

Adverbios evidenciales: ocurrencia y frecuencia en dos 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. Continúan se presentan 3 ejemplos de oraciones donde los mecanismos con valor evidencial se señalan en cursiva:

- i. *Por lo visto* no habrá navidad este año
- ii. *Resulta* que el furgón es más caro que el colegio
- iii. *Obviamente* apagué el gas antes de salir.

En iii, el mecanismo utilizado corresponde a un adverbio. Esta clase de palabras ha sido ampliamente discutida en la bibliografía ya que “se trata de una categoría muy heterogénea, cuya caracterización resulta conflictiva desde diversos puntos de vista, ya sea en lo relativo a la definición del concepto mismo de adverbio y a la tipología adverbial, a su descripción morfológica o a su sintaxis” (Torner, 2016, p. 380). Uno de los problemas categoriales en torno a los adverbios surge si se considera que la morfología del español permite adverbializar adjetivos –usualmente calificativos- mediante la adición de la terminación *-mente*. Debido a esto, esta categoría se encuentra constantemente en crecimiento, volviendo compleja la tarea de delimitar las funciones específicas de los adverbios. Aproximación computacional a este fenómeno otorga la posibilidad de procesar grandes cantidades de texto, para la identificación y descripción de oraciones que posean adverbios con valor evidencial. Estudiar el contexto en que se usan los adverbios evidenciales permite ilustrar las redes semánticas que se construyen en torno a estos modificadores oracionales.

Una de las clasificaciones más aceptadas es la que se refiere a la distinción del elemento oracional modificado por el adverbio, considerando que este “es externo al significado de la proposición; esto es, no se interpreta en relación con la predicación, sino que modifica la oración tomada como un todo” (Torner, 2016, p. 387). De la batería funcional que proponer Torner (2016) para clasificar los adverbios, el presente estudio se propone describir en dos

corpus del español -CORDIAL¹ y PRESEEA²- los contextos de ocurrencia de tres clases de adverbios, enlistados a continuación:

Adverbios de modalidad:

a) Reforzadores y restrictivos del valor de verdad:

- **Intensionales (restrictivos del valor de verdad):** *Aparentemente, hipotéticamente, nominalmente, presumiblemente, presuntamente, pretendidamente, supuestamente, teóricamente, virtualmente.*
- **Evidenciales (reforzadores del valor de verdad):** *Ciertamente, evidentemente, incuestionablemente, indiscutiblemente, indudablemente, obviamente, realmente, verdaderamente.*

Adverbios de enunciación:

a) **Orientados hacia el hablante:** *Francamente, honestamente, honradamente, sinceramente.*

Para el autor, sólo los reforzadores de verdad poseen valor evidencial, no obstante, en la presente investigación se consideran también los orientados al hablante y los intensionales, ya que en otros estudios -Rodríguez Ramalle (2017), Figueras (2017), Rodríguez Espiñeira (2017) ~~en La evidencialidad en español: teoría y descripción- aúnan~~ estos adverbios para estudiar su posible valor evidencial.

Para hacer una descripción más adecuada, primero se cuantifican los adverbios evidenciales, para luego contrastar la ocurrencia de estas palabras en los dos corpus previamente mencionados. Posteriormente, se caracterizan los elementos adyacentes a los adverbios evidenciales con el fin de identificar la existencia –o no- de patrones de uso para estos adverbios.

1. Metodología

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 (2019) 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” (2019, 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.

2. Metodología

¹ <http://lablita.it/app/cordial/corpus.php>

² <https://preseea.linguas.net/>

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 poseen la opción de descargar los textos completos. En [GitHub](https://github.com/djeastm/Spanish_Corpus_Analysis_Project)³ se pueden encontrar muestras de algunos de los corpus ya mencionados. Para este trabajo se seleccionaron el CORDIAL (Corpus Oral Didáctico Anotado Lingüísticamente) y PRESEEA (Proyecto para el estudio sociolingüístico del español de España y América).

2.1. Pre-procesamiento de los corpus.

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 –con *Regular expressions*⁴–, dividir el texto en oraciones –con *String*⁵–, 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 el número de *tokens* y *types*. *Types* son las palabras que ocurren sólo una vez en el corpus, mientras que *tokens* es el número total de palabras. Estas cifras nos permiten obtener la relación *type/token* (TTR), una medida simple para identificar la complejidad de un texto (Kettunen, 2014).

2.2. Búsqueda y etiquetado de datos.

Se vuelve a usar *re* con el fin de contabilizar todos los adverbios terminados en *–mente*, y a su vez los adverbios evidenciales seleccionados para este estudio. La finalidad de esto es identificar la proporción entre el total de adverbios construidos con *–mente* vs el total de adverbios evidenciales. Se generan entonces dos diccionarios, uno donde se almacenan todos los adverbios terminados en *–mente* y sus ocurrencias. El otro, con las ocurrencias de los adverbios evidenciales.

A continuación, se usa *Networkx*⁶ para identificar las co-ocurrencias de los evidenciales en sus oraciones, tomando las 5 palabras más cercanas, estén estas delante o detrás en la cadena de palabras. Con *Matplotlib*⁷ se grafican estas co-ocurrencias para visualizar qué palabras son más usadas con determinados adverbios.

³ https://github.com/djeastm/Spanish_Corpus_Analysis_Project

⁴ <https://docs.python.org/3/library/re.html>

⁵ <https://docs.python.org/3/library/string.html>

⁶ <https://networkx.org/documentation/stable/tutorial.html>

⁷ <https://matplotlib.org/2.0.2/index.html>

Una vez hecho esto, se utiliza la librería *Spacy*⁸ para etiquetar las palabras con información gramatical. Este etiquetado se utiliza para identificar la cantidad de adverbios totales – terminen o no en *–mente–*.


3. Descripción de los datos.

En este apartado se presenta una descripción de los corpus seleccionados. La tabla 1 contiene información respecto al tipo de registro, zona geográfica donde se tomó la muestra, *tokens* y *types* –junto al TTR–, y la cantidad de oraciones totales:

	CORDIAL	PRESEEA
TIPO DE REGISTRO	Esponáneos: 30% Esponáneos semidirigidos: 68 % Formales: 2%	Entrevista semidirigida: 100%
¿DÓNDE SE OBTUVO LA MUESTRA?	Madrid	España y América Latina. Principales ciudades (50 regiones diferentes)
TOKENS	118.592	1.581.464
TYPES	10.220	62.864
TTR	0.086	0.039
CANTIDAD DE ORACIONES	8.619	44.418

Tabla 1.

Descripción de los corpus CORDIAL y PRESEEA.

La primera observación que surge es el tamaño dispar de las muestras, ahora bien, cabe acar que el contraste realizado en este estudio no es de carácter cuantitativo, sino en cuanto al sentido de uso de los adverbios evidenciales. Es por ello que se trabaja con ambos corpus a pesar de su diferencia, ya que la finalidad es describir los contextos de ocurrencia, independiente de la cantidad de ocurrencias. Otra observación interesante, es el bajo valor que el TTR arroja sobre los textos. En la actualidad todavía se debate si el TTR es en realidad un cálculo válido para identificar la variabilidad léxica de un texto (Kettunen, 2014), no obstante, el carácter oral de ambos corpus podría explicar el valor tan bajo que arrojó el cálculo. Aunque el lenguaje en las entrevistas –se presume– es más cuidado que en registros espontáneos, el discurso oral en pocos casos posee preparación, por ello, la alta repetición de palabras se vuelve un fenómeno esperable. En las figuras 1 y 2, se muestra la distribución de clases de palabras en cada corpus estudiado:

⁸ <https://spacy.io/usage>

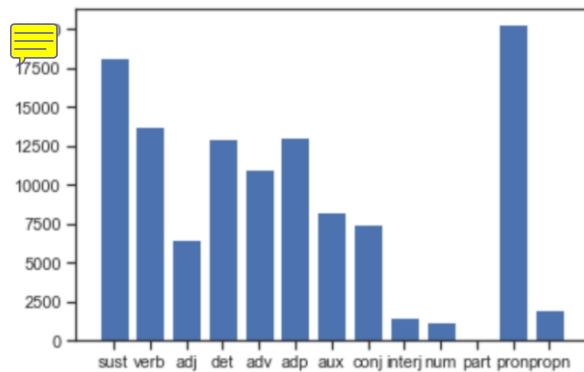


Fig 1.

Distribución clases de palabras en CORDIAL

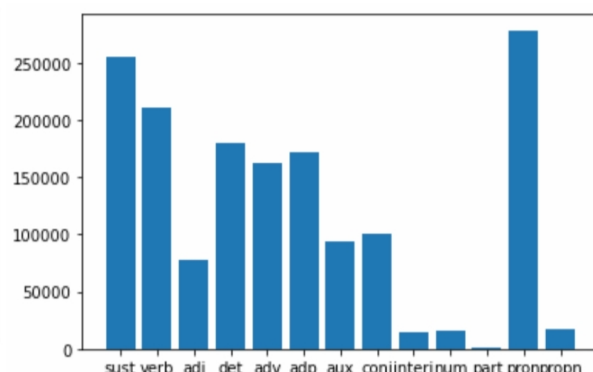


Fig2.

Distribución clases de palabras en PRESEEA

La distribución entre ambos corpus es bastante similar. En el corpus PRESEEA los adverbios constituyen un décimo del total de palabras, y en CORDIAL esta proporción se mantiene. La segunda tabla se limita a mostrar la información en torno a la categoría Adverbio: Se describen *tokens* y *types* de todos los adverbios encontrados, *tokens* y *types* de los adverbios terminados en *-mente*, ocurrencias de adverbios evidenciales y oraciones que contienen adverbios evidenciales:

	CORDIAL	PRESEEA
ADVERBIOS	11.000 - 267	161.869 - 1.735
ADV. TERMINADOS EN -MENTE	560 - 134	5.285 - 449
ADV. EVIDENCIALES	94	786
ORACIONES CON ADV. EVID.	91	714

Tabla 2.

Descripción de los adverbios buscados en CORDIAL y PRESEEA.

Las muestras se mantienen similares en cuanto a proporciones de uso, ya que en los dos corpus un sexto de los adverbios es utilizado con valor evidencial. Se observa, además, que la cantidad de adverbios evidenciales es mayor que la cantidad de oraciones con Adv. Evid. Cuestión que implica que en más de un ejemplo se puede encontrar más de un evidencial por oración.

4. Análisis de co-ocurrencias

Finalmente, en la tabla 3 se detalla la cantidad de ocurrencias para cada adverbio en estudio y las palabras más cercanas: En rosado evidenciales, celeste para intensionales, morado para orientados hacia el hablante:

	CORDIAL		PRESEEA	
Ciertamente	0		3	'También', 'con', 'empezaron', 'pero', 'y'
Evidentemente	17	'Pues', 'no', 'porque', 'pero', 'es'	46	'Pues', 'pero', 'sí', 'porque', 'y'
Indiscutiblemente	0		6	'Que', 'sea', 'pues', 'necesitas', 'eeh'

Indudablemente	2	'Hhh', 'en', 'que', 'la'	18	'Hay', 'que', 'pero', 'sí', 'la'
Obviamente	3	'Idea', 'de', 'sabían', 'hacer', 'lugar'	92	'Pues', 'que', 'no', 'y', 'bueno'
Realmente	41	'Es', 'no', 'porque', 'que', 'todo'	501	'Que', 'no', 'y', 'pues', 'porque'
Verdaderamente	14	'Yo', 'porque', 'que', 'supone', 'una'	21	'Que', 'era', 'necesita', 'me', 'te'
Aparentemente	2	'Pero', 'eso', 'rentabilidad', 'palpable'	8	'Hacía', 'decidió', 'por', 'actitud', 'respetuosa'
Supuestamente	0		37	'Que', 'y', 'ahí', 'ya', 'en'
Teóricamente	1		2	'Desconocidos', 'que', 'aunque', 'teo', 'ahora'
Francamente	2	'Tiempo', 'y', 'o'	5	'Pues', 'también', 'a', 'recuerdo', 'por'
Honestamente	0		10	'No', 'hijos', 'que', 'bien', 'nos'
Sinceramente	10	'O', 'creo', 'pues', 'yo', 'no'	37	'Digo', 'no', 'sí', 'que', 'te'

Tabla 3.

Ocurrencias y co-ocurrencias de adverbios evidenciales, intensionales y orientados al hablante en CORDIAL y PRESEEA.


El primer comentario sobre la tabla 3 radica en la ausencia de algunos adverbios propuestos en el punto 1. En los 3 sub grupos de adverbios sucedió que hubo adverbios– honradamente, francamente, nominalmente, presumiblemente, pretendidamente, incuestionablemente- que no presentaban realizaciones. En segundo lugar, se encuentra la alta presencia de la conjunción subordinante *que* acompañando a los evidenciales en ambos corpus. La presencia de esta partícula es señal de que las oraciones generadas con estos adverbios corresponden a estructuras complejas del español, ya que esta conjunción suele ser introductora de oraciones subordinadas o relativas. Puede observarse también la marcada presencia de otras conjunciones: *pero*, *pues*, *y*, *o*. Las cuales poseen una función sintáctica coordinante dentro de la oración, o semánticamente, expresan cierta causa-efecto. Muchos de estos adverbios se refieren al valor de verdad de la proposición en la medida que lo dicho se considera una verdad compartida por los hablantes, entonces, el uso de estas palabras acompaña esta noción de lo que dicho es algo natural, observable, deducible del mundo.


La presencia del pronombre personal *yo* se repite en *sinceramente*-orientado al hablante- y *verdaderamente*-evidencial puro-. En un idioma como el español donde el sujeto explícito es optativo, se hace relevante comentar por qué dos adverbios que difieren en función propician en los hablantes el uso de sujeto explícito, especialmente, un sujeto que involucra directamente al hablante en lo dicho. Una inferencia es que la verdad y la sinceridad son semánticamente muy cercanas, no obstante, esta idea no sirve para explicar porque *honestamente* no propicia la implicación directa del hablante en lo dicho.

Un caso especial de comentar es el del adverbio *evidentemente*, ya que las co-ocurrencias fueron casi similares en ambos corpus. Debido a que estos adverbios obedecen a procesos de creación de la lengua (Adj+’-mente’), es posible que algunos de estos evidenciales posean


‘reglas’ de uso más estandarizadas que otros. Si en dos corpus distintos en cuanto a registros y zonas geográficas se repite la forma de usar una palabra, es pertinente pensar que los hablantes poseen nociones similares sobre el uso y significado del evidencial *evidentemente*.

5. Conclusiones, problemas y proyecciones

Sobre lo observado en esta investigación, cabe señalar que las categorías hasta ahora planteadas e los adverbios van bien encaminadas, no obstante, los límites son difusos debido a las múltiples opciones de uso que ofrece la lengua. Si se estudian los contextos que rodean las palabras los sentidos que estas poseen pueden verse modificados. En cuanto a la cantidad de adverbios puramente evidenciales –es decir, los reforzadores del valor de verdad– que ocurren en los corpus, se puede apreciar que su frecuencia de uso es alta, lo que mantiene el foco de estudio sobre la sistematicidad de la evidencialidad.

Cabe señalar algunos  que surgieron durante el proceso. En primer lugar, aunque los corpus fueron pre-procesados, en los resultados finales aparecen elementos que no son palabras (hhh). En segundo lugar, debido a que son transcripciones provenientes de la oralidad, la puntuación no siempre es transparente. Hay oraciones cortadas, cuyo sentido no logra recuperarse, como también existen oraciones extremadamente largas. En el etiquetado con *Spacy*, cuando se revisó manualmente la asignación palabra-categoría, de 100 entradas, 3 poseían error.

Sobre los adverbios descritos en la bibliografía que no aparecieron en los corpus – honradamente, francamente, nominalmente, presumiblemente, pretendidamente, incuestionablemente– queda la tarea pendiente de revisar corpus basados en textos escritos, para identificar si este grupo posee más frecuencia en otros contextos.

Retomando el valor evidencial que se codifica en el español, cabe recordar que los adverbios evidenciales corresponden a un tipo de mecanismo lingüístico de una lista más extensa (vale decir, verbos con valor evidencial, construcciones, etc.). Este dominio aún está en estudio, y una aproximación computacional al panorama completo es uno de los caminos que los estudios gramaticales y el procesamiento de texto podrían seguir ro camino que mezcla gramática diacrónica y análisis de datos mediante computadora puede ser la reconstrucción de los procesos de gramaticalización de los adverbios terminados en *–mente*.

6. Bibliografía

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

González Ruiz, R., Izquierdo Alegría, D. y Loureda Lamas, Ó. (eds.) (2016): La evidencialidad en español: teoría y descripción, Madrid y Fráncfort del Meno, Iberoamericana/ Vervuert.

Kettunen, K. (2014). Can type token ratio be used to show morphological complexity of languages?. *Journal of Quantitative Linguistics*, 21(3), 223-245.

Lee, C. (2019). How are 'immigrant workers' represented in Korean news reporting?—A text mining approach to critical discourse analysis. *Digital Scholarship in the Humanities*, 34(1), 82-99.

Torner, S. (2016). Adverbio. In *Enciclopedia de Lingüística Hispánica* (pp. 380-392)