First Approach to Automatic Text Simplification in Basque

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

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

Abstract

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

1. Introduction

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

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

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

2. Typology of Basque

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

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

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

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

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

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

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

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

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

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

3. Corpora analysis

We have used two corpora for this task: EPEC: *Euskararen Prozesamendurako Errenferentzia Corpusa*-Reference Corpus for the Processing of Basque (Aduriz et al., 2006a) and *Consumer* corpus (Alcázar, 2005).
EPEC corpus contains 300 000 words written in Standard Basque and it is tagged at morphological, syntactical levels (dependency-trees) (Aranzabe, 2008), and semantic level: word senses according to Basque WordNet and Basque Semcor (Agirre et al., 2006) and thematic roles in (Aldezabal et al., 2010). It is being tagged too at the pragmatic level: discourse markers (Iruskieta et al., 2011) and anaphora (Aduriz et al., 2006b).
Consumer corpus[3] is used in machine translation since the texts it contains are written in four languages (Spanish, Basque, Catalonian and Galician). It is a specialised corpus, compliling texts published the *consumer* magazine: critics, product comparison and so on.
The main characteristic of those corpora is that they contain authentic text.
In order to study the structures that should be simplified in Basque, to get better results in advanced application such as machine translation, we have taken the longest sentences from both corpora. We based our hypothesis on the results obtained by the machine translation system developed in our group when translating sentences of different length (Labaka, 2010). The results show that, the longer sentence longer, the higher error rate in Basque Spanish translation (table 1). The error rate used for scoring the results is HTER (Human-targeted Translation Error Rate) (Snover et al., 2006).

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

Table 1: Sentence length and error rate in MT

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

4. Simplification Process

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

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

Table 2: Number of sentences and sentence length in Corpora

1. **Spliting:** Make as many new sentences as clauses out of the original.

2. **Reconstruction:** Two operations take place:

   (a) Removing no longer needed morphological features like complementisers (comp). Being Basque an agglutinative language we have to remove parts of words and not a whole word in case of finite verbs.

   (b) Adding new elements like adverbs or paraphrases. The main goal is to maintain the meaning.

3. **Reordering:** Reorder the elements in the new sentences, and ordering the sentences in the text.

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

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

5. System Architecture

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

1. The first step will be to evaluate the complexity of the text by means of a system already developed by our group for the auto-evaluation of essays *Idazlanen Autoebaloaziorako Sistema (IAS)* module (Castro-Castro et al., 2008). This module examines the text in order to determine its complexity based on several criteria such as the clause number in a sentence, types of sentences, word types and lemma number among others.

2. Once a sentence has been categorised as complex in the previous step, *Mugak* module (a system created in our group for detecting chunks and clauses) (Arrieta, 2010) will help us in the task of splitting long sentences into simple ones. *Mugak* is a general purpose clause identifier that combines rule-based and statistical-based clause identifiers previously developed for Basque. It works on the basis of the output produced by several tools implemented in our group[4]:

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

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

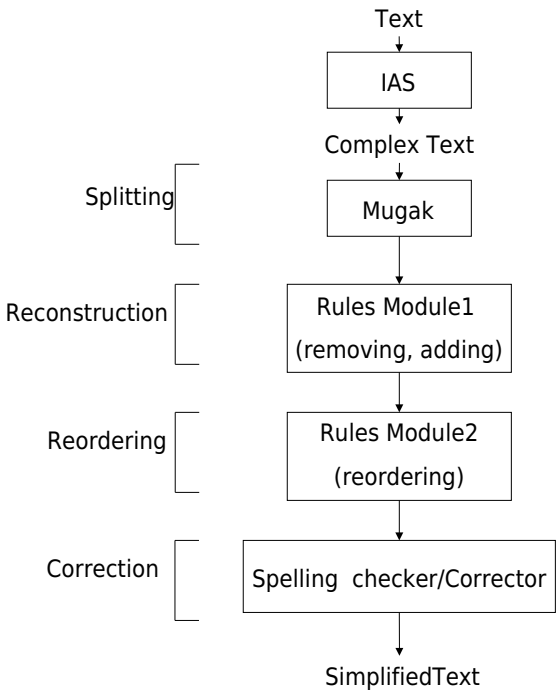

Figure 1: The architecture of system

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

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

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

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

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

Figure 2: A rule for an adverbial temporal sentences

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

6. Treated Phenomena

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

6.1. Apposition and parenthetical structures

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

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

2. (a) We remove the apposition out of the original sentence.

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

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

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

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

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

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

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

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

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

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

b. *Jose Maria Aznar Espainiako presidentea da.*

Jose Maria Aznar is President of Spain.

c. *Javier Arenas PPko idazkari nagusia da.*

Javier Arenas is Secretary-general of PP.

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

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

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

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

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

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

b. *Sabino Badajozetik fitxatua da.*

Sabino is signed up from Badajoz.

c. *Moha Barcelona Btik fitxatua da.*

Moha is signed up from Barcelona B.

d. *Aitor Ocio Athleticek utzita da.*

Aitor Ocio is transfered from Athletic.

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

6.2. Relative clauses

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

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

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

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

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

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

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

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

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

[6] Phonetically null but syntactically active element

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

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

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

6.3. Adverbial temporal clauses

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

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

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

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

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

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

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

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

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

a. *Erabakia hartu behar izan zuen.*

   'He/she/it needed to decide.'

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

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

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

7. Example

We will explain here the process explained in section 4. Sentence (9) has the three phenomena we have presented in this paper. The changes we want to point out are underlined. We use the glosses in order to illustrate the morphological process properly, when needed.

Let us explain some morpho-syntactic aspects of the sentence (9) before showing the simplification steps:

There are 5 verbs in sentence (9), and each one builds a clause. The main verb is *da*, therefore it builds the main clause. The verb *dute* is main too, but in our analysis system it is dependent on the substantive it is referring to as apposition. The periphrastic verbs *igurtzitzen ditugunean* and *sortzen den* build subordinated clauses, and contrary to *gertatu* they are inflected. The non-inflected verb *gertatu* will be simplified although it is not treated in this approach. It will be treated when we treat non-inflected verbs.[7]

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

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

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

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

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

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

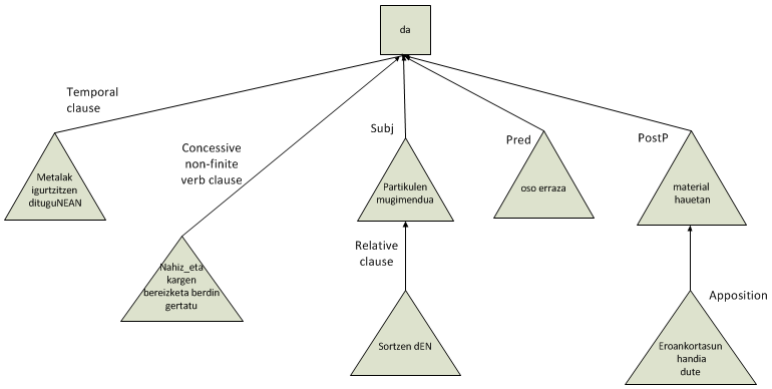

Figure 3: Tree of original sentence in example (9)

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

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

4. **Correction and Adequation:**

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

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

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

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

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

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

8. Conclusions

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

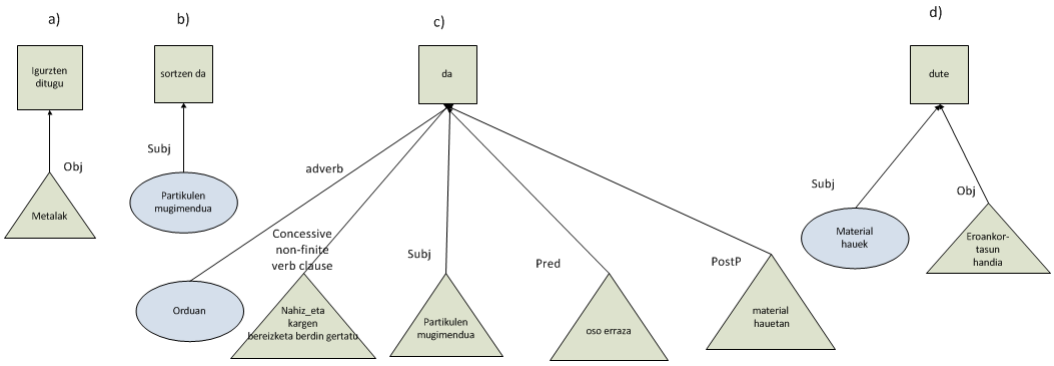

Figure 4: Tree of simplified sentences in example (10)

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

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

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

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

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

9. Acknowledgements

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

10. References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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