Morfeus+: Word parsing in Basque beyond morphological segmentation

Itziar Aduriz
University of Barcelona

Jose M. Arriola
University of the Basque Country

Xabier Artola
University of the Basque Country

Zuhaitz Beloki
Elhuyar Fundazioa

Nerea Ezeiza
University of the Basque Country

Koldo Gojenola
University of the Basque Country

Abstract

This work describes the formalization of a word structure grammar that represents the complex morphological and morphosyntactic information embedded within the word forms of an agglutinative language (Basque), giving a comprehensive linguistic description of the main morphological phenomena, such as affixation, derivation, and composition, and also taking into account the modeling of both standard and non-standard words. We have identified the relevant issues to be addressed in the representation of such a grammar.

Word Structure 13.3 (2020): 283–315
DOI: 10.3366/word.2020.0172

www.euppublishing.com/word

We also present the development of Morfeus+, a tool for the analysis of unrestricted texts, testing its applicability and showing that its coverage is wide and robust, allowing the efficient processing of big volumes of data.

This paper describes a mature system that has required several person/years and that tries to integrate a rigorous linguistic specification together with more practical implementation matters, such as the appropriate treatment of unknown words in unrestricted texts.

Keywords: word structure grammar, morphosyntactic analysis, computational treatment of variants, derivation and composition, morphologically rich languages

1. The need for a word structure grammar for processing morphologically rich languages

The main goal of this paper is to highlight the issues encountered when analyzing the internal structure of words in Basque (ISO 639–1: eu), a highly agglutinative language. The topics covered will include the interrelation between syntax and morphology and its formalization to represent the linguistic information embedded within word forms. We will base our study on data from corpora, analyzing composition, derivation, and linguistic variants.

Basque presents a complex intraword morphological structure based on morpheme agglutination.[1] In this respect, in Zuñiga & Fernández (2019: 185) we have the following description: "Basque morphology is largely agglutinative, i.e., it is predominantly concatenative and of separative exponence (except in the person-number inflection of verbs), with some flexivity (i.e., the allomorphy found in inflectional phenomena is not purely phonological) in both the verbal and nominal domains." Besides, the SOV pattern, together with agglutination and ergativity are perhaps the most characteristic features of Basque, or at least the most often mentioned ones (Manterola, 2008).

Following Alegria *et al.* (1996) we can say that Basque can be considered as a morphologically rich language. Prepositional functions are realized by case suffixes inside word forms, Basque presenting a relatively high capacity to generate inflected word forms. For instance, from one noun entry a minimum of 135 inflected forms can be generated. Moreover, while 77 of them are simple combinations of number, determination, and case marks, not capable of further inflection, the other 58 are word forms ending with one of the two possible genitives (possessive and locative) or with a sequence composed of a case mark and a genitive mark. If the latter is the case, then by adding again the same set of morpheme combinations (135) to each one of those 58 forms

[1] There is some work comparing Basque morphology with other languages. For instance in the third chapter of Salaburu & Alberdi (2012), "Basque and Romance Languages: Languages with Different Structures", we have a comparison with Romance languages by I. Zabala and I. San Martin. Moreover, Ormazabal (1992) did a comparison of general morphological processes in three languages: Basque, English, and Spanish.

a new, complete set of forms could be recursively generated. This kind of construction, revealing a noun ellipsis inside a complex noun phrase, could be theoretically extended *ad infinitum*, although in practice it is not usual to find more than two levels of this kind of recursion in a single word form.

Word formation is also very productive, as it is very usual to create new compounds as well as derivatives. As a result of the wealth of information contained within word forms, complex structures have to be built to represent complete morphological information at word level. For instance, in *mendian*, Basque for 'in the mountain', *mendi* stands for 'mountain', *-a* for the determiner (translatable as 'the'), and *-n* for the locative case (translatable as 'in'). In other cases like *alaba*, Basque for 'the daughter' (*alaba+a*), and *alabek*, Basque for 'the daughters' in the ergative case (*alaba+ek*), where some coalescence between stem and suffixes occur, the morphological segmentation process deals with the interactions between morpheme boundaries.

The determiner, number, and case morphemes are appended to the last element of the noun phrase and always occur in this order. In the previous example, the *-n* suffix added to the stem *mendi* 'mountain' assigns the verb complement function (locative) to the word. Regarding verbal morphology, the main verbs as well as the auxiliaries have been stored in sublexicons and, by means of morphotactics, we define which morphemes can be combined with the verb entries. For main verb forms, verbal inflection is represented as aspect and factitive morphemes. For auxiliary verbs, although they could be decomposed into morphemes, as they form a closed and relatively reduced set, they have been stored in sublexicons, thus easing the analysis process and reducing the number of morpheme sublexicons.

One of our focus points in this paper are morphosyntactic rules. The rules work on the output of morphological segmentation (Aduriz *et al.*, 2000), and the work done in the formalization of Basque morphology according to the two-level model of computational morphology (Koskenniemi, 1983).

In this paper, we focus on both morphology and morphosyntax, since in highly agglutinative languages like Basque (Zuñiga & Fernández, 2019; Manterola, 2008), Turkish, Hungarian, or Finnish, it is difficult to separate one from the other, and so, it is necessary to thoroughly parse all the information found at word level in the first stage of language processing (Sak *et al.*, 2011; El-Haj *et al.*, 2014; Haverinen *et al.*, 2014). We consider morphosyntactic parsing as the first phase of shallow syntactic analysis. In the remainder of the paper, we will use the term morphosyntactic analysis to refer to the formalization of the morphological structure of word forms.

Apart from the description and formalization of a language processor that will be an essential tool in practical applications, we believe that descriptive linguistics can also benefit from this kind of morphosyntactic analysis and its formal representation, as this formalization could help to fully understand the involved linguistic phenomena.

Regarding previous efforts on morphological analysis tools, we can distinguish two main groups of processors:

- General tools that include the full range of basic linguistic processors, including sentence splitting, tokenization, tagging, and syntax, among more specific ones like

Named Entity Recognition. Among these, we can name two widely used and robust general NLP tools: (1) Freeling (Padró & Stanilovsky, 2012), which uses a dictionary-based approach combined with a probabilistic prediction of unknown word categories, and that has been tested on a range of languages, including Catalan, English, Spanish, or Russian; and (2) OpenNLP[2], whose morphological processor, based on a part-of-speech (POS) tagger, is not well suited for morphologically rich languages, as it is based on the simplest options of no dictionary (the tagger guesses the category of each word) or adding a dictionary of (word form, tag) pairs. From our point of view, neither of these solutions seems satisfactory for morphologically rich languages because, if we only take into account standard word lemmas, Basque (for instance) can produce a plethora of different (correct) word forms. Furthermore, these tools would be more problematic with non-standard texts.

- Tools specifically designed with morphologically rich languages in mind. Among these, we can mention (1) the morphological processor for Turkish introduced by Sak *et al.* (2011), based on a stochastic transducer, which presents a coverage of 96.7% on a text corpus collected from online newspapers; (2) the system introduced by Şahin *et al.* (2013) also for Turkish, which implements a complementary unknown-word analyzer by making use of wildcard entries; and (3) the Czech morphological processor (Hajič, 2004; Spoustová *et al.*, 2007), which includes a guesser for unknown tokens. In the later case, the base of the morphological analyzer is a large dictionary containing more than 350,000 entries, and the authors state that, on average, 2.5% of all word forms are unrecognized, most of them foreign proper names and typos.

For the present study, we have developed a module that deals with all the phenomena involved in the morphological description of Basque word forms, integrating it in Morfeus+, a robust corpus processing tool. Specifically we have dealt with the following issues:

- The manual compilation of a robust lexicon by a group of linguists and computer scientists over several years.
- The specification of a complete set of linguistic principles for representing all the morphological and morphosyntactic phenomena: on the one hand, those related to the merging of multiple values for case, number, and definiteness (inflection) and, on the other hand, the description of the internal structure of words, as well as the relevant syntactic and semantic features that correspond to all the elements as a whole.
- The treatment of derivatives and compounds, producing analyses with rich information and a completely defined structure, which combines morphological, syntactic, and semantic features.
- The incorporation of a new method to describe orthographic and dialectal variants (see Section 3.4). We deal with variants in inflectional morphology as well as with

[2] https://opennlp.apache.org/index.html (accessed: 2019-07-01).

lexical variants (produced due to dialectal usage, competence errors, use of non-standard forms, etc.). The representation of the word links the variants with their corresponding standard forms at a morpheme level. In this way, Morfeus+ gives information about the status of each word with respect to the standard entries contained in the *Hiztegi Batua*, the Unified Basque Dictionary (UBD) (Euskaltzaindia, the Basque Language Academy, 2000).

Regarding variants, their morphological treatment is integrated into the morphological processor, instead of using a closed list of variants. This opens up the way to the recognition of complex combinations of variants, compounds, and derivatives (hereafter referred to as hybrids), which other systems usually treat by means of probabilistic or guessing techniques. In our approach, the coverage is extended considerably, as the morphological processor is able to correctly analyze different combinations of these phenomena. For instance *oxijeno-hornitzailea* (Basque for 'the oxygen supplier') contains a non-lexicalized compound word form, composed of a spelling variant (*oxijeno*) and a non-lexicalized derivative (*hornitzailea*, Basque for 'the supplier').

Moreover, in this work we adopt a completely revamped encoding of linguistic annotations by using a stand-off markup format based on XML (Artola *et al.*, 2009)[3].

The tool presented here is currently being used in a wide set of tools and applications that process Basque, including a spelling checker, syntactic corpus processing (Otegi *et al.*, 2017), machine translation (Artetxe *et al.*, 2016), text simplification (Gonzalez-Dios *et al.*, 2018), sentiment analysis (Alkorta *et al.*, 2018), biomedical text processing (Perez-de-Viñaspre *et al.*, 2018) and discourse processing (Bengoetxea & Iruskieta, 2018). Finally, we have tested it against several corpora, to determine its usefulness in real texts, and we have carried out an evaluation on a large corpus, in order to assess the validity and expressiveness of the system.

The remainder of the paper is organized as follows: after an overview of related work in Section 2, we will give some basic background on Basque morphology and specific topics such as composition, derivation, and the treatment of variants in Section 3. Section 4 will be devoted to the representational issues of the word structure grammar. In Section 5 we will provide a corpus-based analysis of the morphosyntactic phenomena under consideration. Finally, in Section 6 we will draw some conclusions, suggesting avenues for future developments.

2. Related work

In this section, we will review relevant work corresponding to both linguistic formalizations of morphology and its major computational developments. Although most published studies have focused exclusively on one of these two areas, we believe

[3] Stand-off markup or annotation, as opposite to inline markup (or annotation), is the kind of markup that resides in a location different from the one the data being described by it reside.

that they have much in common, and that interesting advances can be made trying to keep both linguistic insight and computational efficiency and preciseness at the same time.

Three main aspects are defined for morphological processing (Ritchie *et al.*, 1992): a) morphophonology (or morphographemics) defines the segmentation of words into morphemes and the changes in their combination, b) morphotactics defines the sequential combinatorics of different sets of lexical units (usually by means of continuation classes), and c) morphosyntax (or word structure grammar), responsible for putting the information together. Implemented systems differ regarding their use (or not) of these three elements.

From the computational point of view, Koskenniemi (1983) defined two-level morphology for morphographemics and morphotactics, which has been successfully applied to a wide variety of languages including Basque (Alegria *et al.*, 1996). Karttunen (1994) improved the two-level model compiling two-level rules into lexical transducers, also increasing the expressiveness of the model.

The morphological analyzer created by Ritchie *et al.* (1992) does not adopt any finite-state mechanism to control morphotactic phenomena, but their two-level implementation incorporates straightforward morphotactics instead, and the segmentation step yields multiple hypotheses for segmentation that will be later discarded by a posterior unification phase. This approximation would be highly inefficient for agglutinative languages, as it would create many nonsensical interpretations that would subsequently be rejected. They use a word structure grammar for both morphotactics and feature combination. Following a similar approach, Trost (1990) makes a proposal to combine two-level morphology and non-sequential morphotactics. The PC-Kimmo-V2 system (Antworth, 1994) presents an architecture for morphological analysis of English, using a finite-state segmentation phase before applying a unification-based grammar.

Earlier descriptions of Basque morphological analysis with finite-state techniques are given by Aduriz *et al.* (1993) and Alegria *et al.* (1996). A later implementation using the Xerox/PARC compilers is described in Alegria *et al.* (2002). Aduriz *et al.* (2000) present a model for designing a full morphological analyzer for Basque, integrating the two-level formalism and a unification-based formalism. They propose separating the treatment of sequential and non-sequential morphotactic constraints. Sequential constraints are applied in the segmentation phase, and non-sequential ones in the final feature-combination phase, using a word-level unification grammar. Early application of sequential morphotactic constraints during the segmentation process makes feasible an efficient implementation of the full morphological analyzer.

Oflazer (1999) presents a more radical approach for the treatment of Turkish, applying directly a dependency-parsing scheme to morpheme groups, that is, totally merging morphology and syntax. Although a similar model could be applied to Basque, many applications are word-based and need full morphological parsing of each word form (Karlsson *et al.*, 1995).

Taking a view from the linguistic side, Aduriz *et al.* (2000) followed an approach similar to classical morphology in the sense that they were mainly concerned with the

arrangement of morphemes in a particular order. In this arrangement affixes have the same status as words and they are stored in the lexicon. In contrast, this paper will deal with the arrangement of the features that will be assigned to the overall analysis of word forms. For this purpose, once the sequential rules determine the combination of lemmas and suffixes, we define the features that must be promoted as top-level features, thereby describing the whole word form.

Haspelmath & Sims (2010) distinguish two models for morphology. In the morpheme-based model, morphological rules combine morphemes in the same way that syntactic rules combine words. Although some authors claim that all the phenomena can be described by a pure concatenative approach, Haspelmath and Sims are of the view that this poses considerable difficulties in expressing hierarchical structure. In their word-based model, morphology is described by word schemas that represent the features common to morphologically related words, a solution akin to feature structures and unification. Aronoff & Fudeman (2011) also emphasize the importance of giving a hierarchical structure to word components, especially for derivation and composition, by means of tree diagrams that distinguish the scope to which each phenomenon is applied. In the same way, Bender (2013) considers morphosyntax critical for extracting sentence meaning, presenting a varied set of examples that show how the internal structure of words can have effects on phonology, syntax, and semantics. Similarly, Würzner & Hanneforth (2013) propose a model for the morphological analysis of German that goes beyond a flat structure. Their approach uses a morphological analyzer of German based on weighted finite-state transducers to segment words into lexical units and a probabilistic context-free grammar trained for the parsing step, assigning hierarchical structures to complex words.

Recently, there has been a surge of interest in automatic methods for the processing of morphology, using either statistical methods or deep learning algorithms (Goodfellow *et al.*, 2017). These methods claim to be effective for a rapid deployment of morphological processors for multiple languages, with promising results on the experiments performed so far (Cotterell *et al.*, 2017).

The development of machine learning tools for morphology has proved successful in several cases, with good results on multiple and varied languages for the proposed tasks, as the learning of several nominal and verbal paradigms, or predicting inflected forms from a given sample. Although the opening of this new research avenue has shown high performance with the proposed datasets (Lee & Goldsmith, 2016; Cotterell *et al.*, 2017), there are still some important issues that need further research:

- The experiments have been performed on a reduced set of phenomena, such as a limited set of noun inflections or verbal paradigm completions, being far from a complete morphological system for any given language.
- The proposed tests have only used a restricted set of morphological information, not covering the full morphological description of a language. Looking at the datasets in Cotterell *et al.* (2017) we see that the data include several of the most regular phenomena, albeit trying also to consider the highest number of irregularities inside

each selected paradigm. In contrast, the objective of our work is the construction of a high-quality and robust system.
- Even when using the successful string-to-string neural methods (Goodfellow *et al.*, 2017), every phenomenon would need a dedicated set of training data, which seems a costly enterprise.

As a comparison, we have analyzed the results corresponding to the morphological analysis of Basque at the *Shared Task on Universal Morphological Reinflection in 52 languages* (Cotterell *et al.*, 2017). We can see that, for Task 1 (automatic inflection of forms, given sparse training data), the best system obtained a per-form accuracy of 89.0% when trained with 1,000 samples, far from the results given by our system, and without taking into account that the shared task only concentrated on a reduced set of inflections for nouns and verbs, far from the requirements of a production tool like the one presented in this paper.

Other approaches try to use unsupervised learning of morphology that would allow us to avoid the bottleneck of knowledge acquisition. For example, Khaliq & Carroll (2013) present a morphological analyzer of Arabic by inducing a lexicon of root and pattern templates from an unannotated corpus using maximum entropy modeling. Although the results are encouraging, reaching an accuracy of 87.2%, they are still far from those obtained using knowledge-based approaches.

In this respect, our work describes a robust and full-fledged system instead, that has been developed under a knowledge-based approach.

To summarize the previous discussion, several proposals have been given for the linguistic description of morphological processes, and most of them apply: (1) concatenative approaches for segmentation, which are treated efficiently and elegantly using two-level descriptions; and (2) solutions that deal with the hierarchical structure of several phenomena, which can be dealt with using syntactic context-free rules.

From the computational point of view, most works, for practical reasons, have not tackled in depth the problem of having rich morphological information and, as a result, most of the currently employed systems, such as morphological analyzers integrated into taggers or syntactic analyzers, only make use of superficial information like part-of-speech tags or flat lists of multiword lexical units, as in the widely used Stanford CoreNLP tool (Toutanova & Manning, 2000; Klein & Manning, 2003).

3. Composition, derivation, variants, and ambiguity

In this section we will firstly present the differences between the theoretical and the practical points of view with regard to the linguistic phenomena under consideration. Following this, we will deal with composition and derivation and, in so doing, we will briefly describe some characteristics of lexical morphology. We will then present the phenomenon of variants and standardization of the language.

Lexical morphology (derivation or composition), as well as inflection, occurs by adding morphemes and/or lexemes to the stem of a word. As a result, complex structures

have to be built to represent complete morphological information at word level. Basque is characterized by highly productive lexical morphology that may produce a large number of words for a given base form. Furthermore, there is the problem of variants due to the fact that Basque is still involved in both a normalization process and the standardization of the language (Hualde & Zuazo, 2007). For this reason you can find in texts a significant number of non-standard word forms, among which we consider not only out-of-vocabulary words (OOV), but also linguistic variants.

Note that many of these variants can also be derivatives, compounds, or both. In fact, these phenomena may occur with all kinds of morphemes, both lemmas as well as lexical or inflectional suffixes. Therefore, we need a processing model and an architecture that explicitly integrates this range of information into the morphological analysis as a basic step for further processing stages, especially for parsing and semantics.

Every morphological unit contains many types of linguistic information, including:

- Basic morphosyntactic features, such as main category (or part of speech), subcategory, case, number, gender, tense, aspect, or subordination type (for verbs).
- Features that serve as a reference to the UBD. These features reflect the corresponding dictionary entry, the homograph number, and information that defines whether the entry is a normalized standard entry. Thus, every text element can be linked to its corresponding entry in EDBL, a general lexical database (Aldezabal *et al.*, 2001), or otherwise it is made explicit that the word does not have a correspondence, that is, it could be either a variant, a misspelling, a neologism, or a foreign term.
- Features that serve to describe variants of standard entries. For these entries, the feature structure will present the variant's lemma itself as well as its corresponding standard entry, if any. The lexicon contains a list of typical errors, deviations, and dialectal variants, which will allow the linking of many non-standard word forms with their standard entries. These features are especially important for measuring the degree of standardization or dialectal variability of different text corpora.

In the next subsections we will describe several linguistic phenomena that have been taken into account when modeling a morphological module for Basque.

3.1 The phenomena treated and their context of application

Many authors of computational NLP systems, especially for morphologically rich languages, have recognized "the lexical challenge" that appears when extending the coverage of the systems to obtain a robust and accurate prediction of any word form (Tsarfaty *et al.*, 2013; Seeker & Kuhn, 2013). This has often been named as the OOV problem (Goldberg & Elhadad, 2013). For example, Foster (2010) and Foster *et al.* (2011) examined the usability of an English parser and POS tagger on texts written in Twitter, and found that the accuracy of the POS tagger diminishes drastically from 96.3% to 84.1%. These studies imply that a robust and accurate morphological analysis is essential for the success of any NLP application.

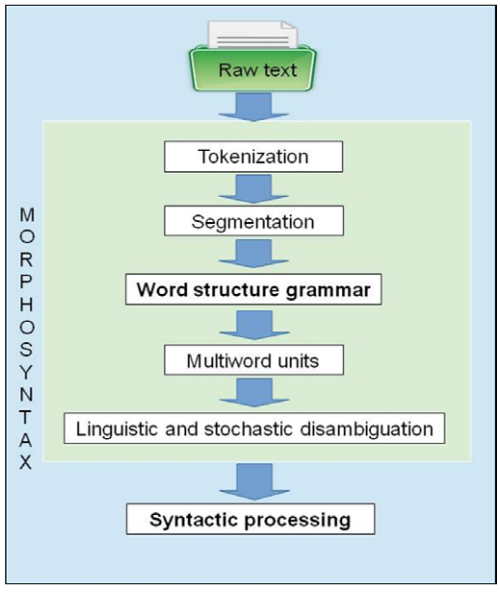

Figure 1: The morphosyntactic analysis chain

In our case, morphosyntactic processing will follow the analysis chain described in Figure 1. It is composed of the following language-processing tools:

- Tokenization.
- Segmentation. This module is based on a set of two-level rules compiled into finite-state transducers. It corresponds to the concatenative part of morphology.
- Word structure grammar. This module named Morfeus+ combines information from multiple morphemes, also taking into account phenomena such as derivation, composition, and dialectal and orthographic variation. This module deals with the internal hierarchical structure of words, producing the set of features corresponding to the word form as a whole.
- Recognition of multiword units.
- Disambiguation. Due to the high word-level ambiguity of Basque, each token receives many morphological interpretations, which are resolved by a module that combines linguistic disambiguation rules with a stochastic model (Ezeiza *et al.*, 1998).
- Syntactic processing.

The morphological segmentation of words is performed in three main phases and gives, as a result, all the possible analyses of each word in the text:

- The analysis of standard forms. In this phase, the processor is able to analyze and generate standard-language word forms based on a general lexicon and the corresponding rules for morphotactics and morphophonological changes.

Table 1: Distribution of tokens in increasingly bigger subsets of the EPEC training corpus

Tokens	13,507	25,812	37,233	49,255	67,450	130,422
% of corpus	10.36	19.79	28.55	37.77	51.72	100.00
Standard (%)	74.81	76.59	77.35	78.41	79.10	78.73
Variant (%)	0.63	0.62	0.62	0.54	0.54	0.58
OOV (%)	1.60	2.38	2.93	2.84	2.71	2.80
Other (%)	22.97	20.41	19.10	18.21	17.64	17.90

- The analysis and normalization of linguistic variants. The lexicon used in this phase includes extra entries for variants, linked to their corresponding standard forms, and rules to effectively treat high-frequency morphophonological changes according to dialectal uses and competence errors.
- The analysis of words based on lemmas not belonging to the previous lexicons. In this phase, the guesser uses a lexicon simplified by allowing only open categories (nouns, adjectives, verbs, and so on) and any combination of characters as lemmas. These generic lemmas are combined with affixes related to open categories in order to capture as many morphologically significant features as possible.

In this three-phase architecture, each word will be processed by the first analyzer, which is able to produce at least one valid segmentation according to the previously defined order.

In order to illustrate this process, we will examine EPEC (Aduriz *et al.*, 2006), an automatically annotated corpus for Basque, from which some distributional figures will be shown. EPEC was initially a 50,000-word sample collection of written standard Basque that has since been extended to 300,000 words. The collection of texts was obtained from the *Statistical Corpus of 20th Century Basque*[4]. This corpus has been automatically processed and the results have been manually revised.

In order to measure the incidence of each type of word, we have analyzed the distribution of tokens in the training part of EPEC. We have divided it into increasingly bigger subsets to verify whether the proportions remain stable with respect to the number of tokens in the corpus. Table 1 shows the actual figures: the first row gives the exact number of tokens[5] used in each subset; the second row shows the percentage of tokens in the corpus used for each subset; the subsequent rows give the percentage of tokens for each type of word (standard, variant, and OOV); and the last row shows the percentage that corresponds to other kinds of tokens (including punctuation marks, separators, etc.).

[4] http://xxmendea.euskaltzaindia.eus/Corpus/ (accessed: 2019-07-01).

[5] The criterion to create the subsets to train and test was the number of words; therefore the total number of tokens is higher.

Table 2: Ambiguity measures in the segmentation output for the EPEC test corpus

	Distribution	Ambiguity Rate	Interpretations per token	Recall	Precision	F-score
Standard	78.11	82.38	3.62	99.41	27.46	43.03
Variant	0.74	76.92	3.64	76.72	20.81	32.65
OOV	3.15	99.41	15.70	91.76	5.84	10.98
Words	82.00	82.98	4.08	98.91	24.11	38.91
Average	100.00	68.05	3.53	99.11	28.08	43.76

As can be observed in the table, the number of non-standard words (variants and OOVs) adds up to 3.38% of tokens (around 4% of words). However, we have observed that when the source of the collection does not follow the standards of the language or when they include a larger number of dialectal variants, neologisms, or OOV proper names, the proportion of non-standard words can increase up to 10%.

Among non-standard words, the range of dialectal variants and competence deviations may vary depending on the source of the corpus, as will be seen in Section 5. In the case of EPEC, we can observe in Table 1 that the number of variants is not very high (0.58% of tokens). However, it is essential to link them to their corresponding entries in the standard lexicon, whenever possible. Incorporating these links into computer-based applications, such as spelling correctors, is helpful in promoting the use and learning of the standard language. In addition, as opposed to the guesser, the morphological segmentation of variants includes closed categories in the lexicon, allowing the assignment of appropriate analyses to words that otherwise would not be assigned their correct interpretations.

Table 2 shows ambiguity measures of morphological segmentation for EPEC. The performances of the three segmentation phases are given separately, and the last row accounts for all tokens, including other kinds of tokens, usually unambiguous, such as separators and punctuation marks. Taking all the tokens into account, 68.05% of them are ambiguous with an average of 3.53 analyses. If we consider only text words (82% of the tokens), 82.98% of them are ambiguous, with 4.08 interpretations assigned to each on average. Alternatively, we have used recall, the number of correct interpretations out of the total number of tokens, to measure the performance of the processor. The error rate is relatively low, less than 1%, although the performance is obviously lower for variants and OOV words than for standard ones.

We want to point out that the guesser poses an overgeneration problem, as it has to consider all the possible valid segmentations for all the open categories, making OOVs artificially ambiguous. In order to reduce the artificially produced high number of analyses, we apply some context-free disambiguation heuristics. Table 3 presents the new figures taken after this procedure has been applied to the EPEC test corpus. The drop from 15.70 (see Table 2) to 6.56 interpretations (see Table 3) in OOV words is the main

Table 3: Ambiguity measures after context-free procedures for the EPEC test corpus

	Distribution	Ambiguity Rate	Interpretations per token	Recall	Precision	F-score
Standard	78.11	81.75	3.55	99.41	28.00	43.70
Variant	0.74	76.47	3.51	75.00	21.36	33.26
OOV	3.15	99.22	6.56	78.60	11.98	20.79
Average	100.00	67.54	3.19	98.68	30.98	47.15

improvement. Of course, there is a loss in recall, which is less than 0.5% taking all the information into account, but this reduction in ambiguity helps in better disambiguating each word in context.

3.2 Derivation

There are about 80 derivational affixes in Basque. Some of them are not productive in any way, and they are a source of lexicalized derivatives (those having an entry in the lexical database). Some others, however, are very common and productive, and need to be processed by the analyzer. We have selected the most productive affixes (about 20) to treat the widest range of non-lexicalized derivatives in corpora. Most of the derivation affixes change the lexical category or the subcategory of the base form, but also semantic features must be dealt with in order to correctly analyze derivative words.

```
Input: dokumentalgilea
Output:
  Segmentation
     (dokumental)          BASE  N C -ANIM
     (-gile)               LEX   N C DOER +ANIM
     (-a)                  INFL  ABS S +DEF (@OBJ @SUBJ @PRED)
  Morphosyntactic Analysis
     [dokumental + -gile + -a]
                           N C DOER +ANIM ABS S +DEF (@OBJ @SUBJ @PRED)
```

Example 1 Segmentation and morphosyntactic analysis of *dokumentalgilea*.

Based on this, the analysis for derivatives deals with the following features: part of speech of the derivative, subcategory of the derivative, category and lexical information of the base form, case, determiner, number, and syntactic function. For instance, the derivative *dokumentalgilea* (Basque for 'the documentary maker', see Example 1) contains a base form, in this case a noun, *dokumental* (Basque for 'documentary'), and a derivational suffix, *-gile* (cf. *egin* 'to make', *egile* 'maker'), where the suffix is taken as the head in the sense that it subcategorizes the base forms that can be attached to it and their corresponding features. The suffix *-gile* is added to nouns, creating words with the meaning 'someone whose job is to make N', that is, a suffix which denotes the maker or

doer of something (N, the noun). According to this, the morphological rules will build up the word-level information of derivatives:

- **Structure of the word form**: base form (*dokumental*) + lexical suffix (*-gile*) + inflection suffix (*-a*).
- **POS of the derivative word form**: noun (N). In this case the lexical suffix does not change the part of speech or category of the base form for the resulting derivative word.
- **Subcategory of the derivative word form**: common noun (C). Similarly, the lexical suffix keeps the subcategory of the base form.
- **Semantic features of the derivative word form**: an animate 'doer' (+ANIM) that makes an artifact. The meaning or semantics of the word is changed with respect to the base form by means of the lexical suffix *-gile*.
- **The values for case, number, and definiteness**: absolutive (ABS) and definite (+DEF) singular (S).
- **The surface syntactic tags**: the word is ambiguous with respect to syntactic function tags: @OBJ (object), @SUBJ (subject), or @PRED (predicative).

In *dokumentalgilea*, the suffix *-gile* selects the POS of the nominal base form (N). As a result of the derivation process the POS of the base form is still a noun, but the agentive ('doer') information is added by the derivative and this information will be coded as semantic information. The *-a* morpheme corresponds to the absolutive case, singular and definite, acting as object, subject, or predicate.

3.3 Composition

Composition in Basque, as well as derivation, is an intrinsic or inherent path to word generation (Euskaltzaindia, 1987, 1991, 2014; Euskara Institutua). The most common definition of composition focuses on the union of two or more independent elements which may belong to different grammatical categories and also have different semantic features. In Basque the most common compounds are created joining nouns (N), adjectives (ADJ), and verbs (V)[6] (see Example 2).

N+N = *zabor-hipoteka* (*zabor* 'rubbish' + *hipoteka* 'mortgage')
'subprime-mortgage'
ADJ+ADJ = *zuri-urdin* (*zuri* 'white' + *urdin* 'blue') 'white-blue'
N+ADJ = *ilegorri* (*ile* 'hair' + *gorri* 'red') 'red-headed'
V+N = *jarleku* (*jar* 'to seat' + *leku* 'place') 'seat'

Example 2 Examples of the most common composition schemes.

[6] The Basque Language Academy distinguished 17 composition types (Euskaltzaindia, 1992).

Many of these compounds are lexicalized as subentries in dictionaries, so they will also be stored in EDBL. In the case of non-lexicalized compounds, Morfeus+ deals with the most frequent ones, that is, hyphenated N+N compounds , where N corresponds to either a noun or a nominalized verb (nominalization realized by means of derivation morphemes that convert a verb into a noun, e.g. 'destruction' = 'destroy' + '-tion').

In the well-known discussion about the relation between morphology and syntax, it is commonly accepted that composition is very close to syntax (Scalise & Vogel, 2010). In this way, since Basque is syntactically characterized as a right-headed language, the main information of the compound is taken from the second element.

For example, the compound *zabor-hipoteken* (*zabor* 'rubbish' + *hipoteka* 'mortgage'+ *-en* 'of the', Basque for 'of the subprime mortgages'), is an N+N compound where the left element *zabor* depends semantically and syntactically on the lexical element situated to its right, *hipoteka*, which is the head.

The morphosyntactic rules code the various features conveyed by each component of the compound in order to build up its word-level information as indicated by Example 3:

- **Structure of the word form**: noun (*zabor*) + hyphen + noun (*hipoteka*) + inflection suffix (*-en*).
- **POS of the compound**: noun (N).
- **Subcategory of the compound**: common noun (C).
- **Semantic features of the word form**: non-animate (-ANIM)
- **Values for case, number, and definiteness**: genitive (GEN), definite (+DEF), plural (PL).
- **Surface syntactic information tags**: the syntactic function is given by a suffix attached to the last element of the compound (the *-en* genitive marker), with the functions @<NCOMPL (left-headed noun complement) and @NCOMPL> (right-headed noun complement).
- **Type of compound**: noun + noun (N+N). This feature is obtained taking into account the structure of the compound.

```
Input: zabor-hipoteken
Output:
  Segmentation
     (zabor)              BASE  N C -ANIM
     (-)                  HYPHEN
     (hipoteka)           BASE  N C -ANIM
     (-en)                INFL  GEN PL +DEF (@<NCOMPL @NCOMPL>)
  Morphosyntactic Analysis
     [zabor + - + hipoteka + -en]
                          N C -ANIM GEN PL +DEF N+N (@<NCOMPL
                          @NCOMPL>)
```

Example 3 Segmentation and morphosyntactic analysis of *zabor-hipoteken*.

3.4 Variants

There are two principal causes for the existence of variants in Basque. One of these is competence errors. As already mentioned, the language is still involved in a process of both normalization and standardization, which started in 1968. In this year the Basque Academy took the first steps in the creation of a language standard, called *Euskara Batua* (Basque for 'Unified Basque'). Therefore, we should also take into account information about competence errors when analyzing Basque texts.

The other source of variants is dialectal usage. Standard Basque co-exists with the main five dialects in the territory and sometimes interference occurs. To achieve good coverage, we included the principal dialectal variants in EDBL, where they are related to their correspondent standard form(s).

For instance, *biyotzetikan* (Basque for 'from the heart') contains a non-standard lemma *biyotz* (variant of *bihotz*, 'heart') and a non-standard suffix corresponding to the ablative case, *-tikan* (dialectal variant of the standard ablative suffix *-tik*). The analyzer will connect the non-standard components (-STD) of the word to their corresponding standards (STD): the standard lemma for *biyotz* is *bihotz* and the standard suffix for *-tikan* is *-tik*. The analysis rules link EDBL with the lexical repository of standard words, and the rules also deal with the base form features in order to build up the word-level information of variants as illustrated by Example 4:

- **Structure of the word form**: base form (*biyotz*, non-standard variant of *bihotz*, **-**STD) + inflection suffix (**-*0***, definite singular) + inflection suffix (*-tikan*, non-standard variant of the ablative morpheme *-tik*, **-**STD).
- **Linking the word form with its corresponding standard**: *bihotzetik*, surrounded by slashes, represents the corresponding standard form.
- **POS of the word form**: noun (N).
- **Subcategory of the word form**: common noun (C).
- **Semantic features of the word form**: non-animate (-ANIM).
- **Values for case, number, and definiteness**: ablative (ABL), singular (S), and definite (+DEF).
- **Linking the variant suffix with its corresponding standard**: the variant *-tikan* is linked to its standard form *-tik* (surrounded by slashes), also signaling the phenomenon involved (DIAL, for dialectal). This deviation is encoded in the lexical database (EDBL).

```
Input: biyotzetikan
Output:
  Segmentation
    (biyotz) /bihotz/          -STD   BASE   N C -ANIM
    (-0)                       INFL   S +DEF
    (-tikan) /-tik/            -STD   INFL   ABL DIAL @ADLG
  Morphosyntactic Analysis
    [biyotz + -0 + -tikan] /bihotzetik/
                               -STD N C -ANIM ABL S +DEF DIAL @ADLG
```

Example 4 Segmentation and morphosyntactic analysis of *biyotzetikan*.

- **The surface syntactic information tag**: the syntactic function @ADLG (adverbial complement) is conveyed by the *-tikan* ablative case marker.

As a result of applying the analysis rules, we have a common non-standard noun linked with its corresponding standard word form, in ablative case (conveyed by a non-standard suffix), singular, definite, non-animate, and acting as adverbial.

3.5 Hybrids

In the previous sections we have described derivation, composition, and variants. Now, we want to show that these phenomena can occur in a word at the same time. For instance, derivatives may be part of compounds, compounds may take derivational suffixes, and variation phenomena can appear in both derivatives and compounds. These hybridization phenomena increase processing complexity and present interesting examples.

In *oxijeno-hornitzailea* (Basque for 'the oxygen supplier', see Example 5), we have an N+N compound where the left element is *oxijeno*, a spelling variant of standard *oxigeno*, and the second element is the lexicalized derivative *hornitzailea*. This form is analyzed as an N+N compound. The hyphen is recognized as well and it is treated as a lexical element:

- **Structure of the word form**: noun (*oxijeno*, non-standard variant of *oxigeno*, **-**STD) + hyphen + derivative noun (*hornitzaile*) + inflection suffix (*-a*).
- **POS of the compound**: noun (N).
- **Subcategory of the compound**: common noun (C).
- **Semantic features of compound**: agent (DOER).
- **Values for case, number, and definiteness**: absolutive (ABS), definite (+DEF), singular (S).
- **Surface syntactic information tags**: the syntactic function is given by a case marker suffix attached to the compound (*-a*) with the functions subject (@SUBJ), object (@OBJ), and predicative (@PRED).
- **Type of compound**: noun + noun (N+N). This feature is obtained taking into account the structure of the compound. The second noun indicates that there is a human agent that SUPPLIES something.

```
Input: oxijeno-hornitzailea
Output:
  Segmentation
    (oxijeno) /oxigeno/ -STD     BASE N C -ANIM
    (-)                 HYPHEN
    (hornitzaile)       BASE N C DOER +ANIM
    (-a)                INFL ABS S +DEF (@OBJ @SUBJ @PRED)
  Morphosyntactic Analysis
    [oxijeno + - + hornitzaile + -a]
                        N C DOER +ANIM ABS S +DEF N+N (@OBJ @SUBJ @PRED)
```

Example 5 Segmentation and morphosyntactic analysis of *oxijeno-hornitzailea*.

The power of the presented approach does not lie in the mere inclusion of non-standard entries in the lexicon, which has been the simplest option, already applied in several systems, but in the union of those non-standard lexicons with finite-state morphology and the word grammar, so allowing the recognition of hundreds of variants of those non-standard elements. This way, not only *oxijeno* can be recognized, but any of its inflected forms, including any form obtained by composition or derivation. Moreover, any other frequent misspelling involving the *j*/*g* change can be recognized without including an exhaustive list of possible variant lemmas replacing *g* by *j*. This also applies to all the other frequently misspelled pairs of consonants used to design the morphophonological transformations of variants, which are mainly based on orthographic differences between Basque and Spanish.

4. The word structure grammar

We have developed Morfeus+, a processing module that deals with all the phenomena involved in Basque word forms from the morphological point of view (Aduriz *et al.* 2000), and incorporated it in a robust corpus processing tool. In this section we will focus on the design of a unification-based word structure grammar, which combines the information conveyed by the different lemmas and morphemes that compose a given word form.

In the rest of this section, we will start presenting the representation schema adopted for dealing with the data interchanged between the linguistic processing tools, a stand-off schema based on XML and annotation standards (see Section 4.1). Then, Section 4.2 will present an overview of the main rules developed for the word processing grammar, and the main principles guiding their design.

4.1 Representational issues

AWA (Annotation Web Architecture: Artola *et al.*, 2009) is a data model for representing linguistic annotations, designed to serve as a schema for the annotation of a very broad range of linguistic phenomena. All the modules of the analysis chain described in Section 3.1 communicate with each other using AWA annotations.

The model follows a stand-off annotation schema, by means of which linguistic information attached to text anchors is represented separately using TEI-encoded typed feature structures (FS)[7]. Different FS types have been defined to represent different types of linguistic content. Among them, the type that represents the morphological analysis of a word form has the following features: the word form itself, its lemma, its variant lemma if it has one, top-level features (features of the word form as a whole), and a sequence of components used to properly represent intraword ellipsis (each component consists of a sequence of lemma parts and a sequence of other morphemes).

[7] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/FS.html (accessed: 2019-07-01).

```
analysis:
   lemma: oxigeno-hornitzaile
   variant: oxijeno-hornitzaile
   form: oxijeno-hornitzailea
   top-level-features: {Lemma-in-UBD: -, Variant-in-UBD: -,
     Lemma-Variant-Link-in-UBD: -, CAT: N, SUBCAT: C, ANIM: -,
     CAS: ABS, NUM: S, DEF: +, SFL: [@OBJ @SUBJ @PRED],
     COMP1: oxigeno, COMP2: hornitzaile, COMP: N+N}
   components:
      - lemma-parts:
          - entry: {spelling: oxigeno, STD-in-UBD: +}
            variant: {spelling: oxijeno, STD-in-UBD: -}
            link-in-UBD: -
        features: {CAT: N, SUBCAT: C, ANIM: -, ERROR-CODE: PHON}
          - entry: {spelling: -, STD-in-UBD: -}
            features: {CAT: HYPHEN}
          - entry: {spelling: hornitzaile, STD-in-UBD: +}
            features: {CAT: N, SUBCAT: C, ANIM: -}
        morphemes:
          - entry: {spelling: 0, STD-in-UBD: -}
            features: {CAT: INFL, NUM: S, DEF: +}
          - entry: {spelling: a, STD-in-UBD: -}
            features: {CAT: INFL, CAS: ABL, SFL: [@OBJ @SUBJ @PRED]}
```

Figure 2: YAML version (http://yaml.org/) of the FS corresponding to the morpho-syntactic analysis of *oxijeno-hornitzailea*

The FS in Figure 2 corresponds to the analysis of the non-lexicalized compound *oxijeno-hornitzailea*[8]. As explained in Section 3.5, this is a variant form of the standard *oxigeno-hornitzailea*, whose lemma is *oxigeno-hornitzaile* (please note that in the FS the variant lemma —*oxijeno-hornitzaile*— is also explicitly specified). The example represents the morpheme structure of the word: three lemma morphemes (the hyphen is also taken as part of the lemma) and the inflection morphemes (*0*, conveying information on number and definiteness, and *-a*, absolutive case).

Regarding the top-level features, we can observe in the example that some values, such as the list of syntactic functions, case, number, and definiteness, among others, are promoted from their corresponding morphemes to the top-level feature set. In other cases, top-level values are the result of a more complex calculation. For instance, the fact that *oxijeno-hornitzailea* is a noun results from the fact that the word-form is composed of two nouns linked by means of a hyphen; and so on. Features whose names contain `in-UBD` indicate whether their corresponding lemma and/or variant exists (`Lemma-in-UBD`, `Variant-in-UBD`, and `STD-in-UBD`) or whether they are explicitly interlinked in the `UBD` (`Lemma-Variant-Link-in-UBD` and `link-in-UBD`).

[8] In the example we use the more compact YAML notation for the sake of readability.

4.2 Design and implementation of the word structure grammar

We have redesigned and reimplemented the grammar developed by Aduriz *et al.* (2000) to address the problems described in Section 3. We chose a basic unification formalism, the PATR formalism (Shieber, 1986), for the definition of the morphosyntactic rules. There were two main reasons for this choice: (1) the formalism, being based on unification, is adequate for the treatment of complex phenomena (e.g., agreement of constituents in case, number, and definiteness) and complex linguistic structures and constraints, as is our case; and (2) simplicity.

As we stated in the introduction, our proposal separates sequential morphotactics (that is, which sequences of morphemes can or cannot combine with each other to form valid words), which will be recognized by the two-level system by means of continuation classes, from non-sequential morphotactics such as word-internal long-distance dependencies that are controlled by the word structure grammar. In fact, they make explicit the constraints specified implicitly as continuation classes in the two-level system.

The new grammar for Basque consists of 43 rules, compared to 25 rules in Aduriz *et al.* (2000). The grammar tackles the issues described in Section 3 and extends the linguistic descriptions of varying phenomena, especially with respect to robustness, in order to be able to analyze big corpora. There are 11 rules that deal with the merging of case, number, and definiteness morphemes and their combination with the main categories, 26 rules for the description of verbal subordination morphemes, 2 general rules for derivation (one for affixes and another one for suffixes), 2 rules for composition, 2 for conditional affixes, 1 rule for internal word ellipsis, and another one to deal with the degree of comparison of adjectives (comparative and superlative). The present work started from the design of the lexical database and continued to an overall rewriting of the grammar in order to cope with the new features, especially with respect to the treatment of composition, derivation, and variants. The general linguistic principles used to define unification equations in the rules are the following:

1. Information conveyed by the lemma. The main category and semantic features are promoted from the lemma.
2. Information coming from case suffixes. Case suffixes provide information on case, number, and syntactic function. For example, the singular ablative case is given by the suffix *-tik* in *bihotz+tik* (*bihotzetik*, 'from the heart').
3. Noun ellipsis. When an intraword nominal ellipsis occurs, the part of speech of the whole word is expressed by a composed category, which indicates the presence of the ellipsis and the category of the word. This way, while *mendi+a* corresponds to 'the mountain', adding an ellipsis null morpheme (*Ø*) after a genitive morpheme (*-ko*), as in *mendi+ko+Ø+a*, means '(the one) of the mountain', introducing an elliptical element ('the one') that corresponds to a noun.
4. Subordination morphemes. When a subordination morpheme is attached to a verb, the verb and its features are promoted as well as the subordinate relation and the syntactic function conveyed by the morpheme.

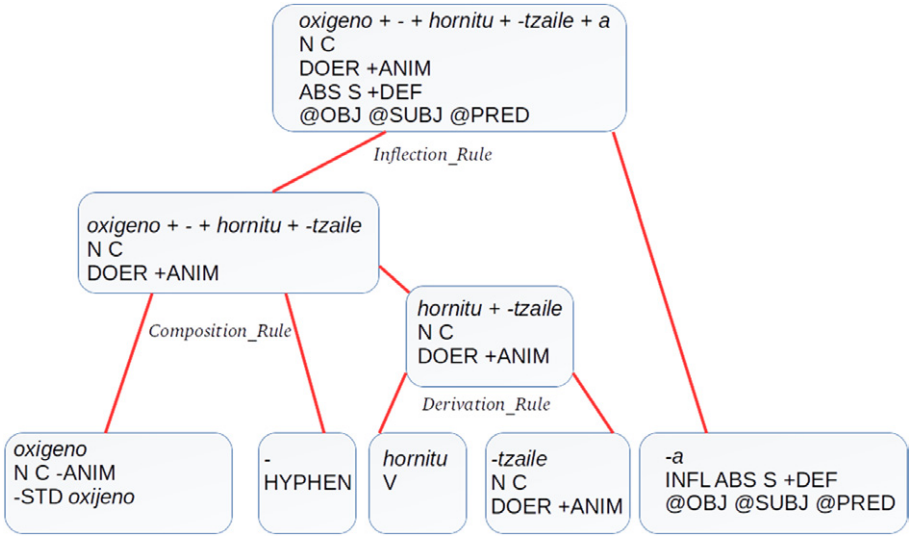

Figure 3: Analysis of *oxijeno-hornitzailea* (variant + compound + derivative)

5. Degree morphemes. They are attached to adjectives, past participles, and adverbs.
6. Derivation. Derivational suffixes select the category of the base form, and in most cases they make the derivative belong to another category. For instance, the suffix *-tzaile* ('doer') is applied to verbs and the derived word is a noun (see *hornitzaile*, 'supplier' in Figure 3). The analysis also contains information about the standard dictionary entry corresponding to the base form.
7. Composition. Basque is a right-headed language, so the main information of the compound is taken from the rightmost element. The analysis in this case contains information about the standard dictionary entries corresponding to the base form.
8. Treatment of variants and non-standard forms. Apart from the variants of basic word forms, variants can interfere with all the standard linguistic phenomena, as inflection, derivation, and composition. For example, there could be non-standard lemmas, but variants could also happen as dialectal inflectional morphemes.

Table 4 presents a description of the word structure grammar, with examples of every set of rules, classified according to their function. This grammar gives a complete description of Basque morphology, and takes into account regular and irregular morphological phenomena, as well as the treatment of variants and non-standard phenomena. Figure 4 presents a simplified version of a rule, where unification equations describe the constraints and relations among the set of features employed. Although this rule only gives an account of the main characteristics of the morphological phenomenon described, the actual rules are more complex, detailing all the linguistic processes that are involved (general context for rule application, exceptions and feature interactions).

Table 4: Description of the word structure grammar

Phenomena	Rule patterns	Examples
Inflection for all categories (noun, pronoun, adjective, determiner and adverbs): **14 rules**.	X + morpheme (where X in {noun/adj/adv/verb/pron})	*gizon* + *-a* → *gizona* man + the → the man *mendi* + *-eta* + *-raino* → *mendietaraino* mountain + the (PL) + up to → up to the mountains
Verbal paradigms: **23 rules**.	verb + morpheme	*egin* + *-tea* → *egitea* do + -ing → doing *ekar* + *-tze* + *-agatik* → *ekartzeagatik* bring + having + for → for having brought
Derivation: **2 rules**.	X + derivational morpheme	*hornitu* + *-tzaile* → *hornitzaile* provide + DOER→ provider
Composition: **2 rules**.	X + X (where X is {noun/verb})	*zine* + *egile* → *zine-egile* movie + maker → movie-maker
Gradation (adjective, determiner, adverb, noun, pronoun and verbs): **1 rule**.	X + morpheme (where X is {noun/adj/adv/verb/pron})	*handi* + *ago* → *handiago* big + -er → bigger
Ellipsis (all categories): **1 rule**.	X + genitive + ellipsis-morpheme (where X is {noun/adj/adv/verb/pron})	*mendi*+ *-ko* + *-a* → *mendikoa* mountain + of + the → (the one) of the mountain *dator*+ *-en* + *-a* → *datorrena* comes + that + one → (the one) that comes

4.3 A complex example

Figure 3 shows the morphosyntactic analysis of *oxijeno-hornitzailea* (see Example 5), which exemplifies the interaction of several phenomena. The first one corresponds to a typical error consisting of the interchange of the letters *g* and *j*, giving *oxijeno* from the standard base form *oxigeno*. The word also contains the formation of a derived word,

```
rule INFLECTION1 X0 ---> X1 X2
% lexical plural, e.g. scissors = scissors + s(+plural)
        X1/feats/category      =        noun
        X1/feats/subcategory   not in   [person-name, loc-name]
        X2/feats/category      =        infl
        X2/feats/num           =        p
        X1/feats/plu           =        "+"
        X0/form                =        X1/form
        X0/feats/category      =        X1/feats/kat
        X0/feats/subcategory   =        X1/feats/azp
        X0/feats/plu           =        X1/feats/plu
        X0/feats/case          =        X2/feats/case
        X0/feats/num           =        X2/feats/num
        X0/feats/definiteness  =        X2/feats/definiteness
        X0/lemma               =        X1/lemma
        ...
```

Figure 4: Example rule (X0 = parent, X1 = left child, X2 = right child)

hornitzaile ('provider'), coming from *hornitu* ('to provide') and the derivational suffix *-tzaile*[9]. Composition and inflection give the final word form.

5. Corpus analysis of the data on morphosyntactic phenomena

In this section we describe a corpus-based analysis of the morphosyntactic phenomena presented above. Firstly, we will describe in detail the analysis of the phenomena described in the previous sections based on EPEC, the manually processed corpus, and then we will give some figures taken from a larger corpus, to see to what extent the distributions of the phenomena are similar to those observed in EPEC.

Firstly, Table 5 shows the actual number of tokens analyzed based on the rules and lexicons defined for derivation and composition, and on those for the hybridization of both, along with their evaluation. It can be seen that they account for 2.01% of the tokens, most of them are analyzed using derivation, and some of them present both phenomena together.

Table 6 shows that derivation is mainly observed in the segmentation of standard words, having a residual impact on OOV words. Having a derivational interpretation does not necessarily mean that this segmentation is the one selected for disambiguation purposes: (1) there might be an alternative lexicalized interpretation, which would be preferred: *aniztasun* (Basque for 'diversity') *versus anitz* (Basque for 'many') + *-tasun* (a noun-forming suffix denoting quality); or (2) due to overgeneration, there might be interpretations that have nothing to do with this phenomenon: OOV proper names such as *Baztarrika* and *Irazoki* can be segmented as *baztarri+ka* and *irazo+ki*, because of the

[9] In this example, the derivative *hornitzaile* is not taken as a lexicalized lexeme, but analyzed as the combination of the verb *hornitu* and the suffix *-tzaile*.

Table 5: Summary of the distribution of the phenomena in the EPEC test corpus

	Tokens	Percentage in corpus	Recall
Derivation	1,044	1.60	97.41
Composition	263	0.40	95.82
Hybridization	18	0.03	100.00
Total/Average	1,307	2.01	97.02

Table 6: Derivatives on the EPEC test corpus

	Tokens	Correct	Recall
Standard	1,008	993	98.51
Variant	27	23	85.19
OOV	9	2	22.22
Average	1,044	1,018	97.41

short length of the affixes, although the analyses based on derivation are not, in this case, the correct ones.

In addition, most of the OOV words contain very short affixes, such as the adverb-producing *-ki* and *-ka*. Many of the occurrences studied in EPEC using these affixes are adverb+adverb compounds, usually repeated adverbs used to emphasize their meaning (reduplication). For instance, *presaka* (Basque for 'quickly') may be repeated and hyphenated, as in *presaka-presaka*, meaning 'very quickly'. Unfortunately, although the use of a derivational affix in both elements of a compound is foreseen, this kind of compound (adverb+-+adverb) is not described by the morphotactics, hence the system is not able to segment them adequately (*presaka+-+presaka*). Instead, the analyzer makes use of the derivation mechanism to analyze them as OOV words, giving an adverb interpretation as one of the possible readings. For the example at hand, the segmentation is *presaka-presa+ka*, assigning an adverb analysis and *presaka-presa* as lemma. Therefore, short affixes seem to be prone to errors and we should be very careful when it comes to applying them when analyzing OOV words.

On the other hand, we have extracted the analyses produced based on derivation in order to measure its contribution to the process. As we have said, some of the words are lexicalized (see Table 7) and have their own entry in the lexicon, therefore having analyses equivalent to those using derivational affixes. In our corpus they comprise 58% of the words (608 out of 1,044). Unfortunately, other words have no alternative equivalent option; in particular, 391 tokens in our corpus (see rows 2 and 3 in Table 7). Among them, the main group (quantifiers in Table 7) is composed of ordinal and

Table 7: Evaluation of derivation on the EPEC test corpus

	Tokens	# of correct with derivation	# of correct without derivation
Lexicalized	608	608	608
Quantifiers	287	287	0
Other non-lexicalized	104	78	0
Other correct readings	45	45	45
Total	1,044	1,018	653

distributive quantifiers, which are formed from the affixes *-garren*[10] and *-na*[11] respectively. It is important to highlight that only derivation can produce correct segmentations of these quantifiers.

Quite the opposite applies to most of the other cases of non-lexicalized derived words, to which our system assigns derivation-based readings, among others (other non-lexicalized in Table 7). Some of them, namely nouns and adjectives, might be analyzed using the OOV module, even though they could be assigned a larger number of interpretations, as we have seen before. Bearing in mind that this module only uses open categories in segmentation, generating the correct analysis is not assured for this subset of words. Finally, it is remarkable that only 45 words, recognized by applying derivational rules, have another different reading in context, although derivation is also possible.

To summarize, our approach contributes to the correct processing of an open set of words, namely quantifiers, which cannot be included one by one in the lexicon or added to the OOV processor without dramatically increasing the number of interpretations assigned to them. In addition, it can be said that it handles the most productive affixes with very good results both in coverage (there is no manually added derivational interpretation in the corpus) and in recall, since only 2.5% of the words are incorrectly analyzed: (1044–1018) out of 1044. It goes without saying that short affixes represent a difficult challenge for the segmentation of OOV words.

Regarding composition, this phenomenon is less frequent than derivation. Table 8 shows actual figures taken from EPEC. The errors detected during the evaluation are mainly due to one of the following reasons: (1) emphatic hyphenated duplication of adverbs and adjectives analyzed as N+N compounds; or (2) the remaining incorrectly analyzed words require the redefinition of some continuation classes and/or

[10] For instance, *laugarrena*, Basque for 'the fourth one': *lau* ('four') + *-garren* (ordinal) + *-a* ('the', determiner).

[11] For instance, *launa*, Basque for 'four each': *lau* ('four') + *-na* (distributive).

Table 8: Evaluation of composition on the EPEC test corpus

	Tokens	Correct	Recall
Standard	212	207	97.64
Variant	6	5	83.33
OOV	45	41	91.11
Average	263	253	96.21

Table 9: Evaluation of hybridization on the EPEC test corpus

	Tokens	Correct	Recall
Standard	17	17	100.00
Variant	1	1	100.00
Average	18	18	100.00

morphophonological rules for OOV in order to be able to assign the correct interpretations.

To complete the evaluation of EPEC, Table 9 shows the figures for hybrids. There are very few examples in this corpus, but, as detailed in Table 9, all of them are correctly treated. Additionally, we can say that most of them have an equivalent lexicalized interpretation.

As mentioned at the beginning of this section, we have also analyzed a larger corpus to confirm that the distribution of the phenomena remains in comparable proportions when the size of corpus is drastically increased. For that purpose we have used the *Observatory of the Lexicon* corpus (OLC)[12] with 33.1 million tokens (27.1 million words). This corpus has not been manually revised and, therefore, the measures are given only as a general view of the phenomena presented in this paper.

The texts in OLC have been collected from diverse sources, mainly journalistic. Table 10 shows the actual distribution of word forms in the corpus. The main sources are *Berria*, the Basque newspaper (http://berria.eus), which contributes almost 40% of the tokens, and the website of the Basque Public TV (http://eitb.eus), which provides nearly 27%. Both sources use style manuals including the latest decisions of the Basque Academy and, therefore, they are considered good sources for the standard language. The rest of the sources are far smaller, and they increase variability in the corpus, as they do not necessarily follow so strictly the standardization recommendations and norms of the Academy.

[12] http://lexikoarenbehatokia.euskaltzaindia.net/cgi-bin/kontsulta.py (accessed: 2019-07-01).

Table 10: Distributions of tokens in OLC among sources and according to the lexicon used to analyze them

Source	Tokens	Percentage in OLC	Variants (%)	OOV (%)
Berria	13,218,363	39.86	0.49	3.46
eitb.eus	8,980,455	27.08	0.85	5.72
Argia	3,108,010	9.37	0.45	2.25
DiarioVasco	2,266,213	6.83	1.30	4.53
ETB documentaries	1,754,479	5.29	0.48	4.94
Consumer	1,395,147	4.21	0.29	1.85
Jakin	1,089,530	3.29	0.74	3.94
Elhuyar	587,171	1.77	0.24	3.00
Sustraia	362,231	1.09	0.26	1.43
Deia	278,169	0.84	1.79	8.35
Kresala	107,991	0.33	1.09	4.04
Chiloe	14,688	0.04	1.03	3.91
Totals/Average	33,162,447	100.00	0.64	4.07

Table 10 also shows the distribution of non-standard words according to the source of texts. They amount to 4.71% of tokens on average (variants + OOV), ranging from 1.69% (*Sustraia*) to 10.14% (*Deia*) depending on the sources. However, we have to say that both ends of the range correspond to small subsets of the corpus.

With regard to variants, they constitute only 0.64% of the tokens on average, which might seem residual or insignificant. However, as the morphological segmentation for OOVs only considers open categories and generic lemmas, skipping the treatment of variants might imply the assignment of incorrect interpretations. Moreover, as we have seen before in this section, variants may also contain derivation and composition in their morphological structure that might not be taken into account if they are not correctly segmented, especially if the correct POS is missing.

Table 11 compares the proportions of tokens in OLC and in EPEC that have been analyzed through derivation, composition, and the hybridization of both. In general, it can be said that the proportions remain almost the same across corpora. Concerning composition, this phenomenon is less frequent in OLC than in EPEC. This is an expected result as the hyphen in compounds is optional (non-hyphenated compounds are not recognized as compounds by Morfeus+).

With respect to derivation, we have computed the number of words that are analyzed using both derivation and their lexicalized information (lexicalized in Table 12). As we have seen for EPEC that ordinal and distributive quantifiers are only correctly analyzed

Table 11: Summary of the distribution of the phenomena in OLC and EPEC

	Tokens	Percentage in OLC	Percentage in EPEC (Table 5)
Derivation	543,626	1.64	1.60
Composition	94,320	0.28	0.40
Hybridization	8,170	0.02	0.03
Total/Average	33,162,447	1.95	2.01

Table 12: Evaluation of derivation in OLC

	Tokens	Tokens with no derivation
Lexicalized	361,146	361,146
Quantifiers	114,126	0
Total	475,272	361,146

through derivation, we have also included them in Table 12. Even if the rest of the words analyzed using derivation were considered to be wrongly treated, the results show that the number of words that cannot be properly analyzed in another way is not negligible. Moreover, quantifiers play a crucial role in syntax. If the morphosyntactic analyzer was not capable of assigning the correct interpretations to this type of word, the syntactic analyzer in the processing chain would not be able to build the appropriate parse.

Finally, we have examined the hybridization of the three phenomena, namely composition, derivation, and variants. Less than 10% of the compounds (8,170 out of 94,320) appear combined with derivatives, that is, where derivation is present in one or more elements of the compound. Even though composition is not a high-frequency phenomenon (0.28% of the tokens in OLC, see Table 11), it increases the complexity of the analysis significantly. In addition, there are some examples in which the three phenomena are present in the same word, among which we can find the one described in Example 5.

6. Conclusions and future work

We have presented the design and development of a system for morphosyntactic analysis of a morphologically rich language, Basque, which combines a complete linguistic formalization of the phenomena involved in the formation of words, including a large-scale lexicon, with a robust data-processing tool.

Our two main contributions are: first, (1) our system gives a comprehensive linguistic description of the main morphological phenomena, such as affixation, derivation, and composition. This description takes into account the modeling of standard and out-of-vocabulary words, including dialectal and orthographical variants, and their

linking, when appropriate, to their corresponding standard entries in a lexical database. Altogether, this gives a wide-coverage linguistic specification of Basque morphology, from a theoretical as well as from an applied point of view. Secondly, (2) the linguistic specifications have been implemented in Morfeus+, a tool capable of analyzing unrestricted texts. We have tested the applicability of the tool on a big corpus of varied genres, and the tool has been used for the analysis of high volumes of text, showing that its coverage is wide and robust and allowing the efficient processing of large volumes of data. The number of OOV words amounts to approximately 4% of the tokens on average in OLC ranging from 1.43% to 8.35% depending on the type of text (see Table 10), and this fact implies that, unless there is an explicit commitment to deal with OOV words, any posterior process such as syntactic parsing, named-entity recognition, or semantic processing will suffer the problem of unknown words, showing a poor coverage and harming the usability of any language processing tool.

Overall, the system presented has incurred a significant workload in both research and development, which can be distributed equally on both the linguistic side (design and development of grammar and lexicons) and the software-engineering side (finite-state implementation, unification-based parser, annotation and representation issues, and corpus processing tool).

Regarding the portability of the present approach to other languages, we think that the overall design and architecture of our solution can be inspiring and adapted to other languages, especially to those that share many features with Basque, such as complex morphology and agglutination. However, obtaining a robust and high-performance tool will also require a good deal of work defining lexicons and the word grammar that describes the language in question.

To summarize, we can state that a comprehensive linguistic morphological description together with a robust implementation are the keys to have a working tool capable of successfully dealing with unrestricted written texts, not only tackling a full linguistic description of Basque morphology, but also presenting solutions for practical aspects related to robustness on a working implementation. Other types of implementations that mainly try to incorporate vast lexica can be effective with languages of simple morphology, but these models suffer lack of coverage when applied to morphologically richer languages. For this reason, the system presented can be a model for a wide set of languages whose characteristics are far from mainstream languages such as English or Spanish. Although an important part of the work is intrinsically related to the processing of a specific language, we think that the general model and architecture can be of interest to many researchers and developers of basic-language processing tools for different languages.

References

Aduriz, Itziar & Agirre, Eneko & Alegria, Iñaki & Arregi, Xabier & Arriola, Jose Mari & Artola, Xabier & Díaz de Ilarraza, Arantza & Ezeiza, Nerea & Maritxalar, Montse & Sarasola, Kepa & Urkia, Miriam. 1993. A morphological analysis based method for spelling correction. In

Proceedings of the European Association for Computational Linguistics, EACL 93. Utrecht, Netherlands.

Aduriz, Itziar & Agirre, Eneko & Aldezabal, Izaskun & Alegria, Iñaki & Arregi, Xabier & Arriola, Jose Mari & Artola, Xabier & Gojenola, Koldo & Sarasola, Kepa & Urkia, Miriam. 2000. A word-grammar based morphological analyzer for agglutinative languages. *Proceedings of the International Conference on Computational Linguistics. COLING 2000, 1–7*. Saarbrücken, Germany.

Aduriz, Itziar & Aranzabe, María Jesús & Arriola, Jose Mari & Atutxa, A. & Díaz de Ilarraza, Arantza & Ezeiza, Nerea & Gojenola, Koldo & Oronoz, Maite & Soroa, Aitor & Urizar, Ruben 2006. Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing. In Wilson, Andrew & Rayson, Paul & Archer, Dawyin (eds.), *Corpus linguistics around the world* (Language and computers, vol. 56), 1–15. Amsterdam: Rodopi.

Aldezabal, Izaskun & Ansa, Olatz & Arrieta, Bertol & Artola, Xabier & Ezeiza, Aitzol & Hernández, G. & Lersundi, Mikel 2001. EDBL: A general lexical basis for the automatic processing of Basque. In *IRCS Workshop on Linguistic Databases*. Philadelphia, USA.

Alegria, Iñaki & Artola, Xabier & Sarasola, Kepa & Urkia, Miriam. 1996. Automatic morphological analysis of Basque. *Literary & Linguistic Computing* 11(4). 193–203.

Alegria, Iñaki & Aranzabe, Maxux & Ezeiza, Aitzol & Ezeiza, Nerea & Urizar, Ruben 2002. Using finite state technology in natural language processing of Basque. In Watson, Bruce & Wood, Derick (eds.), *LNCS: Implementation and application of automata* 2494, 1–11.

Alkorta, Jon & Gojenola, Koldo & Iruskieta, Mikel 2018. SentiTegi: Building a semantic oriented Basque lexicon. In *Computación y sistemas* 22(4). 1295–1306.

Antworth, Evan L. 1994. Morphological parsing with a unification-based word grammar. In *North Texas Natural Language Proceesing Workshop*. Austin, TX: The University of Texas at Austin.

Aronoff, Mark & Fudeman, Kirsten 2011. *What is morphology?* Malden, MA: Wiley-Blackwell.

Artola, Xabier & Díaz de Ilarraza, Arantza & Soroa, Aitor & Sologaistoa, Aitor. 2009. Dealing with complex linguistic annotations within a language processing framework. In *IEEE Transactions on Audio, Speech and Language Processing* 17(5). 904–915.

Artetxe, Mikel & Labaka, Gorka & Saedi, Chakaveh & Rodrigues, João & Silva, João & Branco, António & Agirre, Eneko. 2016. Adding syntactic structure to bilingual terminology for improved domain adaptation. In *Proceedings of the 2nd Deep Machine Translation Workshop (DMTW 2016)*, 39–46. Lisboa, Portugal.

Bender, Emily M. 2013. *Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax*. San Rafael, CA: Morgan & Claypool.

Bengoetxea, Kepa & Iruskieta, Mikel 2018. A supervised central unit detector for Spanish. *Procesamiento del Lenguaje Natural* 60. 29–36.

Khaliq, Bilal & Carroll, John. 2013. Unsupervised induction of Arabic root and pattern lexicons using machine learning. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, 350–356. Hissar, Bulgaria.

Cotterell, Ryan & Kirov, Christo & Sylak-Glassman, John & Walther, Géraldine & Vylomova, Ekaterina & Xia, Patrick & Faruqui, Manaal & Kübler, Sandra & Yarowsky, David & Eisner, Jason & Hulden, Mans 2017. CoNLL SIGMORPHON 2017 Shared Task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, 1–30. Vancouver, Canada.

El-Haj, Mahmoud & Kruschwitz, Udo & Fox, Chris 2014. Creating language resources for under-resourced languages: Methodologies, and experiments with Arabic. *Language Resources and Evaluation* 49(3). 549–580.

Euskaltzaindia. 1987. *Hitz-elkarketa/1*. Bilbao. https://www.euskaltzaindia.eus/dok/iker_jagon_tegiak/6913.pdf

Euskaltzaindia. 1991. *Hitz elkarketa/3*. Bilbao. https://www.euskaltzaindia.eus/dok/iker_jagon_tegiak/11582.pdf

Euskaltzaindia. 2000. *Hiztegi Batua*. Euskera 5. http://www.euskaltzaindia.net/hiztegibatua/.

Euskaltzaindia. 2014. *Hitz elkarketa/2*: Bilbao. https://www.euskaltzaindia.eus/dok/iker_jagon_tegiak/77118.pdf

Euskara Institutua. S.a. *Hitz eratorpena*, Sareko Euskal Gramatika. http://www.ehu.eus/seg.

Ezeiza, Nerea & Aduriz, Itziar & Alegria, Iñaki & Arriola, Jose Mari M. & Urizar, Ruben 1998. Combining stochastic and rule-based methods for disambiguation in agglutinative languages. In *Proceedings of the International Conference on Computational Linguistics and Annual Conference of the Association for Computational Linguistics COLING-ACL'98*, 380–384. Montreal, Canada.

Foster, Jennifer. 2010. "cba to check the spelling": Investigating parser performance on discussion forum posts. In *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 381–384. Los Angeles, CA.

Foster, Jennifer & Çetinoglu, Özlem & Wagner, Joachim & Le Roux, Joseph & Hogan, Stephen & Nivre, Joakim & Hogan, Deirdre & Van Genabith, Josef. 2011. #hardtoparse: POS tagging and parsing in twitterverse. In *Proceedings of the 5th AAAI Conference on Analyzing Microtext*, 20–25. San Francisco, USA.

Goldberg, Yoav & Elhadad, Michael. 2013. Word segmentation, unknown-word resolution, and morphological agreement in a Hebrew parsing system. *Computational Linguistics* 39(1). 121–160.

Gonzalez-Dios, Itziar & Aranzabe, María Jesús & Díaz de Ilarraza, Arantza. 2018. The corpus of Basque simplified texts (CBST). *Language Resources and Evaluation* 52. 217–247.

Goodfellow, Ian & Bengio, Yosua & Courville, Aaron. 2017. *Deep learning*. Cambridge, MA: The MIT Press.

Hajič, Jan. 2004. *Disambiguation of rich inflection (computational morphology of Czech)*. Prague: Charles University Press.

Haspelmath, Martin & Sims, Andrea. 2010. *Understanding morphology*. London: Routledge.

Haverinen, Katri & Nyblom, Jenna & Viljanen, Timo & Laippala, Veronika & Kohonen, Samuel & Missilä, Anna & Ojala, Stina & Salakoski, Tapio & Ginter, Filip. 2014. Building the essential resources for Finnish: The Turku Dependency Treebank. *Language Resources & Evaluation* 48. 493.

Hualde, José Ignacio & Zuazo, Koldo. 2007. The standardization process of the Basque language. *Language Problems and Language Planning* 31(2). 143–168.

Karlsson, Fred & Voutilainen, Atro & Heikkilä, Juha & Anttila, Arto. 1995. *Constraint Grammar: A language-independent framework for parsing unrestricted text*. Berlin: Mouton de Gruyter.

Karttunen, Lauri. 1994. Constructing lexical transducers. In *Proceedings of the International Conference on Computational Linguistics, Coling 1994*, 406–411. Kyoto, Japan.

Klein, Dan & Manning, Christopher D. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 423–430. Sapporo, Japan.

Koskenniemi, Kimmo. 1983. *Two-level morphology: A general computational model for word-form recognition and production*. Helsinki: University of Helsinki Department of General Linguistics.

Laka, Itziar. 1996. *A brief grammar of euskara, the Basque language*. Euskararako Errektoreordetza, UPV/EHU. https://www.ehu.eus/eu/web/eins/a-brief-grammar-of-euskara

Lee, Jackson L. & Goldsmith, John A. 2016. Linguistica 5: Unsupervised learning of linguistic structure. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 22–26. San Diego, CA.

Manterola, Julen. 2008. Is Basque an agglutinative language? A proposal for the diachrony of nominal morphology. In *Basque Studies Symposium*, May 2008, Santa Barbara, United States.

Oflazer, Kemal. 1999. Dependency parsing with an extended finite state approach. In *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics*, College Park, MD. Published in: *Computational Linguistics* 19(4). 2003. 515–544.

Ormazabal, Javier. 1992. On the structure of complex words: The morphology-syntax interplay. *Anuario del Seminario de Filología Vasca "Julio Urquijo"* 26(3). 725–765.

Otegi, Arantxa & Imaz, Oier & Díaz de Ilarraza, Arantxa & Iruskieta, Mikel & Uria, Larraitz. 2017. ANALHITZA: A tool to extract linguistic information from large corpora in humanities research. *Procesamiento del Lenguaje Natural* 58. 77–84.

Padró, Lluís & Stanilovsky, Evgeny. 2012. FreeLing 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, 2473–2479. Istanbul, Turkey.

Perez-de-Viñaspre, Olatz & Oronoz, Maite & Elvira, Natalia. 2018. KabiTermICD: Nested term based translation of the ICD-10-CM into a minor language. In *MultilingualBIO: Multilingual Biomedical Text Processing*. LREC Workshop, Miyazaki, Japan.

Ritchie, Graeme D. & Pulman, Stephen G. & Black, Alan W. & Russell, Graham J. 1992. *Computational morphology: Practical mechanisms for the English lexicon*. Cambridge, MA: The MIT Press.

Şahin, Muhammet & Sulubacak, Umut & Eryiğit, Gülşen. 2013. Redefinition of Turkish morphology using flag diacritics. In *Proceedings of The Tenth Symposium on Natural Language Processing (SNLP-2013)*. Phuket, Thailand.

Salaburu, Pello & Alberdi, Xabier. (eds.). 2012. *The challenge of a bilingual society in the Basque Country*. Reno, NV: Center for Basque Studies, University of Nevada.

Sak, Haçim & Güngör, Tunga & Saraçlar, Murat. 2011. Resources for Turkish morphological processing. *Language Resources & Evaluation* 45. 249.

Scalise, Sergio & Vogel, Irene. (eds.). 2010. *Cross-disciplinary issues in compounding*. Amsterdam: Benjamins.

Seeker, Wolfgang & Kuhn, Jonas. 2013. Morphological and syntactic case in statistical dependency parsing. *Computational Linguistics* 39(1). 23–55.

Shieber, Stuart M. 1986. *An introduction to unification-based approaches to grammar*. Stanford, CA: CSLI.

Spoustová, Drahomíra & Hajič, Jan & Votrubec, Jan & Krbec, Pavel & Květoň, Pavel. 2007. The best of two worlds: Cooperation of statistical and rule-based taggers for Czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*, 67–74. Prague, Czech Republic.

Trost, Harald. 1990. The application of two-level morphology to non-concatenative German morphology. In *Proceedings of the 13th International Conference on Computational Linguistics, COLING'90*, 371–376. Helsinki, Finland.

Tsarfaty, Reut & Seddah, Djamé & Kübler, Sandra & Nivre, Joakim. 2013. Parsing morphologically rich languages. *Journal of Computational Linguistics* 39(1). 15–22.

Toutanova, Kristina & Manning, Christopher D. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. EMNLP/VLC*, 63–70. Hong Kong, China.

Würzner, Kay-Michael & Hanneforth, Thomas. 2013. Parsing morphologically complex words. In *Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing*, 39–43. Saint Andrews, Scotland.

Zuñiga Fernando & Fernández, Beatriz. 2019. Grammatical relations in Basque. In Witzlack-Makarevich, Alena & Bickel, Balthasar (eds.), *Argument selectors: A new perspective on grammatical relations*, 185–211. Amsterdam: Benjamins.