09
Others	17.57	19.74
Discursive	14.96	15.86
Lexical	6.70	11.03
Correction	0.39	1.29

Table 7.14 – Results of the transformation types in both approaches

In our opinion, the importance of the syntax when simplifying texts is underlined as it is the most used transformation type in both approaches. Except for the syntactical and lexical transformations, there is no big difference between the approaches in the other transformation types. The translator has given more importance to the syntactic transformations (almost eight points of difference) while the teacher has given it to lexis (more than four points of difference). We would like also to remark the importance of the morphological transformations. Transformations tagged as other should also be analysed in the future.

Let us show now the results of the split operations. The split depending on the strength that has been most used by the translator is the soft (74.06 %) while the most used by the teacher is the hard (69.03 %). As we can see both approaches differ absolutely at this point.

Looking at the phenomena that have been split, the coordination has been the most (the translator 39.17 % and the teacher 45.13 %), followed by the adverbial clauses (the translator 19.16 % and the teacher 16.81 %). All the results are to see in Table 7.15.

Split phenomena	Translator	Teacher
Coordination	39.17	45.13
Adverbial clauses	19.16	16.81
Relative clauses	16.25	11.50
Apposition/ Parentheticals	10.83	7.96
Noun clauses	7.50	0.00
Postposition	3.75	3.54
Others	3.33	15.05

Table 7.15 – Results of the splitting operation according to the phenomena in both approaches

We also have analysed the types of the adverbial clauses that have been split and these results are shown in Table 7.16.

The most split adverbial clauses by the translator have been the conditional (23.91 %) and the causal clauses (21.74 %). Causal clauses (42.11 %) have also been the most simplified by the teacher together with the temporal clauses (26.32 %).

We also have analysed the percentage of split subordinate clauses taking into account their number in the original texts. To perform this experiment, we have used the automatic linguistic analysis and profiling of *ErreXail*. These results are shown in Table 7.17.

Split (adverbial)	Translator	Teacher
Conditional	23.91	0.00
Causal	21.74	42.11
Modal	17.39	5.23
Temporal	13.04	26.32
Concessive	10.87	15.79
Purpose	6.52	10.52
Comparative	6.52	0.00

Table 7.16 – Results of the splits adverbial clauses in both approaches

Subordinate type	Number (orig.)	Split (trans.)	Split (teach.)
Noun clause	162	11.11	0.00
Modal	69	11.59	1.45
Relative	57	66.67	22.81
Conditional	57	19.30	0.00
Temporal	34	17.65	14.71
Causal	23	43.48	34.78
Purpose	20	15.00	10.00
Modal-temporal	17	0.00	0.00
Concessive	5	100.00	60.00

Table 7.17 – Proportion of the split subordinate clauses

The clauses that have mainly been split by the translator are the concessive clauses (100.00 %), relative clauses (66.67 %) and causal clauses (43.48 %). The teacher has also split those clauses mainly but the ranking is different: concessive clauses (60.00 %) causal clauses (34.78 %) and relative clauses (22.81 %). The proportion of temporal clauses is also similar (translator 17.65 % and teacher 14.71 %). Neither of two split the modal-temporal clauses.

Another macro-operation that has been widely used is the insert. The results of the three insert types are shown in Table 7.18.

Insert operations	Translator	Teacher
Non required	44.39	57.89
Funct_NS	42.15	30.99
Ellided_morph	13.45	11.11

Table 7.18 – Results of the insert types in both approaches

The non required inserts have been the most used insert type in both approaches (translator 44.39 % and teacher 57.89 %). The inserts that have

been used in the creation of new sentences (Funct_NS) are in the second position in both approaches and the recovery of the ellided elements was the less used (it seems that this phenomenon is not so frequent). Although the ranking of the insert types is the same in both approaches, there are big differences in the use of them.

According to the treatment of the information we have distinguished two delete operations. Those where information has been omitted are the 25.56 % in the translator's texts and the 30.37 % in the teacher's texts. The deletes of functional words are 74.44 % in the translator's texts and 69.36 % in the teacher's texts. That is, in both approaches the most of the deletes do not imply information loss. These results are shown in Table 7.19.

Delete operations	Translator	Teacher
Delete information	25.56	30.37
Delete functional words	74.44	69.36

Table 7.19 – Results of the delete in both approaches

The deletes where information has been lost require a deeper analysis, and from that analysis, we will see if any categorisation could be made. On the other hand, the deletes of functional is a closed group and in Table 7.19 we show the functional deletes that have been performed. In both approaches the functional words that have been mainly deleted are coordinate conjunctions, punctuation and discourse markers.

Delete functional words types	Translator	Teacher
Coordinate conjunction	54.48	33.08
Punctuation	23.88	34.59
Discourse marker	14.93	24.06
Other	6.71	8.27

Table 7.20 – Results of the delete of functional words in both approaches

The results of the reordering operations are shown in Table 7.21. The most used reordering in both approaches has been the reordering of phrases, although it has been broader used by the teacher (78.95 %) than the translator (43.20 %). The second most used by the translator has been the movement of phrases into new sentences (41.98 %) while the teacher has changed the ordering of clauses (13.16 %).

It will be interesting to analyse how these reordering operations have been performed. In the case of phrasal reordering operations, we should see if they

Reordering operations	Translator	Teacher
Reord_Phrase	43.20	78.95
Reord_NS_Phrases	41.98	7.89
Reord_Clause	13.58	13.16
Reord_Aux	1.23	0.00

Table 7.21 – Results of the reordering in both approaches

have moved the phrases to fulfil the canonical word order or in case of the clauses if their movements are according to the positions we have found in our quantitative corpus analysis (Chapter 4).

The results of rest of the macro-operations (no_operation, merge and other) are shown in Table 7.22. Except for the no_operation, the other operations do not reach the 1 %.

Other macro-operations	Translator	Teacher
No_operation	3.53	6.20
Merge	0.40	0.22
Other	0.10	0.00

Table 7.22 – Results of the other macro-operations in both approaches

The sentences where no_operation has been applied need also another analysis to know why they have not been simplified. In our opinion, the merge has not been performed because it is an operation that is more related to summarisation than to simplification.

Comparison among approaches

In order to summarise the results, we are going to point out what we have found in common in both approaches and compare with ours. The most performed macro-operation has been the transformation and the most used transformation type has been the syntactic. The need of correction has also been indicated. The phenomena that have been mainly split are the coordination and the adverbial clauses. Among the types of the subordinate clauses and taking into account the numbers of the original texts, the concessive, the causal and the relative clause have been the most split.

These points we have mentioned agree with the simplification we have proposed in previous chapters: we have decided to perform syntactic simplification and we have concentrated on coordination and adverbial clauses in

our split and reconstruction operations. In our proposal we have also pointed out the need of correction.

Among other operations that are common in both approaches, we find the non required inserts. That phenomenon is considered as future work for us, when the coreference system that will carry out that task be ready and when we will be able to insert information from other sources (e.g. Wikipedia). The functional deletes, although they have not been treated as a category in our approach, it is true that some of them such as the punctuation or the conjunction deletion are taken into account in our syntactic simplification rules. We also have considered the phrase reordering, and although we plan to use neurolinguistic studies and the canonical word ordering, we think we also should make a corpus analysis.

Summing up, we think that what we have found in common in both approaches is linked to what we propose or what is planned for the future work. We also want to underline the incidence of the morphological transformations and we think we should also bear in mind. Anyway, there is still a lot of to learn from these approaches.

So, in the the list of simplification operations (Appendix C) to enlarge manually the corpus (extension phase) following points should be taken into account: syntactic transformations should be performed, concentrating on the splitting of coordinate and concessive, causal and relative clauses. Non required information should be also added, that is, elided subjects, objects and so on should be recovered.

To finish with this subsection, we want to mention that our tagging scheme has been effective to tag and analyse the texts and to compare the approaches. We are sure, however, that it can be still further developed with another phase of the analysis.

Comparison to other languages

When we have performed the comparison of the annotation schemes we have seen that the schemes for Italian and Spanish are quite similar to ours, at least at macro-operation level. So, in this subsection we present our results compared to those languages at that level and we will also try to compare the subsequent levels. We will also relate our results to the ones got in Brazilian Portuguese but that comparison is more difficult due to the facts that there is no annotation scheme and that they give the results according to the simplification levels.

The macro-operations that have been the most used in Spanish and in Italian are transformations, delete and insert (Bott and Saggion, 2014; Brunato et al., 2015). And those three macro-operations are the same that the teacher has mainly used. On the other hand, the translator has used in the second position the splits.

Looking at the percentages, the reordering operations performed in Basque and the insert are quite similar to the *teacher* and *terence* corpora. The proportion is smaller in the Spanish corpus. The less used macro-operation is also the same in the three languages: the merge or fusion. It is remarkable that the split has been broader used in Basque. The data used to make this comparison is presented in Table 7.23 and we have got the results for Spanish and Italian from these works: Bott and Saggion (2014) and Brunato et al. (2015), respectively.

Macro-operation	**Italian**		**Spanish**	**Basque**	
	Terence	**Teacher**		**Translator**	**Teacher**
Transformation	48.18	47.76	39.02	24.92	33.62
Split	1.71	2.06	12.20	23.55	12.30
Insert	18.72	15.66	12.60	21.88	18.61
Delete	21.94	25.32	24.80	17.66	20.78
Reordering	8.65	7.89	2.85	7.95	8.27
No_operation	-	-	-	3.53	6.20
Merge	0.81	1.30	0.81	0.40	0.22
Other	-	-	-	0.10	0.00

Table 7.23 – Comparison of macro-operations across languages

If we go a level down and analyse the transformations, we see that the most performed transformation types are lexical in Italian and Spanish. In the Brazilian Portuguese the lexical substitution has also been the most used when simplifying from original to natural simplification (Caseli et al., 2009). In our corpus, however, syntactical transformations have been the most applied.

Looking at the split phenomena, coordination has also been the most split in Spanish like in our corpus. It is difficult for us to compare other operations with the available data. Nevertheless, there is no doubt to confirm that transformations play an important role in text simplification.

7.5 Summary

In this chapter we have presented the corpus of Basque simplified texts ETSC-CBST. We have developed an annotation scheme where different macro-operations and operations have been compiled to know which happens when simplifying texts and also, to compare them across approaches. Although the first aim of this corpus was to compare our approach with other approaches, we have to mention that we have got a lot of useful information for a further development of the system and we can still learn from this corpus. Anyway, we also have put a basis with the common phenomena that will serve to enlarge the corpus. In Figure 7.2 we have added ETSC-CBST to the contributions.

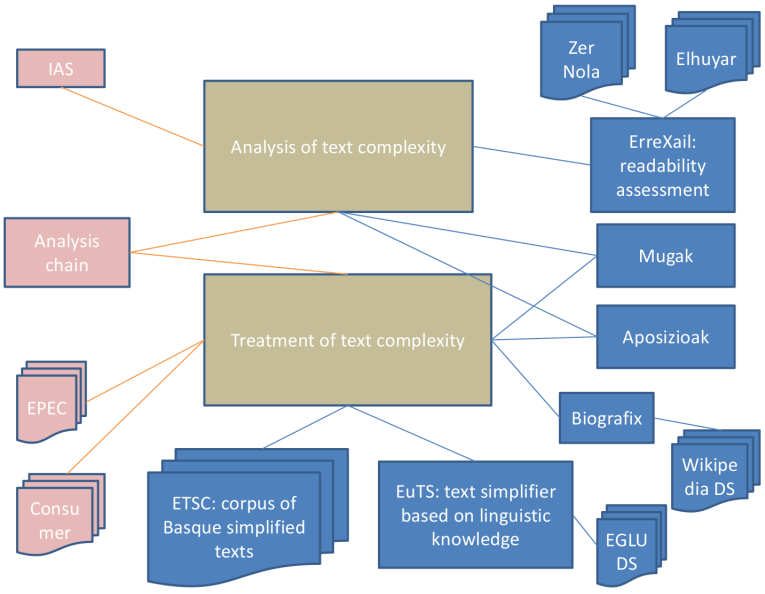

Figure 7.2 – Resources and tools used during thesis, and the contributions

CONCLUSION

8 Conclusion and Future Work

8.1 Introduction

In this thesis we have covered to research lines in NLP: Readability Assessment and Automatic Text Simplification. Our main work has been to analyse the complexity and the simplification of Basque written text based on other languages and to the extent possible, to bring Readability Assessment and Automatic Text Simplification to Basque NLP.

As we have pointed out in the motivation of the beginning of this report (Section 1.2), we have tried to solve the problems the long and/or complex sentences cause in NLP advanced applications and to offer to people learning Basque easier texts, analysing the reusability of the tools and resources developed by the Ixa research group. To that end, we have set two scopes: i) to make an analysis of the complexity and do simplification proposals and ii) to provide the readability assessment and text simplification system the required linguistic information. In the following sections we will present the contributions we have done to accomplish these scopes together with the open research lines and the future work.

8.2 Contributions

We will explain the main contributions of this thesis based on the groups of research questions presented in the introduction and in the research lines we have covered.

8.2.1 Analysis of text complexity and readability assessment

We have established the criteria to consider from a syntactic point of view a text as complex and we have **defined the Basque complex structures**. In order to define the Basque complex structures, we have based on the works done in other languages to begin our linguistic analysis. In our linguistic analysis the structures are found in coordinate clauses, subordinate clauses, apposition and parenthetical structures have been considered as complex. We have also taken into account the postpositional phrases that express thoughts or statements. So that these structures can be simplified, we have marked out that they should have a minimum length of two chunks plus the verb.

To assess the readability automatically, we have **implemented the readability assessment system ErreXail**. *ErreXail* determines whether texts are simple or complex based on 94 linguistic features and using the SMO classifier (Platt, 1998). In addition to that, we have made a list of the features that distinguish between simple and complex texts in the experiments performed with *ErreXail*.

This group of research questions has been addressed in Chapters 2 and 4.

8.2.2 Treatment of text complexity and automatic text simplification

To treat the text complexity we have **made the linguistic design of the EuTS system**. The *EuTS* system means to be a rule-based system that will apply the rules presented on our linguistic analysis. It performs two simplification types: syntactic substitution simplification and syntactic simplification. In the former, low frequency syntactic structures are substituted with more frequent ones and in the latter, structural changes are performed in order to get rid of the complex structures following the syntactic simplification process. To carry out those simplification types we have defined 5 operations: the shallow syntactic substitutions in the the syntactic substitution simplification and the splitting, reconstruction, reordering and correction and adequation in the syntactic simplification. To avoid the monotony that the added elements may cause in the reconstruction operation we have proposed the alternative added elements.

The target audience of *EuTS* system is quite open but we have defined three simplification levels for people learning Basque and NLP advanced ap-

plications. So, we have defined the three simplification levels levels for these target audiences: shallow syntactic simplification, natural simplification and absolute simplification. The shallow syntactic simplification level is intended to people who have no problem with Basque grammar but do not know dialectical and diachronic variations and to NLP tools or systems that have not got the dialectical and diachronic structures in their training. The natural simplification level is intended to people with and intermediate level of Basque and to NLP tools that get better results with shorter sentences. The absolute simplification level is intended to beginners and to NLP tools that process one verb per sentence.

As case study we have **implemented Biografix**. *Biografix* is a multilingual tool that simplifies parenthetical structures that contain biographical information based on the rules proposed in our linguistic analysis. With this tool we have seen that our operations and rules and useful for syntactic simplification. In addition to that, we also have checked that the rules defined for Basque can be also applied in other languages.

This group of research questions has been addressed in Chapters 5 and 6.

8.2.3 Resources

To confront our approach to other simplification approaches, we have **built the ETSC-CBST corpus**, a corpus of manually simplified texts. There, we have compiled three original texts and two simplified versions of each one. To analyse them, we have developed an annotation scheme. Based on that annotation-scheme we have analysed the operations that are performed when texts are manually simplified following different approaches. When possible, we have also related our work to the work done in other languages.

In relation to the basic tools, we have **created Mugak and Aposizioak**. To simplify the texts, basic tools that perform the analysis of the texts are required. For syntactic simplification purposes, tools that detect the clause and appositive boundaries are undeniable. Our main contributions to the basic tools are the improvement of the *MuGa* grammar that is included in *Mugak* and the development of the apposition detector *Aposizioak*. To evaluate these tools we also have created two gold standards. In addition to them, we also have **created corpora and datasets**. To train *ErreXail* we have compiled two corpora: Elhuyar T-comp and Zernola T-simp. The Elhuyar T-comp corpus has also been used in the linguistic analysis. To develop parts of *EuTS* and *Biografix* we also have compiled two datasets: Wikipedia DS

eta EGLU DS.

This group of research questions has been addressed in Chapters 5 and 7.

8.2.4 Comparison to other languages

After having analysed the works done in other languages, Basque needs the same resources (corpora and basic tools) as other lesser resourced languages. The challenge that Basque poses is that simplification rules are based on morphological features and this should be taken into account when implementing the modules of the architecture. But, in fact, as we have seen in the case study of parenthetical structures, rules defined for Basque can be applied to other languages. This group of research questions has been addressed through the report.

The main contributions of this thesis are presented in blue colour in Figure 8.1 together with the resources we have at the very beginning.

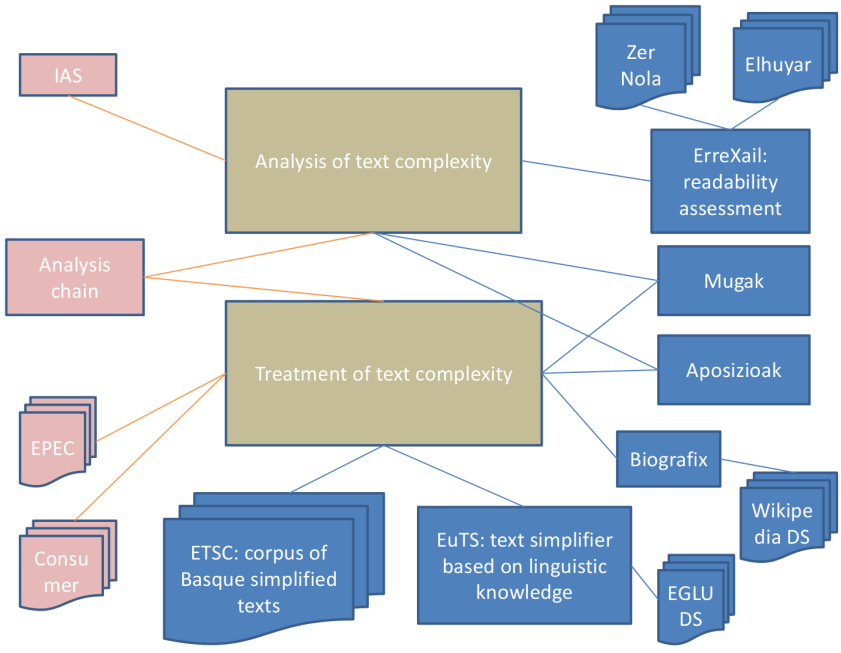

Figure 8.1 – Resources and tools used during thesis, and the contributions

So, as we have seen along this report, if we want to get simpler texts, first we perform i) the automatic linguistic analysis of the text. Second, we apply ii) the readability assessment system, and in the case the text be complex, finally, we apply iii) the automatic simplification. This process is shown in Figure 8.2.

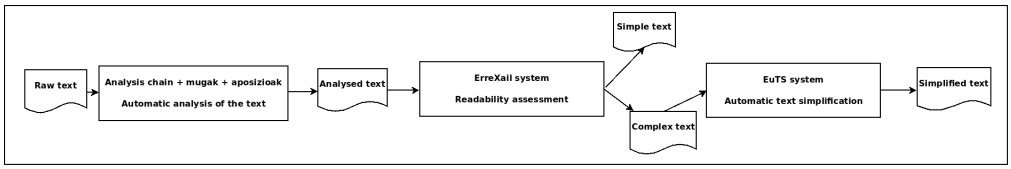

Figure 8.2 – Summary of the process to simplify a text

Now, we come back to the example presented in the introduction in Table 1.1 and in addition to the translation of the original sentence we give the simplified sentetences based on the contributions of this thesis and their machine tranlations[1] in Table 8.1.

Original sentence	**Translation of the original sentence**
1962an Charles De Gaulle eta Konrad Adenauer Bonnen elkartu zirenean 55 miloi lagun bizi ziren herrialde horretan, eta 47 milioi Frantzian.	Charles De Gaulle and Konrad Adenauer in Bonn, when 55 million people were living together in this country, and 47 million in France.
Simplified sentences	**Translation of the simplified sentences**
1962an Charles De Gaulle eta Konrad Adenauer Bonnen elkartu ziren. Orduan 55 miloi lagun bizi ziren herrialde horretan, eta 47 milioi Frantzian.	Charles De Gaulle and Konrad Adenauer in Bonn in 1962, attended the event. Then, 55 million people were living in the country, and 47 million in France.

Table 8.1 – Machine translations of an original and its respective simplified sentences

As we saw in the translation of the original sentences, the elementes of the clauses were mixed and the verb was not correctly translated. However, in the simplified sentences the verb *elkartu* has been translated as "attended

[1]The machine translation was done with Google Translate https://translate.google.es/ in February, 2013.

the event", which is not the perfect translation but it is acceptable. There are no other mistakes apart from a punctuation error and the use of a different determinant in the second simplified sentence (it should have been "in that country" instead of "in the country") in the simplified sentences. So, we think that if sentences are simplified before they are translated automatically, the quality of the output will be better.

8.3 Open research lines and future work

The research lines open in this thesis can be further exploited and we even could reuse them for other purposes.

- **Readability assessment**
  - Adding new features: more linguistic features can be added to *ErreXail* like the tags of the semantic analysis or frequency lists.
  - Classifying more readability levels: if we got more corpora, other readability levels could be assessed, e.g. the levels of Common European Framework of Reference for Languages.
  - Domain adaptation: it could also be trained with other domains using the texts of the Wikipedia and Vikidia.
  - Stylistic analysis of texts: the linguistic monitoring or profiling that is performed by *ErreXail* can be used for texts stylistics.
- **Automatic text simplification**
  - Implementation and evaluation: Looking at *EuTS*, we should finish its implementation and perform an evaluation.

    In addition to the intrinsic and extrinsic evaluations, we could perform cognitive experiments with users and target audience.

    The evaluation can also show us that improvements should be done in the basic tools. Indeed, the improvements of the basic tools like *Mugak* and *Aposizioak* is also a continuous and ceaseless work.

  - Complete the linguistic analysis: The implementation and evaluation may bring a redefining of the rules. The rules could also be adapted to domain specific needs and to geographical origin.

    To enlarge the linguistic analysis, more kind of parenthetical structures can be analysed, and as we saw in the results of ETSC-CBST, we should analyse postpositional structures. If we got bigger corpora, we could also broaden the quantitative corpus analysis of adverbial clauses and analyse the internal word and noun phrase ordering in the sentences..

  - Additional features: In ATS, in addition to integrating the lexical simplification, we can enrich the text by means of links to Wikipedia, maps, websites or adding definitions.

    Apart from that, instead of basing on the syntactical analysis, we can experiment with the semantic analysis.

    Appart from the three simplification levels we have presented, we can also provide tailored or customised simplification, where only needed or required phenomena will be simplified. In this case, rules will be applied depending on the needs of the target audience.

    In the M-Xuxen module of the system, the grammar checker and the coreference resoulution system (Soraluze et al., 2015) can also be integrated.

- **Analysis of manually simplified texts**

  - We should go a level deeper in the analysis of the ETSC-CBST corpus and also, we should make it bigger.
  - The results of this analysis can also be the basis to set priorities in the system.

In this Ph.D. thesis we analysed the Basque complex structures and we have open the way to simplify them automatically.

BIBLIOGRAPHY

Bibliography

Aduriz, I. (2000). *EUSMG: Morfologiatik sintaxira murriztapen gramatika erabiliz [EUSMG: From Morphology to Syntax using Constraint Grammar]*. PhD thesis, University of the Basque Country (UPV/EHU).

Aduriz, I., Agirre, E., Aldezabal, I., Alegria, I., Ansa, O., Arregi, X., Arriola, J. M., Artola, X., Díaz de Ilarraza, A., Ezeiza, N., Gojenola, K., Maritxalar, M., Oronoz, M., Sarasola, K., Soroa, A., Urizar, R., and Urkia, M. (1998). A framework for the automatic processing of Basque. In *Proceedings of Workshop on Lexical Resources for Minority Languages. First LREC Conference. Granada.*

Aduriz, I., Agirre, E., Aldezabal, I., Alegria, I., Arregi, X., Arriola, J. M., Artola, X., Gojenola, K., Maritxalar, M., Sarasola, K., and Urkia, M. (2000). A Word-grammar Based Morphological Analyzer for Agglutinative Languages. In *COLING, 18th International Conference on Computational Linguistics*. Universitaet des Saarlandes, Saarbrueucken, Germany, Morgan Kaufmann.

Aduriz, I., Aranzabe, M. J., Arriola, J. M., Atutxa, A., Díaz de Ilarraza, A., Ezeiza, N., Gojenola, K., Oronoz, M., Soroa, A., and Urizar, R. (2006a). Methodology and Steps Towards the Construction of EPEC, a Corpus of Written Basque Tagged at Morphological and Syntactic levels for Automatic Processing. *Language and Computers*, 56(1):1–15.

Aduriz, I., Aranzabe, M. J., Arriola, J. M., Díaz de Ilarraza, A., Gojenola, K., Oronoz, M., and Uria, L. (2004). A Cascaded Syntactic Analyser for Basque. *Computational Linguistics and Intelligent Text Processing*, pages 124–134.

Aduriz, I., Arrieta, B., Arriola, J. M., Díaz de Ilarraza, A., Izagirre, E., and Ondarra, A. (2006b). Muga Gramatikaren Optimizazioa [Optimization of the Clause Boundary Grammar]. Technical report, UPV/EHU/LSI/TR 9-2006.

Aduriz, I., Arriola, J. M., Artola, X., Díaz de Ilarraza, A., Gojenola, K., and Maritxalar, M. (1997). Morphosyntactic Disambiguation for Basque Based on the Constraint Grammar Formalism. In *Proceedings of Recent Advances on NLP (RANLP)*, pages 282–288, Tzigov Chark, Bulgary.

Aduriz, I., Arriola, J. M., Gonzalez-Dios, I., and Urizar, R. (2015). Funtzio Sintaktikoen Gold Estandarra eskuz etiketatzeko gidalerroak [Guidelines to Annotate the Gold-standard of Syntactic Functions]. Technical report, University of the Basque Country (UPV/EHU) UPV/EHU/LSI/TR 01-2015.

Aduriz, I. and Díaz de Ilarraza, A. (2003). Morphosyntactic Disambiguation and Shallow Parsing in Computational Processing of Basque. In *Inquiries into the lexicon-syntax relations in Basque*. Servicio Editorial de la Universidad del Pais Vasco-Euskal Herriko Unibertsirareko Argitalpen Zerbitzua.

Agirrezabal, M., Gonzalez-Dios, I., and Lopez-Gazpio, I. (2015). Euskararen Sorkuntza Automatikoa: lehen urratsak [Automatic Generation of Basque: First Steps]. In *I. Ikergazte Nazioarteko ikerketa euskaraz Kongresuko artikulu-bilduma*, pages 15–23.

Al-Subaihin, A. A. and Al-Khalifa, H. S. (2011). Al-Baseet: A proposed Simplification Authoring Tool for the Arabic Language. In *International Conference on Communications and Information Technology (ICCIT)*, pages 121–125.

Alcázar, A. (2005). Towards Linguistically Searchable Text. In *Proceedings of BIDE Summer School of Linguistics*.

Aldabe, I., Gonzalez-Dios, I., Lopez-Gazpio, I., Madrazo, I., and Maritxalar, M. (2013). Two Approaches to Generate Questions in Basque. *Procesamiento del Lenguaje Natural*, 51:101–108.

Aldezabal, I., Ansa, O., Arrieta, B., Artola, X., Ezeiza, A., Hernández, G., and Lersundi, M. (2001). EDBL: a General Lexical Basis for the Automatic Processing of Basque. In *Proceedings of the IRCS Workshop on linguistic databases*.

Aldezabal, I., Aranzabe, M. J., Atutxa, A., Gojenola, K., Sarasola, K., and Igone, Z. (2003). Hitz-hurrenkeraren azterketa masiboa corpusean [Massive Analysis of Word Ordering in Corpus]. Technical report, University of the Basque Country (UPV-EHU).

Alegria, I., Ansa, O., Artola, X., Ezeiza, N., Gojenola, K., and Urizar, R. (2004). Representation and Treatment of Multiword Expressions in Basque. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, pages 48–55. Association for Computational Linguistics.

Alegria, I., Ezeiza, N., Fernandez, I., and Urizar, R. (2003). Named Entity Recognition and Classification for Texts in Basque. In *II Jornadas de Tratamiento y Recuperación de Información, JOTRI, Madrid*.

Aluísio, S. M., Specia, L., Pardo, T. A., Maziero, E. G., and Fortes, R. P. (2008a). Towards Brazilian Portuguese Automatic Text Simplification Systems. In *Proceedings of the eighth ACM symposium on Document engineering*, DocEng '08, pages 240–248, New York, NY, USA. ACM.

Aluísio, S. M., Specia, L., Pardo, T. A. S., Maziero, E. G., Caseli, H. M., and Fortes, R. P. M. (2008b). A Corpus Analysis of Simple Account Texts and the Proposal of Simplification Strategies: First Steps Towards Text Simplification Systems. In *Proceedings of the 26th annual ACM international conference on Design of communication*, SIGDOC '08, pages 15–22, New York, NY, USA. ACM.

Amoia, M. and Romanelli, M. (2012). SB: mmSystem-Using Decompositional Semantics for 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 482–486. Association for Computational Linguistics.

Angrosh, M. and Siddharthan, A. (2014). Text Simplification Using Synchronous Dependency Grammars: Generalising Automatically Harvested Rules. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 16–25, Philadelphia, Pennsylvania, U.S.A. Association for Computational Linguistics.

Aranberri, N., Labaka, G., Díaz de Ilarraza, A., and Sarasola, K. (2014). Comparison of Post-editing Productivity between Professional Translators and Lay Users. In *Proceedings of Third Workshop on Post-Editing Technology and Practice*, pages 20–33.

Aranzabe, M. J. (2008). *Dependentzia-ereduan oinarritutako baliabide sintaktikoak: zuhaitz-bankua eta gramatika konputazionala [Syntactic Resources based on the Dependency Model: the Treebank and the Computational Grammar]*. PhD thesis, University of the Basque Country (UPV/EHU).

Aranzabe, M. J., Díaz de Ilarraza, A., and Gonzalez-Dios, I. (2012a). First Approach to Automatic Text Simplification in Basque. In Rello, L. and Saggion, H., editors, *Proceedings of the Natural Language Processing for Improving Textual Accessibility (NLP4ITA) workshop (LREC 2012)*, pages 1–8.

Aranzabe, M. J., Díaz de Ilarraza, A., and Gonzalez-Dios, I. (2013). Transforming Complex Sentences using Dependency Trees for Automatic Text Simplification in Basque. *Procesamiento de Lenguaje Natural*, 50:61–68.

Aranzabe, M. J., Kepa, B., de Ilarraza Arantza, D., Nerea, E., Goenaga, I., and Gojenola, K. (2012b). Combining Rule-Based and Statistical Syntactic Analyzers. In *Proceedings of the ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages (SP-Sem-MRL 2012), pp. 48-54, Association for Computational Linguistics (ACL), USA, ISBN: 978-1-937284-30-5, July 12, 2012, Jeju Island, Republic of Korea.*

Arrieta, B. (2010). *Azaleko sintaxiaren tratamendua ikasketa automatikoko tekniken bidez: euskarako kateen eta perpausen identifikazioa eta bere erabilera koma-zuzentzaile batean [The Treatment of the Surface Syntax with*

Machine Learning Techniques: the Identification of Basque Chunks and their Use in a Comma Checker]. PhD thesis, University of the Basque Country (UPV/EHU).

Bach, N., Gao, Q., Vogel, S., and Waibel, A. (2011). TriS: A Statistical Sentence Simplifier with Log-linear Models and Margin-based Discriminative Training. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 474–482.

Baeza-Yates, R., Rello, L., and Dembowski, J. (2015). CASSA: A Context-Aware Synonym Simplification Algorithm. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1385, Denver, Colorado. Association for Computational Linguistics.

Barlacchi, G. and Tonelli, S. (2013). ERNESTA: A Sentence Simplification Tool for Children's Stories in Italian. In *Computational Linguistics and Intelligent Text Processing*, pages 476–487. Springer.

Bautista, S., Drndarevic, B., Hervás, R., Saggion, H., and Gervás, P. (2012). Análisis de la Simplificación de Expresiones Numéricas en Español mediante un Estudio Empírico. *Linguamática*, 4(2):27–41.

Bautista, S., Hervás, R., Gervás, P., Power, R., and Williams, S. (2013). A System for the Simplification of Numerical Expressions at Different Levels of Understandability. In *Natural Language Processing for Improving Textual Accessibility (NLP4ITA 2013)*, pages 10–19.

Bautista, S. and Saggion, H. (2014). Making Numerical Information more Accessible. The Implementation of a Numerical Expression Simplification System for Spanish. In François, T. and Bernhard, D., editors, *International Journal of Applied Linguistics. Special Issue on Recent Advances in Automatic Readability Assessment and Text Simplification*, volume 165, pages 299–323. John Benjamins Publishing Company.

Bawakid, A. and Oussalah, M. (2011). Sentences Simplification for Automatic summarization. In *IEEE 10th International Conference on Cybernetic Intelligent Systems (CIS)*, pages 59–64. IEEE.

Beigman Klebanov, B., Knight, K., and Marcu, D. (2004). Text Simplification for Information-Seeking Applications. *On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE*, pages 735–747.

Bengoetxea, K. (2014). *Estaldura zabaleko euskararako analizatzaile sintaktiko estatistikoa [The Statistical Parser of Basque with Extensive Coverage]*. PhD thesis, University of the Basque Country (UPV/EHU).

Bengoetxea, K. and Gojenola, K. (2007). Desarrollo de un analizador sintáctico estadístico basado en dependencias para el euskera. *Procesamiento del lenguaje natural*, 39:5–12.

Biran, O., Brody, S., and Elhadad, N. (2011). Putting it Simply: a Context-Aware Approach to Lexical Simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 496–501, Portland, Oregon, USA. Association for Computational Linguistics.

Bott, S., Rello, L., Drndarevic, B., and Saggion, H. (2012a). Can Spanish Be Simpler? LexSiS: Lexical Simplification for Spanish. In *Proceedings of COLING*, pages 357–373.

Bott, S. and Saggion, H. (2011). An Unsupervised Alignment Algorithm for Text Simplification Corpus Construction. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, MTTG '11, pages 20–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bott, S. and Saggion, H. (2012). Automatic Simplification of Spanish Text for E-accessibility. In *Computers Helping People with Special Needs*, pages 527–534. Springer.

Bott, S. and Saggion, H. (2014). Text Simplification Resources for Spanish. *Language Resources and Evaluation*, 48(1):93–120.

Bott, S., Saggion, H., and Figueroa, D. (2012b). A Hybrid System for Spanish Text Simplification. In *Third Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, pages 75–84, Montreal, Canada.

Brouwers, L., Bernhard, D., Ligozat, A.-L., and François, T. (2012). Simplification syntaxique de phrases pour le français. In *Actes de la Conférence Conjointe JEP-TALN-RECITAL, Montpellier, France*, pages 211–224.

Brouwers, L., Bernhard, D., Ligozat, A.-L., and Francois, T. (2014). Syntactic Sentence Simplification for French. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 47–56, Gothenburg, Sweden. Association for Computational Linguistics.

Brunato, D., Dell'Orletta, F., Venturi, G., and Montemagni, S. (2015). Design and Annotation of the First Italian Corpus for Text Simplification. In *The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015*, pages 31–41.

Candido, Jr., A., Maziero, E., Gasperin, C., Pardo, T. A. S., Specia, L., and Aluísio, S. M. (2009). Supporting the Adaptation of Texts for Poor Literacy Readers: a Text Simplification Editor for Brazilian Portuguese. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, EdAppsNLP '09, pages 34–42, Stroudsburg, PA, USA. Association for Computational Linguistics.

Canning, Y. and Tait, J. (1999). Syntactic Simplification of Newspaper Text for Aphasic Readers. In *ACM SIGIR'99 Workshop on Customised Information Delivery*, pages 6–11. Citeseer.

Carroll, J., Minnen, G., Canning, Y., Devlin, S., and Tait, J. (1998). Practical Simplification of English Newspaper Text to Assist Aphasic Readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10. Citeseer.

Caseli, H. M., Pereira, T. F., Specia, L., Pardo, T. A. S., Gasperin, C., and Aluísio, S. (2009). Building a Brazilian Portuguese Parallel Corpus of Original and Simplified Texts. In *the Proceedings of CICLing*, pages 59–70.

Castro-Castro, D., Lannes-Losada, R., Maritxalar, M., Niebla, I., Pérez-Marqués, C., Alamo-Suarez, N. C., and Pons-Porrata, A. (2008). A Multilingual Application for Automated Essay Scoring. In *Lecture Notes in Advances in Artificial Intelligence - LNAI 5290 - IBERAMIA*, pages 243–251. Springer New York.

Chandrasekar, R., Doran, C., and Srinivas, B. (1996). Motivations and Methods for Text Simplification. In *Proceedings of the 16th Conference*

on Computational Linguistics - Volume 2, COLING '96, pages 1041–1044, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chandrasekar, R. and Srinivas, B. (1997). Automatic Induction of Rules for Text Simplification. *Knowledge-Based Systems*, 10(3):183–190.

Chen, H.-B., Huang, H.-H., Chen, H.-H., and Tan, C.-T. (2012). A Simplification-Translation-Restoration Framework for Cross-Domain SMT Applications. In *COLING*, pages 545–560.

Chung, J.-W., Min, H.-J., Kim, J., and Park, J. C. (2013). Enhancing Readability of Web Documents by Text Augmentation for Deaf People. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, WIMS '13, pages 30:1–30:10, New York, NY, USA. ACM.

Coster, W. and Kauchak, D. (2011). Learning to Simplify Sentences Using Wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, MTTG '11, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Crossley, S. A., Allen, D., and McNamara, D. S. (2012). Text Simplification and Comprehensible Input: A case for an Intuitive Approach. *Language Teaching Research*, 16(1):89–108.

Daelemans, W., Höthker, A., and Sang, E. T. K. (2004). Automatic Sentence Simplification for Subtitling in Dutch and English. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1045–1048.

Damay, J. J. S., Lojico, G. J. D., Lu, K. A. L., Tarantan, D. B., and Ong, E. C. (2006). SIMTEXT. Text Simplification of Medical Literature. In *3rd National Natural Language Processing Symposium - Building Language Tools and Resources*, pages 34–38.

De Belder, J., Deschacht, K., and Moens, M.-F. (2010). Lexical Simplification. In *Proceedings of Itec2010: 1st International Conference on Interdisciplinary Research on Technology, Education and Communication*.

De Belder, J. and Moens, M.-F. (2010). Text Simplification for Children. In *Proceedings of the SIGIR workshop on accessible search systems*, pages 19–26.

Doi, T. and Sumita, E. (2004). Splitting Input Sentence for Machine Translation Using Language Model with Sentence Similarity. In *Proc. of the 20th international conference on Computational Linguistics*.

Drndarević, B., Štajner, S., Bott, S., Bautista, S., and Saggion, H. (2013). Automatic Text Simplification in Spanish: A Comparative Evaluation of Complementing Modules. In *Computational Linguistics and Intelligent Text Processing*, pages 488–500. Springer.

Erdozia, K. (2006). Euskarazko hitz hurrenkera desberdinak prozesatzen [Processing Different Word Orders in Basque]. In Fernandez, B. and Laka, I., editors, *Essays in Honour of Professor Eguzkitza.* University of the Basque Country Press.

Euskaltzaindia (1999). V, (Mendeko perpausak-1, osagarriak, erlatiboak, konparaziozkoak, ondoriozkoak) [V (Subordinate Clauses-1, Completive, Relative, Comparative, Consecutive]. In *Euskal Gramatika Lehen Urratsak [Basque Grammar First Steps]*. Euskaltzaindia, Bilbo.

Euskaltzaindia (2002). *Euskal Gramatika Laburra: perpaus bakuna.* Euskaltzaindia, Bilbo.

Euskaltzaindia (2005). VI, (Mendeko perpausak-2, baldintzazkoak, denborazkoak, helburuzkoak, kausazkoak, kontseziozkoak eta moduzkoak) [VI (Subordinate Clauses-2, Conditional, temporal, Purpose, Causal, Concessive and Modal]. In *Euskal Gramatika Lehen Urratsak [Basque Grammar First Steps]*. Euskaltzaindia, Bilbo.

Euskaltzaindia (2011). VII, (Perpaus jokatugabeak: denborazkoak, kausazkoak eta helburuzkoak, baldintzazkoak, kontzesiozkoak, moduzkoak, erlatiboak eta osagarriak) [VII (Subordinate Clauses-2, temporal, Causal and Purpose, Conditional, Concessive, Modal, Relative and Completive]. In *Euskal Gramatika Lehen Urratsak [Basque Grammar First Steps]*. Euskaltzaindia, Bilbo.

Evans, R. J. (2011). Comparing Methods for the Syntactic Simplification of Sentences in Information Extraction. *Literary and linguistic computing*, 26(4):371–388.

Ezeiza, N. (2002). *Corpusak ustiatzeko tresna linguistikoak. Euskararen etiketatzaile morfosintaktiko sendo eta malgua.* PhD thesis, University of the Basque Country (UPV/EHU).

Fajardo, I., Tavares, G., Ávila, V., and Ferrer, A. (2013). Towards text simplification for poor readers with intellectual disability: When do connectives enhance text cohesion? *Research in Developmental Disabilities*, 34(4):1267–1279.

Feblowitz, D. and Kauchak, D. (2013). Sentence Simplification as Tree Transduction. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 1–10, Sofia, Bulgaria. Association for Computational Linguistics.

Fernandez Gonzalez, I. (2012). *Euskarazko Entitate-Izenak: identifikazioa, sailkapena, itzulpena eta desanbiguazioa.* PhD thesis, University of the Basque Country (UPV/EHU).

Gasperin, C., Maziero, E., Specia, L., Pardo, T. A., and Aluisio, S. M. (2009). Natural Language Processing for Social Inclusion: a Text Simplification Architecture for Different Literacy Levels. *the Proceedings of SEMISH-XXXVI Seminário Integrado de Software e Hardware*, pages 387–401.

Gojenola, K. (2000). *Euskararen sintaxi konputazionalerantz. Oinarrizko baliabideak eta beren aplikazioa aditzen azpikategorizazio-informazioaren erauzketan eta erroreen tratamenduan [Towards the Computational Syntax of Basque. Basic Resources and their Application in the Extraction of Verb Subcategorisation Information and Error Treatment].* PhD thesis, University of the Basque Country (UPV/EHU).

Gonzalez-Dios, I. (2011). Euskarazko egitura sintaktikoen azterketa testuen sinplifikazio automatikorako: Aposizioak, erlatibozko perpausak eta denborazko perpausak [Analysis of Basque Syntactic Structures for Automatic Text Simplification]. Master's thesis, University of the Basque Country (UPV/EHU).

Gonzalez-Dios, I. (2013). Euskarazko testuen sinplifikazio automatikoa [Automatic Simplification of Basque Texts]. Abstract and oral presentation.

Gonzalez-Dios, I. (2014a). Euskarazko testuak errazten: euskal testuen sinplifikazio automatikoa [Making Basque Texts Easier: Automatic Simplification of Basque Texts]. In Aduriz, I. and Urizar, R., editors, *Euskal Hizkuntzalaritzan egungo zenbait ikerlerro. Hizkuntzalari euskaldunen I. topaketa*, pages 135–149. Udako Euskal Unibertsitatea.

Gonzalez-Dios, I. (2014b). Simplificación automática de textos en euskera [automatic simplification of basque texts]. In Ureña López, L. A., Troyano Jiménez, J. A., Ortega Rodríguez, F. J., and Martínez Cámara, E., editors, *Actas de las V Jornadas TIMM, Cazalla de la Sierra, España, 12-JUN-2014, publicadas en http://ceur-ws.org*, pages 45–50.

Gonzalez-Dios, I., Aranzabe, M. J., and de Ilarraza, A. D. (2015a). Simplifying Basque Texts: the Shallow Syntactic Substitution Simplification. In *Proceedings the 7th Language & Technology Conference.*, pages 450–454.

Gonzalez-Dios, I., Aranzabe, M. J., de Ilarraza, A. D., and Soraluze, A. (2013a). Detecting Apposition for Text Simplification in Basque. In *Computational Linguistics and Intelligent Text Processing*, pages 513–524. Springer.

Gonzalez-Dios, I., Aranzabe, M. J., and Díaz de Ilarraza, A. (2013b). Testuen sinplifikazio automatikoa: arloaren egungo egoera [Automatic Text Simplification: State of Art]. *Linguamática*, 5(2):43–63.

Gonzalez-Dios, I., Aranzabe, M. J., and Díaz de Ilarraza, A. (2014a). Making Biographical Data in Wikipedia Readable: A Pattern-based Multilingual Approach. In *Proceedings of the Workshop on Automatic Text Simplification - Methods and Applications in the Multilingual Society (ATS-MA 2014)*, pages 11–20, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Gonzalez-Dios, I., Aranzabe, M. J., and Díaz de Ilarraza, A. (2015b). Perpaus adberbialen agerpena, maiztasuna eta kokapena EPEC-DEP corpusean [Presence, frequency and Position of Basque Adberbial Clauses in The BDT corpus]. Technical report, University of the Basque Country (UPV/EHU) UPV/EHU/LSI/TR 02-2015.

Gonzalez-Dios, I., Aranzabe, M. J., Díaz de Ilarraza, A., and Salaberri, H. (2014b). Simple or Complex? Assessing the Readability of Basque Texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 334–344, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Hidalgo, V. (1999). Hitz ordenaren estatistikak euskaraz [Statistics on the Order of Words in Basque]. *Anuario del Seminario de FilologÃŋa Vasca Julio de Urquijo*, 23(2):393–451.

Hualde, J. I. and Ortiz de Urbina, J., editors (2003). *A Grammar of Basque*. Mouton de Gruyter.

Inui, K., Fujita, A., Takahashi, T., Iida, R., and Iwakura, T. (2003). Text Simplification for Reading Assistance: A Project Note. In *Proceedings of the second international workshop on Paraphrasing-Volume 16*, pages 9–16. Association for Computational Linguistics.

Iruskieta, M., Aranzabe, M. J., Díaz de Ilarraza, A., Gonzalez-Dios, I., Lersundi, M., and Lopez de Lacalle, O. (2013). The RST Basque TreeBank: an Online Search Interface to Check Rhetorical Relations. In *Proceedings of the 4th Workshop RST and Discourse Studies*, pages 40–49.

Jauhar, S. K. and Specia, L. (2012). UOW-SHEF: SimpLex–Lexical Simplicity Ranking based on Contextual and Psycholinguistic Features. 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 477–481. Association for Computational Linguistics.

Johannsen, A., Martínez, H., Klerke, S., and Søgaard, A. (2012). EMNLP@ CPH: Is Frequency all there is to Simplicity? 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 408–412. Association for Computational Linguistics.

Jonnalagadda, S. and Gonzalez, G. (2010a). BioSimplify: an Open Source Sentence Simplification Engine to Improve Recall in Automatic Biomedical Information Extraction. In *AMIA Annual Symposium Proceedings*, volume 2010, pages 351–355. American Medical Informatics Association.

Jonnalagadda, S. and Gonzalez, G. (2010b). Sentence Simplification Aids Protein-Protein Interaction Extraction. *Arxiv preprint arXiv:1001.4273*.

Jonnalagadda, S., Tari, L., Hakenberg, J., Baral, C., and Gonzalez, G. (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.

Kajiwara, T. and Yamamoto, K. (2015). Evaluation Dataset and System for Japanese Lexical Simplification. In *Proceedings of the ACL-IJCNLP 2015 Student Research Workshop*, pages 35–40, Beijing, China. Association for Computational Linguistics.

Kandula, S., Curtis, D., and Zeng-Treitler, Q. (2010). A Semantic and Syntactic Text Simplification Tool for Health Content. In *AMIA Annual Symposium Proceedings*, volume 2010, pages 366–370. American Medical Informatics Association.

Kauchak, D. (2013). Improving Text Simplification Language Modeling Using Unsimplified Text Data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1537–1546, Sofia, Bulgaria. Association for Computational Linguistics.

Keskisärkkä, R. (2012). Automatic Text Simplification via Synonym Replacement. Master's thesis, Linköping.

Klerke, S. and Søgaard, A. (2012). DSim, a Danish Parallel Corpus for Text Simplification. In Calzolari (Conference Chair), N., Choukri, K., Declerck, T., Uğur Doğan, M., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 4015–4018, Istanbul, Turkey. European Language Resources Association (ELRA).

Klerke, S. and Søgaard, A. (2013). Simple, Readable Sub-sentences. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 142–149, Sofia, Bulgaria. Association for Computational Linguistics.

Laka, I. (1996). A Brief Grammar of Euskara, the Basque Language.

Leroy, G., Endicott, J. E., Kauchak, D., Mouradi, O., and Just, M. (2013). User Evaluation of the Effects of a Text Simplification Algorithm Using Term Familiarity on Perception, Understanding, Learning, and Information Retention. *Journal of medical Internet research*, 15(7).

Ligozat, A.-L., Garcia-Fernandez, A., Grouin, C., and Bernhard, D. (2012). ANNLOR: A Naïve Notation-system for Lexical Outputs Ranking. 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 487–492. Association for Computational Linguistics.

Ligozat, A.-L., Grouin, C., Garcia-Fernandez, A., and Bernhard, D. (2013). Approches à base de fréquences pour la simplification lexicale. In *Actes TALN-RÉCITAL 2013*, pages 493–506. ATALA.

Lozanova, S., Stoyanova, I., Leseva, S., Koeva, S., and Savtchev, B. (2013). Text Modification for Bulgarian Sign Language Users. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 39–48, Sofia, Bulgaria. Association for Computational Linguistics.

Madrazo, I. (2014). Testuen irakurgarritasuna neurtzeko sailkatzaile automatikoa [An Automatic Classiffier of Text Legibility]. Master's thesis, University of the Basque Country (UPV/EHU).

Max, A. (2006). Writing for Language-Impaired Readers. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 3878 of *Lecture Notes in Computer Science*, pages 567–570. Springer Berlin Heidelberg.

Medero, J. and Ostendorf, M. (2011). Identifying Targets for Syntactic Simplification. In *Proceedings of the SLaTE 2011 workshop*, pages 69–72.

Minard, A.-L., Ligozat, A.-L., and Grau, B. (2012). Simplification de phrases pour l'extraction de relations. In *Proceedings of the Joint Conference JEP-TALN-RÉCITAL 2012, volume 2: TALN*, pages 1–14, Grenoble, France. ATALA/AFCP.

Mitkov, R. and Štajner, S. (2014). The Fewer, the Better? A Contrastive Study about Ways to Simplify. In *Proceedings of the Workshop on Automatic Text Simplification - Methods and Applications in the Multilingual Society (ATS-MA 2014)*, pages 30–40, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Narayan, S. and Gardent, C. (2014). Hybrid Simplification using Deep Semantics and Machine Translation. In *the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 435–445.

Narayan, S. and Gardent, C. (2015). Unsupervised Sentence Simplification Using Deep Semantics. *arXiv preprint arXiv:1507.08452*.

Nunes, B. P., Kawase, R., Siehndel, P., Casanova, M. A., and Dietze, S. (2013). As Simple as It Gets - A Sentence Simplifier for Different Learning Levels and Contexts. In *Advanced Learning Technologies (ICALT), 2013 IEEE 13th International Conference on*, pages 128–132.

Ondarra, A. (2003). Murriztapen Gramatikaren sintaxia. EUSMG optimizatzen. Esaldi-mugak [The Syntax of Constraint Grammar. Optimising EUSMG. Clause Boundaries]. Master's thesis, University of the Basque Country (UPV/EHU).

Ong, E., Damay, J., Lojico, G., Lu, K., and Tarantan, D. (2007). Simplifying Text in Medical Literature. *J. Research in Science Computing and Eng*, 4(1):37–47.

Paetzold, G. (2015). Reliable Lexical Simplification for Non-Native Speakers. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 9–16, Denver, Colorado. Association for Computational Linguistics.

Paetzold, G. and Specia, L. (2015). LEXenstein: A Framework for Lexical Simplification. In *Proceedings of ACL-IJCNLP 2015 System Demonstra-*

tions, pages 85–90, Beijing, China. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.

Paetzold, G. H. and Specia, L. (2013). Text Simplification as Tree Transduction. In de Computaçao, S. B., editor, *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, pages 116–125.

Peng, Y., Tudor, C. O., Torii, M., Wu, C. H., and Vijay-Shanker, K. (2012). iSimp: A Sentence Simplification System for Biomedical Text. In *IEEE International Conference onBioinformatics and Biomedicine (BIBM)*, pages 1–6. IEEE.

Peng, Y., Tudor, C. O., Torii, M., Wu, C. H., and Vijay-Shanker, K. (2014). iSimp in BioC Standard Format: Enhancing the Interoperability of a Sentence Simplification System. *Database: The Journal of Biological Databases and Curation*.

Petersen, S. E. and Ostendorf, M. (2007). Text Simplification for Language Learners: A Corpus Analysis. In *In Proceedings of Workshop on Speech and Language Technology for Education. SLaTE*, pages 69–72. Citeseer.

Platt, J. C. (1998). Fast Training of Support Vector Machines using Sequential Minimal Optimization. In Schölkopf, B., Burges, C. J. C., and Smola, A. J., editors, *Advances in Kernel Methods-Support Vector Learning*. MIT Press.

Poornima, C., Dhanalakshmi, V., Anand, K., and Soman, K. (2011). Rule based Sentence Simplification for English to Tamil Machine Translation System. *International Journal of Computer Applications*, 25(8):38–42.

Rennes, E. and Jönsson, A. (2015). A Tool for Automatic Simplification of Swedish Texts. In *Nordic Conference of Computational Linguistics NODALIDA 2015*, pages 317–320.

Rybing, J., Smith, C., and Silvervarg, A. (2010). Towards a Rule Based System for Automatic Simplification of texts. In *The Third Swedish Language Technology Conference (SLTC 2010)*, pages 17–18.

Saggion, H., Bott, S., and Rello, L. (2013). Comparing Resources for Spanish Lexical Simplification. In *SLSP 2013: 1st International Conference on Statistical Language and Speech Processing*, pages 1–12. Springer.

Saggion, H., Bott, S., and Rello, L. (2016). Simplifying Words in Context. Experiments with two Lexical Resources in Spanish. *Computer Speech & Language*, 35:200 – 218.

Saggion, H., Gómez-Martínez, E., Etayo, E., Anula, A., and Bourg, L. (2011). Text Simplification in Simplext: Making Text More Accessible. *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 47:341–342.

Saggion, H., Štajner, S., Bott, S., Mille, S., Rello, L., and Drndarevic, B. (2015). Making It Simplext: Implementation and Evaluation of a Text Simplification System for Spanish. *ACM Trans. Access. Comput.*, 6(4):14:1–14:36.

Scarton, C., de Oliveira, M., Candido Jr, A., Gasperin, C., and Aluísio, S. M. (2010). SIMPLIFICA: a Tool for Authoring Simplified Texts in Brazilian Portuguese Guided by Readability Assessments. In *Proceedings of the NAACL HLT 2010 Demonstration Session*, pages 41–44. Association for Computational Linguistics.

Seretan, V. (2012). Acquisition of Syntactic Simplification Rules for French. In Calzolari (Conference Chair), N., Choukri, K., Declerck, T., Uğur Doğan, M., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 4019–426, Istanbul, Turkey. European Language Resources Association (ELRA).

Shardlow, M. (2012). Bayesian Lexical Simplification. Technical report, Short Taster Research Project. The University of Manchester.

Shardlow, M. (2013). A Comparison of Techniques to Automatically Identify Complex Words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 103–109, Sofia, Bulgaria. Association for Computational Linguistics.

Sheremetyeva, S. (2014). Automatic Text Simplification For Handling Intellectual Property (The Case of Multiple Patent Claims). In *Proceedings of*

the Workshop on Automatic Text Simplification - Methods and Applications in the Multilingual Society (ATS-MA 2014), pages 41–52, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Siddharthan, A. (2002). An Architecture for a Text Simplification System. In *Proceedings of the Language Engineering Conference (LEC'02)*, pages 64–71, Washington, DC, USA. IEEE Computer Society.

Siddharthan, A. (2006). Syntactic Simplification and Text Cohesion. *Research on Language & Computation*, 4(1):77–109.

Siddharthan, A. (2010). Complex Lexico-Syntactic Reformulation of Sentences using Typed Dependency Representations. In *Proceedings of the 6th International Natural Language Generation Conference*, pages 125–133. Association for Computational Linguistics.

Siddharthan, A. (2011). Text Simplification using Typed Dependencies: A Comparison of the Robustness of Different Generation Strategies. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 2–11. Association for Computational Linguistics.

Silveira Botelho, S. and Branco, A. (2012). Enhancing Multi-document Summaries with Sentence Simplification. In *ICAI 2012: International Conference on Artificial Intelligence*.

Sinha, R. (2012). UNT-SimpRank: Systems for Lexical Simplification Ranking. 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 493–496. Association for Computational Linguistics.

Soraluze, A., Arregi, O., Arregi, X., and Díaz de Ilarraza, A. (2015). Coreference Resolution for Morphologically Rich Languages. Adaptation of the Stanford System to Basque. 55:23–30.

Specia, L. (2010). Translating from Complex to Simplified Sentences. *Computational Processing of the Portuguese Language*, pages 30–39.

Specia, L., Aluísio, S. M., and Pardo, T. A. (2008). Manual de Simplificação Sintática para o Português. Technical Report NILC-TR-08-06, São Carlos-SP.

Specia, L., Jauhar, S. K., and Mihalcea, R. (2012). Semeval-2012 Task 1: English Lexical Simplification. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, pages 347–355.

Srivastava, J. and Sanyal, S. (2012). Segmenting Long Sentence Pairs to Improve Word Alignment in English-Hindi Parallel Corpora. In *Advances in Natural Language Processing*, pages 97–107. Springer.

Štajner, S. (2014). Translating Sentences from 'Original' to 'Simplified' Spanish. *Procesamiento del Lenguaje Natural*, 53:61–68.

Štajner, S. (2015). *New Data-Driven Approaches to Text Simplification*. PhD thesis, University of Wolverhampton.

Štajner, S., Calixto, I., and Saggion, H. (2015). Automatic Text Simplification for Spanish: Comparative Evaluation of Various Simplification Strategies. In *Proceedings of Recent Advances in Natural Language Processing*, pages 618–626.

Štajner, S., Drndarevic, B., and Saggion, H. (2013). Corpus-based Sentence Deletion and Split Decisions for Spanish Text Simplification. *Computacion y Systemas*, 17(2):251–262.

Štajner, S. and Saggion, H. (2015). Translating from Original to Simplified Sentences using Moses: When does it Actually Work? In *Proceedings of Recent Advances in Natural Language Processing*, pages 611–617.

Stenetorp, P., Pyysalo, S., Topic, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). BRAT: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, pages 102–107.

Temnikova, I. (2012). *Text Complexity and Text Simplification in the Crisis Management Domain*. PhD thesis, University of Wolverhampton.

Temnikova, I., Orasan, C., and Mitkov, R. (2012). CLCM - A Linguistic Resource for Effective Simplification of Instructions in the Crisis Management Domain and its Evaluations. In Calzolari (Conference Chair), N., Choukri,

K., Declerck, T., Uğur Doğan, M., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3007–3014, Istanbul, Turkey. European Language Resources Association (ELRA).

Thomas, S. R. and Anderson, S. (2012). WordNet-Based Lexical Simplification of a Document. In *Empirical Methods in Natural Language Processing*, pages 80–88.

Tur, G., Hakkani-Tur, D., Heck, L., and Parthasarathy, S. (2011). Sentence Simplification for Spoken Language Understanding. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5628–5631. IEEE.

Urizar, R. (2012). *Euskal lokuzioen tratamendu konputazionala [Computational Treatment of Basque Locutions]*. PhD thesis, University of the Basque Country (UPV/EHU).

Vertan, C. and von Hahn, W. (2014). Making Historical Texts Accessible to Everybody. In *Proceedings of the Workshop on Automatic Text Simplification - Methods and Applications in the Multilingual Society (ATS-MA 2014)*, pages 64–68, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Vickrey, D. and Koller, D. (2008). Sentence Simplification for Semantic Role Labeling. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-2008: HLT)*, pages 344–352.

Vu, T. T., Tran, G. B., and Pham, S. B. (2014). Learning to Simplify Children Stories with Limited Data. In *Intelligent Information and Database Systems*, pages 31–41. Springer.

Woodsend, K. and Lapata, M. (2011). Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 409–420, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wubben, S., van den Bosch, A., and Krahmer, E. (2012). Sentence Simplification by Monolingual Machine Translation. In *Proceedings of the 50th*

Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, pages 1015–1024. Association for Computational Linguistics.

Yatskar, M., Pang, B., Danescu-Niculescu-Mizil, C., and Lee, L. (2010). For the Sake of Simplicity: Unsupervised Extraction of Lexical Simplifications from Wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 365–368, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zabala, I. (2000). Hitz-hurrenkera euskara tekniko-zientifikoan [Word Ordering in Scientific Basque]. In *Ekaia: Euskal Herriko Unibertsitateko zientzi eta teknologi aldizkaria,*, volume 12, pages 143–166.

Zhu, Z., Bernhard, D., and Gurevych, I. (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.

APPENDIX

A

Structures of Adverbial Clauses

In this appendix we present the list of the structures of adverbial clauses we have made from the grammar *Euskal Gramatika Lehen Urratsak* (EGLU) ('Basque Grammar First Steps') (Euskaltzaindia, 1999, 2005, 2011). We present the structures according to their adverbial type and the finiteness of the verb.

Finite temporal
-enean
-ela
-elarik
Noiz eta...bait-/-en
-enetan
-en bakoitzean
-en guztietan
-en aldikal
-en aldiro
Zenbat aldiz -en... hainbat aldiz
-eneko
-en orduko
-en bezain laster
-en bezain sarri
-en bezain agudo
-en bezain fite
-en ber
-enaz batera
-en baino lehen
-en aurrean
-en aitzinean
-en ondoan
-en ostean
-enetik
-enez gero
-enik...-ra
-en arte
-eno
-eino
-en bitartean
-en artean
-en arteko
-enerako
-en heinean
-en momentuan

Non-finite temporal
-tzean
-tzerakoan
-tzekoan
-tzearekin
-tzeari
-tzerat
-tu(k)eran
-tu aldiro
-tu bakoitzean
-tu guztian
-tu ahala
-tu arau
-tzerako
-tu orduko
-tu bezain laster
-tu bezain pronto
-tu eta berehala
-tu eta laster
-tuaz
-tzearekin bat(era)
-tu berri(t)an
-tu ahala
-tu arau
-tu baino lehen
-tu aurretik
-tu aitzinean
-tu gabe
-tu eta
-tu eta gero
-tuta
-tutakoan
-tu ondoren
-tu ondoan
-tu ostean

-tu(a)z
-tuz gero
-turik
-tu arte
-tu artean
-tu bitartean
-tzeraino
-tu aitzin
-tu osteko
-tu bezperan
-tu ondotik
-tzear

Finite causal

-elako((t)z)
-elakoan
bait-
... eta
zeren eta ...(bait-/-(e)n)
zeren zergatik
Ezen (...) (bait-)
-enez gero(ztik)
-enez
-en legez
nola/zelan... (-en/bait)
-ela kausa
-ela bide
-ela medio

Non-finite causal

-tzeagatik
-tzearren
-turik
-tutakoan
-tuz
-tuta
-tuz gero
-tzearekin

Finite concessive

Nahiz (eta) -(en/-ela/ba-)
-en arren
ba- (...) ere

Non-finite concessive

nahiz eta ... -tu/0
-tu arren
-tugatik
-tuta (gabe/ezta) ere
-turik (gabe/ezta) ere
-tuz gero ere
-tzearren
-tuz gero
-ik ere

Finite modal

-ela
-elarik
-en moduan/ra
-en arabera(n)
-en eran/ra
-en antzera
-en moldean/ra
-en gisan/ra
-en bezala
-en legez
-enez

Non-finite modal

-tuz
-tuta
-turik
-tu gabe
-tu barik
-tu ezta
-tu ordez
-tu ordean
-tu aginean
-tu aginik
-tu ahala
-tu arau
-tu beharrean
-tu nahirik
-tu nahian
-tu gurarik
-tu ezinik
-tu ezinda
-tu ezinean
-tu beharrez
-tzeko zorian
-tu hurran
-tzeko moduan
-tzeko gisan
-tzeko eran
-tzeko maneran
-tzeko moldean
-tzekotan
-tu bezala
-tzeke
-era
-tzeaz
-tzeaz gain
-tik

Finite purpose

subjuntive

Non-finite purpose

-tzeko(tz)
-tzekotzat
-tzearren
-tzeagatik
-tze alde(ra)
-tzekotan
-tzeko asmotan
-tzeko intentziotan
-tzeko intentzioarekin
-tzera

Finite consecutive

(...) (non) ... bait-
(...) (ezen) ... -en

Finite conditional

ba-
non ez... -(e)n

Non-finite conditional

-tuz gero((z)tik)
-tu ezkero(an/k/tino/((z)tik))
-tu ezik
-tu ezean
-tzekotan
-tzekoz
-tzez gero
-tuenean
-tzera(t)

Syntactic Simplification Rules

In this appedix we present the syntactic simplification rules for Basque. The rules are presented in tables according to their phenomena or rule-set: coordination (Table B.1), relative clauses (Table B.2), noun clauses (Table B.3), apposition (Table B.4), parenthetical structures (Table B.5) and adverbial clauses (Table B.6).

The columns of the tables represent the following:

- Type: the type of the phenomena
- Structure: the target structure
- Remove: the relation mark (complementiser, case marker, postposition...) that should be removed (Relation_Marks_List)
- Add: the added element (First element, Added_Elements_List)
- Add2, Add3 and Add4: the alternative added elements (From the second element on, Added_Elements_List)
- Where add?: The clause where the added element should be added
- Ordering: The ordering of the simplified sentences (Reordering_List)
- Notes

Type	**Structure**	**Remove**	**Add**	**Where add?**	**Ordering**	**Notes**
Copulative	eta	eta	ø		coord1-coord2	
Adversative	baina	baina	Baina	2.coord	coord1-coord2	
Optative	edo	edo	Edo	2.coord	coord1-coord2	
Yustaposition	;	;	ø		coord1-coord2	
Yustaposition	,	,	ø		coord1-coord2	

Table B.1 – Syntactic simplification rules for Basque (coordination)

Type	**Structure**	**Remove**	**Add**	**Where add?**	**Ordering**	**Notes**
Common relatives	-en	-en	antecedent + determinant	beginning of the main	subordinate-$_{orig}$-main$_{orig}$	adjust case markers; if determinant, do not add it; if named-entity, do not add determinant
Zein relatives	zein	zein	antecedent + determinant	beginning of the subordinate	main$_{orig}$-subordinate-$_{orig}$	adjust case markers

Table B.2 – Syntactic simplification for Basque (relative clauses)

Type	**Structure**	**Remove**	**Add**	**Add2**	**Where add?**	**Ordering**	**Notes**
Completive	-ela	-ela	honako hau	ø	before the main verb	main_{orig}-$\text{subordinate}_{orig}$	pronomina change; verb person change; punctuation: finish the main with colon, subordinate between quotation marks
Completive	-enik	-enik	honako hau	ø	before the main verb	main_{orig}-$\text{subordinate}_{orig}$	pronomina change; verb person change; punctuation: finish the main with colon, subordinate between quotation marks
Completive	-tzen	-tzen	honako hau	ø	before the main verb	main_{orig}-$\text{subordinate}_{orig}$	if predicative function, do not simplify; finish main clause with colon, begin subordinate clause with lower case
Completive	-tzera	-tzera	honako hau	ø	before the main verb	main_{orig}-$\text{subordinate}_{orig}$	if predicative function, do not simplify; finish main clause with colon, begin subordinate clause with lower case

(Continued on next page)

Type	**Structure**	**Remove**	**Add**	**Add2**	**Where add?**	**Ordering**	**Notes**
Completive	-tzeko	-tzeko	honako hau	ø	before the main verb	main_{orig}-$\text{subordinate}_{orig}$	if predicative function, do not simplify; finish main clause with colon, begin subordinate clause with lower case
Completive	-tzea + atzizkia	-tzea + atzizkia	honako hau	ø	before the main verb	main_{orig}-$\text{subordinate}_{orig}$	if predicative function, do not simplify; finish main clause with colon, begin subordinate clause with lower case
Completive	-tzeari	-tzeari	honako honi	ø	before the main verb	main_{orig}-$\text{subordinate}_{orig}$	finish main clause with colon, begin subordinate clause with lower case
Completive	-tzerik	-tzerik	honako hau	ø	before the main verb	main_{orig}-$\text{subordinate}_{orig}$	if predicative function, do not simplify; finish main clause with colon, begin subordinate clause with lower case
Completive	-tu izana	-tu izana	honako hau	ø	before the main verb	main_{orig}-$\text{subordinate}_{orig}$	if predicative function, do not simplify; finish main clause with colon, begin subordinate clause with lower case

(Continued on next page)

Type	**Structure**	**Remove**	**Add**	**Add2**	**Where add?**	**Ordering**	**Notes**
Completive	-tu izanari	-tu izanari	honako honi	ø	before the main verb	main_{orig}-$\text{subordinate}_{orig}$	finish main clause with colon, begin subordinate clause with lower case
Indirect questions	-en	-en	honako hau	ø	before the main verb	main_{orig}-$\text{subordinate}_{orig}$	delete *ea*; pronomina change; verb person change; punctuation: finish the main with colon, subordinate between quotation marks
Indirect questions	galdetza-ileak ... -en	-en	honako hau	ø	before the main verb	main_{orig}-$\text{subordinate}_{orig}$	put interrogatives at the beginning of the subordinate clause; delete *ea*; pronomina change; verb person change; punctuation: finish the main with colon, subordinate between quotation marks
Indirect questions	non-finite	ø	ø	ø	ø	ø	DO NOT SIMPLIFY
Other	-enez + reporting verb	-enez	honako hau	ø	before the main verb	$\text{subordinate}_{orig}$-$\text{main}_{orig}$	pronomina change; verb person change; punctuation: finish the main with colon, subordinate between quotation marks

(Continued on next page)

Type	**Structure**	**Remove**	**Add**	**Add2**	**Where add?**	**Ordering**	**Notes**
Postpositional	-en arabera	-en arabera	honako hau	dio/te	before the main verb; txert2 perpaus amaieran	subordinate$_{orig}$-main$_{orig}$	word in genitive must be animated; if genitive removed, add ergative; be careful with sing/pl when chosing the verb; punctuation: finish the main with colon, subordinate between quotation marks
Postpositional	-en araberan	-en araberan	honako hau	dio/te	before the main verb; txert2 perpaus amaieran	subordinate$_{orig}$-main$_{orig}$	word in genitive must be animated; if genitive removed, add ergative; be careful with sing/pl when chosing the verb; punctuation: finish the main with colon, subordinate between quotation marks
Postpositional	-en arau	-en arau	honako hau	dio/te	before the main verb; txert2 perpaus amaieran	subordinate$_{orig}$-main$_{orig}$	word in genitive must be animated; if genitive removed, add ergative; be careful with sing/pl when chosing the verb; punctuation: finish the main with colon, subordinate between quotation marks

(Continued on next page)

Type	**Structure**	**Remove**	**Add**	**Add2**	**Where add?**	**Ordering**	**Notes**
Postpositional	-en arauaz	-en arauaz	honako hau	dio/te	before the main verb; txert2 perpaus amaieran	subordinate$_{orig}$-main$_{orig}$	word in genitive must be animated; if genitive removed, add ergative; be careful with sing/pl when chosing the verb; punctuation: finish the main with colon, subordinate between quotation marks
Postpositional	-en arauka	-en arauka	honako hau	dio/te	before the main verb; txert2 perpaus amaieran	subordinate$_{orig}$-main$_{orig}$	word in genitive must be animated; if genitive removed, add ergative; be careful with sing/pl when chosing the verb; punctuation: finish the main with colon, subordinate between quotation marks
Postpositional	-en ereduz	-en ereduz	honako hau	dio/te	before the main verb; txert2 perpaus amaieran	subordinate$_{orig}$-main$_{orig}$	word in genitive must be animated; if genitive removed, add ergative; be careful with sing/pl when chosing the verb; punctuation: finish the main with colon, subordinate between quotation marks

(Continued on next page)

Type	**Structure**	**Remove**	**Add**	**Add2**	**Where add?**	**Ordering**	**Notes**
Postpositional	-en eredura	-en eredura	honako hau	dio/te	before the main verb; txert2 perpaus amaieran	subordinate$_{orig}$-main$_{orig}$	word in genitive must be animated; if genitive removed, add ergative; be careful with sing/pl when chosing the verb; punctuation: finish the main with colon, subordinate between quotation marks
Postpositional	-en hitzetan	-en hitzetan	honako hau	esan du/te	before the main verb; txert2 perpaus amaieran	subordinate$_{orig}$-main$_{orig}$	word in genitive must be animated; if genitive removed, add ergative; be careful with sing/pl when chosing the verb; punctuation: finish the main with colon, subordinate between quotation marks

(Continued on next page)

Type	Structure	Remove	Add	Add2	Where add?	Ordering	Notes
Postpositional	-en adier-azpenetan	-en adier-azpenetan	honako hau	adierazi du/te	before the main verb; txert2 perpaus amaieran	subordinate$_{orig}$-main$_{orig}$	word in genitive must be animated; if genitive removed, add ergative; be careful with sing/pl when chosing the verb; punctuation: finish the main with colon, subordinate between quotation marks

Table B.3 – Syntactic simplification rules for Basque (noun clauses)

Type	Structure	Remove	Add	Where add?	Ordering	Notes
Inside NP	entity + explicative appositive	No longer needed case markers	da/dira	in the subordinate clause	main$_{orig}$-subordinate$_{orig}$	internal ordering: entity+ explicative appositive + verb; be careful with sing/pl when chosing the verb
Inside NP	explicative appositive + entity	No longer needed case markers	da/dira	in the subordinate clause	main$_{orig}$-subordinate$_{orig}$	internal ordering: entity + explicative appositive + verb; be careful with sing/pl when chosing the verb

(Continued on next page)

Type	**Structure**	**Remove**	**Add**	**Where add?**	**Ordering**	**Notes**
Appositive NP		No longer needed case markers	da/dira	in the subordinate clause	main_{orig}-$\text{subordinate}_{orig}$	internal ordering: entity + explicative appositive + verb; be careful with sing/pl when chosing the verb

Table B.4 – Syntactic simplification rules for Basque (apposition)

Type	**Structure**	**Remove**	**Add**	**Add2**	**Where add?**	**Ordering**	**Notes**
Place information		inessive in the main clause	dago/daude	inessive	in the subordinate clause	main_{orig}-$\text{subordinate}_{orig}$	internal ordering: place + parent. place + verb; be careful with sing/pl when chosing the verb
Birth information		No longer needed case markers	jaio zen	inessive	in the subordinate clause	main_{orig}-$\text{subordinate}_{orig}$	internal ordering: entity + birthday + , + birthplace + verb; add inessive in place and date
Death information		No longer needed case markers	hil zen	inessive	in the subordinate clause	main_{orig}-$\text{subordinate}_{orig}$	internal ordering: entity + death day + , + death place + verb; add inessive in place and date

Table B.5 – Syntactic simplification rules for Basque (parenthetical structures)

Type	**Structure**	**Remove**	**Add**	**Add2**	**Add3**	**Add4**	**Where add?**	**Ordering**	**Notes**
Temporal	-enean	-enean	Orduan	Une hartan	Aldi berean	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-ela	-ela	Orduan	Une hartan	Aldi berean	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-elarik	-elarik	Orduan	Une hartan	Aldi berean	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	noiz eta... bait-	noiz eta... bait-	Orduan	Une hartan	Aldi berean	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	noiz eta... -en	noiz eta... -en	Orduan	Une hartan	Aldi berean	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tzean	-tzean	Orduan	Une hartan	Aldi berean	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tzerakoan	-tzerakoan	Orduan	Une hartan	Aldi berean	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tzekoan	-tzekoan	Orduan	Une hartan	Aldi berean	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tzearekin	-tzearekin	Orduan	Une hartan	Aldi berean	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	

(Continued on next page)

Type	**Structure**	**Remove**	**Add**	**Add2**	**Add3**	**Add4**	**Where add?**	**Ordering**	**Notes**
Temporal	-tzeari	-tzeari	Orduan	Une hartan	Aldi berean	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tzerat	-tzerat	Orduan	Une hartan	Aldi berean	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tueran	-tueran	Orduan	Une hartan	Aldi berean	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tukeran	-tukeran	Orduan	Une hartan	Aldi berean	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-enetan	-enetan	Une horietan guztietan	Aldiro	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-en bakoitzean	-en bakoitzean	Une horietan guztietan	Aldiro	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-en guztietan	-en guztietan	Une horietan guztietan	Aldiro	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-en aldikal	-en aldikal	Une horietan guztietan	Aldiro	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-en aldiro	-en aldiro	Une horietan guztietan	Aldiro	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	

(Continued on next page)

Type	**Structure**	**Remove**	**Add**	**Add2**	**Add3**	**Add4**	**Where add?**	**Ordering**	**Notes**
Temporal	zenbat aldiz -en... hainbat aldiz	zenbat aldiz -en... hainbat aldiz	Une horietan guztietan	Aldiro	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tu aldiro	-tu aldiro	Une horietan guztietan	Aldiro	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tu bakoitzean	-tu bakoitzean	Une horietan guztietan	Aldiro	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tu guztian	-tu guztian	Une horietan guztietan	Aldiro	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tu ahala	-tu ahala	Une horietan guztietan	Aldiro	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tu arau	-tu arau	Une horietan guztietan	Aldiro	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-eneko	-eneko	Orduko	Segidan	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-en orduko	-en orduko	Orduko	Segidan	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tzerako	-tzerako	Orduko	Segidan	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	

(Continued on next page)

Type	**Structure**	**Remove**	**Add**	**Add2**	**Add3**	**Add4**	**Where add?**	**Ordering**	**Notes**
Temporal	-tu orduko	-tu orduko	Orduko	Segidan	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-en bezain laster	-en bezain laster	Une horretan bertan	Orduko	Segidan	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-en bezain sarri	-en bezain sarri	Une horretan bertan	Orduko	Segidan	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-en bezain agudo	-en bezain agudo	Une horretan bertan	Orduko	Segidan	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-en bezain fite	-en bezain fite	Une horretan bertan	Orduko	Segidan	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-en ber	-en ber	Une horretan bertan	Orduko	Segidan	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-enaz batera	-enaz batera	Une horretan bertan	Orduko	Segidan	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tu bezain laster	-tu bezain laster	Une horretan bertan	Orduko	Segidan	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tu bezain pronto	-tu bezain pronto	Une horretan bertan	Orduko	Segidan	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tu eta berehala	-tu eta berehala	Une horretan bertan	Orduko	Segidan	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	

(Continued on next page)

Type	**Structure**	**Remove**	**Add**	**Add2**	**Add3**	**Add4**	**Where add?**	**Ordering**	**Notes**
Temporal	-tu eta laster	-tu eta laster	Une horretan bertan	Orduko	Segidan	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tuaz	-tuaz	Une horretan bertan	Orduko	Segidan	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tzearekin bat	-tzearekin bat	Une horretan bertan	Orduko	Segidan	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tzearekin batera	-tzearekin batera	Une horretan bertan	Orduko	Segidan	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tu berrian	-tu berrian	Une horretan bertan	Orduko	Segidan	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tu berritan	-tu berritan	Une horretan bertan	Orduko	Segidan	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tu ahala	-tu ahala	Une horretan bertan	Orduko	Segidan	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tu arau	-tu arau	Une horretan bertan	Orduko	Segidan	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-en baino lehen	-en baino lehen	Gero	Ondoren	Ostean	ø	beginning of the main	main$_{orig}$-subordinate$_{orig}$	
Temporal	-en aurrean	-en aurrean	Gero	Ondoren	Ostean	ø	beginning of the main	main$_{orig}$-subordinate$_{orig}$	

(Continued on next page)

Type	**Structure**	**Remove**	**Add**	**Add2**	**Add3**	**Add4**	**Where add?**	**Ordering**	**Notes**
Temporal	-en aitzinean	-en aitzinean	Gero	Ondoren	Ostean	ø	beginning of the main	$main_{orig}$-$subordinate_{orig}$	
Temporal	-tu baino lehen	-tu baino lehen	Gero	Ondoren	Ostean	ø	beginning of the main	$main_{orig}$-$subordinate_{orig}$	
Temporal	-tu aurretik	-tu aurretik	Gero	Ondoren	Ostean	ø	beginning of the main	$main_{orig}$-$subordinate_{orig}$	
Temporal	-tu aintzinean	-tu aintzinean	Gero	Ondoren	Ostean	ø	beginning of the main	$main_{orig}$-$subordinate_{orig}$	
Temporal	-tu gabe	-tu gabe	Gero	Ondoren	Ostean	ø	beginning of the main	$main_{orig}$-$subordinate_{orig}$	
Temporal	-tu orduko	-tu orduko	Gero	Ondoren	Ostean	ø	beginning of the main	$main_{orig}$-$subordinate_{orig}$	Ambiguous
Temporal	-tzerako	-tzerako	Gero	Ondoren	Ostean	ø	beginning of the main	$main_{orig}$-$subordinate_{orig}$	Ambiguous
Temporal	-en ondoan	-en ondoan	Ondoren	Ostean	ø	ø	beginning of the main	$subordinate_{orig}$-$main_{orig}$	
Temporal	-en ondoren	-en ondoren	Ondoren	Ostean	ø	ø	beginning of the main	$subordinate_{orig}$-$main_{orig}$	
Temporal	-en ostean	-en ostean	Ondoren	Ostean	ø	ø	beginning of the main	$subordinate_{orig}$-$main_{orig}$	

(Continued on next page)

Type	**Structure**	**Remove**	**Add**	**Add2**	**Add3**	**Add4**	**Where add?**	**Ordering**	**Notes**
Temporal	-tu eta	-tu eta	Ondoren	Ostean	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tu eta gero	-tu eta gero	Ondoren	Ostean	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tuta	-tuta	Ondoren	Ostean	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tu ondoan	-tu ondoan	Ondoren	Ostean	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tu ostean	-tu ostean	Ondoren	Ostean	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tuz	-tuz	Ondoren	Ostean	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tuaz	-tuaz	Ondoren	Ostean	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tuz gero	-tuz gero	Ondoren	Ostean	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-turik	-turik	Ondoren	Ostean	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-enetik	-enetik	Ordutik	Une hartatik	Harrezkero	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	

(Continued on next page)

Type	**Structure**	**Remove**	**Add**	**Add2**	**Add3**	**Add4**	**Where add?**	**Ordering**	**Notes**
Temporal	-enez gero	-enez gero	Ordutik	Une hartatik	Harrezkero	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-enik -ra	-enik -ra	Ordutik	Une hartatik	Harrezkero	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tuz gero	-tuz gero	Ordutik	Une hartatik	Harrezkero	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	Ambiguous
Temporal	-en arte	-en arte	Ordura arte	Orduraino	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tu arte	-tu arte	Ordura arte	Orduraino	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tu bitartean	-tu bitartean	Ordura arte	Orduraino	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tzeraino	-tzeraino	Ordura arte	Orduraino	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-eno	-eno	Bitartean	Artean	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-eino	-eino	Bitartean	Artean	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-en bitartean	-en bitartean	Bitartean	Artean	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	

(Continued on next page)

Type	**Structure**	**Remove**	**Add**	**Add2**	**Add3**	**Add4**	**Where add?**	**Ordering**	**Notes**
Temporal	-en artean	-en artean	Bitartean	Artean	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-en arteko	-en arteko	Bitartean	Artean	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Temporal	-tu bitarte	-tu bitarte	Bitartean	Artean	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	Ambiguous
Temporal	-tu bitartean	-tu bitartean	Bitartean	Artean	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	Ambiguous
Temporal	-tu artean	-tu artean	Bitartean	Artean	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	Ambiguous
Causal	-elako	-elako	Horregatik	Hori dela eta	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Causal	-elakoz	-elakoz	Horregatik	Hori dela eta	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Causal	-elakotz	-elakotz	Horregatik	Hori dela eta	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Causal	-elakoan	-elakoan	Horregatik	Hori dela eta	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Causal	bait-	bait-	Horregatik	Hori dela eta	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	

(Continued on next page)

Type	**Structure**	**Remove**	**Add**	**Add2**	**Add3**	**Add4**	**Where add?**	**Ordering**	**Notes**
Causal	zeren eta ... bait-	zeren eta ... bait-	Horregatik	Hori dela eta	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Causal	zeren eta ... -en	zeren eta ... -en	Horregatik	Hori dela eta	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Causal	- tzeagatik	- tzeagatik	Horregatik	Hori dela eta	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Causal	-tzearren	-tzearren	Horregatik	Hori dela eta	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Causal	bait-	bait-	Izan ere	ø	ø	ø	beginning of the main	main$_{orig}$-subordinate$_{orig}$	Explicative
Causal	... eta	... eta	Izan ere	ø	ø	ø	beginning of the main	main$_{orig}$-subordinate$_{orig}$	Explicative
Causal	zeren eta ... bait-	zeren eta ... bait-	Izan ere	ø	ø	ø	beginning of the main	main$_{orig}$-subordinate$_{orig}$	Explicative
Causal	zeren eta ... -en	zeren eta ... -en	Izan ere	ø	ø	ø	beginning of the main	main$_{orig}$-subordinate$_{orig}$	Explicative
Consessive	nahiz eta... -en	nahiz eta -en	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	beginning of the main	subordinate$_{orig}$-main$_{orig}$	

(Continued on next page)

Type	Structure	Remove	Add	Add2	Add3	Add4	Where add?	Ordering	Notes
Consessive	nahiz eta... -ela	nahiz eta -ela	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	beginning of the main	subordi-nate$_{orig}$-main$_{orig}$	
Consessive	nahiz eta... ba-	nahiz eta ba-	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	beginning of the main	subordi-nate$_{orig}$-main$_{orig}$	
Consessive	nahiz... -en	nahiz -en	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	beginning of the main	subordi-nate$_{orig}$-main$_{orig}$	
Consessive	nahiz... -ela	nahiz -ela	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	beginning of the main	subordi-nate$_{orig}$-main$_{orig}$	
Consessive	nahiz... ba-	nahiz ba-	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	beginning of the main	subordi-nate$_{orig}$-main$_{orig}$	
Consessive	-en arren	-en arren	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	beginning of the main	subordi-nate$_{orig}$-main$_{orig}$	
Consessive	ba- ere	ba- ere	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	beginning of the main	subordi-nate$_{orig}$-main$_{orig}$	

(Continued on next page)

Type	**Structure**	**Remove**	**Add**	**Add2**	**Add3**	**Add4**	**Where add?**	**Ordering**	**Notes**
Consessive	nahiz eta -tu	nahiz eta -tu	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Consessive	-tu arren	-tu arren	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Consessive	-tuagatik	-tuagatik	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Consessive	-tuta ere	-tuta ere	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Consessive	-tuta gabe ere	-tuta gabe ere	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Consessive	-tuta ezta ere	-tuta ezta ere	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Consessive	-tzearren	-tzearren	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	beginning of the main	subordinate$_{orig}$-main$_{orig}$	

(Continued on next page)

Type	Structure	Remove	Add	Add2	Add3	Add4	Where add?	Ordering	Notes
Consessive	-tuz gero	-tuz gero	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Consessive	-ik ere	-ik ere	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Modal	-ela	-ela	Hala	Horrela	Era berean	Modu horretan	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Modal	-elarik	-elarik	Hala	Horrela	Era berean	Modu horretan	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Modal	-en moduan	-en moduan	Hala	Horrela	Era berean	Modu horretan	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Modal	-en modura	-en modura	Hala	Horrela	Era berean	Modu horretan	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Modal	-en antzera	-en antzera	Hala	Horrela	Era berean	Modu horretan	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Modal	-en bezala	-en bezala	Hala	Horrela	Era berean	Modu horretan	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Modal	-en legez	-en legez	Hala	Horrela	Era berean	Modu horretan	beginning of the main	subordinate$_{orig}$-main$_{orig}$	

(Continued on next page)

Type	**Structure**	**Remove**	**Add**	**Add2**	**Add3**	**Add4**	**Where add?**	**Ordering**	**Notes**
Modal	-enez	-enez	Hala	Horrela	Era berean	Modu horretan	beginning of the main	subordinate$_{orig}$-main$_{orig}$	without reporting verbs
Modal	-tuz	-tuz	Hala	Horrela	Era berean	Modu horretan	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Modal	-tuta	-tuta	Hala	Horrela	Era berean	Modu horretan	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Modal	-turik	-turik	Hala	Horrela	Era berean	Modu horretan	beginning of the main	subordinate$_{orig}$-main$_{orig}$	
Modal	-tu gabe	-tu gabe	Hala	Horrela	Era berean	Modu horretan	beginning of the main	subordinate$_{orig}$-main$_{orig}$	polarity change in the subordinate clause
Modal	-tu barik	-tu barik	Hala	Horrela	Era berean	Modu horretan	beginning of the main	subordinate$_{orig}$-main$_{orig}$	polarity change in the subordinate clause
Modal	-tu ezta	-tu ezta	Hala	Horrela	Era berean	Modu horretan	beginning of the main	subordinate$_{orig}$-main$_{orig}$	polarity change in the subordinate clause
Modal	-tu ordez	-tu ordez	Hala	Horrela	Era berean	Modu horretan	beginning of the main	subordinate$_{orig}$-main$_{orig}$	polarity change in the subordinate clause
Modal	-tu ordean	-tu ordean	Hala	Horrela	Era berean	Modu horretan	beginning of the main	subordinate$_{orig}$-main$_{orig}$	polarity change in the subordinate clause
Modal	-tzeke	-tzeke	Hala	Horrela	Era berean	Modu horretan	beginning of the main	subordinate$_{orig}$-main$_{orig}$	polarity change in the subordinate clause

(Continued on next page)

Type	Structure	Remove	Add	Add2	Add3	Add4	Where add?	Ordering	Notes
Modal	-tu beharrean	-tu beharrean	Hala	Horrela	Era berean	Modu horretan	beginning of the main	subordinate$_{orig}$-main$_{orig}$	polarity change in the subordinate clause
Modal	-tu aginean	-tu aginean	Hala	Horrela	Era berean	Modu horretan	beginning of the main	subordinate$_{orig}$-main$_{orig}$	add *ia* in the subordinate clause
Modal	-tu aginik	-tu aginik	Hala	Horrela	Era berean	Modu horretan	beginning of the main	subordinate$_{orig}$-main$_{orig}$	add *ia* in the subordinate clause
Modal	-tzeko zorian	-tzeko zorian	Hala	Horrela	Era berean	Modu horretan	beginning of the main	subordinate$_{orig}$-main$_{orig}$	add *ia* in the subordinate clause
Modal	-tu hurran	-tu hurran	Hala	Horrela	Era berean	Modu horretan	beginning of the main	subordinate$_{orig}$-main$_{orig}$	add *ia* in the subordinate clause; Ambiguous
Modal	-tu beharrean	-tu beharrean	Hala	Horrela	Era berean	Modu horretan	beginning of the main	subordinate$_{orig}$-main$_{orig}$	add *behar izan* in the subordinate clause
Modal	-tu beharrez	-tu beharrez	Hala	Horrela	Era berean	Modu horretan	beginning of the main	subordinate$_{orig}$-main$_{orig}$	add *behar izan* in the subordinate clause
Modal	-tu nahirik	-tu nahirik	Hala	Horrela	Era berean	Modu horretan	beginning of the main	subordinate$_{orig}$-main$_{orig}$	add *nahi izan* in the subordinate clause
Modal	-tu nahian	-tu nahian	Hala	Horrela	Era berean	Modu horretan	beginning of the main	subordinate$_{orig}$-main$_{orig}$	add *nahi izan* in the subordinate clause

(Continued on next page)

Type	**Structure**	**Remove**	**Add**	**Add2**	**Add3**	**Add4**	**Where add?**	**Ordering**	**Notes**
Modal	-tu gu-rarik	-tu gu-rarik	Hala	Horrela	Era berean	Modu horre-tan	beginning of the main	subordi-nate$_{orig}$-main$_{orig}$	add *nahi izan* in the subordinate clause
Modal	-tu ezinik	-tu ezinik	Hala	Horrela	Era berean	Modu horre-tan	beginning of the main	subordi-nate$_{orig}$-main$_{orig}$	add *ezin izan* in the subordinate clause
Modal	-tu ezinda	-tu ezinda	Hala	Horrela	Era berean	Modu horre-tan	beginning of the main	subordi-nate$_{orig}$-main$_{orig}$	add *ezin izan* in the subordinate clause
Modal	-tu ezinean	-tu ezinean	Hala	Horrela	Era berean	Modu horre-tan	beginning of the main	subordi-nate$_{orig}$-main$_{orig}$	add *ezin izan* in the subordinate clause
Modal	-tu ahala	-tu ahala	Hala	Horrela	Era berean	Modu horre-tan	beginning of the main	subordi-nate$_{orig}$-main$_{orig}$	add impf aspect to the verb of the subordinate clause
Modal	-tu arau	-tu arau	Hala	Horrela	Era berean	Modu horre-tan	beginning of the main	subordi-nate$_{orig}$-main$_{orig}$	add impf aspect to the verb of the subordinate clause
Modal	-tzeko moduan	ø	ø	ø	ø	ø	ø	ø	DO NOT SIM-PLIFY
Modal	-tzeko gisan	ø	ø	ø	ø	ø	ø	ø	DO NOT SIM-PLIFY
Modal	-tzeko eran	ø	ø	ø	ø	ø	ø	ø	DO NOT SIM-PLIFY
Modal	-tzeko maneran	ø	ø	ø	ø	ø	ø	ø	DO NOT SIM-PLIFY
Modal	-tzeko moldean	ø	ø	ø	ø	ø	ø	ø	DO NOT SIM-PLIFY

(Continued on next page)

Type	**Structure**	**Remove**	**Add**	**Add2**	**Add3**	**Add4**	**Where add?**	**Ordering**	**Notes**
Modal	-tu bezala	ø	ø	ø	ø	ø	ø	ø	DO NOT SIMPLIFY
Modal	-tzekotan	ø	ø	ø	ø	ø	ø	ø	Only syntactic substitution simplification
Consecutive	non... bait-	non... bait-	Ondorioz	Beraz	Hortaz	Honenbestez	beginning of the subordinate	$main_{orig}$-$subordinate_{orig}$	quantifier change
Consecutive	ezen... bait-	ezen... bait-	Ondorioz	Beraz	Hortaz	Honenbestez	beginning of the subordinate	$main_{orig}$-$subordinate_{orig}$	quantifier change
Consecutive	non... -en	non... -en	Ondorioz	Beraz	Hortaz	Honenbestez	beginning of the subordinate	$main_{orig}$-$subordinate_{orig}$	quantifier change
Consecutive	ezen... -en	ezen... -en	Ondorioz	Beraz	Hortaz	Honenbestez	beginning of the subordinate	$main_{orig}$-$subordinate_{orig}$	quantifier change
Purpose	subjuntiboa	subjuntiboa	nahi izan	gura izan	ø	ø	in the subordinate clause	$main_{orig}$-$subordinate_{orig}$	nominalise the verb of the subordinate; aditz laguntzailea ezabatu
Purpose	-tzeko	-tzeko	nahi izan	gura izan	ø	ø	in the subordinate clause	$main_{orig}$-$subordinate_{orig}$	put the verb of the subordinate as participle

(Continued on next page)

Type	**Structure**	**Remove**	**Add**	**Add2**	**Add3**	**Add4**	**Where add?**	**Ordering**	**Notes**
Purpose	-tzekotz	-tzekotz	nahi izan	gura izan	ø	ø	in the subordinate clause	$main_{orig}$-$subordinate_{orig}$	put the verb of the subordinate as participle
Purpose	-tzekotzat	-tzekotzat	nahi izan	gura izan	ø	ø	in the subordinate clause	$main_{orig}$-$subordinate_{orig}$	put the verb of the subordinate as participle
Purpose	-tzearren	-tzearren	nahi izan	gura izan	ø	ø	in the subordinate clause	$main_{orig}$-$subordinate_{orig}$	put the verb of the subordinate as participle
Purpose	-tzeagatik	-tzeagatik	nahi izan	gura izan	ø	ø	in the subordinate clause	$main_{orig}$-$subordinate_{orig}$	put the verb of the subordinate as participle
Purpose	-tze alde	-tze alde	nahi izan	gura izan	ø	ø	in the subordinate clause	$main_{orig}$-$subordinate_{orig}$	put the verb of the subordinate as participle
Purpose	-tze aldera	-tze aldera	nahi izan	gura izan	ø	ø	in the subordinate clause	$main_{orig}$-$subordinate_{orig}$	put the verb of the subordinate as participle
Purpose	-tzekotan	ø	ø	ø	ø	ø	ø	ø	Only syntactic substitution simplification
Purpose	-tzeko asmotan	ø	ø	ø	ø	ø	ø	ø	DO NOT SIMPLIFY

(Continued on next page)

Type	**Structure**	**Remove**	**Add**	**Add2**	**Add3**	**Add4**	**Where add?**	**Ordering**	**Notes**
Purpose	-tzeko intentziotan	ø	ø	ø	ø	ø	ø	ø	DO NOT SIMPLIFY
Purpose	-tzeko intentzioarekin	ø	ø	ø	ø	ø	ø	ø	DO NOT SIMPLIFY
Conditional	ba- (realpresent)	ba-	Demagun	Suposa dezagun	Kasu horretan	ø	add and add2 in the subordinate; add3 in the main	subordinate$_{orig}$-main$_{orig}$	delete *baldin*
Conditional	ba- (realpast)	ba-	Demagun	Suposa dezagun	Kasu horretan	ø	add and add2 in the subordinate; add3 in the main	subordinate$_{orig}$-main$_{orig}$	delete *baldin*
Conditional	ba- (unreal)	ba-	Bestela	ø	ø	ø	beginning of the main	subordinate$_{orig}$-main$_{orig}$	polarity change in the subordinate clause; delete *baldin*
Conditional	non eta ez... -en	ø	ø	ø	ø	ø	ø	ø	Correlations
Conditional	non ez... -en	ø	ø	ø	ø	ø	ø	ø	Correlations
Conditional	-tuz gero	ø	ø	ø	ø	ø	ø	ø	Only syntactic substitution simplification
Conditional	-tuz geroz	ø	ø	ø	ø	ø	ø	ø	Only syntactic substitution simplification

(Continued on next page)

Type	**Structure**	**Remove**	**Add**	**Add2**	**Add3**	**Add4**	**Where add?**	**Ordering**	**Notes**
Conditional	-tuz geroztik	ø	ø	ø	ø	ø	ø	ø	Only syntactic substitution simplification
Conditional	-tu ezkero	ø	ø	ø	ø	ø	ø	ø	Only syntactic substitution simplification
Conditional	-tu ezkeroan	ø	ø	ø	ø	ø	ø	ø	Only syntactic substitution simplification
Conditional	-tu ezkerok	ø	ø	ø	ø	ø	ø	ø	Only syntactic substitution simplification
Conditional	-tu ezkerotino	ø	ø	ø	ø	ø	ø	ø	Only syntactic substitution simplification
Conditional	-tu ezkeroz	ø	ø	ø	ø	ø	ø	ø	Only syntactic substitution simplification
Conditional	-tu ezkeroztik	ø	ø	ø	ø	ø	ø	ø	Only syntactic substitution simplification
Conditional	-tu ezik	ø	ø	ø	ø	ø	ø	ø	Only syntactic substitution simplification
Conditional	-tu ezean	ø	ø	ø	ø	ø	ø	ø	Only syntactic substitution simplification
Conditional	-tzekotan	ø	ø	ø	ø	ø	ø	ø	Only syntactic substitution simplification

(Continued on next page)

Type	Structure	Remove	Add	Add2	Add3	Add4	Where add?	Ordering	Notes
Conditional	-tzekoz	ø	ø	ø	ø	ø	ø	ø	Only syntactic substitution simplification
Conditional	-tzez gero	ø	ø	ø	ø	ø	ø	ø	Only syntactic substitution simplification
Conditional	-tuenean	ø	ø	ø	ø	ø	ø	ø	Only syntactic substitution simplification
Conditional	-tzera	ø	ø	ø	ø	ø	ø	ø	Only syntactic substitution simplification
Conditional	-tzerat	ø	ø	ø	ø	ø	ø	ø	Only syntactic substitution simplification

Table B.6 – Syntactic simplification rules for Basque (adverbial clauses)

C

Compulsory Operations to Enlarge the ETSC-CBST Corpus

In this appendix we present the operations that should be performed compulsorily to enlarge manually the ETSC-CBST corpus.

- Perform above all syntactic transformations
- Split coordinate clauses and create new sentences out of them
- Split adverbial clauses, paying attention specially to concessive and causal clauses, and create new sentences out of them
- Split relative clauses and create new sentences out of them
- Split postpositional phrases and create new sentences out of them
- Give complementary information (make the information explicit, give definitions)
- Make explicit ellided arguments (coreference resolution)
- Correct the errors in the original texts

The writing of this thesis
was ended on 17th May, 2016.