A modular grammar-helping tool for Basque: work in progress

Izaskun Aldezabal
IXA NLP group
University
of the Basque Country
izaskun.aldezabal@
@ehu.eus

Jose Mari Arriola
IXA NLP group
University
of the Basque Country
josemaria.arriola
@ehu.eus

Ainara Estarrona
IXA NLP group
University
of the Basque Country
ainara.estarrona
@ehu.eus

Abstract

In this article, we explain the first steps towards a grammar-helping tool for Basque from a ruled-based approach. Specifically, we show the first steps carried out for helping with verb agreement, some of the difficulties encountered, which linguistic issues arise when new rules are designed, and future perspectives.

1 Introduction

This article concerns the ongoing work of a Constraint Grammar (vislcg3) (Bick and Didriksen, 2015) based tool for helping with useful information for dealing with verb agreement in sentences. The evaluation report of the Basque Government (Government, 2017) about grammar competence at Primary School includes verb agreement and incorrect use of ergative as grave errors if they occur repeatedly. Based on this fact, the purpose of this work is twofold: a) detecting agreement errors and give help with that kind of grammatical information; b) helping to develop a system to certify the Basque level automatically, a similar approach to Hancke et al. (2012). For the first purpose, we follow similar steps proposed in DanProof (Bick and Didriksen, 2015; Antonsen et al., 2009). Concerning the second goal, the plan is to collaborate with HABE (Institute for Adult Literacy and Basque Learning) which certify Basque levels.

The underlying ideas for both goals are extending grammatical knowledge of the student and helping to certify the language level corresponding to each student. In this paper, we will focus on the detection of some agreement errors.

SAROI (Oronoz et al., 2010) is one of the first tools for detecting syntactic errors used in Basque based on the rule-based approach. Wiechetek (2017) gives an overview of Constraint Grammar-based grammar checkers for many languages.

In the preliminary study presented here, we have started using the information provided by the auxiliary verb in the sentence. In Basque, the auxiliary verb carries information, among others, about the arguments of the verb, including the subject, the object and the indirect object; whether they are first, second or third person, and whether they are singular or plural (Laka, 1996). The auxiliary verb must keep the agreement with such arguments so that the sentence is grammatical. However, it is known that errors that disturb the syntax and semantics of the whole sentence of running texts go beyond the morphological concordance between the auxiliary verb and the mentioned three arguments. For instance, for the verb erosi ('to buy') we find examples like:

(1) *Mikelek tomateak 5 eurogatik erosi ditu*
Mikel-Erg tomatoes-Abs 5 euro-Mot buy have
'Mikel has bought tomatoes for 5 euros'

In (1) where the argument *eurogatik* 'for 5 euros' expressing "asset" with the *-gatik* ('for') motivative case is not considered suitable with the verb *erosi* 'to buy' (surely used incorrectly by the interference of semantic equivalents of the Spanish preposition *por*). To deal with this type of errors, other kinds of linguistic resources are needed, such as verb lexicons containing information regarding valency and semantics of arguments, what we find in the Basque Verb Index (BVI) (Estarrona et al., 2018). Based on the information containing in this lexicon for the verb *erosi* 'to buy', we are able to determine that the third argument expressing "asset" is realized with the inessive case instead of the motivative one.

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0

In this line, Wiechetek (2017) managed to detect valency errors based on a deep syntactic and semantic analysis using Constraint Grammar. For the future, we plan to reuse the BVI lexicon following the same idea.

In the current approach, we have implemented the first module of agreement rules using auxiliary information, and we have studied the frame and argument structure needed for a more global approach.

The paper is organized as follows. Section 2 deals with the adopted methodology and development phase (the corpus, grammar formalism and design principles), Section 3 describes the preliminary evaluation and, Section 4 explains the further steps and future work. Finally, Section 5 will present some conclusions.

2 Methodology

In this section, we present the initial steps of our methodology.

2.1 Compiling available corpora

For the construction of the grammar, we have used a fragment of annotated corpora with agreement error tags (Aldabe et al., 2007). The error corpus for developing the grammar contains 8.368 words and the corpus for testing contains 14.257 words. It is a heterogeneous corpus, containing different types of texts such as abstracts of final degree reports of university students, compositions of Basque learners of intermediate and high level, compositions of students of Basque for special purposes etc.

We have used the 8.368 word sample for developing the grammar and the other 14.257 word sample for testing and controlling false positives. For the later goal, we have also used a sample of EPEC, the Reference Corpus for the Processing of Basque (Aduriz et al., 2006), available in the Ixa group. The sample contains the 10 most frequent verbs in EPEC (covering the 85% of the corpus).

2.2 Analyzing available corpora

As starting point, we use the output of the morphological analyzer (naki Alegria et al., 1996) with all the analyses. We did not use the disambiguation module because it could eliminate correct information that might be needed later to find the error.

2.3 Designing initial grammar

The initial grammar covers maintaining agreement between finite verbs and subjects and objects. In addition, the tool provides possible correct alternatives for repairing those agreement errors. The system uses morphological information, and has a special focus on finite verbs, because we get basic information for checking the verb agreement with subject and object from them. For instance, in (2):

(2) *Diseinu inteligentearen bultzatzaileak beste bide batetik sartu nahi dute kreazionismo*
Design intelligent-Gen the prime movers-Erg another way from-Abl to lead wanted creationism-Abs
'The prime movers of the intelligent design wanted to lead creationism from another way'

Bultzatzaileak 'the prime movers' (with the *–ak* ergative third person singular or absolutive third person plural mark) is grammatically incorrect, because it does not agree with the auxiliary *dute* which demands ergative case, third person and plural.

This mistake is also common in native Basque speakers, specially writing. In these cases, we attach advice tags to finite verbs involved in the agreement error and the words with the incorrect morphological case for the agreement. For instance, we add to the word containing the error *bultzatzaileak* a helping message, such as "take care of the agreement for ergative plural" as shown in the next rule example:

(3) ADD (%Take_care_of_agreement_ERG_PL)
TARGET (ERG) IF (0 ERG-SING) (NOT *1 ERG-PL) (NOT *1 (NR_HAIEK)) (*1 (NK_HAIEK-K) BARRIER (NK-HARK));

The above rule attaches to the singular ergative *bultzatzaileak* the helping message, if there is an auxiliary verb that needs third person plural ergative (NK_HAIEK-K) and there is not an auxiliary verb that demands third person plural absolutive (NR_HAIEK) and the checking is delimited by an auxiliary verb that involves third person singular ergative (NK-HARK).

The current version of the grammar only handles agreement errors of sentences where finite verbs are involved.

3 Preliminary evaluation

The initial grammar rules to find errors describe the conditions for valid structures for sentences where finite verbs are involved, and if these conditions are not accomplished the error tags are added.

In order to evaluate the grammar, as mentioned we have used the hand-annotated corpus (14.257 words). We chose to evaluate the agreement of absolutive and ergative cases. In this section, we give a preliminary evaluation:

- **Annotated errors correctly detected**: for the absolutive case the 50% of the errors are detected correctly. Concerning ergatives, we are able to detect correctly the 28% of the annotated errors. We consider as erroneous annotations when the error tag is assigned to a correct auxiliary verb and to a correct word containing absolutive or ergative case.

Apart from uncorrect annotations there have been detected a big amount of false positives.

From a qualitative point of view the main difficulties encountered by our grammar are the following:

- **False positives**: most of the false positives encountered are due to the ellipsis of the grammatical objects or subjects. In these cases the helping messages are unnecesary because there is not an error. But the messages are just attached to the auxiliary.
- **Complex constructions**: dealing with some subordinating sentences is challenging in the case that the barriers are properly established. We need to improve barriers with a more systematic treatment.
- **Ambiguity of the input**: in the initial approach, we have used the output of the morphological analyzer with all the information, but the preliminary evaluation show us the need of an adaptation of the POS disambiguation module in order to discard verb/noun ambiguity, but maintaining cases.
- **Linguistic issues**: dealing with errors where –ak (absolutive plural / ergative singular) case is involved. For instance in (4):

(4) *Tabernak izugarrizko kutxak egiten dituzte*
Bars-Erg-S/Nom-Pl great takings obtain
'Bars obtain great takings'

This kind of errors would ideally be solved with more complex knowledge. Therefore, in these cases we just can give as advice that the ergative plural is missing according to the auxiliay verb.

- **Bad or incomplete rules**: in some cases we should refine our rules, because we have not taken into account some grammatical possibilities of the language.

In order to improve the ongoing grammar we need more corpora for a more exhaustive analysis.

4 Next steps and future workPreliminary evaluation

Considering the preliminary evaluation and the difficulties encountered, we have in mind the following steps:

- Try to find a solution to the phenomena explained in the previous section.
- Extend these small-scale studies on certain error types to a large-scale analysis of real word student's errors, compiling the learner's corpora for each level.
- Analyze if this kind of agreement errors appear in all levels
- Include the BVI information in the grammar and in the analyzed corpora, and see in which extend could improve the results.

5 Conclusions

This paper has presented a preliminary constraint grammar for helping Basque students with grammatical agreement. The preliminary evaluation indicates the main strategies to improve the results.

The grammar can be in principle reused for other applications that do not necessarily have anything to do with error detection, such as Intelligent Computer-Assisted Language Learning (ICALL) systems.

Acknowledgments

The research leading to these results has been carried out as part of the *DeepReading: Mining, Understanding, and Reasoning with Multilingual Content* project (RTI2018-096846-B-C21 (MCIU/AEI/FEDER, UE)).

References

Itziar Aduriz, Maria Jesus Aranzabe, Jose Maria Arriola, Aitziber Atutxa, Arantza Diaz de Ilarraza, Nerea Ezeiza, Koldo Gojenola, Maite Oronoz, Aitor Soroa, and Ruben Urizar. 2006. Methodology and steps towards the construction of epec, a corpus of written basque tagged at morphological and syntactic levels for the automatic processing. *Corpus Linguistics Around the World. Book series: Language and Computers*, 56:1–15.

Itziar Aldabe, Bertol Arrieta, Arantza Diaz de Ilarraza, Montse Maritxalar, Ianire Niebla, Maite Oronoz, and Larraitz Uria. 2007. Basque error corpora: a framework to classify and store it. In *Proceedings of the 4th Corpus Linguistic Conference*, Birmingham, UK.

Iñaki Alegria, Xabier Artola, and Kepa Sarasola. 1996. Automatic morphological analysis of basque. *Literary and Linguistic Computing*, 11(4):193–203.

Lene Antonsen, Saara Huhmarniemi, and Trond Trosterud. 2009. Constraint grammar in dialogue systems. In *Northern European Association for Language Technology, NEALT*, pages 13–21.

Eckhard Bick and Tino Didriksen. 2015. Cg-3 beyond classical constraint grammar. In *Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015*, pages 31–39, Vilnius, Lithuania.

Ainara Estarrona, Izaskun Aldezabal, and Arantza Diaz de Ilarraza. 2018. https://doi.org/s10579-018-9440-0 How the corpus-based basque verb index lexicon was built. *Language Resources and Evaluation. First Online 05 December 2018*, pages 1–23.

Basque Government. 2017. *Idazmenaren ebaluazioa. Testuak zuzentzeko irizpideak*. ISEI-IVEI, Hezkuntza Saila, Eusko Jaurlaritza.

Julia Hancke, Detmar Meurers, and Sowmya Vajjala. 2012. Readability classification for german using lexical, syntactic, and morphological features. In *Proceedings of the 24th international conference on computational linguistics (COLING)*, pages 1063–1080, Mumbai, India.

Itziar Laka. 1996. *A Brief Grammar of Euskara, the Basque Language*. Office of the Vice-Rector for the Basque Language, UPV/EHU.

Maite Oronoz, Arantza Diaz de Ilarraza, and Koldo Gojenola. 2010. Design and evaluation of an agreement error detection system: Testing the effect of ambiguity, parser and corpus type. *Proceedings of the 7th international conference on Advances in natural language processing (IceTAL 2010)*, 6233:281–292.

Linda Wiechetek. 2017. *When grammar can't be trusted - Valency and semantic categories in North Sami syntactic analysis and error detection*. Ph.D. thesis, UiT The arctic university of Norway.