Lang Resources & Evaluation (2018) 52:217–247
https://doi.org/10.1007/s10579-017-9407-6

The corpus of Basque simplified texts (CBST)

Itziar Gonzalez-Dios[1] · **María Jesús Aranzabe**[1] · **Arantza Díaz de Ilarraza**[1]

Published online: 18 November 2017

Abstract In this paper we present the corpus of Basque simplified texts. This corpus compiles 227 original sentences of science popularisation domain and two simplified versions of each sentence. The simplified versions have been created following different approaches: the structural, by a court translator who considers easy-to-read guidelines and the intuitive, by a teacher based on her experience. The aim of this corpus is to make a comparative analysis of simplified text. To that end, we also present the annotation scheme we have created to annotate the corpus. The annotation scheme is divided into eight macro-operations: delete, merge, split, transformation, insert, reordering, no operation and other. These macro-operations can be classified into different operations. We also relate our work and results to other languages. This corpus will be used to corroborate the decisions taken and to improve the design of the automatic text simplification system for Basque.

Keywords Text simplification · Monolingual parallel corpora · Annotation scheme · Basque

1 Introduction

In the information society millions of texts are produced every day, but not all the texts are easy to understand for certain people due to their complexity. Adapting these texts manually is a difficult and expensive task. For that reason, research on text simplification and automatic evaluation of complexity has gained attention in the last years. A way to comprehend which knowledge lies under simplification

✉ Itziar Gonzalez-Dios
itziar.gonzalezd@ehu.eus

[1] Ixa NLP Group, University of the Basque Country (UPV/EHU), Manuel Lardizabal Pasealekua, 1, 20018 Donostia, Spain

strategies and how to evaluate their complexity is to analyse corpus of simplified texts.

Corpora of simplified text can be understood as text collections where each original text has its simplified counterpart. These texts form what can be called a monolingual parallel corpus, since most of the sentences in each version should be related. The goal of this kind of corpora is, therefore, to compile simplified versions of a text that vary according to their difficulty.

The simplified texts can be oriented to different levels and target audiences and can be created following either intuitive approaches or structural approaches (Crossley et al. 2012). On the one hand, intuitive approaches rely on the experience and intuition of the teacher or the expert who is simplifying the text. On the other hand, structural approaches are used to create graded readings. This way, predefined word and structure lists are used to adapt the texts to the required level. In this approach, readability formulae are also used to check the complexity of the texts candidate to be simplified. Readability formulae take into account features such as syllable, word and sentence length or lexical lists, to mention a few. These criteria are close to those that are used when designing the rules to be implemented in knowledge-based automatic text simplifications systems.

The corpus we are presenting here is the corpus of Basque simplified texts (CBST), or *Euskarazko Testu Sinplifikatuen Corpusa* (ETSC) in Basque. The aim of CBST is to make an analysis of the characteristics of simplified texts in Basque, compare them with those found in simplified text for other languages, and analyse the results structural and intuitive simplification strategies produce. With that aim in mind, we have chosen 227 sentences in the domain of science popularisation and two language experts with different backgrounds have simplified them. We have manually analysed the simplified texts and identified quantitatively and qualitatively the similarities and differences found. In addition, an annotation scheme has been proposed to analyse and compare them.

This corpus will also be used to evaluate the decisions taken so far in the design of the automatic simplification system for Basque (Gonzalez-Dios 2016). Indeed, we want to see if the common results or similarities of both approaches have been considered in the annotation process. The results of the comparison between both approaches will also be used to improve the system. To our knowledge, this is the first corpus in Basque where simplification strategies have been annotated and analysed and one of the first corpora where the same text has been simplified following different approaches.

This paper is structured as follows: in Sect. 2 we present the related work. In Sect. 3 the corpus building and annotation are explained. In Sect. 4 we describe the annotation scheme. In Sect. 5 we give the annotation results and trends. Finally, Sect. 6 presents some conclusions and future work.

2 Related work

In this section we expose the notion of text complexity related to readability assessment and text simplification. We also describe corpora of simplified texts, and corpora that compile simple and complex texts. Finally, we present the resources for Basque.

The analysis of text complexity is very important in human communication and human–computer interaction. Particularly, providing graded or adapted texts to audiences such as people with impairment, low-literate or foreign language learners help them to get access to the information.

To measure text complexity, several approaches have been proposed. From a psycho- and neurolinguistic point of view, Rosenberg and Abbeduto (1987) designed a seven level scale (D-scale) to measure the indicators of linguistic performance in English of mildly retarded adults. D-scale has been revised by Covington et al. (2006) and automated by Lu (2009). Phenomena such as subordination (level 6 in D-scale) and several different embeddings in a single sentence (level 7 in D-scale) are to find in the highest levels of the D-scale. Other studies have focused on, e.g. to know how the referential processing (Warren and Gibson 2002) or the noun phrase types (Gordon et al. 2004) affect sentence complexity. In Basque Neurolinguistics the relative clauses (Carreiras et al. 2010), the internal word reordering (Laka and Erdozia 2010) and the phrasal length (Ros et al. 2015) have been studied so far in relation to sentence complexity.[1]

The study of text complexity in the educational domain has focused on readability assessment. The readability of the texts has been studied over decades and applied by means of formulae such as Flesh (Flesch 1948), Dale-Chall (Chall and Dale 1995) and Gunning FOG index (Gunning 1968). These formulae take into account raw features (word and sentence number), lexical features and word frequencies and are language-dependent.

Readability assessment has also been treated from a computational point of view. Computing facilities and Natural Language Processing (NLP) applications make possible a more sophisticated (taking into account more features) and faster analysis of the complexity. Usually, an analysis of several linguistic and statistical features such as word types, dependencies or n-grams is performed and then machine learning techniques are applied in order to determine the complexity grade of the text. Surveys about readability assessment techniques can be found at DuBay (2004), Benjamin (2012) and Zamanian and Heydari (2012).

Reducing the complexity of the texts to the required level of the target is the task of Text Simplification (TS). This can be done following intuitive or structural approaches. In NLP, Automatic Text Simplification (ATS) aims to automatise or semi-automatise this task. To build these systems, rule-based strategies or data-driven approaches are followed. While the former has been the strategy used in the early works and in lesser resourced languages, the latter has been more frequent in

[1] According to Carreiras et al. (2010) subject relative clauses are harder to process than object relative clauses. Laka and Erdozia (2010) claim that the canonical word order of Basque (SOV) is processed faster and with greater ease and Ros et al. (2015) find a long-before-short preference and tendency, when the constituent is long, to place the verb in a sentence-medial position.

the last years for English. Detailed surveys about ATS can be found in the works by Gonzalez-Dios et al. (2013), Shardlow (2014) and Siddharthan (2014). In both approaches corpora of simplified texts are needed (not necessarily parallel) (1) to write and revise the rules and (2) to learn them automatically or establish weights and priorities among them.

In order to perform simplification studies, corpora of simplified texts are usually needed. These monolingual parallel corpora contain aligned texts of different complexity: there is usually the original or complex text and its simplified version or versions. Corpora of simplified texts have been built for languages such as English (Petersen and Ostendorf 2007; Pellow and Eskenazi 2014; Xu et al. 2015), Brazilian Portuguese (Caseli et al. 2009), Spanish (Bott and Saggion 2011, 2014; Štajner 2015), Danish (Klerke and Søgaard 2012), German (Klaper et al. 2013) and Italian (Brunato et al. 2015). The aims of building these corpora are (1) to study the process of simplifying texts, and (2) to use them as resources to build machine learning systems and evaluations.

The strategies to create the simplified texts are different in the mentioned corpora. In the case of Petersen and Ostendorf (2007), their corpus has been built by a literacy organization (Literacyworks[2]) whose target audience is language learners and adult literacy learners. Xu et al. (2015) present the Newsela corpus which is motivated by the Common Core Standards guidelines (the English level required for each grade). Each text of the Newsela corpus has associated with four simplifications (each one corresponding to a language level) proposed by professional editors. The Brazilian Portuguese corpus (Caseli et al. 2009) compiles texts from a newspaper which edits, for each text, its corresponding simplified version for children. In this corpus two levels of simplification are compiled: natural simplification and strong simplification. The process of simplification is performed by linguist experts in text simplification. The same happens in the Danish corpus referred to in Klerke and Søgaard (2012) that has been created by journalists trained in simplification. In that corpus, the texts are simplified targeting reading-impaired adults and adults learning Danish. The Spanish corpus (Bott and Saggion 2011, 2014; Štajner et al. 2013; Štajner 2015) has been created following easy-to-read guidelines adapted for people with cognitive disabilities. The German corpus (Klaper et al. 2013) is built with texts from websites that have been adapted to people with disabilities. The Italian corpus (Brunato et al. 2015) is divided into two sub-corpora created under a different simplification approaches: the Terence sub-corpus, targeted towards children, follows the structural approach and the Teacher sub-corpus follows the intuitive approach, has been simplified by teachers. Finally, Pellow and Eskenazi (2014) present a corpus of everyday documents and plan to enlarge the corpus using crowdsourced simplifications.

To analyse these corpora common statistics (e.g. average sentence length) and readability assessment measures have been used. These statistics, however, do not reflect directly the changes or operations that are performed to simplify the texts. This is done by annotating the changes performed when simplifying. To our knowledge, the operations performed in the simplification are only presented in the case of the Brazilian Portuguese corpus (Caseli et al. 2009), the Spanish corpus

[2] http://literacynet.org/cnnsf/index_cnnsf.html (2004–2007) (last accessed 11th April, 2016).

(Bott and Saggion 2014) and the Italian corpus (Brunato et al. 2015) but only in the cases of the Spanish and Italian corpora, these operations are organised in annotation schemes.

Apart from the simplified corpora, monolingual corpora containing complex or normal texts and simple texts have also been used in readability assessment and in automatic text simplification. These corpora (Brouwers et al. 2014; Coster and Kauchak 2011; Dell'Orletta et al. 2011; Hancke et al. 2012) contain instances of normal or complex language and simple language, but these texts are not related. That is, although the texts may be about the same topic the simple texts has not been created/simplified from the normal or complex ones. We consider these corpora as monolingual non-parallel corpora. To create the non-parallel corpora, resources like simple Wikipedia, Vikidia, newspapers or magazines for children have been used. These corpora can give us models in order to determine simple or normal/complex languages in order to determine which structures can be used in simple or normal/complex texts.

Concerning Basque, we would like to point out two resources: (1) the Elhuyar and the Zernola corpora used in training of the readability assessment for Basque *ErreXail*[3] (Gonzalez-Dios et al. 2014) and (2) the Basque Vikidia.[4] The Elhuyar corpus and the Zernola corpus compile texts from the science popularization domain; the former is for adults and the latter for children. We can consider this resource as a non-parallel monolingual corpus. The Basque Vikidia is a collaborative project to create an encyclopaedia for children aged 8–13 which was launched in the summer of 2015. Nowadays, it has around 350 articles and according to its promoter most of them are translations from other Vikidias. So, the corpus Zernola and the Basque Vikidia can be considered as instances of simple language.

3 Corpus building and annotation

The original texts we have used to be simplified are part of the Elhuyar corpus that was used to train the *ErreXail* system (Gonzalez-Dios et al. 2014). We selected 227 sentences corresponding to long texts from different topics: social sciences, medicine and technology. We decided to use long texts instead of short ones to see the continuity of the simplification operations on the same topic. We differentiated between three phrases to create the corpus:

1. *Starting phase* a text from each topic has been simplified to see whether these texts fit for this task. A list of basic operations (changes carried out to create the simplified text) performed has been created based on that simplification and on other languages. This list of operations and brief description of them builds the

[3] *ErreXail* (Gonzalez-Dios et al. 2014) classifies texts as simple or complex based on 96 linguistic features and machine learning techniques. Its main function is to determine which texts should be simplified.

[4] https://eu.vikidia.org/wiki/Azala (last accessed 18th March, 2016).

CBTS-annotationScheme-v0. Operations such as split clauses, substitute synonyms, or reorder clauses are defined. In total, there are 16 operations.

2. *Comparison phase* a text of each topic has been given to two different persons in order to be simplified: a court translator who has never worked on simplification before and a language teacher who used to simplify texts for learners of Basque as a foreign language. The translator was given easy-to-read guidelines and the operations covered by CBTS-annotationScheme-v0 annotation scheme to help her (structural approach). These guidelines were inspired by Mitkov and Štajner (2014): use simple and short sentences, resolve anaphora, use only high frequency words, use always the same word to refer to a concept. Based on the analysis of the previous phase, we also added 4 criteria to the guidelines: (1) keep the logical and chronological ordering, (2) recover elided arguments (if needed), (3) recover elided verbs, (4) and use only one finite verb in each sentence. The teacher followed her intuition and experience (intuitive approach). This phase has different aims:

   a. Look for common criteria when simplifying
   b. Compare structural and intuitive approaches
   c. Improve the CBTS-annotationScheme-v0 with new operations or specify them

   To achieve these aims, quantitative and qualitative analyses of the corpus have been performed until the definitive annotation scheme has been created. The outcome of this phase is the corpus and the annotation scheme (CBTS-annotationScheme-v1) we are presenting in this paper. At this phase, we also compare our annotation scheme to the schemes in other languages (Sect. 4.2).

3. *Extension phase* the corpus will be enlarged applying the common criteria.

The comparison phase of the annotation process is divided in two sub steps:

1. *Exploratory analysis of the tagging* we tagged the texts at paragraph level based on the operation list extracted from the starting phase. We identified and classified the new phenomena that were not covered (classified as others) in the CBTS-annotationScheme-v0 and we created a new set of operations (CBTS-annotationScheme-v1). This improved set has 31 operations and it is divided in lexical, syntactic and discourse level operations. We also detected several operations to get information about how to treat the ellipsis and the treatment of the information contained in the sentences. We compared the CBTS-annotationScheme-v1 to the Italian operations and annotation scheme (Brunato et al. 2015) as it was the one that fitted best to our study.
2. *Definitive analysis of the tagging* we tagged and analysed the texts at sentence level, following the definitive annotation scheme (see Sect. 4). The tool we used to annotate the corpus is Brat (Stenetorp et al. 2012).

In Fig. 1 we can see an example of an annotated text. Texts are presented and divided into 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.

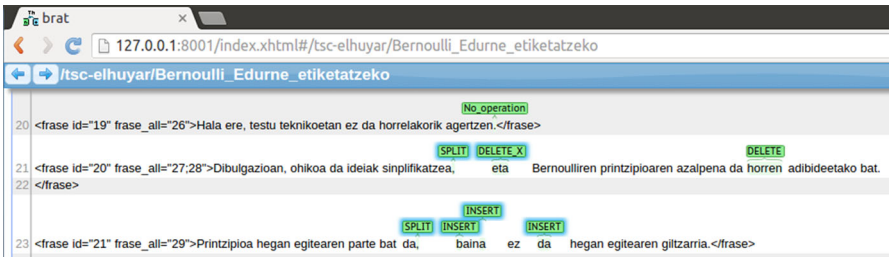

Fig. 1 A part of the text annotated with Brat

In the following section we present our annotation scheme expressed by means of macro-operations and operations.

4 Annotation scheme

In this section we expose our annotation scheme and the comparison to annotations schemes in other languages.

4.1 Annotation scheme for Basque

The annotation scheme we present is organised in eight macro-operations: delete, merge, split, transformation, insert, reordering, no_operation and other. In the following subsections we go through these macro-operations and describe the criteria taken and the operations involved. The examples are given in Basque and English (sometimes, the English translations may sound unnatural or ungrammatical but we have taken this decision to be able to illustrate the Basque phenomena properly). The cue words of the operation we are describing in each case will be underlined in both cases (Basque and English, as mentioned before). The different operations presented in this scheme are based on the annotation of the corpus; the structure of the annotation scheme has also been compared to the Spanish (Bott and Saggion 2014) and the Italian (Brunato et al. 2015) annotation schemes.

4.1.1 Delete

A *delete* operation is performed when some elements are eliminated from the original text. We distinguish two types of deletions based on the criterion of the nature of information contained in the deleted element:

- 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 1, the relative clause "*sortzen den*" (that is created) containing a piece of information (maybe not relevant) has been deleted. The

Table 1 Examples of delete operations

Operation	Original	Simplified
delete-info	*Sortzen den aldea oso handia da*	*Aldea oso handia da*
	The part that is created is very big	The part is very big
delete-functional	*Eta beste edozein hegazkinekin ere gauza bera gertatzen da*	*Beste edozein hegazkinekin ere gauza bera gertatzen da*
	And it also happens with any other kind of plane	It also happens with any other kind of plane

deleted element 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 1, we consider that the deletion of the *eta* (and) conjunction does not delete information; so, we tagged it as *delete-functional*.

4.1.2 Merge

When a *merge* operation is performed elements are fused; that is, a clause or a sentence has been created after having joined other clauses or sentences. 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 2, two sentences have been merged to create one, using as a link the pronoun in the genitive case "*haien*" (their). In this case, the *merge* has been

Table 2 Examples of merge operations

Operation	Original	Simplified
Merge	*Adibide bat gaur egungo hegazkin komertzialen hegoak dira. Haien diseinua plano aerodinamiko superkritikoan oinarrituta dago*	*Gaur egungo hegazkin komertzialen hegoen diseinua plano aerodinamiko superkritikoan oinarritzen da*
	The wings of the modern commercial planes are an example. Their design is based on the supercritical airfoil	The design of the wings of the modern commercial planes is based on the supercritical airfoil

performed by means of a coreference resolution, since the pronoun has been substituted with its referent to link the sentences.

4.1.3 Split

The *split* is the operation where clauses, phrases or words are divided with the aim of creating new sentences. We distinguish different types of splits based on two criteria:

- Strength: soft and hard. The *soft split* occurs when a new sentence has been delimited by a comma or a semicolon and the *hard split* happens when the new simplified sentence has been delimited by a full stop.
- Phenomena: coordination, noun-clauses, relative clauses, adverbial clauses (and different adverbial types), appositions, and postpositions are the phenomena we took into account.

In Table 3 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 referred to the coordination.

4.1.4 Transformation

Transformations represent the change of a word, a phrase or a structure. The criterion we have used to classify the *transformation* operations is the type: lexical,

Table 3 Examples of split operations

Operation	Original	Simplified
split-hard-coordination	*Dibulgazioan, ohikoa da ideiak sinplifikatzea, eta Bernoulliren printzipioaren azalpena da horren adibideetako bat*	*Dibulgazioan ohikoa da ideiak sinplifikatzea. Bernoulliren printzipioaren azalpena da adibideetako bat*
	It is normal to simplify the ideas in the science popularisation, and the explanation of Bernoulli's principle is an example of that.	It is normal to simplify the ideas in the science popularisation. Bernoulli's principle is an example of that.
split-soft-coordination	*Hortik aurrerako azalpena konplexua da, eta hegalari batetik bestera asko aldatzen da*	*Hortik aurrerako azalpena konplexua da*; *hegalari batetik bestera asko aldatzen da*
	From that on, the explanation is complex, and it changes considerably from one flyer to another	From that on, the explanation is complex; it changes considerably from one flyer to another

morphological, syntactic, discursive and corrections. In addition, combinations of these can happen. These are their distinguished transformation operations:

- Lexical: *Subst_Syn* (synonym substitution) and *Subst_MultiWord* (substitution of phrases)
- Morphological: *Pas2Act* (passive → active or impersonal → personal), *Fin2-NonFin* (finite verb → non-finite verb), *NonFin2Fin* (non-finite verb → finite verb), *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 (marked coreference resolution) and *Connect_Disc* (Change of discourse marker)
- Correction: *Correction* (correction of orthographic or grammatical mistakes)
- Combinations: *Reform* (reformulation or paraphrasing) and *Other_Subst* (other kind of transformations)

Examples of the *transformation* operations are shown in Table 4. It is possible that some instances represent more than one operation. Indeed, it is difficult to find examples with one operation only.

4.1.5 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 take into account 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: where the ellipsis is marked morphologically (*elided_morph*) or not (*non_required*).

Those are the three types of insertions we distinguish:

- *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 if a split has not been performed. In the example presented in Table 5, the coordinated apposition has been split and the verb "*da*" (is) has been added to create the simplified sentences out of the appositions.
- *Elided_morph:* verbs or nouns that are marked morphosyntactically (there is a morphological mark of the ellipsis in the word, usually a determinant) but have been made explicit. This operation happens in the former original sentence. In

Table 4 Examples of transformation operations

Operation	Original	Simplified
LEXICAL		
Subst_Syn	*ahaleginetan*	*lanetan*
	in the efforts	in the works
Subst_MultiWord	*urteetan zehar*	*urtero*
	through the years	every year
MORPHOLOGICAL		
Pas2Act	*ikusi da*	*ikusi dute*
	it has been seen	they have seen
Fin2NonFin	*hegazkin horiei airean eusten dien printzipio fisikoa*	*hegazkin horiei airean eusteko printzipio fisikoa*
	the physical principle that keeps those planes in the air	the physical principle to keep those planes in the air
NonFin2Fin	*Airea beherantz bultzatuta*	*Airea beherantz bultzatzen da*
	pushing down the air	the air is pushed down
Subst_Per	*orduan odolean begiratzen dugu*	*orduan odolean begiratzen dute*
	so, we look in the blood	so, they look in the blood
Verb_Feats	*gai izango litzateke*	*gai izango da*
	they might be able	he will be able
SYNTACTIC		
Clause2Phrase	*Jatorri genetikoa duten minbizi gehienetan*	*Jatorri genetikodun minbizi gehienetan*
	in the most of the cancers that have genetic origin	in the most of the cancers with genetic origin
Phrase2Clause	*bakoitzak oso diseinu ezberdinarekin*	*Bakoitzak bere diseinua du*
	each one with its different design	each one has its own design
Ind2Dir_Speech	*familian zenbat kasu dauden galdetzen dugu*	*zenbat kasu daude familian?*
	we ask how many cases there are in the family	how many cases are there in the family?
Dir2Ind_Speech	*horiekin "ez da eragozten" minbizia sortzea*	*horiekin ez dela galarazten minbizia sortzea*
	the creation of 'is not impeded' with those	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*
	that that flux makes more pressure to the wing downwards than upwards	the flux makes more pressure to the wing downwards than upwards

Table 4 continued

Operation	Original	Simplified
Main2Sub	*Familia barruan minbizi horietako kasu asko dituzten pertsonak iristen dira kontsultara*	*Mujikak esan du kontsultara etortzen direla familia bereko pertsonak*
	People that have those cancer cases in the family arrive at the consultation	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*	*angelu horren inguruan irauten badu*
	while it lasts around that angle	if it lasts around that angle
Sub2Coor	*Hartara, mutazioa identifikatuta*	*Hartara, mutazioa identifikatzen dugu*
	Thus, identified the mutation	Thus, we identify the mutation
DISCOURSE		
Coref	*Mende hartan*	*XVIII. mendean*
	in that century	in 18th century
Connect_Disc	*beraz*	*ondorioz*
	thus/therefore	as a result of
CORRECTION		
Correction	*abiadura (…) izan beharko luke*	*abiadurak (…) izan beharko luke*
	the speed (abs) should have	the speed (erg) should have
COMBINATION		
Reform	*Zama guztiarekin, 573 tonara irits daiteke*	*Zama guztiarekin, 573 tona pisatzen du gutxi gorabehera*
	With all the load, it can arrive to 573 tones	With all the load, it weights 573 tones approximately
Subst_Other	*hegaldiaren azalpenetik*	*hegaldiaren azalpenean*
	from the explanation of the flight	in the explanation of the flight

the example of Table 5, 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*: elided arguments, adjectives, adverbs, sentences or whatever is understood taking the context into account but that have been inserted to make the meaning clearer. This operation also happens in the former original sentence. In the example of Table 5, the subject[5] "*proteinak*" (the proteins) has been added. In this case the insert happens because of the coreference resolution.

[5] Basque is a pro-drop language and it is possible to make the ellipsis of subject, direct objects and indirect objects.

Table 5 Examples of insert operations

Operation	Original	Simplified
Funct_NS	*(…) Antonio Cantó dibulgatzaile eta hegazkinetan adituak*	*Antonio Cantó dibulgatzailea da; Antonio Cantó hegazkinetan aditua da.*
	Science populariser and expert on planes Antonio Cantó (….)	Antonio Cantó is a science populariser; Antonio Cantó is an expert on planes.
Ellided_morph	*endometrioko minbiziaren pronostikoa obulutegietakoa baino askoz ere hobea izaten da*	*endometrioko minbiziaren pronostikoa obulutegietako minbiziaren pronostikoa baino askoz ere hobea izaten da*
	the prognosis of the endometrial cancer is so much better than the ovarian	the prognosis of the endometrial cancer is so much better than prognosis of the ovarian cancer
Non-required	∅ *Eraldatuta badaude*	*Proteinak eraldatuta badaude*
	If (they) are transformed	If the proteins are transformed

4.1.6 Reordering

In the *reordering* operation the order of the elements is altered. We find different types of *reordering* operations and the criteria are: (1) element that has been moved (phrase, clause or auxiliary verb) and (2) the place to where it has been moved (inside a former original sentence or from a former original sentence to a new sentence). These are the *reordering* operations we find:

- *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 to a different position in sentence. This is the case of emphatisations or when negative verbs are changed into positive.
- *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 set back 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 6.

Table 6 Examples of reordering operations

Operation	Original	Simplified
Reord_Phrase (Phrases)	*(…) argitu du Bachiller astronomoak*	*Bachiller astronomoak argitu du (…)*
	(…) has clarified the astronomer Bachiller	The astronomer Bachiller has clarified (…)
Reord_Clause (Clauses)	*Aireak hegazkinaren inguruan duen jokabidea zoruak alda dezake, hegaldia oso baxua denean*	*Hegaldia oso baxua denean zoruak hegazkinaren inguruko airearen jokabidea alda dezake*
	The soil can change the behaviour that the air has around the plane, when the flight is very low	When the flight is very low, the soil can change the behaviour that the air has around the plane
Reord_Aux (Auxiliary verbs)	*Orain dela 25 urte, berriz, eguzki-sistemako planetak baino ez ziren ezagutzen*	*Orain dela 25 urte, berriz, eguzki-sistemako planetak bakarrik ezagutzen ziren*
	25 years ago, on the contrary, only planets in the solar system known were	25 years ago, on the contrary, only planets in the solar system were known
Reord_NS_Phrases (Phrases in new sentences)	*Hala ere, adituek aurreikusten dute planetagaien % 90, gutxi gorabehera, benetako planetak izango direla*	*Hala ere, planetagaien % 90, gutxi gorabehera, benetako planetak izango dira; hala aurreikusi dute adituek*
	However, the experts foresee that more or less the 90% of the candidates to be planets is going to be real planets	However, more or less the 90% of the candidates to be planets is going to be real planets; so foresee the experts

4.1.7 No_operation and other

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 that we can explore why they have not been simplified.

We can also find other operations not covered by this annotation scheme or that are tricky to classify. In these cases, the macro-operation used is *other* and it will be used as little as possible. The sentences with this tag will be further analysed.

4.2 Comparison of the Basque annotation scheme to annotation schemes for other languages

In Table 7, we sum up the macro-operations covered in our annotation scheme together with the criteria and sub-criteria we have taken to classify the operations.

After having detailed our annotation scheme, 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 8 we sum up the terms used in these works and our equivalents.

Table 7 Annotation scheme of Basque

Macro-operation	Criteria	Sub-criteria
Delete	Information	Information vs. functional
Merge		
Split	Strength	Hard vs. soft
	Phenomena	Coordination, adverbial clauses, relative clauses, apposition/parentheticals, noun clauses, postposition, others
Transformation	Linguist level	Lexical, morphological, syntactic, discourse, correction, other
Insert	Ellipsis type	Marked morphologically vs. not marked
	Place	Former original sentence vs. new sentence
Reordering	Element	Phrase, clause, auxiliary verb
	Place	Former original sentence vs. new sentence
No_operation		
Other		

Table 8 Terminology used in different annotation schemes

Basque	Italian	Spanish
Macro-operations	Classes	First dimension
Operations	Sub-classes	Second dimension

Let us begin by explaining the similarities and the differences found in relation to the Italian annotation scheme. At macro-operation level, we have defined the same macro-operations, the only difference being that we have grouped those cases that cannot be classified properly with 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 a title), two macro-operations 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 their results 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*.

5 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 both approaches. These results and comparisons will help us to establish common criteria to be applied in the implementation of the automatic text simplification system for Basque (Aranzabe et al. 2012).

Before we discuss the results, we will show the details of the CBST corpus. We recall that CBST is formed by 227 original sentences from long texts of the Elhuyar corpus and two different simplifications of each sentence. 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 sentence and word number of each text on the corpus can be seen in Table 9. We also show the average sentence length of each text.

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 cases simplified by the translator with the structural approach but that tendency occurs only in one of the texts simplified by the teacher (intuitive approach). The average sentence length is reduced in all the simplified versions, above all in the intuitive approach.

Let us give an overview of the corpora simplified manually in other languages. In Table 10, we indicate the language and the reference for the corpus, the number of articles comprised and the number of sentences and words we can find in the original documents and in the simplified ones and average sentence length.

If we compare the size of CBST, it is in general smaller than the others. 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. The average sentence length is also reduced in all corpora.

Table 9 Sentence and word number and average sentence length in the original and simplified texts

Text	Version	Sentences	Words	Average sentence length
Bernoulli (technology)	Original	89	1446	16.25
	Structural	123	1472	11.97
	Intuitive	105	1253	11.94
Etxeko (medicine)	Original	70	1535	21.93
	Structural	84	1611	19.18
	Intuitive	105	1608	15.29
Exoplanetak (social science)	Original	68	1512	22.24
	Structural	75	1608	21.44
	Intuitive	96	1258	13.10
Total	Original	227	4493	19.79
	Structural	282	4691	19.63
	Intuitive	276	4119	14.92
Total	Corpus	785	13,303	16.95

5.1 Alignment

The aim of alignment process of the corpus is basically to know which sentences of the simplified texts have been created out of each original sentence. We have aligned the sentences manually before the annotations as Brunato et al. (2015), but there are other methods like the cardinality property[6] defined by Caseli et al. (2009) and the automatic alignment and manual revision by Bott and Saggion (2011). So, we have 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 simplified one 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 11.

Most of the sentences have been aligned in 1:1 scale. The second most used scale has been 1:2. The 1:3 and 2:1 scales are less frequent in both approaches. Other scales cover the cases where a sentence has been aligned with more than three sentences or to half sentences. 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 (Teacher) and 1:2 in the structural approach (Terence) and in Spanish 1:N in both corpora.

[6] They take into account how many sentences are produced by each operation.

Table 10 Sentence and word number and average sentence length in the original and simplified texts in other corpora

Language/corpus	Articles	Sentences		Words		Average length sentence	
	Docs	Original	Simplified	Original	Simplified	Original	Simplified
English Petersen and Ostendorf (2007)	104	2539	2459	41,982	29,584	16.50	12.00
Brazilian Portuguese Caseli et al. (2009)	104	2116	3104 (natural simplification) 3537 (strong simplification)	41,897	43,013 (natural) 43,676 (strong)	19.80	13.85 (nat.) 12.35 (str.)
Spanish Bott and Saggion (2011)	–	110	145	2456	1840	34.64	12.44
Danish Klerke and Søgaard (2012).	3701	48,186	62,365	–	–	17.30	11.10
Italian Terence Brunato et al. (2015)	–	1036	1060	–	–	–	–
Italian Teacher Brunato et al. (2015)	24	–	–	–	–	–	–

Table 11 Alignment results

Scale	Structural	Intuitive
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 12 Results of the macro-operations in both approaches

Macro-operations	Structural	Intuitive
Transformation	24.92 (254)	33.62 (309)*
Split	23.55 (240)*	12.30 (113)
Insert	21.88 (223)*	18.61 (171)
Delete	17.66 (180)	20.78 (191)
Reordering	7.95 (81)	8.27 (76)
No_operation	3.53 (36)	6.20 (57)*
Merge	0.40 (4)	0.22 (2)
Other	0.10 (1)	0.00 (0)

5.2 Incidence of macro-operations and operations

We are going to present the incidence of the operations performed to create the simplified texts. We will start the description of the results of the macro-operations in general (Table 12). In parentheses we show the raw number each macro-operation has carried out.

The asterisk in Table 12 shows the statistically significant differences between approaches. The differences between the both approaches in the macro-operations transformation (p value: 0.03668), *split* (p value: $< 2.2e{-}16$), *insert* (p value: 0.01245) and *no_operation* (p value: 0.002526) are statistically significant. The test we carried is the paired t-test and it has been applied at sentence level using the programming language R. In the null hypothesis we assumed that all the means are equal and in the alternative we assumed that they are different (two-paired). No test was carried for the operations *merge* and *other* because there are not enough data points.

Transformation is the most frequent macro-operation (24.92% in the structural approach and 33.62% in the intuitive). The second most used operation differs in the approaches: the translator has used the *split* (23.55%) in the structural approach while the teacher tends to use the *delete* operation (20.78%) in the intuitive. We think that the predominance of the *split* in the translator's simplification is

influenced by the guidelines she received where it was stated to use one verb per sentence. The less frequent macro-operations are *merge* and *other* in the both approaches. The sentences which have not been simplified (*no_operation*) are also more frequent in the intuitive approach (6.20%) than in the structural approach (3.53%). The percentages of *reordering, insert* and *delete* are quite similar. Finally, the *split* has been used more times in the structural than in the intuitive with a difference of more than 10 points.

It is predictable to find that the *transformation* is the most used macro-operation. We have to take into account that it incorporates many different operations, and that simplification is also considered as rewriting, and many of the rewriting operations are usually transformations.

5.2.1 Transformation

The most frequent transformation operation found in the structural approach is *Sub2Main* (48.50%) and the *reformulation* (19.09%) has been the most used in the texts of the intuitive. With these results, we see that there is a tendency to convert subordinate clauses into main clauses in the structural approach while a broader variety of operations has been used in the intuitive. Sorting the *transformation*s according to their type (Table 13), we see that in both approaches the most used *transformation*s are the syntactic transformations. The least used is *correction*.

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. Syntactic transformations have been given importance (almost eight points of difference) in the structural approach while lexis has been given in the intuitive (more than four points of difference). We would like also to mention the importance of the morphological transformations. Transformations tagged as *other* should also be analysed in the future.

5.2.2 Split

Let us show now the results of the *split* operations. The *split* depending on the strength that has been most used in the structural approach is the *soft split* (74.06%) while the most used in the intuitive is the *hard split* (69.03%). These results show that both approaches differ absolutely at this point. This leads us to analyse[7] more carefully the average sentence length of the simplified texts taking into account the *soft splits* (Table 14) where, as expected, the average sentence length decreases above all in the structural approach.

Looking at the phenomena that have been split, coordination has been the most (structural: 39.17% and intuitive: 45.13%), followed by the adverbial clauses (structural: 19.16% and intuitive: 16.81%). All the results can be seen in Table 15.

[7] To recalculate the average sentence length with *soft split*s we have also considered the clauses delimited with semicolons as sentences.

Table 13 Results of the *transformation* types in both approaches

Transformation type	Structural	Intuitive
Syntactic	41.34 (105)	33.01 (102)
Morphological	22.05 (56)	19.09 (59)
Others	14.57 (37)	19.74 (61)
Discursive	14.96 (38)	15.86 (49)
Lexical	6.70 (17)	11.03 (34)
Correction	0.39 (1)	1.29 (4)

Table 14 Average sentence length taking into account the different *split* operations

Text	Version	Average sentence length	Average sentence length with soft splits
Bernoulli (technology)	Original	16.25	
	Structural	11.97	9.03
	Intuitive	11.94	9.94
Etxeko (medicine)	Original	21.93	
	Structural	19.18	10.39
	Intuitive	15.29	14.01
Exoplanetak (social science)	Original	22.24	
	Structural	21.44	9.67
	Intuitive	13.10	10.84

We have also analysed the types of adverbial clause that have been split and these results are shown in Table 16.

The most split adverbial clauses in the structural approach have been the conditional (23.91%) and the causal clauses (21.74%). Causal clauses (42.11%) have also been the most simplified in the intuitive 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 17.

There is the tendency to split relative clauses and causal clauses in both approaches. The proportion of temporal clauses is also similar (structural: 17.65% and intuitive: 14.71%). Modal-temporal clauses were not split in any of the approaches.

5.2.3 Insert

Another macro-operation that has been widely used is the *insert*. The results of the three *insert* types are shown in Table 18.

Table 15 Results of the *split* operation according to the phenomena in both approaches

Split phenomena	Structural	Intuitive
Coordination	39.17 (94)	45.13 (51)
Adverbial clauses	19.16 (46)	16.81 (19)
Relative clauses	16.25 (39)	11.50 (13)
Apposition/parentheticals	10.83 (26)	7.96 (9)
Noun clauses	7.50 (18)	0.00 (0)
Postposition	3.75 (9)	3.54 (4)
Others	3.33 (8)	15.05 (17)

Table 16 Results of the *split*s adverbial clauses in both approaches

Split (adverbial)	Structural	Intuitive
Conditional	23.91 (11)	0.00 (0)
Causal	21.74 (10)	42.11 (8)
Modal	17.39 (8)	5.26 (1)
Temporal	13.04 (6)	26.32 (5)
Concessive	10.87 (5)	15.79 (3)
Purpose	6.52 (3)	10.53 (2)
Comparative	6.52 (3)	0.00 (0)

Table 17 Proportion of the split subordinate clauses

Subordinate type	Number (orig.)	Split (structural)	Split (intuitive)
Noun clause	162	11.11 (18)	0.00 (0)
Modal	69	11.59 (8)	1.45 (1)
Relative	57	66.67 (38)	22.81 (13)
Conditional	57	19.30 (11)	0.00 (0)
Temporal	34	17.65 (6)	14.71 (5)
Causal	23	43.48 (10)	34.78 (8)
Purpose	20	15.00 (3)	10.00 (2)
Modal-temporal	17	0.00 (0)	0.00 (0)
Concessive	5	100.00 (5)	60.00 (3)

The *non-required insert*s have been the most used *insert* type in both approaches (structural: 44.39% and intuitive: 57.89%). In the guidelines to perform the structural approach, the translator was recommended to cover all the possible arguments, but as we see the teacher, following her intuition, performs the same operation. The functional insert operations that have been used in the creation of new sentences are in the second position in both approaches and the recovery of the

Table 18 Results of the *insert* types in both approaches

Insert types	Structural	Intuitive
Non-required	44.39 (99)	57.89 (99)
Funct_NS	42.15 (94)	30.99 (53)
Ellided_morph	13.45 (30)	11.11 (19)

Table 19 Results of the *delete* types in both approaches

Delete types	Structural	Intuitive
Delete information	25.56 (46)	30.37 (58)
Delete functional words	74.44 (134)	69.36 (133)

morphologically marked elided elements was the least 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.

5.2.4 Delete

Regarding the treatment of the information, we have distinguished two *delete* operations. Those where information has been omitted are 25.56% in the structural approach and 30.37% in the intuitive. The *deletes* of functional words are 74.44% in the structural and 69.36% in the intuitive. That is, in both approaches most of the *deletes* do not imply information loss. These results are shown in Table 19.

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 words is a closed group and in Table 20 we show the f*unctional delete*s that have been performed. In both approaches the functional words that have been mainly deleted are coordinate conjunctions, punctuation and discourse markers.

5.2.5 Reordering

The results of the *reordering* operations are shown in Table 21. The most used *reordering* in both approaches has been the reordering of phrases, although it has been more broadly used by the teacher in the intuitive approach (78.95%) than the translator in the structural (43.20%). The translator was told in the guidelines to keep a canonical and chronological reordering, so in the future we plan to corroborate if these movements have been performed to fulfil this guideline. The second most used in the structural approach has been the movement of phrases into new sentences (41.98%) while the ordering of clauses (13.16%) has been changed in the intuitive.

Table 20 Results of the *delete of functional words* in both approaches

Delete functional word types	Structural	Intuitive
Coordinate conjunction	54.48 (73)	33.08 (44)
Punctuation	23.88 (32)	34.59 (46)
Discourse marker	14.93 (20)	24.06 (32)
Other	6.71 (9)	8.27 (11)

Table 21 Results of the reordering in both approaches

Reordering types	Structural	Intuitive
Reord_Phrases	43.20 (35)	75.00 (57)
Reord_NS_Phrases	41.98 (34)	11.84 (9)
Reord_Clause	13.58 (11)	13.16 (10)
Reord_Aux	1.23 (1)	0.00 (0)

Table 22 Results of the other macro-operations in both approaches

Other macro-operations	Structural	Intuitive
No_operation	3.53 (36)	6.20 (57)
Merge	0.40 (4)	0.22 (2)
Other	0.10 (1)	0.00 (0)

5.2.6 Other Macro-operations

The results of the rest of macro-operations (*no_operation, merge* and other) are shown in Table 22. Except for the *no_operation*, the other operations do not reach 1%.

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 operation has not been performed because it is an operation that is more related to summarisation than to simplification.

5.3 Discussion

In order to summarise these results, we are going to point out what we have found in common in both approaches. 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 subordinate clauses, and taking into account the numbers of the original texts, the ones which

have been split most are the causal and the relative. Among other operations that are common in both approaches, we find *non-required inserts* (elided elements that are understood taking the context into account), *functional deletes*, and *phrase reordering*.

The points we have mentioned agree with the simplification study for Basque and with the future work proposed by Aranzabe et al. (2012). In this study, syntactic simplification is treated and, as we have seen, here most of the transformations have been syntactical. The *split* is also important in the design of the system for Basque. The reordering of phrases is defined and a correction module is also foreseen in the system. In relation to future work it is planned to recover the elided elements (non-elided elements) and the functional deletes are included in the reconstruction rules of the system (Gonzalez-Dios 2016). These common operations will also serve as the basis for new guidelines when enlarging the corpus. We also think that these operations should be given more weight in the automatic text simplification system for Basque. That is, syntactic simplification should be more important than the lexical and the rules related to the coordination and adverbial clauses should have priority.

As an example of the macro-operations and operations presented above, let us go through three sentences to see the effect of the different simplifications. In the original sentence in Table 23 we present a sentence where there is a modal non-finite clause ("*Airea beherantz bultzatuta*"), there is an inversion of the order of the elements in the verb ("*egiten dute hegan*"[8]) and the subject is in the last position ("*hegazkinek*").

In the structural approach, broadly explained, the subordination has been removed and the order of the elements has been changed to become canonical. Exactly, (1) a *soft split* has been performed, (2) the non-finite clause has been converted into finite (*NonFin2Fin*), (3) the subject has been moved before the verb (*Reord_Phrase*), (4) the ordering of the verb has been presented as the common one, that is a lexical verb + auxilary verb (*Reord_Aux*) and (5) a discourse marker "*orduan*" ('then') has also been inserted to recover the modal relation (*Funct_NS*). In the intuitive approach, a simpler sentence has been created without structural changes[9] at syntactic level: the subordinate clause has been moved back (*Reord_Clause*), and the subject and the verb have been reordered as in the intuitive approach following the canonical order (*Reord_Phrase*).

In the example of Table 24 shorter sentences have also been created in both approaches. In the original sentence there is coordinated clause whose first coordinate has a non-finite temporal clause ("*Teleskopioak izarretara zuzentzean*") that includes a relative clause ("*guregandik urrun dauden"*).

In the structural approach, (1) the coordinates have undergone a *soft split*, (2) the discourse marker "*orduan*" (then) has been inserted in the second one, (3) the conjunction "*eta*" (and) has been deleted (d*elete functional words*), (4) there has been reformulation of the verb "*ez dira gai (…) ikusteko*" (they are not able to

[8] In Basque the 'fly' means *hegan egin,* which literally translated is 'fly do' (do fly).

[9] For text simplification purposes, we have defined structural changes as the cases where the depth of the syntactic tree has been altered (Gonzalez-Dios 2016).

Table 23 An example of a simplification from the text "Bernoulli" in both approaches

Original	Structural	Intuitive
Airea beherantz bultzatuta egiten dute hegan hegazkinek	*Airea beherantz bultzatzen da; orduan hegazkinek hegan egiten dute*	*Hegazkinek hegan egiten dute airea beherantz bultzatuta*
Pushing down the air does fly the planes	The air is pushed down; then the planes fly	The planes fly pushing down the air

Table 24 An example of a simplification from the text "Exoplanetak" in both approaches

Original	Structural	Intuitive
Teleskopioak guregandik urrun dauden izarretara zuzentzean, ordea, izarren argitasunak itsutu egiten ditu, eta ez dira gai inguruko planetak ikusteko	*Teleskopioak urrutiko izarretara zuzentzean, ordea, izarren argitasunak itsutu egiten ditu; orduan, ezin dituzte inguruko planetak ikusi*	*Teleskopioak guregandik urrun dauden izarretara zuzentzen dira; orduan izarren argitasunak itsutu egiten ditu. Ondorioz, teleskopioak ez dira gai inguruko planetak ikusteko*
When directing the telescopes to the stars that are far from us, however, the starlight blinds them and they are not able to see the surrounding planets	When directing the telescopes to the distant stars however, the starlight blinds them; then, they cannot see the surrounding planets	The telescopes are directed to the stars that are far from us; then, the starlight blinds them. So, the telescopes are not able to see the surrounding planets

see) → "*ezin dituzte (…) ikusi*" (they cannot see) and 5) the relative clause in the first coordinate has been transformed into an adjective (*clause2phrase*).

In the intuitive approach, (1) the *split* in the temporal clause has been soft but (2) the one in the coordinate has been hard. In order to recover the relation lost when splitting the temporal clause, (3) the discourse marker "*orduan*" ('then') has been inserted (*Funct_NS*) and after the split in the coordinate clause, (4) the discourse marker "*ondorioz*" ('so') has been inserted.

When analysing the corpus, we have seen that in the case of coordinate clauses they were mainly split and the conjunction was deleted. In the sentence presented in Table 25, however, this has not been performed in the intuitive approach. This leads us to think that we should also analyse the phenomena taking into account the surrounding context.

Looking at the simplification outputs presented in the examples presented in the Tables 23, 24, and 25, we have seen that the main effect is the shortening of the sentence length by means of different operations. Not only does the split play a role in it, converting clauses into phrases, but the reformulations may also have an effect. In Table 23 we have seen that the reordering of the elements can also play a crucial role. These effects, in our opinion, are not only related to Basque, since e.g. short sentences also incarnate an important characteristic of simplicity in other languages.

Table 25 An example of a simplification of a coordinate clause

Original	Structural	Intuitive
Izan ere, inbertsio handiak egin behar izan dituzte, eta ahalmen handiko gailuak ibiltzen dituzte misio horietan	*Izan ere, inbertsio handiak egin behar izan dituzte; ahalmen handiko gailuak ibiltzen dituzte misio horietan*	*Izan ere, inbertsio handiak egin behar izan dituzte, eta ahalmen handiko gailuak ibiltzen dituzte misio horietan*
In fact, they had needed to do big investments, and they use powerful tools in those missions	In fact, they had needed to do big investments; they use powerful tools in those missions	In fact, they had needed to do big investments, and they use powerful tools in those missions

We want to mention also that both using short sentences and keeping the chronological and canonical order were recommended in the guidelines given to the translator to simplify the texts in the structural approach.

When we performed the comparison of the annotation scheme, we found that the schemes for Italian and Spanish are quite similar to ours, at least at macro-operation level. Therefore, 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 those in Brazilian Portuguese. This comparison is, however, more difficult due to the following: (1) there is no structured annotation scheme[10] of the simplification operations, although they show a list of them, and (2) the results are given according to the simplification levels natural and strong.

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). These three macro-operations are the same ones that the teacher has mainly used in the intuitive approach. Looking at the percentages, the reordering operations performed in Basque and the *insert* are quite similar to the Italian Teacher and Terence corpora. The proportion is smaller in the Spanish corpus. The least used macro-operation is also the same in the three languages: the *merge* or *fusion*. It is remarkable that the split has been more widely used in Basque. If we compare between approaches and languages, we see that both in the Italian (Teacher) and in the Basque intuitive approaches, there is a tendency to perform *deletes*, while in the structural approaches (Terence and Basque structural) the tendency is to perform *inserts*. The data used to make this comparison with Spanish (Bott and Saggion 2014) and Italian (Brunato et al. 2015) is presented in Table 26.

Looking at the results of Brazilian Portuguese (Caseli et al. 2009), the *split* is also the second most applied operation in Brazilian Portuguese when simplifying from original to natural simplification, covering 34.17% of the operations. The *reordering of clauses* (inversion of clause reordering) is 9.30% original to natural simplification and the operation *joining sentences* (related to our *merge*) is also unusual in original to natural simplification (0.24%).

[10] In Caseli et al. (2009) it is presented a XCES annotation scheme as a corpus coding standard, not as an abstraction of different simplification operations.

Table 26 Comparison of macro-operations across languages

Macro-operation	Italian		Spanish	Basque	
	Terence	Teacher		Structural	Intuitive
Transformation	48.18 (1183)	47.76 (811)	39.02	24.92 (254)	33.62 (309)
Split	1.71 (43)	2.06 (35)	12.20	23.55 (240)	12.30 (113)
Insert	18.72 (460)	15.66 (266)	12.60	21.88 (223)	18.61 (171)
Delete	21.94 (539)	25.32 (430)	24.80	17.66 (180)	20.78 (191)
Reordering	8.65 (212)	7.89 (134)	2.85	7.95 (81)	8.27 (76)
No_operation	–	–	–	3.53 (36)	6.20 (57)
Merge	0.81 (20)	1.30 (22)	0.81	0.40 (4)	0.22 (2)
Other	–	–	–	0.10 (1)	0.00 (0)

If we go a level down in the annotation and analyse the types of the transformations, we see that contrary to the results in the CBST corpus, the most performed type is 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). We think that this difference may be related to the domain. Looking at the split phenomena, coordination has also been the most split in Spanish, as in the Basque corpus. It is difficult for us to compare other operations with the available data.

6 Conclusion and future work

In this paper we have presented the corpus of Basque simplified texts (CBST) which compiles different simplification approaches of the texts. We have developed an annotation scheme where different macro-operations and operations have been compiled to know what happens when simplifying texts and also, to compare them across approaches. This 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.

Although the first aim of the CBST was to make an analysis of the simplified texts, we have to mention that we have obtained useful information for the evaluation and further development of the system for Basque (Aranzabe et al. 2012) by giving more importance to the common operations (performing syntactic transformations, splitting coordination and the adverbial clauses, correction…) found here. Moreover, we can still learn from this corpus by analysing further, for example, the information deletes or the movements performed in the reordering operations to see if they fit to the canonical order or to a discourse level theory like RTS (Mann and Thompson 1988) or Centering Theory (Grosz et al. 1995). We also plan to analyse the *split* operations to identify if there are other reasons or patterns beyond the guideline ‘use one verb per sentence’ that was recommended in the structural approach.

We have also compared the CBST to corpora in other languages. Although CBST is in general smaller, the alignment and macro-operations results are similar to those of the other languages. It will also be interesting to compare those corpora in depth to find more common criteria or universal criteria when simplifying texts.

Nevertheless, we have also created a basis with the common phenomena that will serve as guidelines for the expansion of the CBST. To enlarge 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 also be added, that is, elided subjects, objects and so on should be recovered. Those are, indeed, the points we have found in common in both approaches.

In relation to the comparison to other languages we will like also to perform cross-genre analysis to see if the same macro-operations are performed in the same genres. To that end, we are simplifying texts of educational and journalistic domains while we also enlarge the corpus.

Acknowledgements Itziar Gonzalez-Dios's work was funded by a Ph.D. grant from the Basque Government and a postdoctoral grant for the new doctors from the Vice-rectory of Research of the University of the Basque Country (UPV/EHU). We are very grateful to the translator and teacher that simplified the texts. We also want to thank Dominique Brunato, Felice Dell'Orletta and Giulia Venturi for their help with the Italian annotation scheme and their suggestions when analysing the corpus and Oier Lopez de Lacalle for his help with the statistical analysis. We also want to express our gratitude to the anonymous reviewers for their comments and suggestions. This research was supported by the Basque Government (IT344-10), and the Spanish Ministry of Economy and Competitiveness, EXTRECM Project (TIN2013-46616-C2-1-R).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Aranzabe, M. J., Díaz de Ilarraza, A., & Gonzalez-Dios, I. (2012). First approach to automatic text simplification in basque. In L. Rello, & H. Saggion (Eds.), *Proceedings of the natural language processing for improving textual accesibility (NLP4ITA) workshop (LREC 2012)* (pp. 1–8).

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

Bott, S., & Saggion, H. (2011). An unsupervised alignment algorithm for text simplification corpus construction. In *Proceedings of the workshop on monolingual text-to-text generation, Association for Computational Linguistics, Stroudsburg, PA, USA, MTTG '11* (pp. 20–26).

Bott, S., & Saggion, H. (2014). Text simplification resources for Spanish. *Language Resources and Evaluation, 48*(1), 93–120.

Brouwers, L., Bernhard, D., Ligozat, A. L., & 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), Association for Computational Linguistics, Gothenburg, Sweden (pp. 47–56).

Brunato, D., Dell'Orletta, F., Venturi, G., & Montemagni, S. (2015). Design and annotation of the first Italian corpus for text simplification. In *The 9th linguistic annotation workshop held in conjunction with NAACL 2015*.

Carreiras, M., Duñabeitia, J. A., Vergara, M., de la Cruz-Pavía, I., & Laka, I. (2010). Subject relative clauses are not universally easier to process: Evidence from Basque. *Cognition, 115*(1), 79–92.

Caseli, H. M., Pereira, T. F., Specia, L., Pardo, T. A. S., Gasperin, C., & Aluísio, S. (2009). Building a Brazilian Portuguese parallel corpus of original and simplified texts. In *Proceedings of CICLing* (pp. 59–70).

Chall, J. S., & Dale, E. (1995). *Readability revisited: The new Dale–Chall readability formula*. Northampton: Brookline Books.

Coster, W., & Kauchak, D. (2011). Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human language technologies: Short papers* (Vol. 2, pp. 665–669).

Covington, M. A., He, C., Brown, C., Naçi, L., & Brown, J. (2006). How complex is that sentence? A proposed revision of the Rosenberg and Abbeduto D-Level Scale. CASPR Research Report 2006-01. Athens, GA: The University of Georgia, Artificial Intelligence Center.

Crossley, S. A., Allen, D., & McNamara, D. S. (2012). Text simplification and comprehensible input: A case for an intuitive approach. *Language Teaching Research, 16*(1), 89–108.

Dell'Orletta, F., Montemagni, S., & Venturi, G. (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, Association for Computational Linguistics, Stroudsburg, PA, USA, SLPAT '11* (pp. 73–83).

DuBay, W. H. (2004). *The principles of readability*. Costa Mesa, CA: Impact Information.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221–233.

Gonzalez-Dios, I. (2016). Euskarazko egitura konplexuen analisirako eta testuen sinplifikazio automatikorako proposamena/Readability assessment and automatic text simplification. The analysis of Basque complex structures. PhD Thesis, University of the Basque Country (UPV/EHU).

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., 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).

Gordon, P. C., Hendrick, R., & Johnson, M. (2004). Effects of noun phrase type on sentence complexity. *Journal of Memory and Language, 51*(1), 97–114.

Grosz, B. J., Weinstein, S., & Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics, 21*(2), 203–225.

Gunning, R. (1968). *The technique of clear writing*. New York: McGraw-Hill.

Hancke, J., Vajjala, S., & Meurers, D. (2012). Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of COLING 2012, the 24th international conference on computational linguistics: Technical papers* (pp. 1063–1080).

Klaper, D., Ebling, S., & Volk, M. (2013). Building a German/simple German parallel corpus for automatic text simplification. In *Proceedings of the second workshop on predicting and improving text readability for target reader populations, Association for Computational Linguistics, Sofia, Bulgaria* (pp. 11–19).

Klerke, S., & Søgaard, A. (2012). DSim, a Danish parallel corpus for text simplification. In N. Calzolari (Conference Chair), K. Choukri, T. Declerck, M. Ugur Dogan, B. Maegaard, J. Mariani, et al. (Eds.),. *Proceedings of the eight international conference on language resources and evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey* (pp. 4015–4018).

Laka, I., & Erdozia, K. (2010). Linearization references given "Free Word Order"; Subject preferences given ergativity: A look at Basque. In E. Torrego (Ed.), *Festschrift for Professor Carlos Piera*. Oxford: Oxford University Press.

Lu, X. (2009). Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics, 14*(1), 3–28.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text, 8*(3), 243–281.

Mitkov, R., & 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), Association for Computational Linguistics and Dublin University* (pp. 30–40).

Pellow, D., & Eskenazi, M. (2014). An open corpus of everyday documents for simplification tasks. In *Proceedings of the 3rd workshop on predicting and improving text readability for target reader populations (PITR), Association for Computational Linguistics, Gothenburg, Sweden* (pp. 84–93).

Petersen, S. E., & Ostendorf, M. (2007). Text simplification for language learners: A corpus analysis. In *Proceedings of workshop on speech and language technology for education*. SLaTE, Citeseer (pp. 69–72).

Ros, I., Santesteban, M., Fukumura, K., & Laka, I. (2015). Aiming at shorter dependencies: The role of agreement morphology. *Language, Cognition and Neuroscience, 30*(9), 1156–1174.

Rosenberg, S., & Abbeduto, L. (1987). Indicators of linguistic competence in the peer group conversational behavior of mildly retarded adults. *Applied Psycholinguistics, 8*(1), 19–32.

Shardlow, M. (2014). A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications (IJACSA), 4*(1), 58–70.

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

Štajner, S. (2015). New data-driven approaches to text simplification. PhD Thesis, University of Wolverhampton.

Štajner, S., Drndarevic, B., & Saggion, H. (2013). Corpus-based sentence deletion and split decisions for Spanish text simplification. *Computación y Sistemas, 17*(2), 251–262.

Stenetorp, P., Pyysalo, S., Topic, G., Ohta, T., Ananiadou, S., & Tsujii, J. (2012). BRAT: A web-based tool for NLP-assisted text annotation. In *Proceedings of the demonstrations session at EACL 2012*.

Warren, T., & Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition, 85*(1), 79–112.

Xu, W., Callison-Burch, C., & Napoles, C. (2015). Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics, 3,* 283–297.

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