Proyecto de investigación sobre las Variedades Vernáculas Malagueñas (VUM, HUM-392)

DISPOCEN. MUCHO MÁS QUE UN PROGRAMA PARA EL CÁLCULO DE LA DISPONIBILIDAD LÉXICA

Juan Andrés Villena Ponsoda (vum@uma.es)

Antonio Manuel Ávila Muñoz (amavila@uma.es)

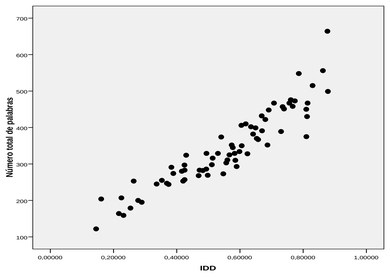
UNIVERSIDAD DE MÁLAGA

Resumen

DispoCen es un sistema para el análisis de la disponibilidad y la centralidad léxica. Aunque existen programas específicos para el cálculo de los citados índices, estos suelen restringir en exceso las posibilidades de análisis y explotación de los datos, bien porque se trata de herramientas obsoletas, bien porque sus códigos son excesivamente cerrados e inaccesibles.
DispoCen está basado en una librería de herramientas en *R* que pone al alcance de quienes estudian el léxico el desarrollo de múltiples aplicaciones y modelos originales. *DispoCen* incluye los códigos necesarios para ejecutar los análisis, con lo que se potencia la necesaria replicabilidad que favorece el trabajo autónomo de la comunidad investigadora. Para facilitar el acceso al sistema, también se ha generado una sencilla utilidad gráfica que permite el acceso a los análisis más usuales. Como muestra de las posibilidades de *DispoCen*, incluimos un apartado específico con propuestas de análisis realizadas con filtros sociológicos.

Aplicaciones (I) . Capacidad léxica individual

Nuestro modelo de representación del lenguaje considera el lexicón como una característica construida por los propios hablantes. Así, el núcleo del léxico de un prototipo determinado (en nuestro caso, presentado en forma de estímulo cognitivo que genera listas de palabras) estaría disponible para todos los sujetos. Por lo tanto, a cualquier individuo se le supone el acceso al léxico que forma parte del núcleo del prototipo, con lo que la caracterización de su diversidad léxica vendrá determinada, especialmente, por aquella parte del léxico que es más específica y menos compartida con el resto de la población. Según este planteamiento, las distintas listas de disponibilidad léxica individuales pueden considerarse complementarias entre sí y, por ello, nuestra propuesta abre nuevas vías de análisis cuantitativo que, hasta el momento, no se habían explorado: la hipótesis es que si un individuo actualiza palabras con un bajo índice de disponibilidad debe tener más diversidad léxica. Parece lógico suponer que, si las palabras que proporciona son menos disponibles, debe de tener más fácil acceso a aquellas otras compartidas por todos, además de a otras menos generalizadas, con lo que su diversidad léxica debería considerarse mayor. Se observa una relación directa entre el número total de palabras aportadas por cada individuo y la especificidad de los términos que aporta (Índice de Descentralización, IDD).

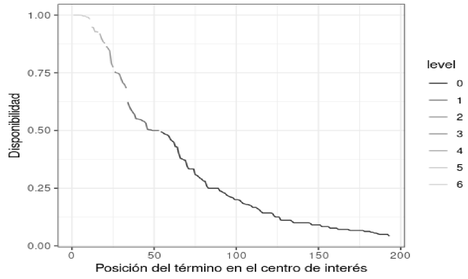


Introducción

Presentamos aquí el programa *DispoCen* para el cálculo de la disponibilidad léxica, la centralidad léxica y las diferentes funciones que se derivan de ambas. Hemos creado una librería de herramientas en *R* que permite, a través de programación funcional, implementar múltiples modelos y aplicaciones, entre las que destaca, precisamente, el desarrollo inmediato de las funcionalidades señaladas (disponibilidad, centralidad). Sin embargo, estas funciones de aplicación directa representan tan solo un subconjunto de las posibilidades que el sistema ofrece.
Se obvia aquí conscientemente la exposición de los fundamentos teóricos y metodológicos de la disponibilidad y la centralidad léxica, pues este póster va dirigido a investigadores especializados en el estudio del léxico. Consideramos que la comunidad investigadora demandaba un contexto actualizado a partir del cual realizar sus análisis y organizar los datos de manera autónoma. La obsolescencia de algunos programas previos para el cálculo de la disponibilidad, o el acceso restringido a otros, justifica la necesidad de una herramienta como la que presentamos en este trabajo.
Una de las ventajas del sistema que se ha creado es que nos ha permitido incluir en el mismo documento el análisis realizado junto con el código necesario para su ejecución, con lo que se potencia la replicabilidad, característica cada vez más exigida en el ámbito académico. Se consigue así que otros investigadores puedan replicar —tanto para verificar como para desarrollar— el trabajo realizado.
Para facilitar el proceso de instalación y ejecución de *DispoCen*, hemos montado un vídeo explicativo alojado en *Youtube* donde se detallan cada uno de los pasos necesarios para que la herramienta pueda usarse de manera eficaz (https://m.youtube.com/watch?v=IU5VfUvG4Ag).

Aplicaciones (II). Niveles de centralidad

Tamaño del conjunto de elementos que se debería establecer para configurar el núcleo de un centro de interés (o cuáles son los elementos más prototípicos de un centro de interés) Se proporciona una herramienta que etiqueta los términos por niveles de centralidad. El nivel 0 correspondería a aquellos elementos que no pertenecen al núcleo, es decir, aquellos términos que no son generalmente accesibles para el conjunto analizado. Los niveles 1, 2, 3 y sucesivos representarían un mayor grado de centralidad y aproximación al centro de interés.

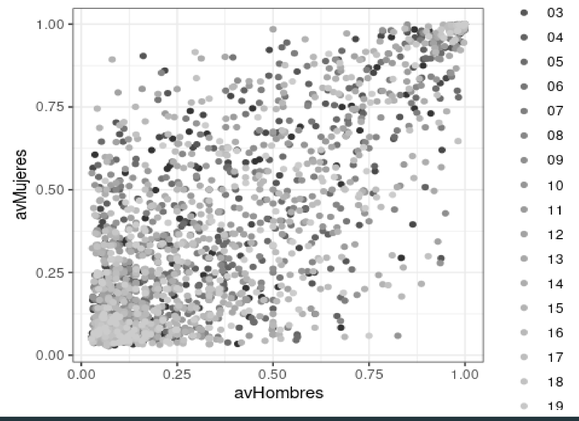


Antecedentes. La disponibilidad léxica

Nuestro modelo supera las limitaciones observadas en trabajos previos y sugiere una aproximación constructiva que revisa el concepto de 'disponibilidad léxica' en el que se basa. La disponibilidad se construye a partir de una metodología consolidada en el ámbito de la investigación sociolingüística y psicolingüística, donde está comprobada su utilidad a pesar de su simplicidad para extraer la información lingüística. Cuando se le pide a una persona que escriba listas de léxico disponible, se le ofrece un estímulo que activa un mecanismo cognitivo individual relativo al procesamiento de información. Se supone que aquellas formas que aparecen en las primeras posiciones de las listas son las más disponibles en la categoría cognitiva aludida por el estímulo inicial. De la consideración de un conjunto de listas particulares —donde se pondera el número de veces que aparecen las palabras y la posición que estas ocupan en los listados— resultará el índice de disponibilidad léxica: los términos más disponibles son los más frecuentes en los primeros lugares de los listados analizados (López-Morales, 1989; Strassburger-Frías y López-Chávez, 2000).

Aplicaciones (III). Análisis sociolingüísticos

Análisis por estratos y grupos sociales. Análisis comparativos. Ejemplos de uso. En el gráfico se representa la dispersión de términos por rango de disponibilidad en hombres y mujeres. Aunque ambos grupos comparten la mayoría de los términos muy disponibles y poco disponibles. Sin embargo, los términos que aparecen en los diversos niveles intermedios parecen variar de forma muy significativa.



Dispocen. Exposición, justificación y contextualización del Sistema

El modelo propuesto usa este método y lo relaciona con los avances de un ámbito específico de las matemáticas —la Teoría de los conjuntos difusos (Zadeh, 1065; Zimmermann, 2001)— que ha resultado natural y adecuado ya que facilita el empleo de herramientas contrastadas para la interpretación y el estudio de los datos obtenidos. El sistema se ha desarrollado en *R* a través del entorno comercial *Rstudio*. A diferencia de soluciones anteriores basadas en programas autónomos que calculaban la disponibilidad, nuestra propuesta permite, además, la integración de todo el sistema de gestión de datos. De esta manera, los resultados obtenidos a partir de las herramientas proporcionadas se incorporan al sistema, con lo que se proporciona la integración de datos en nuevos procesos y la aplicación personal de los resultados que se obtienen en función de las demandas específicas de los usuarios. Empleamos repositorios de código como *GitHub* que permiten la gestión y el control automatizado de las diferentes versiones. Es posible llevar a cabo el control exhaustivo de las modificaciones que se van realizando, sin necesidad de descargar e instalar continuamente diferentes actualizaciones de los mismos programas. Con dos líneas de código se asegura la instalación de la última versión disponible. Además, este modelo de intercambio de código permite que los proyectos no desaparezcan, ya que se pueden generar ramas de cualquier repositorio de código para continuar su desarrollo, independientemente de los creadores originales. Entre las múltiples herramientas construidas sobre *R* hemos encontrado muy interesantes y productivas las que conforman el *Universo Tidyverse*. Se trata de un conjunto de librerías que facilitan y permiten una manipulación expresiva de los datos, basada en el concepto de 'tubería'. Este tratamiento se construye mediante la secuenciación de manipulaciones que van encadenándose. De esta manera, el resultado de una etapa de análisis constituye la fuente de datos para la siguiente. Esta característica, junto con operadores adaptados a este tipo de trabajo, proporciona una forma potente y cómoda de llevar a cabo transformaciones que, de otra forma, serían muy complejas. El tratamiento de datos presentado en este trabajo se realiza utilizando estas herramientas.

Conclusiones

DispoCen es una utilidad para el cálculo de la disponibilidad y la centralidad léxica. Pero sus posibilidades van mucho más allá de las utilidades señaladas. Al estar generada con una librería de herramientas en *R* por medio de programación funcional, sus aplicaciones son más abiertas y extensas. De momento, el investigador de la disponibilidad y la centralidad léxica encontrará en *DispoCen* una herramienta actualizada que sustituya, quizás, a programas obsoletos o de acceso restringido comunes hasta ahora para esta función.
DispoCen abre una serie de posibilidades que hasta ahora eran inaccesibles: desde la verificación de los pasos realizados hasta el desarrollo y adaptación a cada situación particular de análisis gracias a la replicabilidad que ofrece esta aproximación.

Biibliografía

Ávila-Muñoz, Antonio-Manuel y Sánchez-Sáez, José-María 2011. La posición de los vocablos en el cálculo del índice de disponibilidad léxica: procesos de reentrada en las listas del léxico disponible de la ciudad de Málaga. *ELUA. Estudios de Lingüística* 25: 45-74.
Ávila-Muñoz, Antonio-Manuel 2016. Can speakers' virtual lexical richness be calculated? Individual and social determining factors. *Spanish in Context*, Volume 13/2, pp. 285 – 307..
Ávila-Muñoz, Antonio-Manuel y Villena-Ponsoda, Juan-Andrés (eds.) 2010. *Variación social del léxico disponible en la ciudad de Málaga*. Sarriá: Málaga.
Strassburger, Carlos y López-Chávez, Juan 2000. El diseño de una fórmula matemática para obtener un índice de disponibilidad léxica confiable. Anuario de Letras: Lingüística y filología, 38, pp. 227-251.
.