3. Características de las etiquetas y marcas del corpus

3.1. Etiquetas o marcas para la transliteración de los materiales usados en Preseea-Tunja-Co

Por tratarse de un corpus oral es necesario estandarizar etiquetas y marcas para trasliterar o transcribir las entrevistas. Al ser un proyecto coordinado dentro del PRESEEA, en este capítulo se presentan el sistema de marcas y etiquetas. Lo que presentamos a continuación esta disposición en la web www.linguas.net/preseea[11].

Este documento se ha elaborado tomando como base el trabajo de Antonio Ávila, Matilde Vida y Mari Cruz Lasarte «Propuesta de transliteración y etiquetado del Macrocorpus PRESEEA» (2006). Una vez difundida esta propuesta entre los equipos que integran el proyecto y experimentadas las ventajas y las dificultades de su aplicación en el proceso de investigación, la asamblea general de PRESEEA decidió elaborar un documento de marcas y etiquetas mínimas, de uso obligado en todos los materiales que integran PRESEEA y destinados a los fines comunes del proyecto. Para la redacción de este documento, se ha contado con informes y propuestas aportados por los equipos de las siguientes ciudades: Alcalá de Henares, Barcelona, Bogotá, Medellín, México D.F., Palma de Mallorca y Santiago de Compostela, además de Málaga. La aplicación de estas normas comunes para el marcado y etiquetado de los textos de PRESEEA no ha de impedir su complementación o ampliación con otras normas que sirvan a los intereses específicos de cada equipo de la red, para lo cual se requiere simplemente que los criterios comunes y los específicos no sean incompatibles metodológicamente.

En la elaboración de este documento de marcas y etiquetas obligatorias han participado las siguientes personas:

---

[11] PRESEEA (2008) Marcas y etiquetas mínimas obligatorias. Vers.1.2, 17/02/2008. http://www.linguas.net/preseea

Antonio Ávila (redactor de documento base, consultor de XML y revisor)

Laura Camargo (revisora de marcas y etiquetas; consultora de cuestiones discursivas)

Ana M. Cestero (revisora de marcas y etiquetas; consultora de cuestiones conversacionales) M. Claudia González Rátiva (revisora de marcas y etiquetas)

M. Clara Henríquez (revisora de marcas y etiquetas)

M. Cruz Lasarte (redactora del documento base y revisora)

Pedro Martín Butragueño (revisor de marcas y etiquetas; consultor de cuestiones fónicas)

Francisco Moreno Fernández (redactor del documento y revisor) Montserrat Recalde (revisora de marcas y etiquetas)

Guillermo Rojo (asesor para cuestiones informáticas, consultor de XML y revisor) José María Sánchez Sáez (consultor de XML y revisor)

María Sancho Pascual (creadora de macro de etiquetas y revisora)

Antonio Torres (revisor de marcas y etiquetas)

Victoria Vázquez (revisora de marcas y etiquetas) Matilde Vida (redactora del documento base y revisora)

Juan Villena Ponsoda (consultor de XML y revisor)

Los materiales reunidos dentro de PRESEEA y destinados a los fines comunes del proyecto se presentan en dos tipos formatos:

1. Archivos sonoros

2. Archivos de texto en transcripción ortográfica enriquecida

Los archivos sonoros ofrecen la grabación de las entrevistas semidirigidas realizadas en las comunidades de habla objeto de investigación sociolingüística (www.linguas.net/preseea). Los archivos de texto ofrecen la transcripción ortográfica enriquecida de las entrevistas grabadas. Estos archivos de texto se presentan en dos formatos, preparados con fines diferentes:

Texto etiquetado, destinado a su uso informático con diversos fines.

Texto sin etiquetar, destinado a su publicación impresa y a la lectura convencional.

Este documento de «Marcas y etiquetas mínimas obligatorias» ofrece la información esencial sobre cómo elaborar y presentar los archivos de texto en transcripción ortográfica enriquecida.

A. TEXTOS ETIQUETADOS

Formato

Los textos etiquetados han de elaborarse y almacenarse en formato electrónico, en archivos de texto con extensión .TXT. Para su creación y manejo se requiere un editor de textos.

La denominación de los archivos ha de seguir unas pautas comunes. Se utiliza un sistema que comienza con cuatro caracteres identificativos de la comunidad estudiada (p.e. MALA: Málaga), seguidos del código sociolingüístico del informante (p.e. H23. Sexo/género: H(ombre), M(ujer); grupo de edad: 1, 2, 3; nivel educativo: 1, 2, 3) y de un número de tres cifras, comprendido entre el 001 y el número máximo de informantes utilizados (p.e. 054 ó 108). Cada equipo podrá adjudicar a los archivos otros códigos y numeraciones para sus fines particulares, pero los materiales presentados para el proyecto común deberán ajustarse al sistema general de identificación.

El posterior tratamiento informático de los textos etiquetados se hará utilizando el lenguaje de codificación XML, versión adaptada y simplificada de SGML, que facilitará el intercambio de datos y la recuperación selectiva de información.

Configuración de los textos

Los textos etiquetados presentan dos partes bien diferenciadas: la cabecera y el texto propiamente dicho. Para la elaboración de la cabecera, se hará uso de una plantilla que habrá de rellenarse con los datos específicos.

Cabecera

La cabecera está formada por una serie de campos que proporcionan información sobre los siguientes aspectos:

Datos del propio archivo.

Datos de la grabación de la entrevista.

Datos sobre la transcripción y revisión de la entrevista.

Datos sobre los hablantes participantes en la entrevista.

Esos datos han de ajustarse a un formato común que garantice la homogeneidad de su tratamiento por parte de todos los equipos investigadores y los procesos comunes. Además, la plantilla tendrá un formato compatible con XML, para asegurar la

recuperación de la información. Los datos específicos se cumplimentan entre las comillas correspondientes en los espacios que aparecen sombreados en esta plantilla de muestra.

<Trans audio_filename=»MALA_H23_001.mp3" xml:lang=»español»>

<Datos clave_texto=»MALA_H23_001" tipo_texto=»entrevista_semidirigida»>

<Corpus corpus=»PRESEEA» subcorpus=»ESESUMA» ciudad=»Málaga» pais=»España»/>

<Grabacion resp_grab=»Matilde Vida» lugar=»domicilio informante»

duracion=»07'16"» fecha_grab=»1998 01 01" sistema=»WAV»/>

<Transcripcion resp_trans=»Matilde Vida» fecha_trans=»1999 01 01" numero_palabras=»1586"/>

<Revision num_rev=»1" resp_rev=»Antonio Ávila» fecha_rev=»2005 01 02"/>

<Revision num_rev=»2" resp_rev=»Juan Villena» fecha_rev=»2007 01 03"/></Datos>

<Hablantes>

<Hablante id=»hab1" nombre=»MALA_H23_001" codigo_hab=»I» sexo=»hombre» grupo_edad=»2" edad=»44" nivel_edu=»alto» estudios=»derecho» profesion=»abogado» origen=»Málaga» papel=»informante»/>

<Hablante id=»hab2" nombre=»Matilde Vida» codigo_hab=»E» sexo=»mujer» grupo_edad=»2" edad=»36" nivel_edu=»alto» estudios=»filología» profesion=»profesora» origen=»Málaga» papel=»entrevistador»/>

<Hablante id=»hab3" nombre=»LPM» codigo_hab=»A1" sexo=»mujer» grupo_edad=»2"

edad=»40" nivel_edu=»alto» estudios=»filología» profesion=»profesora» origen=»Málaga» papel=»audiencia»/>

<Relaciones rel_ent_inf = «desconocidos» rel_inf_aud1=»desconocidos» rel_ent_aud1=»conocidos» rel_inf_aud2=»no» rel_ent_aud2=»no»/>

</Hablantes> </Trans>

Es importante tener en cuenta que, en XML, las mayúsculas y las minúsculas son diferentes, por lo que su uso dentro de la cabecera debe ajustarse al modelo que acaba de exponerse. A continuación se presentan los datos contenidos en la plantilla con las explicaciones necesarias y los requisitos que han de cumplirse.

Datos del propio archivo

nombre del archivo de audio (audio_filename): MALA_H23_001.mp3

clave de texto (clave_texto): MALA_H23_001 [Código con formato general de PRESEEA: código de

ciudad – 4 caracteres –, código de informante – sexo/género, grupo de edad, nivel educativo – número

de entrevista, dado por cada equipo]

tipo de texto: entrevista_semidirigida corpus: PRESEEA

subcorpus: ESESUMA [nombre del corpus de cada equipo, si lo hay]

ciudad: Madrid país: España

Datos sobre la grabación de la entrevista responsable de grabación (resp_grab): Matilde Vida lugar de grabación: domicilio informante

duración: 07'16"

fecha de grabación (fecha_grab): 1998 01 01 [las fechas que aparecen en la cabecera han de tener el formato aaaa mm dd]

sistema: WAV [tipo de archivo de grabación original; si la grabación se hizo en cinta, debe anotarse

«analógico»]

Datos sobre la transcripción y revisión de la entrevista responsable de transcripción (resp_trans): Matilde Vida

fecha de transcripción (fecha_trans): 1999 01 01 [las fechas que aparecen en la cabecera han de tener el formato aaaa mm dd]

extensión (numero_palabras): 1586 [el número de palabras corresponde al texto transcrito, excluidas la cabecera completa y las etiquetas]

revisión 1 (resp_rev): Antonio Ávila

fecha revisión 1 (fecha_rev): 2005 01 02 [las fechas que aparecen en la cabecera han de tener el

formato aaaa mm dd]

revisión 2 (resp_rev): Juan Villena

fecha revisión 2 (fecha_rev): 2007 01 03 [las fechas que aparecen en la cabecera han de tener el formato aaaa mm dd]

Datos sobre los hablantes participantes en la entrevista

nombre de informante (nombre): MALA_H23_001 [se utilizará el mismo código que figura en el campo «clave de texto»; en este caso, cada equipo podría añadir, tras un nuevo guion bajo, otro código que permitiera su identificación con fines particulares]

código hablante (codigo _hab): I [los códigos de hablante son I (Informante), E (Entrevistador) y A1 (Audiencia)

sexo: hombre [hombre | mujer]

grupo de edad (grupo_edad): 2 [1|2|3]

edad: 44

nivel educativo (nivel_edu): alto [bajo|medio|alto]

estudios: derecho profesión: abogado origen: Málaga papel: informante

nombre de entrevistador (nombre): Matilde Vida código hablante (codigo_hab): E

sexo: mujer [hombre | mujer]

grupo de edad (grupo_edad): 2 [1|2|3]

edad: 36

nivel educativo (nivel_edu): alto [bajo|medio|alto]

estudios: filología

profesión: profesora origen: Málaga papel: ntrevistador

nombre de audiencia 1: [puede haber más de un interlocutor como audiencia; los datos que se desconozcan reciben el valor «desc» (desconocido); si no existe audiencia, se elimina de la cabecera el segmento correspondiente al «hab3»]

código hablante (código_hab): A1 [en caso de haber más de un hablante como «audiencia», pueden utilizarse los códigos A2, A3, ...]

sexo: mujer [hombre |mujer] grupo de edad (grupo_edad): 2 [1|2|3 | desc]

edad: 40

nivel educativo (nivel_edu): alto [bajo|medio|alto |desc]

estudios: filología [también podría ser «desc»] profesión: profesora [también podría ser «desc»] origen: Málaga [también podría ser «desc»] papel: audiencia relación EI (rel_ent_inf): desconocidos [relación entre entrevistador e informante: conocidos | desconocidos]

relación IA1 (rel_inf_aud1): desconocidos [relación entre informante y audiencia: conocidos | desconocidos|no] [si no existe A1, debe anotarse la opción «no»]

relación EA1 (rel_ent_aud1): conocidos [relación entre entrevistador y audiencia: conocidos | desconocidos|no] [si no existe A1, debe anotarse la opción «no»]

relación IA2 (rel_inf_aud2): desconocidos [relación entre informante y audiencia: conocidos | desconocidos|no] [si no existe A2, debe anotarse la opción «no»]

relación EA2 (rel_ent_aud2): conocidos [relación entre entrevistador y audiencia: conocidos | desconocidos|no] [si no existe A2, debe anotarse la opción «no»] [para la cabecera, no se tendrá en cuenta la posible presencia de otras personas que actúen como «audiencia»]

Etiquetado de texto

El texto correspondiente a cada entrevista ha de transcribirse en ortografía convencional, incluida la acentuación. Las palabras que presenten elisiones se completan en su escritura.

No obstante, el texto transcrito también debe ajustarse a ciertos criterios gráficos que no son los habituales en la escritura ordinaria, sino que responden a unas necesidades propias de la representación escrita de la lengua oral. Estos criterios afectan especialmente a la puntuación y a la forma de representar las pausas, pero también a otros aspectos, que quedan aquí convenientemente precisados.

El marcado y etiquetado del texto debe hacerse utilizando una serie de signos, marcas, etiquetas y convenciones que serán de uso común y obligatorio por parte de los equipos integrantes de PRESEEA.

Es muy importante que los investigadores conozcan cuáles son los objetivos generales del marcado y etiquetado de textos, puesto que esos objetivos son los que deben orientar las decisiones que se tomarán en los casos de duda o de dificultad imprevista. Ha de tenerse en cuenta que es imposible prever en un documento de estas características todas las circunstancias posibles en el proceso de transcripción y etiquetado de un texto, por lo que resulta mprescindible recurrir a unos objetivos comunes y unos criterios generales.

Esos criterios generales, a partir de los cuales se han seleccionado las marcas y etiquetas mínimas y obligatorias de PRESEEA, se pueden resumir en los siguientes puntos:

Las marcas y etiquetas han de reflejar aquellos aspectos fónicos y ambientales cuyo conocimiento resulte imprescindible para una correcta comprensión del contenido de la entrevista.

Las marcas y etiquetas han de ser económicas y deben evitar las redundancias y los esfuerzos desproporcionados en el proceso de transcripción.

Las marcas y etiquetas deben buscar la fidelidad y la objetividad de lo registrado en la entrevista, minimizando o eliminando el componente interpretativo, especialmente en lo relacionado con la intención comunicativa de los hablantes.

Las marcas y etiquetas no pretenden reflejar toda la información que ofrece un archivo sonoro, cuya audición será imprescindible para proceder al análisis de determinados aspectos lingüísticos y comunicativos.

Las marcas y etiquetas han de ser compatibles con tratamientos informáticos orientados a diversos fines (p.e. alineamiento texto voz; integración de los textos en el «Corpus del Español del siglo XXI»).

Las marcas y etiquetas no tienen como finalidad señalar posibles objetos de estudio.

A continuación se presentan los signos, convenciones, marcas y etiquetas que han de aplicarse en la transcripción de los textos que integran el proyecto PRESEEA. Junto a cada uno de ellos, aparece una breve descripción de su función; también se marca, cuando es pertinente, si deben aparecer antes (A) o después (D) de la secuencia etiquetada o si reflejan un hecho producido en el instante (I) en que se anota la etiqueta.

Las etiquetas propiamente dichas pueden clasificarse de la siguiente forma: de ruidos, fónicas, léxicas, de dinámica discursiva, de lengua y de transcripción. Las etiquetas siempre van precedidas y seguidas de un espacio, excepto la etiqueta <alargamiento/>. Tras el cuadro general, se hacen otras aclaraciones y comentarios complementarios.

Marcas y etiquetas comunes de PRESEEA

ORTOGRAFÍA Y PUNTUACIÓN

¡!	Enunciados exclamativos
¿?	Enunciados interrogativos
/	Pausa mínima
//	Pausa
:	Tras código de hablante (I: E: A1:)
Mayúsculas	Inicial de nombres propios y siglas
Elementos cuasi léxicos funcionales	Interjecciones; apoyos.Escritura ortográfica (ah, ay, aha, mmm, eeh, pff, bah)
Onomatopeyas	Escritura ortográfica (zas, bum, plas)

ETIQUETADO DE RUIDOS

<ruido = « «/>	Ruido, con especificación de tipo (p.e. <ruido =»chasquido boca»/>) I
<ruido_fondo> </ruido_fondo>	Ruido continuo de fondo AD
<risas = « «/>	Risas, con especificación de emisor/es (p.e. <risas =»E»/>, <risas = «todos»/>) I
<entre_risas> </entre_risas>	Risas simultáneas con el habla AD
<registro_defectuoso></registro_defectuoso>	Fragmento de la grabación de mala calidad AD
<interrupción_de_grabación/>	Interrupción de la grabación I

ETIQUETADO FÓNICO

<énfasis> </énfasis>	Fragmento con pronunciación claramente enfática AD
<alargamiento/>	Alargamiento de sonido D (sin espacios)
<silencio/>	Silencio de un segundo o más I
<palabra_cortada/>	Palabra cortada D
<vacilación/>	Vacilación; titubeo breve I
<sic> </sic>	No es descuido de transcripción AD
<ininteligible/>	Fragmento ininteligible I

ETIQUETADO LÉXICO

<término> </término>	Lexía claramente usada como uso especializado AD
<extranjero> </extranjero>	Extranjerismo (excepto usos de la L2 del hablante) AD
<siglas = [ ]> </siglas>	Siglas; incluye pronunciación AD

ETIQUETADO DE DINÁMICA DISCURSIVA

<cita></cita>	Cita, estilo directo AD
<simultáneo></simultáneo>	Solapamiento (traslape). También se usa en turnos deapoyo, si fuera necesario AD

ETIQUETADO DE LENGUA

<lengua = « «> </lengua>	Cambio de lengua (léxico, oracional, ...), especialmente L2 del hablante, con indicación desarrollada de lengua(p.e. <lengua = «gallego»> </lengua> AD

ETIQUETADO DE TRANSCRIPCIÓN

<transcripción_dudosa> </transcripción_dudosa>	Transcripción dudosa para transcriptor y revisores AD
<tiempo = « «/>	Anotación de minuto y segundo de grabación. (p.e.<tiempo = «02:45»/>) I
<observación_complementaria = « «/>	Observación complementaria I

Comentarios complementarios

- Los dos puntos (:) se utilizan para marcar inicio de turno, tras el código del hablante.

- Las mayúsculas solamente se utilizan en la inicial de nombres propios y en las siglas.

- La pausa mínima en ocasiones es casi imperceptible; puede representar una pausa que coincida con una frontera de entonación. El silencio es claramente perceptible (generalmente más de un segundo). La pausa es intermedia entre los anteriores.

- No se utiliza punto a final de turno. La ausencia de punto equivale a abandono o suspensión, sin especificarse si ha sido voluntario o involuntario. Si el final de un turno resulta de una interrupción por parte del interlocutor y se corta una palabra, se marca <palabra_cortada/>, sin embargo no se marca nada si corta una oración.

- Las pausas entre turnos se marcan al final del turno, no al inicio del siguiente.

- Los elementos cuasi léxicos o paralingüísticos se representan en ortografía ordinaria, de acuerdo con las convenciones más habituales, cuando existan (uhum, ah, eeh, bah). Los ejemplos del cuadro no constituyen una lista cerrada.

- La etiqueta <vacilación/> se usa para marcar pequeños titubeos que no constituyen palabras.

- La etiqueta <sic> no marca usos erróneos o vulgarismos, ni llama la atención sobre formas especiales; solamente se usa cuando la transcripción pudiera interpretarse como error del transcriptor.

- La etiqueta <término> se aplica exclusivamente cuando el hablante haga uso de un tecnicismo, claramente especializado.

- La etiqueta <cita> se usa tanto para las citas textuales como para el estilo directo, usos que en la ortografía convencional se marcarían mediante comillas.

- El solapamiento o traslape de turnos (<simultáneo>) se marca cuando comienza y cuando termina en el discurso del hablante y también se marca cuando comienza y cuando termina en el discurso del interlocutor.

- La etiqueta <tiempo = « «/> se inserta siempre a comienzo de turno, en lugares correspondientes a intervalos de dos minutos aproximadamente o cuando se inicia una fase diferenciada en el desarrollo de la entrevista.

- La etiqueta de <observación_complementaria = « «/> puede incluir cualquier información no prevista en la relación mínima de etiquetas.

- Los equipos que trabajen en comunidades bilingües pueden incluir en la relación de etiquetas otras que tengan que ver con el uso de dos lenguas en la conversación, además del cambio de una lengua a otra, ya recogido en la serie común. En este caso, será imprescindible una descripción detallada del significado de cada etiqueta.

B. TEXTOS SIN ETIQUETAS

La versión de los textos sin etiquetar tiene como finalidad la publicación impresa de los materiales y su lectura convencional. La presentación de un texto sin etiquetar consiste básicamente en disponer una cabecera con los datos esenciales del informante, el número de entrevista, la fecha de la grabación y el texto de la transcripción desprovisto de las etiquetas, tanto las de apertura y cierre, como las aisladas, excepto <risas = « «/> y <silencio/>

Finalmente, se informa sobre los grupos de trabajo que conforman el macro corpus PRESEEA Internacional:

CÓDIGOS DE CIUDADES DE PRESEEA

PRESEEA - ALCALÁ DE HENARES - ES	ALCA
PRESEEA - BARCELONA - ES	BARC
PRESEEA - BARRANQUILLA - CO	BARR
PRESEEA - BOGOTÁ - CO	BOGO
PRESEEA - CÁDIZ - ES	CÁDI
PRESEEA - CARACAS - VE	CARA
PRESEEA - CIPOLLETTI - AR	CIPO
PRESEEA - CULIACÁN - MX	CULI

PRESEEA - GRANADA - ES	GRAN
PRESEEA - GUATEMALA - GU	GUAT
PRESEEA - LAS PALMAS - ES	LASP
PRESEEA - LÉRIDA/LLEIDA - ES	LERI
PRESEEA - MADRID - ES	MADR
PRESEEA - MÁLAGA - ES	MALA
PRESEEA - MEDELLÍN - CO	MEDE
PRESEEA - MÉRIDA - MX	MERI
PRESEEA - MÉXICO - MX	MEXI
PRESEEA - MIAMI - EU	MIAM
PRESEEA - MONTERREY - MX	MONR
PRESEEA - MONTEVIDEO - UR	MONV
PRESEEA - OVIEDO - ES	OVIE
PRESEEA - PALMA DE MALLORCA - ES	PALM
PRESEEA - PEREIRA - CO	PERE
PRESEEA - QUITO - EC	QUIT
PRESEEA - SAN JUAN - PR	SANJ
PRESEEA - SAN MIGUEL - SA	SANM
PRESEEA - SANTIAGO DE CHILE - CH	SCHI
PRESEEA - SANTIAGO DE COMPOSTELA - ES	SCOM
PRESEEA - SEVILLA - ES	SEVI
PRESEEA - TUNJA - CO	TUNJA
PRESEEA - VALENCIA - ES	VALE
PRESEEA - VALLEDUPAR - CO	VALL
PRESEEA - VALPARAÍSO - CH	VALP
PRESEEA - ZARAGOZA - ES	ZARA