TRANSFORMACIONES EN SECUENCIAS NUMÉRICAS PARA EL ESTUDIO DE UNIDADES LINGÜÍSTICAS Y HABLANTES EN CORPORA LINGÜÍSTICOS

Perman, Santiago
Departamento de Letras, FHUC, UNL
Director: Horacio Miguel, Dotti

Área: Humanidades

Palabras clave: cuantificadores, secuencias, corpus

INTRODUCCIÓN

El presente texto abordará formas de obtener secuencias ordenadas de valores numéricos, a partir de textos verbales escritos o transcritos desde su origen oral. También se indicará su interés para diversas áreas de la investigación lingüística, como la psicolingüística. La exploración de estas transformaciones ocurre en el marco de una investigación léxica y sintáctica sobre los cuantificadores – como *todos*, *muchos*, *algunos*, *un*, *la mitad de*, etc. – en el discurso de menores en etapa escolar, especialmente menores con Trastorno de Desarrollo del Lenguaje (TDL) en contraste con aquellos de desarrollo típico, i.e. sin afecciones de desarrollo, teniendo en cuenta variables como el nivel socioeconómico, la edad y el sexo. Lo que se presentará a continuación es solamente la porción de una investigación mayor ligada a un proyecto de doctorado y a adscripciones pasadas, y centrada en los cuantificadores y las dificultades de lenguaje. Por lo tanto, no se desarrollarán todos los enfoques y metodologías sobre los objetos de estudio, sino varios pertinentes al tema del título.

En esta ocasión, bajo la hipótesis de que los cuantificadores varían en costo de procesamiento cognitivo según su complejidad semántica y morfológica, se podría esperar alguna tendencia de patrones antecediendo o procediendo la aparición de dichos ítems léxicos en la secuencia producida a partir de un texto. Si la tendencia no fuera registrada, las conclusiones preliminares son también interesantes. Pues o las propiedades de los cuantificadores poseen una influencia limitada al muy corto alcance sobre el resto de los elementos, o su costo de procesamiento no se traslada a la etapa de linearización del procesamiento lingüístico, o las dimensiones de las métricas, en que los cuantificadores ciertamente se encuentran como los otros ítems, no aparentan particularidades al respecto.

OBJETIVOS

-Aportar evidencias sobre patrones relativos a la aparición de cuantificadores en mediciones textuales.

Título del proyecto: TRANSFORMACIONES EN SECUENCIAS NUMÉRICAS PARA EL ESTUDIO DE UNIDADES LINGÜÍSTICAS Y HABLANTES EN CORPORA LINGÜÍSTICOS.
Instrumento: CAID
Año de convocatoria: 2023
Organismo financiador: UNL
Directora: Hracio Dotti

-Aportar nuevas métricas en relación con métricas ya establecidas para dilucidar las dimensiones del lenguaje que representa cada una.

METODOLOGÍA

La metodología predominante en esta dimensión de análisis es cuantitativa. Ello no quita la existencia de criterios cualitativos para la elección de unidades de análisis y la producción de inferencias. Por el contrario, las transformaciones parten de una abstracción del fenómeno lingüístico, en este caso de elementos categoriales a numéricos, a partir de alguna propiedad que podríamos caracterizar como cualitativa, por ejemplo, el reemplazo de palabras por sus frecuencias de aparición. Las abstracciones así producidas conllevan la pérdida de información sobre la multiplicidad de significados, referencias, circulación, interpretaciones para pasar a un simple vector de valores, o varios. Los abordajes cuantitativos tienen relativas limitaciones en el tratamiento de relaciones entre elementos cuya vecindad no es inmediata y, crucialmente, no parecen captar puntualmente la existencia de elementos y operaciones implícitas a las facetas sensoriomotoras del lenguaje, i.e. lo expresado. Pero a su vez implican una focalización y formalizaciones de otra forma inexistentes, las cuales, eventualmente, pueden servir a inferencias interesantes sobre el uso del lenguaje por un tipo de sujeto, las características de una corriente estética, o sobre el uso de una unidad lingüística, hasta inclusive del lenguaje como tal.

Sabemos que diferencias de resultados a partir de métricas sirven para diferenciar sujetos en psicolingüística, o como indicios para atribuir autorías a textos anónimos. Esos resultados suelen consistir en un único número, un resultado global o parcial. En su conjunto, los resultados de distintas muestras pueden ser utilizados con un test estadístico o un modelo de regresión para decidir si las cifras son lo suficientemente claras, con la subjetividad y arbitrariedad que dicho proceso implica. Pero si el tratamiento cuantitativo en lingüística es marginal y frecuentemente ancilar, menos común es un tratamiento donde el investigador indaga propiamente sobre la descripción cuantitativa del objeto, su desenvolvimiento y las formas, si las hay, que se generan al aplicar una transformación. Consecuentemente, más escasas son las explicaciones acerca de la aparición de patrones. Por ello, en esta dimensión de análisis procuramos inicialmente un abordaje descriptivo y gráfico que, eventualmente, podría aportar información tanto sobre dichos patrones como sobre las características de los sujetos, o las unidades lingüísticas que involucran.

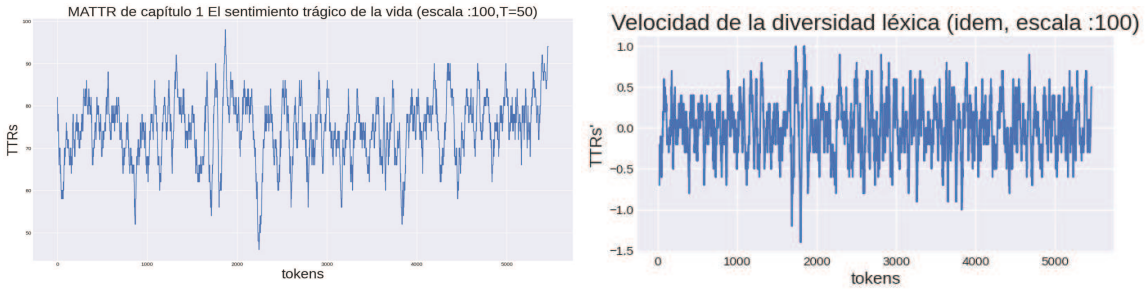
Por cuestiones de espacio, solamente trataremos del MATTR para la diversidad léxica (Covington y McFall, 2010) y los motivos [motifs] (Köhler, 2015). Ambas transformaciones tienen la ventaja de requerir poco acondicionamiento de las muestras, faceta que suele ser la más ardua en la pesquisa cuantitativa.

El primero (MATTR) es utilizado en psicolingüística y neurolingüística, con sujetos con, por ejemplo, afasias, trastorno de desarrollo del lenguaje (TDL) (Charest et al. 2020), o en estudios de adquisición del lenguaje. El TDL, uno de los objetos de la investigación más amplia en que se enmarca la presente, se caracteriza por dificultades y retrasos en la utilización del lenguaje, sea en la semántica, sintaxis, pragmática, etc. que pueden extenderse hasta la adultez (Bishop et al., 2016). Ahora bien, para calcular el MATTR se toma recursivamente de un texto una selección continua de palabras de tamaño T, en el orden lineal. Existe un seleccionador W de palabras contiguas que selecciona porciones de un texto. Las palabras dentro de esa selección serán entendidas en dos formas: los *tokens*

son todas y cada una de las palabras expresadas en el enunciado; los *types* son las especies abstractas de las palabras. Por ejemplo, “perro, gato, perro” es un enunciado con tres tokens y dos types, pues “perro” se repite. El ratio de la cantidad de types sobre la cantidad de tokens es el TTR, un número entre 0.0 y 1.0; mientras más cercano a 1.0, más diversidad léxica. Una vez que el TTR es obtenido, se guarda en una lista y *W* se mueve un número *P* de pasos en la dirección natural del texto, normalmente uno. Se vuelve a realizar las mismas operaciones hasta que el texto ha sido recorrido prácticamente en su totalidad. Tenemos entonces una lista con los valores de cada TTR en forma ordenada, una secuencia. El MATTR es el promedio de todos los TTRs, pero también nos interesa el vector con sus valores secuenciales. En pseudocódigo:

```
texto = “A A B C D A C”
T = 3
W = seleccionador()
TTRs = []
W(texto, T)
-> [A A B] C D D D :: TTRs[0.66]
->A [A B C] D D D :: TTRs[0.66, 1.0]
...
-> A A B C [D D D] :: TTRs[0.66, 1.0, 1.0, 0.66, 0.33]
TTRs.promedio()
->0.73
```

Aquí tenemos dos gráficos sobre un texto escrito. La secuencia de TTRs (*T*=50) y la aceleración de la diversidad léxica – la rapidez con que la velocidad de cambio entre valores de TTRs muda. La velocidad, a la que se ha aplicado un suavizado con promedios (n=20), conserva ruido; eventualmente podría aplicarse algún tratamiento.



Formas de estudio de los cuantificadores y los sujetos que los usan aquí es el análisis de estadísticas descriptivas y la procura de oscilaciones y recurrencias. Podemos esperar diferentes tipos de resultados: que no haya orden aparente, sino pura aleatoriedad; que la presencia de ciertos cuantificadores produzca un descenso de la diversidad léxica – pensando que los cuantificadores también afectan dicha medición — o lo opuesto. Paralelamente, podría esperarse que exista una dinámica diferente entre la producción de sujetos con TDL o sospechas de TDL y aquella de sujetos de control.

La segunda transformación que mencionamos anteriormente, el motivo [motif], es más bien una familia de transformaciones; se ha utilizado en ciencias de la computación, lingüística cuantitativa, y recientemente en estudios sobre la traducción por intérpretes (Liang et al., 2019). Dependiendo de la unidad que se tome, sus variantes miden la extensión de aumentos monotónicos en algún plano del lenguaje en su orden lineal – o inverso –, como se explicará luego. Los motivos lingüísticos, término acuñado en analogía a los motivos musicales (Köhler, 2015), han sido menos usados en psicolingüística que el TTR. Cada porción es la serie de valores numéricos en dirección ascendente, cada vez que el valor desciende comienza una nueva porción dentro del motivo. Por ejemplo, el motivo L –

por “length”, extensión – de letras por palabra en el enunciado “Martina compró un cuaderno y quiere escribir cuentos en él” se transforma a la siguiente secuencia de porciones: (7)-(6)-(2-8)-(1-7-8)-(7)-(2-2); las cantidades de miembros por porción son 1-1-2-3-1-2. Este es uno de los motivos más sencillos y posee limitaciones obvias como el pareo entre fonos y grafos. Pero existe gran versatilidad al poder utilizarlos con sílabas, morfemas, palabras, cláusulas, unidades del discurso, categorías léxicas, frecuencias léxicas, duración de la pronunciación y otras dimensiones más como el vector de TTRs producido por el MATTR. Además, los motivos pueden aplicarse recursivamente resultando en vectores de menor extensión. En el ejemplo anterior, el motivo LL sería: (1-1-2-3)-(1-2); y las cantidades de miembros por porción serían 4-2.

¿Cómo se pueden emplear los motivos para estudiar los cuantificadores? Si lo medimos por cantidad de constituyentes por palabras, sea por letras, sílabas o morfemas, esperaríamos que los cuantificadores encabecen porciones, pues la mayoría son breves. Ahora bien, aspectos de interés primario serían cuál es la extensión de cada porción cuando encabezada por un cuantificador y cuál es la extensión de las porciones que las rodean. Ello podría otorgar indicios sobre patrones posiblemente ligados a costos de procesamiento, inicialmente esperando que mientras más complejo el cuantificador, menor la porción en que se inserta – lo que coincidiría con la tendencia de Menzerath, por la cual el tamaño de las unidades varía en forma inversa al tamaño de sus constituyentes respectivos. Otros motivos de interés son el motivo F – de frecuencia – léxica, y el motivo R – de repetición. Este último toma como valores categorías léxicas y finaliza las porciones ni bien se repite uno de sus constituyentes; por ejemplo, usando una red neuronal entrenada en POS tagging, “Un perro y un gato se juntaron a comer en el patio” se transforma en (DET-N-CONJ)-(DET-N-PRON-V-PREP)-(V-PREP-DET-N); y luego en 3-5-4.

CONCLUSIONES Y PROSPECTIVAS

Es trabajo en producción la implementación de algunas de las transformaciones aquí mencionadas para estudiar textos orales producidos por menores en proceso escolar de diversas características, y textos escritos por adultos. Se espera que la diversidad de hablantes y la diversidad de formatos arrojen resultados más sustanciosos, si existieran patrones.

BIBLIOGRAFÍA

-**Bishop**, **D.V.M.**, **Snowling**, **M.J.**, **Thompson**, **P. A.**, **Greenhalgh**, **T.**, **CATALISE consortium** (2016) “CATALISE: A Multinational and Multidisciplinary Delphi Consensus Study. Identifying Language Impairments in Children” en *PLOS ONE*, 11(7).

-**Covington**, **M. A.**, **McFall**, **J. D.** (2010) “Cutting the Gordian Knot: The Moving-Average Type-Token Ratio (MATTR)” en *Journal of Quantitative Linguistics*, 17:2, 94-100.

-**Charest**, **M.**, **Skoczylas**, **M. J.**, **Schneider**, **P**. (2020) “Properties of Lexical Diversity in the Narratives of Children With Typical Language Development and Developmental Language Disorder” en *American Journal of Speech-Language Pathology*, 29(4):1866-1882.

-**Köhler**, **R**. (2015) “Linguistic Motifs” en *Sequences in Language and Text*. De Gruyter Mouton, Berlín/Boston.

-**Liang**, **J**., **Lv**, **Q**., **Liu**, **Y**. (2019) “Quantifying Interpreting Types: Language Sequence Mirrors Cognitive Load Minimization in Interpreting Tasks” en *Frontiers in Psychology*, Volume 10.