Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 103-111 recibido 1-05-2009; aceptado 5-06-2009

Tratamiento del léxico del euskara occidental basado en la división de radical y desinencia para reconocimiento de habla dialectal

Stem and ending division based treatment of Western Basque lexicon for dialectal speech recognition

Igor Odriozola
UPV/EHU
Urkixo zum., z/g
igor.odriozola@ehu.es

Eva Navas
UPV/EHU
Urkixo zum., z/g
eva.navas@ehu.es

Jon Sánchez
UPV/EHU
Urkixo zum., z/g
jon.sanchez@ehu.es

Inma Hernáez
UPV/EHU
Urkixo zum., z/g
inma.hernaez@ehu.es

Resumen: En este artículo se presenta una primera aproximación para tratar el reconocimiento de habla dialectal en euskara, basada en la división de los elementos del diccionario en radicales y desinencias. De este modo se logra, por un lado, disminuir el tamaño del diccionario al tratar los casos gramaticales aglutinantes del euskara como un grupo finito de desinencias, y, por otro, tratar las diferentes variantes fonéticas y fonológicas que presentan dichos casos gramaticales en las distintas hablas pertenecientes al dialecto occidental. En este artículo se muestra el procedimiento seguido en los experimentos y los resultados obtenidos.
Palabras clave: Radical y desinencia, variaciones fonéticas, ASR dialectal, euskara occidental.

Abstract: In this paper a first approach based in the division of the dictionary elements into stems and endings is introduced to deal with Basque dialectal speech recognition. In this way, two objectives are achieved: on the one hand, the great dictionary decrease due to the treatment of the agglutinative grammatical cases of Basque as a finite group of endings; on the other hand, the treatment of the phonetic and phonological variants that show these grammatical cases in the different forms of the western dialect. In this paper, the procedure used in the experiments and the results obtained are shown.
Keywords: Stems and endings, phonetic variations, dialectal ASR, western Basque.

1 Introducción

El euskara es una lengua preindoeuropea aglutinante y altamente flexiva que, según las últimas propuestas dialectológicas (Zuazo, 2003), está compuesta de seis dialectos principales. El dialecto occidental se habla en toda la provincia de Bizkaia, en el norte de Araba y en la parte más occidental de Gipuzkoa, por lo que es uno de los dialectos que más hablantes tiene junto con el central, que corresponde a la parte restante de Gipuzkoa, al sudoeste de Lapurdi y al noroeste de Nafarroa Garaia.

El proceso de estandarización del euskara se inició 1968, por lo cual el euskara estándar o *batua* es, hoy por hoy, un concepto que la mayoría de los hablantes nativos no han asimilado para la versión hablada, aunque, en cuanto a la comunicación escrita, su uso es general en todos los ámbitos. Hay que señalar, además, que el euskara batua se creó con el objetivo de unificar el euskara escrito. Por lo tanto, no es de extrañar que los dialectos estén presentes en toda clase de comunicación oral, salvo en las más formales. Por otro lado, hay una idea muy extendida entre los teóricos de la dialectología (Zuazo, 2000), cuya premisa principal para la correcta formación del euskara batua en los próximos años es la promoción de los dialectos.

El euskara batua está basado, por cuestiones literarias, en los dialectos más centrales, y los dialectos de los extremos, entre ellos el occidental, quedaron en cierta manera relegados. Tanto es así, que el dialecto occidental tiene diferencias muy notables con respecto al estándar, sobre todo a nivel morfofonológico. Por tanto, un sistema de reconocimiento diseñado para el euskara batua

ISSN: 1135-5948

tiene obstáculos insalvables para reconocer el dialecto occidental (y sus subdialectos).

En la actualidad y como resultado de diversas iniciativas, hay varias bases de datos acústicas diseñadas para el reconocimiento del euskara estándar, por ejemplo, la base de datos SpeechDat_eu FDB1060 disponible en ELRA[1]. Existen otras bases de datos para entrenar y testear sistemas de reconocimiento, pero no están disponibles por el momento. En lo que se refiere al euskara dialectal, las bases de datos existentes se han diseñado para múltiples propósitos (sobre todo, lingüísticos), y no son apropiadas para ser usadas en sistemas de reconocimiento; además, tampoco se prevé que en el futuro se diseñen bases de datos dialectales para reconocimiento.

Esta situación nos llevó a analizar el comportamiento del reconocimiento del euskara occidental usando una base de datos de euskara estándar. Para dicho fin, se consideró la división de las entradas lexicales del diccionario del sistema en radicales y desinencias, ya que, de esa forma, se pueden abordar dos aspectos de la lengua: por una parte, la alta aglutinación de morfemas que el euskara en sí posee y, por otra, las variaciones subdialectales existentes dentro del euskara occidental. Cabe destacar que el dialecto occidental no es un dialecto en absoluto homogéneo, y puede clasificarse a su vez en varios subdialectos. En (Gaminde, 2002) se explican las características más importantes del euskara occidental, dividido en 8 subgrupos, atendiendo a motivos morfofonológicos.

2 Características fonológicas destacables del euskara occidental

En este apartado se presentan las características morfofonológicas más importantes del dialecto occidental del euskara, en contraposición con el estándar. Como se ha mencionado en el apartado anterior, el dialecto occidental es el dialecto que más variaciones posee y, por lo tanto, es el más complejo, debido a un gran número de reglas fonológicas y realizaciones fonéticas que no aparecen en otros dialectos.

2.1 Estructura morfológica general del euskara

El euskara es una lengua postposicional y aglutinante, rica en flexiones, tanto en la formación de verbos como de los sintagmas casuales. La descripción de la morfología del euskara es un tema muy extenso, por lo cual en este apartado se analizarán sólo los aspectos más relevantes. Primero, se expondrá el verbo perifrástico, para transmitir una idea de la complejidad flexiva del euskara, y, después, se describirán los sintagmas casuales, los cuales serán directamente tratados en este artículo.

Los verbos perifrásticos están formados de la siguiente manera:

[*radical verbal* + *sufijo aspectivo*] [*verbo aux*.]

El verbo auxiliar es altamente flexivo. Contiene las marcas de tiempo, modo y caso (absolutivo, ergativo o dativo). Por ejemplo, en euskara estándar «esaten zenizkidaten» (*me los decíais*) está formado por «esan» (*decir*, radical verbal «esa») + «ten» (aspecto frecuentativo), y el verbo auxiliar «zen+i+zki+da+Ø+te+n», donde «zen» es la marca de ergativo de segunda persona; «i» es la radical verbal del verbo auxiliar; «zki» es la marca de absolutivo plural; «da» es la marca de dativo de primera persona singular; «Ø» es la marca de indicativo; «te» es la marca de ergativo plural; y «n» es la marca de pretérito. Debido a esta complejidad, es habitual que en los trabajos de procesamiento de lenguaje natural se opte por tratar los verbos auxiliares como formas indivisibles, en vez de dividirlos en sus morfemas constituyentes (Alegria et al., 1996).

Los sintagmas casuales se construyen de acuerdo con el siguiente esquema:

radical nom. + (*artículo*) + (*número*) + *caso*(*s*)

donde los elementos entre paréntesis no tienen por qué aparecer siempre.

Es fundamental llevar a cabo todos los pasos de dicha secuencia para formar el sintagma casual; por ejemplo, «mutilarengana» (*hacia el chico*) está formado por la secuencia «mutil+a+r+en+ga+n+a», donde «mutil» (*chico*) es el radical nominal; «a» es el artículo singular; «r» es una consonante epentética de unión; «en» la marca de genitivo; «gan» la marca de ser animado; y «a» la marca de adlativo. Hay 17 casos en euskara, y algunos de ellos pueden concatenarse, luego el número de combinaciones posibles por lema es muy alto.

2.2 Estructura morfofonológica de la adaptación del artículo singular

Ésta es la característica más importante que se ha abordado en este trabajo. El artículo singular

[1] catalog.elra.info/product_info.php?products_id=5

«a» es siempre el primer morfema que se añade al lema, y los sufijos posteriores dependen del modo en que el artículo se adapta al lema. En euskara estándar, el resultado de la adaptación radical+artículo tiene un solo resultado, pero, en los dialectos, debido a procesos fonológicos de asimilación y disimilación —y subsiguiente *feeding* (Kiparsky, 1968)—, la adaptación del artículo depende del último fonema del lema. Estas adaptaciones, sin embargo, no son homogéneas en el área donde se habla el dialecto occidental, y, atendiendo a esta característica, dicha área se puede clasificar en ocho regiones subdialectales.

Los finales resultantes para cada fonema final más artículo se muestran en la Tabla 1, tanto para cada una de las ocho regiones subdialectales como para el estándar (los símbolos utilizados para representar los fonemas son los de SAMPA —*Speech Assessment Methods Phonetic Alphabet*—; véase http://aholab.ehu.es/sampa_basque.htm). Nótese que «-e+a» indica un lema terminado en «e» más el artículo «a». Las combinaciones de la columna «-C+a» (lema acabado en consonante m) tienen dos realizaciones en casi todas las regiones, debido a un proceso de asimilación vocálica: si la última vocal anterior a la consonante es una vocal cerrada (/i/ ó /u/), el artículo se cierra a /e/.

	-a+a	-e+a	-i+a	-o+a	-u+a	-C+a
A	[ea]	[ea]	[ie]	[oa]	[ue]	[Ca]/[Ce]
B	[ia]	[ia]	[ie]	[oa]	[ue]	[Ca]/[Ce]
C	[ie]	[ie]	[iZe]	[oa]	[ue]	[Ca]/[Ce]
D	[ia]	[ia]	[iZe]	[oa]	[ue]	[Ca]/[Ce]
E	[ie]	[ie]	[iZe]	[ue]	[ue]	[Ca]/[Ce]
F	[i]	[i]	[iZe]	[u]	[u]	[Ca]/[Ce]
G	[ia]	[ia]	[iZa]	[ua]	[ua]	[Ca]
H	[e]	[e]	[i]	[o]	[u]	[Ca]/[Ce]
Estándar	[a]	[ea]	[ia]	[oa]	[ua]	[Ca]

Tabla 1: Realizaciones fonéticas de las diferentes combinaciones de fonema final más artículo, por región (A-H) y en estándar

En algunos de los grupos radical+artículo no se encuentra la variación correspondiente al estándar, el cual no presenta ningún cambio fonológico; por ejemplo, en el caso «-a + a» hay cinco combinaciones diferentes para las ocho regiones, pero ninguna es la misma que en el estándar: «neska+a» (*la chica*), según la región, adopta las formas «neskea», «neskia», «neskie», «neski», «neske», pero nunca «neska», como en batua (Oñederra, 2005). Cabe decir, además, que la adaptación del artículo no sólo afecta a la realización del artículo, sino que también cambia el último fonema del lema (además de introducir, como en el caso «-i + a», una consonante extra). Por consiguiente, aparecen varios alomorfos, tanto para el lema como para el artículo, fenómeno que sólo ocurre en el euskara estándar para el caso -/a/, cuando se añade la marca del artículo plural. En las Tablas 2 y 3 se muestran los alomorfos de los lemas y de los artículos, respectivamente.

Último fonema de lema	Alomorfos
-/a/	-[e], -[i]
-/e/	-[e], -[i]
-/i/	-[i]
-/o/	-[o], -[u]
-/u/	-[u]
-/C/	-[C]

Tabla 2: Alomorfos de los lemas, debido a la adaptación del artículo al lema

Región	Alternancias fonéticas del artículo
A	[a] ~ [e]
B	[a] ~ [e]
C	[a] ~ [e] ~ [Ze]
D	[a] ~ [e] ~ [Ze]
E	[a] ~ [e] ~ [Ze]
F	[a] ~ [e] ~ [Ze] ~ [Ø]
G	[a] ~ [Za]
H	[a] ~ [e] ~ [Ø]
Estándar	[a]

Tabla 3: Alomorfos del artículo, debido a la adaptación del artículo al lema

En cuanto a los lemas, se observa que los lemas terminados en -/a/, -/e/ y -/o/ tienen dos alomorfos cada uno, y el resto sólo uno (en euskara estándar todos tienen un único alomorfo). En cuanto al artículo, aparecen cuatro realizaciones fonéticas diferentes en la región F, y como mínimo se producen dos realizaciones; en euskara batua sólo una realización es posible.

2.3 Los participios verbales

Una cantidad importante de verbos ha sido tomada prestada (en su forma participial) de las

lenguas vecinas, especialmente de las lenguas romances y, hoy día, del español. En español hay dos terminaciones posibles para los participios verbales: «-ado» e «-ido». Los verbos que proceden de uno con participio acabado en «-ado» pueden adoptar las siguientes realizaciones fonéticas, debido a procesos fonéticos: -[atu], -[au], -[eu], -[a]; y los que provienen de uno con participio terminado en «-ido» tienen las realizaciones -[itu], -[idu], -[iu]. Cada realización es específica de la región.

En este artículo hemos denominado *Verbos recientes* a aquellos que son prestados de lenguas vecinas y son susceptibles de sufrir dichos cambios fonéticos, y *Verbos antiguos* a aquellos que no son prestados y no sufren dichos cambios.

2.4 Otros cambios fonéticos

En el dialecto occidental existen varios cambios fonéticos respecto al estándar; por ejemplo, hoy día, no existe el fonema fricativo postalveolar sordo /s'/, el cual ha sido sustituido por el fonema [s], fricativo alveolar sordo. Por consiguiente, tanto el /s'/ como el /s/ del euskara estándar se pronuncian [s] en el dialecto occidental. El fenómeno de la palatalización de consonantes siguientes al fonema /i/ es también muy fuerte en el dialecto occidental, y surgen realizaciones que no aparecen en ningún otro dialecto.

Este tipo de cambios son, principalmente, de fonema a fonema.

3 Descripción de las bases de datos utilizadas

Tal y como se ha explicado en la introducción, el objetivo de este trabajo es utilizar una base de datos acústica de euskara estándar para entrenar el sistema de reconocimiento, y realizar el test utilizando habla dialectal. Por lo tanto, se han elegido dos bases de datos para los experimentos presentados en este artículo: SpeechDat_eu (Hernáez et al. 2003) y Bizkaifon (Castelruiz et al., 2004). La primera base de datos está diseñada para el euskara batua y contiene grabaciones realizadas sobre la red de telefonía fija. Esta base de datos es la que se ha utilizado como punto de partida para construir los modelos acústicos de referencia. La segunda base de datos contiene sólo habla dialectal del euskara occidental y es la utilizada para realizar la evaluación de los experimentos descritos en el siguiente apartado.

Hay diferencias notables entre las dos bases de datos, como puede apreciarse en la Tabla 4.

	SpeechDat_eu	Bizkaifon
Tipo de habla	Euskara estándar	Dialecto occidental
Canal	Telefonía fija	Micrófono (cinta digital y magnét.)
Frecuencia de muestreo	8 kHz	8, 16 y 32 kHz
Ficheros de audio	5.200	11.868
Lexicón	3.968 palabras estándar	8.199 formas dialectales (5.224 estándar)
Préstamos no admitidos	0 (0%)	993 (12,38%)
Género de locutores	Hombres: 45,28% Mujeres: 55,72%	Hombres: 4,70% Mujeres: 95,30%

Tabla 4: Diferencias más importantes entre las bases de datos SpeechDat_eu y Bizkaifon

La base de datos SpeechDat_eu contiene grabaciones de 1.060 locutores (480 hombres, 580 mujeres) grabados sobre la red de telefonía fija, y cumple las condiciones especificadas en el proyecto SpeechDat (Winski, 1997). Aunque la base de datos contiene, sobre todo, euskara estándar, una pequeña parte de las grabaciones se hicieron para habla dialectal, por medio de cuestionarios 'dialectales'. Esto se hizo para obtener grabaciones que tuvieran una representación del inventario completo de los sonidos del euskara de Hego Euskal Herria. De hecho, el fricativo prepalatal sonoro [Z] se realiza, principalmente, como consecuencia de las reglas morfofonológicas descritas anteriormente (véase la Tabla 1), y no pertenece al inventario de sonidos del euskara batua. En el lexicón de la SpeechDat_eu este sonido está considerado como una pronunciación alternativa.

La base de datos Bizkaifon incluye 11.868 ficheros de audio de palabras aisladas. Como se puede observar en la Tabla 4, la mayoría de los locutores son mujeres, y las grabaciones se han obtenido por diferentes medios (diferentes micrófonos, cinta magnética, cinta digital...). Además, han sido digitalizadas con diferentes frecuencias de muestreo. Antes de comenzar los experimentos, todos los ficheros se adaptaron a la frecuencia de muestreo de 8 kHz.

Cada fichero de Bizkaifon contiene una transcripción ortográfica de la forma dialectal y otra de su equivalente en estándar. El primer problema con el que nos enfrentamos fue la existencia de palabras dialectales que no están aceptadas en el euskara estándar, principalmente porque son préstamos relativamente recientes y de uso local. Por ejemplo, la palabra dialectal «abuelie» proviene del español «abuela», el cual no se acepta en el estándar, en el que se utilizan los tradicionales «amama» o «amona», de uso general. En este trabajo, se ha optado por eliminar todos los ficheros de audio que contuvieran préstamos no admitidos (1.275), con el fin de centrarnos únicamente en los problemas que causan las variaciones fonológicas. Así, se redujo el número de ficheros a 10.593, entre los cuales hay 7.029 formas dialectales diferentes que se corresponden con 4.721 palabras del euskara estándar. De esas 4.721 palabras, sólo 1.143 coinciden con su versión dialectal.

La base de datos ha sido etiquetada morfológicamente usando un procedimiento semiautomático. Se utilizó un analizador morfológico de euskara estándar (Ezeiza, 1998) junto con las transcripciones de euskara estándar, para, así, obtener las etiquetas morfológicas para las formas de palabra dialectales. Posteriormente, las etiquetas se revisaron manualmente. La distribución morfológica de la base de datos Bizkaifon se muestra en la Tabla 5.

Categoría gramatical	Número de ficheros	Número de formas dialectales
Nombres + adj (art. sg.)	8.239	5.466
-a	1.874	1.508
-e	898	681
-i	1.290	826
-o	1.252	889
-u	746	474
-C	2.179	1.088
Nombres + adj (art. pl.)	345	253
Verbos (participios)	1.435	859
Verbos recientes	510	267
Verbos antiguos	925	592
Otros	574	451
Total	10.593	7.029

Tabla 5: Análisis morfológico de la base de datos Bizkaifon

Los nombres y los adjetivos se han clasificado en el mismo grupo, ya que se comportan de forma similar en el contexto de palabras aisladas. Este grupo, por su cantidad, forma el núcleo de la base de datos, y, tal y como se observa en la tabla, la mayoría de las palabras de este grupo tienen añadido el artículo singular; de hecho, sólo en raras ocasiones sucede que los sintagmas nominales no lleven artículo, lo cual nos indica el diferente significado y uso del artículo en euskara en comparación con las lenguas circundantes. El resto de las palabras de este grupo formado por nombres y adjetivos son palabras que no contienen artículo o con artículo plural (en el cual no surgen apenas cambios fonéticos, a diferencia de lo que ocurre con el artículo singular). Además, existen también participios verbales, tanto de verbos recientes como antiguos, aunque en menor cantidad.

4 Experimentos

El objetivo de los experimentos presentados en este apartado es conseguir un reconocedor de palabras aisladas de euskara dialectal, utilizando modelos acústicos entrenados para el euskara estándar. Por lo tanto, el fin es obtener una salida en euskara estándar, sea cual sea la variante dialectal de entrada (véase la Figura 1).

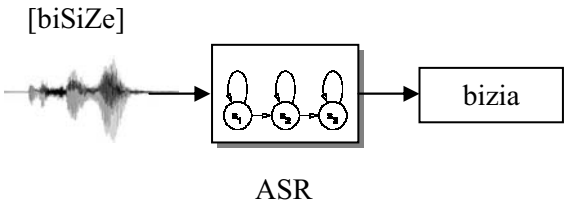

Figura 1: Descripción de la entrada (dialectal) y salida (estándar) del sistema de reconocimiento

Todos los experimentos de reconocimiento de este trabajo se han llevado a cabo utilizando el software de reconocimiento *HTK Speech Recognition Toolkit* (Young et al., 2000). En este apartado, se describe, en primer lugar, un experimento preliminar utilizando el diccionario de pronunciaciones del euskara estándar, seguido de un experimento con el diccionario dialectal creado a partir de las transcripciones dialectales obtenidas manualmente. En el apartado 4.2 se presenta y analiza el tratamiento del léxico basado en la división de radical y desinencia, y, finalmente, en el apartado 4.3 se realiza un análisis de los resultados.

4.1 Experimento preliminar

Como punto de partida para el desarrollo de un sistema de reconocimiento de palabras aisladas para el dialecto occidental, consideramos los experimentos de evaluación realizados con la base de datos SpeechDat_eu para el euskara estándar. El experimento de palabras aisladas descrito en (Hernáez et al., 2003) presenta un WER (*Word Error Rate*) de 17,20% con un diccionario de 3.968 entradas léxicas (LE, *Lexical Entry*) para el subcorpus de palabras fonéticamente ricas. Los modelos acústicos comprenden el inventario de sonidos del euskara estándar, y están basados en modelos de Markov para trifonemas con 3 estados y mezcla de 32 gaussianas. Los parámetros acústicos utilizados fueron los primeros 13 coeficientes MFCC y sus primeras y segundas diferencias (un total de 39 parámetros obtenidos cada 10 ms para un ancho de trama de 25 ms).

Con los 4.721 lexemas en estándar (transcripciones ortográficas en estándar) y utilizando un transcriptor basado en reglas grafema-fonema (G2P) se ha creado un diccionario de euskara estándar, utilizando el alfabeto fonético SAMPA. Dicho conversor G2P se vale de las recomendaciones de pronunciación dictadas por la Academia de la Lengua Vasca (Euskaltzaindia, 1998).

El resultado de testear los ficheros de audio de Bizkaifon utilizando un diccionario de euskara estándar y modelos acústicos generados a partir de los ficheros de SpeechDat_eu produce un WER de 46,27%. Era de esperar un resultado tan pobre, por tres razones:

- El lexicón dialectal tiene 7.029 formas diferentes, sólo 1.143 (16,26%) coinciden con su correspondiente versión estándar.
- La existencia del sonido /Z/ en el inventario de sonidos del dialecto occidental y su alta frecuencia de aparición en Bizkaifon. Este sonido no fue explícitamente modelado al generar los modelos acústicos de la SpeechDat_eu.
- Las importantes diferencias acústicas entre ambas bases de datos, lo que, sin duda, tiene una influencia significativa en los resultados.

Para conocer mejor la influencia de estos factores en los experimentos de reconocimiento, se realizó un experimento de reconocimiento 'dialectal'. Primero, se creó un diccionario dialectal con las 7.029 formas dialectales como lexemas obteniendo su transcripción fonética por medio del conversor G2P del laboratorio. Además, se creó el modelo acústico del fonema /Z/ partiendo de los ficheros de la SpeechDat, aunque solamente con 71 muestras. Finalmente, se testearon los 10.593 ficheros de audio de palabras aisladas de Bizkaifon.

En este reconocimiento dialectal, se obtuvo un WER de 23,14%. El diccionario se ha generado utilizando sólo las formas dialectales que pertenecen a la base de datos, luego este porcentaje representa el límite superior para los siguientes experimentos. Por tanto, aunque las diferencias acústicas de ambas bases de datos sean considerables, el resultado es alentador.

Para completar el análisis de errores, se llevó a cabo un experimento más, adaptando los modelos acústicos de la SpeechDat_eu con los ficheros de audio de Bizkaifon. Se reentrenaron los modelos de trifonemas hasta que se obtuvo un WER estable de 2,66%. Este resultado demuestra que las diferencias acústicas entre las dos bases de datos son muy significativas. No obstante, para el siguiente experimento se han utilizado exclusivamente los modelos acústicos originales.

4.2 Experimento con división de léxico en radical y desinencia

Aunque en el contexto de reconocimiento de palabras aisladas puede ser factible el uso de un diccionario que contemple todas las variaciones fonéticas y morfofonológicas dialectales, es necesario adoptar una estrategia de unidades menores que la palabra, sobre todo en el caso de lenguas aglutinantes, para la evolución a un sistema de reconocimiento de habla continua.

El carácter aglutinante del euskara permite dividir los sintagmas casuales (y los verbos, pero en este trabajo sólo se consideran los participios, los únicos que se encuentran en la base de datos) para tratarlos como radicales y desinencias. En euskara hay 17 casos en 4 contextos numéricos diferentes (singular, plural, plural próximo e indefinido), luego tenemos 68 desinencias distintas que, concatenadas en segundo grado, dan 275 casos en total (Alegria et al., 1996). Además, el artículo tiene una media de 4.17 alomorfos por lema en el dialecto occidental; por tanto, la cantidad de desinencias fonéticamente diferentes en el dialecto occidental asciende, en el segundo grado de concatenación, hasta 1.146,75 desinencias por lema sustantivo. Es por eso que es tan importante considerar

unidades subpalabra en euskara y, sobre todo, en el dialecto occidental.

Último fonema de lema	radical	desinencias	
		eusk. occ.	bat.
-/a/	lema – /a/	-[ea],-[ia],-[ie],-[i],-[e]	-[a]
-/e/	lema – /e/	-[ea],-[ia],-[ie],-[i],-[e]	-[ea]
-/i/	lema	-[Ø],-[Za],-[Ze],-[e]	-[a]
-/o/	lema – /o/	-[oa],-[ua],-[ue],-[u],-[o]	-[oa]
-/u/	lema	-[Ø],-[a],-[e]	-[a]
-/C/	lema	-[a],-[e]	-[a]

Tabla 6: Radicales y desinencias considerados para los nombres y adjetivos

Considerando como radical la parte del lema que, cuando se flexiona, no varía, en los lemas de los nombres que acaban en -/a/, -/e/ y -/o/ no abarcan el último fonema (véase la Tabla 2), mientras que en los casos restantes el radical y el lema coinciden. La Tabla 6 muestra los radicales y las desinencias que se han considerado en este trabajo para el grupo de nombres y adjetivos con artículo singular; por ejemplo, si consideramos el lema «neska» que acaba en -/a/, el radical es «nesk» (lema – /a/), y con las desinencias que le corresponden (-«ea», -«ia», -«ie», -«i», -«e») se obtienen todas las variedades posibles del apartado 2.2.

Categoría gramatical	Cant.	Prop.	Cant.	Div.?
nombres	5.098	singular	4.854	S
		plural	137	N
		indefinido	107	N
adjetivos	621	singular	612	S
		plural	5	N
		indefinido	4	N
verbos	859	participio reciente	267	S
		participio antiguo	592	N
adverbios	281			N
sintagmas postp.	124			N
otros	46			N

Tabla 7: Contenido de la base de datos Bizkaifon, desde un punto de vista morfológico

La base de datos Bizkaifon está formada por 7.029 formas dialectales diferentes, de las cuales 5.733 se han dividido en radical +desinencia. En la Tabla 7 se muestra el contenido de la base de datos desde el punto de vista morfológico con mayor detalle. También se indica qué palabras han sido divididas.

Para permitir que cada radical vaya seguido de las desinencias alomórficas que le corresponden, se ha generado la gramática mostrada en la Figura 2. Además, se han tenido en cuenta los contextos de interior de palabra para modelar correctamente las transiciones de radical a desinencia (Rotovnik et al., 2007).

Por otra parte, en el diccionario del sistema se han introducido, como alternativas, los cambios fonéticos del apartado 2.4, para paliar los efectos de dichos cambios fonema a fonema.

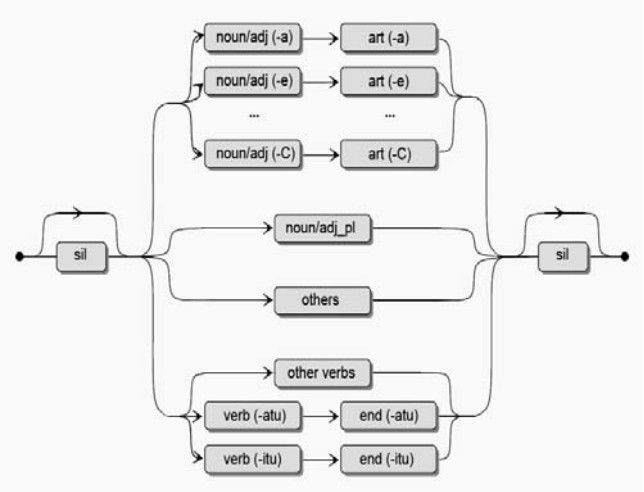

Figura 2: Esquema de la gramática utilizada para generar las combinaciones de radical+desinencia adecuadas

4.3 Resultados

Los resultados obtenidos se muestran en la Tabla 8. En dicha tabla, se han reflejado los resultados obtenidos con y sin contexto de transición del interior de palabra. Como puede apreciarse en la última fila, el uso del contexto del interior de palabra mejora mucho los resultados. Así, el WER total que se obtiene es de 28,82%, el cual, desde el 46,27% inicial, se aproxima notablemente al límite superior teórico de 23,14% (véase el apartado 4.1). La diferencia restante puede atribuirse tanto a los cambios fonéticos que no han sido modelados, como a la introducción de confusión en el sistema, debido a que el sistema, en este último

experimento, contempla todas las posibilidades que un lexema puede tener y no sólo las que están presentes en la base de datos, lo cual disminuye en cierta medida su eficacia.

	Sin contexto	Con contexto
ER de radicales	41,64	25,77
ER de desinencias	19,20	9,43
ER de desinencias en radicales correctos	1,37	1,65
ER de lexemas no divididos	27,22	37,15
WER total	39,77	28,82

Tabla 8: Resultados del experimento basado en la división del lexema en radical+desinencia

En la primera y segunda fila de la tabla se muestran, respectivamente, la tasa de error (ER, *Error Rate*) de radicales y desinencias.

En la tercera fila se muestran las desinencias erróneamente reconocidas para aquellos radicales bien reconocidos. Es obvio que es una cifra muy baja, ya que a cada grupo de lemas le corresponde en la salida una sola desinencia en estándar (salvo en los casos en que los nombres y los verbos comparten el mismo radical).

La cuarta fila de la tabla muestra los errores de las palabras que no han sido divididas. Se observa que el hecho de considerar el contexto del interior de palabra reduce notablemente la tasa de error de los radicales y desinencias, pero da peores resultados para las palabras no divididas. Esto es debido a que la consideración de los contextos añade cierta confusión al reconocimiento de dichas palabras. Además, el análisis de los errores nos ha permitido apreciar que la mayoría de los errores en palabras no divididas sucede en verbos antiguos, los cuales son palabras muy cortas (4 ó 5 fonemas).

5 *Discusión final de los resultados*

En este trabajo se ha utilizado una base de datos acústica del euskara estándar para reconocer el habla del dialecto occidental. Como muestra el experimento preliminar, este dialecto necesita un tratamiento específico para abordar, por una parte, la característica fuertemente aglutinante del euskara, y, por otra, las variaciones fonéticas características del dialecto occidental.

La división de los lexemas del diccionario del sistema en radical+desinencia permite que dicho diccionario no crezca de acuerdo con el número de desinencias (recuérdese que en el euskara estándar son posibles 275 casos en la concatenación de segundo grado), además de un procesamiento de los datos mucho más rápido. Por otra parte, se pueden modelar los diferentes alomorfos de las desinencias (para el artículo singular hay 4,17 alomorfos por lema), con lo que se obtiene una mejora del 17,45% en el resultado del reconocimiento y nos acercamos al límite de error de 23,14% en el WER (véase el apartado 4.1). Teóricamente, hay aún un 5,66% de mejora posible, pero hay que tener en cuenta la confusión que se introduce en el sistema debido a que, en la aproximación de división de lexemas, se contemplan todas las posibilidades de formación de cada palabra. Por tanto, en la práctica, siempre habrá un margen de error que impedirá llegar exactamente a dicho límite superior.

Otro elemento importante que ha tenido gran influencia en los experimentos son las diferencias acústicas de las dos bases de datos, que son, tal y como se ha demostrado, las causantes principales de que el límite teórico se sitúe en 23,14% para los experimentos.

El siguiente paso a este trabajo es la extensión de esta aproximación al habla continua, para abrir el rango de aplicaciones. El euskara batua tiene pocas recomendaciones para su pronunciación, y se habla, sobre todo, en situaciones muy formales. Por tanto, es muy frecuente encontrar habla dialectal en programas de radio y televisión. El hecho de considerar tanto habla dialectal como estándar podría mejorar notablemente el rendimiento de los sistemas de reconocimiento.

El mayor problema con el que nos encontramos para avanzar por esa línea es la falta de bases de datos dialectales, por lo cual es interesante la posible adaptación de este sistema para obtener una salida dialectal. De este modo, se podría crear un transcriptor semiautomático (con ayuda de correcciones manuales), para generar nuevas bases de datos y utilizar los textos dialectales así obtenidos para explorar otras técnicas de extracción de conocimiento de forma estadística.

6 *Agradecimientos*

Este trabajo ha sido parcialmente financiado por el Ministerio de Educación y Ciencia dentro del proyecto AVIVAVOZ (TEC2006-13694-C03-02, www.avivavoz.es) y por el Gobierno Vasco en su subvención a grupos de investigación del sistema universitario vasco (IT-444-07).

Bibliografía

Alegria, I., Artola, X., Sarasola, K. and Urkia, M. 1996. Automatic morphological analysis of Basque. *Literary & Linguistic Computing* 11(4):193-203.

Castelruiz, A., Sánchez, J., Zalbide, X., Navas, E., Gaminde, I. 2004. Description and Design of a WEB Accesible Multimedia Archive. En *Proceedings of 12th IEEE Mediterranean Electrotechnical Conference, MELECON*, páginas 681-684, Dubrovnik.

Euskaltzaindia (Academia de la Lengua Vasca), 1998. Euskara batuaren ahoskera zaindua. En *Euskaltzaindiaren Arauak*, páginas 805-808, Bilbao.

Ezeiza, N., Aduriz, I., Alegria, I., Arriola, J.M., Urizar, R. 1998. Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages. En *Proceedings of COLING-ACL'98*, páginas 379-384, Montreal.

Gaminde, I. 2002. Bizkaiko euskararen ezaugarri fonologiko batzuen inguruan. *Euskalingua* 1:4-14.

Hernáez, I., Luengo, I., Navas, E., Zubizarreta, M., Gaminde, I., Sánchez, J. 2003. The Basque Speech_Dat (II) Database: A Description and First Test Recognition Results. En *Proceedings of EUROSPEECH*, páginas 1549-1552, Geneva.

Kiparsky, P. 1968. Linguistic Universals and Linguistic Change. En *Universals in Linguistic Theory*, páginas 170-202, eds. Bach and Harms.

Oñederra, M. L. 2005. Fonologiaren mugak: alabea eta birjinak elexan. En *Nerekin yaio nun. Txillardegiri omenaldia*, páginas 379-397, IKER 17, Euskaltzaindia, Bilbao.

Rotovnik, T., Sepesy Maucec, M, Kacic, Z. 2007. Large vocabulary continuous speech recognition of an inflected language using stems and endings. Speech Communication 49(6):437-452.

Winski, R. 1997. Definition of corpus, scripts and standards for Fixed Networks. LE2-4001-SD1.1.1, Informe técnico.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P. 2000. *The HTK Book*, Cambridge University, Cambridge.

Zuazo, K. 2003. *Euskalkiak, Herriaren lekukoak*. Elkar argitaletxea, Donostia.

Zuazo, K. 2000. *Euskararen sendabelarrak*. Alberdania argitaletxea, Irun.