DE UN SISTEMA DE ANOTACIÓN SINTÁCTICA PARA LENGUA ESCRITA A OTRO PARA LENGUA ORAL[1]

FROM A SYNTACTIC ANNOTATION SYSTEM FOR WRITTEN LANGUAGE TO ANOTHER FOR SPOKEN LANGUAGE

María Paula Santalla del Río
Universidade de Santiago de Compostela
Eva María Domínguez Noya
Instituto da Lingua Galega / Centro Ramón Piñeiro para a investigación en humanidades

RESUMEN
En el curso de la producción manual de una muestra de corpus oral sintácticamente analizada, explicamos cómo adaptamos un sistema de anotación sintáctica para lengua escrita preexistente. Constatamos que lo que requiere modificación en ese sistema no exige la adición de etiquetas o estructuras sintácticas nuevas, sino la incorporación de mecanismos para el tratamiento de lo que es propio de la lengua oral. Sin partir de una delimitación previa de las unidades de análisis, procedemos reconociéndolas a medida que analizamos sintácticamente, al margen de la segmentación prosódica e identificando las estructuras sintácticas, de entre las recogidas en nuestro sistema de anotación, más amplias posible; integramos asimismo en el análisis, por medios diversos, las denominadas disfluencias orales (repeticiones, truncamientos y reformulaciones), y, reutilizando las etiquetas que para ello ya utilizábamos en la anotación de lengua escrita, damos cuenta de la mayor presencia en la oralidad de marcadores, conectores y vocativos.

PALABRAS CLAVE: corpus oral, disfluencias, marcas de oralidad, anotación sintáctica, adaptación de tagset

ABSTRACT
In the course of the manual production of a syntactically analyzed spoken corpus sample, we explain how we adapt a pre-existing system of syntactic annotation for written language. We found that what requires modification in that system does not implies the addition of new tags or syntactic structures, but the incorporation of mechanisms for the treatment of what is proper to the oral language. Without starting from a prior delimitation of the units of analysis, we proceed recognizing

[1] Esta investigación ha podido beneficiarse de la financiación FEDER/Ministerio de Ciencia, Innovación y Universidades – Agencia Estatal de Investigación/ESLORA+ (FFI2017-86379-P). María Paula Santalla forma asimismo parte del grupo Gramática del Español de la Universidad de Santiago de Compostela, beneficiario de la ayuda 'Consolidación 2020 GRC GI-1372 Gramática do español' de la Consellería de Cultura, Educación e Universidades de la Xunta de Galicia (ED431C 2020/21).

them as we parse, regardless of the prosodic segmentation and identifying, among those collected in our annotation system, the most extensive syntactic structures possible; we also integrate into the analysis, by various means, the so-called oral dysfluencies (repetitions, truncations and reformulations), and, reusing the labels that we already used in the annotation for this in written language, we account for the greater presence in the orality of markers, connectors and vocatives.

KEYWORDS: spoken corpus, dysfluencies, syntactic annotation, orality marks, tagset adaptation

1. INTRODUCCIÓN

El desarrollo de corpus orales en español comenzó hace más de medio siglo con el *Proyecto de estudio coordinado de la norma lingüística culta del español hablado en las principales ciudades de Iberoamérica y de la Península Ibérica* (Lope Blanch, 1986). Desde entonces otros corpus orales de español se han sumado al anterior (PRESEEA, COSER, C-ORAL ROM, VAL-ES-CO, o las secciones orales de CREA o de CORPES). Hay también en la actualidad corpus analizados sintácticamente disponibles para el español: AnCora, IULA Spanish LSP Treebank, UAM Spanish Treebank o CSA. Pero todos estos corpus son de lengua escrita, sin que, sin embargo, se puedan encontrar para el español corpus orales sintácticamente anotados, aunque sí se dispone ya de ellos en otras lenguas. Si bien podrían citarse más, nos limitamos aquí a aquellos a los que nos referimos a lo largo de esta exposición[2]: para el portugués, el corpus C-ORAL (Bick, 2012; Bick et al., 2013); para el inglés, el Switchboard (Meteer, 1995; Taylor, 2003); el Corpus Gesproken Nederlands, CGN, para el holandés (Oostdijk, 2000; Oostdijk et al., 2002; Van Der Wouden, 2003; Schuurman et al., 2004); el Tartu University Corpus of Spoken Estonian, TUCSE, para el estonio (Müürisep y Uibo, 2006); el Corpus of Spontaneous Japanese, CSJ, para el japonés (Uchimoto et al., 2006), y el Arabic Treebank, AT, para el árabe (Maamouri et al., 2010).

Entre los corpus orales desarrollados para el español no hemos citado arriba el corpus ESLORA (Vázquez et al., 2020), un corpus oral del español de Galicia que consta en su versión actual – 2.0 de 2020 – de 768.005 palabras procedentes de la transcripción ortográfica de 60 horas de entrevistas semidirigidas y 20 horas de conversaciones de hablantes gallegos grabadas entre los años 2007 y 2015. El corpus, que alinea la transcripción con el audio para facilitar el acceso inmediato a este, fue enriquecido con la anotación morfosintáctica automática y la lematización de los textos que lo integran.

En el seno del proyecto en el que se desarrolla este corpus, a fin de avanzar en el conocimiento de la lengua y la disposición de materiales del registro oral con los que favorecer la realización de descripciones gramaticales que tengan en cuenta también la oralidad, uno de los objetivos actuales de trabajo consiste en analizar sintácticamente una pequeña muestra del corpus, unas 50.000 palabras. Se trata de un objetivo modesto, sin duda en términos de volumen, que responde a los hechos de que ni hay herramientas de análisis automático para español oral, ni hay un corpus

[2] Páginas de información y/o consulta de estos corpus se recogen en el apéndice. Aquí hacemos referencia a las publicaciones en que bien se describe el proceso de desarrollo de estos corpus desde un punto de vista global, bien en concreto de su anotación sintáctica, que es nuestra prioridad.

de entrenamiento que pueda servir para entrenar un posible analizador, ni queremos tampoco sacrificar profundidad de análisis. Nuestro objetivo, no tan modesto a la luz de estas circunstancias, se concreta entonces en obtener ese corpus de entrenamiento, en obtenerlo de modo manual y en adaptar para ello un sistema y una guía de anotación sintáctica previos (los aplicados al corpus CSA).

En este artículo se describe la primera fase de investigación para la obtención de ese resultado. En la sección 2 se describe el proceso de anotación: procedimiento, recursos previos y recursos adicionales, o adaptaciones de los existentes, requeridos por la oralidad. En las secciones siguientes se describen más detalladamente esas novedades o adaptaciones: en la sección 3, las que tienen que ver con la identificación de las unidades de análisis; en la sección 4, con las marcas de oralidad; en la sección 5, con marcadores, vocativos, conectores y tópicos, y en la sección 6, con las disfluencias propias de la oralidad. En la sección 7 se presentan las conclusiones del trabajo.

2. EL PROCESO DE ANOTACIÓN

2.1. Procedimiento y herramientas

La anotación sintáctica del fragmento del corpus ESLORA que está previsto anotar en esta fase preparatoria, una entrevista, se lleva a cabo de manera íntegramente manual. Con este fin, se ha utilizado, con leves adaptaciones para su uso en la anotación de lengua oral, la herramienta de anotación sintáctica manual PaME (http://galvan.usc.es/drasae), creada en el marco del proyecto DRASAE.[3] La apariencia de la pantalla principal para la edición de secuencias en PaME es como se ve en la figura 1.

Figura 1. La herramienta de análisis manual PaME

Su funcionamiento es muy simple: al seleccionar un subsegmento de un segmento (la secuencia inicial o un segmento ya subsiguientemente identificado en ella), al clicar en el botón con el símbolo + a la derecha de ese segmento, el subsegmento seleccionado se copia como hijo del segmento en el que hemos pulsado el botón +, en una caja en el nivel de profundidad siguiente. A continuación, el anotador puede seleccionar en las

[3] <https://gramatica.usc.es/proxectos/drasae/>

correspondientes ventanas desplegables la función y el tipo de unidad del segmento en cuestión. Dos huecos de información adicionales permiten, el primero, añadir índices, con significados diversos, a los segmentos a los que afectan, y, el segundo, señalar posibles situaciones de ambigüedad. El resto de los botones finales junto al botón + permiten otras operaciones de movimiento de los segmentos o su eliminación. Actualmente debemos decir que PaME, aunque ha resultado y resulta de suma utilidad para poder llevar a cabo el ensayo cuya descripción nos ocupa, es hoy una herramienta que próximamente planeamos hacer más visual y amigable para la codificación de algunas circunstancias propias de la oralidad y de ESLORA en concreto; por lo que algunas de las soluciones adoptadas a las que aquí haremos referencia (los modos en que se reflejan en estos momentos en la anotación el tratamiento de las repeticiones o las reformulaciones, por ejemplo) pueden sufrir alteraciones en un futuro no lejano: no así el hecho más importante de que esas circunstancias se marquen como tales en el conjunto de la anotación sintáctica y, según los casos, sean, además, internamente, objeto de una anotación estrictamente sintáctica (estructural y funcional).

2.2. El sistema de anotación

Para la anotación sintáctica de ESLORA se utilizó, adaptado como explicamos más adelante, el sistema de anotación[4] desarrollado en el marco del proyecto DRASAE ya citado. Un ejemplo de una unidad de ESLORA analizada se muestra en la figura 4. Cada línea horizontal es un segmento identificado en el análisis de una secuencia. Para ese segmento, se recoge esencialmente el nivel de análisis (cuyo reconocimiento se apoya también en el sangrado), el segmento mismo, su función (si la hay) y su unidad. Se trata, como puede observarse, de un árbol tradicional, solo que la representación jerarquizada de los nodos que lo constituyen se orienta en vez de verticalmente, esto es, de arriba a abajo, horizontalmente, esto es, de izquierda a derecha. Como sistema de anotación desarrollado en el proyecto DRASAE para la anotación de lengua escrita, este sistema consta de 42 etiquetas para la especificación de función (tales como 'Sujeto', 'Complemento directo' o 'Modificador') y 32 para la de unidad sintáctica (tales como 'Estructura coordinada', 'Cláusula', 'Frase nominal' o 'Sustantivo'). Los rasgos principales de esta anotación son:

1. 'Análisis constitutivo exhaustivo'. Proporcionamos estructuras constitutivas en forma de árbol, identificando a lo largo de ellas todos los segmentos de las secuencias analizadas desde el más amplio, la secuencia misma, hasta las unidades léxicas, y asignándole a cada uno de esos segmentos, con salvedades motivadas por la necesidad de dar cabida en la anotación a cuestiones no sintácticas, etiquetas de función y unidad.
2. 'Respeto del carácter presencial y el orden secuencial de los constituyentes'. Los argumentos no explícitos no están representados en el análisis y el orden de las secuencias no es nunca alterado.
3. 'Consistencia, simplicidad, familiaridad y tradición': la primera en dos sentidos: el mismo fenómeno es siempre anotado de la misma manera a lo largo del corpus y fenómenos parcialmente semejantes son anotados dando cuenta de esa semejanza. La segunda, porque entendemos que cuantas menos etiquetas y recursos adicionales de la anotación (parentización, sangrado, coindización,

[4] <https://gramatica.usc.es/wiki/drasae>

etc.), mejor. Las dos últimas para que la anotación plasme un análisis que resulte a los usuarios lo más familiar posible en cuanto a segmentación y jerarquización, así como a discriminación y denominación de funciones y unidades sintácticas: con este objetivo, sin renunciar a apartarnos de esa descripción donde lo encontramos oportuno por razones lingüísticas o estrictamente representacionales, codificamos un análisis sintáctico fundamentalmente tradicional, entendiendo por tal el que hace la RAE en su *Nueva gramática de la lengua española* (RAE 2009).

4. 'Atención especial a la identificación y análisis de estructuras suprapredicativas'. Esto es, a estructuras que ponen en relación, prioritaria, pero no exclusivamente,[5] estructuras predicativas, estas a su vez configuradas en torno a un verbo que funciona como predicado y elementos en torno a él que funcionan como sujeto, complemento directo, etc. Las estructuras suprapredicativas son estructuras adversativas, condicionales, comparativas, etc. que sustentan el número finito y restringido de relaciones lógicas que a los hablantes les interesa expresar entre lo que por medio de esas estructuras ponen en conexión.

2.3. La oralidad

El tipo de análisis descrito en la sección anterior se hacía operativo en la lengua escrita con la que se trabajaba en el proyecto DRASAE en el marco delimitado por la puntuación ortográfica, que, de acuerdo con ese método de trabajo, es la clave sobre la que se identifican las que se consideran unidades de referencia de la lengua escrita en cuestión y, en consecuencia, también para el análisis sintáctico. No es posible, sin embargo, en la lengua oral que ilustra ESLORA hacer operativo ese análisis sobre la base de ese mismo marco de delimitación de unidades, ya que la introducción de puntuación con ese significado está ausente de la transcripción llevada a cabo de las entrevistas y conversaciones recogidas en el corpus, para las que la transcripción marca, en cambio, las pausas, con distintas longitudes, y los silencios (vid. sección 3).

La delimitación de unidades, que no puede basarse como en la lengua escrita en una puntuación inexistente o muy reducida en la transcripción, es, pues, el primer y principal condicionante a la hora de anotar sintácticamente lengua oral. Pero no es el único. Junto a ella hay que tener en cuenta la presencia de marcas de oralidad (codificaciones de eventos que suceden durante el discurso hablado [risas, ruidos, etc.] o lo condicionan de algún modo [solapamientos]), la proliferación en lengua oral de marcadores, conectores, vocativos o topicalizaciones (especialmente las exentas) y la huella que el proyecto constantemente revisado y modificado que es el discurso hablado deja en él, esto es, las disfluencias orales (repeticiones, reformulaciones y truncamientos).

Incluso en estas circunstancias, que lo dificultan enormemente, nuestro propósito es ofrecer para la lengua oral de ESLORA el mismo tipo de análisis (que alcance a las estructuras suprapredicativas) y con el mismo nivel de detalle que para la lengua escrita. Es decir, ante, por ejemplo,

[5] Efectivamente, las referidas son estructuras cuya manera de relacionar segmentos donde se muestra más rentable es operando sobre estructuras con verbo, pero queremos insistir en que no está limitada su aparición a circunstancias de esta clase, pues pueden también relacionar estructuras frasales tanto nominales como adverbiales.

yo noté mucha diferencia [pausa/] [alargamiento]pero[/alargamiento] [alargamiento]tú[/alargamiento] date cuenta que yo me vine a Santiago [pausa_larga/] [ruido tipo="chasquido boca"/] cuando [alargamiento]tenía[/alargamiento] [silencio/] pues [pausa/] aún no [palabra_cortada]ll[/palabra_cortada] aún no cumpliera los veinte años [pausa/]

queremos analizar todo lo que creemos que es sintaxis (no diferente de la que se puede encontrar en lengua escrita) sin que lo que es propio de la lengua oral lo oculte. Si, de hecho, eliminamos los elementos que podrían tener este efecto (pausas y silencios, especificaciones que caracterizan a parte del discurso [alargamiento], eventos [ruido], conectores [*pues*], truncamientos [*ll*] y reformulaciones [*cuando tenía – aún no ll – aún no cumpliera*], lo resultante es una secuencia gramatical que podría también haber sido encontrada en un texto escrito, una secuencia que queremos identificar y analizar en detalle como lo habríamos hecho si efectivamente así hubiera sido:

yo noté mucha diferencia pero tú date cuenta que yo me vine a Santiago cuando aún no cumpliera los veinte años

No solo eso, sino que cuando lo que es propio de la lengua oral tiene también calado sintáctico, lo cual ocurre especialmente en las partes rechazadas de las reformulaciones, también queremos dar cuenta de ello (*cuando tenía*, aunque inacabado, es, creemos, una estructura de interés sintáctico que no queremos dejar fuera del análisis).

Si comparamos este planteamiento con los del CGN o del Switchboard, también constitutivos, vemos, por ejemplo en el segundo, que el análisis es, en primer lugar, más esquemático, tal y como puede observarse en la figura 2. Así, la secuencia *{D Well,} [I, + I] think it's a pretty goog idea. /* correspondiente al código de hablante A4, se identifica como una S – *Sentence* – en la que se da cuenta de la repetición de *I* como *Sujeto,* fenómeno codificado entre corchetes en la transcripción que se mantiene asimismo en el nivel de anotación sintáctica.[6] No se señala, sin embargo, que esa cláusula subordinada desempeña la función de CD.

```
A.4: {D Well, } [ I, + I ] think it's a pretty good idea. /

((CODE SpeakerA4.))
((S (INTJ Well)
    ,
    (EDITED (RM [)
        (NP-SBJ I)
        ,
        (IP +))
    (NP-SBJ I)
    (RS ])
    (VP think
      (SBAR 0
          (S (NP-SBJ it)
            (VP 's
              (NP-PRD a
                  (ADJP pretty good)
                  idea)))))
```

Figura 2. Muestra del nivel de análisis sintáctico tomada del Penn Treebank

[6] Se marca entre corchetes la repetición íntegra, en la que el signo + separa aquello que se repite o reformula de lo formulado inicialmente.

En sentido contrario, debe valorarse especialmente que en este corpus sí se da cabida a elementos cuasiléxicos, repeticiones y correcciones[7] en la anotación morfológica y sintáctica. Es de notar, sin embargo, y frente a lo que veremos que sucede en ESLORA, que estos fenómenos se codifican ya en el nivel de transcripción, lo cual facilita enormemente la posterior anotación gramatical semiautomática.

Desde el punto de vista del detalle en la jerarquización, el corpus C-ORAL, al codificar, frente a los anteriores, un análisis dependencial, no es contrastable con ESLORA; desde el punto de vista funcional, en cambio, ambos corpus son comparables. Por otro lado, el análisis en este corpus no se aplica sobre lo que, tras una fase intermedia de marcación que opera sobre la transcripción, está contenido entre caracteres no alfanuméricos, las comillas angulares, o que está codificado con símbolos como '&' o '+': estas marcaciones contienen a los distintos fenómenos característicos de la oralidad, palabras cortadas o retracciones, a cuyo análisis se renuncia por completo en C-ORAL, simplemente manteniéndolos en el texto a modo de metaetiquetas. Ilustra este proceder un brevísimo segmento tomado de Bick (2012: 6) en el que se observa cómo el cuasiléxico (*hhh*) y la palabra cortada (*&dire*), cuya transcripción ortográfica se recoge en la primera línea, se transforman en el nivel de análisis gramatical dispuesto verticalmente en la metaetiqueta '<nonword>':

*GIL: hhh eu tenho &dire
<GIL:>
<nonword:hhh>
eu [eu] PERS M/F 1S NOM @SUBJ
tenho [ter] <fmc> V PR 1S IND VFIN @FMV
<nonword:&dire>

Estas diferencias en el detalle en un sentido u otro del análisis, que no es tan minucioso en estos corpus como en ESLORA, son probablemente deudoras del método de análisis, que es semiautomático en ambos casos, y no estrictamente manual como sucede en ESLORA. Recorremos a partir de ahora más en profundidad los modos en que en ESLORA se enfrenta la anotación sintáctica teniendo en cuenta lo que es propio de la oralidad.

3. UNIDADES DE ANÁLISIS

Mientras que el texto de la lengua escrita proporciona marcas de párrafo, puntos, puntos y coma y otros signos de puntuación a partir de los cuales se segmentan las unidades sintácticas, el texto correspondiente a la lengua oral, al menos el de ESLORA, carece de puntuación, con la salvedad de los signos de exclamación e interrogación, por lo que, como ya indicamos más arriba, a priori, la delimitación de la unidad de análisis sintáctico se complica. Cuando el análisis sintáctico está previsto desde el principio, en la transcripción se hace ya el esfuerzo de identificar las que vayan a ser las unidades del análisis sintáctico, que trabaja sobre ellas. Si no es así, cuando el análisis va a ser automático o semiautomático, se introduce una fase de identificación de esas unidades. Si, en cambio, va a ser manual, como en el caso de ESLORA, puede todo hacerse al tiempo.

[7] Puede verse una pequeña muestra de los tres niveles de codificación (disfluencias, análisis morfológico y análisis sintáctico) en <https://web.archive.org/web/20131109202842/http://www.cis.upenn.edu/~treebank/>.

Todos los corpus cuyo desarrollo hemos revisado, el CGN, el CSJ y los corpus C-ORAL y Switchboard (así como los que siguen la estela de estos, AT y TUCSE), que llevan a cabo análisis semiautomático, delimitan los enunciados en la fase de transcripción. Lo ilustramos por medio del C-ORAL y el Switchboard, que para esa delimitación emplean las siguientes marcas específicas: '//' y '/', que señalan, respectivamente en C-ORAL y el Switchboard, enunciados completos, y '+' y '-/', que indican que están inconclusos. C-ORAL utiliza una tercera marca para distinguir las pausas breves:[8] '/', que posteriormente se desambiguan como rupturas o no de enunciado. Constituyen una muestra de enunciados en el corpus C-ORAL (1) y en el Penn Treebank[9] (2) los siguientes ejemplos:

(1) *GIL: <eu &a [/2] eu acho que e> esse [/2] e esse aqui o' // <&he> +
(2) B.3: {D Well } what do you think about the idea of, {F uh, } kids having to do public service work for a year? / Do you think it's a , -/

Frente a esto, en ESLORA el registro oral textual se fracciona en la transcripción exclusivamente conforme a criterios prosódicos, de modo que una etiqueta de pausa larga, un silencio o un cambio de turno establecen la existencia de fragmentos prosódicos diferentes. Estos fragmentos no pueden por sí solos ser el objeto del análisis sintáctico, puesto que, incluso aunque coincidan, que no siempre, con segmentos sintácticos, separan con mucha frecuencia segmentos que configuran unidades sintácticas más amplias, estructuras frasales, predicativas o, especialmente y de más interés para nosotros, suprapredicativas.

Así, por ejemplo, la secuencia "Si es un día normal y no salgo o no tengo ningún tipo de reunión o estoy en casa, a las ocho y media nueve hacemos como una especie así de medio cena, y si hay algo interesante en la tele lo vemos", podría ser razonablemente analizada como una 'Estructura condicional' cuyos CONDICIONANTE – "Si es un día normal y no salgo o no tengo ningún tipo de reunión o estoy en casa" – y CONDICIONADO – "a las ocho y media nueve hacemos como una especie así de medio cena, y si hay algo interesante en la tele lo vemos" – están ambos desempeñados por una 'Estructura aditiva' con dos MIEMBROs. No obstante, la segmentación prosódica proporcionada por la transcripción, de ser lo determinante para la identificación de unidades sintácticas, impediría en ESLORA el reconocimiento de esa estructura, ya que esa unidad está segmentada en el corpus en cuatro fragmentos prosódicos distintos, en virtud de la pausa larga que figura al final de cada uno de ellos:

si es un día [alargamiento]normal[/alargamiento] [pausa_larga/]
a las [pausa_larga/]
y no [alargamiento]salgo[/alargamiento] o no tengo ningún tipo de [alargamiento]reunión[/alargamiento] o estoy en [alargamiento]casa[/alargamiento] o tal [pausa/] pues eso [pausa/] pues a [alargamiento]las[/alargamiento] ocho y media nueve [pausa/] [palabra_cortada]m[/palabra_cortada] hacemos como una especie así de medio cena [pausa_larga/]
y eeh [pausa/] si hay algo interesante en la tele lo vemos [pausa_larga/]

[8] En su homologación con el registro escrito, en el nivel de anotación sintáctica las marcas '//', '+' y '/' se convierten respectivamente en los signos de puntuación ';', '…' y ','.
[9] Tomamos este ejemplo del nivel de codificación de disfluencias referido más arriba.

Así pues, en ESLORA-CSA debe ser el anotador quien delimite la unidad sintáctica, proponiéndose, como principio fundamental de esa delimitación, el reconocimiento de los segmentos más amplios posible, atravesando para ello, si es preciso, las fronteras de la segmentación prosódica o del turno de habla. Las unidades que el anotador identifique podrán, así, atravesar turnos sucesivos o no del mismo hablante o de hablantes distintos, tendrán en cuenta, pero no se someterán a ella, la prosodia y dependerán de modo importante de la existencia de nexos que evidencien explícitamente su existencia. Así, un fragmento prosódico puede constituir una unidad sintáctica, como sucede con (3); esta puede estar conformada por varios segmentos prosódicos sucesivos, como ocurre en (4), o incluso la unidad sintáctica puede estar constituida por un segmento de uno de los fragmentos, atravesar un turno de palabra simultáneo y terminar en el siguiente fragmento, como tiene lugar en (5):[10]

(3) trabajo en la universidad [alargamiento]en[/alargamiento]
[alargamiento]en[/alargamiento] la administración [pausa_larga/]

(4) [ruido tipo="chasquido boca"/] eehmm me levanto sobre las siete de la mañana [silencio/]
más o menos [pausa_larga/]
desayuno [pausa/] que es lo primero que hago [pausa_larga/]
[risa/] después me ducho [pausa/] me arreglo [pausa/] y me voy a trabajar [pausa_larga/]

(5) ~~[ruido_inicio tipo="ruido de fondo"/]muy bien [pausa/] vaya día que está[ruido_fin tipo="ruido de fondo"/]~~ tenemos un día
~~sí [pausa/] voy a poner [pausa_larga/]~~
espantoso horrible y asqueroso pero bueno [silencio/]

Asimismo, en el sistema aplicado en ESLORA-CSA, a la hora de analizar un segmento sintáctico tenemos en cuenta las respuestas parciales a una pregunta y las coconstrucciones,[11] esto es, aquellas situaciones en las que un hablante termina lo iniciado por otro hablante, de modo que el análisis practicado evidencia la relación, sobreponiéndose a la sobresegmentación prosódica y discursiva, aunque por el momento no se explicita ese vínculo mediante ningún otro código. El análisis es capaz de reflejar, por lo tanto, que una unidad sintáctica se extiende a lo largo de varios fragmentos prosódicos, que pueden ser contiguos o no y pertenecientes al mismo hablante o no. Así, la primera parte de una unidad en esa circunstancia se identifica, tanto en su función como en su tipo de unidad, con el nombre que corresponda seguido de I (por ejemplo, EFECTO I, 'Estructura condicional I') y de un número arábigo que indica el número de fragmentos a lo largo de los cuales esa función/unidad se extiende. Todas las partes siguientes de la unidad fragmentada, no solo la segunda, llevarán el número romano II (por ejemplo, EFECTO II, 'Estructura condicional II'), pero el número arábigo subsiguiente constará en este caso de 2 dígitos, de los cuales el primero indicará la posición que ocupa el fragmento en la serie, y el segundo, el número total de fragmentos de la serie. Así, como se muestra en la figura 3, *me levanto sobre las siete de la mañana [silencio/]* del nivel 0 ocupa la posición primera de los cuatro fragmentos a lo largo de los cuales se

[10] El formato es nuestro para indicar que el texto tachado no forma parte de la unidad sintáctica que ejemplificamos.

[11] En el Penn Treebank no se codifican las coconstrucciones, de forma que la intervención del hablante 2 cuenta como unidad independiente y la del hablante 1 se marca como inconclusa.

extiende la estructura aditiva recogida arriba en (4), y *sobre las siete de la mañana [silencio/]* es el 'Complemento circunstancial I' desempeñado a su vez por la 'Frase preposicional I', que termina de construirse en el fragmento siguiente con el adverbio *más o menos*:

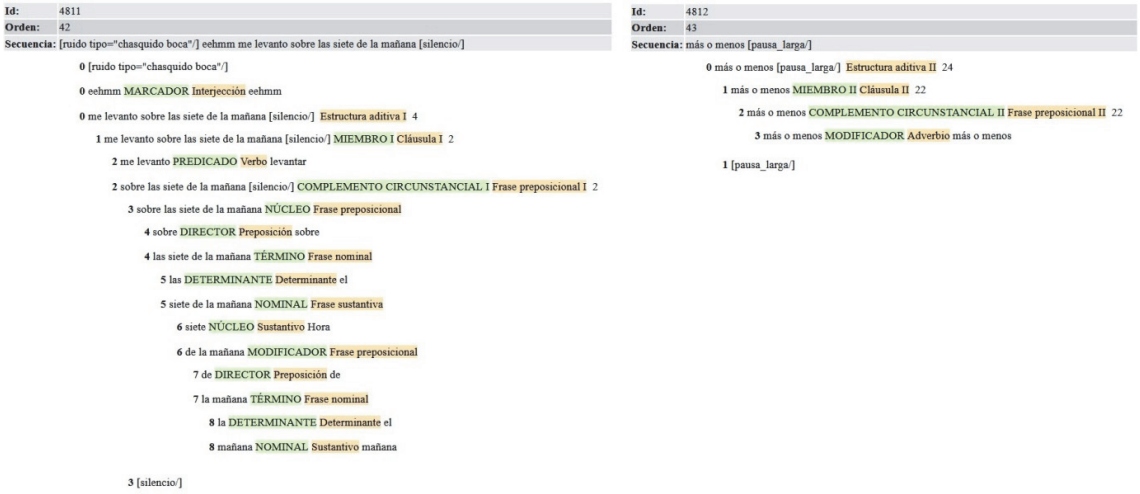

Figura 3. Muestra de funciones y unidades partidas en ESLORA-CSA

Esta marcación se lleva a cabo de modo que eliminando los I y el dígito subsiguiente de la primera referencia a una unidad fragmentada, así como completos todos los niveles de análisis en los que aparezcan funciones y unidades que acaben en II, obtengamos las unidades fragmentadas como las habríamos analizado de no estarlo. Con la ventaja de que se mantiene la congruencia entre la segmentación estrictamente prosódica y la sintáctica. La marcación diseñada, con todo, se considera provisional, ligada a la evolución de la herramienta de edición. Al igual que es provisional la ausencia de información sobre qué hablante emite cada fragmento o la carencia de marcas que señalen los solapamientos, marcajes estos últimos que sí se codifican en la transcripción y de cuya información dispondremos cuando contemos con una herramienta de análisis que nos permita más funcionalidades.

4. MARCAS DE ORALIDAD

Las marcas de oralidad (introducidas en la transcripción como etiquetas XML) contienen información sobre hechos no verbales que suceden durante el discurso transcrito o lo caracterizan de algún modo. Evidentemente estas marcas deben subsistir en su lugar en cualquier versión lingüísticamente anotada del corpus (como sucede en ESLORA y en otros corpus orales sintácticamente anotados que hemos revisado), aunque obviamente no tienen que recibir anotación de ninguna clase.

Para facilitar la integración de ESLORA en la herramienta de análisis PaME, fue necesario en primer lugar transformar las comillas angulares (i.e. <ininteligible/>, <pausa/>, etc.), presentes en las etiquetas XML de los fragmentos prosódicos transcritos de la entrevista semidirigida elegida, en corchetes (i.e. [ininteligible/], [pausa/], etc.). Cuando el análisis sintáctico se haya completado, en una etapa de posprocesado ulterior, se procederá nuevamente a la reconversión a formato XML de las marcas ahora visibles con corchetes.

Por otra parte y también debido a la obsolescencia de la herramienta de análisis, se hizo asimismo necesario prescindir, por el momento, de la información sobre qué hablante emite cada fragmento, así como de la que concierne a los solapamientos.

Respecto a las demás marcas de oralidad que se utilizan en la trascripción y a cómo las tratamos en el nivel de análisis sintáctico, debemos distinguir las marcas simples de las dobles. A las simples – ininteligible, algunos tipos de ruido, pausa, pausa larga, silencio, risa, etc. – se les aplica el mismo criterio que veremos (sección 5) que se les aplica a conectores, marcadores y demás elementos que no consideramos sintácticos: se sitúan donde aparecen en el nivel más alto en el que se los puede colocar de modo que por su causa no se rompa ningún segmento identificable como un segmento sintáctico. De este modo, el *chasquido de boca* o *ehmm,* un MARCADOR, de la figura 3 aparecen en el nivel 0, precediendo al segmento sintáctico que se identifica a continuación, mientras que la etiqueta de *silencio* se sitúa en el nivel 3, porque ese es el nivel en el que se sitúa el MODIFICADOR *más o menos* que se encuentra presente en el segmento prosódico siguiente, el 43, y que completa la frase preposicional *sobre las siete de la mañana más o menos* iniciada en el fragmento 42. Situar *[silencio/]* en otro nivel rompería esa frase preposicional.

En el caso de las marcas con inicio y cierre – palabra cortada, énfasis, ficticio, ruido de fondo, cita, risa, etc. – , si la etiqueta abarca solo a una palabra o un constituyente, ubicamos tanto su apertura como su cierre en el segmento al que abracen, como puede observarse en la marca *ficticio* de la figura 5. Si por el contrario la marca de oral afecta a más de un constituyente, la etiqueta de apertura se sitúa junto a la primera palabra abrazada y la de cierre junto a la última. Así, en la figura 4, donde seleccionamos un segmento de un análisis que contiene una marca de oral doble, la parte del inicio de *risa* abraza al 'Predicado' *vivía* y la de cierre al 'Complemento preposicional' *allí.*

4 cuando yo [risa_inicio/]vivía allí [risa_fin/] COMPLEMENTO CIRCUNSTANCIAL Cláusula

5 cuando COMPLEMENTO CIRCUNSTANCIAL Adverbio cuando

5 yo SUJETO Pronombre yo

5 [risa_inicio/]vivía PREDICADO Verbo vivir

5 allí [risa_fin/] COMPLEMENTO PREPOSICIONAL Adverbio allí

Figura 4. Muestra de empleo de la marca de oralidad doble en ESLORA-CSA

5. PUNTUACIÓN, VOCATIVOS, CONECTORES, MARCADORES Y TÓPICOS

Los reunidos bajo este epígrafe, al contrario que los anteriores, son elementos verbales no exclusivos de los corpus orales, porque, aunque en la lengua escrita con distinta frecuencia, también están presentes en ella. No son, sin embargo, elementos sintácticos, sino de carácter discursivo, que, a pesar de ello, hay que integrar en la anotación sintáctica. La postura más radical con respecto a esta integración de los corpus que hemos revisado la representa ESLORA: todos estos elementos son tratados exactamente igual que las marcas de oral vistas en la sección 4, sin ligarlos, por tanto, de ningún modo a una unidad sintáctica con la que se pueda entender que configuran una unidad discursiva – no estrictamente sintáctica – oral, la que podría

considerarse la unidad de referencia para el estudio del habla.[12] Otros corpus orales proceden de modos distintos para integrar en su anotación sintáctica estos elementos, que en ellos sí quedan ubicados con respecto a las unidades estrictamente sintácticas analizadas, constituyendo con ellas unidades más amplias que podrían ser consideradas las unidades de referencia para el estudio del discurso oral: el CGN, por ejemplo, tiene una unidad discursiva (DU) en el interior de la que algunos de los elementos de esta clase constituyen la función *satélite*, otros corpus (C-ORAL, Swichboard), aunque no tienen una unidad discursiva, no al menos tan explícitamente declarada como tal, se esfuerzan por determinar junto a qué segmento sintáctico se tienen que analizar los elementos en cuestión. A continuación desarrollamos en más detalle cómo sucede lo aquí planteado de modo general.

Con respecto a ESLORA, es preciso comenzar aclarando que ninguno de los términos distintos de puntuación que se recogen en el encabezado de esta sección se marcan en el nivel de la transcripción. Es en la fase de análisis sintáctico donde vamos a identificarlos y a tratarlos según indicamos. Dado que entendemos que puntuación,[13] conectores, marcadores, vocativos y (algunos) tópicos son ajenos al segmento sintáctico que queremos analizar, a fin de eludir la toma de decisiones sobre con qué segmentos del discurso guardan estos elementos una relación más estrecha, configurando junto a ellos las que deben ser unidades de referencia del habla, todos estos elementos se ubican en el análisis siguiendo el mismo principio puramente formal con el que se ubicaban las marcas de oral: considerando el 0 el nivel de análisis más alto, se los sitúa en el nivel de análisis más alto en el que se los puede colocar de forma que no rompan ningún constituyente sintáctico.

A los vocativos se les asigna la etiqueta de función VOCATIVO, se indica la clase de palabra que los desempeña y se señala su lema. Frente a los 22 casos que se registran en la obra analizada en CSA, *Crónica de una muerte anunciada*, desempeñados principalmente por sustantivos (– Cristóbal – *gritó – :*), pero también por adjetivos (– *Suéltala,* blanco – *le ordenó en serio* – .), en la muestra analizada de corpus oral solo encontramos 2 casos, figura 5, ambos además caracterizados con la marca *ficticio,* ya que para proteger la identidad de los hablantes se procedió a la anonimización del audio y de la transcripción (Vázquez Rozas et al., 2020).

Secuencia: ah [pausa/] [ficticio]Marcos[/ficticio] [pausa/] [ficticio]Ana[/ficticio] [pausa/]

0 ah MARCADOR Interjección ah

0 [pausa/]

0 [ficticio]Marcos[/ficticio] VOCATIVO Sustantivo Marcos

0 [pausa/]

0 [ficticio]Ana[/ficticio] VOCATIVO Sustantivo Ana

0 [pausa/]

Figura 5. Ejemplo de vocativo extractado de ESLORA-CSA

[12] En este sentido ESLORA se aparta también de CSA y del sistema de anotación de lengua escrita que de aquel hereda y que adapta. En CSA, efectivamente, la segmentación basada en la puntuación ubica estos elementos con respecto a lo estrictamente sintáctico.

[13] Los signos de apertura y cierre de admiración e interrogación, que son los únicos signos de puntuación que se utilizan en ESLORA.

También los corpus con los que contrastamos ESLORA-CSA etiquetan vocativos. Así, el Corpus C-ORAL les asigna la etiqueta @VOK (eu já vi ele, *Carlos Henrique*), mientras que en el corpus de conversaciones telefónicas, siguiendo el manual del Treebank, se les asigna la etiqueta -VOC y no se coindexan con el sujeto explícito (*Mike,* would you please close the door?).[14]

Por lo que respecta a los marcadores, la presencia de estos es abrumadora en ESLORA-CSA como demuestra el contraste entre los 20 casos recogidos en la *Crónica* (1 adverbio [*mejor*], 2 frases interjectivas [*gracias por todo*; *ojalá te maten*], 2 frases sustantivas [*feliz año nuevo*; *hijos de puta*], 15 interjecciones [*ah, anda, ave María Purísima, ay, carajo* (4)*, collons de déu, Dios Santo, Dios mío, mierda* (2), *por el amor de Dios, Santo Dios*]) frente a los 388 anotados en los fragmentos analizados de SCOM_M32_025: *eeh* (102), *bueno* (68), *mmm* (51), *claro* (44), *eh* (32), *mm* (22), *eehmm* 14, *hola* (7), *hm* (7), *vale* (4), *aha* (4), *vamos* (3), *pff* (3), *hombre* (3), *qué tal* (2), *mira* (2), *bah* (2), *ah* (2), *mhm* (2), *hm hm* (2) y con 1 ocurrencia *qué hay, mejor, la verdad, jo, hmm, fff, encantada, ehm, clac, buf, ay, adiós, aah y aaah*.

Categorialmente los marcadores están desempeñados en su mayoría por interjecciones, si bien aparecen también realizados por adverbios (*claro, mejor, sí*), frase nominal (*la verdad*) o adjetivo (*encantada*).

En lo tocante a los conectores, la diferencia entre el registro escrito y el oral no es tan acusada: 183 casos en la *Crónica* – *pero* (54), *sin embargo* (28), *en cambio* (11), *de modo que* (9), *en realidad* y *en efecto* (8); *y* (7); *por su parte, al contrario* y *además*, (6) apariciones cada uno; *más aún* y *así que*, (4); *entonces, en todo caso* y *de todos modos*, (3); *sobre todo, pues, por último, por el contrario* y *no obstante*, (2); *que, primero, por supuesto que, por supuesto, más bien, es que, es decir, en fin, en efecto, después, debió ser que, debe ser que, de manera que*, (1) – mientras que en ESLORA se recogen 413 casos en esos fragmentos iniciales con la siguiente distribución: *y* (110), *pues* (51), *sí* (45), *o sea* (34), *no* (32), *entonces* (33), *así* (22), *tal* (15), *pero* (13); *que* y *por ejemplo* (8); *porque* (7), *ya* (6), *además* (5), *muy bien* (4), *yo qué sé* (3), *efectivamente* y *después* (2); *verdad*, *un decir, si, sabes, o sea que, más o menos*, *eso, es que, en fin, de todas maneras, de hecho, bien, aparte* (1).

El corpus Switchboard codifica ya en el nivel de transcripción ortográfica los elementos cuasiléxicos, los marcadores discursivos e incluso las conjunciones coordinantes. Para ello recogen entre llaves la marca asociada a cada tipo y el término concreto de que se trate, con las siguientes correspondencias:

Filler: {F *uh*}, *um, huh, oh*, etc.
Explicit editing term: {E *I mean*}, *sorry, excuse me*, etc.
Discourse marker: {D *you know*}, *well*, *so*, *like*, etc.
Coordinating conjunction: {C *and*}, *and then, but, because*, etc.

En el análisis sintáctico, como se aprecia en la figura 2 para *well,* los elementos paralingüísticos se analizan al mismo nivel que los constituyentes entre los que aparecen y se caracterizan en general con la función INTJ (*Interjection*), aplicada también a otras formas como *yes, okay, uh, please, no, alas*, etc., aunque algunos conectores como *you know* son analizados asignándoles un nodo específico denominado PRN, que se reserva en el Penn Treebank para los incisos.

[14] Contrástese este hecho con la coindización que sí realizan con el predicado para los tópicos, lo cual incide en el carácter más ajeno a la sintaxis de los vocativos en comparación con los tópicos.

Frente al Penn Treebank, en el corpus C-ORAL los cuasiléxicos no se codifican en el nivel de transcripción. Se detectan en una fase de preprocesado mediante la consulta al lexicón y se les asigna la etiqueta 'non-word'. En el nivel de anotación sintáctica la etiqueta 'non-word', junto con el elemento paralingüístico, se transforma en metaetiqueta, de modo que se mantienen en el nivel de análisis sintáctico, pero no se les proporciona ningún tipo de análisis.

Por último, los corpus con los que contrastamos ESLORA-CSA también etiquetan tópicos. En el CGN, los tópicos son por excelencia los satélites a los que nos referimos más arriba. En el Corpus C-ORAL, cuando se topicaliza un constituyente, se asigna la etiqueta sintáctica @TOP (*Esse negócio*, não gosto dele), mientras que en el corpus de conversaciones telefónicas, siguiendo el manual del Treebank, los tópicos se anotan con la etiqueta -TPC y se coindizan con el predicado (*Marching past the reviewing stand* were musicians). También el CSJ habilita una manera de anotar las topicalizaciones. Nosotros, sin embargo, una etiquetación para esta circunstancia no la habíamos considerado hasta que en las muestras de lengua oral nos topamos con varias tematizaciones que por su forma eran sintácticamente no integrables en la secuencia en la que aparecen, como es el caso de *Santiago* en el fragmento *pero bueno* Santiago *[pausa/] bah [pausa/] va habiendo así una [alargamiento] ofertita [/alargamiento] [ruido tipo="palmada"/] [silencio/].*

Esta es una zona todavía en experimentación en ESLORA-CSA. A este respecto, estamos considerando si anotar con la función que correspondería en la secuencia y una marca que señale la topicalización sintácticamente no integrable o si analizar como 'TÓPICO', con una función más netamente discursiva. A fin de evidenciar y hacer recuperable la estructura argumental del verbo, provisionalmente nos hemos decantado por la primera opción. Así, por ahora, el *Santiago* anterior está etiquetado como complemento circunstancial y el índice 77, que caracteriza la tematización cuando la forma, además de la posición y la prosodia, con la pausa, es inadecuada a la función que tendría el elemento en la secuencia.

6. DISFLUENCIAS: REPETICIONES, REFORMULACIONES Y TRUNCAMIENTOS

Por fin, propio de la lengua oral es aquello a lo que da lugar el hacerse de la misma mientras sucede, su planificación repensada y alterada durante el discurso mismo, fenómenos cuya integración en la anotación sintáctica de un corpus oral es el objeto de esta sección.

De las disfluencias características del habla, en la fase de transcripción ortográfica del corpus ESLORA solo se codifican los truncamientos a nivel de palabra con la etiqueta <palabra cortada></palabra cortada>, envolviendo al segmento de que se trate. Ni los truncamientos que afectan a unidades superiores a la palabra, ni las repeticiones o reformulaciones se marcan de modo alguno en la transcripción. Frente a este proceder, en otros corpus se codifican ya en el nivel de transcripción los distintos tipos de disfluencias, que después, sin embargo, no siempre son igualmente tratadas por todos ellos en la anotación sintáctica: en algunos casos son obviadas en el análisis, en otros son también objeto de él. Entre los que proceden del primer modo, se encuentra el C-ORAL, que, manteniéndolas como metaetiquetas, elimina del análisis superficial los distintos tipos de retracciones, premarcadas manualmente entre corchetes con señalamiento de su punto de inicio. Por ejemplo, en la unidad "<eu &a [/2] eu acho que é> esse [/2] é esse aquí o' // <&he>", el preprocesador transforma el falso inicio (*eu &a [/2]*) y la repetición (*é> esse [/2]*),

respectivamente, en las metaetiquetas <retract:eu_&a> y <retract:e>_esse>, que permanecen sin tratar en el análisis sintáctico. De manera semejante procede el CGN, que también deja fuera del grafo de análisis sintáctico correcciones o retracciones.

Frente a ellos, la labor llevada a cabo en el Switchboard destaca en este sentido, sobre todo si tenemos en cuenta que el proyecto data de la década de los 90 del siglo pasado. En él, las disfluencias anotadas en la fase de transcripción entre corchetes, que señalan además el punto en el que se reinicia el discurso, se mantienen y se analizan en la fase de análisis sintáctico, como se observa en la repetición del pronombre *I* en la figura 2 recogida más arriba. Como en el Switchboard, también en el CSJ las disfluencias presentes en el corpus tienen un hueco en la representación dependencial que constituye la anotación sintáctica del corpus.

En ESLORA-CSA hemos creado una etiqueta de unidad, denominada 'Tentativa idéntica', con la que identificar las repeticiones. Estas se analizan al nivel de lo repetido y se caracterizan con 'Tentativa idéntica', no le asignamos función y la desglosamos en tantos elementos constitutivos como la integren, dando cuenta del tipo de palabra y lema de que se trate en cada caso. Recogemos una muestra en la figura 6, donde la 'Tentativa idéntica' está conformada por la preposición *en*:

Figura 6. Ejemplo de *Tentativa idéntica* tomado de ESLORA-CSA

En relación con los truncamientos, sea a nivel de palabra, frase, cláusula o estructura, y reformulaciones, de modo provisional, analizamos sintácticamente unos y otras y les asignamos el índice 99 y 88, respectivamente, a fin de diferenciarlos y poder recuperarlos más tarde. Así, si un hablante empieza a hablar, se para dejando lo dicho inacabado, y lo retoma de otra forma, entendemos que estamos ante una reformulación que debe ser analizada y marcada específicamente para permitir su estudio. Por ello, en la unidad sintáctica conformada por la parte sin tachar del fragmento 77, junto con los fragmentos 78, 79, 80 y 81 recogidos a continuación, entendemos que la ANTÍTESIS se reformula, primero, en una cláusula truncada (*[alargamiento]y[/alargamiento] eso que bueno [pausa/] eeh [pausa/] tenía así algunas cositas también [pausa/] de*), por lo que se le asigna el índice 89 que marca ambos fenómenos, reformulación y truncamiento y, a continuación, en una cláusula

partida en dos fragmentos prosódicos (la parte final del 79: *yo recuerdo así de cuando era joven [pausa_larga/],* y todo el fragmento 80), que se caracteriza con el índice 88 que señala la reformulación.

77	~~no [palabra_cortada]ta[/palabra_cortada] tampoco nos podemos quejar [risa_inicio/]mucho [risa_fin/] aquí hay [pausa/] hay otros sitios que están peor [pausa/]~~ yo soy de Lugo [pausa_larga/]
78	y en Lugo la cosa era peor cuando yo [risa_inicio/]vivía allí [risa_fin/] [pausa_larga/] [alargamiento]y[/alargamiento] eso que bueno [pausa/] eeh [pausa/] tenía así
79	algunas cositas también [pausa/] de yo recuerdo así de cuando era joven [pausa_larga/]
80	eeh chavalita [pausa/] vamos [pausa/] ¿no? que [pausa/] cuando yo estaba en Lugo que bueno [pausa/] también había [pausa/] aprovechabas [pausa/] eh cosas ¿no? [pausa_larga/]
81	pero bueno Santiago [pausa/] bah [pausa/] va habiendo así una [alargamiento]ofertita[/alargamiento] [ruido tipo="palmada"/] [silencio/]

La figura 7, que muestra la herramienta de análisis PaME en la que hemos contraído los nodos inferiores al nivel 2 para facilitar la comprensión, refleja el análisis que acabamos de explicar:

Figura 7. Pantalla de introducción de datos en el análisis sintáctico mediante PaME

7. CONCLUSIONES

En este artículo se ha descrito la primera fase de trabajo para la obtención de un sistema de anotación para lengua oral a partir de uno preexistente para lengua escrita. Se ha descrito el proceso de anotación y los principios que lo sustentan, heredados del sistema de anotación previo para lengua escrita que constituye su punto de partida. Se han descrito asimismo las adaptaciones que ese sistema previo ha tenido que sufrir para ser aplicado a lengua oral. Se ha visto que estas conciernen no a la necesidad de dar cuenta de una estructura sintáctica particular de la lengua oral, no al menos en términos globales, sino a lo que es específico de esta por cómo la transcribimos en los corpus que la consignan, sin signos de puntuación y con marcas de oralidad, y por su hacerse mientras se realiza. Por lo primero, la identificación de las unidades de análisis debe llevarse a cabo de manera completamente diferente a

como se lleva a cabo en lengua escrita. Por lo segundo, hay que habilitar modos de marcar y tratar las llamadas disfluencias de la lengua oral: repeticiones, truncamientos y reformulaciones. En este artículo se ha descrito cómo se ha pensado por ahora resolver ambas cuestiones en ESLORA.

REFERENCIAS

AnCora. <http://clic.ub.edu/corpus/es> (acceso: 22/05/2021)

AT. *Arabic Treebank – Broadcast News.* <https://catalog.ldc.upenn.edu/LDC2012T077> (acceso: 09/06/2021)

Bick, E. 2012. *Grammatical annotation of the Portuguese C-ORAL Corpus*. <https://visl.sdu.dk/pdf/coral_palavras.pdf> (acceso: 22/05/2021)

Bick, E., Mello, H., Panunzi, A. y Raso, T. 2013. The annotation of the C-ORAL-BRASIL spoken corpus using an adaptation of the Palavras Parser. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12).* <http://www.lrec-conf.org/proceedings/lrec2012/pdf/586_Paper.pdf> (acceso: 22/05/2021)

C-ORAL-BRASIL. <http://www.c-oral-brasil.org/> Consultable en <https://www.linguateca.pt/acesso/corpus.php?corpus=CORALBRASIL> (acceso: 09/06/2021)

C-ORAL-ROM. <http://www.elda.org/en/proj/coralrom.html > (acceso: 20/05/2021)

CGN. Corpus Gesproken Nederlands. <https://ivdnt.org/images/stories/producten/documentatie/cgn_website/doc_English/topics/index.htm> (acceso: 26/05/2021)

CORPES. *Corpus del español del siglo XXI.* Real Academia Española. <http://rae.es/recursos/banco-de-datos/corpes-xxi> (acceso: 20/05/2021)

COSER. *Corpus oral y sonoro del español rural.* <http://www.corpusrural.es/> (acceso: 20/05/2021)

CREA. *Corpus de referencia del español actual.* Real Academia Española. <https://www.rae.es/banco-de-datos/crea> (acceso: 20/05/2021)

CSA. <http://galvan.usc.es/drasae/> (acceso: 06/06/2021)

CSJ. *Corpus of Spontaneous Japanese.* <https://ccd.ninjal.ac.jp/csj/en/> (acceso: 9/06/2021)

DRASAE. <https://gramatica.usc.es/wiki/drasae> (acceso: 22/05/2021)

ESLORA. *Corpus para el estudio del español oral.* <http://eslora.usc.es> (acceso: 14/05/2021)

IULA Spanish LSP Treebank. <http://www.iula.upf.edu/recurs01_tbk_uk.htm> (acceso: 20/05/2021)

Lope Blanch, J. M. 1986. *El estudio del español hablado culto. Historia de un proyecto.* México: UNAM.

Maamouri, M., Bies, A., Kulick, S., Zaghouani, W., Graff, D. y Ciul, M. 2010. From speech to trees: Applying treebank annotation to Arabic broadcast news. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).* <http://www.lrec-conf.org/proceedings/lrec2010/pdf/558_Paper.pdf> (acceso: 26/05/2021)

Meteer, M. et al. 1995. *The Penn Treebank Project: Dysfluency annotation stylebook for the Switchboard Corpus.* <https://catalog.ldc.upenn.edu/docs/LDC99T42/dflguide.pdf> (acceso: 26/05/2021)

Müürisep, K. y Uibo, H. 2006. Shallow parsing of spoken Stonian using Constraint grammar. En P. J. Henrichsen y P. R. Skadhauge eds. *Treebanking for discourse and speech. Proceedings of NODALIDA 2005. Special session on treebanks for spoken language and discourse. Copenhagen Studies in Language* 32. Samfundslitteratur.

Oostdijk, N. 2000. The Spoken Dutch Corpus. Overview and first evaluation. *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00).* <http://www.lrec-conf.org/proceedings/lrec2000/pdf/110.pdf> (acceso: 22/05/2021)

Oostdijk, N., Goedertier, W., van Eynde, F., Boves, L., Martens, J. P., Moortgat, M. y Baayen, H. 2002. Experiences from the Dutch Corpus project. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02).* <http://www.lrec-conf.org/proceedings/lrec2002/pdf/98.pdf> (acceso: 22/05/2021)

PRESEEA. *Proyecto para el estudio sociolingüístico del español de España y América.* <https://preseea.linguas.net/> (acceso: 20/05/2021)

Real Academia Española. 2009. *Nueva gramática de la lengua española.* Madrid: Espasa-Calpe.

Schuurman, I., Goedertier, W., Hoekstra, H., Oostdijk, N., Piepenbrock, R. y Schouppe, M. 2004. Linguistic annotation of the Spoken Dutch Corpus: If we had to do it all over again. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04).* <http://www.lrec-conf.org/proceedings/lrec2004/pdf/437.pdf> (acceso: 22/05/2021)

Switchboard. Switchboard-1 Telephone Speech Corpus. <https://catalog.ldc.upenn.edu/LDC97S62> (acceso: 09/06/2021)

Taylor, A., Marcus, M. y Santorini, B. 2003. *The Penn Treebank: An overview* <https://www.researchgate.net/publication/2873803_The_Penn_Treebank_An_overview> (acceso: 26/05/2021)

TUCSE. Tartu University Corpus of Spoken Estonian. <http://kodu.ut.ee/~kaili/Korpus/puud> (acceso: 21/05/2021)

UAM Spanish Treebank. <http://www.lllf.uam.es/ESP/Treebank.html> (acceso: 20/05/2021)

Uchimoto, K., Hamabe, R., Mauyama, T., Takanashi, K., Kawahara, T. y Isahara, H. 2006. Dependency-structure annotation to Corpus of Spontaneous Japanese. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'05).* <http://www.lrec-conf.org/proceedings/lrec2006/pdf/298_pdf.pdf> (acceso 26/5/2021)

VAL-ES-CO. *Corpus Val.Es.Co 2.1* <http://www.valesco.es/corpus> (acceso: 20/05/2021)

Van Der Wouden, T., Hoekstra, H., Moortgat, M., Renmans, B. y Schuurman, I. 2002. Syntactic analysis in the Spoken Dutch Corpus (CGN). *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02).* <http://www.lrec-conf.org/proceedings/lrec2002/pdf/71.pdf> (acceso: 22/05/2021)

Vázquez Rozas, V., Barcala, M., Domínguez Noya, E., Fernández Sanmartín, A., Rojo, G. y Santalla, M.ª P. 2020. Codificación y anotación del habla en un contexto bilingüe: El corpus ESLORA de español de Galicia. En A. J. Gallego y F. Roca Urgell eds. *Dialectología digital del español. Anexo Verba* 80: 189–224.