Framework for the Analysis of Simplified Texts Taking Discourse into Account: the Basque Causal Relations as Case Study

Itziar Gonzalez-Dios and **Arantza Diaz de Ilarraza** and **Mikel Iruskieta**
itziar.gonzalezd@ehu.eus; a.diazdeilarraza@ehu.eus, mikel.iruskieta@ehu.eus
University of the Basque Country (UPV/EHU)
IXA Group for NLP
Manuel Lardizabal pasealekua 1. 20018 Donostia, Gipuzkoa

Abstract

Text simplification is crucial for some readers to understand the content of a text. Analyzing simplified texts can help to understand the mechanism hidden in the process of simplification. In this paper we present a research framework to analyze the impact of simplification operations on discourse. To that end, we used the Corpus of the Simplified Basque texts (CBST) and we studied the strategies followed in the simplification of causal relations and their effects at discourse level. From this analysis of the sample we derive that discourse has not been always taken into account which may lead to a lack of coherence in the simplified text.

1 Introduction and Related Work

Text Simplification is a research line that has been important in the educational community (Simensen, 1987; Young, 1999; Crossley et al., 2007) but it is also becoming important in the Natural Language Processing (NLP) community. Therefore, multidisciplinary researchers are working on different ways to make text simplification by automatic or semi-automatic means. This task is known as Automatic or Automated Text Simplification (ATS) and its development has been deeply explained in the literature ((Saggion, 2017)).

In this work, we want to describe a framework to analyze simplified texts taking discourse structure following the Rhetorical Structure Theory (RST)[1] (Mann and Thompson, 1988) into account and answer the following research questions:

– How can we describe the impact of simplification operations in discourse?

– How do simplification operations affect the rhetorical structures of the original texts?

This type of studies need annotated corpora which are expensive, but at the same time, necessary. We can find in the literature corpora available for English (Petersen and Ostendorf, 2007; Xu et al., 2015; Pellow and Eskenazi, 2014), Danish (Klerke and Søgaard, 2012), German (Klaper et al., 2013), Brazilian Portuguese (Caseli et al., 2009), Spanish (Bott and Saggion, 2011), Italian (Brunato et al., 2015) and Basque (Gonzalez-Dios, 2016). In the case of the last three corpora, simplification operations have been annotated and general annotation schemes derived. Besides, from the simplification perspective, Gonzalez-Dios et al. (2016) analyzed in the Basque corpus whether conditional, concessive, purpose, temporal and relative clauses[2] have been simplified or not, and if so, which were the macro-operations that had been performed.

From the discourse perspective, Crossley et al. (2007) analyzed the cohesion of 105 texts taken from seven texts-books aiming beginners of English as a second language with Coh-metrix (Graesser et al., 2004). They focused on the following seven sets: i) causal cohesion, ii) connectives and logical operators, iii) coreference measures, iv) density of major parts of speech measures, v) polysemy and hypernymy measures, vi) syntactic complexity, and vii) word information and frequency measures. They found out among others that original

[1] RST is an approach to describe text coherence by means of coherence relations or rhetorical relations and has been applied to many NLP tasks.

[2] These clauses are the most five predictive features for the readability assessment system for Basque (Gonzalez-Dios et al., 2014) at the syntactic level.

Proceedings of the 6th Workshop Recent Advances in RST and Related Formalisms, pages 48–57,
Santiago de Compostela, Spain, September 4 2017.

Original	Structural	Intuitive
Beraz, hegoaren formak, nahiz eta hegan egitearen lehen arrazoia ez izan, garrantzi handia du, inguruan duen airearen jarioan asko eragiten duelako.	*Beraz, hegoaren formak garrantzi handia du ; izan ere, hegoaren formak inguruan duen airearen jarioan asko eragiten du . Hegoaren forma, ordea, ez da hegan egitearen lehen arrazoia.*	*Beraz, hegoaren formak, nahiz eta hegan egitearen lehen arrazoia ez izan, garrantzi handia du ; izan ere, inguruan duen airearen jarioan asko eragiten du.*
So, the form of the wings, though it is not the main motive of the flying, is very important, because it affects a lot the surrounding air flow.	So, the form of the wings is very important; indeed, the form of the wings affects a lot the surrounding air flow. The form of the wings is not, however, the main motive of the flying.'	So, the form of the wings, though it is not the main motive of the flying, is very important; indeed, it affects a lot the surrounding air flow.

Table 1: The original sentence Bernoulli_80 and its two simplified versions

texts had a higher ratio of causal verbs to causal particles. Therefore, original texts exhibited less causal relations. In the analysis of intuitively simplified texts, Crossley et al. (2012) found out that advanced level texts exhibited less causal cohesion than beginning level texts.

To our knowledge, there is no joint framework to analyze simplified texts taking simplification operations and discourse into account. That is why the aim of this paper is to propose a framework to measure how simplification operations affect relational discourse structure. In this study, we focus on forms used to express causality because reducing causal discourse relations is crucial for people with language disorders. For example, Kong et al. (2017) stated that the coherence of speakers with aphasia tended to miss essential information content. This can be measured because aphasia speakers reduce some RST relations, such as ELABORATION and causal relations in their speech.

This paper is structured as follows: in Section 2 we present the resources needed to perform the analysis; in Section 3, we describe the framework for the analysis; in Section 4, we present the results of the quantitative analysis on the causal relations and in Section 5, we conclude and outline the future work.

2 Resources

In order to perform this study, we have used the Corpus of Basque Simplified Text (CBST). This corpus is a collection of texts divided in 227 sentences of the science popularisation domain. Each original sentence in the corpus has a structurally simplified and an intuitively simplified sentence. In this corpus, the operations performed in order to simplify the sentences have been annotated following an annotation scheme[3] composed by the following eight macro-operations: i) delete, ii) merge, iii) split, iv) transformation, v) insert, vi) reordering, vii) no operation and $viii$) other. These macro-operations involve many operations (Gonzalez-Dios, 2016). In Table 1 we show the original sentence identified as *Bernoulli_80* and its two simplified versions.

To create the cause subcorpus, we extracted semi-automatically the causal clauses as done by Gonzalez-Dios et al. (2016) and then, following the proposal of Iruskieta et al. (2016), we extracted the sentences containing causal discourse markers and causal lexical signals. The main figures of this sample are presented in Table 2.

	Original	Structural	Intuitive
Sentences	69	90	97
Words	1441	1482	1399

Table 2: Sentence and word number in our sample

The number of causal structures found in the original sentences of the CBST is shown according to their type in Table 3: i) syntactically marked causal signals (syntactic), ii) causal signals made explicit by discourse markers (DMs), iii) causal relations signaled with

[3]Note that annotation results may yield subjective idiosyncrasies, due to fact that the corpus is annotated only with one annotator. In our opinion this fact is not a problem for the aim of this paper, because our objective is to explore a methodology to measure a joint analysis between simplification and relational discourse structure. As far as we know, no agreement measures have been given in the annotation process of simplified corpora.

nouns and verbs (Lexical).

Type	Simp.	RST	Joint
Syntactic	17	3	3
DMs	16	3	3
Lexical	32	3	3

Table 3: Number of analyzed causal structures

The additional resources used in this analysis are 1) a study of the frequencies and positions of the adverbial clauses (Gonzalez-Dios et al., 2015) in order to see the frequencies of the syntactic relations; 2) the corpus *Zernola* (Gonzalez-Dios et al., 2014) to see if the syntactic relations are also used in simple texts; and 3) a lemma frequency list (Gonzalez-Dios, 2016) to see the frequencies of the discourse markers and lexical signals.

3 Framework for the Analysis of Simplified Texts

In this section, we present the framework and the annotation required to perform the analysis of simplified texts taking discourse into account.

3.1 Simplification Annotation and Analysis

Following Gonzalez-Dios et al. (2016), we propose to annotate whether the target clauses, in our case the causal relations, have been treated or not (binary tagging). If so, which operations have been performed in each structure. Besides, in this study, we add complementary descriptions such as clause length, syntactic depth (depth of the syntactic tree), surrounding phenomena or frequency information. These are the questions we propose:

a) Simplification treatment and macro-operations:
- Have the syntactic, DMs and lexical signals been treated or not? In the case of the syntactic signals,we also analyze if they have been treated or not according to the causal type defined by *Euskaltzaindia* (Euskaltzaindia, 2011): i) pure causal *-(e)lako* 'because', ii) causal explicative *bait-* 'since' and iii) pseudo-causal *-(e)nez* 'as').
- When the simplification is performed, we ask: which macro-operations have been performed? For each macro-operation, which exact operations? In the case of lexical signals, which operations according to the PoS (verbs or nouns)?

b) Length and depth
- The sentences that have been split are longer than the average sentence length of original clause?
- The sentences that have been split are inside another subordinate clause?

c) Frequencies
- In the case of the syntactic signals, are they also frequent in other corpora? For this analysis, the frequencies of other corpora are needed.
- When performing transformations, have the syntactic, DMs and lexical signals been substituted with a more frequent equivalent one?

d) Ordering
- In the case of the syntactic signals, do the reordering operations suit the word order found in other corpora or the canonical RST relation order?
- Do they suit canonical or stylistic word or sentence orders?

3.2 Discourse Annotation (RST) and Analysis

In the discourse analysis, we want to know if the relations found in the original texts have been kept, modified or deleted in the simplified texts. To that end, we follow this procedure:

- Segmentation: automatic fine-grained discourse segmentation with EusEduSeg (Iruskieta and Zapirain, 2015) and manually corrected following Iruskieta (2014). Output format: RS3.
- Rhetorical structure annotation: manually annotated with RSTTool (O'Donnell, 2000) following a modular and incremental annotation method (Pardo, 2005). Output format: RS3.
- Description if there were maintained or changed the nucleus-satellite order of the relations and the relation names with the Rhetorical DataBase (RhetDB) (Pardo, 2005).

In order to describe the simplification operations at rhetorical structure level, we propose the following questions:

a) Rhetorical relations:
- What kind of rhetorical relations were deleted from the original sentences in the intuitive corpus-set and in the structural corpus-set?

– Which relations have been added for text simplification?

b) Ordering:

– Has the nucleus-satellite order been maintained in rhetorical relations?[4]

3.3 Joint Annotation and Analysis

In order to join both analyses and based on the previous annotation, we propose to analyze the influence of simplification operations in discourse looking at the elementary discourse units (EDU), the central subconstituent (CSC)[5] and the rhetorical relations (RR). Exactly, we look the simplification operations performed which impact have on discourse. So, for each relation we make a description like the one that follows for the structurally simplified sentence presented in Table 1: *i*) an insert (*hegoaren formak* 'the shapes of the wings') has been performed in the clausal proposition; *ii*) two split and three insert operations (*izan ere, Hegoaren forma* 'due to the shape of the wings' and *ordea* 'however') in the surrounding phenomena.

Regarding rhetorical structure, we based on the simplification annotation and in the RST trees like the one presented in Figure 1, where the rhetorical structure (RS-tree) of the original text is shown above and the RS-tree of the structurally simplified text is bellow. There are three main changes in Figure 1: *i*) there is one span missing (4 above and 3 bellow), *ii*) the CAUSE relation is attached directly to the most important EDU of the RS-tree (to the central subconstituent), and *iii*) the CONCESSION relation has a new order (SN above and NS bellow) and is attached to a bigger text span (EDU_{1-2} bellow)[6].

In order to quantify and summarize that, these are the questions we propose:

a) Treatment in simplification:

– Has it been treated or not?

b) Elementary discourse unit (EDU):

– Does the EDU number remain the same? If it changes, which are the changes?

c) Central subconstituent (CSC):

– Are there any changes in the CSC? Which?

d) Rhetorical relations (RR):

– Are the RRs kept? Which ones?

– Are there new RRs?

– Which RRs have been added, modified or deleted?

This way we see how the simplification operations affect discourse.

4 Results of the Quantitative Analysis

In this section, we present the results and analysis of the causal relations (our sample) according to the framework presented in Section 3.

4.1 Results of Simplification Analysis

Treatment and macro-operations: In Table 4 we present the results in relation to the treatment in both simplification approaches. As we can see: *i*) more syntactic signals have been treated in the intuitive approach; *ii*) results in the lexical signals are similar; *iii*) and discourse markers do not seem to be treated in any case.

Treated	**Structural**		**Intuitive**	
Syntactic	47.06	(8/17)	64.71	(11/17)
DMs	25.00	(4/16)	6.25	(1/16)
Lexical	21.21	(7/34)	24.24	(8/34)

Table 4: Percentages and raw numbers of causal relations

Focusing on the different types of causal syntactic signals (Table 5), we see that there is a tendency to treat the pure causal *-(e)lako* 'because' in the structural approach, while explicative *bait-* 'since' is treated in the intuitive approach.

	Structural		**Intuitive**	
Pure *-(e)lako*	55.56	(4/9)	33.33	(3/9)
Explicative *bait-*	40.00	(2/5)	100.00	(5/5)
Pseudo *-(e)nez*	33.33	(1/3)	100.00	(3/3)

Table 5: Treated Clauses according to the causal type in both approaches

Looking at the macro-operations (Table 6) we see that, in our sample, while the syntactic signals undergo split and transformation operations, the discourse markers

[4]This is important as Mann and Thompson (1987) state: "if a natural text is rewritten to convert the instances of non-canonical span order to canonical order, it seldom reduces text quality and often improves it".

[5]The CSC is the salient EDU of a text span.

[6]Other changes were done in signaling the relations: in the signal CAUSE, the causal subordinator *-lako* 'since' was changed into the explicative connector *izan ere* 'since'.
And in the signal CONCESSION the subordinator *nahiz eta ...-n* 'in spite of' was changed into the connector *ordea* 'however'.

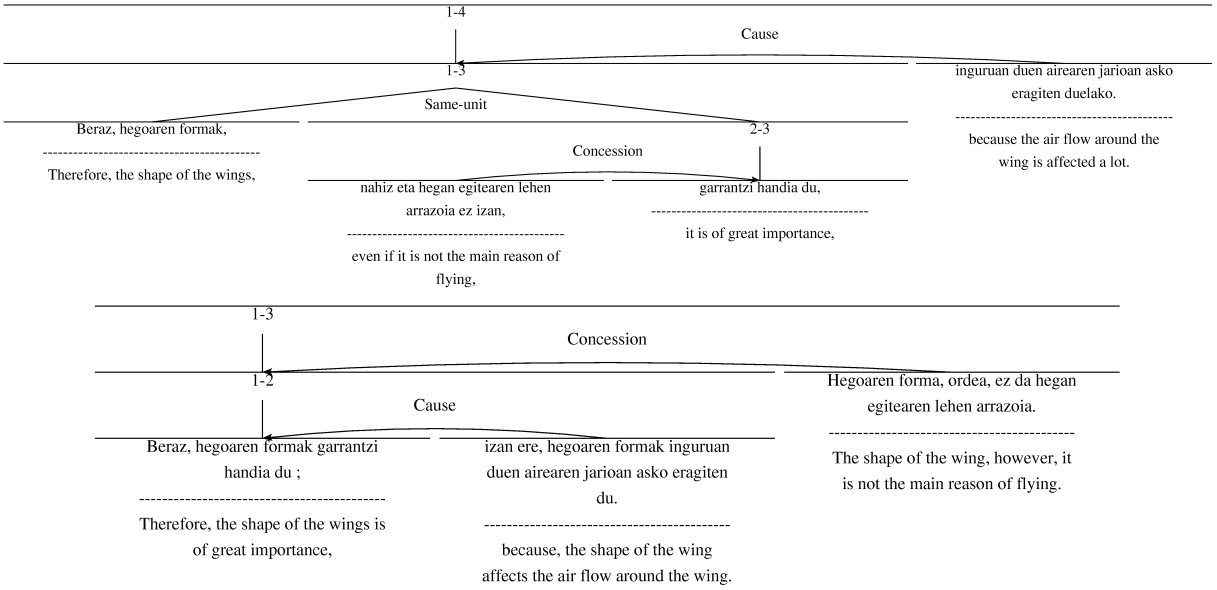

Figure 1: *Bernoulli_80* sentence's original (above) and structural (bellow) RS-trees

undergo transformations (as they are lexical units they cannot undergo splitting operations). The lexical signals undergo split and transformation operations in the structural approach, but only transformations in the intuitive.

Comparing the approaches, it is noticeable that more split operations are performed in the structural approach and more transformations in the intuitive. Exactly, the transformations performed in syntactic signals are: i) transforming a subordinate clause into a main clause ii) reformulations (more than one operations and paraphrases) and ii) changing the syntactic signal.

Regarding discourse markers, the transformation that has been performed is the substitution of a discourse marker for a more frequent one. The other macro-operations are delete and reordering.

In the case of the lexical signals, the operations performed vary according to the PoS. In Table 7 we present figures about the number of operations performed in nouns and verbs.

To summarize the analysis of the operations, we see that some macro-operations are restricted to the relation type and the PoS of it. That is, we see that no split is applied in all causal DMs or in all noun causal signals. For example, in the causal clause of sentence presented in Table 1, an insert has been performed in the structural approach; in the intuitive approach a split, a transformation (subordinate to main clause) and an insert have been performed.

Length and depth: The average length of the causal clauses in our original sample are 7 words[7]. In the intuitive approach, the split operations have been carried out in all the clauses with 7 or more words, but this only happens in 2 out of the 5 split operations carried out in the structural approach. In relation to the depth, two of the split operations in the structural approach were performed in subordinate clauses inside subordinate clauses e.g. a relative clause inside a noun clause.

Frequencies: Related to the description of the syntactic structures contained in the CBST, we have checked if they are also frequent structures in the BDT corpus[8] and in the *Zernola* corpus. As we can see, they are all frequent structures in both corpora (Table 8).

In Table 9 we present some transformation operations involving substitutions. Our analysis lead us to propose some preliminary conclusions: syntactic signals and DMs are not always substituted with more frequent equivalent ones, but with less ambiguous. As we see here, more frequent forms do not always mean simplicity.

Ordering: In relation to the reordering operations, we have analyzed whether the movements carried out

[7]As mentioned before, there are 17 clauses with syntactic relations. The longest of them has 17 words and the shortest 3. The mode is 4 words and the median 6.

[8]We consider a structure as frequent when it has more than 10 % of occurrences in its type.

Macro-oper.	Only split		Only trans		Split+trans		Only others	
Approach	**Str.**	**Int.**	**Str.**	**Int.**	**Str.**	**Int.**	**Str.**	**Int.**
Syntactic	37.5 (3/8)	9.09 (1/11)	37.5 (3/8)	81.82 (9/11)	25.00 (2/8)	9.09 (1/11)	0.00 (0/8)	0.00 (0/11)
DMs	0.00 (0/4)	0.00 (0/1)	25.00 (1/4)	100.00 (1/1)	0.00 (0/4)	0.00 (0/1)	75.00 (3/4)	0.00 (0/1)
Lexical	42.86 (3/7)	0.00 (0/8)	42.86 (3/7)	50.00 (4/8)	0.00 (0/7)	0.00 (0/8)	14.29 (1/7)	50.00 (4/8)

Table 6: Percentages and raw numbers of macro-operations performed in causal relations

Oper.	Split		Trans		Reor.		Delete	
Appr.	Str	Int	Str	Int	Str	Int	Str	Int
Noun	0	0	1	3	1	1	0	1
Verb	3	0	2	1	0	0	0	2

Table 7: Macro-operations performed in the lexical signals according to their PoS

	BDT	Zernola
Pure *-lako*	26.91	28.10
Explicative *bait-*	39.94	46.28
Pseudo *-nez*	23.94	25.62
Others	9.21	0.00

Table 8: Distribution of causal structures in the corpora BDT and Zernola

in the simplified sentences at syntactic level suit the canonical word order or the order of clauses found in EPEC. In our sample no reordering was performed at that level. But, we did find an interesting reordering in the intuitive approach: a stylistic reordering took place in the signals in order to avoid the rear-burden [9].

4.2 Results of Discourse Analysis

In Table 10, we present the results obtained with Rhetorical Database in the different corpus-sets regarding simplification approaches and rhetorical relations. The number (K) of all the relations and the differences (diff.) of each corpus-set: i) relations of the original texts (source text) in the first two columns, ii) relations of the intuitively simplified texts in the following two, and iii) relations of the structurally simplified texts in the last two.

We can observe different simplification strategies in Table 10:

- Less frequent RRs in both simplified datasets: the causal relation RESULT has less frequency in both simplified corpus-sets and CIRCUMSTANCE has also less frequency in both corpus-sets[10].
- More frequent RRs in both simplified datasets: SOLUTIONHOOD, CONCESSION and BACKGROUND are used to simplify texts.
- New RRs in one of the simplified datasets: PURPOSE, RESTATEMENT and MEANS are new relations in the intuitive approach and JOINT and PREPARATION in the structural.[11]

Using RhetDB, we extracted and presented in Table 11 the nuclearity type (SN: satellite first and nucleus after; NS: the other way around, nucleus first and satellite after) of all the hypotactic relations[12] and their frequencies.

Regarding Table 11, we see that the frequency of the causal relations (CAUSE, RESULT and PURPOSE) is bigger in the original subcorpus 0.411 (0.117 for SN and 0.294 for NS),[13] than in the intuitive 0.318 (SN: 0.09 and NS: 0.227) and structural approach 0.3 (SN: 0.00 and NS 0.3). This shows that there are less causal relations in the simplified datasets as also found by Graesser et al. (2004) and Crossley et al. (2012) and the NS order is preferred in the causal subgroup, when any causal relation is maintained.

Another interesting observation is that the NS ordering has been increased in the structural approach,

[9] "(...) "rear burden" (...) [is] the effect that occurs when some key elements for correct processing of the message (e.g. the verb) are pushed towards the end of the sentence, thus delaying and making more difficult the comprehension of the message by the receiver." (Maia-Larretxea, 2015, 68).

[10] Although SAME-UNIT (SU) is not a relation, we report it, because it was also simplified in both corpus-sets.

[11] We think that RRs such as JOINT have appear because discourse was not taken into account when simplifying texts.

[12] Note that all multinuclear or paratactic relations were excluded from this analysis.

[13] The frequencies were normalized, as follows: original cause subgroup SN: the total K of the SN divided by the total K in the subcorpus: (2+1)/(9+8).

Type	Transformation	Explanation
Syntactic	*bait- -> -(e)lako*	causal explicative substituted with a pure causal (less frequent)
DMs	*horrez gain* 'moreover' -> *gainera* 'in addition'	substituted with a more frequent
	bada 'so', 'then', 'well' -> *hala ere* 'however'	substituted with a less frequent, but less ambiguous
Signals	*eragile* 'originator','promoter' -> *arrazoi* 'reason', 'cause', 'motive'	substituted with a more frequent near synonym

Table 9: Transformation operations involving substitutions

Source text		Intuitive		Structural	
Relations	**K**	**K**	**Diff.**	**K**	**Diff.**
Result	3	1	-2	2	-1
Circumstance	3	1	-2	1	-2
*Same-unit (SU)	4	3	-1	2	-2
Solutionhood	1	3	2	2	1
Concession	2	3	1	4	2
Background	1	2	1	2	1
Purpose	0	1	1	0	0
Restatement	0	1	1	0	0
Means	0	1	1	0	0
Preparation	0	0	0	1	1
Joint	0	0	0	1	1
Cause	3	4	1	3	0
Justify	1	1	0	1	0
Condition	1	2	1	1	0
No-conditional	1	1	0	1	0
Elaboration	1	1	0	0	-1
List	3	2	-1	4	1

Table 10: Simplification strategies and rhetorical relations

	Original		Intuitive		Structural	
Relations	**SN**	**NS**	**SN**	**NS**	**SN**	**NS**
Cause	2	1	1	3		3
Justify		1		1		1
Result		3		1		2
Purpose			1			
Condition	1		1	1	1	
No-conditional	1		1		1	
Circumstance	1	2	1		1	
Solutionhood	1		3		1	1
Concession	2		3			2
Background	1		2			2
Restatement				1		
Means			1			
Preparation					1	
Elaboration		1		1		4
Total	9	8	14	8	5	15

Table 11: Nucleus/satellite ordering of the rhetorical relations in the original and simplified datasets

whereas in the intuitive approach the SN was increased (and, therefore, the NS decreased). This change brings the important message to the back of the structure and this way, it is more difficult to maintain all the information needed to understand the sentence in the memory, above all in the case of long sentences.

4.3 Joint Analysis

The results of the joint analysis of our sample are presented in Table 12. First column shows the sentence identifier, second column if it has been treated in simplification or nor, third column the changes performed in EDU frequency,[14] fourth column if the changes were performed in the CSCs, the fifth column if RRs were maintained and the sixth column if RRs were changed.

To underline these results of Table 12 we summarized the most important differences in Table 13. We observe that the simplification operations performed in the intuitive (Int.) and structural (Str.) approaches are similar when simplifying (Simpl.), maintaining or changing the EDUs (Changes in EDUs), performing changes in the CSC and maintaining the RRs. But there is a great difference when they establish a new rhetorical relation (see Table 13), because there are only 3 changed relations (underlined in bold) in common: RESULT > CAUSE, CIRCUMSTANCE > CONDITION and +CONCESSION.

4.4 Concluding remarks

As a conclusion of this joint analysis, we think that rhetorical relations of the original texts were not always

[14]The sign '+' means that there are more EDUs or that some relation was added, whereas the sign '−' means that something is missing, for example '−info' means that there is less information. The sign > means that something at the left was changed by another thing to the right.

Text	Simpl.	Changes in EDUs	Changes in CSC	Maintained RRs	Changed RRs
Etxeko_19_int	YES			List	
Bernoulli_80_int	YES			Concession, Cause	
Exoplanetakv39_int	YES	+EDU		Cause	+Restatement
Exoplanetak33_int	NO			No-conditional	
Etxeko_20_int	NO	−Same-unit	−Same-unit, −info	Circumstance, Result	−Same-unit
Etxeko_28_int	YES			Justify	Result > Solutionhood
Exoplanetak_13_int	YES	+EDU		Condition, Elaboration	**Result** > **Cause**, +Solutionhood
Bernoulli_04_int	YES			Concession	**Circumstance** > **Condition**
Bernoulli_38_int	YES	+EDU	+EDU, −info	Background	+**Concession**
Etxeko_19_est	YES	+EDU	+EDU, −info	List	+Elaboration
Bernoulli_80_est	YES	−Same-unit	−Same-unit	Concession, Cause	−Same-unit
Exoplanetak_39_est	YES				Cause > Joint (NS > NN)
Exoplanetak_33_est	YES	+EDU	−Info	No-conditional, Same-unit	+Concession, +Preparation
Etxeko_20_est	YES		+N	Circumstance, Result	+Contrast
Etxeko_28_est	NO			Justify, Result	
Exoplanetak_13_est	YES			Condition, Elaboration	**Result** > **Cause**
Bernoulli_04_est	YES		CU changed	Concession	**Circumstance** > **Condition**
Bernoulli_38_est	YES	+EDU	+EDU, −info	Background	+**Concession**, +Solutionhood

Table 12: Contingency table of the joint analysis

	Simpl.	EDU	CSC	RR
Int.	7 Yes	3 +EDU	1 −SU −info	12 kept
	2 No	1 −SU	1 +EDU −info	6 changed
Str.	8 Yes 1 No	3 +EDU 1 −SU	2 +EDU −info 1 −info 1 −SU 1 Change the CSC 1 NN	12 kept 9 changed 1 NS > NN

Table 13: Results of the joint analysis

taken into account when simplifying them (most of them were maintained). So, we want to propose for future simplification guidelines that not only lexis or syntax should be taken into account, but also discourse. That is, if in the original text there is a significant discourse relation, it should be kept in the simplified text when it helps comprehension but deleted when it leads to confusion. But the need of the discourse would not be limited to relations but to the overall relational discourse structure when simplifying text manually, the CSC and the same-unit should also be carefully treated.

For automatic texts simplification systems, the detection of the CSC should also be an important step, above all in the cases that the main piece of information should be highlighted. The difficult task of detecting the same-unit constructions could also be interesting, so that they should be deleted as much as possible.

5 Conclusion and Future Work

In this paper, we present a framework for the analysis of simplified texts taking discourse into account. In the simplification analysis, we propose to analyze the treatment and its the macro-operations, the length and depth, the frequencies and the reordering; in the discourse analysis, we propose to segment, annotate and describe the rhetorical relations; and, in the joint analysis, we propose to see the impact of simplification operations on the elementary discourse units, central subconstituents and rhetorical relations. Preliminary results show that this framework is useful to describe the simplified texts and that discourse is not always taken into account when simplifying texts in our datasets with the risk of creating not-coherent simplified texts. We have seen e.g. that some macro-operations such as the split cannot be applied to all the relations and that being more frequent does not involve simplicity as took for granted many times.

Currently, we are searching for more simplified texts in Basque to get more data and asking more people to simplify them, in order to get ride of the possible bias caused by the people who simplified the texts. Moreover, we are annotating in the Corpus of Basque Simplified Texts (CBST) more rhetorical relations to understand or describe all the simplification mechanisms. In the near future, we also want to perform this analysis with entire texts and not only sentences.

Acknowledgments

This study was carried out within the framework of the following projects: IXA group, Research Group (GIU16/16) and TUNER (TIN2015-65308-C5-1-R).

References

[Bott and Saggion2011] Stefan Bott and Horacio Saggion. 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.

[Brunato et al.2015] Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 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.

[Caseli et al.2009] Helena M. Caseli, Tiago F. Pereira, Lucia Specia, Thiago. A. S. Pardo, Caroline. Gasperin, and Sandra Aluísio. 2009. Building a Brazilian Portuguese Parallel Corpus of Original and Simplified Texts. In *the Proceedings of CICLing*, pages 59–70.

[Crossley et al.2007] Scott. A. Crossley, Max M. Louwerse, Philip M. McCarthy, and Danielle S. McNamara. 2007. A Linguistic Analysis of Simplified and Authentic Texts. *The Modern Language Journal*, 91(1):15–30.

[Crossley et al.2012] Scott A Crossley, David Allen, and Danielle S McNamara. 2012. Text Simplification and Comprehensible Input: A case for an Intuitive Approach. *Language Teaching Research*, 16(1):89–108.

[Euskaltzaindia2011] 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.

[Gonzalez-Dios et al.2014] Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Haritz Salaberri. 2014. 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, August. Dublin City University and Association for Computational Linguistics.

[Gonzalez-Dios et al.2015] Itziar Gonzalez-Dios, 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 BDT corpus]. Technical report, University of the Basque Country (UPV/EHU) UPV/EHU/LSI/TR 02-2015.

[Gonzalez-Dios et al.2016] Itziar Gonzalez-Dios, María Jesús Aranzabe, and Arantza Díaz de Ilarraza. 2016. A Preliminary Study of Statistically Predictive Syntactic Complexity Features and Manual Simplifications in Basque. In *Proceedings of the Computational Linguistics for Linguistic Complexity (CL4LC) workshop at Coling 2016*, pages 89–97.

[Gonzalez-Dios2016] Itziar Gonzalez-Dios. 2016. *Euskarazko egitura sintaktiko konplexuen analisirako eta testuen sinplifikazio automatikorako proposamena-Readability Assessment and Automatic Text Simplification. The Analysis of Basque Complex Structures*. Ph.D. thesis, University of the Basque Country (UPV/EHU).

[Graesser et al.2004] 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.

[Iruskieta and Zapirain2015] Mikel Iruskieta and Benat Zapirain. 2015. Euseduseg: A Dependency-based EDU Segmentation for Basque. *Procesamiento del Lenguaje Natural*, 55:41–48.

[Iruskieta et al.2016] Mikel Iruskieta, María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Mikel Lersundi. 2016. Kausazko koherentzia-erlazioen azterketa automatikoa euskarazko laburpen zientifikoetan [Toward a computational approach of causal coherence relations in scientific abstract texts]. *Gogoa*, 14:45–77.

[Iruskieta2014] Mikel Iruskieta. 2014. *Pragmatikako erlaziozko diskurtso-egitura: deskribapena eta bere ebaluazioa hizkuntzalaritza konputazionalean [The Discourse Structure of the Pragmatic Relations: Description and its Evaluation in Computational Lingustics]*. Ph.D. thesis, University of the Basque Country (UPV/EHU).

[Klaper et al.2013] David Klaper, Sarah Ebling, and Martin Volk. 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*, pages 11–19, Sofia, Bulgaria, August. Association for Computational Linguistics.

[Klerke and Søgaard2012] Sigrid Klerke and Anders Søgaard. 2012. DSim, a Danish Parallel Corpus for Text Simplification. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 4015–4018, Istanbul, Turkey, May. European Language Resources Association (ELRA).

[Kong et al.2017] Anthony Pak-Hin Kong, Anastasia Linnik, Sam-Po Law, and Waisa Wai-Man Shum. 2017. Measuring Discourse Coherence in Anomic Aphasia Using

Rhetorical Structure Theory. *International Journal of Speech-Language Pathology*, pages 1–16.

[Maia-Larretxea2015] Julian Maia-Larretxea. 2015. On Criteria of Professionals of the Language about the Back-burden in Basque. *Procedia-Social and Behavioral Sciences*, 212:67–73.

[Mann and Thompson1987] William C Mann and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute.

[Mann and Thompson1988] William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

[O'Donnell2000] Michael O'Donnell. 2000. RSTTool 2.4: a Markup Tool for Rhetorical Structure Theory. In *Proceedings of the first international conference on Natural language generation-Volume 14*, pages 253–256. Association for Computational Linguistics.

[Pardo2005] Thiago Alexandre Salgueiro Pardo. 2005. *Métodos para análise discursiva automática*. Ph.D. thesis, Instituto de Ciências Matemáticas e de Computação.

[Pellow and Eskenazi2014] David Pellow and Maxine Eskenazi. 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)*, pages 84–93, Gothenburg, Sweden, April. Association for Computational Linguistics.

[Petersen and Ostendorf2007] Sarah E. Petersen 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*, pages 69–72. Citeseer.

[Saggion2017] Horacio Saggion. 2017. *Automatic Text Simplification*. Morgan & Claypool.

[Simensen1987] Aud Marit Simensen. 1987. Adapted Readers: How are they Adapted. *Reading in a Foreign Language*, 4(1):41–57.

[Xu et al.2015] Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

[Young1999] Dolly N. Young. 1999. Linguistic Simplification of SL Reading Material: Effective Instructional Practice? *The Modern Language Journal*, 83(3):350–366.