Pliego de prescripciones técnicas

1. Introducción..... 3
2. Consideraciones generales..... 4
2.1. Material proporcionado.....4
2.2. Descripción de la tarea.....5
2.3. Identificación de las personas etiquetadoras..... 5
3. Consideraciones particulares.....5
3.1. Puntuación y Capitalización..... 6
3.2. Expresiones especiales.....6
3.3. Palabras incompletas..... 6
3.4. Auto-correcciones..... 6
3.5. Anuncios.....7
3.6. Idioma.....7
3.7. Hablantes..... 8
3.8. Solapamientos.....8
3.9. Segmentos sin habla o con música.....11
3.10. Textos alternativos.....11
3.11. Nombres propios y palabras extranjeras o en castellano.....13
3.12. Números.....14
3.13. Siglas.....15
3.14. Páginas web.....15
3.15. Palatalizaciones..... 15
4. Entrega.....17

1. Introducción

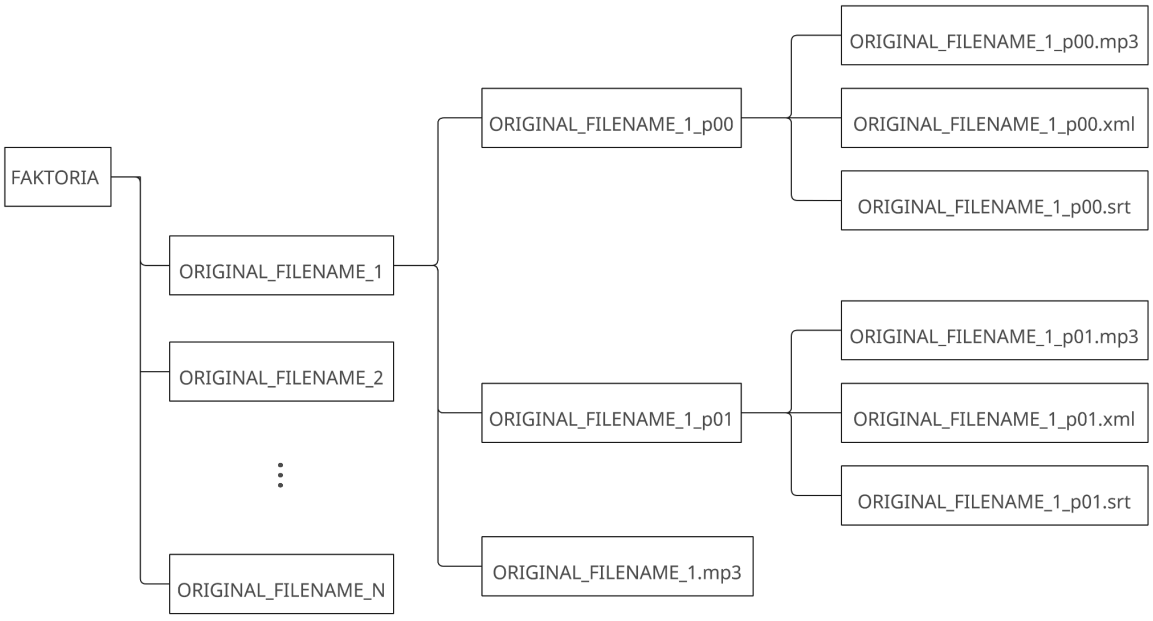
Este documento contiene la descripción de las instrucciones de etiquetado correspondientes al Pliego de Prescripciones Técnicas Particulares del expediente 39/24 PA.

2. Consideraciones generales

2.1. Material proporcionado

Se proporcionará el siguiente material:

- Por un lado, se proporcionarán 51 ficheros de audio en formato .mp3. Cada fichero se corresponde con un programa de radio. La duración de cada programa es de aproximadamente 3 horas. Se proporcionarán más programas en caso de ser necesario.
- Para cada uno de los programas se proporciona el resultado de realizar la transcripción automática con un programa de reconocimiento. Este programa proporciona también información adicional. Por cada programa, existe una carpeta que lleva el nombre del programa con el fichero original del programa completo. Cada fichero '.mp3' queda dividido en dos partes, y cada una se ubica en una subcarpeta diferente. La subcarpeta lleva el nombre del fichero original seguido de _p00 (primera parte) o _p01 (segunda parte). La primera parte (_p00) contiene siempre 1 hora y 50 minutos del programa, y la segunda parte (_p01) contiene el resto del programa, entre 55 minutos y 1 hora y 15 minutos. Cada una de las partes viene con su correspondiente fichero de audio '.mp3' y sus correspondientes transcripciones automáticas en formatos .xml y .srt. La estructura de las carpetas y los ficheros que se proporcionan es la mostrada en la siguiente gráfica:



También se proporcionarán scripts de conversión de formatos .xml al textgrid (del Praat).

2.2. Descripción de la tarea

La tarea principal consiste en corregir las transcripciones de los ficheros XML y añadir las anotaciones indicadas en esta guía de la siguiente manera:

- El fichero de audio queda dividido en segmentos () que podrán ser de habla o eventos no-habla. Cada segmento de habla debe corresponder a una frase completa o incompleta. Cada segmento no-habla se corresponderá con eventos de música, ruido, toses, etc.
- Los segmentos con habla pueden durar hasta un máximo de 30 segundos. Si la frase es más larga hay que dividirla en varios segmentos, (max 30 s). El corte hay que hacerlo en un intervalo de silencio.
- Cada segmento de habla de duración menor de 30 segundos tiene que acabar con alguno de los signos de puntuación siguientes: punto, puntos suspensivos, signo de interrogación, signo de exclamación.
- Hay que alinear el inicio y el fin de segmento. Los tiempos de los segmentos no se deben solapar. El tiempo de inicio (start) de un segmento debe coincidir con el tiempo de inicio de la primera palabra (<word>), y el tiempo de fin (end) de un segmento debe coincidir con el tiempo de fin de la última palabra.
- Las transcripciones automáticas contienen alineamiento a nivel de palabra que no es necesario corregir, salvo si se trata del inicio de la primera y el fin de la última palabra.
- Se deben transcribir segmentos en euskera y en español/castellano pero no en otros idiomas.
- A veces hace falta juntar o dividir las etiquetas de palabras (<word>) para realizar una corrección de la transcripción. En el caso de juntar palabras, mantenemos el tiempo de inicio (start) de la primera palabra y el tiempo de fin (end) de la última palabra (no es necesario corregir la alineación a nivel de cada palabra). En el caso de dividir una etiqueta en varias palabras, se deben mantener los tiempos de inicio y fin, y los demás se pueden poner al azar dentro de los límites de inicio y fin.
- En la etiqueta <word> existe el atributo ‘conf’. Este atributo se puede ignorar.

2.3. Identificación de las personas etiquetadoras

Se debe identificar con un ID único a la persona que ha hecho la anotación de cada fichero. Por eso, junto a los ficheros XML corregidos y anotados debe proporcionarse una lista de los IDs de las personas que han hecho la corrección/anotación. Cada persona debe tener solo un ID.

3. Consideraciones particulares

3.1. Puntuación y Capitalización

- Cuando una frase completa se acaba hay que poner siempre un punto, o signo de interrogación, o signo de exclamación después de la última palabra:

```
<word start="6304.44" end="6304.65">Gaur</word>
<word start="6304.65" end="6304.85">ezin</word>
<word start="6304.85" end="6304.99">dut.</word>
```

- Marcar frases incompletas con puntos suspensivos: "..." unidos a la última palabra del segmento (p.e. cuando se solapan 2 locutores y la frase queda incompleta, v. apartado *Solapamientos*):

```
<word start="6304.44" end="6304.65">Gaur</word>
<word start="6304.65" end="6304.85">ezin...</word>
```

- Usar la puntuación bien: comas, puntos, puntos suspensivos, signos de interrogación/exclamación, guión (v. apartado *Nombres/palabras extranjeros/as o en castellano*). No usar otra puntuación (p.e., dos puntos, punto y coma, etc).
- Usar correctamente mayúsculas y minúsculas.

3.2. Expresiones especiales

- Las palabras que están en el audio pero que no han aparecido en la transcripción automática, hay que transcribirlas e insertarlas, aunque sean interjecciones o palabras cortas tipo 'ondo', 'ba' etc.
- Las expresiones de relleno tipo 'eeeeh', 'aaaaah', 'mmmmm', etcétera tampoco aparecerán en la transcripción automática y hay que insertarlas siguiendo estas reglas:
    - Las palabras eh, ah, oh, etc, siempre con una vocal y con una 'h' siempre al final: #eh#, #ah# etc.
    - Las palabras em, am, um, etc, siempre se transcriben como #mm#.

3.3. Palabras incompletas

- En el caso de palabras incompletas: no transcribir la palabra incompleta, y poner puntos suspensivos a la palabra anterior.

3.4. Auto-correcciones

- En el caso de auto-correcciones, la primera palabra (incorrecta) forma un segmento aparte, y se marca como frase incompleta, seguida con puntos suspensivos. La segunda palabra (correcta) será la primera palabra del segmento siguiente.

- Si la autocorrección se hace cuando se equivocan en una palabra o declinación, etc., en alt1 se usa la palabra entera después de la corrección, y en el valor de la etiqueta <word> lo que corresponde a lo que han dicho. Si se trata de una declinación, se tiene que usar un guión antes:

```
<word start="304.44" end="304.65"  alt1="Herrialdeak">Herrialde  #mm#, -ak</word>

<word start="304.44" end="304.65"  alt1="Hiria">Herria #eh# Hiria</word>
```

3.5. Anuncios

- Se transcriben si son en euskera o castellano

3.6. Idioma

- Para el atributo 'lang', se usan los siguientes códigos de idiomas/dialectos:

Código	**Idioma/Dialecto**
eus	Euskera (no se identifica el dialecto)
biz	Euskera, dialecto de Bizkaia
gip	Euskera, dialecto de Gipuzkoa
ipa	Euskera, dialecto de Iparralde
cas	Castellano
ing	Inglés
fra	Francés
cat	Catalán
gal	Gallego
otro	Otro idioma que no está en la lista o que no se ha podido identificar

- Es imprescindible reconocer si el idioma es castellano o euskera, y si es euskera, en la medida de lo posible, el dialecto. Para los otros idiomas se puede usar el código "otro" si no se puede identificar.

3.7. Hablantes

- En cada fichero, cada hablante o locutor tiene un código que le identifica en el atributo speaker=”codigo_speaker” de la etiqueta . Sin embargo, el mismo código de locutor puede corresponder a dos personas diferentes si se trata de dos ficheros distintos. Si se trata de un presentador del programa, se usa el código “pres” seguido por un número. Esos códigos de presentadores se tienen que mantener en todos los ficheros (p.e., pres2 corresponde siempre a la misma persona en todos los ficheros) :

```
<segment speaker="pres2" lang="eus" start="204.44" end="204.65">

...
</segment>
```

- Si el locutor habla en un dialecto en un segmento, se tiene que marcar en el atributo “lang”:

```
<segment speaker="5" lang="gip" start="204.44" end="204.65">

...
</segment>
```

- Si el mismo locutor en otro segmento habla en otro dialecto/idioma, se debe etiquetar el dialecto del segmento actual. Por ejemplo, si el locutor del ejemplo anterior habla en batua (no se identifica una palabra dialectal en lo que dice) se pone como lang="eus":

```
<segment speaker="5" lang="eus" start="304.44" end="304.65">

...
</segment>
```

3.8. Solapamientos

- Se producen cuando varias personas hablan al mismo tiempo.
- Los solapamientos tienen que ir en segmentos aparte.
- Si dos locutores dicen lo mismo al mismo tiempo, se marcan los dos como locutores con un “+” , p.e. speaker= “1+3” (sin espacios entre los números).
- Si se solapan dos locutores y dicen cosas diferentes, la frase que queda incompleta de cada locutor, hay que indicarla con puntos suspensivos (como se dice en Puntuación y Capitalización:).
- Para el segmento solapado:
    - La transcripción va entre corchetes y la de cada locutor se separa con /

- Si una frase se acaba (frase completa) se usa el punto junto a la última palabra. Si se trata de una frase incompleta, hay que usar los puntos suspensivos.
- Si una frase dentro de un segmento de solapamiento se oye desde el principio, la primera palabra tiene que empezar por mayúscula, y si no se oye desde el principio, se tienen que usar los puntos suspensivos antes de la primera palabra.
- Si hay un trozo no inteligible de cualquiera de los dos locutores, se generan nuevos segmentos, todos ellos de solapamiento.
- Las transcripciones de los trozos no inteligibles son como las normales, con texto '(inint)'.
- En resumen hay 3 casos:

    1. [transcripción 1/transcripción 2]
    2. [transcripción 1/(inint)] o [(inint)/transcripción 2]
    3. [(inint)/(inint)]

    Para estos casos:

    1. Si el solapamiento es continuación de lo que estaba diciendo el primer locutor, se usan puntos suspensivos antes de la primera palabra del primer locutor. Si el final del solapamiento es el comienzo de lo que dice el segundo locutor en el siguiente segmento, hay que poner los puntos suspensivos después de la última palabra del solapamiento del segundo locutor.
    2. Si sólo se entiende lo que dice un locutor, se marca como un nuevo segmento. Para la parte ininteligible se tiene que usar el código "inint" entre paréntesis.
    3. Si se solapan dos locutores pero no se entiende lo que dice ninguno de los dos, el solapamiento tiene que ir en un segmento aparte.

- En todos los segmentos de solapamiento, como atributo de locutor ("speaker") se marcan los dos como locutores con una barra inclinada "/" , p.e. speaker= "1/3" (sin espacios entre los números), siguiendo el orden en que aparecen (en 1/3 primero habla el locutor 1 y luego el locutor 3).

```
Caso 1: Solapamiento donde se entiende lo que dice cada uno:

a. El locutor 1 no se acaba la frase, y no se oye el inicio de la frase del  locutor 3
<segment speaker="1/3" lang="eus" start="204.44" end="204.65">
        <word start="64.44" end="204.65">"[...joan…/…dut…]]"</word>
</segment>



Solapamiento donde se entiende lo que dice cada uno:
```

```
b. El locutor 1 se acaba la frase, y se oye el inicio de la frase del  locutor 3
<segment speaker="1/3" lang="eus" start="204.44" end="204.65">
        <word start="64.44" end="204.65">"[...joan./Baina…]]"</word>
</segment>


Caso 2: Solapamiento donde no se entiende lo que dice el segundo locutor:

El locutor 1 se acaba la frase, y no se oye lo que dice el  locutor 3
<segment speaker="1/3" lang="eus" start="204.44" end="204.65">
        <word start="64.44" end="204.65"> "[...joan./(inint)]"</word>
</segment>


Caso 3: Solapamiento donde no se entiende lo que dice ningún locutor:

<segment speaker="1/3" lang="eus" start="204.44" end="204.65">
        <word start="64.44" end="204.65">"[(inint)/(inint)]"</word>
</segment>
```

- Algunas veces, hay que crear muchos segmentos para la parte solapada. Por ejemplo:
  Segmento1: Habla sólo el locutor 1

  Segmento2: El locutor 1 sigue hablando y se solapa con el locutor 2 (se entiende lo que dice cada uno)

  Segmento3: Los locutores 1 y 2 siguen hablando (se entiende lo que dice el locutor 1)

  Segmento4: Los locutores 1 y 2 siguen hablando no se entiende lo que dice ninguno)

  Segmento5: Los locutores 1 y 2 siguen hablando (se entiende lo que dice el locutor 2)

  Segmento6: Sigue solo el locutor 2.
- Si hay solape de más de dos locutores, se marca segmento con solapamiento y se transcribe como ininteligible ‘(inint)’.

3.9. Segmentos sin habla o con música

Para los segmentos que contienen: música, silencio, ruido, risa, llanto, bostezo etc:

- Música. Dejamos el atributo locutor (“speaker”) vacío, y en el atributo “lang” hay que poner la lengua de la canción si es música con letra. Las canciones no se transcriben. Estos segmentos deben contener una etiqueta <word> con atributos start y end que serán los mismos que los de la etiqueta . El valor de la etiqueta <word> tiene que ser ‘(musika)’ entre paréntesis para música con letra, o ‘(musika ins)’ entre paréntesis, para música instrumental.

- Silencio: Dejamos el atributo locutor (“speaker”) y el atributo “lang” vacíos. Estos segmentos deben contener una etiqueta <word> con atributos start y end que serán los mismos que los de la etiqueta . El valor de la etiqueta <word> tiene que ser una barra baja (“_”).

- Para los demás casos, rellenamos el atributo de locutor (“speaker”) con el código. correspondiente y dejamos el atributo “lang” vacío. En el valor de la etiqueta <word> hay que poner: “(RU)” para ruido, “(RI)” para risa, “(LL)” para llanto, “(BO)” para bostezo, siempre entre paréntesis:

```
Para una canción en euskera:
<segment speaker="" lang="eus" start="63.44" end="204.65">
        <word start="64.44" end="204.65">(musika)</word>
</segment>

Para música instrumental:
<segment speaker="" lang="" start="63.44" end="204.65">
        <word start="64.44" end="204.65">(musika ins)</word>
</segment>

Para un silencio:
<segment speaker="" lang="" start="63.44" end="204.65">
        <word start="64.44" end="204.65">_</word>
</segment>

Para una risa:

<segment speaker="2" lang="" start="63.44" end="204.65">
        <word start="64.44" end="204.65">(RI)</word>
</segment>
```

3.10. Textos alternativos

Algunas palabras se pueden escribir de más de una forma. En este caso hay que añadir un atributo ‘alt1’, ‘alt2’ etc. en la palabra (etiqueta <word>) con el el texto o textos alternativos. Por

ejemplo, con los números (v. apartado *Números*).
En el campo principal de la palabra se transcribe el texto que corresponda más fielmente a lo que se ha pronunciado.

- Si se trata de una pronunciación dialectal en la medida de lo posible se incluirá la transcripción estándar en el texto alternativo::

```
ERRÓNEO:
<word start="6304.44" end="6304.65">jun</word>

<word start="6304.65" end="6305.04">zala</word>

<word start="6304.65" end="6305.04">nijoan</word>

<word start="6304.65" end="6305.04">nitula</word>

<word start="6304.65" end="6305.04">muxu</word>
```

```
CORRECCIONES:

<word start="6304.44" end="6304.65" alt1="joan">jun</word>

<word start="6304.65" end="6305.04" alt1="zela">zala</word>

<word start="6304.65" end="6305.04" alt1="nihoan">nijoan</word>

<word start="6304.65" end="6305.04" alt1="nituela">nitula</word>

<word start="6304.65" end="6305.04" alt1="musu">muxu</word>
```

- Palabras que no existen en euskera. Muchas veces se usan palabras (p.e., del castellano) que no existen en euskera. En estos casos se incluirá la transcripción estándar en el texto alternativo:

```
<word start="6304.44" end="6304.65" alt1="baxua">bajua</word>
```

- Palabra “Eta”: A veces la palabra eta se pronuncia como “ta” o “da”. En el caso de “ta” se tiene que escribir separado. En el caso de “da” se tiene que escribir junto a la palabra anterior en el valor de la etiqueta <word>. En ambos casos en alt1 se incluye la palabra “eta”.:

```
Caso da → va unida a la palabra anterior:
<word start="6304.44" end="6304.65" alt1="berrehun eta hamar">"berrehunda hamar"</word>

Caso ta → va por separado:
<word start="6304.44" end="6304.65" alt1="gorriak eta beltzak">"gorriak ta beltzak"</word>
```

- Palabra “Eta”: A veces la palabra eta se puede pronunciar como, “t” (p.e., Agur t’erdi). En este caso en el valor de la etiqueta <word> se une la t con la palabra siguiente (p.e., terdi), y en alt1 la forma ortográfica (p.e., t’erdi):

```
<word start="6304.44" end="6304.65">Agur</word>
<word start="6304.65" end="6304.75" alt1="t’erdi">terdi</word>
```

3.11. Nombres propios y palabras extranjeras o en castellano

- Los nombres propios en euskera se transcriben en euskera: p.e., Ezcaray → Ezkarai. Para los nombres que pueden corresponder tanto a euskera como a castellano (Carlos/Karlos) se transcriben en euskera: p.e., Carlos→ Karlos. Los nombres que son claramente en castellano se transcriben en castellano: p.e., Jorge→Jorge.
- Palabras y nombres extranjeros: transcribir lo más fiel a la pronunciación del hablante, utilizar la transcripción correcta extranjera en el campo “alt1” y añadir el idioma en el campo ‘lang’.

```
Ejemplo 1:

<word start="1352.00" end="1353.56" lang="ing" alt1="Google"> Guguel</word>
<word start="1352.56" end="1353.96" lang="ing"> Maps</word>

Ejemplo 2:

<word start="1352.00" end="1353.56" lang="ing" alt1="Biden"> Baiden</word>
```

- Para las palabras extranjeras declinadas se usa un guión y no se deja espacio p.e. Google Maps-en:

```
<word start="1352.00" end="1353.56" lang="ing" alt1="Google"> Guguel</word>
<word start="1352.56" end="1353.96" lang="ing"> Maps-en</word>
```

- Palabras o frases de castellano que se usan en euskera (p.e., Es que, vale etc.) se transcriben en euskera → Es ke, bale.
- Para palabras en castellano que se usan en un segmento de euskera (salvo en el caso de nombres propios), se tiene que usar un atributo “lang” dentro de la etiqueta de la palabra (<word>), con el código “cas”.

```
<word start="1352.00" end="1353.56"> baina</word>
<word start="1352.56" end="1353.96"  lang="cas" > nos</word>
<word start="1353.96" end="1354.26"  lang="cas"> fuimos</word>
```

```
```

3.12. Números

En general, el número escrito en cifras se incluye en el atributo ‘alt1’ y la transcripción ortográfica en el valor de la etiqueta.

- Si la transcripción ‘fiel’ no corresponde con la transcripción reglada, entonces la transcripción reglada irá en el campo ‘alt2’.
  Ejemplo:

```
<word start="1352.00" end="1353.56" alt1="217" alt2="berrehun eta hamazazpi">"berrehunda hamazazpi"</word>
```

```
<word start="1352.00" end="1353.56" alt1="8ren" alt2="zortziren">zortzien</word>
```

- Punto decimal: usar lo que se ha pronunciado (‘koma’ o ‘puntu’ o ‘kakotx’) en el valor de la etiqueta <word>. Usar siempre la coma (,) para separar los números en el atributo ‘alt’:

```
- ERRÓNEO:
  <word start="1352.00" end="1353.56">19,99</word>

- Corrección 1:
- <word start="1352.00" end="1353.56" alt1="19,99">"hemeretzi koma laurogeita hemeretzi"</word>
- Corrección 2:
- <word start="1352.00" end="1353.56" alt1="19,99" alt2="hemeretzi koma laurogeita hemeretzi">"hemeretzi koma larogeita hemeretzi"</word>
```

- Puntos separadores de miles, millones etc: no utilizar

- Números ordinales: en el valor de la etiqueta poner la cifra seguida de un punto (8.--> zortzigarren)

```
<word start="1352.00" end="1353.56" alt1="8."> zortzigarren</word>
```

- Números romanos : Usarlo únicamente con nombres propios p.e., Karlos IV, etc. hay que usar los números romanos en alt1:

```
<word start="1352.00" end="1353.56"> Karlos</word>
<word start="1352.56" end="1353.96" alt1="IV.a">laugarrena</word>
```

- Fechas: Con las fechas, usar la transcripción expandida en el campo principal p.e., 2018ko urriaren 11.:

```
<word start="1352.00" end="1353.56"  alt1="2018ko urriaren 11" >”bi mila eta hemezortziko urriaren hamaika"</word>
```

3.13. Siglas

- Las siglas se transcriben como han sido pronunciadas (p.e. EHU→ e hatxe u) . Las siglas ocupan el valor de una sola etiqueta <word>. En el atributo ‘alt1’ se tiene que poner la sigla y en ‘alt2’ la palabra “sigla”.

```
<word start="1352.00" end="1353.56" alt1="EHU" alt2="sigla">"e hatxe u"</word>
```

- Para las siglas declinadas, hay que usar siempre un guión entre la sigla y la declinación sin espacio.

```
<word start="352.00" end="353.56" alt1="PP-ren" alt2="sigla">"pe pe-ren"</word>
```

3.14. Páginas web

- Para páginas web, en texto alternativo hay que poner la dirección de la página, tal como se usaría en un navegador web p.e., alt1= “eitb.eus” y en ‘alt2’ usar siempre la palabra “url” (alt2=”url”), mientras que en el valor de la etiqueta <word> se tiene que usar: a) las letras separadas si se trata de una sigla o signos de puntuación, b) las palabras juntas, si se trata de palabras enteras, c) una combinación de a) y b) :

```
Ejemplo Siglas:
<word start="6304.44" end="6304.65" alt1="eitb.eus" alt2="url">"e i te be puntu eus"</word>

Ejemplo palabras:
<word start="6304.44" end="6304.65" alt1="euskalirratia.eus" alt2="url">"euskalirratia puntu eus"</word>

Ejemplo combinación:
<word start="6304.44" end="6304.65" alt1="euskadi-euskalirratia.eus" alt2="url">"euskadi marra euskalirratia puntu eus"</word>
```

3.15. Palatalizaciones

- Muchas veces se usan palabras palatalizadas con un significado más afectivo. Si existen en el diccionario (p.e., pollit), se incluirá la transcripción en el campo principal:

```
<word start="6304.44" end="6304.65">pollit</word>
```

4. Entrega

Por cada fichero de audio, se debe entregar un archivo XML. Cada XML debe contener la división de cada audio en segmentos (). Cada segmento debe contener la división por palabras (<word>). Por ejemplo:

```
<segment speaker="18" lang="eus" start="46.55" end="47.15">
        <word start="46.55" end="46.82"> Uste</word>
        <word start="46.82" end="47.00"> dut</word>
        <word start="47.00" end="47.15"> baietz.</word>
</segment>
<segment speaker="22" lang="eus" start="47.31" end="48.23">
        <word start="47.31" end="47.78"> Bagoaz</word>
        <word start="47.78" end="48.23"> aurrera.</word>
</segment>
```

- El xml debe contener todos los nuevos atributos (p.e., alt1, alt2, etc.)
- Un fichero de Excel con dos columnas: en la primera columna aparecerá el nombre del fichero mp3 y al lado del nombre del fichero, en la segunda columna, el ID de la persona que ha hecho la anotación/corrección.
- Un fichero de transcripción por cada audio anotado, según el software que se ha usado (p.e., en formato .TextGrid, si se ha usado Praat, o en formato .trs, si se ha usado Transcriber).
- La estructura de las carpetas que se proporciona se debe mantener.
- La carpeta de cada fichero mp3/parte debe contener:
    1. El fichero .mp3 que se ha proporcionado
    2. El fichero .xml corregido y anotado según esta guía
    3. El fichero de transcripción según el software de anotación (p.e. .TextGrid o .trs)

    Los tres ficheros deben tener el mismo nombre con diferente extensión, manteniendo el nombre de cada fichero .mp3 proporcionado