Resultados Preliminares de Decodificación Fonética sobre Distintos Tipos de Habla Espontánea

Doroteo T. Toledano[a], Eduardo Campos Palarea[a], Antonio Moreno Sandoval[b], José Colás Pasamontes[a], Javier Garrido Salas[a]

[a]Human-Computer Technologies Laboratory (HCTLab), Escuela Politécnica Superior.
[b]Laboratorio de Lingüística Informática (LLI-UAM)
Universidad Autónoma de Madrid

{doroteo.torre, eduardo.campos, antonio.msandoval, jose.colas, javier.garrido@uam.es}

Resumen

Este artículo presenta los resultados preliminares de decodificación fonética obtenidos sobre el corpus de habla espontánea C-ORAL-ROM. Se comparan los resultados con los resultados obtenidos sobre la base de datos de habla leída ALBAYZIN, y también se comparan los resultados obtenidos para los distintos tipos de habla espontánea que contiene el corpus C-ORAL-ROM. Como conclusiones más importantes se aprecia que el tipo de habla espontánea considerado tiene una influencia muy importante en los resultados de reconocimiento del habla espontánea, mostrando la voz obtenida a través de medios de comunicación los mejores resultados de decodificación fonética de entre los subtipos de habla espontánea considerados.

1.- Introducción

En la actualidad una de las líneas de investigación más importantes en el campo de las tecnologías del habla, y en particular en el reconocimiento de voz, es la investigación en el procesamiento del habla espontánea. En este momento el National Institute of Standards and Technology (NIST) [3] está desarrollando un programa de investigación denominado Rich Transcription (RT) en el que el procesamiento de habla espontánea es un punto clave. Lamentablemente dicho programa no incluye al español entre uno de sus idiomas de interés, por lo que la investigación en el procesamiento del habla espontánea para el español podría quedar relegada a una segunda velocidad. Existen en la actualidad grupos españoles que están ya realizando investigaciones muy interesantes en el campo [2]. Sin embargo, el área de investigación en procesamiento de habla espontánea en español sigue siendo un campo todavía muy abierto.

En este artículo presentamos los resultados exploratorios de nuestra experimentación con el corpus C-ORAL-ROM [7], un corpus de habla espontánea en español que incluye voz espontánea dividida en diferentes categorías. Este corpus, así como algunas adaptaciones que ha sido necesario realizar sobre él para realizar los experimentos, se describe en la Sección 2. Los experimentos realizados son de decodificación fonética empleando modelos acústicos basados en Modelos Ocultos de Markov (HMMs). Estos modelos acústicos están entrenados sobre un corpus de habla leída en español, ALBAYZIN [4]. La Sección 3 describe el entrenamiento de los HMMs empleados. La Sección 4 muestra los resultados de decodificación fonética obtenidos con los HMMs obtenidos, tanto para voz leída como para voz espontánea, y compara ambos. También se compara en esa sección los resultados obtenidos para los distintos tipos de habla espontánea, tal y como están definidos en el corpus C-ORAL-ROM. Finalmente, la Sección 5 presenta las conclusiones y trabajos futuros.

2.- Descripción del Corpus C-ORAL-ROM

C-ORAL-ROM es un corpus multilingüe que engloba cuatro lenguas romances: italiano, francés, portugués y español. Para nuestro trabajo nos hemos centrado en el corpus español que consta de alrededor de 300.000 palabras. Desde el punto de vista sociolingüístico los hablantes se caracterizan por su edad, sexo, lugar de origen, la educación y la profesión. Desde la perspectiva textual la base de datos se divide en las partes explicadas en la Tabla 1 [8].

<table>
<tr><td colspan="4">Informal
150.000 pal.</td><td>Formal
150.000 pal.</td></tr>
<tr><td colspan="2">Familiar
113.000</td><td colspan="2">Público
37.000</td><td>Formal en contexto natural
65.000</td></tr>
<tr><td rowspan="2">Monólogos
33.000</td><td rowspan="2">Diálogo/ Convers.
80.000</td><td rowspan="2">Monólogos
6.000</td><td rowspan="2">Diálogo/ Convers.
31.000</td><td>Formal en los medios
60.000</td></tr>
<tr><td>Grabaciones telefónicas
25.000</td></tr>
</table>

Tabla 1: Distribución de palabras en C-ORAL-ROM

Como se puede observar la separación principal es equilibrada entre habla formal e informal. El habla informal considera una separación entre habla en contexto familiar/privado y habla en contexto público. El primer grupo está dividido en monólogos (efammn, véase la Tabla 2 para interpretar los códigos), diálogos (efamdl) y conversaciones (efamcv). El segundo grupo está dividido de forma similar en monólogos (epubmn), diálogos (epubdl) y conversaciones (epubcv). En cuanto al habla formal, se ha distinguido entre contexto natural, voz en los medios de comunicación y conversaciones telefónicas. En el primero de estos grupos se incluyen discursos políticos (enatps), debates políticos (enatpd), sermones (enatpr), enseñanza (enatte), exposiciones profesionales (enatpe), conferencias (enatco), en el ámbito de los negocios (enatbu) y voz en ámbitos legales (enatla). La voz en los medios de

comunicación se clasifica en noticias (emednw), deportes (emedsp), entrevistas (emedin), meteorología (emedmt), ciencia (emedsc), reportajes (emedrp) y programas de debate (emedts). En cuanto a las conversaciones telefónicas, no se han considerado subdivisiones.

<table>
<tr><td rowspan="4"><u>Informal</u></td><td rowspan="2">Familiar/Privado
efam</td><td>Monólogo
mn</td></tr>
<tr><td rowspan="2">Diálogo
dl</td></tr>
<tr><td rowspan="2">Público
epub</td></tr>
<tr><td>Conversación
cv</td></tr>
</table>

<table>
<tr><td colspan="3"><u>Formal</u></td></tr>
<tr><td>Formal en contexto natural
enat</td><td>Media
emed</td><td>Teléfono
etelef</td></tr>
<tr><td>Discurso politico
ps</td><td>Noticias
nw</td><td rowspan="8"></td></tr>
<tr><td>Debate politico
pd</td><td>Deportes
sp</td></tr>
<tr><td>Sermones
pr</td><td>Entrevistas
in</td></tr>
<tr><td>Enseñanza
te</td><td>Meteorología
mt</td></tr>
<tr><td>Exposiciones profesionales
pe</td><td>Científico
sc</td></tr>
<tr><td>Conferencias
co</td><td>Reportajes
rp</td></tr>
<tr><td>Negocios
bu</td><td>Debates
ts</td></tr>
</table>

Tabla 2: Distribución del corpus C-ORAL-ROM y sus códigos

Estas divisiones y subdivisiones del corpus C-ORAL-ROM nos permitirán comparar los resultados obtenidos en la decodificación fonética para distintos tipos de habla espontánea, que se expondrán en la Figura 1.

El corpus C-ORAL-ROM contiene un total de 183 grabaciones con una duración total aproximada de 30 horas. Los ficheros de sonido originales son en la mayoría de los casos de más de 10 minutos. Estos ficheros no resultaban apropiados para su procesamiento automático, por lo que se procedió a extraer cada uno de los grupos fónicos del corpus como ficheros independientes. Afortunadamente el corpus C-ORAL-ROM incluye una precisa segmentación manual de todas sus grabaciones en grupos fónicos, segmentación que ha sido esencial a la hora de realizar los experimentos que aquí se describen.

2.1. Transcripción Fonológica

A fin de comparar los resultados de decodificación fonética, en primer lugar necesitamos una transcripción fonológica de referencia. Desafortunadamente, el corpus C-ORAL-ROM no incluye dicha transcripción fonológica, sino únicamente transcripción ortográfica. Por ello ha sido necesario obtener la transcripción fonológica a partir de la ortográfica empleando para ello un sencillo transcriptor basado en reglas. Este transcriptor emplea un conjunto mínimo de fonemas para el castellano (23 fonemas). Evidentemente, un transcriptor tan sencillo no permite obtener una transcripción fonológica correcta en todos los casos. Sin embargo, consideramos que la precisión que alcanza es suficiente para obtener unos resultados de decodificación fonética significativos.

3.- Entrenamiento de los HMMs para Decodificación Fonética

Para el entrenamiento de los modelos ocultos de Markov a partir de la base de datos ALBAYZIN se ha empleado el software HTK [6]. Como *front-end* de extracción de características se ha empleado el *front-end* avanzado para reconocimiento de voz distribuido definido por el estándar ETSI ES 202 050 [5]. Este *front-end* incluye mecanismos de robustez frente a ruido basados en el procesado de las características extraídas de la voz. Básicamente el mecanismo de robustez frente a ruido consiste en un doble filtrado de Wiener que estima y sustrae el espectro de ruido.

El conjunto de fonemas que se ha empleado en todos los experimentos es un conjunto mínimo de 23 fonemas del español. También se consideran modelos para el silencio inicial, silencio final y silencio intermedio. Se han entrenado tanto modelos de fonema independientes del contexto como modelos de fonema dependientes del contexto. Se comenzó entrenando unos modelos semilla independientes del contexto empleando 600 de las 1200 frases de ALBAYZIN segmentadas y etiquetadas fonéticamente (el resto se reservaron para realizar ajustes y evaluación). A continuación, y empleando dichos modelos semilla se procedió a entrenar los modelos independientes del contexto con 3500 pronunciaciones del conjunto de entrenamiento de ALBAYZIN. Se entrenaron modelos de hasta 150 Gaussianas por estado, si bien se observó que los resultados de decodificación fonética mejoraban ya muy ligeramente a partir de unas 65 Gaussianas, por lo que se decidió emplear este número de Gaussianas. A partir de los modelos independientes del contexto se entrenaron modelos dependientes del contexto y se procedió al atado de estados empleando el algoritmo de agrupamiento de estados basado en árbol de decisión. Los modelos dependientes del contexto resultantes del agrupamiento contenían en conjunto un total de 2079 estados. Dado que el conjunto de modelos independientes del contexto incluía un total de 26x3x65=5070 Gaussianas, elegimos utilizar modelos dependientes del contexto de una complejidad similar o algo inferior, a fin de comprobar la bondad de la utilización de modelos dependientes del contexto. Por ello elegimos emplear modelos dependientes del contexto de 2 Gaussianas por estado, lo que implicaba un total de 2079x2=4158 Gaussianas.

4. Resultados de Decodificación Fonética

La prueba de decodificación fonética que realizamos consiste en la evaluación de la precisión de la decodificación fonética de los modelos, es decir, la precisión del reconocimiento de fonemas empleando una gramática sin restricciones (con la excepción de que debe comenzar y terminar en el silencio inicial y final respectivamente). En el caso de los modelos dependientes del contexto, la gramática también exige que se

respeten los contextos fonéticos en la secuencia de fonemas, lo que hace bastante más lento el proceso de decodificación fonética.

Para evaluar los resultados se ha procedido al alineamiento de las cadenas fonéticas obtenidas a partir de la transcripción ortográfica mediante el transcriptor automático al que hacíamos referencia en la Sección 2.1 y calculando el porcentaje de fonemas correctos (%C) y la precisión de la decodificación fonética (%A), definida como el porcentaje de fonemas correctos menos el porcentaje de fonemas insertados.

4.1. Decodificación Fonética de Habla Leída

Es importante tener una idea del grado de precisión que alcanzan los modelos entrenados en la decodificación fonética del habla leída, a fin de establecer el grado de influencia del tipo de habla en la precisión de la decodificación fonética. Por ello hemos realizado una prueba de decodificación fonética sobre un pequeño conjunto de 300 pronunciaciones de la base de datos ALBAYZIN que estaban etiquetadas y segmentadas fonéticamente y que no fueron empleadas en el proceso de entrenamiento de los modelos.

Sobre este conjunto de prueba la decodificación fonética con modelos independientes del contexto obtenía %C = 81.07% y %A = 76.56% cuando se comparaba con la transcripción fonética obtenida de forma automática. A fin de validar los resultados comparando con la transcripción fonética obtenida de forma automática, se han obtenido los resultados comparando también con la transcripción fonética manual que incluía el corpus. Los resultados obtenidos son muy similares, %C = 81.36% y %A = 76.24%, lo que justifica que podamos emplear la transcripción fonética automática en la evaluación de la decodificación fonética del corpus C-ORAL-ROM, para el que no hay todavía disponible una transcripción fonética verificada manualmente.

En el caso de la decodificación fonética con modelos dependientes del contexto, lo que obtenemos comparando con la transcripción fonética automática es %C = 83.88% y %A = 74.55%. En el caso de emplear la transcripción fonética manual los resultados son también muy similares, %C = 83.79%, %A = 73.01%. Los resultados obtenidos con los modelos dependientes del contexto son muy similares a los obtenidos con los modelos independientes del contexto.

4.2. Decodificación Fonética de Habla Espontánea

Al realizar la misma prueba de decodificación fonética sobre todo el corpus C-ORAL-ROM empleando como referencia la transcripción fonológica automática se obtienen resultados mucho más modestos, como cabría esperar. Para el caso de los fonemas independientes del contexto se obtiene %C = 44.00% y %A = 25.71%. Para el caso de los fonemas dependientes del contexto los resultados vuelven a ser muy similares, %C = 43.06%, %A = 25.07%.

Esta reducción en la precisión de la decodificación fonética se debe a la dificultad inherente que presenta la voz espontánea. Sin embargo, sería absurdo pensar que es ése el único factor que provoca esta reducción tan drástica en las tasas de decodificación fonética. Otros factores que ejercen una influencia decisiva en esta reducción son los siguientes:

- La no coincidencia del canal entre la voz empleada para el entrenamiento y la empleada para el reconocimiento. El corpus ALBAYZIN es un corpus de voz microfónica limpia, mientras que el corpus C-ORAL-ROM es un corpus que incluye voz microfónica grabada con distintos tipos de micrófonos en distintos entornos (más o menos ruidosos), voz tomada de los medios de comunicación, voz tomada de grabaciones telefónicas, etc. Esta no coincidencia del canal se mitiga en parte mediante los mecanismos de robustez frente al canal que incluye el *front-end* empleado [5]. Sin embargo, su influencia en los resultados de decodificación puede ser todavía importante.
- La presencia de ruidos con distintas características y niveles en el corpus de reconocimiento. El efecto de este factor también se ve mitigado por el empleo de un *front-end* con mecanismos de robustez frente al ruido [5], pero su influencia también debe ser considerada.
- La no coincidencia de las características de la voz de entrenamiento y de reconocimiento, tanto en tipo de

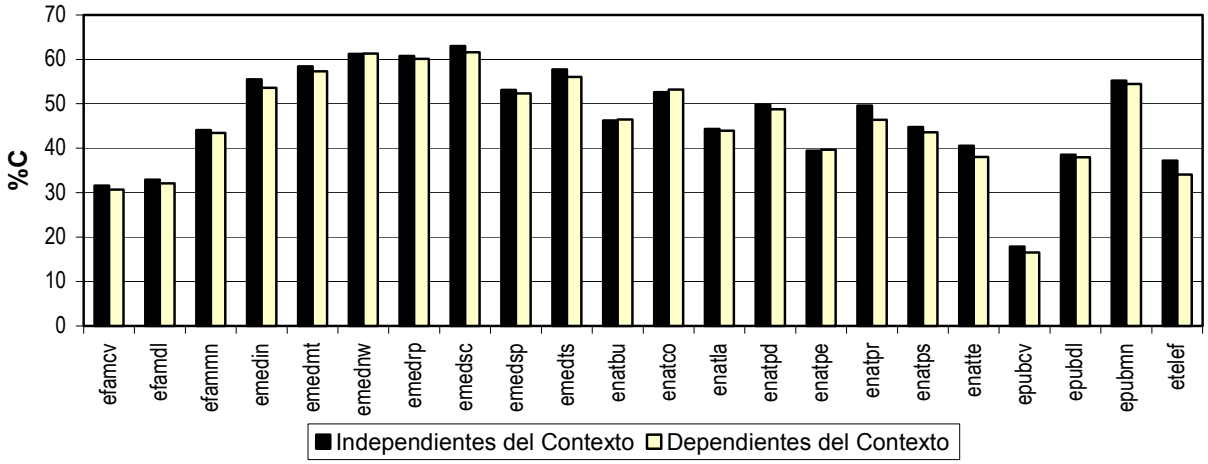

Figura 1: Resultados de decodificación fonética (porcentaje de fonemas correctos, %C) por subtipo de habla espontánea (ver Sección 2) en el corpus de habla espontánea C-ORAL-ROM.

habla como en nivel de ruidos. En cuanto a este factor, es posible que realizando un entrenamiento o una adaptación de los modelos empleando parte del corpus C-ORAL-ROM se consiguiese mejorar los resultados al conseguir modelos más adaptados al habla espontánea y a entornos más ruidosos.

Todos estos factores limitan la utilidad de la comparación entre las tasas de decodificación fonética del apartado 4.1 y del apartado 4.2. Mucho más interesante que esta comparación resulta la comparación entre las tasas de decodificación fonética obtenidas sobre los distintos tipos de habla espontánea que contiene el corpus C-ORAL-ROM.

4.3. Comparativa de Resultados de Decodificación Fonética para Distintos Tipos de Habla Espontánea

La Figura 1 muestra el porcentaje de fonemas correctos para cada uno de los subtipos de habla espontánea que se consideran en el corpus C-ORAL-ROM, y que se describen brevemente el la Sección 2.

La comparación de los modelos acústicos empleados muestra que los resultados obtenidos con los modelos fonéticos independientes del contexto y con los modelos fonéticos dependientes del contexto son similares, siendo algo superiores los obtenidos con los dependientes del contexto. Esta observación podría deberse a que los modelos independientes del contexto son modelos más complejos por incluir, globalmente, un mayor número de Gaussianas que los modelos independientes del contexto. Muy probablemente aumentando la complejidad de los modelos dependientes del contexto se consiga mejorar notablemente los resultados aquí presentados, aunque será a costa de una mayor complejidad de los modelos acústicos.

Resulta muy interesante observar que existe un margen de variación muy amplio entre los distintos tipos de habla espontánea considerados: desde el algo menos del 20% de fonemas correctos para conversaciones informales en contexto público (epubcv) hasta el algo más del 60% de fonemas correctos para los programas de ciencia en medios de comunicación (emedsc).

En general se observa que para las conversaciones y los diálogos (efamcv, efamdl, epubcv y epubdl) se obtienen los resultados más bajos (alrededor del 30% de fonemas correctos para todo el grupo). Otro subconjunto relacionado con los anteriores en que también contiene diálogos o conversaciones es el conjunto de conversaciones telefónicas (etelef) para el que los resultados son similares. En todos estos casos parece evidente que la interacción (con frecuentes solapamientos) entre los hablantes es la causa de la reducida tasa de decodificación fonética. En el caso de las conversaciones telefónicas también existe un claro desacoplamiento entre las características de la voz empleada para entrenar los modelos acústicos y la empleada para realizar las pruebas de reconocimiento.

En cuanto a los monólogos informales, se aprecia que en contexto familiar (efammn) obtienen unos resultados algo superiores a los anteriores (algo más del 40%), mientras que en contexto público (epubmn) obtienen unos resultados claramente superiores (de cerca del 55% de fonemas correctos).

En general, los grupos anteriores correspondían a situaciones informales, y se observa que, con la excepción de los monólogos en contexto público, los resultados son siempre inferiores a los resultados que se obtienen con la voz correspondiente a situaciones formales, ya sean en contexto natural (enat**) o en el contexto de los medios de comunicación (emed**). Comparando estos dos grandes grupos se observa que la voz correspondiente a situaciones formales en contexto natural (enat**) produce resultados bastante inferiores (rondando entre el 40% y el 50% de fonemas reconocidos correctamente) a los que se observan con voz correspondiente a situaciones formales en el contexto de los medios de comunicación (emed**), para la cual los resultados de decodificación fonética se sitúan entre el 50% y el 60% de fonemas reconocidos correctamente.

Dentro del conjunto de habla formal en el contexto de los medios de comunicación se aprecian también diferencias interesantes. Los peores resultados se obtienen con los programas de deportes (emedsp), probablemente debido a una utilización del lenguaje menos cuidada y con vocalizaciones a veces exageradas, así como a solapamientos de los hablantes. Algo mejores son los resultados obtenidos con entrevistas (emedin) donde también se producen solapamientos frecuentes. Le siguen los resultados obtenidos sobre programas de meteorología (emedmt) y de debate (emedts). Finalmente, los mejores resultados se obtienen con los programas de noticias (emednw), de reportajes (emedrp) y, sobre todo, con los programas científicos (emedsc). Cabe suponer que en este tipo de contenidos se produzca un reducido número de solapamientos y que el lenguaje empleado esté especialmente cuidado.

Si comparamos los resultados de decodificación automática con los problemas de transcripción de las grabaciones por parte de los expertos humanos, recogidas en [9], observamos coincidencias muy significativas. En concreto, los transcriptores humanos encontraron serias dificultades con los rasgos típicos de la interacción en la comunicación espontánea: solapamientos, número de palabras por turno y velocidad de habla. En estos casos se cumplía la intuición:

Escala 1: Grado de formalidad

informal media formal

+difícil _____ - difícil

Escala 2: Número de hablantes

conversación diálogo monologo

+ difícil _____ - difícil

En la primera escala, se espera que cuanto más formal es el tipo de habla, más fácil es de transcribir, pues se siguen más los convenciones discursivas y retóricas. El habla es más predecible y la pronunciación es más cuidada.

En la segunda escala se especifica que cuantos más hablantes participan en la grabación, más compleja se hace la

transcripción, al tener que distinguir el turno de cada hablante y la aparición de solapamientos. En los monólogos esta dificultad se reduce al mínimo.

Estos resultados coinciden en gran medida con los obtenidos por el reconocimiento automático: los textos más fáciles de transcribir y segmentar son los de los medios de comunicación, producidos por profesionales del discurso oral espontáneo, que combinan la buena dicción con la experiencia de una elaboración fluida y dentro de la norma lingüística culta. Cuanto más nos acercamos a contextos informales y aumenta el número de participantes, el reconocimiento y la transcripción se hacen más complejas.

5. Conclusiones y trabajos futuros

En este artículo hemos presentado unos resultados preliminares de decodificación fonética sobre habla espontánea y los hemos comparado con los resultados que se obtenían sobre habla leída de las mismas características que la empleada en el entrenamiento de los modelos acústicos. Esta comparación presenta globalmente una reducción de en torno al 50% en términos relativos de la tasa de fonemas reconocidos correctamente cuando pasamos de habla leída a habla espontánea. Aunque la influencia de las características del habla (leída frente a espontánea) sobre los resultados de decodificación fonética es innegable, también es cierto que en los experimentos que hemos realizado influyen también otros factores como la no coincidencia del canal de grabación ni del nivel de ruido entre las grabaciones empleadas para entrenar y las empleadas para obtener los resultados de decodificación fonética. Esto hace que esta comparación sea de utilidad limitada.

Mucho más interesante es la comparación entre los resultados de decodificación fonética obtenidos sobre distintos tipos de habla espontánea. Entre los tipos de habla espontánea analizados, los mejores resultados obtenidos se obtienen con voz obtenida de los medios de comunicación. Para este tipo de habla espontánea los resultados presentan un empeoramiento de sólo un 25% en términos relativos (aproximadamente) frente a los resultados obtenidos sobre habla leída de las mismas características que la empleada para entrenar los modelos acústicos. Esto significa que este tipo de habla espontánea es el más sencillo de procesar de entre todos los analizados. Siguiendo en orden de complejidad de procesado están la voz producida en situaciones formales y en contextos naturales, los monólogos informales, y por último los diálogos y conversaciones informales, en los que probablemente los solapamientos e interrupciones entre los hablantes hagan que la complejidad del procesado de este tipo de habla resulte muy superior a las anteriores. Estos resultados coinciden con la experiencia de los transcriptores humanos.

Evidentemente, el estudio que aquí presentamos es un estudio preliminar. Un estudio más detallado debería tener en consideración de forma más detallada aspectos como la frecuencia de solapamientos, interrupciones y otros fenómenos de habla espontánea, que han sido completamente obviados en este estudio preliminar, y que constituyen la línea de investigación que afrontaremos en un futuro próximo.

6. Referencias

[1] Rabiner L. and Juang B., Fundamentals of Speech Recognition, Prentice Hall,

[2] Luis Javier Rodríguez Fuentes. Estudio y modelización acústica del habla espontánea en diálogos hombre-máquina y entre personas. Tesis Doctoral. Facultad de Ciencia y Tecnología Universidad del País Vasco.

[3] http://nist.gov/speech/tests/rt/

[4] Climent Nadeu. ALBAYZIN, Universitat Politecnica de Catalunya, ETSET New Jersey, 1993.

[5] Aurora Front-End manual. ETSI ES 202 050 V1.1.3 (2003-11).

[6] Young, S. et al., The HTK Book (for HTK Version 3.1), Microsoft Corporation, July 2000.

[7] Cresti & Moneglia (eds.), C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages, Amsterdam, John Benjamins, 2004.

[8] Moreno Sandoval, A. La evolución de los corpus de habla espontánea: la experiencia del LLI-UAM. Actas de la II Jornadas en Tecnologías del Habla, diciembre 2002, Granada.

[9] González Ledesma, A.; De la Madrid, G.; Alcántara Plá, M.; De la Torre, R.; Moreno-Sandoval, A. Orality and Difficulties in the Transcription of Spoken Corpora. Proceedings of the Workshop on Compiling and Processing Spoken Language Corpora, LREC, 2004, Lisboa.