Recursos y métodos de sustitución léxica en las variantes dialectales en euskera

Resources and methods for lexical substitution between Basque dialects

Larraitz Uria
IKER (UMR5478)
IKERBASQUE
larraitz.uria@ehu.es

Mans Hulden
University of Helsinki
Language Technology
mans.hulden@helsinki.fi

Izaskun Etxeberria
IXA taldea
UPV-EHU
izaskun.etxeberria@ehu.es

Iñaki Alegria
IXA taldea
UPV-EHU
i.alegria@ehu.es

Resumen: La coexistencia de cinco idiomas oficiales en la Península Ibérica (euskera, catalán, gallego, portugués y español) nos lleva a buscar la colaboración para compartir los recursos desarrollados en los diferentes idiomas de la región. Sin embargo, dentro de un mismo idioma se puede dar la coexistencia de más de un dialecto y así ocurre con el euskera. Las herramientas desarrolladas para este idioma se han centrado básicamente en el euskera unificado o estándar, de modo que no funcionan correctamente con los dialectos, que son numerosos. Este trabajo se enmarca dentro de la idea de buscar la forma de establecer semiautomáticamente una relación entre el euskera estándar y sus variantes dialectales. Esto permitiría aplicar las herramientas automáticas disponibles a los textos anteriores a la unificación del idioma, pudiendo explotar de forma automática la gran riqueza lingüística que aportan.
Palabras clave: Morfología computacional, reglas fonológicas, programación lógica inductiva, dialectos.

Abstract: The coexistence of five languages with official status in the Iberian Peninsula (Basque, Catalan, Galician, Portuguese, and Spanish), has prompted collaborative efforts to share and cross-develop resources and materials for these languages of the region. However, it is not the case that comprehension boundaries only exist between each of these five languages; dialectal variation is also present, and in the case of Basque, for example, many written resources are only available in dialectal (or pre-standardization) form. At the same time, all the computational tools developed for Basque are based on the standard language ("Batua"), and will not work correctly with other dialects, of which there are many. In this work we attempt to semiautomatically deduce relationships between the standard Basque and dialectal variants. Such an effort provides an opportunity to apply existing tools to texts issued before a unified standard Basque was developed, and so take advantage of a rich source of linguistic information.
Keywords: Computational morphology, phonological rules, inductive logic programming, dialects.

1. Introducción

En el área de la morfología computacional existe una línea de investigación abierta en relación a la forma de combinar las aproximaciones lingüísticas y las basadas en aprendizaje automático. Los métodos basados en aprendizaje automático (Goldsmith, 2001) pueden ser interesantes cuando se requiere un desarrollo rápido y se cuenta con pocos recursos o no se dispone de expertos en el idioma a tratar. Pero si se quiere un analizador que compagine cobertura y precisión, la mejor opción es una descripción basada en un léxico y un conjunto de paradigmas y reglas fonológicas especificados por expertos. Las descripciones basadas en tecnologías de estados finitos son las más populares para este fin (Beesley y Karttunen, 2002).

El desarrollo de las bibliotecas digitales y de la lingüística basada en corpus impli-

ca a menudo el tratamiento de las variantes dialectales y/o diacrónicas del idioma, pero no resulta viable tener que realizar una nueva especificación por cada variante a tratar. Así pues, el objetivo de nuestras investigaciones es inferir la morfología de las variantes, o la equivalencia entre variantes y formas estándar del euskera a partir de un pequeño corpus paralelo variante/estándar, un corpus de la variante y un analizador o reconocedor del estándar.

En el trabajo que presentamos tratamos de inferir métodos de sustitución léxica entre variantes y formas estándar del euskera basándonos en la morfología. Concretamente, nuestros primeros experimentos se centran en el dialecto labortano y el objetivo es la sustitución léxica de las formas propias del dialecto por las correspondientes del euskera estándar. La tarea clave, en una primera fase al menos, es la inferencia de las reglas fonológicas a partir de pares variante-estándar. En este artículo describimos los recursos básicos con los que contamos en nuestra investigación, así como los métodos que estamos experimentando para inferir las reglas.

Aunque los resultados obtenidos en los primeros experimentos son alentadores, todavía deben ser ampliados y mejorados antes de poder integrarlos en herramientas computacionales efectivas.

Las técnicas que describimos son, en su mayor parte, independientes del idioma y además, es de suponer que con cierta adaptación pueden ser aplicadas a otras variantes o registros del idioma (por ejemplo, idioma más informal: email, SMS...).

2. Trabajos relacionados

El problema general de aprendizaje supervisado de las variantes dialectales ha sido discutido en la literatura en varias áreas: fonología computacional, morfología, aprendizaje automático...

Por ejemplo, (Kestemont, Daelemans, y Pauw, 2010) presentan un sistema independiente del idioma que puede "aprender" variaciones intra-lema. El sistema se utiliza para producir una lematización coherente de textos en holandés antiguo sobre un corpus de literatura medieval (Corpus-Gysseling), que contiene manuscritos de fecha anterior al año 1300.

(Koskenniemi, 1991), por su parte, ofrece un esbozo de un procedimiento de inferencia de reglas fonológicas de dos niveles pero sin llegar a automatizarlo.

En un trabajo anterior, (Johnson, 1984) presenta un "procedimiento de inferencia" para el aprendizaje de reglas fonológicas a partir de datos, lo que puede ser considerado un trabajo precursor del algoritmo ILP (*Inductive Logic Programming*) que proponemos entre nuestros métodos.

3. Recursos lingüísticos

Para el aprendizaje o inferencia y para la evaluación se necesitan recursos que deben ser almacenados, testeados y, en su caso, etiquetados. La idea de este trabajo es usar métodos no supervisados o con un mínimo de supervisión, ya que ése es el escenario realista para generar aplicaciones en el área.

De momento vamos a probar distintas técnicas en el contexto de las variaciones dialectales en euskera, pero intentando que los métodos sean, en la medida de lo posible, independientes del idioma.

Para llevar a cabo nuestros experimentos en esta investigación, contamos con tres corpus de origen y características diferentes:

- Corpus de transcripciones en labortano
- Corpus de la Biblia en euskera estándar y labortano
- Corpus de transcripciones en diversos dialectos

3.1. Corpus de transcripciones en labortano

Por una parte, contamos con un corpus paralelo construido en el centro de investigación IKER (UMR5478) de Bayona (Francia) dentro del proyecto TSABL[1]. El objetivo de este proyecto es el estudio de la variación sintáctica de los dialectos del País Vasco al norte de los Pirineos (*Iparralde*). Para ello, se ha creado la aplicación BASYQUE[2], en la que se recogen datos y ejemplos de variantes dialectales que provienen de tres fuentes de información: cuestionarios específicos, vídeos de testimonios grabados en otros proyectos y textos literarios.

Una de las principales razones que nos ha llevado a utilizar los datos recogidos en

[1] *Towards a Syntactic Atlas of the Basque Language*: http://www.iker.cnrs.fr/-tsabl-towards-a-syntactic-atlas-of-.html?lang=fr

[2] http://ixa2.si.ehu.es/atlas2/index.php?lang=eu

BASYQUE es la posibilidad que nos ofrece de crear corpus paralelos. Los cuestionarios y testimonios grabados se transcriben y junto a cada ejemplo o frase dialectal también se especifica la forma estándar que le corresponde. En el caso de los textos literarios escritos en dialecto, también se indica la forma estándar que corresponde a cada frase. Estos corpus paralelos labortano-estándar son los que vamos a utilizar en los experimentos de sustitución léxica.

La aplicación BASYQUE pretende abarcar todos los dialectos y subdialectos de *Iparralde* y para ello la recopilación de los datos se extiende a todo el territorio. Para los experimentos, en cambio, en esta primera fase nos centramos en el dialecto labortano, por lo que hemos empleado los ejemplos y los textos que provienen de las zonas donde se habla dicho dialecto. Y de momento hemos utilizado los ejemplos recogidos mediante los cuestionarios y los textos literarios, ya que las grabaciones de video no están transcritas todavía. Cabe reseñar que dichos corpus están siendo actualizados y ampliados dentro del mencionado proyecto, de modo que los datos presentados en la Tabla 1 corresponden al corpus de transcripciones labortano-estándar disponible en el momento de realizar los experimentos.

	Corpus	**80 %**	**20 %**
N°frases	2.117	1.694	423
N°palabras	12.150	9.734	2.417
Palabras dif.	3.874	3.327	1.243
Pares filtrados	3.610	3.108	1.172
Pares idénticos	2.532	2.200	871
Pares diferentes	1.078	908	301

Tabla 1: Datos correspondientes al corpus labortano-estándar utilizado en los experimentos realizados hasta el momento. La primera columna corresponde al corpus completo. El 80 % ha sido utilizado en la fase de aprendizaje y el 20 % restante en la fase de test.

En la Tabla 2 se presentan varios ejemplos de frases con el fin de que se vea el tipo de diferencias que se pueden encontrar entre el dialecto y el estándar, así como la correspondencia palabra a palabra con que se cuenta en dicho corpus.

Éste es el corpus en el que hemos centrado nuestros primeros experimentos y con el que hemos obtenido los resultados que presentamos en el apartado 5.

Dialecto labortano vs **Euskera estándar**
Leihoa ***estea*** *erreusitu du.*
Leihoa ***ixtea*** *erreusitu du.*
Eni galdegin ***daut*** *100 euro.*
Eni galdegin ***dit*** *100 euro.*
Ez gero uste izan ***nexkatxa guziek*** *tu egiten* ***dautatela****.*
Ez gero uste izan ***neskatxa guztiek*** *tu egiten* ***didatela****.*

Tabla 2: Varios ejemplos de frases en el corpus paralelo labortano-estándar.

3.2. Corpus de la Biblia

Otra fuente de información básica para nuestro trabajo es la Biblia, que está publicada en euskera estándar y también en dialecto labortano, lo que nos proporciona un corpus paralelo bastante mayor que el anterior. La versión de la Biblia en euskera estándar ha sido editada dos veces, en 1994 y en 2004 respectivamente, y existe una versión electrónica en la web (http://www.biblija.net). En cuanto a la versión en dialecto labortano, se trata de una adaptación de la versión estándar realizada por Marcel Etcehandy y publicada en 2007, y dispone también de una versión electrónica (http://amarauna.org/biblia/). Debido a problemas de formato, de momento sólo hemos alineado 9 libros (elegidos al azar) con las características que se reflejan en la Tabla 3.

N°de libros total	76
N°de libros disponible	66
Palabras totales en euskera estándar	545.700
Palabras diferentes	38.069
Libros alineados	9
Palabras totales en libros alineados	104.967
Palabras diferentes en libros alineados	15.007

Tabla 3: Datos correspondientes al corpus de la Biblia y a los libros alineados hasta la fecha.

Este corpus, al ser de mayor tamaño, nos va a permitir realizar experimentos con distintos tamaños de corpus paralelo, y así conseguir estimar correlaciones entre tamaños de

corpus paralelo y calidad de la inferencia, pero todavía no tenemos resultados que mostrar sobre este aspecto ya que estamos en la fase de preparación y obtención de información de este corpus. Por otro lado, a diferencia del corpus descrito en 3.1, en el corpus de la Biblia no hay transcripción palabra a palabra tal y como se puede observar en el pequeño ejemplo[3] que se presenta a continuación, por lo que la obtención del diccionario de palabras equivalentes se prevé más complicada.

- Dialecto labortano:

  > *"**Errana dauzut: ukan indar eta kuraia.** Ez **ikara**, ez **izi**, ni, Jauna, zure Jainkoa, zurekin **izanen** bainaiz **joanen ziren toki guzietan**".*

- Euskera estándar:

  > *"**Kementsu eta adoretsu izateko esan dizut.** Ez **ikaratu**, ez **kikildu**, ni, Jauna, zure Jainkoa, zurekin **izango** bainaiz **zure ibilera guztietan**".*

3.3. Corpus de transcripciones en diversos dialectos

Existen varios proyectos en el País Vasco (Ahotsak.com[4] o EKE.org[5], por ejemplo) que tienen como objetivo recoger el habla tradicional de cada zona, es decir, recopilar y difundir testimonios orales de vasco-parlantes. En ambos proyectos se graban y se recogen conversaciones y/o testimonios de personas que se expresan en su propio dialecto.

Nosotros hemos creado una red de colaboración con Ahotsak.com para poder recopilar y ayudar a transcribir corpus paralelos de variantes dialectales relacionadas con la forma estándar, ya que el objetivo de Ahotsak.com es ir transcribiendo gran parte de los testimonios grabados. Hasta ahora, cuentan con 5.204 pasajes (1.462.555 palabras) transcritos en las formas dialectales. Sin embargo, para facilitar la búsqueda se quiere relacionar cada forma dialectal con su correspondiente estándar, y para hacerlo de forma (semi)automática nos queremos valer de las técnicas que estamos desarrollando y que describimos posteriormente.

Las características de este corpus son en parte equiparables a las del primer corpus descrito, pero con dos diferencias reseñables:

- recoge gran variedad de dialectos, ya que ciertas formas van cambiando casi de pueblo a pueblo (véase el mapa en http://ahotsak.com/herriak/mapa/)
- de momento sólo disponemos de la transcripción de las formas dialectales y queremos obtener de forma (semi)automática las correspondientes formas estándar. Una parte de la investigación que hacemos es determinar el mínimo de trabajo manual (para relacionar las formas estándar con las dialectales) necesario para obtener unos buenos resultados después en la posterior sustitución léxica.

4. Métodos

Nuestra primera aproximación se va a basar en obtener pares de palabras variante/estándar a partir de un corpus paralelo (que quisiéramos minimizar). Para ello reutilizamos lo que hemos llamado métodos básicos. Posteriormente inferiremos reglas fonológicas mediantes dos métodos.

4.1. Métodos básicos

De cara a obtener pares de palabras equivalentes a partir de corpus paralelos vamos a utilizar dos programas: *lexdiff* y Giza++.

El primero, *lexdiff*, ha sido diseñado y utilizado para la migración automática de textos entre diferentes ortografías del portugués (Almeida, Santos, y Simoes, 2010), debido al cambio de norma que se produjo en ese idioma. Este programa trata de identificar la equivalencia de palabras a partir de frases paralelas. Funciona muy bien cuando los textos son equivalentes palabra por palabra, y es por ello que lo hemos utilizado en los experimentos realizados hasta ahora con el corpus de transcripciones labortano-estándar.

Adicionalmente, *lexdiff* también calcula los cambios de ngramas y sus frecuencias, obteniendo resultados de este tipo: `76 ait ->at`; `39 dautz ->diz`; lo que indica que el ngrama `ait` ha cambiado a `at` 76 veces en el corpus y que `dautz` ha cambiado 39 veces a `diz`.

Estos resultados pueden expresar cambios (morfo)fonológicos regulares entre los textos,

[3] El ejemplo corresponde al versículo 9 del capítulo 1 del libro de Josué.

[4] http://www.ahotsak.com/

[5] http://www.eke.org/

y han sido explotados en el primero de los métodos de inferencia que presentamos a continuación.

Giza++[6] es una conocida herramienta para inferir diccionarios, con probabilidades de traducción, a partir de corpus paralelos. Lo queremos comparar con *lexdiff* dado que el corpus de la Biblia con el que contamos es un corpus paralelo divergente y de mayor tamaño, pero todavía no podemos presentar resultados sobre dicha comparación.

4.2. Métodos de inferencia

Estamos experimentando con dos métodos de inferencia:

1. Inferencia de reglas fonológicas basada en substrings
2. Inferencia usando programación lógica inductiva sobre pares de palabras equivalentes

El método *baseline* consiste en aprender las equivalencias de pares diferentes en el corpus de aprendizaje (corpus paralelo) y sustituirlas en el de test, suponiendo que si no se ha aprendido la forma estándar correspondiente a la variante es la propia variante. Este método tiene como resultado buena precisión y baja cobertura. Los dos métodos que proponemos parten de una lista de equivalencia de palabras o de substrings obtenida por las herramientas básicas y tratan de inferir reglas fonológicas de reemplazamiento que puedan ser compiladas por *xfst* de Xerox (Beesley y Karttunen, 2002) o *foma* (software libre, (Hulden, 2009)).

4.2.1. Inferencia de reglas fonológicas basada en substrings.

En principio se basa en los cambios de ngramas que obtiene *lexdiff*. Hay varias formas de transformar esa salida de *lexdiff* en reglas de reemplazamiento que se compilan a transductores finitos. Estamos teniendo en cuenta los siguientes factores:

- Limitar los cambios a tener en cuenta a aquellos que tienen un mínimo de frecuencia (por ejemplo, dos o tres). Si aumentamos el mínimo mejoraremos la precisión, pero perderemos cobertura.
- Limitar el número de reglas que pueden ser aplicadas a la misma palabra. Por ejemplo, la correspondencia *agerkuntza/agerpena* puede expresarse mediante dos reglas: `rkun ->rpen` y `ntza ->na`, pero permitir varios cambios puede producir ruido innecesario y bajar la precisión.
- La forma de aplicar las reglas: secuencialmente o paralelamente.
- Hacer que los cambios sean de longitud mínima y condicionados por el contexto.

4.2.2. Inferencia usando programación lógica inductiva.

El segundo método consiste en los siguientes pasos:

1. Alinear los pares de palabras letra por letra usando la mínima distancia de edición.
2. Extraer un conjunto de reglas fonológicas.
3. Por cada regla, buscar contraejemplos.
4. Buscar la restricción de contexto mínima que resuelva los contraejemplos.

Por ejemplo, si tenemos los pares *emaiten/ematen* e *igorri/igorri*, en el primer paso se detecta el cambio *i/0*, que en el paso dos se convierte en la regla `i ->0`. Pero ese cambio no se puede aplicar con *igorri*, por lo que la regla se transforma para evitar que sea aplicada. Este método tiene la ventaja de explotar las formas que son idénticas en el dialecto y en el estándar.

5. Resultados y trabajos futuros

Hemos centrado los experimentos en el corpus descrito en el apartado 3.1 con el fin de testear y evaluar los métodos descritos en el apartado 4. Los primeros resultados nos muestran una mejora respecto al método *baseline*, pero todavía deben ser mejorados para utilizarlos en herramientas computacionales efectivas.

La Tabla 4 muestra los resultados obtenidos. Dichos resultados corresponden tanto al método *baseline*, como a los mejores resultados obtenidos con cada una de las propuestas de inferencia de reglas descritas y se expresan en términos de precisión (*precision*), cobertura (*recall*) y la medida-F (*F-score*), que es la combinación de ambas. En los tres casos, el proceso de aprendizaje se ha llevado a cabo

[6] http://code.google.com/p/giza-pp/

con el 80 % del corpus, y el test, cuyos resultados son los que see muestran en la Tabla 4, se ha realizado sobre el 20 % restante.

Aunque no se presentan más que los mejores resultados obtenidos con cada método, el número de experimentos realizados con ambos métodos ha sido numeroso, sobre todo con el método de inferencia de reglas basada en substrings, debido a los diferentes factores que se pueden tener en cuenta para inferir las reglas fonológicas. Dichos experimentos nos muestran que:

- Disminuir la mínima frecuencia exigida a un cambio para obtener una regla fonológica a partir de él, aumenta notablemente la cobertura, pero también hace que disminuya la precisión, con lo que el resultado en términos de F-score apenas mejora.
- La aplicación de más de una regla en una palabra no parece aportar incrementos importantes en la mejora de los resultados.
- El modo de aplicación, secuencial o paralelo, de las reglas (cuando se aplica más de una regla en la misma palabra) presenta resultados muy similares, aunque algo mejores si la aplicación es paralela.
- Por último, minimizar la longitud de los cambios y hacer que sean condicionados por el contexto, obtiene claramente mejores resultados.

En los primeros experimentos con este método de inferencia, ya pudimos comprobar que la aplicación exclusivamente de las reglas fonológicas no mejoraba los resultados del método *baseline*, debido a que la precisión era excesivamente baja (para cada término a sustituir, el número de candidatos era a menudo elevado). Ello nos llevó a aplicar un post-filtro al proceso, basado en la frecuencia de los candidatos en euskera estándar[7]. El filtro aplicado es muy simple: si hay más de un candidato se elige el más frecuente, pero a pesar de su simplicidad se mejoran los resultados y se consigue superar el *baseline* tal y como se puede ver en los resultados presentados en la Tabla 4.

[7]La frecuencia de cada término la hemos obtenido de un corpus de un diario de noticias editado en euskera.

	Precision	**Recall**	**F-score**
Baseline	95,62	43,52	59,82
Método 1	75,10	60,13	66,79
Método 2	85,02	58,47	69,29

Tabla 4: Mejores resultados (en términos de *F-score*) obtenidos con ambos métodos de inferencia en los experimentos realizados con el corpus de transcripciones labortano-estándar.

Con respecto al segundo método de inferencia, basado en programación lógica inductiva, los resultados obtenidos han sido mejores, y además, con este método no es necesaria la aplicación del filtro posterior. El motivo fundamental es que este método no sólo utiliza la información de los pares diferentes, sino también la de los pares iguales en el dialecto y en el estándar.

Se puede consultar información más detallada tanto de los métodos propuestos como de la evaluación realizada en (Hulden et al., 2011).

Todavía nos queda mucho trabajo por realizar en el campo de esta investigación. La aplicacion de los métodos descritos al corpus de la Biblia nos va a permitir precisar hasta qué punto es determinante que la transcripción entre dialecto y estándar sea palabra a palabra, y qué tamaño de corpus es necesario para obtener resultados que indiquen que es posible conseguir herramientas automáticas de sustitución léxica.

Además, creemos que los métodos utilizados deben ser combinados con otros que infieran relaciones entre lemas y morfemas (variantes y formas estándar), variantes de paradigmas y que contrasten esas inferencias con corpus de variantes (sin que sean corpus paralelos) más amplios.

Bibliografía

Almeida, J. J, A. Santos, y A. Simoes. 2010. Bigorna–a toolkit for orthography migration challenges. En *Seventh International Conference on Language Resources and Evaluation (LREC2010), Valletta, Malta.*

Beesley, K. R y L. Karttunen. 2002. Finite-state morphology: Xerox tools and techniques. *Studies in Natural Language Processing. Cambridge University Press.*

Goldsmith, J. 2001. Unsupervised learning of the morphology of a natural language. *Computational linguistics*, 27(2):153–198.

Hulden, M., I. Alegria, I. Etxeberria, y M. Maritxalar. 2011. An unsupervised method for learning morphology of variants from the standard morphology and a little parallel corpus. En *(EMNLP workshop) Dialects-2011 — First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*.

Hulden, Mans. 2009. Foma: a finite-state compiler and library. En *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, páginas 29–32, Athens, Greece. Association for Computational Linguistics.

Johnson, Mark. 1984. A discovery procedure for certain phonological rules. En *Proceedings of the 10th international conference on Computational linguistics*, COLING '84, páginas 344–347. Association for Computational Linguistics.

Kestemont, M., W. Daelemans, y G. De Pauw. 2010. Weigh your words—memory-based lemmatization for Middle Dutch. *Literary and Linguistic Computing*, 25(3):287–301.

Koskenniemi, K. 1991. A discovery procedure for two-level phonology. *Computational Lexicology and Lexicography: A Special Issue Dedicated to Bernard Quemada*, páginas 451–446.