ONOMÁZEIN
Revista semestral de lingüística, filología y traducción

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
FACULTAD DE LETRAS

El reconocimiento del habla ante la variedad de una lengua: la africada dento-alveolar andaluza

Speech recognition faced with a language variety: Andalusian dento-alveolar affricate

María Del Saz
Universidad de Santiago de Chile
Chile

ONOMÁZEIN 34 (diciembre de 2016): 278-295
DOI: 10.7764/onomazein.34.17

María Del Saz: Departamento de Lingüística y Literatura, Facultad de Humanidades, Universidad de Santiago de Chile, Chile. | Correo electrónico: maria.delsaz@usach.cl

Fecha de recepción: agosto de 2015
Fecha de aceptación: febrero de 2016

Resumen

La calidad del reconocimiento de voz dentro de las tecnologías del habla depende del tipo de sistema utilizado y de la aplicación para la que se destine. Uno de los problemas que puede presentar es el grado de robustez ante variedades dialectales de una lengua. En este estudio se analiza la tasa de reconocimiento de la aspiración andaluza ante oclusiva dento-alveolar sorda usando Google Voice Search y Windows 7 sin entrenamiento, por una parte, y usando Windows 7 con diferente número de sesiones de entrenamiento con un locutor, por otra. Inicialmente, Google Voice Search tiene mayor robustez ante Windows 7, hecho que Windows 7 compensa a medida que aumentan las sesiones de entrenamiento. En general, se observa que la robustez del reconocimiento se ve afectado por factores como el número de candidatos léxicos posibles y el contexto de las entradas.

Palabras clave: reconocimiento de voz automático; tecnologías del habla; variedades lingüísticas; fonética.

Abstract

The quality of voice recognition within the field of speech technologies depends on the type of system used and the application to which it is destined. One of the problems that it can present is its degree of robustness with dialectal variants of a language. In this study, an analysis is done of the recognition rate of Andalusian aspiration before the voiceless dento-alveolar plosive by using Google Voice Search and untrained Windows 7, on the one hand, and Windows 7 after a number of training sessions with a speaker, on the other hand. Initially, Google Voice Search has greater robustness than Windows 7, which Windows 7 compensates as the number of training sessions increase. In general, it can be observed that recognition robustness is affected by factors such as the number of possible lexical candidates and the context of the entries.

Keywords: automatic speech recognition; speech technologies; linguistic varieties; phonetics.

1. Introducción

El reconocimiento automático de habla se utiliza para numerosas aplicaciones, tales como servicios telefónicos, búsqueda de información en internet o dictado automático. Sus ventajas son numerosas, ya que basa la forma de comunicación humano-máquina en la forma de comunicación humana básica: la comunicación oral. Por tanto, ofrece una interacción más rápida y natural que la forma tradicional, puesto que permite una mayor movilidad al no depender de teclado, incluye acceso remoto y proporciona una mayor eficacia al contar con algoritmos más eficientes, equipos más potentes y modelados más sofisticados. A diferencia de la síntesis de habla, que apareció por primera vez a mediados del siglo XVIII, el reconocimiento de habla no comenzó a intentarse hasta principios del siglo XX (Fandiño, 2005). En este caso, se trata de un sistema que toma sus unidades de reconocimiento de un extenso corpus oral con el que ha sido entrenado, compara la señal de voz recibida con estas unidades y reproduce un texto según el modelo de lenguaje que se le ha incorporado mediante un corpus textual.

Para este uso de las tecnologías del habla, debemos considerar cinco factores fundamentales. El primero de ellos, el locutor, implica tanto variabilidad en la señal acústica a nivel intralocutor como en la señal a nivel interlocutor. Para poder hacer frente a esta característica, necesitamos sistemas muy robustos. A diferencia de la síntesis de habla, que toma sus unidades a partir de un solo locutor, en el reconocimiento de habla podemos encontrar sistemas de dos tipos: dependientes del locutor, donde el sistema se entrena para reconocer a un determinado hablante, e independientes del locutor, donde el sistema está preparado para reconocer a una gran variedad de hablantes de una lengua. En segundo lugar, consideramos la forma de hablar del locutor respecto al tipo de unidades que va a reconocer el sistema. Las más sencillas son las palabras aisladas, seguidas de palabras dentro de una frase, llegando a la más complicada, aunque más natural, que es el habla continua, donde se aúnan efectos coarticulatorios y prosódicos. Unido a este factor encontramos un tercero, que trata sobre el vocabulario recogido por el sistema. Un vocabulario mayor implica un mayor tiempo de tratamiento al tener más palabras con las que contrastar la entrada de voz. Para paliar este efecto, se recurre a un inventario de

alófonos y difonemas, con su frecuencia de aparición, unidos a las reglas fonotácticas de una lengua. Un cuarto factor a considerar es la gramática de la lengua, es decir, las reglas que componen el idioma en los diferentes niveles lingüísticos, las cuales ayudan a los sistemas de reconocimiento de voz al eliminar ambigüedades y reducir el tiempo de tratamiento. Ya por último, también influye el entorno físico en el que se desarrolla la entrada de voz, el ruido ambiental, el ancho de banda del sistema, etc. Para que un sistema sea eficaz, debe ser robusto en la detección de voz, en la reducción de ruido y en la cancelación de eco.

Las primeras técnicas utilizadas para llevar a cabo el reconocimiento de habla fueron topológicas, como el Alineamiento Temporal mediante algoritmos de Programación Dinámica (DTW), que mide las distancias entre patrones. Les siguieron las probabilísticas, como los Modelos Ocultos de Markov (HMN), que emplean un modelo estadístico para cada palabra. Después llegaron las Redes Neuronales Artificiales (RN), basadas en la interconexión de unidades de proceso en paralelo. Por último, se desarrollaron los sistemas basados en el conocimiento, que no solo se usan para el reconocimiento de habla, sino también para la comprensión de habla, agregando además módulos de análisis semántico, pragmático y de conocimiento del mundo (Fandiño, 2005). En cualquier caso, en todos ellos se siguen tres fases: parametrización, entrenamiento y reconocimiento (Hernández, 2003).

Dentro de la fase de parametrización del corpus de entrenamiento, se segmenta y se etiqueta cada sonido y cada unidad de reconocimiento, y se da su transcripción fonética y su representación ortográfica. A su vez, en los diccionarios de pronunciación se incluye la forma canónica de los sonidos junto a las reglas fonéticas para sus variantes. Esta información fonética tiene en cuenta los rasgos distintivos de la lengua, el efecto de la coarticulación y la intra- e intervariabilidad de los locutores, así como el acento de palabra y frase, y la prosodia. De hecho, esta variabilidad es uno de los factores en detrimento de la robustez del sistema (Ferreiros, 2003). Uno de los requisitos para reducir este efecto es que la información proporcionada al sistema en la fase de parametrización "tiene que incluir necesariamente hablantes de las principales variantes dialectales de la lengua" (Llisterri y

otros, 2003: 19). Nogueiras y otros (2002) presentaron un modelo acústico multidialectal para el reconocimiento de las variedades del español, usando diversas bases de datos disponibles para el entrenamiento del sistema, con buenos resultados. En él, se incluyó una transcripción fonética basada en reglas para cada variante (Colombia, Venezuela, España), así como los fonemas comunes a todas y específicos de cada una. Sin embargo, para la variante de España se tomó como base el castellano, no incluyendo variedades dialectales como el andaluz. Aunque en el habla de Venezuela y Colombia se indica que la /s/ en posición final se aspira, dando lugar a [s] → [h], un sistema así podría presentar problemas para reconocer la variante dialectal que se propone estudiar en este artículo, ya que se caracteriza por post-aspiración más que por pre-aspiración e incluso por una articulación africada más que oclusiva.

En este estudio, comprobamos el reconocimiento de palabras que contienen la secuencia <st> en español andaluz occidental, donde la aspiración de /s/ en posición final se refleja en la dental oclusiva sorda /t/ como una post-aspiración (Torreira, 2007; Parrell, 2012), dando lugar a [t^h]. Si vamos un poco más allá, en la ciudad de Sevilla, en concreto, existe una variación emergente de este sonido que es africada, expresada como [tʂ] por Moya Corral (2007) o [t^s] por Ruch (2010), pudiéndose confundir con [tʃ]. "Se trata de un sonido africado, que se articula en la zona dento-alveolar con la parte más adelantada del predorso de la lengua" (Moya Corral, 2007: 457). Este estudio se centra en el reconocimiento automático de palabras que contienen esta variante, comparando, en primer lugar, Google Voice Search y Windows 7 de forma independiente del locutor y, en segundo lugar, la robustez de Windows 7 tras varias sesiones de entrenamiento con un mismo locutor. Aunque nuestro objetivo es conocer el grado de robustez del reconocimiento del habla de estos dos sistemas frente a esta variante fonética, no solo se consideran los componentes fonéticos y fonológicos en el análisis de resultados, sino que también se tienen en cuenta factores morfológicos, sintácticos y semánticos, tales como el número de competencias léxicas que presentan las palabras clave, así como el apoyo gramatical que éstas poseen cuando se incluyen en sintagmas.

2. Metodología

2.1. Sujetos

Para este estudio contamos con dos sujetos, un hombre y una mujer de Sevilla, con una edad media de 32 años, hablantes de andaluz occidental. Los dos participaron en el primer experimento, que compara Google Voice Search con Windows 7, mientras que solo la mujer participó en el segundo experimento, que compara los resultados obtenidos con diferentes sesiones de entrenamiento de reconocimiento de voz de Windows 7.

2.2. Corpus

El corpus utilizado consta de 30 palabras aisladas, bisílabas y llanas (tabla 1). Quince de ellas siguen la estructura CVs.tV, 10 de ellas tienen la estructura CV.tV, y las otras cinco se componen de CV.chV. En el grupo 1, tenemos 5 palabras de cada una de las estructuras, que podrían considerarse pares mínimos (pesto-peto-pecho). En el grupo 2 encontramos 5 palabras de la primera y la segunda estructuras (pasta-pata), mientras que en el grupo 3 solo tenemos 5 palabras de la primera estructura, sin competencia léxica (susto), es decir, donde no encontramos palabras con las que puedan confundirse (*suto* o *sucho*), como en los otros dos grupos.

Tabla 1. Listado de palabras aisladas por estructura y número de competencias léxicas

	CVs.tV	CV.tC	CV.chV
Grupo 1	pista	pita	picha
	mosto	moto	mocho
	asta	ata	hacha
	casta	cata	cacha
	pesto	peto	pecho
Grupo 2	pisto	pito	
	pasto	pato	
	pasta	pata	
	costo	coto	

	mixto	mito
Grupo 3	chiste	
	lista	
	cesta	
	gesto	
	susto	

A su vez, a las 15 palabras que contienen <st> se les añadió un calificativo para ayudar a su identificación semántica, constituyendo los sintagmas de la tabla 2.

Tabla 2. Listado de sintagmas por número de competencias léxicas

Grupo 1	Grupo2	Grupo 3
pista falsa	pisto manchego	chiste malo
mosto de uva	pasto verde	lista de la compra
asta de toro	costo total	cesta de navidad
antigua casta	pasta italiana	bonito gesto
salsa pesto	sándwich mixto	susto de muerte

Por tanto, el corpus utilizado está constituido por 45 secuencias (30 palabras aisladas + 15 sintagmas).

2.3. Variables

Este estudio se centra en el reconocimiento de una variante dialectal de la secuencia <st> del español. Para determinar la robustez ante esta variabilidad, tomamos diversos parámetros. En primer lugar, el sistema utilizado. Primero usamos Google Voice Search y Windows 7 sin entrenamiento, con lo que se trata de sistemas independientes del locutor. Después utilizamos Windows 7 con varias sesiones de entrenamiento (una, tres, cinco, siete) y comparamos el reconocimiento entre ellas. En segundo lugar, también tenemos en cuenta el contexto de las entradas. Para ello, usamos palabras aisladas primero para luego

utilizar las mismas dentro de sintagmas. En un tercer lugar, tenemos en cuenta si las palabras clave tienen otras unidades léxicas con las que competir, clasificadas en tres grupos: tres posibles candidatos (grupo 1), dos posibles candidatos (grupo 2) y candidato único (grupo 3). Por último, tenemos en cuenta el tipo de fonema con el que los sistemas identifican la variante dialectal: como <st>, como <ch>, como <t> o como otro sonido.

2.4. Procedimiento

En el experimento 1, se utilizó un teléfono Samsung Galaxy SIII (GT-I9300), con sistema Android versión 4.3 y sistema de reconocimiento de voz Google Voice Search, con idioma predeterminado a español de España. Se usaron los auriculares con micrófono incorporado que se incluyen con este modelo de teléfono. A su vez, se usó un ordenador Intel Core i3, con placa base Gigabyte H61, disco duro 500GB, 8GB DDR3 Kingston y DVDRW LG 16x, con sistema de reconocimiento de voz de Windows 7 y micrófono externo Realtek. Para el experimento 2, se recogieron datos del dictado de voz tras una, tres, cinco y siete sesiones de entrenamiento con el locutor, con el mismo ordenador y el mismo micrófono que los utilizados en el primer experimento.

3. Resultados

En este apartado se presentan los resultados obtenidos en cuanto al reconocimiento de la variante dialectal de español objeto de estudio. Para ello, se ha recurrido a tests no paramétricos de estadística. Los resultados para las entradas que no contienen la secuencia <st> no se incluyen en este estudio.

3.1. Experimento 1

A priori, tanto Google Voice Search (Z = -5,663; p = 0,000) como Windows 7 sin entrenamiento (Z = -2,897; p = 0,004) reconocieron más palabras clave en los sintagmas que cuando se presentaron aisladas, como vemos en la tabla 3. Sin embargo, con Google Voice Search se obtuvo un mayor índice de reconocimiento tanto en palabras aisladas (U =

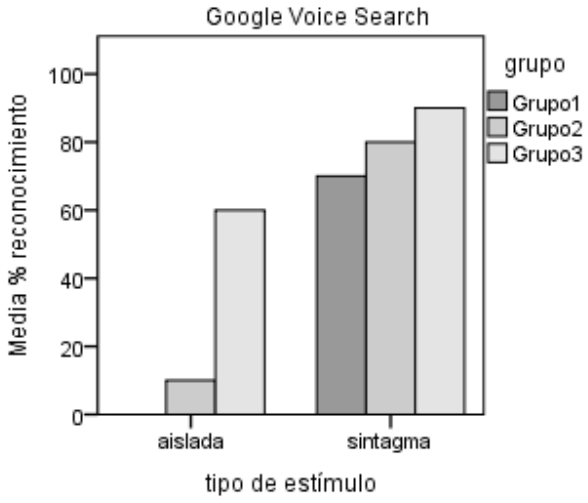
275; p = 0,002) como en sintagmas (U = 0; p = 0,000), comparado con los resultados de Windows 7 sin entrenamiento.

Tabla 3. Porcentajes de reconocimiento de palabras de forma aislada y en sintagmas por Google Voice Search y Windows 7

		estilo de habla			
		aislada		sintagma	
		$\bar{X}$	*SD*	$\bar{X}$	*SD*
sistema	Google	23	30	80	17
	Windows 7	3	8	13	15

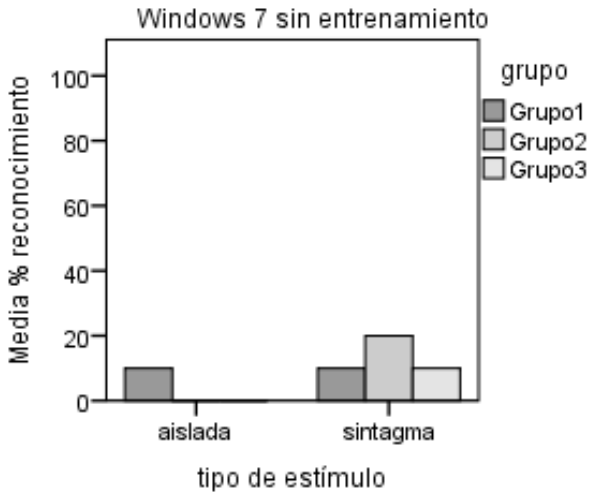
Si desgranamos estos porcentajes en los grupos de palabras mencionados en el apartado anterior, vemos que un gran porcentaje significativo de reconocimiento de palabras aisladas por parte de Google se centra en las del grupo 3 (figura 1), es decir, en aquellas que no tenían competencia léxica, mientras que aquellas que tienen dos competencias léxicas adicionales no se reconocen [$\chi^2(2) = 24{,}323$; p = 0,000). Sin embargo, cuando las palabras se hallan en sintagmas, existe un aumento progresivo de reconocimiento a medida que disminuyen las competencias léxicas, con un mejor resultado en el tercer grupo [$\chi^2(2) = 7{,}250$; p = 0,027].

Figura 1. Reconocimiento con Google Voice Search de palabras aisladas y sintagmas por grupos



En el caso de Windows 7 (figura 2), apenas se aprecia reconocimiento alguno de palabras aisladas, excepto en el primer grupo [$\chi^2(2) = 11{,}600$; $p = 0{,}003$], justo lo contrario que ocurría con Google, mientras que el reconocimiento en sintagmas tiende a ser leve, con mayor incidencia en el grupo en el que existen dos posibilidades léxicas, aunque sin diferencias significativas [$\chi^2(2) = 1{,}450$; $p = 0{,}484$]. Aun así, los porcentajes son muy bajos para este sistema cuando no se ha realizado entrenamiento alguno.

Figura 2. Reconocimiento con Windows 7 de palabras aisladas y sintagmas por grupos



Cuando nos centramos en el reconocimiento de [t^s] como secuencia <st>, independientemente de si la palabra reconocida es la exacta, vemos los siguientes resultados (tabla 4). Google reconoce la mitad de los sonidos clave, seguido de <ch>, aunque esta con menor frecuencia de la esperada[1]. En cambio, Windows 7 tiende a reconocer [t^s] principalmente como <ch>, como otro sonido, o como <t>, mientras que su reconocimiento como <st> es mucho menor que el valor esperado.

[1] La que se esperaría si no hubiese relación entre las variables. Si no guardase relación con el sistema de reconocimiento de voz, los dos presentarían el mismo recuento en la misma proporción que se indica en la muestra considerada en total (Gil Flores y otros, 2011: 140).

Tabla 4. Valores reales y esperados de reconocimiento fonema-grafía con Google Voice Search y Windows 7

		tipo de reconocimiento				
		st	ch	t	otro	Total
Google	Recuento	30	18	4	8	60
	Recuento esperado	19,5	20,0	7,0	13,5	60,0
Windows	Recuento	9	22	10	19	60
	Recuento esperado	19,5	20,0	7,0	13,5	60,0
Total	Recuento	39	40	14	27	120
	Recuento esperado	39,0	40,0	14,0	27,0	120,0

En efecto, existen diferencias significativas entre ambos sistemas cuando se usan con independencia del locutor [$\chi^2(3) = 18{,}761$; $p = 0{,}000$], hallándose una correlación moderada entre el tipo de sistema utilizado y los resultados obtenidos ($C = 0{,}368$; $p = 0{,}000$).

3.2. Experimento 2

En este segundo experimento se comparan los porcentajes de reconocimiento del mismo corpus en base al número de sesiones de entrenamiento de Windows 7 con un mismo locutor. En general, se aprecia un aumento de reconocimiento de voz relacionado con el número de sesiones de entrenamiento, tanto en palabras aisladas [$\chi^2(3) = 47{,}699$; $p = 0{,}000$], como en sintagmas [$\chi^2(3) = 36{,}274$; $p = 0{,}000$], aunque en tres de las cuatro sesiones la identificación en sintagmas es mayor que en palabras aisladas, como podemos ver en la tabla 5. Con una sola sesión de entrenamiento, el reconocimiento tanto de palabras aisladas como en sintagmas es similar.

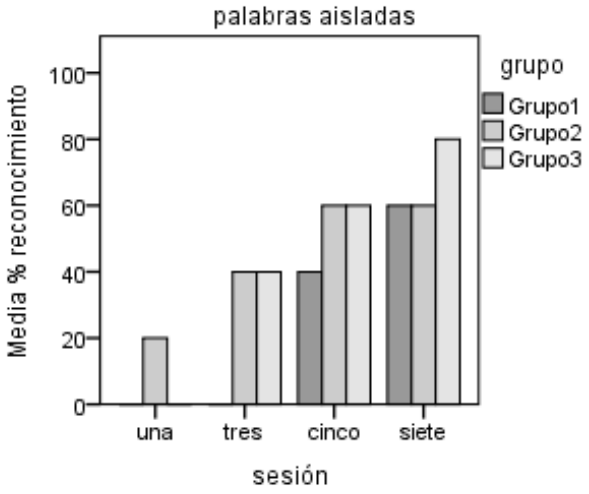
Al fijarnos en los tres grupos de palabras clave (figura 3), vemos que Windows 7 no empieza a reconocer las palabras aisladas del grupo 3 hasta la tercera sesión de entrenamiento, y que no comienza a reconocer las palabras del grupo 1 hasta la quinta sesión. Sin embargo, tras 7 sesiones de entrenamiento no se supera el 60% de

reconocimiento, excepto en el grupo 3, es decir, aquél en el que no existe competencia léxica con la palabra clave.

Tabla 5. Porcentajes de reconocimiento de palabas clave de forma aislada y en sintagmas por número de sesión de entrenamiento con Windows 7

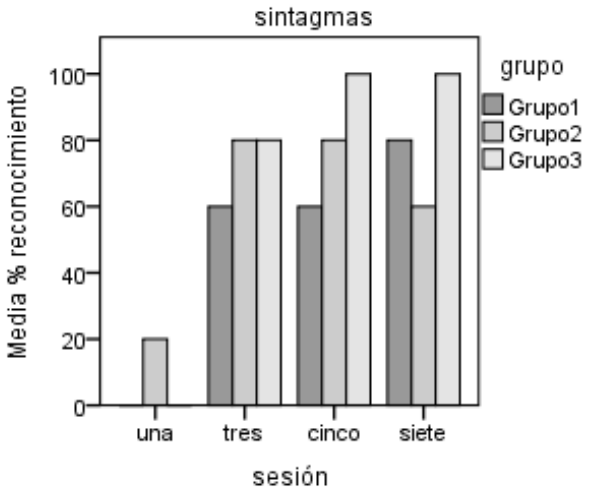
		estilo de habla			
		aislada		sintagma	
		$\bar{X}$	*SD*	$\bar{X}$	*SD*
sesión	una	7	10	7	10
	tres	27	20	73	10
	cinco	53	10	80	17
	siete	67	10	80	17

Figura 3. Reconocimiento de palabras aisladas por grupos de acuerdo al número de sesiones de entrenamiento con Windows 7



En el caso de los sintagmas, Windows 7 solo reconoce algunas palabras clave del grupo 2 tras una sesión de entrenamiento, mientras que ya las reconoce en todos los grupos a partir de la tercera sesión de entrenamiento.

Figura 4. Reconocimiento de palabras en sintagmas por grupos de acuerdo al número de sesiones de entrenamiento con Windows 7



Una vez más, nos fijamos también en el reconocimiento en concreto de [t^s] como <st>, en el que se vuelve a apreciar un cambio en el patrón de identificación a partir de la sesión 3 de entrenamiento. Con una sesión de entrenamiento, Windows 7 reconoce [t^s] mayoritariamente como otro tipo de sonido. A partir de la sesión tres de entrenamiento, el sistema reconoce [t^s] como la secuencia <st> en su mayoría, siendo tras la sesión siete donde vemos un reconocimiento más robusto, con solo tres casos reconocidos como <ch> y ninguno como <t> u otro sonido.

En este caso, las diferencias entre sesiones también son significativas [$\chi^2(9) = 56{,}326$; $p = 0{,}000$] y también están correlacionadas con el número de sesiones de entrenamiento al que se ha sometido al sistema de Windows 7 ($C = 0{,}565$; $p = 0{,}000$).

3.3. Google vs. Windows 7

Para recapitular, en la figura 5 se compara la robustez de Google Voice Search frente a la del dictado por voz de Windows 7 con todas las sesiones de entrenamiento. En cuanto al reconocimiento de palabras aisladas, Google Voice Search supera a Windows 7 sin entrenamiento y con una sesión de entrenamiento. Sin embargo, con tres sesiones de

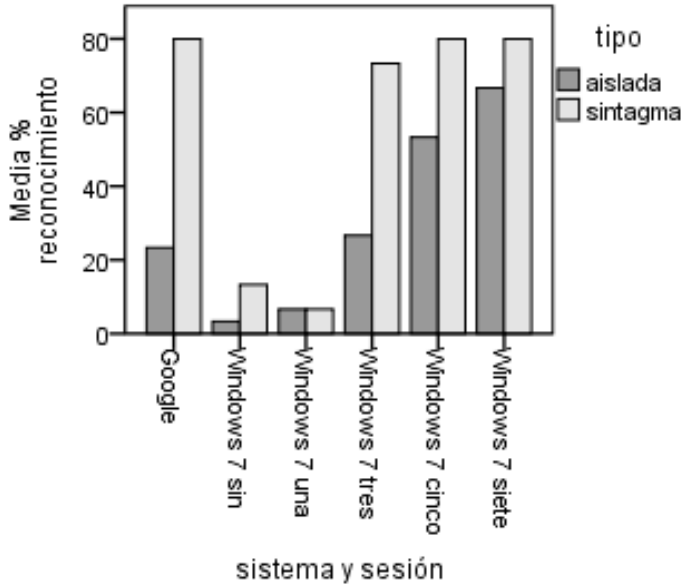
entrenamiento, ya no se aprecian diferencias significativas entre ambos sistemas (U = 187,5; p = 0,334). A partir de cinco sesiones de entrenamiento, Windows 7 supera a Google (U = 87,5; p = 0,001). En el caso de los sintagmas, Google Voice Search vuelve a superar a Windows sin entrenamiento y con una sesión de entrenamiento.

Tabla 6. Valores reales y esperados de reconocimiento fonema-grafía por número de sesiones de entrenamiento con Windows 7

			tipo de reconocimiento				
			st	ch	t	otro	Total
	Sesión1	Recuento	3	9	5	13	30
		Recuento esperado	18,8	4,8	2,3	4,3	30,0
	Sesión3	Recuento	21	5	1	3	30
		Recuento esperado	18,8	4,8	2,3	4,3	30,0
	Sesión5	Recuento	24	2	3	1	30
		Recuento esperado	18,8	4,8	2,3	4,3	30,0
	Sesión7	Recuento	27	3	0	0	30
		Recuento esperado	18,8	4,8	2,3	4,3	30,0
Total		Recuento	75	19	9	17	120
		Recuento esperado	75,0	19,0	9,0	17,0	120,0

De nuevo vemos que aquí tampoco existen diferencias significativas entre los dos sistemas cuando Windows 7 tiene tres sesiones de entrenamiento (U = 175; p = 0,195). Por último, ambos sistemas ofrecen la misma robustez a partir de la quinta sesión de entrenamiento de Windows 7 (U = 225; p = 1) en el reconocimiento de palabras clave en sintagmas.

Figura 5. Comparación entre Google y Windows 7 en el reconocimiento de palabras aisladas y en sintagmas



4. Discusión

A la luz de los resultados obtenidos vemos que, ciertamente, el reconocimiento de palabras aisladas resulta más dificultoso para ambos sistemas que cuando se trata de reconocer estas palabras en sintagmas. Vemos que no solo se trata de reconocer la entrada de voz en cuanto a sus características fonéticas y fonológicas, sino que, una vez ayudados por un análisis morfológico, sintáctico y semántico, hay mayor posibilidad de desambiguar la señal e identificar la palabra correctamente. En el primero de los casos, Google Voice Search ofrece mayor robustez ante la variabilidad dialectal de la señal. En el segundo caso, a partir de la quinta sesión de entrenamiento se reconocen palabras clave en los tres grupos de clasificación cuando se presentan aisladas y a partir de la tercera sesión cuando se presentan en sintagmas, siguiendo un aumento progresivo dependiente del número de sesiones. En cuanto a las competencias léxicas, se demuestra que las palabras que tienen mayor porcentaje de reconocimiento, en general, son aquellas que no tienen otras palabras con las que competir. También, en general, cuando la variante dialectal no se reconoce como <st> existe una gran probabilidad de que se reconozca como <ch>, o bien, como otro

sonido. En este caso, vemos que los sistemas reconocen el componente africado de esta variante, quizás demasiado novedosa para haber sido incorporada a los corpus de entrenamiento.

En un estudio similar, Serrahima (2009) compara el dictado de voz de Dragon Naturally Speaking (Nuance) con el programa de reconocimiento de voz de Windows Vista sin entrenamiento y tras una sesión de entrenamiento. Aunque no encuentra grandes diferencias entre los dos sistemas, ve una mayor eficacia de Windows Vista en el reconocimiento de palabras desconocidas para el sistema. El autor atribuye este hecho a que parece identificar los sonidos de la voz de entrada y agruparlos/distribuirlos en sílabas, para después agrupar las sílabas por palabras, mientras que Dragon Naturally Speaking identifica los sonidos de la voz de entrada, "los divide en palabras, compara cada una de las palabras identificadas con la lista de palabras del propio programa, escoge la que más se parece al sonido identificado y la coloca en su lugar en el texto" (79). Sea como fuere, con el mismo número de sesiones de entrenamiento que en su estudio, el sistema de reconocimiento de voz de Windows 7 no supera la robustez de Google Voice Search en este.

En principio se consideraron tres grupos de palabras clave en base al número de competencias léxicas comprendidas en cada uno de ellos. En el segundo grupo, donde veíamos dos posibles candidatos, Google Voice Search encontró un tercer candidato para las palabras *pasto* y *pasta*: *Pacho*[2] y *Pacha*[3], con las que no contábamos inicialmente. Al tratarse de una búsqueda en internet, las posibilidades de encontrar información son mucho mayores que las contenidas en un sistema de dictado por voz.

Por otra parte, Windows 7 no consiguió identificar *pesto* y *pisto* en ninguna de las ocasiones. En palabras aisladas, llegó a identificar [t^{s}] como la secuencia <st> a partir de la tercera sesión de entrenamiento, aunque no identificó la palabra completa. En concreto, estas dos entradas dieron como resultado *esto* y *visto*, lo cual confirma el reconocimiento de la secuencia <st> como tal, a pesar de no haber identificado las palabras exactas en estos

[2] Municipio de Cundinamarca, capital de la provincia de Rionegro (Colombia).
[3] Sin definición o resultado de búsqueda facilitada. Solo aparece la palabra escrita.

casos. En sintagmas, Windows las identificó sistemáticamente como *pecho* y *dicho*, confundiendo [t^{s}] con [tʃ]. También en palabras aisladas, identificó *mosto* como *mostró* en todas las sesiones, excepto en la sesión sin entrenamiento, que reconoció como *mozo*. Además, identificó *pista* como *pizza* en todas las sesiones menos en la séptima, que la reconoció correctamente. Ante la posibilidad de que estas palabras no estuviesen recogidas en el corpus del sistema, se comprobó dictándolas en castellano, es decir, realizando la /s/ como tal. Efectivamente, estas palabras estaban incluidas en el diccionario de Microsoft Word, aunque, por alguna razón, el sistema no fue capaz de identificarlas con los hablantes de andaluz occidental, aun cuando el entrenamiento mejoró considerablemente los resultados obtenidos para otras palabras.

5. Conclusiones

La robustez de un sistema de reconocimiento de habla se ve mermada ante la variabilidad de la señal en forma de variedad dialectal emergente. Aunque la robustez inicial de un sistema como Google Voice Search sea mayor que la de dictado por voz de Windows 7, tras tres sesiones de entrenamiento se iguala el reconocimiento de ambos sistemas y, a partir de cinco, Windows 7 supera el rendimiento de Google Voice Search. Para futura investigación, queda por ver si estas sesiones de entrenamiento están condicionadas al locutor con el que se entrenó o si, por el contrario, el reconocimiento de sus características dialectales se trasfiere a otros locutores, hablantes de la misma variedad.

6. Bibliografía citada

FANDIÑO RODRÍGUEZ, Deiby Alexander, 2005: *Estado del Arte en el Reconocimiento Automático de Voz. Seminario de investigación*, Universidad de Colombia.

FERREIROS LÓPEZ, Javier, 2003: "¿Qué queremos que sea Tecnología del Habla?", *SEPLN* 31, 375-379.

GIL FLORES, Javier, Javier RODRÍGUEZ SANTERO y Víctor Hugo PERERA RODRÍGUEZ, 2011: *Introducción al Tratamiento Estadístico de Datos mediante SPSS*, Arial S. L.

HERNÁNDEZ GÓMEZ, Luis, 2003: "Modelo de evolución de la Tecnología del Habla, y tendencias futuras", *Procesamiento del Lenguaje Natural* 31, 369-373.

LLISTERRI, Joaquim, Carme CARBÓ, María Jesús MACHUCA, Carme DE LA MOTA, Monserrat RIERA y Antonio RÍOS, 2003: "El papel de la lingüística en el desarrollo de las tecnologías del habla" en Miguel CASAS GÓMEZ y Carmen VARO VARO (eds.): *VII Jornadas de Lingüística*, Cádiz: Servicio de Publicaciones de la Universidad de Cádiz, 137-191.

MOYA CORRAL, Juan Antonio, 2007: "Noticia de un sonido emergente: La africada dental procedente del grupo -st- en Andalucía", *Revista de Filología de La Universidad de La Laguna* 25, 457-465.

NOGUEIRAS, Albino, Mónica CABALLERO y Asunción MORENO, 2002: "Multi-dialectal Spanish SpeechRecognition", *IEEE*, 841-844.

PARRELL, Benjamin, 2012: "The role of gestural phasing in Western Andalusian Spanish aspiration", *Journal of Phonetics* 40(1), 37-45.

RUCH, Hanna, 2010: "Affrication of /st/-clusters in Western Andalusian Spanish: variation and change from a sociophonetic point of view" en Actas de *the Workshop "Sociophonetics, at the crossroads of speech variation, processing and communication"*, Pisa, Italia.

TORREIRA, Francisco, 2007: "Pre- and post-aspirated stops in Andalusian Spanish" en Pilar PRIETO, Joan MASCARO y Maria-Josep SOLÉ (eds.): *Prosodic and Segmental Issues in Romance*, Amsterdam: John Benjamins, 67-82.

SERRAHIMA, Lorenzo, 2009: "Reconocimieno de voz de Windows Vista: ¿mejor, igual o peor que Dragon Naturally Speaking?", *Panace X* 29, 76-79.