TEISEL
Tecnologías para la investigación en segundas lenguas
ISSN 2696-676X
Universitat de Barcelona
Facultat de Filologia i Comunicació

Herramientas para la codificación, el análisis y la explotación de un corpus oral de aprendices francófonos de español

Cristina MUÑOZ DE LA VIRGEN
Institute for American Universities - American College of the Mediterranean
cristina.munoz@iau.edu
https://orcid.org/0000-0001-8816-2992

Resumen: En este trabajo se analiza el uso de tres herramientas digitales para la codificación, el análisis y la explotación de un corpus oral de aprendices de L2. En primer lugar, se presenta la creación, codificación y actualización del corpus, que se compiló con el objetivo de analizar los errores cometidos en el plano oral por parte de un grupo compuesto por 51 informantes franceses, estudiantes de segundo año de Lenguas Extranjeras Aplicadas de la Universidad François Rabelais de Tours (Francia). Una vez conseguido el material auditivo, se procedió a la transcripción y anotación textual codificada, siguiendo la normativa que marca *Text Encoding Initiative*.

Esta investigación se valió de ciertas herramientas digitales, como *XML Formatter* (https://jsonformatter.org/xml-formatter), *XML Validator* (https://codebeautify.org/xmlvalidator), *VS Studio* (https://code.visualstudio.com/), *Git* (https://git-scm.com/), *GitHub* (https://github.com/), *SpaCy* (https://SpaCy.io/), *CATMA* (https://catma.de/) y *Google Colaboratory* (https://colab.google/), que contribuyeron al desarrollo del análisis y la explotación del corpus oral, sirviendo para la configuración y validación de la cabecera del corpus (*XML Formatter - Validator*), para el etiquetado del corpus (*VS Studio*), para el control y almacenaje de las versiones (*Git-GitHub*) y para la actualización del corpus (*SpaCy* permite nuevos análisis, *Google Colaboratory* es un instalador de software usado junto a *SpaCy*, y con *CATMA* se desarrolla el nuevo etiquetado).

Palabras clave: lingüística de corpus; corpus de aprendices; codificación; análisis de datos; L2

Català:

Eines per al codificació, l'anàlisi i l'explotació d'un corpus oral d'aprenents fracnòfons d'espanyol

***Resum**: En aquest treball es presenta la creació, codificació i explotació d'un corpus oral d'aprenents francòfons d''espanyol, destinat a analizar de manera sistemàtica els errors comesos en el pla oral per part d''un grup determinat d'aprenents francòfons d'Espanyol com a Llengua Estrangera. En les línies següents, es detalla l'elaboració d'aquest corpus oral, que va partir de dos qüestionaris enregistrats al llarg d'un semestre i compost per 51 informants francesos, estudiants de segon any de Llengües Estrangeres Aplicades de la Universitat François Rabelais de Tours (França).*

Aquesta investigació es va valer de certes eines digitals, com XML Formatter (https://jsonformatter.org/xml-formatter), XML Validator (https://codebeautify.org/xmlvalidator), VS Studio (https://code.visualstudio.com/), Git (https://git-scm.com/), GitHub (https://github.com/), SpaCy (https://SpaCy.io/), CATMA (https://catma.de/) y Google Colaboratory (https://colab.google/), que van contribuir al desenvolupament de l'anàlisi i l'explotació del corpus oral. servint per a la configuració i validació de la capçalera del corpus (XML Formatter - Validator), per a l'etiquetatge del corpus (VS Studio), per al control i l'emmagatzematge de les versions (Git -GitHub) i per a l'actualització del corpus (SpaCy permet noves anàlisis, Google Colaboratory és un instal·lador de programari usat juntament amb SpaCy i amb CATMA es desenvolupa el nou etiquetatge).

***Paraules clau**: lingüística de corpus; corpus d'aprenents; codificació; anàlisi de dades; L2*

Muñoz de la Virgen, Cristina (2024). Herramientas para la codificación, el análisis y la explotación de un corpus oral de aprendices francófonos de español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_inv03, 1-25. http://doi.org/10.1344/teisel.v3.44741

© La autora | Recepción: 06/11/2023 | Revisión: 18/01/2024 | Aceptación: 09/02/2024 | Publicación: 11/03/2024

ISSN 2696-676X

English:

Coding of an oral corpus of French-speaking learners of Spanish

***Abstract**: This paper presents the creation, coding and exploitation of an oral corpus of French-speaking learners of Spanish, aimed at systematically analyzing the oral errors made by a given group of French-speaking learners of Spanish as a Foreign Language. The following lines detail the elaboration of this oral corpus, which was based on two questionnaires recorded over a semester and composed of 51 French informants, second year students of Applied Foreign Languages at the François Rabelais University (France). Once the auditory material was obtained, we proceeded to transcribe and coded textual annotation, following the standards set by the Text Encoding Initiative.*

This research made use of certain digital tools, such as XML Formatter (https://jsonformatter.org/xml-formatter), XML Validator (https://codebeautify.org/xmlvalidator), VS Studio (https://code.visualstudio.com/), Git (https://git-scm.com/), GitHub (https://github.com/), SpaCy (https://SpaCy.io/), CATMA (https://catma.de/) y Google Colaboratory (https://colab.google/), which contributed to the development of the analysis and exploitation of the oral corpus. serving for the configuration and validation of the corpus header (XML Formatter - Validator), for corpus tagging (VS Studio), for version control and storage (Git-GitHub) and for updating the corpus (SpaCy allows new analyses, Google Colaboratory is a software installer used together with SpaCy and with CATMA the new tagging is developed).

***Keywords**: corpus linguistics; learner corpus; encoding; data analysis; 2L*

Introducción

En el actual contexto de enseñanza mediada por tecnologías, cabe plantearse cómo se desarrolla el aprendizaje de lenguas extranjeras; para ello, los docentes pueden recurrir a un antiguo modelo renovado: el Análisis de Errores Asistido por Computador (*Computer-aided Error-Analysis*, CEA). De acuerdo con este modelo, a lo largo de este trabajo, se presenta la creación, la codificación y el desarrollo de un corpus oral de aprendices francófonos de español como lengua extranjera (ELE), con especial atención a cómo las herramientas digitales pueden ayudar en el análisis y la explotación de los datos (apartado 5).

Siguiendo a Llisterri (2021), es conveniente distinguir entre corpus oral, que cuenta con algún tipo de anotación textual y con la transcripción fonética acompañada del extracto de la grabación, y corpus de la lengua oral, el cual es una transcripción ortográfica enriquecida con aspectos de la lengua oral. El caso que aquí se presenta es un corpus oral de aprendices francófonos de ELE[1], que se configuró mediante una serie de versiones que marcaron su paso del corpus de la lengua oral, en su primera versión, a un corpus oral, anotado y enriquecido con la señal sonora, en su cuarta versión, que se comentarán más adelante.

Para obtener el resultado actual del corpus, se partió de una base de datos digitalizada donde se identificaron y clasificaron los errores etiquetados (anotados), según el criterio específico de la investigación, el cual consistía en el análisis de la interlengua de los informantes en relación con la adquisición del modo subjuntivo, es decir, el estadio de adquisición lingüística en el que se encuentra el sistema verbal de los integrantes en el estudio en un momento determinado del proceso de aprendizaje. En el proceso de etiquetado, se utilizó la anotación descrita en la Iniciativa de

[1] Este corpus está disponible en *Docta UCM* (https://docta.ucm.es/entities/publication/fe061490-2701-41aa-9902-400fd8755e69). Asimismo, el lector podrá escuchar los ejemplos aportados en el documento que acompaña al corpus. Tras la actualización del corpus que se está realizando, se podrá acceder a la producción íntegra de los informantes.

Muñoz de la Virgen, Cristina (2024). Herramientas para la codificación, el análisis y la explotación de un corpus oral de aprendices francófonos de español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_inv03, 1-25. http://doi.org/10.1344/teisel.v3.44741

© La autora | Recepción: 06/11/2023 | Revisión: 18/01/2024 | Aceptación: 09/02/2024 | Publicación: 11/03/2024

ISSN 2696-676X

Text Encoding Initiative Consortium, 2023 (TEI), pero, al mismo tiempo, este sistema se enriqueció con una serie de etiquetas de creación propia, las cuales permitieron un mejor análisis de los objetivos de la investigación. Aun así, el corpus se diseñó sobrepasando las limitaciones de la investigación inicial mediante la creación de etiquetas de carácter abierto, ya que no se limitaron al análisis del modo subjuntivo, sino que se creó una anotación que permitiera personalizar las etiquetas con atributos necesarios para otros usos, como se verá en el apartado 4. Todas las anotaciones realizadas en el corpus son extratextuales, lo cual permite la rápida y fácil modificación o eliminación sin que se altere el resto de los elementos. Así, con este etiquetado, se pudieron llevar a cabo modestas investigaciones sobre otros aspectos, como el análisis de los tiempos de pasado.

En el apartado 2 se explica el contexto previo a la creación del corpus y se ofrece información relevante sobre los informantes y las pruebas. En el apartado 3 se describe el proceso de etiquetado de la cabeza del corpus, para el cual, como se verá, se optó por insertar manualmente las etiquetas, con la ayuda de un editor de XML. En el apartado 4 se describen las etiquetas que se usaron y se crearon para la anotación del corpus, mediante VS Studio. Finalmente, en el apartado 5 se describen las diferentes herramientas digitales que se emplearon a lo largo de la creación y del desarrollo de este corpus oral y se analiza la situación actual en la que se encuentra.

1. Aspectos claves en los corpus orales de aprendices de ELE

El corpus oral de aprendices francófonos que aquí se presentan se configuró siguiendo la estela del Corpus de Aprendices de Español (CAES, https://galvan.usc.es/caes/, Instituto Cervantes, 2014). Este corpus está compuesto por textos escritos procedentes de una gran variedad de aprendices, con estadios de aprendizaje y dominio de la lengua diferentes. Aunque es una valiosa herramienta de análisis, se prefirió confeccionar un corpus más específico (aprendices francófonos de nivel B2, según el *Marco común europeo de referencia para las lenguas*) y dotarlo de un tratamiento oral, es decir, está compuesto por 102 producciones orales relativamente espontáneas de los aprendices, como se verá en los sucesivos epígrafes.

De acuerdo con McEnery y Wilson (2001), Sinclair (2005), Lozano y Mendikoetxea (2013) y Lozano (2022), el corpus oral ha seguido unos parámetros determinados para la codificación y recuperación de los datos, como se verá en el apartado 3, donde se describe la configuración de la cabecera de cada producción de los informantes, y en el apartado 4, donde se detalla el tipo de etiquetas que se emplearon en la anotación del corpus. Estos parámetros permiten el análisis de frecuencia del modo subjuntivo en el plano oral con muestras de uso real del español por parte de los informantes, así como la recuperación de los datos, como se detallarán en el apartado 5. También, se pensó en la representatividad y se analizó la producción oral de todos los estudiantes de segundo año del grado de Lenguas Extranjeras Aplicadas de la Universidad François Rabelais de Tours (Francia), para que los datos fueran extrapolables a todos los estudiantes franceses del mismo grado y año de estudio que se hayan formado en el sistema público francés.

Muñoz de la Virgen, Cristina (2024). Herramientas para la codificación, el análisis y la explotación de un corpus oral de aprendices francófonos de español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_inv03, 1-25. http://doi.org/10.1344/teisel.v3.44741

© La autora | Recepción: 06/11/2023 | Revisión: 18/01/2024 | Aceptación: 09/02/2024 | Publicación: 11/03/2024

ISSN 2696-676X

Aunque no cuenta con un grupo de control, se tiene por referencia la gramática normativa y los fenómenos pragmático-discursivos, ampliamente discutidos en las bases de la investigación de la que se extrae este corpus.

Asimismo, es necesario tener en cuenta las variables que pueden surgir en torno a los aprendices y al tipo de ejercicio que se use para crear el corpus (ya que puede influir en la recogida de los datos), tales como la edad, la lengua nativa, el nivel de competencia lingüística, el entorno de aprendizaje, la exposición a la lengua extranjera, la duración de la prueba, el tipo de tarea o el lugar de realización. Estos factores pueden determinar la utilidad de un corpus general de aprendices de ELE. No obstante, en el corpus que se presenta, la variabilidad de los aprendices es mínima, pues el grupo seleccionado cuenta con características comunes que dotan de cohesión este corpus y realizan el mismo tipo de pruebas simultáneamente, como se comentará en las siguientes líneas en referencia a los informantes y a los cuestionarios.

2. Creación de corpus oral de aprendices de ELE

El primer paso para el diseño de la estructura y el contenido del corpus es la recopilación de los datos. Como se ha mencionado, el corpus se ideó con carácter oral y está formado por dos tipos de pruebas[2]: un cuestionario general para expresar opiniones y preferencias, y un cuestionario con una serie de dilemas para argumentar las respuestas. La elección de estos dos tipos de cuestionarios responde a razones relacionadas con la adquisición lingüística: la expresión de las opiniones y la argumentación supone un nivel intermedio-avanzado de lengua por las estructuras lingüísticas que se requieren. Generalmente, la producción oral se caracteriza por una simplificación del discurso (Briz, 1998), por lo que el uso de estructuras complejas supone un reto en la producción oral de aprendices de ELE; por otro lado, este tipo de cuestionarios permite localizar patrones de error en la adquisición de la lengua oral, identificando los errores más frecuentes de los aprendices, y detectando los que están fosilizados y pueden ser fosilizables. No obstante, conviene tener en cuenta que el hecho de no emplear una determinada estructura no supone su desconocimiento (Lozano, 2021), por lo que la comparación entre el uso del lenguaje en ambas pruebas resulta útil para el análisis de la interlengua de los informantes.

En el diseño de las pruebas, se intentó que las preguntas fueran pertinentes para los informantes en aras de asegurar que todos los participantes tuvieran una opinión formada al respecto. También se consideró que el desarrollo de las preguntas fuera semidirigido, es decir, se procuró que los informantes pudieran usar las estructuras gramaticales con las que más cómodos se sintieran y se pudieran extender el tiempo que desearan en cada pregunta porque, de esta forma, se puede analizar con mayor fiabilidad la interlengua de cada uno de ellos. Sin embargo, en la segunda prueba,

[2] Los cuestionarios de las pruebas se pueden consultar en: https://drive.google.com/file/d/1JPLiePvN78GFHjUZWgLQY5aByL6sShJu/view?usp=sharing

Muñoz de la Virgen, Cristina (2024). Herramientas para la codificación, el análisis y la explotación de un corpus oral de aprendices francófonos de español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_inv03, 1-25. http://doi.org/10.1344/teisel.v3.44741

© La autora | Recepción: 06/11/2023 | Revisión: 18/01/2024 | Aceptación: 09/02/2024 | Publicación: 11/03/2024

ISSN 2696-676X

se seleccionaron dilemas controvertidos que forzaran el uso de más complejas como la negación o las oraciones condicionales de realización poco probable.

La recopilación de los ejercicios tuvo lugar en el laboratorio de idiomas de la universidad, que contaba con un sistema de grabación de alta calidad; además, gracias a que cada informante realizaba las pruebas con auriculares, se conseguía minimizar las interrupciones y molestias entre los compañeros. Cada prueba se destinó un máximo de 30 minutos para su realización.

2.1. Los informantes y el cuestionario

Como ya se ha indicado, en el proceso de elaboración del corpus han participado 51 alumnos franceses pertenecientes al segundo curso del grado de Lenguas Extranjeras Aplicadas de la Universidad de François Rabelais de Tours (Francia). Estos alumnos suponen la totalidad de estudiantes matriculados en el curso, aunque hay que señalar que la cifra excede la aquí citada. El resto de los estudiantes no participaron en el estudio, bien por razones técnicas o bien por su situación personal y laboral, lo cual impedía que se pudiera garantizar el principio de igualdad entre el alumnado.

El francés es la lengua nativa y curricular de todos los informantes y se observa un desequilibrio de género en la creación del corpus, ya que los participantes son 44 mujeres frente a 9 hombres. No obstante, no se ha podido remediar, ya que se trata de un criterio selectivo puramente académico.

El momento en el que se ejecutaron las pruebas tampoco fue casual. Se realizaron en dos momentos diferentes, uno en la semana seis (mitad del semestre) y otro en la semana doce (final del semestre). Desde el comienzo de ese semestre, el alumnado fue preparado para que pudiera dar lo mejor de sí durante la grabación. De esta manera, el temario se centró en los valores del modo subjuntivo en la práctica oral.

La primera prueba era un cuestionario de veintitrés preguntas de carácter general con referencia al pasado, presente, futuro y condicional, incluyendo una complejidad progresiva. Las preguntas se obtuvieron del material didáctico común del Departamento de Español para la preparación de las clases de corte oral, mientras que la segunda prueba trataba de seis dilemas, extraídos de la página web *Haz lo que debas* (https://niaia.es/banco-de-actividades/). Esta wiki pertenece al Instituto Universitario de Ciencias de la Educación de la Universidad Autónoma de Madrid y tiene como objetivo la formación e investigación en la resolución de problemas morales. Con una larga trayectoria (2012), esta web parece la indicada para el planteamiento de dilemas por su aporte didáctico y enseñanza de la ética desde un punto de vista crítico. En la investigación, se consideró necesario este tipo de ejercicio para alentar el uso de determinadas estructuras lingüísticas, como las oraciones subordinadas introducidas por el modificador modal de la negación, las oraciones condicionales o las oraciones causales con una acción verbal hipotética.

Con estas pruebas se consiguió un tiempo de grabación de 30 horas y 41 minutos, divididos de la siguiente manera:

- 19 horas, 35 minutos y 41 segundos para la primera prueba.

Muñoz de la Virgen, Cristina (2024). Herramientas para la codificación, el análisis y la explotación de un corpus oral de aprendices francófonos de español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_inv03, 1-25. http://doi.org/10.1344/teisel.v3.44741

© La autora | Recepción: 06/11/2023 | Revisión: 18/01/2024 | Aceptación: 09/02/2024 | Publicación: 11/03/2024

ISSN 2696-676X

- 12 horas, 5 minutos y 26 segundos para la segunda prueba.

Este tiempo constituye el material auditivo sobre el que se creó el corpus. En el apartado 3, se detalla la cabecera de la intervención de cada informante, nombrando a cada uno de ellos y especificando el tiempo que necesitó para la finalización de la prueba.

2.2. Sistema de versiones en el corpus

El corpus oral de aprendices francófonos se desarrolló en cuatro etapas, que dieron lugar a cuatro versiones. La primera consistió en la transcripción mecanográfica del material auditivo. Aunque existen diversas herramientas para este fin, debido a la naturaleza del material, producciones orales en lengua extranjera, se optó por una transcripción manual que reflejara fehacientemente la producción del informante. La razón por la que se optó por la transcripción manual responde al elevado índice de errores morfosintácticos y discursivos, que se autocorregían cuando se trató de usar el software *Dragon NaturallySpeaking* (https://dragon-naturallyspeaking-premium.softonic.com/), y el deseo por reflejar las marcas propias de la oralidad, tales como las pausas. Para transcribir el contenido de las grabaciones, se siguió el modelo del Corpus Oral de Español como Lengua Extranjera (CORELE, http://cartago.lllf.uam.es/corele/pdf/convenciones.pdf), basado en la transcripción CHAT del proyecto C-Oral-Rom (http://www.lllf.uam.es/ESP/Coralrom.html) y las convenciones del proyecto SPLLOC (http://www.splloc.soton.ac.uk/). Esta versión permitió contar con la primera versión electrónica del corpus.

La segunda versión fue la anotación textual de la primera versión. Para este propósito, se siguieron criterios codificadores basados en la *Text Encoding Initiative* (TEI) y extraídos de la última versión del Consorcio TEI, publicada en 2018 y actualizada recientemente. Esta decisión se basó en el deseo de continuar con el espíritu cooperativo con el que nació la Iniciativa TEI (Vassar College, 1987) para la unificación de los criterios de codificación textual, a fin de que cualquier trabajo pueda ser reutilizado por otros investigadores. No obstante, la naturaleza del corpus de la investigación obligó a crear nuevas etiquetas, siguiendo las normas del metalenguaje XML y recurriendo a los editores que se presentarán en el apartado 5. Este metalenguaje, al ser independiente de plataformas informáticas y fabricantes de software, se ha convertido en la "lingua franca of the data exchange world" (St. Laurent y Fitzgerald, 2005, p. 1). Además, este metalenguaje aúna una serie de características que lo convierten en el apropiado para la creación de corpus porque permite la creación de etiquetas (marcas) propias, la asignación de atributos a las etiquetas, el estándar es de uso público, está basado en texto plano, es multilingüe, modular y ampliable, además, permite la conservación de datos a largo plazo (Fradejas, 2009-2010, pp. 226-227). Esta versión es un primer corpus anotado que permite la recuperación de la información pertinente y proporciona, de forma sistemática y accesible, los datos necesarios para el análisis (McEnery y Hardie, 2012, p. 13).

Muñoz de la Virgen, Cristina (2024). Herramientas para la codificación, el análisis y la explotación de un corpus oral de aprendices francófonos de español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_inv03, 1-25. http://doi.org/10.1344/teisel.v3.44741

© La autora | Recepción: 06/11/2023 | Revisión: 18/01/2024 | Aceptación: 09/02/2024 | Publicación: 11/03/2024

ISSN 2696-676X

La tercera versión trató de una revisión ortográfica anotada, atendiendo a normas de puntuación y estructuración textual, es decir, se trata de una versión revisada y mejorada de la primera anotación, donde se establecieron los párrafos en la intervención de cada informante y las cuestiones ortográficas que se pudieron perder en las versiones anteriores. Esta versión es el resultado final del corpus.

Finalmente, la versión 4 consistió en una revisión intervenida para agilizar la lectura del texto. En esta parte, se eliminaron las anotaciones, las pausas, las reformulaciones y las repeticiones de los informantes, se corrigieron los errores de pronunciación y los calcos léxicos y sintácticos, se ofrecieron alternativas ante una omisión de enunciado que dificulte la comprensión y se optó por una adaptación gramatical de los errores encontrados, a excepción de los verbos en modo subjuntivo usados incorrectamente, ya que se consideraron relevantes para el fin del corpus. El objetivo de esta versión es ofrecer al lector una comprensión rápida del contexto en el que se formula el dato deseado.

3. El esquema de la TEI

Cualquier corpus que se defina por las directrices del Consorcio TEI se estructura mediante un esquema específico en dos partes:

- *Head* (cabecera), que incluye toda la información relativa a la configuración del corpus. Se puede observar en las figuras 1 y 2.
- *Body* (cuerpo), que es el texto codificado en sí.

El esquema de codificación es adaptable a cualquier proyecto, por lo que cuenta con 21 módulos y 505 atributos que se combinan libremente para estructurar el corpus. No obstante, hay cuatro módulos obligatorios para asegurar la internacionalidad del sistema. Son *tei, core, header* y *textStructure*:

- *Tei*: todas las clases de atributos, modelos y las macros utilizadas en el esquema.
- *Core*: elementos básicos para la codificación.
- *Header*: metadatos descriptivos sobre la codificación.
- *TextStructure*: estructura predeterminada de cualquier documento TEI.

Cada producción oral de los informantes comienza con la codificación del elemento <TEI> que indica que se trata de un texto individual, formado por el atributo xmlns y el valor de la web. Para facilitar la visualización, se ha dividido la estructura en dos partes que se pueden observar en las siguientes figuras:

Muñoz de la Virgen, Cristina (2024). Herramientas para la codificación, el análisis y la explotación de un corpus oral de aprendices francófonos de español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_inv03, 1-25. http://doi.org/10.1344/teisel.v3.44741

© La autora | Recepción: 06/11/2023 | Revisión: 18/01/2024 | Aceptación: 09/02/2024 | Publicación: 11/03/2024

ISSN 2696-676X

Figura 1

Esquema de la cabecera. Elaboración propia

Figura 2

Esquema de la cabecera (bis). Elaboración propia

Como se observa, el primer elemento de la cabecera es el <teiHeider>. Se compone de los siguientes atributos:

Muñoz de la Virgen, Cristina (2024). Herramientas para la codificación, el análisis y la explotación de un corpus oral de aprendices francófonos de español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_inv03, 1-25. http://doi.org/10.1344/teisel.v3.44741

© La autora | Recepción: 06/11/2023 | Revisión: 18/01/2024 | Aceptación: 09/02/2024 | Publicación: 11/03/2024

ISSN 2696-676X

- *Type*. Indica el tipo de corpus.
- *Status*. Indica si el texto es nuevo o ha sido modificado.
- *Date.created*. Indica cuándo se creó el texto.
- *Date.updated*. Indica cuándo se modificó por última vez el texto.
- *Id*. Marca la referencia interna del texto y se compone de cuatro letras (identificadoras del título del texto codificado) y cinco números (identificadores del subtítulo codificado), precedidos por la marca *th*, la cual hace referencia a *teiHeader*.

Este elemento se puede ver en la línea (2) de la figura 1.

A continuación, está el elemento <TEI>, que se forma con cuatro componentes:

- Descripción de archivo (<fileDesc>). Cuenta con el atributo id precedido por "fd" y se subdivide en secciones:
    - Información sobre el título (<titleStmt>), que recoge el título del texto, así como al responsable de la transcripción y codificación del mismo. Se ve entre las líneas (5)-(11).
    - Información sobre la edición (<editionStmt>). Lleva un atributo "n", que indica el número de la versión y el de la revisión, y se forma mediante dos subelementos más: <edition> y <date>. El primero indica la versión y su fecha y el segundo, la exportación del texto oral. Se ve entre las líneas (12)-(20).
    - Información sobre la extensión del texto (<extent>). Se ve en la línea (21).
- Descripción del texto fuente (<sourceDesc>). Lleva el atributo id, precedido por "sd". Esta marca se subcategoriza en <biblStruct>, lo cual integra los datos del documento, y <monogr>, que consta de los siguientes elementos:
    - <respStmt>, nombra a los responsables del texto. Se ve entre las líneas (25)-(36). El número (35) contiene la identificación del informante.
    - <title>, nombra el título del texto; lleva el valor "or", que indica oralidad. Se ve en la línea (37).
    - <biblScope>, nombra el formato en el que se encuentra el texto. Se ve en la línea (38).
    - Dentro de la etiqueta <biblStruct> se recoge la procedencia del texto mediante el atributo "idno", como se ve en la línea (39).
- Descripción de la codificación <encodingDesc>. Cuenta con el atributo id. Se ve en la línea (45).
- Descripción del perfil de cada texto <profileDesc>. Este elemento informa sobre la fecha de clasificación del texto, su tipología y el número de hablantes que participan en la grabación. Lleva un identificador de texto precedido por "pd", seguido por las referencias internas, y tres componentes:
    - Creación <creation>.Fecha de clasificación del texto en el corpus. Coincide con la fecha de grabación o la fecha de la primera versión. Se ve entre las líneas (47) - (50).
    - <textClass> Identifica la clase a la que pertenece el texto según la tipología diseñada para el corpus. Lleva el subelemento <catRef>, que contiene un atributo "scheme", cuyo valor es el nombre de la tipología de textos, y un atributo, "target", que recoge un valor asignado al texto según la clasificación. Además, hay un subelemento <keywords> que especifica el nombre en clave del texto. Se ve entre las líneas (51) - (56).
    - Descripción de los hablantes <particDesc>. Se compone de tantos subelementos <person> como hablantes intervengan en el texto. Cuenta con atributos y elementos, pero solo se ha utilizado uno, "id", código que identifica al hablante en sus intervenciones. Se ve en la línea (58).

Muñoz de la Virgen, Cristina (2024). Herramientas para la codificación, el análisis y la explotación de un corpus oral de aprendices francófonos de español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_inv03, 1-25. http://doi.org/10.1344/teisel.v3.44741

© La autora | Recepción: 06/11/2023 | Revisión: 18/01/2024 | Aceptación: 09/02/2024 | Publicación: 11/03/2024

ISSN 2696-676X

El elemento <revisionDesc> sirve para anotar los cambios que se realizan en el corpus. Su componente <change> indica el cambio a través del subelemento <item> y la fecha <date>. Este elemento no aparece en los cuadros, ya que no se realizaron modificaciones significativas en el momento de su validación.

Además, a cada elemento <text> corresponde un texto oral unitario, que comienza con la etiqueta <body> y termina con </body>.

4. El etiquetado del corpus

El etiquetado del corpus categoriza los aspectos puramente lingüísticos. Como se ha dicho, en la codificación de este corpus se han combinado marcas usadas en TEI con otras de creación propia, pero preparadas para ser validadas por la comunidad internacional. El motivo por el cual se decidió utilizar un etiquetado nuevo y exclusivo en este corpus responde a criterios de la propia investigación. Era necesario contar con unas etiquetas que permitieran reflejar los aspectos morfosintácticos y pragmáticos en referencia con la oralidad en la producción de los informantes. Para ello, se siguieron las normas del metalenguaje XML que aplica el Consorcio TEI para que en un futuro esas etiquetas puedan ser incluidas como marcas propias de TEI. Por ejemplo, el Consorcio utiliza el elemento ~~…~~ para señalar que hay algún elemento, una letra, una palabra o una oración, marcado como borrado o superfluo en el texto. A este elemento, se le puede añadir un atributo como *repetition*, lo cual suma información al elemento marcado. La secuencia queda así: <*del type=”repetition”*>…</*del*>. Esta etiqueta es muy productiva en el discurso oral en lengua extranjera, entre otras cosas, por cierta falta en la fluidez y competencia comunicativa. Sin embargo, en el corpus se encontró otro tipo de repetición, cuando el informante se corregía. Se consideró que este hecho es lo suficientemente destacable como para contar con su propio atributo que permitiera la rápida localización en el corpus. No obstante, esa parte corregida sugería que el marcado anterior se debía eliminar; de esta manera, se creó acorde con la etiqueta de repetición: <*del type=”correction”*>… </*del*>.

Del conjunto de las dieciocho etiquetas con el que se puso en marcha el corpus, nueve fueron de creación propia para tratar la adquisición del modo subjuntivo en el plano oral. Aunque originalmente las etiquetas funcionaron para el fin con el que fueron creadas, se reconoce que resultan insuficientes e, incluso, ambiguas para otro tipo de investigación o codificado. Por esa razón, se está llevando a cabo un trabajo de actualización de los atributos de las etiquetas, que resulten más funcionales para explotar y mejorar las futuras investigaciones sobre diferentes aspectos lingüísticos en el proceso de aprendizaje-adquisición del español como lengua extranjera.

Las etiquetas que se encuentran en el corpus se pueden clasificar en cuatro grupos, donde se indicarán las etiquetas de creación propia:

- Etiquetas morfosintácticas:
    - <sic type=”concordance”>…</sic>. Se emplea en los usos erróneos de una forma verbal y es de creación propia. Ejemplo: *Los conocimientos* <*sic type=”concordance”*>*están*</*sic*> *todo*

Muñoz de la Virgen, Cristina (2024). Herramientas para la codificación, el análisis y la explotación de un corpus oral de aprendices francófonos de español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_inv03, 1-25. http://doi.org/10.1344/teisel.v3.44741

© La autora | Recepción: 06/11/2023 | Revisión: 18/01/2024 | Aceptación: 09/02/2024 | Publicación: 11/03/2024

ISSN 2696-676X

el tiempo importantes. Esta etiqueta recoge todos los errores que atañen al sistema verbal, a excepción del subjuntivo.

- <sic type="subjunctive">…</sic>. Se emplea en los usos erróneos del subjuntivo y es de creación propia. Ejemplo: *No pienso que no <sic type="subjunctive">estoy</sic> descontenta.*
- <verb>…</verb>. Se emplea en el uso correcto del subjuntivo y es de creación propia. Ejemplo: *Si <verb>encontrara</verb> una cartera con <num value="1000">mil</num> dólares…*
- <sic type="ellipsis">… </sic>. Se emplea en la supresión de algún elemento en el enunciado y es de creación propia. Ejemplo: *Si quiere realmente <sic type="ellipsis"></sic> su hijo…*

- Etiquetas discursivas:
  - <pause></pause>. Se emplea en las paradas pronunciadas, un fenómeno oral que pretende reflejar la producción del informante. Ejemplo: *Para mí, <rs desc="Político">Churchill</rs> <pause></pause> tendría que avisar a la población.*
  - <unclear>…</unclear>. Se emplea con las palabras o fragmentos que no están claros en la grabación por una pronunciación incorrecta. Ejemplo: *Puede ser un servicio también para la <unclear></unclear>, la ciudad, no solamente para la gente*. Se usa cuando la comprensión es imposible y no se puede transcribir.
  - <del type="uncompleted"></del>. Se emplea en los abandonos de fragmentos u oraciones y es de creación propia. Ejemplo: *Y si el hijo no puede tener confianza en su madre <del type="uncompleted"></del>.*
  - <del type="reformulation"></del>. Se emplea en las reformulaciones de un fragmento u oración y es de creación propia. Ejemplo: *Porque el hospital <pause></pause> <del type="reformulation"></del>, porque la situación de los trabajadores del hospital es importante.* Esta etiqueta es muy productiva en el corpus, ya que recoge un recurso muy común en la oralidad.
  - <del type="hypercorrection">…</del>. Se emplea una repetición de una palabra o fragmento hipercorrigiendo el elemento y es de creación propia. Ejemplo: *Buscar un nuevo <del type="hypercorrection">una nueva</del> sitio.*
  - <del type="correction">… </del>. Se emplea en una repetición de una palabra o fragmento corrigiendo el error y es de creación propia. Ejemplo: *Debe tratarse con una, <del type="correction">con un</del> reforzamiento.*
  - <del type="repetition">…</del>. Se emplea en la repetición de una palabra o fragmento. Ejemplo: *No fumar y, <del type="repetition">y</del> dormir.*
  - <event>… </event>. Se emplea en los elementos fónicos usados en el corpus. Ejemplo: *<event>Risa</event> Pienso que…*
- Etiquetas léxico-semánticas:
  - <foreign lang="…">…</foreign>. Se emplea en las palabras extranjeras que aparecen en el corpus. Ejemplo: *<foreign lang="fr">Action contre la faim</foreign>.*
  - <sic>…</sic>. Se emplea en los errores genéricos detectados en el corpus. Ejemplo: *Tenía <sic>muy</sic> <sic>dineros</sic>.*
  - <sic type="calque">…</sic>. Se emplea en los calcos gramaticales y es de creación propia. Ejemplo: *<sic type="calque">un otro</sic> caso similar.* Esta etiqueta puede incluirse en el grupo morfosintáctico, ya que afecta tanto al léxico como a la sintaxis.
- Etiquetas de referencias:
  - <abbr>…</abbr>. Se emplea con las abreviaturas. Ejemplo: *<foreign lang="fr"> <abbr type ="acronym" repr = "LEA"> LEA </abbr> </foreign>.*

Muñoz de la Virgen, Cristina (2024). Herramientas para la codificación, el análisis y la explotación de un corpus oral de aprendices francófonos de español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_inv03, 1-25. http://doi.org/10.1344/teisel.v3.44741

© La autora | Recepción: 06/11/2023 | Revisión: 18/01/2024 | Aceptación: 09/02/2024 | Publicación: 11/03/2024

ISSN 2696-676X

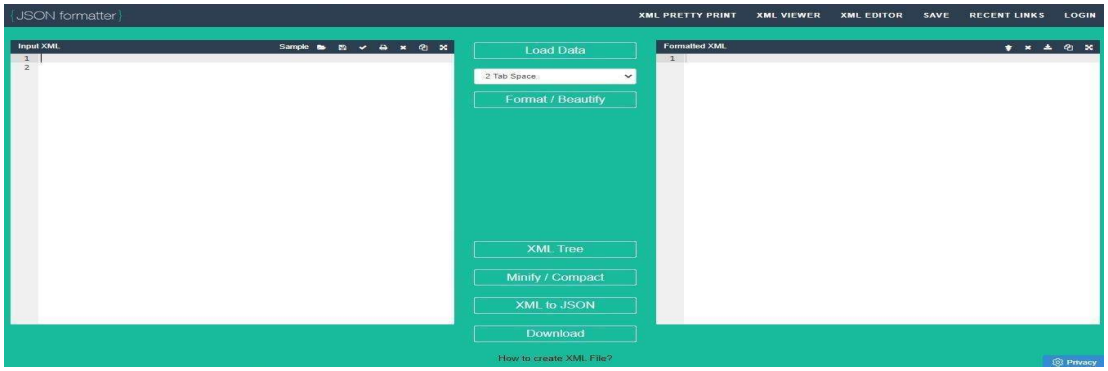
- <rs>…</rs>. Se emplea en el uso de nombres propios. Ejemplo: <*rs desc="Político">Churchill</rs*>.
- <num value="…">…</num>. Se emplea en las cifras que aparecen en el corpus. Ejemplo: <*num value="1">uno</num*>.

5. Empleo de herramientas para la codificación, el análisis y la explotación del corpus

Para desarrollar esta investigación se utilizaron principalmente ocho herramientas que permitieron la codificación, el análisis y la explotación del corpus. En un primer momento, se optó por utilizar el Bloc de Notas (*Notepad*) de Windows (https://es.wikipedia.org/wiki/Bloc_de_notas), ya que está incluido en el sistema operativo y permite el uso del texto plano (.txt), pero se acabó descartando porque no ofrece ninguna herramienta diseñada para facilitar el trabajo con XML y genera problemas cuando se intenta validar algún código. Por ejemplo, uno de los problemas más comunes es el uso del entrecomillado, ya que las comillas dobles o inglesas que por defecto marca Windows no son las mismas que requiere el lenguaje XML, así que cuando se intente validar el código, dará error en cada línea donde se hayan usado. Por ese motivo, se decidió usar la herramienta gratuita *XML Formatter* (https://jsonformatter.org/xml-formatter) para crear la cabecera del texto, según los elementos descritos anteriormente.

Figura 3

Muestra de XML Formatter. Extracción de la web



Esta herramienta es muy simple de usar y ayuda a editar, formatear y analizar los datos XML. Solo hay que cargar el texto etiquetado en la parte de la izquierda y aparecerá duplicado en la parte derecha con los posibles errores señalados. Una vez creada, conviene validarla para subsanar los errores que pudieran surgir. Esta misma herramienta permite la validación del texto, tal y como se aprecia en las figuras 1 y 2, pero, en este punto, se empleó, como garantía de doble comprobación, la herramienta gratuita *XML Validator* (https://codebeautify.org/xmlvalidator).

Muñoz de la Virgen, Cristina (2024). Herramientas para la codificación, el análisis y la explotación de un corpus oral de aprendices francófonos de español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_inv03, 1-25. http://doi.org/10.1344/teisel.v3.44741

© La autora | Recepción: 06/11/2023 | Revisión: 18/01/2024 | Aceptación: 09/02/2024 | Publicación: 11/03/2024

ISSN 2696-676X

Figura 4

Muestra de XML Validator. Extracción de la web

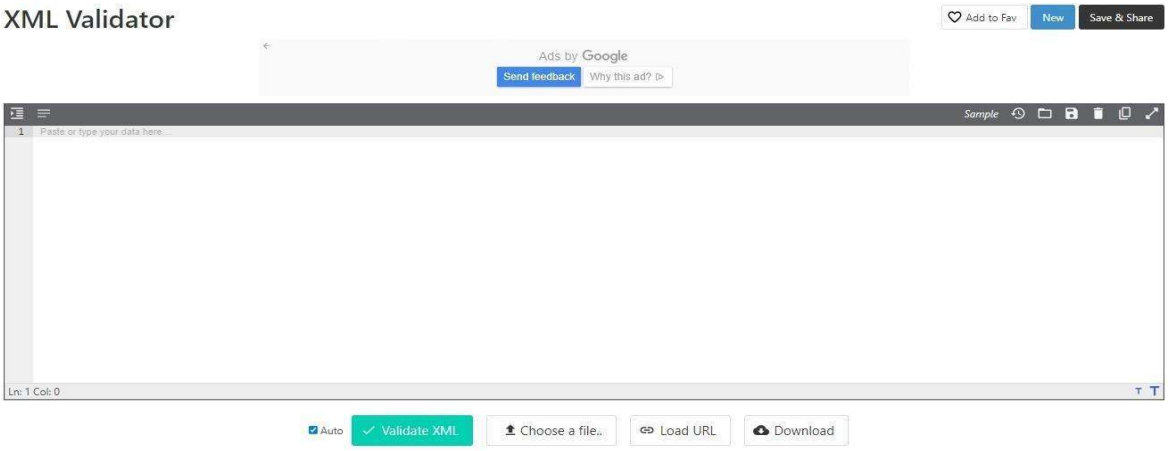


Figura 5

Muestra de validación de la cabecera del corpus con XML Validator. Extracción de la web

Ambas herramientas resultan útiles porque no solo señalan los errores, sino que también guían en la corrección. Es necesario hacer una validación de las anotaciones para asegurarse de que la codificación se ha realizado de manera correcta. Cuando el sistema detecta un error en la escritura de texto, lo señala de dos maneras: un triángulo de advertencia, si el error es producido por un fallo en un atributo (ejemplo, un olvido en las comillas) y con una señal de error roja si falta algún elemento de cierre. Es interesante que ambos sistemas ofrezcan una propuesta al código creado con el objetivo de mejorar algún elemento. Hay que tener en cuenta que las cabeceras de los textos resultan importantes para su correcta identificación en un sistema informático, ya que contienen todos los metadatos asociados al documento digital y permiten observar las diferentes versiones por las que pasa un corpus; de ahí a que se extremen las precauciones que garanticen sus correcciones.

Muñoz de la Virgen, Cristina (2024). Herramientas para la codificación, el análisis y la explotación de un corpus oral de aprendices francófonos de español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_inv03, 1-25. http://doi.org/10.1344/teisel.v3.44741

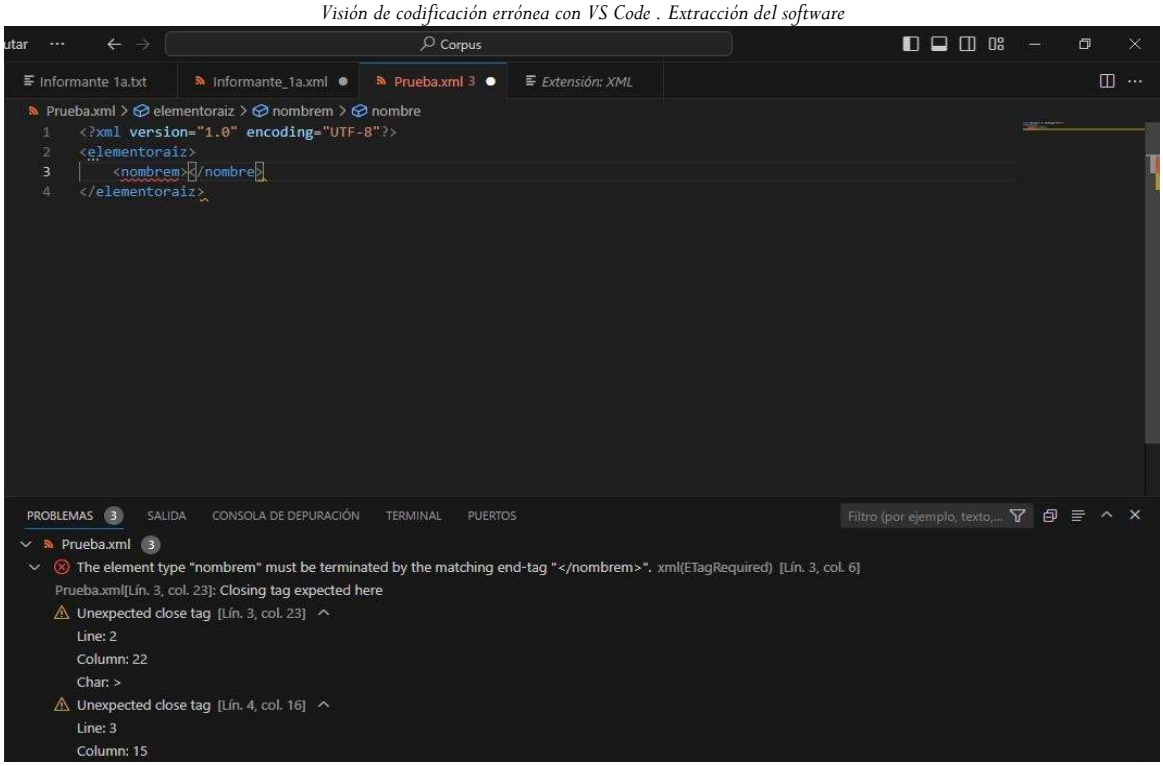
© La autora | Recepción: 06/11/2023 | Revisión: 18/01/2024 | Aceptación: 09/02/2024 | Publicación: 11/03/2024

ISSN 2696-676X

Por otro lado, un software gratuito y de sencillo manejo para la lectura de corpus anotados en TEI es el editor *Visual Studio Code (VS Code)* (https://code.visualstudio.com/). Se trata de un software libre creado por Microsoft y mantenido por la comunidad de programadores, el cual está disponible para Windows, MacOs y Linux. Tras instalar las extensiones, *Paquete de idioma español para VS Code*, la cual permite configurar el software en español, *XML* para ayudar a la edición de la sintaxis en archivos XML y *Open in browser*, que permite abrir cualquier archivo y no únicamente páginas html, el programa permite anotar y validar al mismo tiempo que se elabora. Esta herramienta fue la que se empleó para el texto del corpus. Lo primero que se hizo fue crear una carpeta para almacenar el proyecto. En este caso, se llamó Corpus. Tras eso, en el editor, se fueron creando los archivos, como Informante_1a.xml, que corresponde al primer cuestionario del informante 1 (figura 7). Gracias a la instalación de las extensiones, en el momento en que se abre el tabulador, aparecen varias opciones como *insertar la declaración xml*, imprescindible para la correcta lectura. Este editor también ofrece soluciones a los problemas de lenguaje que van apareciendo conforme se va añadiendo la información al texto y reemplaza automáticamente el error, una vez aceptado. La figura 6 muestra una prueba creada con errores conscientes para mostrar el funcionamiento del editor. En el atributo "nombre", se ha introducido una "m" final que no coincide con el cierre de la etiqueta. En las últimas líneas del documento, se observa un aviso que informa del error.

Figura 6

Visión de codificación errónea con VS Code . Extracción del software



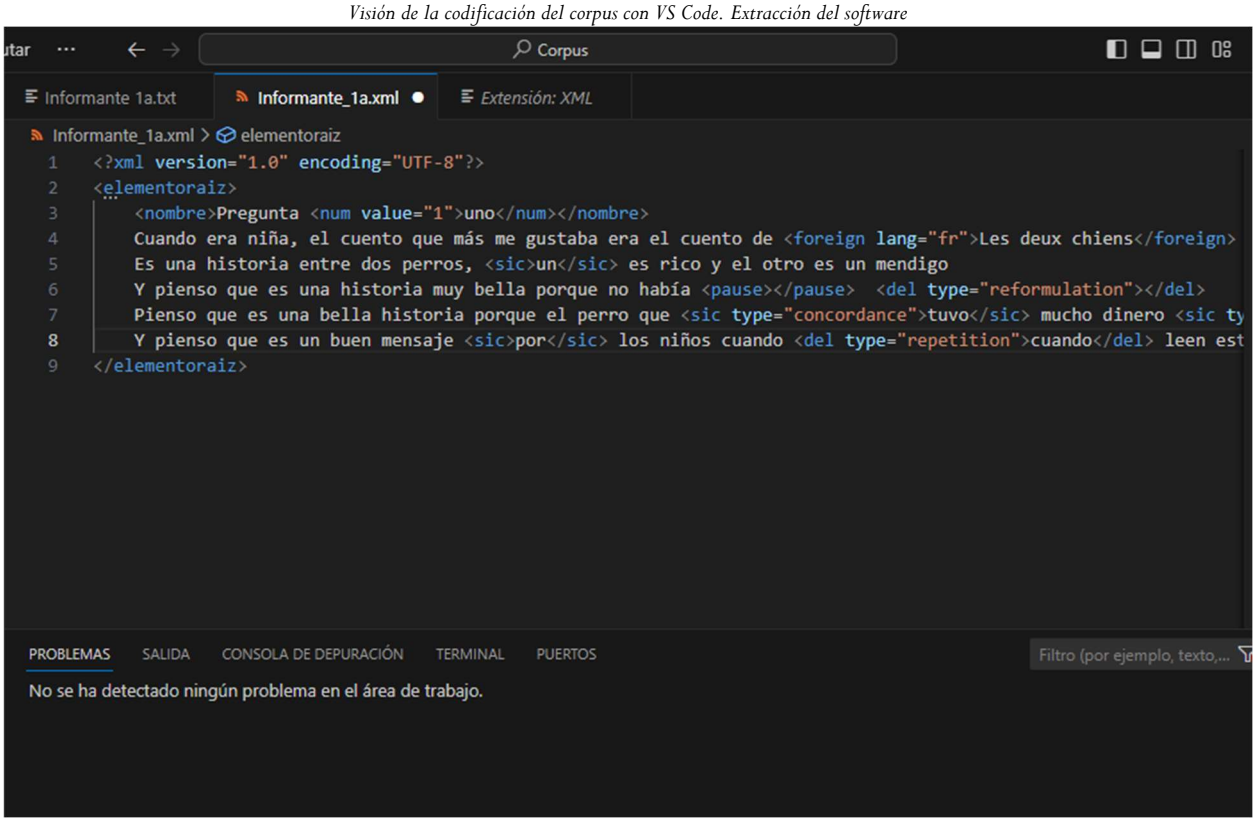
Muñoz de la Virgen, Cristina (2024). Herramientas para la codificación, el análisis y la explotación de un corpus oral de aprendices francófonos de español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_inv03, 1-25. http://doi.org/10.1344/teisel.v3.44741

© La autora | Recepción: 06/11/2023 | Revisión: 18/01/2024 | Aceptación: 09/02/2024 | Publicación: 11/03/2024

ISSN 2696-676X

Figura 7

Visión de la codificación del corpus con VS Code. Extracción del software



Otra aplicación de acceso libre, también empleada para esta investigación, es *Github* (https://github.com/), que permite gestionar y almacenar proyectos; además, tiene un componente de cooperación entre la comunidad y *Git* (https://git-scm.com/).

Git es un sistema de control de versiones distribuido de código abierto desarrollado por Linus Torvalds, el creador de Linux. Se trata de un control de versiones porque permite que el software se pueda descargar para hacer los cambios pertinentes y subir la nueva versión. No se reescribe en el texto original, sino que se crean las versiones del proyecto, las cuales permiten que un desarrollador interesado en un proyecto pueda contribuir con el código del software. *Hub* hace referencia a la comunidad de desarrolladores y la interacción con otros usuarios de la aplicación. Esta aplicación permite trabajar con repositorios locales y en línea, teniendo siempre disponible el trabajo almacenado. Es conveniente descargarla en el ordenador para poder vincularla con la herramienta *VS Code*, mediante las opciones que se dan en la instalación. El sistema de trabajo es el que se describe a continuación:

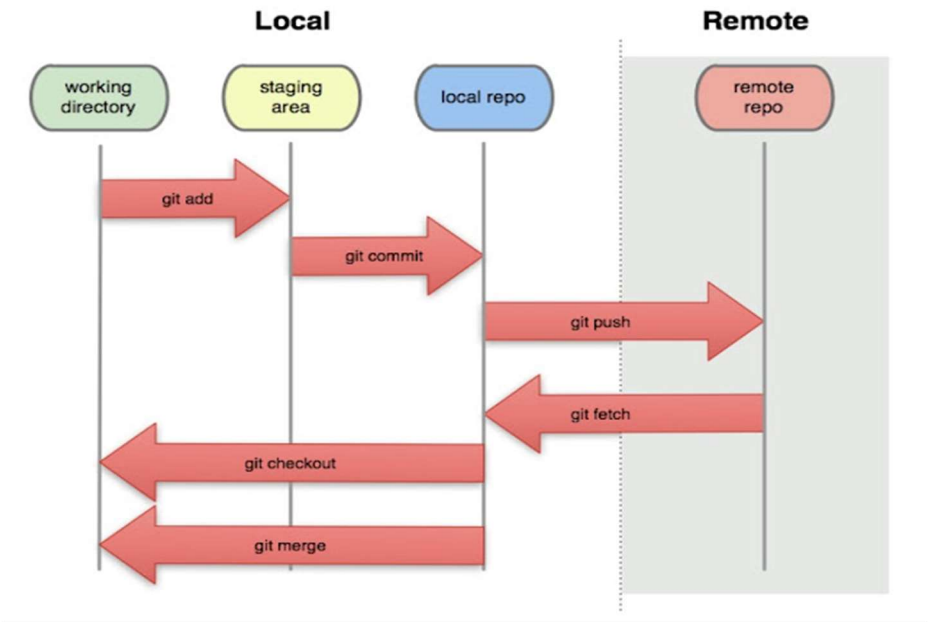
Muñoz de la Virgen, Cristina (2024). Herramientas para la codificación, el análisis y la explotación de un corpus oral de aprendices francófonos de español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_inv03, 1-25. http://doi.org/10.1344/teisel.v3.44741

© La autora | Recepción: 06/11/2023 | Revisión: 18/01/2024 | Aceptación: 09/02/2024 | Publicación: 11/03/2024

ISSN 2696-676X

Figura 8

Flujo de trabajo de Git. Tomada de la web[3]



Una vez descargado Git, se comienza el proyecto en *VS Code*. En el caso del corpus oral que se presenta, se creó una carpeta en Git con las producciones de los informantes en texto plano y se arrastró a *VS Code*. Tras ello, se procedió a la vinculación de ambas aplicaciones mediante el código *git version* en la terminal de *VS Code*. En ese momento, el proyecto se encuentra en *Working directory*, en la primera columna de la figura 8. Para relacionar los archivos de *VS Code* con un repositorio en *Git*, simplemente hay que escribir el código *git init* en el terminal. Eso permitirá respaldar las diferentes versiones que se realicen en el repositorio y visualizar si los archivos se crearon, se modificaron o se eliminaron. En este punto, los archivos pasan a la segunda columna de la figura 8, *Stanging area*, a través del código *git add*. Esta columna es un área temporal que prepara los archivos para ser enviados al repositorio *Local repo*, en la tercera columna con la ayuda del código *git committ*. De esta manera, se crea la primera versión del repositorio. A partir de entonces, cuando se realiza un cambio, se debe guardar realizando el recorrido inverso, esto es, primero se envía al *Stanging area* y, luego, al *Working directory*. Hasta este punto, el trabajo del repositorio está guardado en el almacenaje local, en el ordenador. Es el momento de respaldar la copia del repositorio en remoto con la ayuda de la aplicación *GitHub*. En la web de esta aplicación, cuando se elige la opción de crear un nuevo repositorio, aparecerá un código que habrá que copiar y pegar en la terminal de *VS Studio*. De esta manera, el repositorio del corpus oral que se creó en local ya está disponible en la nube, a salvo de lo que pueda ocurrir con el ordenador donde se inició el trabajo del corpus.

3 https://spaceanalytics.blogspot.com/

Muñoz de la Virgen, Cristina (2024). Herramientas para la codificación, el análisis y la explotación de un corpus oral de aprendices francófonos de español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_inv03, 1-25. http://doi.org/10.1344/teisel.v3.44741

© La autora | Recepción: 06/11/2023 | Revisión: 18/01/2024 | Aceptación: 09/02/2024 | Publicación: 11/03/2024

ISSN 2696-676X

Por otro lado, cabe comentar que se está realizando una actualización del corpus oral para mejorar el etiquetado y, por consiguiente, ampliar las posibilidades de investigación que se puedan realizar una vez terminado. Hasta el momento, se sigue manteniendo un control de versiones en *GitHub* y se continúa trabajando con *VS Studio*. Sin embargo, aunque la finalidad del corpus seguirá siendo la explotación del discurso oral de los aprendices francófonos de ELE, no hay un elemento específico de análisis. Para cumplir con tal fin, la versión final del corpus debería ser lo más detallada posible, por lo que se decidió introducir dos nuevas herramientas: *SpaCy* (https://SpaCy.io/) y *CATMA* (https://catma.de/). Ambas son plataformas abiertas y vinculables con *GitHub*.

SpaCy es una librería creada para facilitar el procesamiento avanzado del lenguaje natural en Python. Esta herramienta permite la anotación morfológica automática del corpus, lo cual resulta útil debido a que se trata de un corpus en lengua extranjera que pretende analizar los procesos de adquisición lingüística en los informantes.

Esta librería permite diversos tipos de instalación. En este caso, se optó por hacerlo a través de *Google Colaboratory* (https://colab.google/) porque su instalación es muy sencilla:

Figura 9

Instalación de SpaCy con Google Colaboratory. Extracción del software

La figura 9 muestra un extracto de la instalación de *SpaCy* mediante *Google Colaboratory*. Simplemente, hay que teclear en Google "SpaCy Google Colaboratory" y el servidor de Google ofrecerá la página de la figura 9. En este momento, se puede instalar y ejecutar desde la propia página o se puede crear una copia en el *Drive* de Gmail (https://www.google.com/intl/es_es/drive/). Naturalmente, también se puede guardar en *GitHub*. Para descargar la herramienta, tan solo hay que ejecutar el botón de *play*, que se ve en la figura 9. Una vez que se complete la descarga, el usuario debe instalar la herramienta ejecutando algunos "*plays*" que aparecerán durante la descarga. Este será el

Muñoz de la Virgen, Cristina (2024). Herramientas para la codificación, el análisis y la explotación de un corpus oral de aprendices francófonos de español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_inv03, 1-25. http://doi.org/10.1344/teisel.v3.44741

© La autora | Recepción: 06/11/2023 | Revisión: 18/01/2024 | Aceptación: 09/02/2024 | Publicación: 11/03/2024

ISSN 2696-676X

momento de especificar el lenguaje en el que se cargará el corpus. Se pueden descargar varios idiomas, como es el caso (se puede ver en la tercera y cuarta línea verde del ejecutador de la figura 9), sin que ello interfiera en el corpus porque en cada sesión en la que se use *SpaCy* se deberá ejecutar la herramienta y seleccionar el idioma. Generalmente, la instalación y ejecución es muy simple y no debería llevar más de algunos minutos; sin embargo, si se ha producido algún cambio en la configuración o requisitos de la herramienta, el preinstalador de *Google Colaboratory* no funcionará. En este caso, se recomienda la subsanación del error a través de *ChatGPT* (https://chat.openai.com/), ya que ofrece soluciones efectivas a los errores de instalación de *SpaCy*. El funcionamiento es igualmente sencillo: habría que copiar la alternativa de código que *ChatGPT* ofrezca para la subsanación del error en *Google Colaboratory* y se volvería a ejecutar hasta que se instale correctamente.

La herramienta de *SpaCy* permite varios usos: desde los más simples como la categoría gramatical de las palabras o la división del documento en oraciones hasta los patrones (*matching*), pasando por varios procesos como:

- Tokenización: separación del texto en partes llamadas *tokens*, que son las unidades mínimas de análisis. Con ello, se puede extraer información diversa (número de palabras, cuántas funcionan como conectores oracionales, cuántas cifras hay en el texto, cuántos verbos aparecen), lo cual puede resultar útil para hacer una valoración de los datos extraídos. En el caso del corpus oral de aprendices francófonos de ELE, la tokenización fue relevante para relativizar los errores de los informantes, extrayendo el porcentaje de verbos mal conjugados sobre el total de verbos, entre otros análisis.
- Lematización: este proceso es interesante para el análisis del léxico, ya que la herramienta ofrece un índice de frecuencias tanto en la palabra como en el lema. También es útil para el estudio de la formación de palabras y la adquisición del léxico.
- Reconocimiento de entidades (NER): este proceso reconoce y clasifica palabras predefinidas del modelo con el que se trabaja. Estas palabras predefinidas son nombres propios de personas, organizaciones, empresas, instituciones, agencias, localizaciones, accidentes geográficos, entre otros. Además, puede extraer información específica del texto y crear un índice de frecuencia; por ejemplo, puede filtrar solo las instituciones que aparecen en el documento y cuál es la que más se repite.

Los patrones ayudan a localizar estructuras específicas en el corpus. Si bien es cierto que, de momento, *SpaCy* no opera con TEI, mediante el uso de los patrones se pueden localizar los elementos necesarios:

Muñoz de la Virgen, Cristina (2024). Herramientas para la codificación, el análisis y la explotación de un corpus oral de aprendices francófonos de español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_inv03, 1-25. http://doi.org/10.1344/teisel.v3.44741

© La autora | Recepción: 06/11/2023 | Revisión: 18/01/2024 | Aceptación: 09/02/2024 | Publicación: 11/03/2024

ISSN 2696-676X

Figura 10

Patrones de SpaCy. Extracción del software

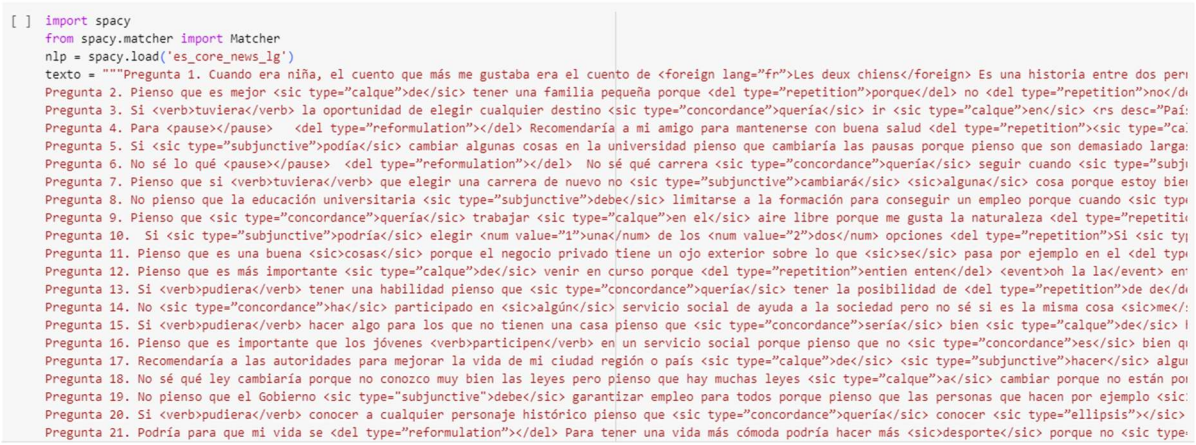


Figura 11

Patrones de SpaCy 2. Extracción del software

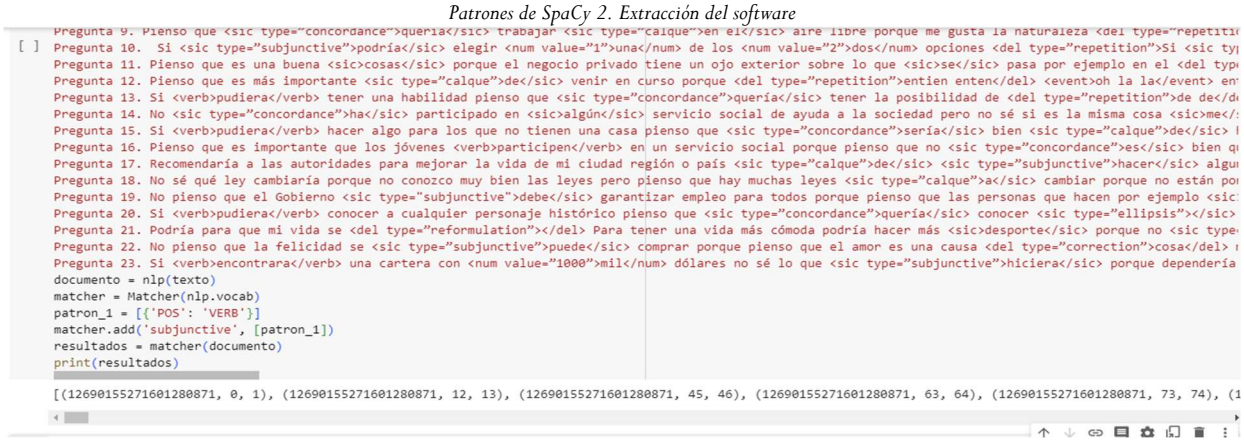
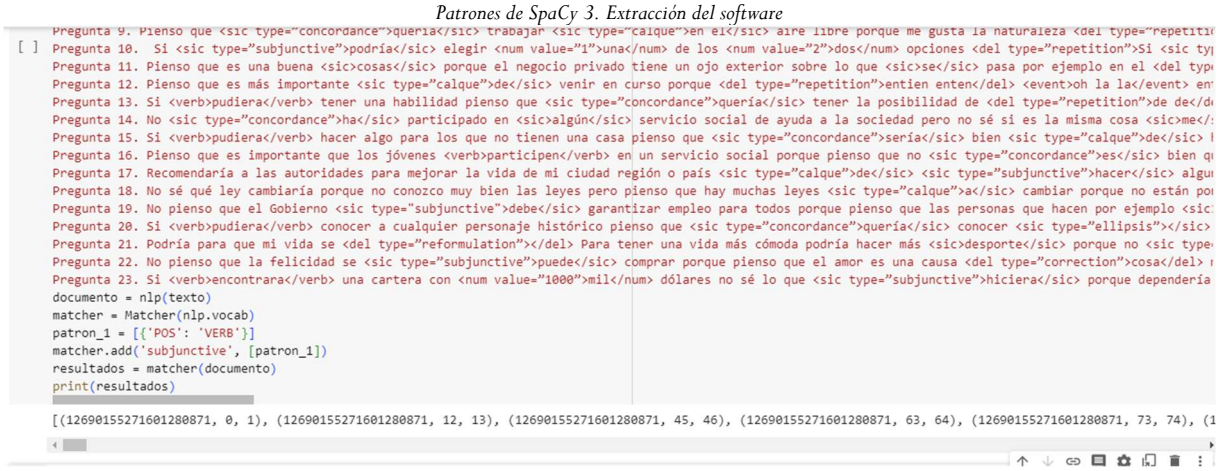


Figura 12

Patrones de SpaCy 3. Extracción del software



Muñoz de la Virgen, Cristina (2024). Herramientas para la codificación, el análisis y la explotación de un corpus oral de aprendices francófonos de español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_inv03, 1-25. http://doi.org/10.1344/teisel.v3.44741

© La autora | Recepción: 06/11/2023 | Revisión: 18/01/2024 | Aceptación: 09/02/2024 | Publicación: 11/03/2024

ISSN 2696-676X

Como se ve en las figuras 10, 11 y 12, se ha recuperado la información anotada bajo la etiqueta subjuntivo (*subjuntive*). El procedimiento ha sido el siguiente:

En primer lugar, se cargó el modelo de trabajo, en este caso, el español. A continuación, se copió y pegó parte del corpus (tan solo la primera actuación del informante 1), tal y como se ve en la figura 10[4]. La primera extracción de la información se ve en la figura 11 en la última línea de la imagen. En este ejemplo, la información que se pretende recuperar es el número de veces que el Informante 1 ha conjugado erróneamente el modo de un verbo. Para ello, se ha utilizado el patrón 1 - *POS:VERB*, donde *POS* hace referencia a la parte del discurso, esto es, el corpus, y *VERB* es el lema a recuperar. En la siguiente línea (*matcher.add*), se especifica qué es lo que se quiere recuperar (*subjunctive*) y, finalmente, en la última línea, se ven las tuplas con la posición en las que aparece en el corpus, lo cual permite localizar los ítems en el corpus y saber cuántos datos se han recuperado. En la figura 11, con el código expuesto, se extraen las muestras de texto anotadas.

Este ejemplo es únicamente una muestra simple del potencial de *SpaCy*. Se pueden escribir diferentes tipos de patrones para recuperar información más concreta, en función de la investigación que se esté realizando.

Igualmente, se ha decidido usar la aplicación *CATMA* (https://catma.de/). Esta interfaz de código abierto permite marcar y analizar textos en la web, sin necesidad de descargar ningún programa, mediante etiquetas definibles por el investigador. Con un uso muy intuitivo, tras la creación de la cuenta, se procede a nombrar el proyecto (*Corpus de aprendices francófonos de español*); a continuación, se agrega los documentos que conforman el proyecto. En este caso, se agregaron los 102 textos de los informantes en formato de texto plano y con las anotaciones de la primera investigación. Una vez creada la colección, es el momento de agregar las etiquetas que se emplearán. La aplicación permite crear etiquetas en función de la necesidad de uso, por lo que no hay que establecerlas todas antes de comenzar la anotación del texto.

La creación de etiquetas está organizada en grupos. En este punto, se ha decidido mantener las cuatro categorías originales de la anotación del corpus: etiquetas morfosintácticas, etiquetas léxico-semánticas, etiquetas pragmático-discursivas y etiquetas de referencia. Dentro de cada grupo, se clasifican las etiquetas que se usarán en la anotación:

[4] Hay que aclarar que en esta figura aparece el corpus con su anotación original, ya que el nuevo etiquetado está aún en proceso.

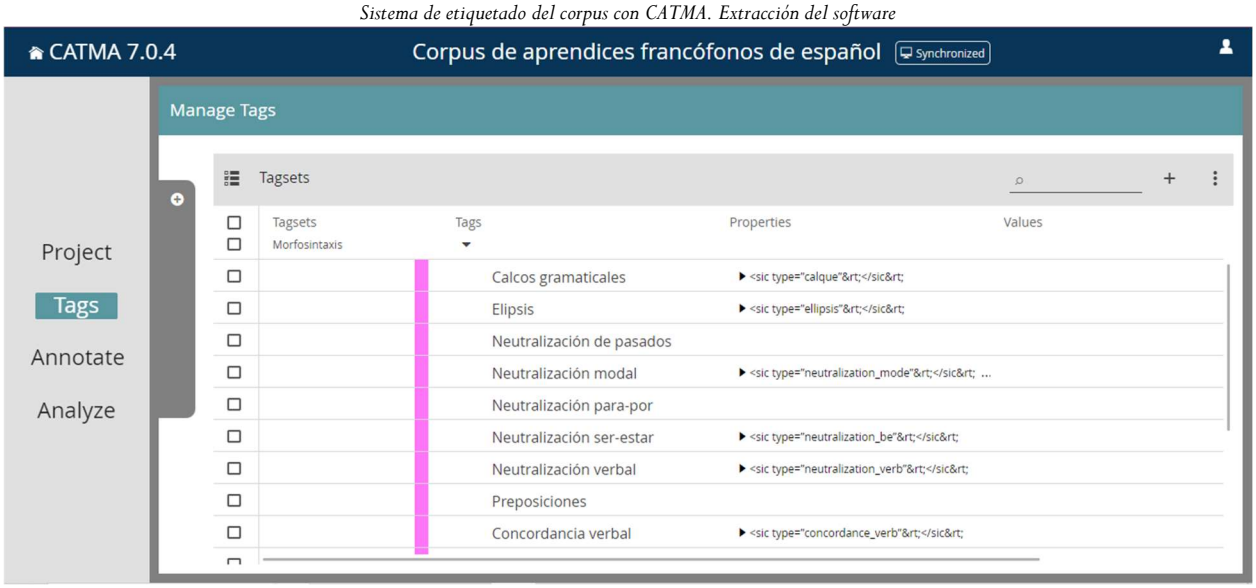
Muñoz de la Virgen, Cristina (2024). Herramientas para la codificación, el análisis y la explotación de un corpus oral de aprendices francófonos de español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_inv03, 1-25. http://doi.org/10.1344/teisel.v3.44741

© La autora | Recepción: 06/11/2023 | Revisión: 18/01/2024 | Aceptación: 09/02/2024 | Publicación: 11/03/2024

ISSN 2696-676X

Figura 13

Sistema de etiquetado del corpus con CATMA. Extracción del software



Aquí se ve una muestra del etiquetado morfológico. A cada etiqueta se le puede añadir una serie de propiedades, las cuales, en este caso, son la anotación en TEI, y se les puede asignar un color. Hasta la fecha, la actualización del etiquetado del corpus ha mantenido las cuatro categorías anteriormente descritas, pero ha aumentado en el número de subcategorías:

- Etiquetas morfosintácticas. Es el grupo con mayor número de cambios. Tan solo se mantiene la etiqueta <*sic type="ellipsis"*>… </*sic*>, pero se modifica la etiqueta <*sic type="subjunctive"*>… </*sic*>, la cual ahora se emplea para los usos correctos del modo subjuntivo y se mantiene en esta categoría también la etiqueta <*sic type="calque"*>…</*sic*>, para los calcos gramaticales. Se añaden las siguientes etiquetas, todas de creación propia:
    - <sic type="neutralization_past">… </sic>. Se emplea en los errores en los que el informante confunde los tiempos de pasado.
    - <sic type="neutralization_mode">… </sic>. Se emplea ante los errores de alternancia modal.
    - <sic type="neutralization_for">… </sic>. Se emplea con los errores neutralizados por la oposición preposicional por-para.
    - <sic type="neutralization_be">… </sic>. Se emplea en los errores de alternancia verbal ser-estar.
    - <sic type="neutralization_verb">… </sic>. Se emplea cuando el informante confunde los tiempos verbales, por ejemplo, cuando usa el condicional en lugar del futuro.
    - <sic type="preposition">… </sic>. Se emplea para el uso incorrecto de las preposiciones.
    - <sic type="concordance_verb">… </sic>. Se emplea para los errores en las designaciones verbales.
    - <sic type="concordance_gender">… </sic>. Se emplea ante los errores de concordancia léxica de género.
    - <sic type="concordance_number">… </sic>. Se emplea ante los errores de concordancia léxica de número.

Muñoz de la Virgen, Cristina (2024). Herramientas para la codificación, el análisis y la explotación de un corpus oral de aprendices francófonos de español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_inv03, 1-25. http://doi.org/10.1344/teisel.v3.44741

© La autora | Recepción: 06/11/2023 | Revisión: 18/01/2024 | Aceptación: 09/02/2024 | Publicación: 11/03/2024

ISSN 2696-676X

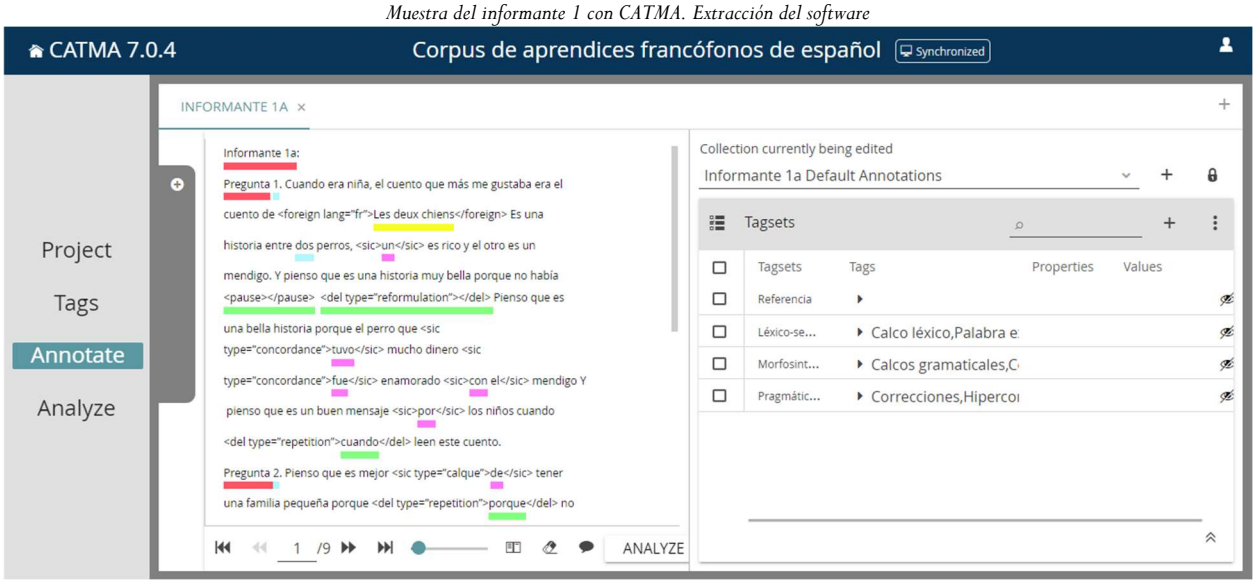
- <sic type="number">… </sic>. Se emplea para los errores en el uso de los determinantes numerales.

- Etiquetas discursivas. Se mantienen las ocho etiquetas descritas y se añade:
  - <phonic_pause></pause>. Se emplea cuando el informante, en lugar de hacer una parada pronunciada e interrumpir el discurso, rellena el espacio con algún sonido fónico. Ejemplo: *Para mí, <phonic_pause>emm, emm</pause> tendría que avisar a la población.*
- Etiquetas léxico-semánticas. Se mantienen *<foreign lang="…">…</foreign>*, para las palabras extranjeras y *<sic type="calque">…</sic>*, para los calcos léxicos. Desaparece *<sic>…</sic>*, al tratar de crear etiquetas más concretas.
- Etiquetas de referencias. Se mantienen las tres etiquetas descritas y se añaden otras tres:
  - <rs desc="informante"></rs>. Se emplea para identificar al informante de la producción.
  - <rs desc="pregunta"></rs>. Se emplea para identificar la parte del discurso en la primera producción.
  - <rs desc="dilema"></rs>. Se emplea para identificar la parte del discurso en la segunda producción.

Con ello, se pretende dar respuesta a la mayor parte de fenómenos encontrados en las producciones de lengua extranjera.

Figura 14

Muestra del informante 1 con CATMA. Extracción del software



En la figura 14, se observa la nueva anotación del informante 1a, es decir, su primera producción. Aunque la herramienta no permite la codificación, parece una buena aplicación para ayuda visual en los datos y el control del etiquetado. Recogiendo las palabras de Chen (2005), CATMA permite la interpretación cómoda de los datos registrados. Hasta el momento, se están empleando cinco colores: rojo (identificación de informantes y producciones), rosa (anotaciones morfosintácticas), celeste (anotaciones de referencia), verde (anotaciones pragmático-discursivas) y amarillo (anotaciones léxico-semánticas).

Muñoz de la Virgen, Cristina (2024). Herramientas para la codificación, el análisis y la explotación de un corpus oral de aprendices francófonos de español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_inv03, 1-25. http://doi.org/10.1344/teisel.v3.44741

© La autora | Recepción: 06/11/2023 | Revisión: 18/01/2024 | Aceptación: 09/02/2024 | Publicación: 11/03/2024

ISSN 2696-676X

Conclusiones

A lo largo de este trabajo se ha presentado la creación, codificación y actualización de un corpus oral de aprendices francófonos de ELE, el cual se diseñó y codificó teniendo en cuenta la finalidad para la que se ideó, esto es, servir de soporte para el análisis de la interlengua de los informantes en relación con la adquisición del modo subjuntivo.

Como se ha visto, el enfoque de esta investigación está basado en datos (*data-driven research*) objetivos, replicables y pertinentes con relación al fenómeno estudiado (Rojo, Palacios, Sampedro y Marsily, 2023), los cuales no se podrían haber obtenido ni gestionado sin la digitalización y codificación del corpus oral. De este modo, las etiquetas insertadas manualmente en el texto permiten localizar y recuperar no solo la información más relevante para la investigación, sino también los errores cometidos por los informantes en el desarrollo de su producción oral, gracias a las herramientas mencionadas. Sin ellas, probablemente muchos de los datos se habrían perdido al no estar identificados y, naturalmente, habría sido imposible recuperarlos para dar un tratamiento integral al análisis de la interlengua.

Sin duda, las tecnologías empleadas en el corpus oral han permitido -y continúan haciéndolo- un tratamiento riguroso de los datos: *XML Formatter* y *XML Validator* aseguran la correcta configuración de la cabecera tanto del corpus como de la intervención de cada informante. A pesar de ser una parte no visible del corpus cuando se visualice en un navegador, resulta fundamental porque contiene toda la información de metadatos pertinente del corpus. Hay que señalar el productivo uso que la herramienta de *VS Studio Code* tiene en el corpus, ya que permite la correcta codificación del corpus y se ocupa de que las etiquetas que se añadan cumplan con los requisitos del lenguaje XML-TEI. Es una herramienta de fácil uso y personalizable con las necesidades de cada proyecto y cada investigador; además, posibilita un almacenamiento de datos ágil, gracias a su integración nativa con *Git* y con el control de versiones de *GitHub*. Estas dos herramientas también tienen un uso destacado en el corpus. La primera, *Git*, sirve de enlace, a través de los comandos mencionados en el apartado 5, entre el trabajo y la actualización del corpus en local y remoto, mientras que *GitHub* asegura que no se pierda la información y mantiene un control de versiones, es decir, la herramienta permite la recuperación de datos que ya han sido actualizados y almacenados, lo cual parece una excelente opción para los casos en los que se haya producido algún tipo de error. En un principio, se apostó, y se sigue apostando, por el mantenimiento del corpus oral en esta plataforma que, además del control de versiones, cumple con la misión de hacerlo público y accesible para otros investigadores tan pronto como la actualización finalice.

La modificación y actualización que se está realizando en el sistema del etiquetado y categorización de nuevos atributos hará posible que se pueda seguir explotando el corpus, de manera más funcional, en futuras investigaciones sobre la adquisición del subjuntivo y otros aspectos lingüísticos, que se ocupen del proceso de aprendizaje-adquisición del ELE. Estos cambios y la posibilidad de que el corpus tenga la capacidad de abrirse a nuevos estudios no serían posibles sin la incorporación de otras dos herramientas: *SpaCy* y *CATMA*. Esta última herramienta ayuda a mantener un control del nuevo etiquetado organizado visualmente mediante colores y categorías, como se ha visto en la figura 14.

Muñoz de la Virgen, Cristina (2024). Herramientas para la codificación, el análisis y la explotación de un corpus oral de aprendices francófonos de español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_inv03, 1-25. http://doi.org/10.1344/teisel.v3.44741

© La autora | Recepción: 06/11/2023 | Revisión: 18/01/2024 | Aceptación: 09/02/2024 | Publicación: 11/03/2024

ISSN 2696-676X

Igualmente, la herramienta ahonda en la posibilidad de realizar otro tipo de análisis, aunque, debido al incipiente estadio de actualización del corpus, aún no se ha profundizado en esta opción.

Por otro lado, *SpaCy* posibilita un etiquetado automático del aspecto morfológico para enriquecer el corpus con este anotado adicional, recuperación estadística de datos como índices de frecuencia o patrones de uso. Es una herramienta potente con la capacidad de procesar rápidamente grandes cantidades de datos y de sencilla instalación, tal y como se ha visto en el apartado 5, a través de *Google Colaboratory*. Parece que la combinación de esta herramienta y la codificación del corpus con el resto de las herramientas mencionadas tendrá la capacidad de dar respuesta a las investigaciones sobre la interlengua de los aprendices francófonos, posibilitando que el corpus oral continúe creciendo.

Finalmente, y de acuerdo con Barros García (2021, p. 231), la visualización de los resultados de las investigaciones en ELE debe contribuir a la accesibilidad de los nuevos conocimientos sobre el proceso de enseñanza-aprendizaje para mantener "la correspondencia real entre la lengua que se enseña y la lengua que se habla". En este sentido, el corpus oral de aprendices de francófonos de ELE cumple la función de conectar ambas "lenguas" y comprobar el proceso de asimilación y patrones de uso de determinados elementos y estructuras sintácticas por parte de los informantes.

Referencias

Barros García, Benamí (2021). Representar visualmente los resultados de la investigación sobre el español LE/L2. En Mar Cruz Piñol (Ed.), *e-Research y español LE/L2. Investigar en la era digital* (pp. 138-163). Routledge. https://doi.org/10.4324/9780429433528

Briz, Antonio (1998). *El español coloquial en la conversación: esbozo de pragmagramática*. Ariel.

Chen, C. 2005. "Top 10 Unsolved Information Visualization Problems". *IEEE Computer Graphics and Applications, 25*(4), 12-16. http://doi.org/10.1109/MCG.2005.91

Fradejas Ruedas, José Manuel (2009-2010). La codificación XML/TEI de textos medievales. *Memorabilia*, *12*, 219-247). https://parnaseo.uv.es/Memorabilia/Memorabilia12/PDFs/Codificacion.pdf

Instituto Cervantes (2014). *Corpus de aprendices de español como lengua extranjera*. https://galvan.usc.es/caes

Llisterri, Joaquim (2021). Corpus para investigar sobre el componente fónico en español como LE/L2. En Mar Cruz Piñol (Ed.), *e-Research y español LE/L2. Investigar en la era digital* (pp. 164-196). Routledge. https://doi.org/10.4324/9780429433528

Lozano, Cristóbal; Mendikoetxea, Amaya (2013). Learner Corpora and Second Language Acquisition: The Design and Collection of CEDEL2. En Ana Díaz-Negrillo, Nicolas Ballier y Paul Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data* (pp. 65–100). John Benjamins. https://doi.org/10.1075/scl.59.06loz

Lozano, Cristóbal (2021). Corpus textuales de aprendices para investigar sobre la adquisición del español LE/L2 En Mar Cruz Piñol (Ed.), *e-Research y español LE/L2. Investigar en la era digital* (pp. 138-163). Routledge. https://doi.org/10.4324/9780429433528

Muñoz de la Virgen, Cristina (2024). Herramientas para la codificación, el análisis y la explotación de un corpus oral de aprendices francófonos de español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_inv03, 1-25. http://doi.org/10.1344/teisel.v3.44741

© La autora | Recepción: 06/11/2023 | Revisión: 18/01/2024 | Aceptación: 09/02/2024 | Publicación: 11/03/2024

ISSN 2696-676X

Lozano, Cristóbal (2022). CEDEL2: Design, compilation and web interface of an online corpus for L2 Spanish acquisition research. *Second Language Research*, *38*(4), 965-983. https://doi.org/10.1177/02676583211050522

MECD (Ministerio de Educación, Cultura y Deporte) (2002). *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación*, Instituto Cervantes / Anaya. https://cvc.cervantes.es/ensenanza/biblioteca_ele/marco/

McEnery, Tony y Wilson, Andrew (2001). *Corpus Linguistics*. Edinburgh University Press.

McEnery, Tony; Hardie, Andrew (2012). *Corpus Linguistics: Method, theory and practice*. Cambridge University Press.

Saint Laurent, Simon y Michael Fitzgerald (2005). *XLM. Pocket Reference*. O'Reilly.

Sinclair, John (2005). How to Build a Corpus. En Martin Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 79-83). Oxbow Books.

Rojo, Guillermo; Palacios, Ignacio; Sampedro, María; Marsily, Aurélie (2023). Los corpus de aprendices de español LE/L2: panorama actual y perspectivas futuras. *Journal of Spanish Language Teaching*, *9*(2), 174-189. https://www.tandfonline.com/doi/full/10.1080/23247797.2022.2157085

TEI CONSORTIUM (2023). *TEI P5: guidelines for electronic text encoding and interchange. Version 4.6.0.* Text Encoding Initiative Consortium. https://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf

Muñoz de la Virgen, Cristina (2024). Herramientas para la codificación, el análisis y la explotación de un corpus oral de aprendices francófonos de español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_inv03, 1-25. http://doi.org/10.1344/teisel.v3.44741

© La autora | Recepción: 06/11/2023 | Revisión: 18/01/2024 | Aceptación: 09/02/2024 | Publicación: 11/03/2024