Una breve introducción a las aplicaciones de la estadística en los estudios del lenguaje

Federico Álvarez

23 de agosto de 2018

A lo largo de su historia la tradición lingüística ha mantenido, mayormente, un enfoque cualitativo en el abordaje de su objeto de estudio. Por diferentes razones (algunas de las cuales pueden rastrearse en la bilbiografía, otras pueden tener más que ver con la apreciación personal de cada investigador) los estudios del lenguaje no han aplicado de manera extensiva herramientas cuantitativas para articular sus hipótesis, a diferencia de la mayoría de las disciplinas científicas (posiblemente la psicolingüística y la neurolingüística sean una excepción, si bien es discutible el grado de participación de los lingüistas en el análisis estadístico de sus experimentos). Esta tendencia se ha ido revirtiendo durante las últimas décadas, y al día de hoy podemos encontrar una gran cantidad de ejemplos de estudios que hacen uso de modelos estadísticos para comprobar sus predicciones e incluso para formularlas. Sin embargo, al menos en nuestro país esta renovación en la metodología no ha sido acompañada por un cambio en los programas de estudios de las carreras de lingüística, lo cual es comprensible en tanto el estudio del lenguaje está aún asociado al estudio de la literatura y no constituye una carrera aparte en la esfera universitaria. Uno de los problemas que se derivan de esta situación es que aquellos lingüistas interesados en adoptar una metodología con una base matemática más sólida no tenemos un andamiaje institucional que facilite el abordaje de la temática. Este trabajo, entonces, se ofrece como un repaso sobre las bases y requisitos de un enfoque cuantitativo, y las áreas donde este tipo de enfoque puede ser de utilidad en el campo disciplinar de la lingüística.

Introducción

Antes de preguntarnos cómo y dónde se usa la estadística en la lingüística, es sensato proponernos una pregunta más básica que es *¿qué es la estadística?* La página en inglés de Wikipedia (2018) propone la siguiente definición:

> La estadística es una rama de la matemática que abarca la recolección, el análisis, la interpretación, la presentación y la organización de datos. Al emplearla en, por ejemplo, un problema científico, industrial o social, es convencional comenzar con una población estadística o un proceso de modelo estadístico a estudiar. Las poblaciones pueden ser temas tan diversos como "todas las personas que viven en un país" o "cada átomo que compone un cristal". La estadística abarca *todos los aspectos de los datos*, incluyendo el planeamiento de la recolección de datos en términos de diseño de encuestas y experimentos (mi traducción).

Esta definición apunta a una serie de cuestiones bastante fundamentales que es necesario plantearse a la hora de abordar una investigación de manera cuantitativa. El propio diseño de la investigación requiere una serie de consideraciones para que las hipótesis sean estadísticamente comprobables, que van desde la correcta operacionalización de las variables a estudiar hasta la propia formulación de la hipótesis. Esto no quiere decir que no haya lugar para la consideración cualitativa en la investigación científica: las consideraciones cualitativas preceden al diseño de los estudios, ya que determinan la selección misma del objeto de estudio, así como de las variables de interés para realizar un análisis, y también determinan la interpretación final de los resultados obtenidos tras la aplicación de un procedimiento estadístico.

¿Qué procesos supone un análisis estadístico?

A grandes rasgos, un estudio estadístico supone la caracterización de un aspecto o conjunto de aspectos (variables) de una población a partir del análisis de la distribución de los mismos en una muestra. En esta sección, voy a describir de manera escueta los procedimientos estadísticos para realizar estas inferencias. En un contexto lingüístico una población podría ser un colectivo de hablantes, el conjunto de las emisiones que realizan, un corpus en particular, la totalidad de realizaciones de una estructura sintáctica y una infinidad más de posibilidades. Cada caso implica una forma particular de definir la población, las variables de interés y la manera de medirlas. Imaginemos un experimento sencillo de decisión léxica, en el que nos interesa comprobar si las personas tardan más en distinguir si una secuencia de caracteres o fonemas es o no una palabra cuando no lo es que cuando sí lo es. En este caso nuestra población sería el conjunto de los hablantes que (suponemos) poseen un mismo conjunto de elementos en su léxico mental, nuestra muestra sería un conjunto de secuencias de caracteres o fonemas, algunas de las cuales son elementos de dicho léxico mental en tanto que otras no, y las variables que nos interesan son el grupo al que pertenezcan las secuencias (palabras / no palabras - variable independiente; VI) y el tiempo que toma para cada hablante decidir si la secuencia presentada pertenece a uno u otro grupo (tiempo de reacción - variable dependiente; VD).

Hipótesis, probabilidades y espacio muestral

Un estudio empírico impone una serie de requisitos sobre la formulación teórica (Gries 2013). Para ajustarse a ellos una hipótesis necesita cumplir con los siguientes criterios:

- Debe ser una *afirmación general* que abarque más que un *evento singular*.
- Debe tener la estructura de una *oración condicional*, o ser parafraseable como una.
- Debe ser *falsable*, o sea, es necesario que exista la posibilidad de concebir eventos o situaciones que contradigan la afirmación; esto a su vez requiere que *exista la posibilidad de comprobarla*, sin embargo estas dos características no son idénticas: hay afirmaciones que son falsables pero no testeables.

Al definir las hipótesis debe definirse también la condición que falsaría cada una, por lo tanto para cada hipótesis propia -hipótesis alternativa H_1- es necesario formular su hipótesis opuesta -hipótesis nula H_0- . El modo en el que nosotros definamos nuestras variables de interés y el procedimiento para medirlas (operacionalización) tiene una consecuencia obvia en la formulación de nuestras hipótesis; siguiendo el ejemplo del experimento de decisión léxica, nosotros podríamos formular una H_1 según la cual el tiempo de reacción es diferente para los estímulos que son palabras y para los que no (nuestra hipótesis también podría ser que la identificación de no palabras tarda más que la de palabras, o al revés, lo que nos daría una *hipótesis direccional*). El conjunto de todos los posibles resultados de nuestro experimento constituye lo que se llama *espacio muestral* (Casella y Berger 2001). Las hipótesis, en términos cuantitativos, representan la asignación de un valor numérico entre 0 y 1 para cada uno de los elementos de nuestro espacio muestral, de forma tal que la suma (o integración) del valor asignado a todos los eventos sume 1. El número asignado a cada elemento del espacio muestral es la *probabilidad* de ese elemento.

Diseño del estudio

No basta con tener hipótesis nula y alternativa y variables operacionalizadas. Los aspectos formales y materiales de la recolección de datos son también extremadamente relevantes. Por ejemplo, para el caso de la decisión léxica, múltiples maneras de extraer información son posibles: ¿Tomamos una muestra por sujeto o varias? ¿y por condición? ¿Todos los sujetos ven los mismos estímulos o hay diferentes condiciones de presentación? De haber diferentes condiciones, ¿todas contienen la misma cantidad de estímulos? ¿Son todos del mismo tipo? Si salimos del caso de la decisión léxica también podemos considerar otras cuestiones como ¿cuántas variables independientes hay en consideración? ¿Cuántas dependientes? ¿Hay alguna relación observable entre ellas? Distintas respuestas a estas preguntas van a suponer el uso de diferentes pruebas.

Además, no todo estudio es un experimento. Diferentes tipos de estudios suponen no sólo una diferencia en la metodología de prueba de hipótesis, sino que además cambia el modo en el que se interpretan las pruebas.

Estas consideraciones sobre los datos no son sólo relevantes a la hora de pensar en cómo probar una hipótesis; son constitutivas respecto de cómo se concibe el objeto de estudio. Sin abogar por ninguna metodología en particular, es importante analizar los métodos de recolección de datos sobre los cuales descansan los supuestos previos sobre el lenguaje empleados para construir las teorías con las que trabajamos (Abney 2011). Aún si halláramos que los procesos son defectuosos, no obstante, eso no implica que los resultados sean falsos ni que las teorías sean insuficientes. Sin embargo sí significa que hay valor en intentar reproducir de las observaciones de una teoría empleando una metodología más rigurosa que la que se utilizó para construirla.

Estadística descriptiva y exploración visual

Cuando poseemos una muestra sobre la cual verificar nuestras hipótesis, es importante reducir la información presente en cada una de las observaciones de nuestra muestra a una cantidad más manejable de valores que sin embargo den debida cuenta de las propiedades de la muestra (Johnson 2014). La estadística descriptiva refiere al conjunto de valores que dan cuenta de las propiedades de una muestra sin, por sí solas, responder preguntas acerca de las propiedades de la población de la que proviene. En general, los valores medidos en una estadística descriptiva dan cuenta de la tendencia central de la muestra (un único valor que resulta representativo de todas las observaciones), la variabilidad observada entre distintos valores de una misma variable, y la asociación entre diferentes variables. La elección de diferentes estadísticos descriptivos, ya sea de tendencia central, dispersión o asociación depende del modo en el que se hayan operacionalizado las variables.

La representación visual de la muestra es también de gran importancia, ya que permite observar patrones que no necesariamente se evidencian en las medidas descriptivas, o permite interpretar de manera rápida e intuitiva dichas medidas. Como todo en estadística, las visualizaciones a emplear están fuertemente asociadas a los tipos de variables de interés, así como a su cantidad.

Estadística inferencial

Este es el paso en el cual se proyecta la información obtenida en la muestra sobre la población de interés. La batería de pruebas estadísticas disponibles representa, en su forma más básica, la asignación de una probabilidad a las observaciones de nuestra muestra tomando en cuenta nuestra hipótesis. De acuerdo con el resultado de la prueba que apliquemos vamos a aceptar nuestra hipótesis o rechazarla en favor de H_0. La aceptación o rechazo de una hipótesis dependen de la probabilidad de obtener las observaciones de la muestra bajo el supuesto de la hipótesis nula, lo cual se conoce como error de tipo 1 (el error de tipo 2 consiste en rechazar erróneamente la hipótesis alternativa) (Arunachalam 2013).

Las diferentes pruebas a realizar dependen del cumplimiento de distintos supuestos sobre las variables, su distribución y su asociación. Las familias de modelos estadísticos se pueden dividir en dos grandes grupos: paramétricos y no paramétricos. Los modelos paramétricos son aquellos cuya estructura está determinada por un conjunto finito de valores, por lo cual su complejidad está delimitada ante un aumento en el tamaño de la muestra. Los modelos no paramétricos, en cambio, no poseen una estructura predefinida. Sin entrar en distinciones, las pruebas paramétricas suelen ser más sencillas en términos de procesamiento, pero suelen tener una mayor cantidad de supuestos sobre la muestra, mientras que las pruebas no paramétricas suelen ser más simples, pero menos sensibles a tendencias en la muestra.

Una cuestión crucial sobre la inferencia es que lo que supone, en su forma más abstracta, es la probabilidad de predecir el futuro. O sea, la idea es que al aceptar una hipótesis estamos comprometiéndonos con la idea de que, ante una eventual nueva muestra de la misma población, pueda predecirse el comportamiento de las observaciones. Esto quiere decir que no nos basta sólo con la adecuación a los datos ya observados: entre dos teorías en competencia, una que explica perfectamente observaciones disponibles pero resulta insuficiente ante la aparición de nuevos datos es menos robusta que una teoría que, sin explicar la totalidad

de la información disponible, realiza predicciones de manera consistente. La inferencia estadística, entonces, supone una formalización sobre la probabilidad de replicar observaciones.

¿En qué áreas de la lingüística puede proponerse un abordaje estadístico?

Hay estudios sobre el lenguaje que parecen prestarse naturalmente a una óptica cuantitativa. Sin embargo, los lugares donde el uso de modelos estadísticos pueden enriquecer nuestro conocimiento sobre el lenguaje son más de los que podríamos suponer si nos guiamos por las teorías más difundidas sobre la gramática. En esta sección paso revista a algunas de estas áreas.

Psicolingüística y neurolingüística - Abordajes experimentales

Tanto en estudios de adquisición como de comprensión y producción de lenguaje, los experimentos realizados suelen estar sujetos a un análisis cuantitativo. De acuerdo al aspecto del lenguaje bajo estudio y a la metodología utilizada las pruebas requeridas van a ser diferentes, pero en general las variables independientes son estructuras de diferentes niveles lingüísticos y/o diferentes grupos dentro de la población y las variables dependientes son, para estudios online, comportamentales (ej.: tiempo de reacción, movimientos sacádicos) o fisiológicas (ej.: ERP, flujo sanguíneo) y para estudios offline pueden ser diferentes cosas como la cantidad de palabras diferentes empleadas para completar un espacio en blanco, la cantidad de errores cometidos en una tarea, y varias más. La población analizada son los hablantes, que pueden estar divididos en diferentes grupos.

Lingüística de corpus

La lingüística de corpus, como disciplina, suele tener un propósito más descriptivo que inferencial: cuando el corpus es la población de interés y ya está disponible en su totalidad, anulando la necesidad de tomar muestras, todo el análisis cuantitativo posible va a ser una caracterización del mismo, sin necesidad de realizar inferencia alguna. Las variables de análisis van a ser básicamente frecuencias, ya se trate de palabras, tipos, lexemas, locuciones, morfemas, estructuras sintácticas, o algún otro constructo teórico. Las medidas más complejas son asociaciones y proporciones entre estas frecuencias (Gries 2009). Obviamente nada impide a priori el uso del corpus como herramienta para realizar inferencias sobre otra población, pero eso ya escaparía al campo de la lingüística de corpus.

Lingüística computacional, procesamiento de lenguaje natural

Estos dos campos, que al utilizar herramientas similares suelen presentarse de manera asociada, pueden distinguirse por diferencias en su propósito: mientras que la lingüística computacional supone el uso de medios computacionales para la comprensión del lenguaje a diferentes niveles de análisis, el procesamiento de lenguaje natural busca optimizar las herramientas disponibles para resolver diferentes problemas utilizando un input lingüístico, aún si las soluciones propuestas no responden al procesamiento que hace un hablante (Cohen 2016). Hay una tradición importante de modelos computacionales basados en teorías formales sobre todos los niveles de la gramática, que emplean sistemas determinísticos de reglas. La disponibilidad de grandes volúmenes de datos y el incremento en la capacidad de cómputo han favorecido el desarrollo de modelos basados en la probabilidad de hallazgo de patrones en corpus de entrenamiento (Klavans y Resnik 1996; Manning y Schütze 1999). Los abordajes computacionales estadísticos abarcan varios niveles de análisis, desde gramáticas probabilísticas que asignan una probabilidad a diferentes estructuras sintácticas y modelos de lenguaje basados en las probabilidades de transiciones entre secuencias de morfemas o palabras a modelos de las temáticas abordadas en un corpus (Blei 2012) o modelos de procesamiento acústico de habla (que no cito por no saber mucho del tema, pero la bibliografía es abundante).

Tipología y lingüística histórica

El artículo de Dunn etal. (2005) muestra el empleo de métodos basados en árboles filogenéticos binarios para la identificación de familias lingüísticas y del grado de semejanza que poseen las distintas lenguas entre sí, basándose en la codificación binaria -presencia o no presencia- de diferentes rasgos gramaticales. Varios métodos de inferencia son empleados para la selección entre los múltiples árboles posibles. Los métodos no se limitan al uso de rasgos gramaticales: es posible el empleo de variables extralingüísticas como la distancia geográfica, u otras variables relacionadas con el lenguaje como la similitud entre palabras a través de cognados.

Teoría lingüística

Al incorporar herramientas computacionales como parte de las opciones disponi- bles al lingüista para desarrollar sus teorías, es posible la generación de hipótesis cuantitativas en todos los niveles de estudio del lenguaje, ya se trate del estudio de la sintaxis a través de juicios de aceptabilidad (C. T. Schütze y Sprouse 2014), semántica(Goodman y Lassiter 2015), o básicamente cualquier otro nivel (Bod, Hay, y Jannedy 2003). De hecho, incluso abordajes teóricos ya consolidados pueden incorporar modelado estadístico a su aparato teórico, como la gramática generativa para modelar adquisición y variación (Yang 2004) o la lingüística cognitiva para modelar fenómenos semánticos, morfológicos y construccionales (Kuznetsova 2013; Milin etal. 2016). Dado que la lógica de los análisis estadísticos se desprende del modo en la teoría misma defina sus variables de interés, en tanto estas definiciones sean formalmente explícitas y tengan procedimientos de medición que permitan la recolección de muestras, todo constructo lingüístico es susceptible de un análisis estadístico.

Conclusión

Espero que al llegar a este punto no resulte abrumadora la cantidad de puntos desde los cuales se puede proponer un abordaje cuantitativo para una problemática lingüística. De hecho, más aún, creo que debería ser reconfortante: sin importar la filiación teórica o el nivel de análisis en el que uno se especialice, es posible utilizar el conocimiento previo como asidero para introducirse en un paradigma más riguroso formal y empíricamente. No hay que entender las opciones disponibles como una limitación que vuelve la perspectiva cuantitativa inabarcable, sino como la presencia de una gran cantidad de puntos de entrada al tema, de forma tal que es posible seleccionar el que sea más afín al bagaje previo de cada uno. La ganancia en precisión teórica que se obtiene con el modelado estadístico del lenguaje representa un avance significativo en el poder explicativo de la teoría lingüística. Abney (2011) dice respecto de la lingüística computacional que:

> El lenguaje es un sistema computacional, y hay una profundidad de entendimiento que es inalcanzable sin un conocimiento profundo de la computación. Pero incluso por sobre eso, el nuevo enfoque [la lingüística computacional] refleja un entendimiento más profundo del método científico y sitúa a la investigación lingüística firmemente dentro del paradigma de investigación intensiva de datos que ha dado en caracterizar a la ciencia moderna (mi traducción).

En tanto el abordaje estadístico está íntimamente ligado al empleo de herramientas computacionales, considero que la afirmación de arriba es más apropiada para referirse al abordaje cuantitativo que al computacional per se. El uso de herramientas computacionales permite una recolección masiva de datos y una aplicación extensiva de análisis de cualquier tipo, por lo que el giro conceptual no se sitúa tanto en la posibilidad de extraer datos sino en las posibilidades de refinamiento teórico que proporciona una metodología cuantitativa, en consonancia con el resto de la ciencia contemporánea.

Licencia

Bibliografía consultada

Abney, Steven. 2011. «Data-intensive experimental linguistics». *Linguistic Issues in Language Technology* 6 (2): 20.

Arunachalam, Sudha. 2013. «Experimental methods for linguists». *Language and Linguistics Compass* 7 (4). Wiley Online Library: 221-32.

Blei, David M. 2012. «Probabilistic topic models». *Communications of the ACM* 55 (4). ACM: 77-84.

Bod, Rens, Jennifer Hay, y Stefanie Jannedy. 2003. *Probabilistic linguistics*. Mit Press.

Casella, George, y Roger L. Berger. 2001. *Statistical Inference*. 2.ª ed. Duxbury Press.

Cohen, Shay. 2016. *Bayesian Analysis in Natural Language Processing*. Synthesis lectures on human language technologies 35. Morgan & Claypool Publishers.

Dunn, Michael, Angela Terrill, Ger Reesink, Robert A Foley, y Stephen C Levinson. 2005. «Structural phylogenetics and the reconstruction of ancient language history». *Science* 309 (5743). American Association for the Advancement of Science: 2072-5.

Goodman, Noah D, y Daniel Lassiter. 2015. «Probabilistic semantics and pragmatics: Uncertainty in language and thought». *The handbook of contemporary semantic theory, 2nd edition. Wiley-Blackwell*.

Gries, Stefan Th. 2009. *Quantitative Corpus Linguistics with R: A Practical Introduction*. 1.ª ed. Routledge.

———. 2013. *Statistics for Linguistics with R: A Practical Introduction*. 2nd rev. ed. Mouton Textbook. Walter de Gruyter.

Johnson, Daniel Ezra. 2014. «Descriptive statistics». *Research methods in linguistics*. Cambridge University Press, 288.

Klavans, Judith, y Philip Resnik. 1996. *The Balancing Act: Combining Symbolic and Statistical Approaches to Language (Language, Speech, and Communication)*. First Edition. Language, Speech, and Communication. The MIT Press.

Kuznetsova, Julia. 2013. «Linguistic profiles: Correlations between form and meaning». *Tromsø: Universitetet i Tromsø PhD thesis*.

Manning, C.D., y H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*.

Milin, Petar, Dagmar Divjak, Strahinja Dimitrijević, y R. Harald Baayen. 2016. «Towards cognitively plausible data science in language research». *Cognitive Linguistics* 27 (4): 507-26.

Schütze, Carson T., y Jon Sprouse. 2014. «Judgment data». *Research methods in linguistics*, 27-50.

Wikipedia. 2018. «Statistics — Wikipedia, The Free Encyclopedia».

Yang, Charles D. 2004. «Universal Grammar, statistics or both?» *Trends in cognitive sciences* 8 (10). Elsevier: 451-56.