A morphological processor based on foma for Biscayan (a Basque dialect)

Iñaki Alegria[1], Garbiñe Aranbarri[2], Klara Ceberio[3], Gorka Labaka[1], Bittor Laskurain [2], Ruben Urizar[1]

[1] IXA research group. University of the Basque Country. 649 PK 20080 Donostia
[2] Eleka S.L. Zelai Haundi kalea, 3 - Osinalde industrialdea - 20170 Usurbil
[3] Elhuyar Foundation. . Zelai Haundi kalea, 3 - Osinalde industrialdea - 20170 Usurbil
E-mail: i.alegria@ehu.es

Abstract

We present a new morphological processor for Biscayan, a dialect of Basque, developed on the description of the morphology of standard Basque. The database for the standard morphology has been extended for dialects and an open-source tool for morphological description named *foma* is used for building the processor. *XuxenB*, a spelling checker/corrector for this dialect, is the first application of this work.

1. Basque and Biscayan morphology

Basque is an agglutinative language with rich morphology. Standard Basque morphology was described by Alegria et al. (1995; 2002) using finite-state morphology.

The Biscayan dialect of Basque (Arejita et al., 2002, 2005), also called Western Basque (Zuazo, 2008), is a dialect of the Basque language spoken in the western part of the Basque speaking area, mainly in the province of Biscay, but also in southwest Guipuzcoa and the Basque speaking areas of Álava. Although it is the most widespread of Basque dialects, it differs considerably from standard Basque, heavily based on the Guipuzcoan dialect.

While the standard written Basque is used in all the levels of education and the media intended for the whole Basque speaking population, there is an increasing interest in using Biscayan in local media and informal communication forums (blogs, chats, phones). It is quite difficult to calculate the number of Biscayan speakers, as the majority of surveys done regarding the use of Basque language focus mainly on Basque speakers as a whole. Nevertheless, according to the estimate made by the Labayru institute (www.labayru.org), a cultural institution working for the promotion of Biscayan, speakers of this dialect amount to about 250,000.

2. *foma*

foma (Hulden, 2009) is a finite-state tool designed for multi-purpose use with explicit support for automata theory research, constructing lexical analyzers for programming languages, and building morpho-phonological analyzers, as well as spell-checking applications.

The compiler allows users to specify finite-state automata and transducers incrementally in a similar fashion to AT&T's *fsm*, Sproat's *Lextools*, the Xerox/PARC finite-state toolkit and the SFST toolkit. One of *foma*'s design goals has been compatibility with the Xerox/PARC toolkit (Beesley and Karttunen, 2003).

foma is licensed/certified under the GNU license: in keeping with traditions of free software, the distribution that includes the source code comes with a user manual and a library of examples. The compiler and library are implemented in C and an API is available (foma. sourceforge.net).

Standard Basque morphology was migrated to open-source finite-state technology using *hunspell* (Nemeth et al., 2004) first and then to *foma* (Alegria et al., 2009). In the last case, it is worth pointing out the process of rewriting parallel two-level phonological rules into sequential replacement rules. Due to the fact that in the context of the sequential rules it is not necessary to take in account the rest of the rules, the use of sequential replacement rules increases flexibility in the description, which allows to include linguistic variants. In contrast, special attention must be taken to place the rules in the correct order..

```
# parallel rules
R->r ; +->0 ;
h->0 || # b e R:r 0:r +:0 _ Vowel;
0->r || R:r _ +:0 (h:0) Vowel;
e->i || # b _ R +:0 Cons;

# sequential rules
h->0 || # b e R + _ Vowel;
0->r || R _ + Vowel;
e->i || # b _ R + Cons;
R->r ; +->0 ;
```

Table 1. Simplified comparison between parallel and sequential rules in Basque morphology.

In Table 1 we provide an example of the differences between both descriptions. The example rules illustrate a number of phenomena. In Basque, the prefix *ber-* (equivalent to English 're-') mostly precedes verbal stems. Before a vowel, the final character of this prefix (*r*) is doubled; this is expressed by the capital `R` (`beR`).

When preceding a consonant the prefix *ber-* changes into *bir-*. In addition to this, if the first character of the verb stem is *h*, this *h* is erased. Thus, the chain *beR+egin* (lexical expression) generates *berregin* 'redo', while *beR+gai+tu* generates *birgaitu* 'rehabilitate', and *beR+has+i* changes into *berrasi* 'restart'.

3. The database for Biscayan

The description of the lexicon and the morphotactics (or word grammar) for standard Basque was carried out using a relational database (Aduriz et al., 1998). Currently, around 100,000 entries are recorded for the description of standard Basque, divided into (a) lexical entries (b) inflectional and derivational morphemes and (c) inflected verb forms. The lexical description needed for the morphological analysis of text words is obtained using an export process from the database.

The database for the Biscayan dialect was built based on the one for standard Basque. When new dialectal variants were added, they were linked to the standard entries, which now became non-standard variants in the dialectal database. For example, the Biscayan verb form *gagoz* ('we are/stay') is stored in the Biscayan database linked to *gaude*, its corresponding entry in standard Basque, which, at the same time, is non-standard in the dialect. These links will be exploited when we want to obtain the standard form from the dialectal one or vice versa. This is very useful for indexation, spelling correction and other applications.

Most of the new entries (1,661) correspond to inflected verb forms, since the main difference between both variants of the language are auxiliary verbs. For instance, the triadic auxiliary verb forms for indicative in Biscayan take the *-eu(t)s-* stem while in standard Basque they take *-i-* e.g. *ekarri deutso / dio* (brought have-it-to_him/her-he/she; 'she has brought it to her').

As for lexical entries (nouns, adjectives, adverbs, lexical verbs, determiners and pronouns), many of the standard words belong to the common core of the language. However, over 900 new lexical entries have been added to the database for Biscayan. Most of the variations are due to phonological phenomena such as *i/u* alternations (*huri / hiri* 'city', *gitxi / gutxi* 'little', *ule / ile* 'hair'), loss of intervocalic nasal in *–sio/-sino* or *-zio/-zino* suffix (*telebisio / telebisino* 'television', *akusazio / akusazino* 'accusation') or final *-a / -e* alternations (*lora / lore* 'flower', *laba / labe* 'oven', *hoba / hobe* 'better').

For those standard entries which do not have an equivalent in Biscayan, we decided to keep them in order to improve coverage. It is very common to find this kind of words in dialectal texts due to the increasing influence of standard Basque over the rest of the dialects especially in written texts.

Regarding inflectional morphemes, just a few changes had to be made. For the sociative case ('with'), the *-gaz* morpheme is used in the Biscayan dialect along with the standard *-rekin* (*amarekin / amagaz* 'with mum'), *-runtz* ('towards') is also acceptable together with the the standard *-rantz* (*etxerantz / etxeruntz* 'towards the house'), verb participles ending in *–atu* become *-au* in their dialectal forms (*errezau / errezatu* 'to pray') and the nominal form of these verbs takes *–eta* along with the standard *-t(z)e* (*errezetan / errezatzen* 'praying').

In order to develop the lexicon for the Biscayan dialect we used the standard version of the analyzer. This was applied on a list of 2775 words made up by the Labayru Institute from a list of dictionary entries and a small corpus written in the dialect. On the one hand, the standard analyzer could not identify 1984 correct Biscayan words in the list, and, on the other hand, 771 words in the list, which were not acceptable in the dialect, were analyzed as correct since there are acceptable in standard Basque.

4. The morphophonological rules for Biscayan

It has been possible to adapt the morphophonological rules for standard Basque to the ones for Biscayan using sequential replacement rules, without changing the original ones.

Ten new sequential rules have been added, most of them close to the lexicon (in the beginning of the rules system before the standard rules) and only one rule in the bottom part of the rules system. Most of these rules deal with several changes in vowels (vowel harmonization) and in sibilants (fricativization of affricates) in the morpheme border.

Two new features have been used: one (5 is used for this feature) intended to give an account of the morphophonological modification taking place when the singular article is added to a word ending in a (*a + a = ea*) as in *alabea*, the 'the daughter' (*alabA + 5a*) and another one for a special ending in verb roots.

In Table 2 an example of the new rules is shown.

```
# final A is realized as e
#  in Biscayan before
#  morpheme border (MM),
#  5 feature and open vowel
A -> e || \i _ MM 5 (E) OpenVowel ;
# alabA+5a:alabea
```

Table 2. Example of additional rules for Byscayan.

An analyzer was built Using *foma*, and the morphological description was tested and debugged based on a corpus of Biscayan.

After refining the lexicon and the rules several times in an iterative way, 93.95% of the dialectal words that the standard processor could not analyze were recognised now, and 97.92% of the non-Biscayan words that the standard processor analyzed as correct were now discarded. Thus, average precision of the analyzer on these problematic words improved up to 95.06%.

5. Tools and application

In addition to the basic analyzer used for spelling, a new transducer has been included (Alegria et al., 2002). It is an enhanced analyzer which links non-standard forms with the corresponding standard ones. This enhanced analyzer is used in spelling checkers to generate proposals for misspelt words. Bearing in mind that many standard Basque forms may be incorrect when writing in the Biscayan dialect, the enhanced analyzer renders it possible to generate the adequate dialectal form for a given standard form when using the speller for Biscayan.

A spelling checker/corrector for Biscayan named XuxenB has been the first application of this morphological description. It was presented in November 2009 and it can be downloaded for free (http://www.azkuefundazioa.org/content/xuxen-bizkaieraz-deskargagarri).

Figures 1 and 2 show practical usage examples of *XuxenB* on MSOffice.

6. Conclusions and future work

A new morphological processor for Biscayan based on the description of the morphology of standard Basque has been presented. The database for the standard morphology has been extended for dialects and an open-source tool for morphological description named *foma* has been used for building the processor. After the evaluation of the tool, *XuxenB*, a spelling checker/corrector for this dialect, has been the first application of this work.

We believe that this methodology can be used for describing dialects of other languages whose finite-state description has been accomplished.

As future work, we intend carry on with the current research in three aspects related to it:

- Integrating all the dialectal variants in a unique database (Aduriz et al., 1998)
- Using the morphological description and the analyzer/generator in speech recognition and synthesis (Hernaez et al., 2003.
- Getting dialectal variants (morphemes, paradigms and rules) based on standard description and dialectal corpora (Rayson et al., 2005). The description developed could be used as a test for evaluation when we try to learn from corpora written in Biscayan.

7. Acknowledgements

This research has been partially supported in part by the Regional Branch of the Basque Government (BerbaTek project in the Etortek call) and by the Azkue Foundation. Thanks to the Labayru Institute for the work carried out during the debugging process.

8. References

Aduriz I., Aldezabal I., Ansa O., Artola X., Díaz de Ilarraza A. 1998. EDBL: a Multi-Purposed Lexical Support for the Treatment of Basque. Proceedings of the First LREC Conference. Vol II. pp 821-826. Granada.

Alegria I., Artola X., Sarasola K., Urkia M. 1996. Automatic morphological analysis of Basque. Literary & Linguistic Computing Vol. 11, No. 4, 193-203. Oxford University Press. Oxford.

Alegria I., Aranzabe M., Ezeiza A., Ezeiza N., and Urizar R. 2002. Using finite state technology in natural language processing of Basque. LNCS 2494: Implementation and Application of Automata.

Alegria I., Etxeberria I., Hulden M., Maritxalar M. 2009. Porting Basque Morphological Grammars to foma, an Open-Source Tool. FSMNLP2009. Pretoria. South Africa.

Arejita A., Legarra H., Oar-Arteta A. 2002. Bizkai euskeraren jarraibide liburua: lehenengo pausuak. Labayru Ikastegia. Bilbao.

Arejita A., Legarra H., Oar-Arteta A. 2005. Bizkai euskeraren jarraibide liburua: bigarren pausuak. Labayru Ikastegia. Bilbao.

Beesley K. R. and Karttunen L. 2003. Finite State Morphology. CSLI Publications, Palo Alto, CA.

Hernaez, I., Luengo, I., Navas, E., Zubizarreta, M., Gaminde, I. and Sanchez, J. 2003. The Basque Speech_Dat-II Database: A Description and First Test Recognition Results. 8th. European Confe. on Speech Communication and Technology.

Hulden M. 2009. Foma: a Finite-State Compiler and Library. EACL 2009. Demo session. pp 29-32.

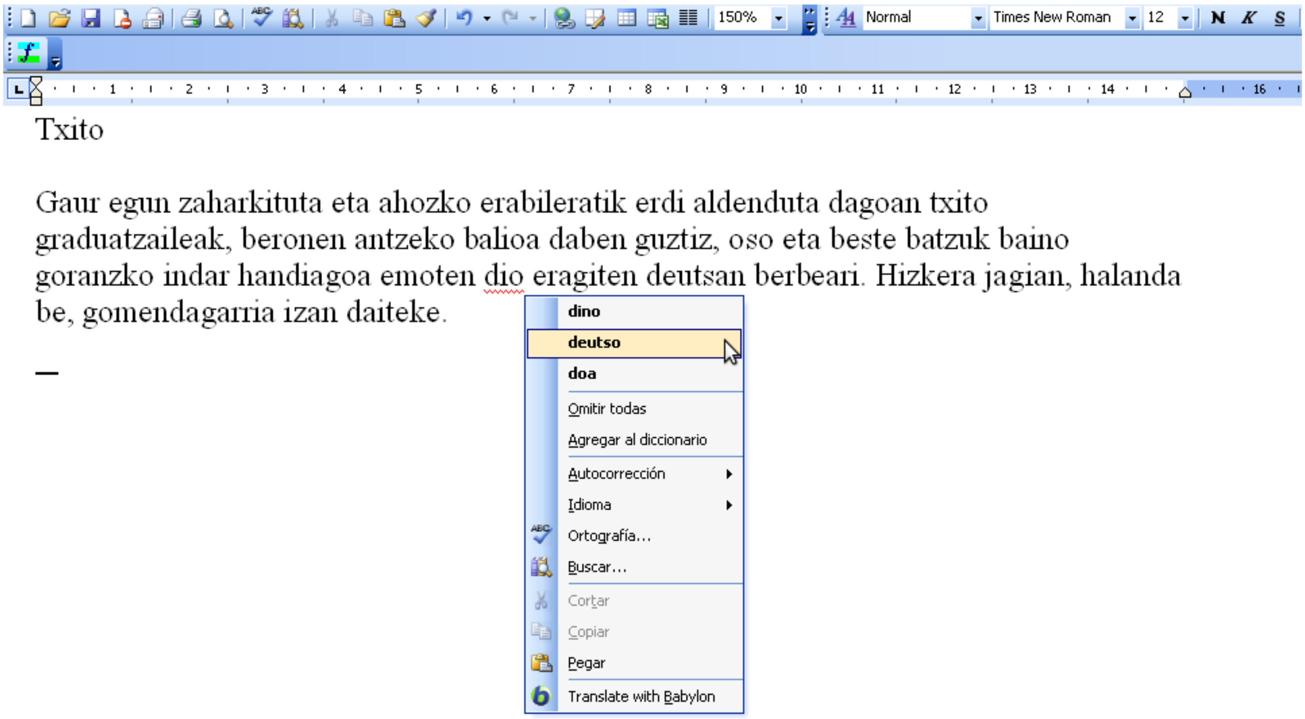
Hulden M. 2009. Fast approximate string matching with finite automata. Proceedings of the SEPLN2009. Donostia. 2009.

Koskenniemi, K 1984. A general computational model for word-form recognition and production. Proceedings of the 10th Conf. on Computational Linguistics, 178—181.

Nemeth V., Tron P., Halacsy A., Kornai A., Rung I. 2004. Leveraging the open source ispell codebase for minority language analysis. Proceedings of SALTMIL, 2004.

Rayson, P. and Archer, D. and Smith, N. 2005. A comparison of the UCREL variant detector and modern spell checkers on English historical corpora. Proceedings of the Corpus Linguistics Conference Series On-Line E-Journal.

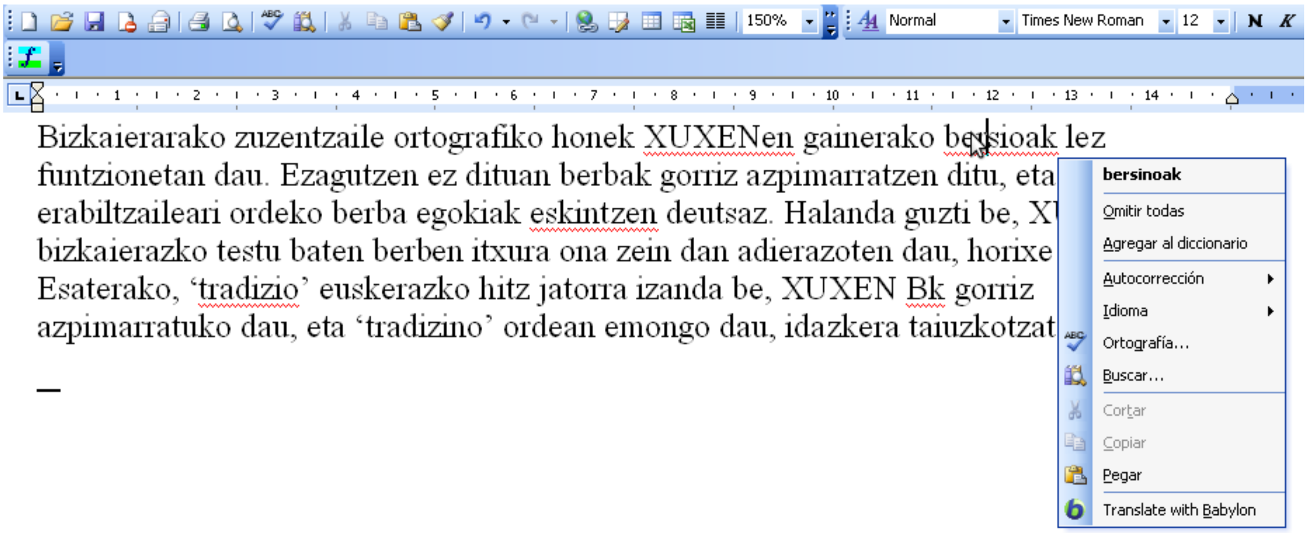
Zuazo K. 2008. Euskalkiak: euskararen dialektoak. Elkar, Donostia.



Txito

Gaur egun zaharkituta eta ahozko erabileratik erdi aldenduta dagoan txito graduatzaileak, beronen antzeko balioa daben guztiz, oso eta beste batzuk baino goranzko indar handiagoa emoten dio eragiten deutsan berbeari. Hizkera jagian, halanda be, gomendagarria izan daiteke.

Figure 1. The Biscayan auxiliary verb *deutso* is proposed as equivalent for the standard *dio*.



Bizkaierarako zuzentzaile ortografiko honek XUXENen gainerako bersioak lez funtzionetan dau. Ezagutzen ez dituan berbak gorriz azpimarratzen ditu, eta erabiltzaileari ordeko berba egokiak eskintzen deutsaz. Halanda guzti be, X bizkaierazko testu baten berben itxura ona zein dan adierazoten dau, horixe Esaterako, 'tradizio' euskerazko hitz jatorra izanda be, XUXEN Bk gorriz azpimarratuko dau, eta 'tradizino' ordean emongo dau, idazkera taiuzkotzat

Figure 2. The Biscayan word *bersinoak* 'versions' is proposed as equivalent for the standard *bertsioak*