Making Biographical Data in Wikipedia Readable: A Pattern-based Multilingual Approach

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

Abstract

In this paper we present *Biografix*, a pattern based tool that simplifies parenthetical structures with biographical information, whose aim is to create simple, readable and accessible sentences. To that end, we analysed the parenthetical structures that appear in the first paragraph of the Basque Wikipedia, and concentrated on biographies. Although it has been designed and developed for Basque we adapted it and evaluated with other five languages. We also perform an extrinsic evaluation with a question generation system to see if *Biografix* improve its results.

1 Introduction and motivation

Parentheticals are expressions, somehow structurally independent, that integrated in a text function as modifiers of phrases, sentences..., and add information or comments to the text. Therefore, it has been argued that they interrupt the prosodic flow, breaking the intonation. According to Dehé and Kavalova (2007), parentheticals can be realised in different ways: one-word parentheticals, sentence adverbials, comment clauses and reporting verbs, nominal apposition and non-restrictive relative clauses, question tags, clauses and backtracking. Besides, the authors argue that sometimes the parentheticals are not related to the host sentence neither semantically nor pragmatically, but they are understood in the text due to the situational context.

Some parentheticals can be the result of a stylistic choice (Blakemore, 2006) and that is the case of parenthetical information found in the first paragraph of some Wikipedia articles. As stated in the Wikipedia guidelines[1] the first paragraph of the articles should contain resuming and important information. That is why the information is there so condensed. Apart from condensing the information parentheticals cause long sentences, which are more difficult to process both for humans and for advanced applications. Moreover, web writting style books (Amatria et al., 2013) suggest not to use parenthetical constructs because they make more difficult the access to the information. Simple wikipedia guidelines[2] recommend also not to use two sets of brackets next to each other.

NLP applications such as question generation systems (QG) for educational domain[3] may fail when finding important information in brackets. For example, if we want to create questions, systems such as the presented in Aldabe et al. (2013) will look for a verb[4]. In the case of parenthetical biographical information there is no verb which makes explicit when the person is born or when she or he died. So, no question will be created based on that information.

The study of parentheticals in Basque has been limited to the analysis of the irony in the narrativity of Koldo Mitxelena (Azpeitia, 2011). In the present study we analyse the parentheticals that are used in the first paragraph of the Basque Wikipedia and developed a rule-based tool *Biografix* to detect these

[1] http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style (last accessed: March, 2014)

[2] http://simple.wikipedia.org/wiki/Wikipedia:Manual_of_Style (last accessed: March, 2014)

[3] Question generation is important in learning technologies (intelligent tutoring systems, inquiry-based environments, and game-based learning environments), virtual environments and dialogue systems among others. http://www.questiongeneration.org/ (last accessed: April, 2014)

[4] Both systems (one chunk-based and another dependency-based) presented in Aldabe et al. (2013) follow the guidelines presented in Rus and Graesser (2009).

Proceedings of the Workshop on Automatic Text Simplification: Methods and Applications in the Multilingual Society, pages 11–20,
Dublin, Ireland, August 24th 2014.

structures and to create new sentences out of them. To be more concrete, we concentrate on biographical information since there are not explicit words in text that give a clue about what type of information it is. Our aim is to make more readable sentences and, consequently, to eliminate the interruption they cause. About the domain of biographies, their automatic generation has been studied (Duboue et al., 2003) in Natural Language Generation (NLG). In this research line, referring expressions to people have been studied for automatic summarisation of news (Siddharthan et al., 2011). The quality of the biographies (linguistic and content) has been recently analysed in the English Wikipedia (Flekova et al., 2014).

We want also to make a first step towards the simplification of Basque Wikipedia, since English simple wikipedia has been a great resource for Text Simplification (TS) and Readability Assessement (RA). Efforts for simple wikipedia have also been made for Portuguese (Junior et al., 2011) using TS techniques. Although *Biografix* has been specially developed for Basque, being pattern-based, we have also evaluated its adaptation to other languages. This work is not limited to wikipedia, *Biografix* can be used on other types of text as well, since these structures can be found in educational texts, newspapers and so on.

This paper is structured as follows: after this introduction we report in section 2 the treatment of parentheticals in TS and in Wikipedia. In section 3 we describe *Biografix* and in section 4 we report its evaluation. Finally, we conclude and outline the future work in section 5.

2 Parenthetical Structures

In this section we report the treatment that parenthentical structures have undergone in TS and other NLP applications. We also describe the parentheticals found in Basque Wikipedia.

Parentheticals have been object of study in TS and three main operations have been proposed: a) parentheticals have been removed out of the texts (Drndarević et al., 2013), b) parentheticals have been removed but they have been kept in another form (Aranzabe et al., 2012; Seretan, 2012) or c) parentheticals have been added to explain the meaning by short paraphrases (Hallett and Hardcastle, 2008) or hyperonyms (Leroy et al., 2013). In any case, it is usually recommended to avoid them (Aluísio et al., 2008). In other NLP applications such as summarisation they are usually removed and even some QG works follow the same strategy, in case they are not relevant (Heilman and Smith, 2010).

2.1 Parenthetical Structures in Basque Wikipedia

Wikipedia guidelines emphasise the importance of the first paragraph. It should indeed contain a summary of the most significant information. To concentrate all the information, stylistic resources such as parenthentical structures are used. The information that is written in brackets in the Basque Wikipedia can be classified in two groups: a) information about people and b) information about concepts. About people biographical data and mandates are usually found and about concepts the etymology of words is frequent. Translations or transliterations of the named entity or the concept is found for both groups.

On the other hand, there are other frequent parenthetical structures that are found in the first paragraph, but they are not written in brackets. This is the case of the nicknames, which are written in commas. This kind of information is also found in other languages. After this analysis, we decided to concentrate on biographical data to create new sentences out of that information.

Biographical data Contrary to English Wikipedia, in Basque Wikipedia the information contained in bracket is, if known, birthplace (town, province, state), date of birth, and if the person is dead, date of death and place of dead. This is the case as well of the Catalan, Spanish, Italian, Portuguese, German and French Wikipedia among others, although sometimes paraphrases are found in brackets. For French there is, for example, more than a way to write the biographical data[5].

In Basque Wikipedia guidelines[6] it is stated that biographical data should be written as in examples 1 and 2. If the person is dead, we see in example 1 that the birth data (town, state and date) and the death data (town, state and date) are linked by a dash.

[5] `http://fr.wikipedia.org/wiki/Wikipedia:Conventions_de_style` (last accessed: March, 2014)
[6] `http://eu.wikipedia.org/wiki/Wikipedia:Artikuluen_formatua` (last accessed: March, 2014)

(1) *Ernest Rutherford, Nelsongo lehenengo baroia, (Brightwater, Zeelanda Berria, 1871ko abuztuaren 30a - Cambridge, Ingalaterra, 1937ko urriaren 19a) fisika nuklearraren aita izan zen.*
'Ernest Rutherford, 1st Baron Rutherford of Nelson, (Brightwater, New Zeeland, 30th August, 1871 - Cambridge, England, 19th October, 1937) was the father of the nuclear Physics.'

And if the person is alive, only birth data (town, province, date) is provided as in example 2.

(2) *Karlos Argiñano Urkiola, nazioartean Karlos Arguiñano grafiaz ezagunagoa, (Beasain, Gipuzkoa, 1948ko irailaren 6a) sukaldari, aktore eta enpresaburu euskalduna da.*
'Karlos Argiñano Urkiola, internationally known with the Karlos Arguiñano spelling, (Beasain, Gipuzkoa, 6th September, 1948) is a basque chef, actor and businessman.'

In both cases, the places (if known) should precede the date and these should be separated by commas. However, biographical data is not frequently written uniformly. Places do not precede the date, the date is incomplete (only year) and sometimes other characters like the question mark appear to denote that the place or the date is known.

Taking into account this guidelines and the articles we have analysed, we have developed *Biografix*, a pattern based tool that detects biographical data and creates new sentences with this information. This tool was originally developed for Basque but it has been adapted to other languages. An adaptation of this tool, moreover, could be used as a first step into Text Summarisation, if we only remove the parentheticals and do not create new sentences.

Biographical information is contained in brackets in other Wikipedias as well but formats may be different. The way of writing, for example, in Catalan, German and Portuguese is similar to Basque. In Spanish, French, and Italian that format is also used but, as mentioned beforehand, other formats are also accepted.

3 Inside *Biografix*

Biografix is a pattern-based tool that simplifies the biographical data and creates new sentences out of that information. Having as an input the example 1 in subsection 2.1, *Biografix* will produce the sentences 3, 4, 5, 6 and 7.

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

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

(5) *Brightwater Zeelanda Berrian dago.*
'Brightwater is in New Zeeland.'

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

(7) *Cambridge Ingalaterran dago.*
'Cambridge is in England.'

So, if the person is dead, *Biografix* will write first the main sentence (3) followed by a new sentence (4) with the information about the birth. If the birthplace is composed by more than a place entity, sentences like (5) will be written. After the birth information, a sentence will contain the information about the death (6). For the cases that more than a place appear, those will be rewritten (7).

If the person is alive like in example 2 in subsection 2.1, the same process will take place, but no death information will appear by creating the new sentences 8, 9 and 10.

(8) *Karlos Argiñano Urkiola, nazioartean Karlos Arguiñano grafiaz ezagunagoa, sukaldari, aktore eta enpresaburu euskalduna da.*
'Karlos Argiñano Urkiola, internationally known with the Karlos Arguiñano spelling, is a basque chef, actor and businessman.'

(9) *Karlos Argiñano 1948ko irailaren 6an Beasainen jaio zen.*
'Karlos Argiñano was born on the 6th of September, 1948 in Beasain.'

(10) *Beasain Gipuzkoan dago.*
'Beasain is in Gipuzkoa.'

So, first, main information will be kept (8) and then the information about the birth will appear (9). As a second place information (the province) original sentence (2), it will be rewritten as well (10).

We have to mention that we use the title of the article as the subject of the sentences containing the biographical information. That is way we see that in sentences 9 and 10 the subject is *Karlos Argiñano* and the subject in sentence 8 is *Karlos Argiñano Urkiola*. We took this decision for cases where the real name of person is not so known, e.g. Cherilyn Sarkisian. Had we used Cherilyn Sarkisian in all the sentences, would someone have known we are talking about Cher?

To carry out these simplifying transformations *Biografix* follows the simplification process explained in Aranzabe et al. (2012):

- **Splitting:** In this stage we get the parts of the sentences we are going to work with. To that end, three steps take place: a) the parenthetical structure is removed from the original sentence; b) the type of parenthetical expression is checked looking at whether there are birth and death data or only the former; c) dates and places are split. We use simple patterns to detect the dates and the places. As it is possible to find more than a place, they will be split by the commas. This stage is common for all the languages.

- **Reconstruction:** The new simplified sentences are created in this stage. This part is language-dependent, since we add the verbs, determinants, prepositions and case markers. In the case of Basque we also remove the absolutive case that is found in some articles[7]. Anyway, we create three kind of sentences that are common for all the languages with the constructs obtained in the splitting stage: a) sentences indicating birth data, b) sentences indicating death data and c) sentences indicating place specifications. The main sentence will be kept as in the original version (the parenthetical has been removed in the splitting stage).

- **Reordering:** The sentences will be ordered in text. First, the main sentence; second, the information about the birth; if there is more than a place, the following sentences will contain that information (place specifications); third, the information about the death (if dead) and finally, the death place specifications.

- **Correction:** The aim of this stage is to check if there are any mistake in the new sentences and to correct them. As one of our goals is to know the correctness of *Biografix*'s output this stage has not been implemented yet.

Biografix has been designed for Basque and then the reconstruction stage has been adapted to other 7 languages: French, German, Spanish, Catalan, Galician, Italian and Portuguese. To develop the Basque version we implemented the guidelines in Wikipedia (see subsection 2.1) and we used a small corpus of 50 sentences to find possible cases, where the guidelines are not fulfilled. These 50 sentences were randomly crawled.

For other languages, we did not make any change in the splitting stage but for German. According to German Wikipedia guidelines birth and death data are separated by a semicolon and not by a dash. Although French, Spanish and Italian have other options to express the biographical information between bracket we did not implement them. Our aim is not to create a tool specially for these languages, but to see if the design for Basque can be applicable to other languages. That is why, the adaptations to other languages are available at our website[8], if someone wants to improve them.

[7]The absolutive case is used according to the format of the date.

[8]`https://ixa.si.ehu.es/Ixa/Produktuak/1403535629`

Other improvements could be done in the reconstruction stage. To rewrite the sentences we have used the most familiar past tense in each language. The only exception was French. The most familiar past tense according to the context is the *passé composé* but this tense requires the agreement of the gender between subject and verb[9]. As the *passé simple* is not very familiar we decided to use the present tense to avoid the concordance problem. So, this could be one of the things to take into account for future developers.

No other changes should be done in the reordering stage but the correction has to be adapted to each language. No training was performed for the other languages. Only 3-5 sentences were used to check that there were no errors.

4 Evaluation

In order to evaluate *Biografix* we crawled the first sentence of 30 Wikipedia articles. The method to select these articles was the following: a) we used CatScan V2.0β[10] to get a list of the Biographies in Basque Wikipedia; b) we randomised that list and make another list to see which articles were written in 8 languages (Basque, Catalan, French, Galician, German, Italian, Portuguese and Spanish); c) we selected the first 32 articles. The first two articles were used to explain and train the annotators. The final test-sample had, therefore, 30 items.

Having that sample, we performed two evaluations: a manual evaluation (section 4.1) and a extrinsic evaluation with a question generation system (section 4.2).

4.1 Manual evaluation

The manual evaluation was carried out for 6 languages: Basque, Catalan, French, Galician, German and Spanish. 10 linguists took part in the evaluation process and they evaluated three aspects of the task: the original sentences (*JatTestua*), *Biografix* performance (*Prog*) and the grammaticality of the new simplified sentences (*Gram*). In total they answered nine yes/no questions. This evaluation method we are proposing is useful to perform an error analysis and find out which are the weak points of our tool.

To evaluate the performance and the adaptation of *Biografix* we chose six languages according to the format of the biographical data: i) Basque (the language *Biografix* has been designed for) ii) Catalan (same format as Basque), iii) German (same format but a slightly variation), iv) Spanish (same format as Basque but other options as well), v) French (same format as Basque in one of the parenthetical formats and other options), vi) Galician (without defined format). Portuguese and Italian were not evaluated because their case studies were already evaluated with Catalan and Spanish. All the sample were evaluated by two annotators except for Catalan and Galician, because Catalan has the same case study as Basque and Galician has not a predefined format that could cause confusion.

Questions concerning the original sentences (*JatTestua*) Three questions were presented in regards to the original sentence in Wikipedia. The aim is to know if the original sentences do have parenthetical structures and therefore, how many of them are candidates to simplification (coverage).

1. Are there parenthetical structures written between brackets?

2. Is the sentence grammatically correct and standard?

3. Is the punctuation correct?

We asked about the grammaticality and the punctuation of original sentences (correctness) because it was shown in Aldabe et al. (2013) that many source sentences were incorrect and that fact decreased the performance of the question generators and the correctness of the created questions.

[9] e.g. *Cher est née en Californie.*, but *Ernest Rutherford est né en Angleterre*.

[10] `http://tools.wmflabs.org/catscan2/catscan2.php` (last accessed: March, 2014)

Questions concerning the performance of *Biografix* (*Prog*) Four questions were designed to check if *Biografix* carries out the process it has been implemented for (precision).

1. Have parenthetical structures been removed?
2. Is all the information kept?
3. Taking into account the original sentence, is all the information correct?
4. Is there new information?

Second and third questions are essential to know if at rewriting in the reconstruction stage no information has been omitted or changed. The aim of the forth question is to know, for example, if sentences with other kind of information like translations have been added and treated as biographical or if a sentence referring to the death of a living person has been created.

Questions concerning the grammaticality of the new simplified sentences (*Gram*) Two questions were prepared to check the correctness of the simplified questions, since to create correct sentences is very important to understand the text. These questions should be answered for each simplified sentence (grammatical precision).

1. Is the sentence grammatically correct and standard?
2. Is the punctuation correct?

If these questions get negative results, we cannot forget that in our simplification study we consider the correction as a last step. This way, the output of *Biografix* will be checked and, were there any mistakes, they would be corrected.

4.1.1 Results of the manual evaluation

In table 1 we present the results obtained in the manual evaluation and it shows the results considering the following measures:

1. The coverage is the percentages out of 30 (the size of the sample) *Biografix* processed, that is, the sentences that had parentheticals.
2. The correctness is the percentage of the source sentences whose grammar and punctuation is correct.
3. The recall is the division between the number of the created simple sentences and the number of the sentences it should have created taking into account all the information in the original sentences.
4. The precision is the division between the correct performed, that is, all the *Prog* questions have been correctly answered and the processed sentences. We call this precision at performance.
5. The grammatical precision is the correctly created sentences among the created sentences.

In the second-last column we show the κ agreement of the evaluators (Cohen, 1960). As we have few examples, the expected agreement is very high and it causes low scores. That is the reason why we also show the percentage agreement (observed agreement) in the last column.

Taking a look at the results for Basque, we see that Biografix is able to create almost all the sentences (recall: 0.94) and that they are correct (grammatical precision: 0.87), although there are little problems keeping all the information and keeping it right (precision: 0.79). Taking into account that the percentage of the correct source sentences is low (82.76), we follow Aldabe et al. (2013) and recalculate the results without the incorrect sentences. This way, recall is 0.93, precision is 0.80, grammatical precision is 0.88. As we see, results do not vary that much, since the grammaticality of the source sentence has only influence in the first of the created sentences. About the agreement between annotators, we see that κ is really low (0.37) due to the few disagreements that annotators had above all about the grammar. However, the observed agreement is high (90.63).

Language	Coverage	Correctness	Recall	Precision	Gram. Prec.	κ	%
Basque	97.00	82.76	0.94	0.79	0.87	0.37	90.63
Catalan	93.33	98.21	0.77	0.53	0.78	-	-
French	73.00	88.64	0.80	0.18	0.37	0.39	85.06
Galician	43.00	88.46	0.76	0.15	0.62	-	-
German	100	100.00	0.78	0.60	0.78	-	100
Spanish	100	85.00	0.71	0.33	0.67	0.52	88.76

Table 1: Results of *Biografix* language by language

In the case of Catalan, we see that *Biografix* is not able to create as many sentences as information in the original source (recall: 0.77) and this tendency occurs in the other languages as well. Precision at performance goes down (0.53) due to added and lost information but grammatical precision is acceptable (0.78). We think, that this is a quite satisfactory adaptation.

The results for French indicate that something went wrong. There is more than a way to express the biographical information and, as expected, the performance goes down. The precision is very low (0.18) due to the fact that a lot of information is lost and as sometime paraphrases do appear in the original sentence, this fact implies grammatical error. Anyhow, the recall is acceptable (0.80) and that is a good starting point for the further development of French version. The average of the obtained κ measures is really low (0.39) and that is why having few instances Cohen's kappa penalises the disagreement too much.

The case of Galician is quite different. It is not stated in the guidelines how biographical data should be written and the parentheticals we found are few (coverage: 43.00) and different from the Basque. However, we wanted to try *Biografix* and what we see is, that, although its precision at performance is really low (0.15), the created sentences are quite correct (0.62). We think the Galician Wikipedia should be analysed thoroughly and then *Biografix* should be adjusted.

The German version of *Biografix* was able to simplify all the sentences found in the test-sample and its recall is high (0.88). Its weak point is the precision at performance (0.60), as in other languages, due to the fact that the second question of *Prog* is not satisfied. The sentence it creates are quite acceptable (0.71) as well. Surprisingly, both linguists agreed in all the cases and questions. So, we conclude that the German adaptation was successful.

Finally, in the case of the Spanish adaptation, we see that the precision is very low (0.33) since there was an important information loss. However, the grammatical precision (0.67) is acceptable. Although κ is higher (0.52) than in other languages, observed agreement is not far from Basque (88.76). It is remarkable as well that being Spanish a long time normalised language only the 85.00 % of the source sentence are correct and that although there are other formats to express the biographical information the coverage is absolute (100.00).

The main disagreement was found when evaluating the grammar and the punctuation due to different criteria of the annotators. For some of them sentences without verb were correct because they considered that there was an elided verb. In our opinion, as we are trying to simplify, we think that all the sentences should have a finite verb. Annotators did not have to much trouble to answer the four *Prog* questions, so we think that this is a good methodology, and, moreover, it makes easy to perform error analysis. We want to point out that κ has not been the best measure but we have used it as we consider that it is a standard to measure data reliability.

To conclude, we find that there is room to improve the versions in other languages, above all trying not to lose information but the adaptation of *Biografix* has been a good starting point. In fact, the adaptation has been quite satisfactory for German and Catalan, because they share the format with Basque but they should be further analysed. As foreseen, the languages with different formats like Galician, Spanish and French require a bigger analysis.

4.2 Extrinsic evaluation

To evaluate the performance of *Biografix* throughout another NLP advanced application, we used the web application *Seneko* (Lopez-Gazpio and Maritxalar, 2013)[11], the application of the chunk-based question generation system for educational purposes presented in Aldabe et al. (2013). This kind of evaluation was only performed for Basque.

We ran *Seneko* with the original sentences and the simplified sentences. The number of the generated questions is presented in table 2. We break down the results on the basis of the case markers as well. In agglutinative languages like Basque case markers are the morphemes that express the grammatical functions.

Source file	Total	Absolutive	Inessive	Genitive	Other
Original sentences	34	23	7	2	2
Simplified sentences	142	65	66	8	3

Table 2: Questions generated by *Seneko* using the original and the simplified sentences

Using as input the original sentences *Seneko* is able to create 34 questions, more or less a question per sentence. 23 of them have been generated for the absolutive case, that is, for the subject and the predicative, and only 7 of them have been generated for the inessive. Taking into account that we are working with biographical information, this is a bad result because the inessive case in Basque is used to express time and place relations. That is, the inessive is used to create questions with the question words *When* and *Where*. On the other hand, using as source the simplified sentences, 65 questions have been generated for the absolutive and 66 for the inessive. This way, we see that using *Biografix*'s output *Seneko* has been able to generate questions about place and time expressions.

Next, in 11 and 12 we show an example of the questions generated by *Seneko*. In 11 we find that using the original input it was only able to create a question, and it makes no sense but using the simplified text (example 12) *Seneko* creates two correct questions.

(11) a. **Source text:** *Eduardo Hughes Galeano (Montevideo, 1940ko irailaren 3a -) Uruguaiko kazetari eta idazlea da.*
'Eduardo Hughes Galeano (Montevideo, 3rd of September, 1940 -) is an Uruguayan journalist and writer.'

b. **Generated question:** *Nor da Eduardo Hughes Galeano Montevideo 1940ko irailaren 3a?*
'Who is Eduardo Hughes Galeano Montevideo 3rd of September, 1940?'

(12) a. **Simplified text:** *Eduardo Hughes Galeano Uruguaiko kazetari eta idazlea da. Eduardo Galeano 1940ko irailaren 3an Montevideon jaio zen.*
'Eduardo Hughes Galeano is an Uruguayan journalist and writer. Eduardo Hughes Galeano was born the 3rd of September, 1940 in Montevideo.'

b. **Generated questions:** *Nor jaio zen 1940ko irailaren 3an Montevideon? Non jaio zen Eduardo Galeano 1940ko irailaren 3an?*
'Who was born on the 3rd of September, 1940 in Montevideo? Where was born Eduardo Hughes Galeano on the 3rd of September, 1940?'

This way, we conclude that *Biografix* is an useful tool to improve the performance of question generation systems like *Seneko*.

5 Conclusion and future work

In this paper we have presented *Biografix*, a tool that detects parenthetical structures and simplifies the biographical data in order to create new more readable sentences. Although *Biografix* has been

[11] `http://ixa2.si.ehu.es/seneko/` (last accessed: March, 2014)

designed and developed for Basque, we have applied it to the parenthetical biographical information written in other seven languages: French, German, Spanish, Catalan, Galician, Italian and Portuguese. The results of the evaluation show that the Basque version obtains very good results but the adaptations should be further developed. Anyway, good results have been obtained for Catalan and German and promising for Spanish and French. Besides, we have shown its validity through an extrinsic evaluation with *Seneko*, a question generation system. These systems are important for the educational domain, and the improvement *Biografix* offers is considerable. Although we have used Wikipedia to develop and evaluate *Biografix*, it can be used for other kind of text with parenthetical biographical information.

For the future, we plan to continue analysing and implementing rules for other kind of parenthetical structures like etymology, translations of named entities or mandates of relevant people. We also plan to link the entities to the their articles in Wikipedia to offer additional information. Patterns could also be improved using previously developed analysers or tools, but this way the splitting stage will become language-dependent. Moreover, we cannot forget that this work is included in the main framework of the TS system for Basque that we are developing and this is another step towards the main aim of getting easier and more readable Basque texts.

Acknowledgements

Itziar Gonzalez-Dios's work is funded by a PhD grant from the Basque Government. We thank Aitor Soroa for his help with the language links in Wikipedia. A great part of this work would have not been able without the collaboration of the linguists Itziar Aduriz, Izaskun Aldezabal, Begoña Altuna, Nora Aranberri, Klara Ceberio, Ainara Estarrona, Mikel Iruskieta, Mikel Lersundi and Uxoa Iñurrieta. We also do appreciate the help Ander Soraluze offered during the implementation of *Biografix* and Oier Lopez de Lacalle for his quick tutorial on R. This research was supported by the Basque Government (IT344-10), and the Spanish Ministry of Science and Innovation, Hibrido Sint project (MICINN, TIN2010-202181).

References

Itziar Aldabe, Itziar Gonzalez-Dios, Iñigo Lopez-Gazpio, Ion Madrazo, and Montse Maritxalar. 2013. Two Approaches to Generate Questions in Basque. *Procesamiento del Lenguaje Natural*, 51:101–108.

Sandra M. Aluísio, Lucia Specia, Thiago A. S. Pardo, Erick G. Maziero, Helena M. Caseli, and Renata P. M. Fortes. 2008. A corpus analysis of simple account texts and the proposal of simplification strategies: first steps towards text simplification systems. In *Proceedings of the 26th annual ACM international conference on Design of communication*, SIGDOC '08, pages 15–22, New York, NY, USA. ACM.

Xabier Amatria, Urtzi Barrenetxea, Irene Fernández, Rakel Olea, Joseba Uskola, and Izaskun Zuntzunegi. 2013. *Komunikazio elektronikoa. IVAPen gomendioak web-orriak idazteko*. IVAP.

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

Agurtzane Azpeitia. 2011. Enuntziatu parentetikoak: Koldo Mitxelenaren intentzio ironikoaren ispilu. *Gogoa*, 10(1&2).

Diane Blakemore. 2006. Divisions of labour: The analysis of parentheticals. *Lingua*, 116(10):1670–1687. Language in Mind: A Tribute to Neil Smith on the Occasion of his Retirement.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20:37–46.

Nicole Dehé and Yordanka Kavalova. 2007. Parentheticals. An introduction. In Nicole Dehé and Yordanka Kavalova, editors, *Parentheticals*, pages 1–22. John Benjamins Publishing Company.

Biljana Drndarević, Sanja Štajner, Stefan Bott, Susana Bautista, and Horacio Saggion. 2013. Automatic Text Simplification in Spanish: A Comparative Evaluation of Complementing Modules. In *Computational Linguistics and Intelligent Text Processing*, pages 488–500. Springer.

Pablo A. Duboue, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. 2003. PROGENIE: Biographical Descriptions for Intelligence Analysis. In *Proceedings of the 1st NSF/NIJ Conference on Intelligence and Security Informatics*, ISI'03, pages 343–345, Berlin, Heidelberg. Springer-Verlag.

Lucie Flekova, Oliver Ferschke, and Iryna Gurevych. 2014. What Makes a Good Biography?: Multidimensional Quality Analysis Based on Wikipedia Article Feedback Data. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 855–866, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Catalina Hallett and David Hardcastle. 2008. Automatic Rewriting of Patient Record Narratives. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Joseph Maegaard, Benteand Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Michael Heilman and Noah A Smith. 2010. Extracting Simplified Statements for Factual Question Generation. In *Proceedings of QG2010: The Third Workshop on Ques-tion Generation*, page 11.

Arnaldo Candido Junior, Ann Copestake, Lucia Specia, and Sandra Maria Aluísio. 2011. Towards an on-demand simple portuguese wikipedia. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 137–147. Association for Computational Linguistics.

Gondy Leroy, David Kauchak, and Obay Mouradi. 2013. A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty. *International Journal of Medical Informatics*, 82(8):717–730.

Iñigo Lopez-Gazpio and Montse Maritxalar. 2013. Web application for Reading Practice. In IADAT, editor, *IADAT-e2013: Proceedings of the 6th IADAT International Conference on Education*, pages 74–78.

Vasile Rus and Arthur C. Graesser. 2009. The Question Generation Shared Task and Evaluation Challenge. In *The University of Memphis. National Science Foundation*.

Violeta Seretan. 2012. Acquisition of Syntactic Simplification Rules for French. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Advaith Siddharthan, Ani Nenkova, and Kathleen McKeown. 2011. Information Status Distinctions and Referring Expressions: An Empirical Study of References to People in News Summaries. *Comput. Linguist.*, 37(4):811–842, December.