Automatic Conversion of the Basque Dependency Treebank to Universal Dependencies

Maria Jesus Aranzabe, Aitziber Atutxa, Kepa Bengoetxea, Arantza Diaz de Ilarraza, Iakes Goenaga, Koldo Gojenola, Larraitz Uria
IXA NLP Research Group
University of the Basque Country (UPV/EHU)
E-mail: koldo.gojenola@ehu.eus

Abstract

This work describes the process of automatically converting the Basque Dependency Treebank to Universal Dependencies (UD). Our objective is to develop a set of conversion rules that will automatically transform the original treebank to UD. Basque is a morphologically rich and agglutinative language, which presents different challenges for the conversion from the initial annotation scheme to UD. We will illustrate the steps pursued and the main difficulties we have encountered. As a main conclusion we can say that, although the Basque original treebank was in accord with many UD guidelines, the process was not trivial, converting around 80% of the tokens.

1 Introduction

In this work we describe the conversion of the Basque Dependency Treebank (BDT) to Universal Dependencies (UD) [1, 2, 3, 4]. Although the Basque original treebank was in accord with many UD guidelines, the conversion process presents different challenges. We will try to give a general overview of the process but we will also concentrate on the phenomena where we found some difficulties, specially ellipsis, copulative sentences or multiword units. The Basque language can be described as a morphologically rich, agglutinative language with a high capacity of generating inflected word-forms, with free constituent order of sentence elements. It can be considered a head-final language, as the syntactic head of phrases is located at the end of the last word of the phrase, in the form of a suffix. BDT [5] is a pure dependency treebank from its original design, annotated in the CoNLL-X format, and it shares with UD a lexicalist hypothesis in syntax, where dependencies occur between whole individual wordforms. Under this lexicalist approach, each word shows several morphosyntactic associated features, corresponding to affixes (prefixes and suffixes) attached to the base forms, such as case (there are 14 morphological cases in Basque), number, definiteness or type of subordinate sentence

Table 1: Mapping between BDT and UD for POS tags and dependency relations

Type of mapping (BDT → UD)	POS tags	Dependencies
1:1	ADJ, ADB, ...(13 categories)	15 dependencies
1:2	Det → DET/NUM Noun → DET/NUM/PROPN	cmod → advcl/acl detmod → det/nummod
1:5		ncmod → advmod/amod/det/nmod/neg

(adversative, conditional, ...). These suffixes usually appear as separated word-forms in non agglutinative languages. The last version of BDT contains 150,000 tokens forming 11,225 sentences, with 1.3% of non-projective arcs. BDT encodes 16 different POS and 28 dependencies, an extended inventory based on [6].

2 Description of the Automatic Conversion Process

UD covers three levels of annotation: part of speech (POS) [3], morphosyntactic features [4] and dependency labels [1]. The first step of the conversion process consisted of analyzing BDT and UD[1] guidelines in order to find the correct mapping of each Basque tag or dependency label. Mapping POS and morphosyntactic features was a quite straightforward step, described in subsection 2.1. Regarding the conversion of dependencies, there are several phenomena that are worth mentioning, which are presented in the following subsections.

2.1 Conversion of POS and morphosyntactic features

Table 1 presents the main differences between the set of POS tags used in BDT and those in UD. The table shows, in its second column, that several of the BDT POS tags have a unique correspondence in UD. However, there are different cases where the mapping is not direct, because a part of speech tag must be mapped to several UD POS tags, depending on other aspects, such as morphological features. This happens with determiners and nouns. On the other hand, there are cases when two different BDT tags are mapped to the same UD POS tag, as in the case of main verbs, which in UD have a unique category (VERB), while there are two tags for Basque main verbs, depending on whether the verb must be accompanied by an auxiliary or it is a *compact* verb where the main verb contains inflectional suffixes corresponding to the auxiliary. This distinction is missed in the UD POS tag, although it can be recovered from the morphosyntactic tags.

Regarding the set of morphosyntactic features, it can be considered the easiest step, as the inventory of UD features was compiled over a big set of dependency treebanks and annotation guidelines [4]. The main differences can be related to differences of specificity, either from BDT or UD, where one of the descriptions gives a more ample set of values for a given category (e.g., the UD guidelines present a wider spectrum of values for numerals, compared to BDT).

[1] http://universaldependencies.github.io/docs/

2.2 Conversion of Dependencies

Table 1 shows in its third column that, although most of the dependencies are mapped in a straightforward manner, some other are more complex, as in the case of the non-clausal modifier (ncmod) relation in BDT, which is mapped to 5 different relations in UD. Apart from this fact, there have been some other aspects that are presented in the following paragraphs.

Morphological ellipsis

Basque allows the formation of ellipsis inside a wordform, by means of a subordinated relative clause or a genitive, as in

dakarrena (the one that (he) brings) = *dakarren* (that brings (he)) + *-a* (the one)

This wordform presents an example of a relative clause that, when combined with a definite article, forms an ellipsis. As the wordform must be assigned a unique part of speech, it could correspond to either a verb from its original root or a noun, taking its function into account (the whole word acts as an object). Figure 1 shows an example of a sentence that illustrates this phenomenon. The figure shows that this word depends on the main verb by means of a *dobj* relation, which seems contradictory since the word is marked as a verb. Figure 2 shows a sentence parallel to that of Figure 1, but without the ellipsis. In the example, the wordform *Gizonak* (the man) acts as a subject of the subordinated verb, which in turn modifies *gauza* (the thing) by a relative clause (relcl) dependency, and this will be the direct object of the main verb.

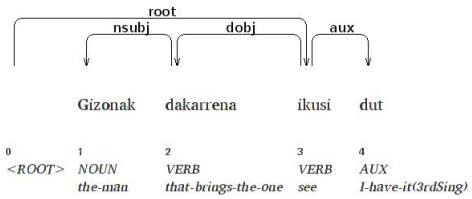

Figure 1: Example of an elliptical relative sentence inside a nominal wordform (*I have seen the one that the man brings*).

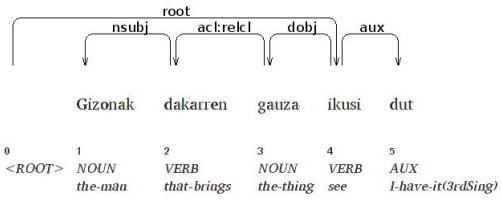

Figure 2: Example of a non-elliptical sentence parallel to the one in Figure 1 (*I have seen the thing that the man brings*).

Universal Dependency annotation follows a lexicalist view of syntax, which

means that dependency relations hold between *words* as in figure 1. Under this view the parallelism that should hold between figure 1 and figure 2 disappears. Universal Dependencies allow some exceptions to the lexicalist view such as Spanish clitics. Up to present, there is agreement on the fact that the lexicalist view should be followed avoiding splitting as much as possible. Figure 3 presents a possible solution to the problem showed in Figure 1, by separating the verbal and nominal information inside the wordform *dakarrena*. This way, the analysis in the figure is symmetric to that in figure 2, and the verb/noun dichotomy present in figure 1 is solved. Although Basque presents a high rate of morphological ambiguity, we think that the splitting could be done automatically.

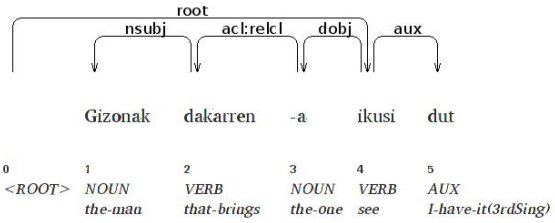

Figure 3: Alternative analysis of the sentence in Figure 1

Multiwords (MWs)

The BDT guidelines allow to agglutinate several wordforms in MWs. Although there are many different combinations creating multiwords, we only transformed the most frequent combinations of POS and CPOS (coarse POS) tags, accounting for 2/3 of the total number of MWs, and leaving the rest for future work. The transformation consists of recovering the original wordforms with their corresponding POS, CPOS and features, assigning at the same time the dependency. An aspect that deserved a careful study was to determine the head and the dependent(s) of each MW. This was easy for compounds, but more difficult with NEs and complex postpositions, as in some of them the words can be inflected, giving different options for choosing the head and dependent. There are three types of MWs in BDT:

- Compounds.
- Named entities (NE), including person, location, organization and undefined (for other types of NEs). These MWs present different patterns for the conversion, as there are a variety of types of elements, such as nouns, adjectives, adverbs, and numerals.
- Complex postpositions like *mendiaren gainean*:

  mendiaren gainean (on top of the mountain) = *mendiaren* (of the mountain) + *gainean* (on top)

  In this example, both wordforms are inflected with the genitive case and the innessive case, respectively. Although at first sight it could be stated that

mountain could be the head of the MW unit, the genitive acts as a complement and suggests that *top* is the head.

Coordination

There are several ways of coding coordinated structures, depending on the head of the coordination structure. In BDT the conjunction is the head, while in UD the first argument of the conjuctions acts as the head of the whole structure. Allowing the conjunction to be the head of the coordination as in BDT can better represent certain scope phenomena and ellipsis occurring through coordination, because the UD specification for coordination, attaching all the elements to the first conjunct, loses some scope information present in the original BDT such as, for example, in figure 4, when a modifier is a dependent of the whole coordinated sequence.

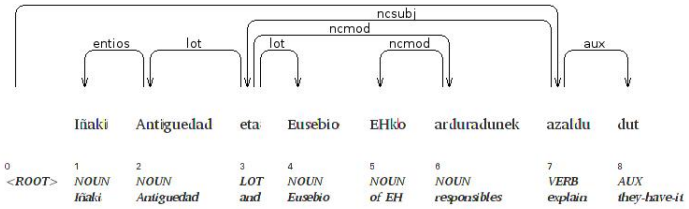

Figure 4: Analysis in BDT where the conjunction *eta* is the head and the scope of the modifier (*the responsibles of EH* linked by a *ncmod* dependency relation) applies over the whole coordination structure (*Iñaki Antiguedad and Eusebio the responsibles of EH have explained (it)*).

In addition, allowing the conjunction to be the head of the coordination favours representing coordinative ellipsis as, for example in figure 5, where two sentences are linked by a coordination conjunction (*eta*), and the second sentence does not contain a main verb (ellipsis). As shown in figure 6, the parallelism occurring in coordinate ellipsis did not get captured after the converstion to UD. One way of solving it could be to add some especificity over the *conj* relation for capturing the symmetry, as presented in figure 7. No decision has been taken in the UD community so far, and coordinate ellipsis remains problematic. In fact, figure 6 is the actual conversion for the original BDT sentence (see figure 5).

Copulative sentences

Although the UD guidelines only allow copulative sentences using *be* as the copula (this restriction is an open issue in the UD community), in Basque several verbs can take part in these sentences, and they need additional analysis. Usually there is agreement between the copulative modifier and the subject, whereas with predicative verbs the modifier is adverbial and does not show agreement.

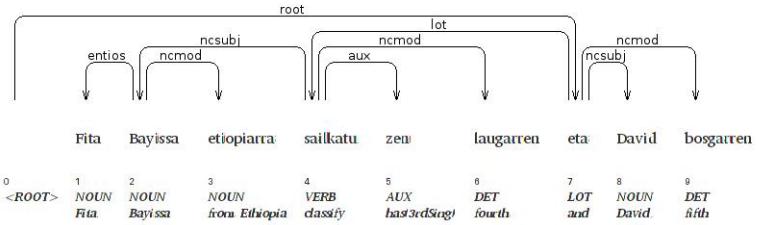

Figure 5: Analysis in the BDT where the conjunction is the head *and* and acts as a place holder for ellipsis (*Fita Bayissa from Ethiopia has classified fourth and David fifth*).

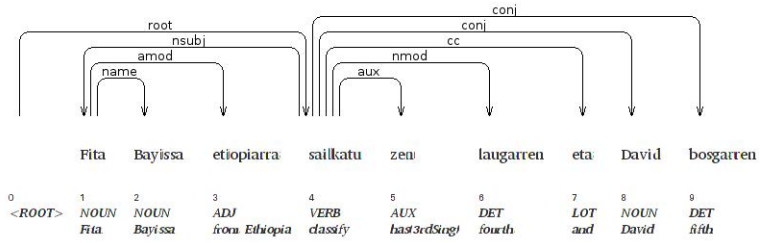

Figure 6: Analysis after the UD conversion where the first conjunct is the head of the coordination.

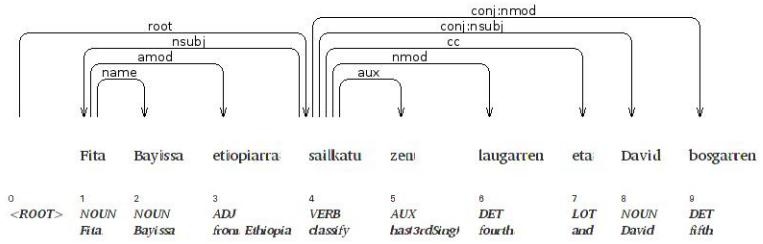

Figure 7: Alternative analysis after the UD conversion where the first conjunct is the head of the coordination.

3 Results

The above presented criteria were transformed in a set of scripts for the automatic conversion from BDT to UD. The order of transformations is not trivial, since changing a part of the treebank can have consequences on subsequent conversions. For example, converting some dependencies needs an examination of the original BDT tags, and for this reason we had to maintain both the original tags together with the UD tags. Generally, more abstract conversions should be applied first, such as the transformation of coordinated sentences, because changing lower level constructions could give erroneous results.

For each phenomena mentioned in subsection 2.2 we first performed a quantitative and qualitative study, and oriented our study towards the design of a set of rules dealing with the most frequent patterns, giving priority to coverage, but without compromising precision, that is, we did not convert any instance not covered in the patterns. This process will leave out a subset of sentences of each phenomena, which are left as future work. A potential side effect will be that some low-frequency phenomena will not be covered by the UD treebank.

As a result of the previously described conversion steps, we have obtained a UD based Basque treebank containing 121,000 tokens, which represents around 80% of the sentences in the BDT. On one hand, this can be seen as a succesful accomplishment, since the conversion rules were designed taking a conservative approach, with the aim of achieving high precision and not leaving any room for conversion errors. On the other hand, the set of remaining sentences correspond to either special cases not accounted by the conversion rules or other types of less frequent phenomena which have not been dealt with at the moment.

4 Conclusion

Although the annotation of the Basque Dependency Treebank (BDT) is in accord with most of the UD guidelines (for example, taking content words as heads), the conversion has been a complex task, from the relatively direct mappings of POS tags to more complex phenomena like ellipsis, copulative sentences or multiwords. At the moment, a set of sentences (120,000 tokens) has been successfully converted, but there are some issues that need to be addressed in order to convert the remaining part of BDT.

Overall, we can state that, except for several phenomena where we have found some difficulty, the automatic conversion process is feasible. We can also say that some of the problematic issues are shared in several cases with typologically similar languages like Finnish or Turkish, and in this respect they can serve to adapt the UD guidelines in order to generalize over the whole set of languages involved.

Acknowledgments

This research was supported by the the Basque Government (IT344-10, DETEAMI), the University of the Basque Country (INF13/59) and the Spanish Ministry of Science and Innovation (MINECO TIN2013-46616-C2-1-R).

References

[1] de Marneffe, Marie-Catherine, Dozat, Timothy, Silveira, Natalia, Haverinen, Katri, Ginter, Filip, Nivre, Joakim, Manning, Christopher D. (2014) *Universal Stanford dependencies: A cross-linguistic typology. Proceedings of the Language Resources and Evaluation Conference (LREC14)*. Reykjavik, Iceland

[2] Mcdonald, Ryan, Nivre, Joakim, Quirmbach-brundage, Yvonne, Goldberg, Yoav, Das, Dipanjan, Ganchev, Kuzman, Hall, Keith, Petrov, Slav, Zhang, Hao, Täckström, Oscar, Bedini, Claudia, Bertomeu Castelló, Núria, Lee, Jungmee (2013) *Universal Dependency Annotation for Multilingual Parsing*, In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp. 92–97, Sofia, Bulgaria

[3] Petrov, Slav, Das, Dipanjan, McDonald, Ryan (2012) *A Universal Part-of-Speech Tagset* In Calzolari, Nicoletta, Choukri, Khalid, Declerck, Thierry, Uğur Doğan, Mehmet, Maegaard, Bente, Mariani, Joseph, Moreno, Asuncion, Odijk, Jan, Piperidis, Stelios (eds.) European Language Resources Association (ELRA) *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC12).* Istanbul. Turkey

[4] Zeman, Daniel (2008) *Reusable Tagset Conversion Using Tagset Drivers. Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC08)*, pp. 28–30, Marrakech, Morocco

[5] Aduriz, Itziar, Aranzabe, Maria Jesus, Arriola, Jose Mari, Atutxa, Aitziber, Diaz de Ilarraza, Arantza, Garmendia, Aitzpea, Oronoz, Maite (2003). *Construction of a Basque Dependency Treebank*. Treebanks and Linguistic Theories, pp. 201-204. Vaxjo, Sweden.

[6] Carroll, John, Briscoe, Ted, Sanfilippo, Antonio (1998). *Parser Evaluation: a Survey and a New Proposal*, *Proceedings of the First International Conference on Language Resources and Evaluation (LREC98).*, pp.474-454. Granada, Spain

[7] Mel'čuk, Igor (1988) *Dependency Syntax: Theory and Practice*. State University of New York Press.

[8] Zeman, Daniel, Mareček, David, Popel, Martin, Ramasamy, Loganathan, Štěpánek, Jan, Žabokrtský, Zdeněk, Hajič, Jan (2012) *HamleDT: To Parse or Not to Parse?* In Calzolari, Nicoletta, Choukri, Khalid, Declerck, Thierry, Uğur Doğan, Mehmet, Maegaard, Bente, Mariani, Joseph, Moreno, Asuncion, Odijk, Jan, Piperidis, Stelios (eds.) *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC12)*, pp. 2735-2741, Istanbul, Turkey