# RECONOCIMIENTO AUTOMÁTICO DEL HABLA EN EUSKERA EMPLEANDO INFORMACIÓN LINGÜÍSTICA

*Víctor G. Guijarrubia, M. Inés Torres*

Departamento de Electricidad y Electrónica
Universidad del País Vasco, Apartado 644, 48080 Bilbao. Spain
{vgga,manes}@we.lc.ehu.es

**RESUMEN**

El presente trabajo presenta experimentos preliminares en reconocimiento automático del habla en euskera haciendo uso de información lingüística, más concretamente, de lemas. El uso de esta información permite reducir el vocabulario de la tarea y agilizar el proceso de reconocimiento, manteniendo las prestaciones de un sistema estándar basado en palabras.

## 1. INTRODUCCIÓN

En el presente trabajo pretendemos adentrarnos en el reconocimiento automático del habla en euskera empleando un cierto conocimiento lingüístico.

El euskera es un lenguaje minoritario, pero oficial en el País Vasco. Se trata de un idioma pre-indoeuropeo de origen desconocido. Comunmente se ha asociado con el alemán, debido a que, como este, presenta una alta declinación, tanto en nombres como en verbos. Por ejemplo, *a casa* se traduce como *etxeRA*, *a las casas*, como *etxeETARA* y *a casas*, como *etxeTARA*. Es por esto que siempre se ha pensado que el uso de conocimiento lingüístico, como el proporcionado por los lemas, podría ser de gran ayuda y se hayan producido grandes estudios en el desarrollo de analizadores morfosintácticos [1] y lematizadores [2]. El empleo de este tipo de información ha reportado mejoras en otras aplicaciones relacionadas con las tecnologías del habla, como por ejemplo en sistemas de traducción estadística texto-texto de castellano a euskera [3].

En este trabajo, se muestran experimentos de reconocimiento automático del habla que hacen uso del conocimiento proporcionado por los lemas disponibles para las palabras de una base de datos meteorológica.

En lo que sigue, el resto del artículo está organizado de la siguiente manera: en la sección 2 se describen los métodos empleados para hacer uso de ese conocimiento lingüístico; en la sección 3 se describe la base de datos empleada en la experimentación, tanto la parte textual empleada para entrenar los modelos de lenguaje como la parte de voz empleada en experimentos de reconocimiento automático del habla; en la sección 4 se describe la experimentación llevada a cabo y los resultados obtenidos; en la sección 5 se encuentran los agradecimientos y por último, en la sección 6 se dan las conclusiones y las líneas de actuación futuras para completar este trabajo.

Este trabajo ha sido subvencionado por la Universidad del País Vasco bajo el proyecto 9/UPV 002224.310-15900/2004 y el CICYT bajo el proyecto TIN2005-08660-C04-03

## 2. HACIENDO USO DE LOS LEMAS

Un lema es cada una de las formas que, en ciertas lenguas, presenta un radical para recibir los morfemas de flexión. Para hacer uso de la información suministrada por los lemas, se han seguido dos líneas de actuación: usarla diréctamente para agrupar palabras bajo un lema común y emplearla para obtener una serie de raices y sufijos típicos en la tarea de aplicación.

Por un lado, se usaron los lemas diréctamente, haciendo un sistema de reconocimiento donde las unidades léxicas no eran las palabras sino los lemas. En este caso, las palabras se introdujeron como pronunciaciones alternativas del lema dentro del léxico. En la experimentación descrita en el presente trabajo, a cada pronunciación se le asignó la probabilidad del unigrama, es decir, el número de veces que aparece dicha palabra lematizada como un cierto lema dividido por el número total de apariciones de dicho lema.

Por otro lado, se pretendió obtener los sufijos más comunes que aparecen en la lengua vasca, restringidos a la tarea de la aplicación, para emplearlos como palabras independientes en un sistema de reconocimiento automático del habla.

Al estar la base de datos disponible tanto en forma de palabras como en forma de lemas, fue posible agrupar todas las palabras correspondientes a un mismo lema. Una vez obtenidas las agrupaciones por lema, la separación en cada agrupación se realizó tomando la raíz común. De este modo, de una sola palabra se pasó a dos: raíz y sufijo diferenciador. De este proceso se obtuvieron, por ejemplo, casos como los mostrados en las figura 1 y 2, donde se muestran los casos de las palabras agrupadas bajo los lemas *agertu* (aparecer) y *aldatu* (cambiar) y se puede observar que de ocho palabras se pasa a siete, dos raices y cinco sufijos.

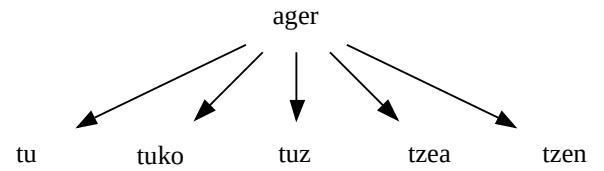


**Figura 1**. *Ejemplo de segmentación de las palabras cuyo lema es agertu (*aparecer*).*

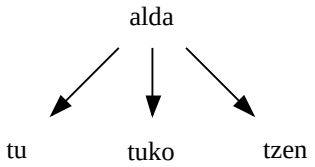


**Figura 2**. *Ejemplo de segmentación de las palabras cuyo lema es aldatu (*cambiar*).*

Aplicando estas separaciones, se obtuvo un nuevo corpus con el que entrenar un modelo de la aplicación.

## 3. BASE DE DATOS

La experimentación descrita en este trabajo se desarrolló empleando la base de datos METEUS [4], un corpus formado por predicciones meteorológicas diarias tomadas del Instituto Vasco de Meteorología durante 28 meses. Se tomaron las predicciones tanto en castellano como en euskera, obteniéndose así un corpus bilíngüe, planeado para realizar experimentación de traducción automática. Posteriormente este corpus fue dividido, de manera aleatoria, en un conjunto de entrenamiento y otro de test y se lematizo, permitiendo trabajar tanto con palabras como con lemas.

En la tabla 1 pueden verse las características más importantes del conjunto de entrenamiento en euskera de esta base de datos. Se da la información tanto para el caso de palabras como para el de lemas y raices más sufijos.

Cabe destacar especialmente el aumento significativo en la talla del vocabulario en euskera. Para esta misma tarea en castellano, el vocabulario no alcanzaba las 700 palabras. Sin embargo, para euskera es nétamente superior, debido a que, como ya se ha mencionado, el euskera es un idioma áltamente declinado. Esta alta declinación queda reflejada observando la talla del vocabulario en los otros

| | Palabras | Lemas | Raices+Sufijos |
|---|---|---|---|
| Frases | 14615 | | |
| Nº palabras | 154778 | | |
| Vocabulario | 1098 | 426 | 588 |
| Longitud media | 10.59 | | |

**Tabla 1**. Características más importantes del corpus de entrenamiento en euskera de la base de datos.

| | |
|---|---|
| Número de locutores | 36 (20 hombres + 16 mujeres) |
| Origen | 17 lengua materna<br>19 lengua no materna |
| Dialecto | 22 *batua*<br>10 vizcaino<br>4 guipuzcoano |
| Número de elocuciones | 1800 |
| Frases diferentes | 500 |
| Tamaño | 3.5 horas |
| Longitud media elocución | 7 segundos |

**Tabla 2**. Características más importantes de la base de datos METEUS de voz en euskera.

dos casos. Como se puede observar, el número de lemas es inferior a la mitad del número de palabras (y del mismo orden que para el castellano, con 414 lemas) y para la metodología basada en raices y sufijos comunes, se produce una disminución en la talla del vocabulario de más de 500 palabras, debido a que muchas palabras comparten los mismos sufijos debido a esta alta declinación.

Posteriormente se decidió grabar el corpus de test tanto para euskera como para castellano. En total, se emplearon 36 locutores bilingües en la grabación. Del conjunto de test se tomaron las frases distintas no contenidas en el entrenamiento, 500 en total. Estas 500 frases diferentes se dividieron en bloques de 50 frases y cada locutor grabó las frases de uno de los bloques, resultando en que los seis primeros bloques fueron grabados por cuatro locutores diferentes, mientras que los otros cuatro, únicamente por tres locutores. En total, el corpus de voz está compuesto por algo de tres horas para cada uno de los idiomas.

En la tabla 2 pueden verse las características más importantes de la parte de voz de esta base de datos, incluyendo una descripción de la distribución de los locutores participantes en el corpus de voz. Estos datos están referidos únicamente a la parte de euskera.

## 4. RESULTADOS EXPERIMENTALES

### 4.1. Condiciones experimentales

Para realizar la experimentación, la base de datos fue parametrizada en 12 coeficientes cepstrales en frecuencia Mel, con sus correspondientes derivadas y segundas derivadas, energía y derivada de la energía. Se definieron por tanto cuatro representaciones acústicas. La longitud de la ventana de análisis fue de 25 milisegundos y el desplazamiento de la ventana, de 10 milisegundos.

Cada unidad fonética se modeló mediante modelos ocultos de Markov continuos de tres estados, con arcos de izquierda a derecha de un estado al siguiente y autolazos en cada estado. Se emplearon modelos de 32 gaussianas por estado y representación acústica. Se emplearon un total de 35 unidades fonéticas [5, 6].

Para representar los modelos de lenguaje empleados

| | Tasa de error |
|---|---|
| PALABRAS | 5.10 |
| LEMAS | 16.31 |
| RAICES+SUFIJOS | 5.93 |

**Tabla 3**. Tasa de error obtenida para el sistema base basado en palabras, el sistema que hace uso de los lemas diréctamente y el sistema basado en las raices y los sufijos obtenidos por medio de los lemas.

| | Coste temporal |
|---|---|
| PALABRAS | 1 |
| LEMAS | 1.05 |
| RAICES+SUFIJOS | 0.95 |

**Tabla 4**. Coste temporal del proceso de reconocimiento.

en el sistema de reconocimiento, en lugar de emplear, como es típico, modelos de *n-gramas*, se emplearon *k-testables en sentido estricto* [7], ya que mantienen las restricciones sintácticas del lenguaje. Para los experimentos de la siguiente sección se emplearon modelos con $k = 3$.

### 4.2. Resultados de la experimentación

En la tabla 3 pueden verse los resultados de reconocimiento obtenidos para los tres sistemas estudiados, los dos basados en los lemas así como el sistema basado en palabras. De igual manera, la tabla 4 muestra el coste temporal del proceso de reconocimiento para los tres casos.

Se observa que el sistema basado en los lemas funciona bastante peor. No solo se produce un aumento considerable en la tasa de error, sino que además es ligéramente más lento. Este aumento en la tasa de error puede ser debido a que, al usar lemas, hay más posibles transiciones entre lemas que entre palabras, con lo que se produce una dispersión a nivel de modelo de lenguaje. Además, se pierde la relación entre las palabras y se le pide a los modelos acústicos que diferencien entre palabras generalmente similares dentro de un lema.

Para el caso basado en raices y sufijos, las tasas de error obtenidas son similares que para el caso base de usar palabras. Como contrapartida, al reducir tan drásticamente la talla del vocabulario, el tiempo de cómputo del sistema de reconocimiento se reduce en algo más de un 5 %, que aproximadamente supone ahorrarse de media unas cuatro décimas de segundo en cada elocución reconocida. Es decir, sacrificamos algo de precisión a costa de hacer el sistema más rápido. Esto puede ser de especial interés en tareas de mayor dificultad y con un vocabulario de mayor tamaño, en el que el efecto de reducción de la talla del vocabulario podría ser mucho mayor y, por consiguiente, se prodría aligerar sensiblemente el proceso de reconocimiento.

Hay que hacer mención también en otro de los problemas o ventajas que puede acarrear la metodología basada en obtener los sufijos comunes. Un sistema de reconocimiento del habla genera siempre frases formadas por palabras del vocabulario. En este caso, al emplear raices y sufijos en el vocabulario, es necesario deshacer la separación al finalizar el reconocimiento. En ese proceso, puede ocurrir que se generen palabras que no existen en el vocabulario original, ya que se puede reconocer un sufijo tras una raiz no observada en entrenamiento debido al suavizado del modelo de lenguaje. Este efecto puede ser positivo o negativo. Por un lado, podría darse el caso en que esos sufijos sirviesen para generalizar una raiz y darle opciones no vistas en un entrenamiento con palabras. Por ejemplo, siguiendo con los ejemplos de las figuras 1 y 2, podría darse el caso, como el mostrado en la figura 3, de que para el lema *gainditu* (superar) se hubiesen encontrado las palabras *gainditu* y *gainditukо*, es decir, al segmentar se obtendría la raíz *gaindi* y los sufijos *tu* y *tuko*. Debido a que para *agertu* y *aldatu* se han observado otros sufijos extra, sería posible que se reconociesen también *gaindituz*, *gainditzea* y *gainditzen*, palabras totálmente correctas, y que el sistema base basado en las palabras no podría hacer. Por otra lado, siguiendo con la misma idea, se podrían generar también palabras que no existen en el idioma, produciendo un efecto indeseado en un sistema de reconocimiento automático del habla. Concretamente, en la experimentación se observaron más de 100 palabras no presentes en el vocabulario original, de las cuales únicamente en torno a una quincena son correctas.

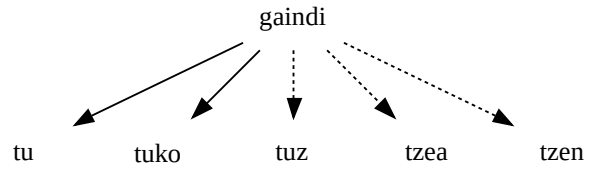


**Figura 3**. *Ejemplo de posible generalización que se podría obtener para otros lemas, como por ejemplo gainditu (*superar*). En trazo continuo se muestran los sufijos supuestamente encontrados en el entrenamiento y en discontinuo, los posibles sufijo extra que se podrían reconocer sin haber sido observados en entrenamiento, obteniendose palabras totálmente válidas.*

Para evitar estos efectos se siguio una aproximación sencilla preliminar en la que se impedían transiciones potencialmente problemáticas. El vocabulario de la tarea está formado por tres tipos de palabras: palabras no segmentadas, raices y sufijos. A las palabras no segmentadas se le impiden transitar a los sufijos. A las raices sólo se les permiter transitar a los sufijos y a los sufijos, tanto a las palabras no segmentadas como a las raices. Para las transiciones de raices a sufijos sólo se permitieron aquellas observadas en el entrenamiento, es decir, se impidió el tipo de generalización positiva descrita en anterioridad por medio de la figura 3. Los resultados obtenidos siguiendo

| Tasa de error | Coste temporal |
|---|---|
| 6.31 | 0.85 |

**Tabla 5**. Tasa de error y coste de procesado obtenidos cuando se impiden transiciones potencialmente problemáticas en el sistema basado en raices y sufijos.

esta aproximación se muestras en la tabla 5.

Se observa un aumento de las tasas a pesar de impedir transiciones a priori no posibles. A pesar de ello, esta modificación forzada del suavizado del modelo de lenguaje provoca un amplio descenso en los costos de procesado, que podría ser interesante en tareas de mayor dificultad. En este caso, únicamente aparecen cuatro palabras no presentes en el vocabulario original, que corresponden a raices que se reconocen al final de la elocución.

## 5. AGRADECIMIENTOS

Queremos agradecer al grupo *Ametzagaiña* y, en particular, a Josu Landa por proveernos con el corpus METEUS lematizado, que hizo posible este trabajo.

## 6. CONCLUSIONES Y TRABAJO FUTURO

En el presente trabajo se ha presentado una metodología para hacer reconocimiento automático del habla en euskera haciendo uso de información lingüística proporcionada por la base de datos lematizada. Para el caso del sistema basado en separar las palabras en raices y sufijos, se obtienen unos resultados similares que al sistema base basado en palabras, pero al reducir la talla del vocabulario se produce una disminución en el tiempo de cómputo del sistema de reconocimiento. En el sistema basado en agrupar palabras por lemas, se produce una dispersión a nivel de modelo de lenguaje y se pierde la relación entre las palabras, resultando en un aumento no sólo de la tasa de error, sino también del coste temporal.

En el futuro se desearía aplicar esta metodología a una tarea de mayor complejidad y con un vocabulario más amplio, en la que la reducción de la talla del vocabulario sería más drástica y se pudiese obtener un beneficio mayor. Este estudio permitiría además dar una valoración definitiva al uso de esta metedología. También sería interesante estudiar otros posibles métodos para impedir la aparición de palabras inexistente en el idioma, sin por ello limitar la posibilidad de generar palabras válidas por medio de uniones de raices y sufijos no vistos en entrenamiento.

## 7. BIBLIOGRAFÍA

[1] I. Aduriz, M. Aranzabe, J. Arriola, A. Díaz de Ilarraza, K. Gojenola, M. Oronoz, y L. Uria, "A cascaded syntactic analyser for basque," in *Computational Linguistic and Intelligent Text Processing*, Berlin, 2004, pp. 124–135, Springer-Verlag, LNCS 2945.

[2] Itziar Aduriz, Izaskun Aldezabal, Iñaki Alegria, Xabier Artola, Nerea Ezeiza, y Ruben Urizar, "Euslem: A lemmatiser/tagger for basque," in *Proceedings of EURALEX'96, Part I*, Goteborg (Sweden), 1996, pp. 17–26.

[3] A. Pérez, I. Torres, y F. Casacuberta, "Towards the improvement of statistical translation models using linguistic features," in *Proceedings of the FinTAL - 5th International Conference on Natural Language Processing. Lecture Notes in Computer Science 4139*, Turku, Finland, 23-25 August 2006, pp. 716–725.

[4] A. Pérez, I. Torres, F. Casacuberta, y V. Guijarrubia, "A Spanish-Basque weather forecast corpus for probabilistic speech translation," in *5th SALTMIL Workshop on Minority Languages*, Genoa, Italy, 23 May 2006, pp. 99–101.

[5] V. Guijarrubia, I. Torres, y L.J. Rodríguez, "Evaluation of a spoken phonetic database in basque language," in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, 2004, vol. 6, pp. 2127–2130.

[6] Karmele López de Ipiña, Inés Torres, Lourdes Oñederra, y Luis Javier Rodriguez, "Selection of sublexical units for continuous speech recognition of basque," in *Proc. of International Conference of Spoken Language Processing*, Beijing, 2000.

[7] I. Torres y A. Varona, "k-tss language model in a speech recognition system," *Computer Speech and Language*, pp. 15(2):127–149, 2001.