Simplifying Basque Texts: the Shallow Syntactic Substitution Simplification

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

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

Abstract

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

1. Introduction and related work

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

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

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

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

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

2. Simplification framework: levels and operations

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

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

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

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

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

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

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

3. Shallow Syntactic Substitution Simplification

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

3.1. Motivation

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

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

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

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

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

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

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

Table 1: Frequency list of non-finite purpose clauses

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

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

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

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

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

3.2. Methodology

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

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

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

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

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

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

3.3. Corpus and approach

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

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

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

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

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

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

Table 3: Summary of the structures and rules

3.4. Evaluation and error analysis

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

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

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

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

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

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

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

Table 5: Simplicity of the sentences evaluated by linguists

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

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

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

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

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

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

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

4. Conclusion and future work

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

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

Acknowledgments.

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

5. References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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