Alumna: Margarita Maturana

Análisis computacional de datos lingüísticos

Profesor: Javier Vera

Adverbios evidenciales: ocurrencia y coocurrencia en tres corpus del español.

1. Presentación del problema y objetivos

La evidencialidad es un rasgo gramatical, cuya obligatoriedad de uso varía de lengua en lengua. Si bien el fenómeno ha sido ampliamente estudiado (Chafe y Nichols (1986), Givón (2001), Aikhenvald (2003), De Haan (2005), etc.), aún es objeto de estudio el funcionamiento sistemático de la evidencialidad para el español. Este procedimiento “no atañe a cualquier marcador de sinceridad, sino a aquellos que en su uso contextual destacan la objetividad y evidencia de lo que se está diciendo, y, por tanto, su fuente enunciativa deja de ser la del yo-hablante.” (Albelda y Cestero, 2011, p. 13). Existen variados mecanismos en la lengua que codifican, entre muchos otros, un valor evidencial; determinados tiempos verbales, algunos verbos, adverbios, ciertas construcciones propias del idioma, etc.

Un grupo especial de palabras en el español son los adverbios. Este gran grupo de palabras aún genera dudas dentro de los gramáticos, ya que en la categoría ‘adverbio’ se agrupan una serie de palabras que poseen funcionamientos y significados muy variados. Si a esto se le suma que el sufijo ‘-mente’ puede ser agregado a casi cualquier adjetivo convirtiéndolo en adverbio, el resultado es una mezcolanza de elementos.

Bajo las problemáticas categoriales anteriormente mencionadas, este estudio se levanta como un estado de la cuestión respecto a los adverbios evidenciales terminados en –mente en determinados corpus del español. Para esto, se contrastará la ocurrencia de adverbios evidenciales en tres corpus: CORDIAL, CORLEC y PRESEEA. Para ello, se debe caracterizar los elementos adyacentes a los adverbios evidenciales con el fin de identificar la existencia –o no- de patrones de uso para estos adverbios. Finalmente, se identificará su lugar en la oración (A-B-C). Este último paso se vuelve pertinente ya que se ha asociado fuertemente el uso de evidenciales con la fuerza ilocutiva del enunciado y/o la modificación total del valor de verdad de la cláusula, no obstante “afirmar que existen diferentes ámbitos de actuación nos permitiría dejar de lado la idea extendida en la bibliografía de que todos los contenidos evidenciales operan en el nivel supraoracional ilocutivo” (González Vázquez, 2016, pp. 56).

Finalmente, cabe señalar que se utilizaron las categorías propuestas por Torner (2005) para identificar los adverbios cuyo contenido semántico posea una función evidencial. Si bien, el autor propone que solo los reforzadores del valor de verdad (“obviamente”, “incuestionablemente”, “evidentemente”, etc.) se relacionan con el valor evidencial, para esta exploración se consideraron también los restrictores del valor de verdad (“aparentemente”, “presumiblemente”, “presuntamente”, etc.).

1. Metodología

En primer lugar, se realizó un sondeo de los corpus de español con acceso liberado en internet: CDE, CORDE, Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico, CHILDES, TalkBank, Corpus de Referencia del Español Actual, C-ORAL-ROM, cespla, CORPES, por mencionar algunos. Si bien la mayoría ofrece formas de consulta online en sus corpus, solo algunos poseían la opción de descargar los textos completos. Desde el repositorio <https://github.com/djeastm/Spanish_Corpus_Analysis_Project> se pueden obtener muestras de algunos de los corpus ya mencionados. Para este trabajo se seleccionaron el CORDIAL (Corpus Oral Didáctico Anotado Lingüísticamente), CORLEC (Corpus Oral de Referencia de la Lengua Española Contemporánea) Y PRESEEA (Proyecto para el estudio sociolingüístico del español de España y América).

Una vez seleccionados los corpus, el primer paso es preprocesar los textos que se encuentran en formato .txt. El preprocesamiento consiste en pasar todas las palabras a minúscula, eliminar caracteres y espacios en blanco sobrantes, dividir el texto en oraciones, y este a su vez en palabras. La división anteriormente desarrollada permite contar la cantidad de palabras y oraciones de cada corpus, lo que más adelante permite encontrar tokens y types.

Una vez hecho esto, se utiliza la librería Spacy para etiquetar las palabras con información gramatical. Este etiquetado se utiliza para identificar la cantidad de adverbios totales, pero, además, otorga información sobre las categorías de las demás palabras. Se genera un diccionario cuyos keys son los adverbios terminados en –mente, y sus values, la cantidad de apariciones en el texto. Se usa un segundo diccionario que contiene sólo los adverbios evidenciales objeto de estudio de la presente investigación, para así poder calcular la razón entre total de adverbios, adverbios terminados en –mente y adverbios evidenciales terminados en –mente.

Finalmente, se tabula y grafica la información con Pandas y Matplotlib con la finalidad de hacer más amigable la observación de datos.

Existe un problema metodológico relacionado a la posibilidad de demostrar computacionalmente que los adverbios en estudio están, efectivamente, operando como evidenciales. Aunque es una problemática engorrosa, Lee (2021) propone el uso de text mining para realizar análisis crítico del discurso. El autor señala que las técnicas de análisis computacional deben ser incorporadas al ACD ya que “uncovering the underlying topical structure of textual data” (2021, p. 83). Mediante la identificación de colocaciones, co-ocurrencias y topic modeling, el autor genera redes semánticas que permiten identificar nociones asociadas a la palabra “migrante”. Esta metodología será replicada en este trabajo con el fin de identificar el contexto oracional de los adverbios evidenciales. El contexto nos permitiría identificar el uso evidencial ya que la tendencia hasta ahora observada indica que los evidenciales aparecen en oraciones donde existen más elementos que modalizan lo dicho.

1. Descripción de los datos:

3.1. CORDIAL

3.1.1. Descripción general

El Corpus Oral Didáctico Anotado Lingüísticamente es un proyecto que transcribe 10:19:36 horas de grabación, correspondientes a 240 sesiones de texto y audio. Se realiza en la ciudad de Madrid, y los tipos de texto que lo constituyen se distribuyen de la siguiente manera: Espontáneos 30%, espontáneos semidirigidos 68 %, formales 2% Variedad geográfica Recogido (Martínez, 2012).

El conteo de tokens indica que el texto posee 118.592 palabras, de las cuales 10.220 son types. La razón entre types y tokens es 0.08617781975. Posee 8.619 oraciones, y las palabras que las componen están distribuidas de la siguiente manera:

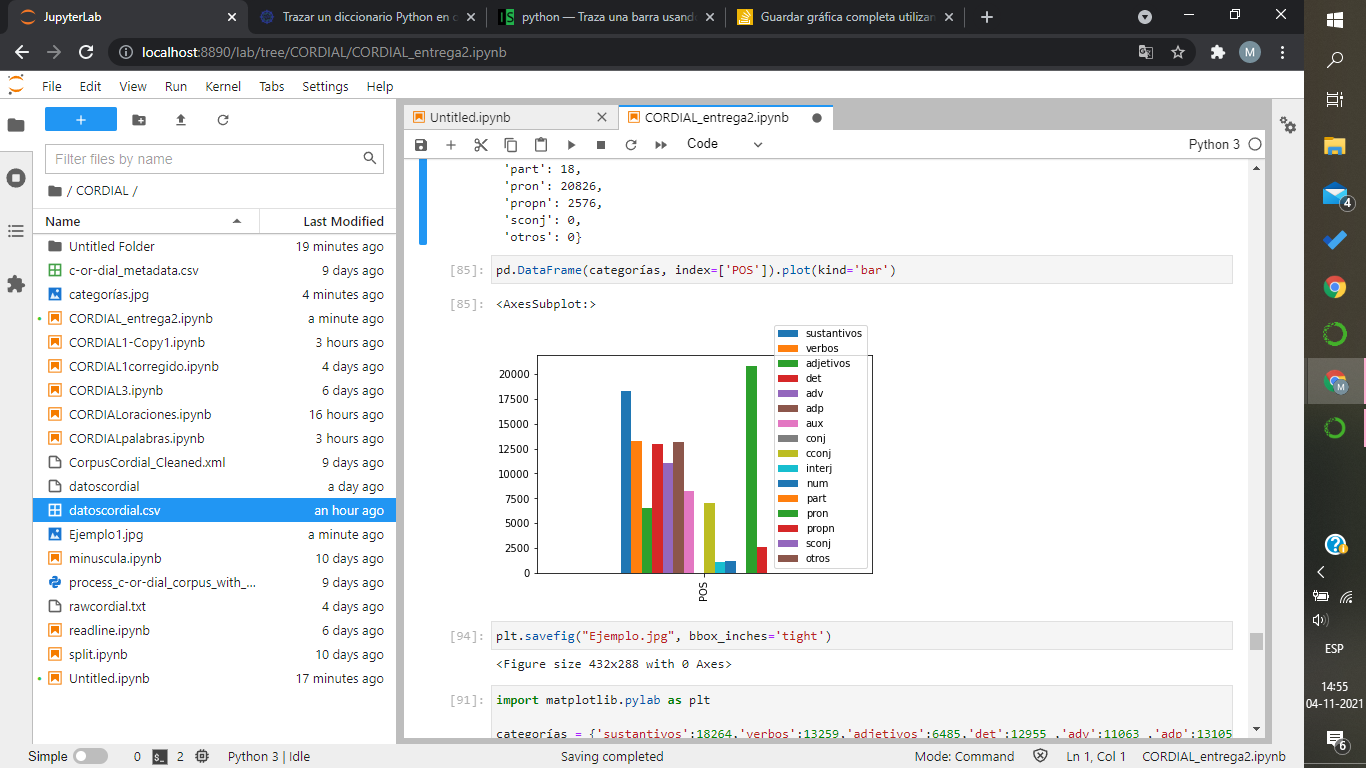


Tabla 1

Distribución de palabras en CORDIAL

3.1.2. Adverbios

El etiquetado realizado por Spacy identificó 11.063 adverbios, de los cuales, 560 coinciden morfológicamente por su terminación en –mente. Este total fue convertido en un diccionario para identificar cuyos keys son los adverbios sin repetirse, y los values la cantidad de veces que aparecen. La información puede observarse en la tabla 2:

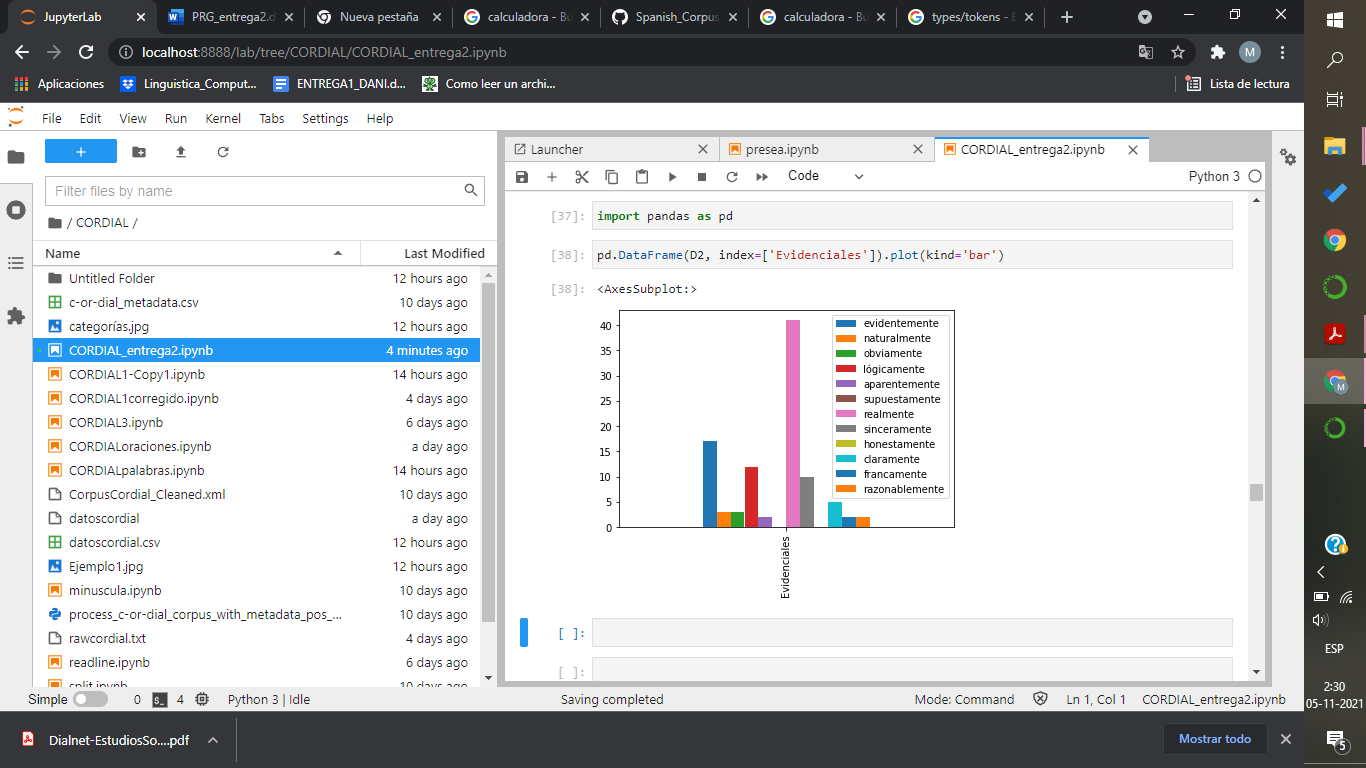


Tabla 2

Adverbios evidenciales en CORDIAL

1. Proyecciones

Todavía queda por pre-procesar e indagar en los corpus CORLEC y PRESEEA, ya que estos presentan una mayor dificultad para ser limpiados. Así mismo, queda pendiente aplicar text mining analysis en las oraciones con adverbios evidenciales. De estas mismas oraciones, se pueden estudiar sus colocaciones y coocurrencias.

1. Bibliografía

Aho, A. V. (1991). Algorithms for finding patterns in strings, Handbook of theoretical computer science (vol. A): algorithms and complexity. MIT Press, Cambridge, MA.(RE)

Albelda, M. & Cestero, A. M. (2011). De nuevo, sobre los procedimientos de atenuación lingüística. *Español actual: Revista de español vivo*, Vol. 96, 9-40.

Gónzalez Vázquez, M. (2016). La naturaleza y función de la evidencialidad en español. *Lingüística iberoamericana*, Vol. 60, 49-74.

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Martínez, C. N. (2012). C-Or-Dial:(corpus oral didáctico anotado lingüísticamente). Liceus, Servicios de Gestión y Comunicación.

pandas-dev/pandas: Pandas (Nº de versión latest). (2020). Windows. Zenodo.

Torner Castells, S. (2005). Aspectos de la semántica de los adverbios de modo en español. Universitat Pompeu Fabra.