Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 125-132 recibido 01-12-2016 revisado 12-01-2017 aceptado 09-02-2017

Building the Gold Standard for the Surface Syntax of Basque

Construcción de un Gold Standard para la Sintaxis Superficial del Euskera

Itziar Aduriz[1], María Jesús Aranzabe[2], José María Arriola[2]*
Arantza Díaz de Ilarraza[2], Itziar Gonzalez-Dios[2], Ruben Urizar[2]
IXA Research Group
[1]University of Barcelona
[2]Universtity of the Basque Country UPV/EHU
* josemaria.arriola@ehu.eus

Resumen: En este artículo presentamos el proceso de construcción de SF-EPEC, un corpus de 300.000 palabras, sintácticamente anotado, que pretende ser un Gold Standard para el pocesamiento sintáctico superficial del euskera. En primer lugar, describimos el conjunto de etiquetas diseñado para este propósito; siendo el euskera una lengua aglutinante, en ocasiones hemos tenido que crear etiquetas sintácticas compuestas. Asimismo, se detallan las distintas fases en la construcción de SF-EPEC.
Palabras clave: Sintaxis superficial, *gold standard*, euskera, anotación de corpus

Abstract: In this paper, we present the process in the construction of SF-EPEC, a 300,000-word corpus syntactically annotated that aims to be a Gold Standard for the surface syntactic processing of Basque. First, the tagset designed for this purpose is described; being Basque an agglutinative language, sometimes complex syntactic tags were needed. We also account for the different phases in the construction of SF-EPEC.
Keywords: Surface syntax, gold standard, Basque, corpus annotation

1 Introduction

Corpora are essential resources in linguistics research. As stated by Sampson (2011), the use of corpora in language research allows a better understanding of language complexity particularly on syntactic issues.

The development of data-driven language processors requires large amounts of texts manually tagged at different levels, which are called gold standard corpora. These are also used to evaluate the output of rule-based processors comparing their results with the gold standard annotation.

Important efforts have been devoted to the construction of syntactically annotated gold standards in several languages such as English (Marcus, Marcinkiewicz, and Santorini, 1993; Silveira et al., 2014), Spanish (Mille et al., 2009), German (Scheible et al., 2011), Norwegian (Solberg et al., 2014), Swedish (Nilsson and Hall, 2005), or Finnish (Voutilainen, Purtonen, and Muhonen, 2012).

Similarly, our effort was led to annotate syntactically the Reference Corpus for the Processing of Basque EPEC (Aduriz et al., 2006a). This syntactically annotated corpus, hereafter SF-EPEC, aims to be a Gold Standard for the development and evaluation of shallow syntactic analyzers for Basque. Specifically, SF-EPEC has as an immediate goal the evaluation of SF-Grammar, a rule-based surface syntactic analyzer for Basque (Arriola, 2015).

Previously, Aduriz and Díaz de Ilarraza (2013) established the theoretical and practical issues for the shallow syntactic annotation in Basque. The annotation process of SF-EPEC was largely inspired in Voutilainen, Purtonen, and Muhonen (2012). The authors specify different steps for the process of corpora annotation, which include tasks such us (i) specifying a tentative annotation model and guidelines; (ii) applying the model to a large sample of example sentences and if necessary refining the model and the guidelines; or (iii) evaluating the applicability by means of the double-blind annotation routine.

Likewise, the methodology for the annotation of SF-EPEC comprised the following

ISSN 1135-5948

steps:

1. A random sample of full sentences –consisting of 3% of the corpus–was extracted for it to be manually annotated.
2. During the annotation of this sample, a discussion phase took place so as to decide how to annotate some specific phenomena. An annotation guideline was drawn up with the decisions taken.
3. Then, taking into account the redefined tagset and the annotation guidelines, three different coders annotated a sample corpus of about 11,500 ambiguous tokens in parallel, and the inter-annotator agreement was measured.
4. Finally, the whole corpus was annotated by two linguists.

After introducing our strategy for building the Gold Standard for surface syntax and related work, Section 2 explains the basic resources for this syntactic annotation. In Section 3, we describe the tagset designed to annotate syntactic functions. Section 4 is devoted to the manual annotation, i.e. the discussion phase and inter-annotator agreement. Finally, some conclusions are presented in Section 5.

2 *Framework for the annotation*

The IXA research group[1] is working on a robust parsing scheme that provides syntactic annotation in an incremental fashion (see Figure 1).

The information contained in the lexical database for Basque EDBL (Aldezabal et al., 2001) constitutes the basis for our analyzers. It consists of 121,823 entries divided into (i) dictionary entries, (ii) inflected verb forms, and (iii) dependent morphemes, all of them with their respective morphological information.

In the morphosyntactic analysis, first a tokenizer divides the text into a sequence of tokens. Then, the robust morphological analyzer MORFEUS (Alegria et al., 1996) gives to each word form every possible analysis, without taking into account the context in which it appears; that way each word form of the whole corpus is assigned its corresponding analysis at the segmentation level.

[1] http://ixa.eus

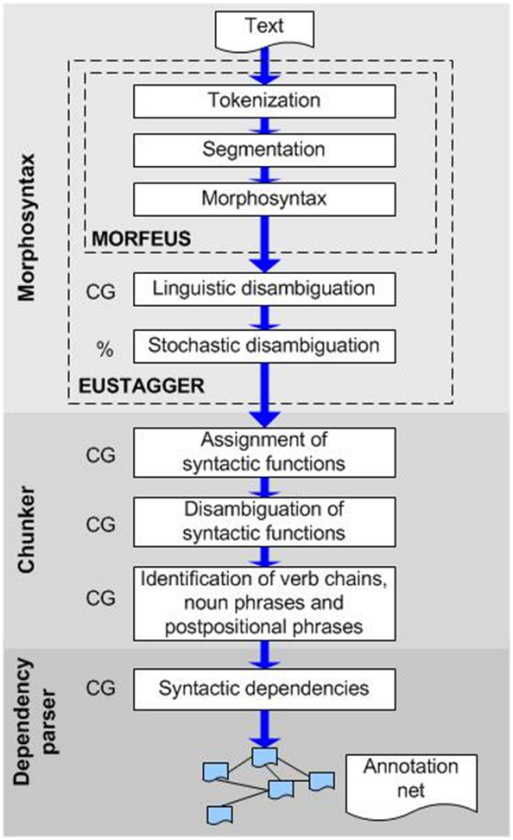

Figure 1: General framework

```
"<$.>" <PUNT_PUNT>"
"<Zalantzak>"
  "zalantza" IZE ARR DEK ABS NUMP MUGM
  "zalantza" IZE ARR DEK ERG NUMS MUGM
  "zalantza" IZE ARR DEK ERG MG
"<argitu>"
  "argitu" ADI SIN AMM PART ASP BURU
  "argitu" ADI SIN AMM PART
"<zituzten>"
  "*edun" ADL B1 NR_HK NK_HK ERL MEN ERLT
  "*edun" ADL B1 NR_HK NK_HK ERL MEN ZHG
  "*edun" ADL B1 NR_HK NK_HK
  "ukan" ADT B1 NR_HK NK_HK ERL MEN ERLT
  "ukan" ADT B1 NR_HK NK_HK ERL MEN ZHG
  "ukan" ADT B1 NR_HK NK_HK
```

Figure 2: Morphological analysis of the sentence *Zalantzak argitu zituzten* 'They clarified the doubts'

Figure 2 shows the analysis provided by MORFEUS for the sentence *Zalantzak argitu zituzten* 'They clarified the doubts' expressed in a Constraint Grammar (CG) style, in which every word form is associated with one or more reading lines. Each line corresponds to a possible interpretation, which provides the word form's lemma, part-of-speech, number, case markers, definiteness and other

morphological information.

Then, the lemmatizer-tagger EUSTAGGER (Aduriz et al., 2003) performs the automatic disambiguation at two levels: first, a rule-based disambiguation is carried out and then the stochastic disambiguation is applied (see Figure 3).

```
"<Zalantzak>"
  "zalantza" IZE ARR DEK ABS NUMP MUGM
"<argitu>"
  "argitu" ADI SIN AMM PART ASP BURU
"<zituzten>"
  "*edun" ADL B1 NR_HK NK_HK
```

Figure 3: Morphological disambiguation

After performing the morphological disambiguation, the next step is to assign the corresponding syntactic tag to each word form. Typically, inflectional suffixes and syntactic functions are closely related in Basque, and therefore most suffixes in the lexical database are assigned their corresponding syntactical function(s) (see Section 3.1). As a result, the output of the morphological analyzer displays these syntactic tags. The syntactic tags at this level refer to shallow syntactic functions. The symbol `@` precedes the abbreviation for the syntactic function. For example, the tags `@OBJ`, `@SUBJ` or `@PRED` stand for object, subject and predicative respectively (Figure 4).

```
"<Zalantzak>"
  "zalantza" IZE ARR DEK ABS NUMP MUGM @OBJ
  "zalantza" IZE ARR DEK ABS NUMP MUGM @SUBJ
  "zalantza" IZE ARR DEK ABS NUMP MUGM @PRED
"<argitu>"
  "argitu" ADI SIN AMM PART ASP BURU
"<zituzten>"
  "*edun" ADL [...] ERL MEN ERLT @+JADLAG_IZLG>
  "*edun" ADL [...] ERL MEN ZHG @+JADLAG_MP_OBJ
  "*edun" ADL B1 NR_HK NK_HK
```

Figure 4: Syntactic tags

However, some word forms lack any suffix or have a suffix with no specify syntactic function as in the past participle *argitu* 'clarified' in the sentence in Figure 4. Those word forms that are not given a syntactic tag by the morphological analyzer are assigned one to their analysis through CG mapping rules (Aduriz and Díaz de Ilarraza, 2013). Similarly, disambiguation is carried out in the case of word forms having more than one possible syntactic function e.g. *zalantzak* and *zituzten* in the sentence in Figure 4. This is also done through a CG grammar (Aduriz, 2000; Arriola, 2015). In the final output, each word form in the sentence keeps a single morphological analysis and a single syntactic tag as shown in Figure 5.

```
"<Zalantzak>"
  "zalantza" IZE ARR DEK ABS NUMP MUGM @OBJ
"<argitu>"
  "argitu" ADI SIN AMM PART ASP BURU @-JADNAG
"<zituzten>"
  "*edun" ADL B1 NR_HK NK_HK @+JADLAG
```

Figure 5: Syntactic disambiguation

Also, a chunk parser provides a partial constituent analysis (Aduriz et al., 2006b) and finally a dependency parser establishes the dependency links (Aranzabe and Díaz de Ilarraza, 2009).

SF-EPEC Gold Standard is aimed to be an essential resource for the evaluation and consequent improvement of the CG grammars that allocate syntactic tags to the word forms in a text.

3 Syntactic tagset

Following the CG formalism, the annotation of syntactic functions in CG is based on the word, understood as the content between two blanks. With this in mind, the main feature of the annotation is that all words need to be provided with a syntactic label (Karlsson et al., 1995). An obvious consequence of this requirement of the CG parser was that, apart from the traditional syntactic functions, specific labels needed to be created for words which in principle do not have 'traditional' syntactic information, such as elements of some multiword expressions.

Being Basque an agglutinative postpositional language, often the syntactic function of a word is given by the suffix attached to it such as a case marker (see Section 3.1a). In (1), the ergative case added to the stem *etxe* 'house' assigns the subject function to the word (*etxeek*).

(1) *etxe-ek*
house-the.PL.ERG
'the houses' (SBJ)

Moreover, subordinating morphemes can be added to finite or non-finite verb forms as well as to main or auxiliary verbs in such a way that each subsequent morpheme gives a piece of the syntactic information. Complex syntactic tags are used for this purpose (see Section 3.1c). For instance, the suffix *-takoan* ('once')

added to the past participle form of a verb allocates the word a complex tag indicating "non-finite verb, subordinate clause functioning as verb complement" as *bukatutakoan* in (2).

(2) *buka-tu-takoan*
finish-ed-once
'once finished' (NFIN SUBR ADV)

Besides, some independent function words—e.g. coordinators (3) or sentence connectors—hold a syntactic function which is inherent to the parts of speech they belong to (see Section 3.1b).

(3) edo
or
'or' (CONJ)

However, not all the lexical words in a sentence are inflected in Basque. For example, it is typically the last element in the noun or postpositional phrase that takes the case marker. In (4), the demonstrative in the final position of the PP takes the inessive case, but the rest of the lexical words (*igande* 'Sunday', *euritsu* 'rainy', *ilun* 'dark') are devoid of a case marker.

(4) *igande euritsu eta ilun hartan*
Sunday rainy and dark that.INE
'in that rainy and dark Sunday'

Therefore, the words lacking a case marker are added a function tag through mapping rules. Some of these syntactic tags are the same as the ones designed for the database, but others are new. In particular 23 specific labels needed to be created for words which in principle do not have 'traditional' syntactic information e.g. elements of some multiword expressions (see Section 3.2). For instance, for the multiword sentence connector *hala eta guztiz ere* 'despite everything' (see (5)), the words *hala, eta*, and *guztiz* are allocated the tag `@HAOS>` denoting they are just components of multiword expression, while the last element *ere* is assigned the function tag for sentence connector `@LOK`.

(5) *hala eta guzti-z ere*
like.that and all-INS too
'despite everything'

Furthermore, some additional tags had to be created during the manual annotation process for specific cases (see Section 4.2).

Bearing all this in mind, we have divided the syntactic tagset developed for the labeling of the Basque corpus in three groups, depending on the step in which they are applied:

- Syntactic tags derived from the lexical database (explained in Section 3.1).
- Tags allocated through mapping rules during the assignment of syntactic functions (explained in Section 3.2).
- Tags created during the manual annotation process for specific cases (Section 4.2).

3.1 Tags from the lexical database

The analyses produced by the morphosyntatic analyzer for Basque MORFEUS are accomplished based on the information included in the lexical database for Basque EDBL. Each entry in EDBL is kept along with its morphosyntactic information. 19 different syntactic tags are used in the lexical database. The following entries holding a syntactic tag:

a) Case markers. As said before, often the syntactic function of a word in a Basque sentence is given by a suffix attached to it such as a case marker. In Table 1 we present some examples of suffixes and their assigned syntactic function.

Function	Meaning	Suffix holding the function
`@SUBJ`	Subject	Ergative and absolutive
`@OBJ`	Direct object	Absolutive
`@ZOBJ`	Indirect object	Dative
`@ADLG`	Verb complement	Locative, directional, origin, comitative, instrumental, cause, goal...
`@PRED`	Predicative	Absolutive
`@IZLG>`	Left noun complement	Genitive locative and genitive

Table 1: Syntactic functions associated to case markers

(6) *Gutun-ek egi-a esan zuten.*
letters-ERG truth-ABS say.PFV AUX.3PL.PST
`@SUBJ` `@OBJ` `@-JADNAG` `@+JADLAG`
SBJ DO NF.VB FIN.AUX
'The letters told the truth.'

In (6), we show which the function tags would be for each word in the sentence *Gutunek egia esaten zuten* 'The letters told the truth'.

b) Some function words. Sentence connectors (*halere*, 'however'), independent

subordinators (*arren*[2], 'although'), and coordinators (*eta*, 'and') hold a syntactic function which is inherent to the part of speech they belong to (Table 2).

Function	Meaning
@LOK	Sentence connector
@PJ	Coordinator
@MP	Independent subordinator

Table 2: Syntactic tags corresponding to function words

We show an example of coordination in (7).

(7) *Peio* *eta* *Iñaki* *hemen* *dira.*
Peio.ABS and Iñaki.ABS here are
@SUBJ @PJ @SUBJ @ADLG @+JADNAG
SBJ CONJ SBJ ADV NF.VB
'Peio and Iñaki are here.'

c) Dependant subordinators (MP) are suffixes which can be added to finite auxiliary verbs (@+JADLAG), non-finite main verbs (@-JADNAG) or finite synthetic verbs (@+JADNAG). The syntactic function that the subordinator assigns to the subordinate clause (verb complement, noun complement, object...) is added to the previous verb-type tag, thus making up a complex tag with the combination of the three elements. For instance, Table 3 shows some complex tags for finite main verbs.

Function	Meaning
@+JADNAG_MP_ADLG	Finite main verb, subordinate clause functioning as a verb complement
@+JADNAG_MP_SUBJ	Finite main verb, subordinate clause functioning as a subject
@+JADNAG_MP_OBJ	Finite main verb, subordinate clause functioning as a direct object

Table 3: Dependant subordinator

In (8), the word *bukatutakoan* 'when finished' holds the complex tag @-JADNAG_MP_ADLG, which stands for "non-finite main verb, subordinate clause functioning as verb complement".

(8) *Buka-tu-takoan* *joan-go* *gara.*
finish-ed-when go.FUT 1PL.PRS
@-JADNAG_MP_ADLG @-JADNAG @+JADLAG
NF.SUBR.ADV NF.VB FIN.AUX
'When finished, we will leave.'

3.2 Tags added in assignment phase

The word forms that are assigned no syntactic tag by the morphological analyzer MORFEUS are allocated one through CG mapping rules (Section 2). Some of the syntactic tags added by this grammar are the same designed for the database (see Section 3.1), but others were created for this stage. In Table 4, we can find some examples of new syntactic tags added in the assignment phase.

Function	Meaning
@KM>	Modifier of the word containing the case marker
@<IA	Postmodifier
@IA>	Premodifier
@<ID	Right determiner
@ID>	Left determiner
<@GRAD	Right grader
@GRAD>	Left grader
@ADILOK>	First element of compound verb
<@ADILOK	Last element of compound verb
@HAOS>	Element of a multiword expression

Table 4: Examples of new syntactic tags added in the assignment phase

Unlike the tags derived from the lexical database, the tags added in the assignment phase need syntactic context to be assigned, and that is why they are attached on the outcome of MORFEUS.

Some words are allocated tags which have no conventional syntactic information. For example, as stated before, it is the last element in the noun or postpositional phrase that takes the case mark in Basque (see (4)). The tag @KM> is assigned to all the nouns lacking a case mark in the phrase, as in *igande* 'Sunday' in (9). Also, the tags @<IA or @IA> are added to noun postmodifiers and premodifiers respectively; for instance, the adjectives *euritsu* 'rainy' and *ilun* 'dark' in (9) get the tag @<IA for postmodifier.

(9) *igande* *euritsu* *eta* *ilun* *hartan*
Sunday.ø rainy.ø and dark.ø that.INE
@KM> @<IA @PJ @<IA @ADLG
HEAD POSTMOD CONJ POSTMOD ADV
'in that rainy Sunday'

Also, some components of compound verbs—such as *min* 'pain' in *min egin*, 'to

[2]Almost all subordinators in Basque are morphemes attached to either finite or non-finite verb forms (see Section 3.1b). Just a few, such as the adversative conjunction *arren* 'despite', are written separately.

hurt' lit. 'do-harm'—are assigned new tags (@ADILOK> or @<ADILOK) since the morphosyntactic information of the compound is usually given by one of the elements. Also, components of other multiword expressions (e.g. *ziur aski*, 'most probably') are added the tag @HAOS> indicating the following element carries the syntactic tag corresponding to the whole expression (see (10)).

(10) *Ziur* *aski* *min* *egin-go* *di-zu.*
sure very harm do-FUT 3SG.SBJ-2SG.IO
@HAOS> @ADLG @ADILOK> @-JADNAG @+JADLAG
> ADV > NF.VB FIN.AUX
'Most probably s/he will hurt you.'

4 Manual annotation

In order to build up the Gold Standard for syntactic functions SF-EPEC we used EPEC, the Reference Corpus for the Processing of Basque (Aduriz et al., 2006a). EPEC is a 300,000-word collection of texts written in standard Basque, which is intended to be a reference corpus for the development and improvement of several NLP tools for Basque. Although small, it is strategic for a less-resourced language like Basque.

EPEC was first morphologically analyzed by means of MORFEUS and then manually disambiguated (Aldezabal et al., 2007). The process of the annotation of the syntactic functions in SF-EPEC consisted in either selecting the correct syntactic function from the different ones provided by the morphological analyzer MORFEUS or adding the correct syntactic tag whenever MORFEUS provided no function or none of the ones provided was correct.

For example, the absolutive case may function either as a subject in intransitive sentences, as an object, or as a predicate. Thus, the word form *zalantzak* in Figure 4 is allocated three syntactic tags. However, in the specific context of *zalantzak* in (11), the correct syntactic function for the annotator to choose would be 'direct object' (absolutive plural).

(11) *Zalantz-ak* *argi-tu* *zituzten.*
doubt-ABS.PL clarify-PTCP 3PL.SUJ·3PL.DO.PST
@OBJ @-JADNAG @+JADLAG
DO NF.VB FIN.AUX
'They clarified the doubts.'

The manual annotation took place in three different stages: the discussion phase, the inter-annotator agreement phase and the annotation of the whole corpora.

4.1 Discussion phase

In order to define the tagset and the criteria for the annotation, a random sample of full sentences—comprising 3% of the corpus—was extracted for it to be manually annotated. The annotation was carried out by a linguist, and the doubts arising during the process were discussed by two more linguists with experience in NLP annotation tasks. Decisions were taken so as to decide how to annotate some specific phenomena. As a result, an annotation guideline was drawn up with the decisions taken in the discussion phase (Aduriz et al., 2015) .

Many of the decisions taken in this stage involved the use of tags previously defined. For instance, we found out that for some tokens the syntactic tags provided by the analyzer did not correspond to their real functions in some specific contexts. In order to solve this problem, the tags were manually added. Sometimes existing tags were added to the tokens, for example, in some multiword expressions such as complex postpositions (*-ren aurrean*, 'in front of') or complex subordinators (*-n arte*, 'until').

(12) *etxe-a-ren* *aurre-a-n*
house-the-GEN front-the-INE
'in front of the house'

In (12) (*etxearen aurrean*, 'in front of the house') the morphological analyzer allocates the noun complement function (@IZLG>) to the first word containing the genitive case (see Section 3.1a). However, in the example above the genitive is part of a complex postposition (*-ren aurrean*, 'in front of') so in the manual annotation the @KM> tag was added to the token containing the genitive case, indicating that it is the following token—*aurrean* 'in front' containing the inessive case marker—that allocates the syntactic function corresponding to the complex postposition.

Similarly, in complex subordinators such as *-n arte* 'until' (see (13)), the subordinator *-n* attached to the finite auxiliary verb is automatically assigned the subordinator function. However, in the manual annotation the @KM> tag was added to this token, indicating that it is the following word that holds the syntactic function corresponding to the whole complex subordinator.

(13) *etorr-i* *de-n* *arte*
come-PTCP has-SUBR until

'until s/he has come'

4.2 New tags

For cases that had not been foreseen in the initial annotation scheme, two new tags were created: `@IS` (isolated noun phrase) and `@FSG` (no syntactic function).

Most of the cases in which a new tag was needed corresponded to phrases belonging to verbless incomplete structures in which it was impossible to determine the syntactic function of the phrase. For isolated noun phrases in contexts such as titles, bibliographical references, mathematical formulae, vocatives, parenthetical structures, dates and places in brackets... the tag 'noun phrase' (`@IS`)[3] was created based on the tagset in the parser Palavras (Bick, 2000).

Also, some tokens such as the numbers in item lists or section headings do not hold a syntactic function. Therefore, the tag @FSG (no syntactic function) was created to annotate these tokens or similar ones that are devoid of a syntactic function.

4.3 Inter-annotator agreement

In order to evaluate the consistency of the annotation guidelines generated so far and the reliability of our corpus, three linguists —two of them with a long experience in several NLP tasks—annotated a part of the corpus. The inter-annotator agreement was measured using Fleiss' kappa (Fleiss, 1971) obtaining 0.945. The observed agreement was 93%. This result shows that our guidelines are clear enough and our tagging is consistent. Besides, they show the reliability of our corpus since agreement is very important for the production of representative text corpora with high-quality linguistic annotation.

Then, we examined the cases in which the annotators disagreed. In 48.55% of the cases, the disagreement was related to the use of the new tags. The disagreements related to the conflictive cases of complex postpositions and complementisers and multiword units were not very common, 7.24% and 2.90% respectively. Nevertheless, 69.56% of the cases where disagreement was found were covered by the guidelines. This suggests that annotators sometimes tended to follow their expertise and intuition rather than the guidelines.

[3] `@IS` stands for Basque *Izen Sintagma* 'Noun Phrase'.

Finally, the whole corpus was annotated bearing in mind all the decisions taken and the expertise gained in the previous stages.

5 Conclusion

Although time-consuming and costly, Gold Standard corpora are essential to develop data-driven language processors as well as to evaluate the output of rule-based processors.

In this paper, we have presented the process in the construction of SF-EPEC a syntactically annotated corpus of 300,000 words aimed to be a Gold Standard for the surface syntactic processing of Basque. Previous to the annotation, a linguistically motivated tagset was designed to account for the morphosyntactic complexity of the Basque language.

The inter-annotator agreement obtained (93%) shows that the tagset developed as well as the criteria established for the annotation are quite sound, and therefore the corpus obtained will be a reliable reference corpus.

Acknowledgments

PROSA-MED: Procesamiento semántico textual avanzado para la detección de diagnósticos, procedimientos, otros conceptos y sus relaciones en informes Médicos (TIN2016-77820-C3-1-R).

References

Aduriz, I. 2000. *EUSMG: Morfologiatik sintaxira Murriztapen Gramatika erabiliz.* Ph.D. thesis, University of the Basque Country (UPV/EHU).

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

Aduriz, I., M. J. 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 Automatic Processing. *Language and Computers*, 56(1):1–15.

Aduriz, I., M. J. Aranzabe, J. M. Arriola, and A. Díaz de Ilarraza. 2006b. Sintaxi

Partziala. In B. Fernández and I. Laka, editors, *Andolin gogoan: Essays in Honour of Professor Eguzkitza*. pages 31–49.

Aduriz, I., J. M. Arriola, I. Gonzalez-Dios, and R. Urizar. 2015. Funtzio Sintaktikoen Gold Estandarra eskuz etiketatzeko gidalerroak. Technical report, University of the Basque Country (UPV/EHU) UPV/EHU/LSI/TR 01-2015.

Aduriz, I. and A. Díaz de Ilarraza. 2013. Morphosyntactic Disambiguation and Shallow Parsing in Computational Processing of Basque. *Anuario del Seminario de Filología Vasca "Julio de Urquijo"*, pages 1–21.

Aldezabal, I., O. Ansa, B. Arrieta, X. Artola, A. Ezeiza, G. Hernández, and M. Lersundi. 2001. EDBL: a General Lexical Basis for the Automatic Processing of Basque. In *Proceedings of the IRCS Workshop on linguistic databases*, pages 1–10.

Aldezabal, I., K. Ceberio, I. Esparza, A. Estarrona, J. Etxeberria, M. Iruskieta, E. Izagirre, and L. Uria. 2007. EPEC (Euskararen Prozesamendurako Erreferentzia Corpusa) segmentazio-mailan etiketatzeko eskuliburua. Technical report, University of the Basque Country (UPV/EHU) UPV/EHU/LSI/TR 11-2007.

Alegria, I., X. Artola, K. Sarasola, and M. Urkia. 1996. Automatic Morphological Analysis of Basque. *Literary and Linguistic Computing*, 11(4):193–203.

Aranzabe, M. J. and A. Díaz de Ilarraza. 2009. Análisis sintáctico computacional del euskera mediante una gramática de dependencias. In *Actas del XI Simposio Internacional de Comunicación Social*, pages 316–320. Centro de Lingüística Aplicada.

Arriola, J. M. 2015. Different Issues in the Design and Implementation of a Rule Based Grammar for the Surface Syntactic Disambiguation of Basque. In *Proceedings of the Workshop on "Constraint Grammar-methods, tools and applications" at NODALIDA 2015*, number 113, pages 1–9. Linköping University Electronic Press.

Bick, E. 2000. *The Parsing System Palavras. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Ph.D. thesis, Aarhus University Press.

Fleiss, J. L. 1971. Measuring Nominal Scale Agreement among many Raters. *Psychological bulletin*, 76(5):378–382.

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

Marcus, M. P., M. A. Marcinkiewicz, and B. Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330.

Mille, S., A. Burga, V. Vidal, and L. Wanner. 2009. Towards a Rich Dependency Annotation of Spanish Corpora. *Procesamiento del Lenguaje Natural*, (43):325–333.

Nilsson, J. and J. Hall. 2005. Reconstruction of the Swedish Treebank Talbanken. Technical report, Växjö University, Sweden. School of Mathematics and Systems Engineering. MSI report 05067.

Sampson, G. 2011. A Two-way Exchange between Syntax and Corpora. In V. Vander, S. Zyngier, and G. Barnbrook, editors, *Perspectives on Corpus Linguistics*, volume XVI, 256. pages 197–211.

Scheible, S., R. J. Whitt, M. Durrell, and P. Bennett. 2011. A Gold Standard Corpus of Early Modern German. In *Proceedings of the ACL-HLT 25th Linguistic Annotation workshop*, pages 124–128. Association for Computational Linguistics.

Silveira, N., T. Dozat, M.-C. de Marneffe, S. R. Bowman, M. Connor, J. Bauer, and C. D. Manning. 2014. A Gold Standard Dependency Corpus for English. In *Proceedings of LREC 2014, the Ninth International Conference on Language Resources and Evaluation*, pages 2897–2904.

Solberg, P. E., A. Skjærholt, L. Øvrelid, K. Hagen, and J. B. Johannessen. 2014. The norwegian dependency treebank. In *Proceedings of LREC'14, the Ninth International Conference on Language Resources and Evaluation*, pages 789–795.

Voutilainen, A., T. Purtonen, and K. Muhonen. 2012. Outsourcing Parsebanking: The FinnTreeBank Project. In *Shall We Play the Festschrift Game? Essays on the Occasion of Lauri Carlson's 60th Birthday*. Springer, pages 117–131.