Toward a punctuation checker for Basque

I. Aldezabal, M.J. Aranzabe, B. Arrieta (1), M. Maritxalar, M. Oronoz

Affiliation: IXA Group (http://ixa.si.ehu.es)
University of the Basque Country (UPV/EHU)
Postal address: Faculty of Computer Science
649 p.k. 20080 Donostia
(Basque Country)
Tel./fax: +34 943 015 061 / +34 943 219 306
E-mail (1): bertol@si.ehu.es

Until some years ago, researchers in computational linguistics have ignored punctuation. Nevertheless, since the publication of Nunberg's monograph [Numberg G., 1990], punctuation works have increased [Bayraktar M. *et al.*, 1998] [Hardt D., 2001] [Pala K. *et al.*, 2003], and, recently, it is used more and more for different tasks of Natural Language Processing. Our research group[1] has been working in the area of NLP for the last 14 years, and as a result of the research and tools developed, some commercial applications have been built, such as XUXEN (a spell checker for Basque) [Aduriz *et al.*, 1997], and bilingual or monolingual dictionaries that are all integrated in different text processors. Nowadays, among other things, we are working in the prototype of a syntactic checker that would include a robust punctuation checker. We think that this punctuation checker would allow segmenting quite easily a text into clauses and sentences[2], and, consequently, it would facilitate the detection of some other syntactic errors. In fact, we think that a complete understanding of written language would be impossible if punctuation marks were not taken into account.

However, the task of developing a punctuation checker implies an additional problem: the fact that the punctuation rules are not *totally* established. In general, there is no problem when using the full stop, the question mark or the exclamation mark. Furthermore, the errors related to them (putting or not the initial question or exclamation mark depending on the language, for instance) are not so complex to treat. In fact, we have already defined some rules to solve some of them. In contrast, comma is the most polyvalent and, thus, the less defined punctuation mark. The matter is that there are not very fixed rules about the comma. It exists some intuitive and generally accepted rules, but they are not used in a standard way. That is why in the first stage of this study we have specially stressed on the treatment of this sign.

In Basque, this problem gets even more evident, since the standardisation and normalisation of the language began about twenty-five years ago. Thus, we contacted an expert in the area, Juan Garzia[3], who has written a book about syntax and writing [Garzia J., 1997] where the Basque punctuation is treated widely. This author has developed a complete theory for punctuation in Basque, and he has done an effort to spread it over many important Basque environments, as *Berria* (the only newspaper in Basque), UPV/EHU (University of the Basque Country), etc. Although we will follow mostly Garzia's theory, it has to be said that the principles that we are going to describe in this paper are generally accepted in the administration and the teaching community.

Summarising, in this paper we present i) the working procedure with the experts in the area, mainly with Garzia, ii) the basic concepts and principles for the correct use of comma, and iii) the steps followed for the formalisation and implementation of the extracted principles, with the aim of doing them more declarative and applicable to NLP tasks.

The working procedure with the experts

Getting information from experts and formalising it for NLP tasks was not easy. Firstly, we did several meetings with Garzia to extract all the information of his theory, and summarise and schematise it from a computer science point of view. To do this, we used a cyclic methodology. We formed a group of 5 people —3 computer scientists and 2 Basque linguists of our NLP group—, and all of us went to the meeting with the expert. There, he explained us his theory, and after, we got together to discuss our notes and write the pertinent report. In this report, we wrote, on one hand, the information extracted in a schematised and formalised way, and, on the other hand, the doubts we had or the things that were not clear in order to ask the expert in the next meeting. Then, we started the cycle from the beginning, doing another meeting with the expert, and so on, until the final report was written.

After this, we wanted to contrast this information with another expert in the area, and we contacted Joxe Ramon Etxebarria[4]. Essentially, both Garzia and Etxebarria, have the same opinion. The main difference lies in the use of comma in short sentences, as we will see in the next section.

[1] http://ixa.si.ehu.es/

[2] A clause is a group of words containing a verb. Sentences contain one or more clauses [Collins, 1995].

[3] Juan Garzia is a Basque writer, a translator and an expert in syntax and punctuation. Besides, he works in the Institute for Basque in the University of the Basque Country (UPV/EHU).

[4] Joxe Ramon Etxeberria is the proofreader of the Summer Basque University (UEU)

The use of comma: the *focus* as the key

Before starting with the theory, it is important to explain the concepts of *focus* and *topic*, which in Basque are defined by the order of the elements in the sentence.

The *focus* is the most important part of the sentence, the part of the sentence that is wanted to remark. In Basque, it goes almost always just before the verb. The *topic*, on the contrary, is the part of the sentence that gives the theme you are writing about. It usually goes at the beginning of the sentence, introducing it.

The most important rule of the theory is based on the *focus* and it says the following: a comma must not be put between the *focus* and the verb of the sentence. And this simple asseveration is the crux of the matter. All the theory is based on it. Let's see an example:

Oinez	joan ginen.
(on foot)	(we went)
FOCUS	VERB

Figure 1: A comma must not be put between the *focus* and the verb of the sentence

The theory also assumes that the *topic* of the sentence has to be followed by a comma when it goes before the verb:

Atzo	,	oinez	joan ginen.
(Yesterday)		(on foot)	(we went)
TOPIC	,	FOCUS	VERB

Figure 2: The *topic* of the sentence has to be followed by a comma when it goes before the verb

Nevertheless, Joxe Ramon Etxebarria claims that this law may be relaxed: he says that the first comma may be removed if the mentioned *topic* is short enough and if there is no need for disambiguation (see figure 3). But this principle seems too subjective. And this is why Juan Garzia restricts this principle and proposes the obligatory comma after the *topic*, although he thinks that this comma is optional in some special cases. In fact, he has built a theory where all cases are contemplated and there is no place for subjective matters. So, he claims that, in this case, it is not necessary to remove the comma. In this aspect, we agree with Juan Garzia: if we want to do a punctuation checker, we need principles that can be converted into fixed rules. However, transforming Joxe Ramon Etxeberria's asseveration into an objective principle would be very useful to get a more flexible punctuation checker and to let the user decide in this concrete case. With this aim, we have resolved that a *topic* is short enough or that there is no need for disambiguation, when neither the *topic* nor the *focus* is a clause. So, we have decided to allow the user putting or not the corresponding comma after the *topic*, when the mentioned *topic* and the following *focus* are not clauses. In other words: a comma should be between the *topic* and the *focus*, if any of them is a clause (see figures 3, 4 and 5).

Atzo	oinez	joan ginen	zinera.
(Yesterday)	(on foot)	(we went)	(to the cinema)
Non clausal TOPIC	non clausal FOCUS	VERB	TOPIC

Figure 3: Example of a non clausal *topic* and its punctuation

Autoa izan arren	,	oinez	joan ginen	zinera.
(In spite of having a car)		(on foot)	(we went)	(to the cinema)
Clausal TOPIC	,	non clausal FOCUS	VERB	TOPIC

Figure 4: Example of a clausal *topic* and its punctuation

Atzo	,	lehengo astean erositako autoarekin	joan ginen	zinera.
(Yesterday)		(with the car bought last week)	(we went)	(to the cinema)
Non clausal TOPIC	,	clausal FOCUS	VERB	TOPIC

Figure 5: Example of a non clausal *topic*, but a clausal *focus*, and its punctuation

With regard to the part of the sentence that is after the main verb, the theory is more flexible. If there is not any clause after the main verb, the comma will be optional. But if there is a clause after the main verb, (no matter if the clause has the *topic* function or the *focus* function), a comma should be just before the clause, and a punctuation mark after the clause. See the following figure:

Oinez	joan ginen	(,)	zinera	,	berandu genbiltzan arren	.
(On foot)	(we went)		(to the cinema)		(although we were late)	
Non clausal FOCUS	MAIN VERB	(,)	NOT A CLAUSE	,	CLAUSE	.

Figure 6: Example of the punctuation of the words that are after the main verb

The other golden rule of the theory (and with a general acceptation in many other languages) says that linking words (i.e. *however*, *on the other hand*…) have to go between a comma and a punctuation mark. Let us see some correct examples with linking words (in bold):

Bitan esan nizun	;	**hala ere**	,	gustora nago.
(I told you twice)		(**however**)		(I am happy)
	;	LINKING WORD	,	
Etxean geratuko naiz	,	**bestela**	.	
(I am going to stay at home)		(**otherwise**)		
	,	LINKING WORD	.	
Bera	,	**hala ere**	,	ez zen garaiz iritsi.
(He)		(**however**)		(didn't come on time)
	,	LINKING WORD	,	

Figure 7: Linking words have to go between a comma and a punctuation mark

The theory is more complex, but this is essentially what it says. Furthermore, Garzia assures that one or more clauses on a same sentence do not change the theory. He claims that a clause does not have to be necessarily restricted by a comma, just because being a clause. Remember that the *focus*, even if it is a clause, it is not followed by a comma if it is before the verb. Let us see an example of a complex sentence in Basque with more than one clause and its corresponding translation to English, with their punctuation. Note that the *focus* is not followed by any comma, in spite of being a clause.

Euria egin zuen egunean	,	zinera joan baino lehen	deitu nizun	,	etxean zeunden ala ez jakitearren.
(The day that it rained)		(before going to the cinema)	(I called you)		(to know whether you were at home)
clausal TOPIC	,	clausal FOCUS	VERB	,	clausal TOPIC

Figure 8: Example of the punctuation of a sentence with more than one clause

The rules and their implementation

Once the information to create the rules was extracted, the second step was to write and implement them. We chose the Constraint Grammar formalism [Karlsson et al., 1995], since it is the one we also use for parsing. All the rules have already been written, but only some of them have been implemented (see some, in figure 9). We hope to finish this implementation in the next months.

```
MAP (&DEV_PUNCT_1_1) TARGET LINKING_WORD IF (0 LINKING_WORD)
                            (-1 COMMA)
                            (NOT 1 PUNCTUATION_MARK);
MAP (&DEV_PUNCT_1_2) TARGET LINKING_WORD IF (0 LINKING_WORD)
                            (NOT -1 PUNCTUATION_MARK)
                            (1 COMMA);
MAP (&DEV_PUNCT_1_3) TARGET LINKING_WORD IF (0 LINKING_WORD)
                            (NOT -1 COMMA)
                            (NOT 1 COMMA);
```

Figure 9: Three Constraint Grammar rules for detecting incorrect punctuation in linking words

With these concrete rules, we will be able to detect the sentences that do not agree with the principle of linking words. For example, if we have the following sentence,

*Bera hala ere, ez zen garaiz iritsi.
*(He however, didn’t come on time.)

the second rule would be applied, because there is a linking word in the sentence followed by a comma, but it is not preceded by any punctuation mark. This would be the output we would obtain:

*Bera hala ere **&DEV_PUNCT_1_2** , ez zen garaiz iritsi.

This way, we would know that there is one comma left around the linking word. And the tag itself would show us where the comma has to go: in this example, just before the linking word.

Using the same procedure, we want to develop more rules in order to complete the punctuation checker with all the principles explained in the previous section.

Apart from this, we have done several successful tries to integrate some simple rules of style and punctuation —Visual Basic made programs— in the Microsoft grammar API (cgapi.h). For instance, we have integrated the program that detects the lower case letter after a full stop, the one that detects the inexistent space after a punctuation mark, the one that detects a non-closed parenthesis, etc.

Conclusions and future work

This work presents a theory for Basque punctuation, mainly regarding the comma, and the first steps toward a punctuation checker. As the punctuation system is not totally established in hardly any languages, and less in Basque, this task has an additional difficulty. In fact, we have had to contact different experts and contrast their opinions in order to get all the required information.

In the future, we have planned to go on making rules and integrating them in the grammar API of Microsoft and in other text processors as Open Office’s. Moreover, we think that if we manage to correct the punctuation in sentences, we will be able to do a good segmentation of sentences and clauses. Furthermore, if we do a good segmentation of sentences and clauses, we will be able to detect more syntactic errors and facilitate the way to the future syntactic corrector for Basque.

Acknowledgements

We would specially like to thank Juan Garzia. His work has been essential for this research. We would also like to thank Joxe Ramon Etxeberria for his contribution. Finally, thanks to Eli Pociello for helping us writing the final version of the paper.

This research is supported by the University of the Basque Country (9/UPV00141.226-14601/2002) and the Ministry of Industry of the Basque Government (XUXENG project, OD02UN52)).

References

[Aduriz I., Alegria I., Artola X., Ezeiza N., Sarasola K., Urkia M., 1997] **“A spelling corrector for Basque based on morphology”**. *Literary & Linguistic Computing, Vol. 12, No. 1. Oxford University Press. Oxford. 1997.*

[Bayraktar M., Say B., Akman V., 1998] **“An Analysis of English Punctuation: The Special Case of Comma**“. *Intl. J. of Corpus Linguistics, 1998*

[Collins, 1995] **“English Dictionary”**. *HarperCollins Publishers Ltd. 1995. Glasgow. Great Britain.*

[Garzia, J., 1997] **“Joskera lantegi”**. *HAEE-IVAP*

[Hardt D., 2001] **“Comma checking in Danish”**. *In Rayson P., Wilson A., McEnery T., Hardie A. & Khoja S. (eds.) Proceedings of the Corpus Linguistics 2001 Conference. University Centre for Computer Corpus Research on Language. Lancaster University.*

[Karlsson F., Voutilainen A., Heikkila J., Anttila A., 1995] **“Constraint Grammar: Language-independent System for Parsing Unrestricted Text”**. *Mouton de Gruyter, Berlin.*

[Numberg G., 1990] **“The linguistics of punctuation”**. *CSLI Lecture Notes No. 18, University of Chicago Press.*

[Pala K., Rychlý P., Smrz P., 2003] **“Text Corpus with Errors”**. *In Proceedings of the 6th Conference on Text, Speech and Dialogue. 2003. Ceske Budejovice, Czech Republic.*