ISSN 2796-8952

CUADERNOS DEL GRUPO DE ESTADÍSTICA PARA EL ESTUDIO DEL LENGUAJE

Volumen I

Julia Carden (ed.)
Ezequiel Koile (ed.)
Analí Taboh (ed.)
Federico Alvarez
Santiago Gualchi
Mercedes Güemes
Laura Tallon
Charo Zacchigna

GESEL

Corrección y edición
Julia Carden
Ezequiel Koile
Analí Taboh

Cuadernos del Grupo de Estadística para el Estudio del Lenguaje, I volumen / Federico Alvarez; Santiago Gualchi; Mercedes Güemes; Laura Tallon y Charo Zacchigna.

1era edición.
25 de mayo, 217; primer piso.
Ciudad Autónoma de Buenos Aires, Argentina.
Publicación digital, diciembre 2021.
ISSN: 2796-8952

CUADERNOS DEL GRUPO DE ESTADÍSTICA PARA EL ESTUDIO DEL LENGUAJE

Volumen I

Federico Alvarez, Santiago Gualchi , Mercedes Güemes, Laura Tallon y Charo Zacchigna.

Editado por Julia Carden, Ezequiel Koile y Analí Taboh.

ISSN 2796-8952

Prólogo

Sobre este cuaderno

El Cuaderno del Grupo de Estadística para el Estudio del Lenguaje - Volumen I es una obra colectiva. Los distintos capítulos que lo integran son el resultado de una recopilación bibliográfica llevada a cabo dentro del marco del subsidio FiloCyT "Métodos cuantitativos en el estudio del lenguaje". Cada capítulo fue elaborado para acompañar una reunión del equipo en la que se trataron los distintos temas con mayor profundidad, a modo de presentación oral. Este volumen no pretende ser una obra acabada sobre los temas de estadística que se abordan, sino una guía para la lectura de bibliografía más compleja y completa sobre el tema. Por este motivo, estos capítulos deben ser acompañados de la lecturas que se sugieren en las referencias.

Estos cuadernos fueron concebidos para aportar material de consulta sobre estadística aplicada al estudio del lenguaje que, al menos en español, sigue siendo un área de vacancia bibliográfica. El objetivo es que proporcionen información, acceso a nuevos materiales y una explicación breve y simple sobre los puntos esenciales que deben ser conocidos para quienes se introduzcan en el estudio cuantitativo del lenguaje y no posean una base firme en matemática.

Capítulo I

Estadística para el estudio del lenguaje: una introducción

Federico Alvarez

A lo largo de su historia, la tradición lingüística ha mantenido un enfoque preponderantemente cualitativo en el abordaje de su objeto de estudio. Por distintas razones (algunas de las cuales pueden rastrearse en la bibliografía, otras pueden tener más que ver con la apreciación personal de cada investigadore), las teorías sobre el lenguaje no han aplicado de manera extensiva herramientas cuantitativas para articular sus hipótesis, cosa que sí ha ocurrido en la mayoría de las disciplinas científicas. Esta tendencia se ha ido modificando durante las últimas décadas, y al día de hoy podemos encontrar una gran cantidad de estudios que hacen uso de modelos estadísticos para formular y verificar sus predicciones. Sin embargo, este viraje en la metodología no ha tenido aún un impacto fuerte en la formación de les lingüistas, y tampoco hay demasiado material disponible para iniciarse en el estudio de la estadística que haga foco en su aplicación en las ciencias del lenguaje. Uno de los problemas que se derivan de esta situación es que aquelles lingüistas interesades en adoptar una metodología con una base formal[1] más sólida no necesariamente tienen un andamiaje institucional que facilite el abordaje de la temática. Este trabajo, entonces, se ofrece como un repaso sobre las bases y los requisitos de un enfoque cuantitativo y las áreas donde este tipo de enfoque puede ser de utilidad en el campo disciplinar de la lingüística, con el propósito de servir como guía en el aprendizaje teórico de la estadística.

1. ¿Qué es la estadística?

Es fácil dar por sentado que sabemos qué es la estadística, al punto que varios textos introductorios al tema no proveen ninguna definición del término. Más aún, al contrastar las definiciones que brindan distintes autores nos encontramos con tantas discrepancias como coincidencias.

Freedman et al. (1998) ofrecen la siguiente definición (trad. mía):

> La estadística es el arte de proponer conjeturas numéricas sobre preguntas desconcertantes.

Esta definición, aunque escueta, nos indica que vamos a trabajar con números con el objetivo de responder preguntas. Una cosa importante que falta es una idea sobre cómo se vinculan las conjeturas con las preguntas. La siguiente definición, de Levshina (2015), dice (trad. mía):

> [...] la estadística es un conjunto de herramientas para describir y analizar datos.

Nuevamente una definición bastante escueta. Aquí nos presenta un concepto central para la disciplina que es el de **dato**. Al hacer estadística necesariamente estamos trabajando en función de datos, ya sea para

[1] El término "formal" no refiere a una oposición entre forma y significado sino al empleo de abstracciones con el fin de explicitar la relación entre los conceptos teóricos y datos empíricos de manera tal que cada observación se pueda interpretar de manera unívoca.

describirlos o para analizarlos. La próxima definición, de Downey (2011), aúna un poco algunos elementos de las dos anteriores (trad. mía):

> La estadística es la disciplina dedicada a usar muestras de datos para respaldar afirmaciones sobre poblaciones. La mayoría del análisis estadístico está basado en la probabilidad, debido a lo cual estos temas suelen presentarse conjuntamente.

Aquí podemos ver un vínculo explícito entre la noción de dato y el objetivo de responder preguntas. Al mismo tiempo, se introducen varios conceptos importantes que también tendremos que definir para orientar nuestro aprendizaje: **muestra**, **población** y **probabilidad**. Finalmente, la siguiente definición, de la página en inglés de Wikipedia (2018), es la más abarcativa (trad. mía):

> 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 modelo estadístico a estudiar. Las poblaciones pueden ser tan diversas como "todas las personas que viven en un país" y "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.

Esta definición apunta a una serie de cuestiones fundamentales que es necesario plantearse para abordar una investigación de manera cuantitativa. Hace mucho énfasis en el concepto de dato, alude al concepto de población e introduce el concepto de **modelo**.

Provisionalmente, vamos a definir los conceptos que aparecieron en estas definiciones para construir una definición operativa de estadística que los tome en cuenta.

- **Dato**: Unidad de observación sobre un fenómeno empírico determinado.
- **Muestra**: Conjunto de observaciones.
- **Población**: Grupo que representa los objetos bajo estudio. Dependiendo del estudio, la población puede ser cosas muy diferentes como "los hablantes de español rioplatense", "los conectores adversativos" o "las oraciones que usan verbos intransitivos". La población es aquello sobre lo cual nos estamos haciendo una pregunta que queremos responder a través de la recolección y el análisis de observaciones.
- **Modelo**: Constructo diseñado para dar cuenta de algún aspecto de la realidad de manera esquemática. Es un conjunto de supuestos explícitos sobre aspectos puntuales de fenómenos de interés.
- **Probabilidad**: Estudio de **eventos aleatorios**. Es una forma de representar numéricamente nuestras intuiciones y nuestros supuestos sobre las chances de que ocurran ciertos eventos y no otros. Normalmente pensamos informalmente en la verosimilitud de ciertos sucesos en relación a otros, más o menos verosímiles. La teoría de la probabilidad nos permite hacer afirmaciones cuantitativas sobre estas ideas.

Con este conjunto de definiciones podemos pasar a proponer una definición operativa de la estadística. Para nuestros propósitos, vamos a considerar que:

> La estadística es la disciplina que abarca todos los aspectos del estudio de observaciones empíricas en función de nuestras creencias sobre fenómenos particulares. Supone la

ISSN: 2796-8952

> recolección de datos de una población de interés, su descripción y análisis en función de modelos probabilísticos, y su subsecuente interpretación.

Esta definición implica que desde el momento en el que decidamos usar observaciones empíricas para responder una pregunta sobre un fenómeno de interés vamos a estar dentro del marco de la estadística, y los procedimientos que elijamos para obtener observaciones, las decisiones sobre cómo vamos a observar el fenómeno, la construcción de modelos que lo expliquen, la formulación de preguntas sobre dichos modelos y la interpretación de los datos con el fin de obtener una respuesta son todos aspectos importantes en la práctica. Esto no quiere decir que no haya lugar para las consideraciones cualitativas sino todo lo contrario: el análisis cualitativo precede al diseño de los estudios, ya que determina la selección misma del objeto de estudio, así como de las variables de interés para realizar un análisis, y también determina la interpretación final de los resultados obtenidos.

2. ¿Cómo hacemos estadística?

De manera muy general, 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. Los distintos objetivos de cada estudio resultan en dos modos complementarios de analizar los datos, **exploratorio** y **confirmatorio**.

Une puede explorar las observaciones a su disposición con el objetivo general de encontrar patrones, pautas que se repitan y den lugar a preguntas sobre el fenómeno observado. Según Behrens (1997), el análisis exploratorio supone el trabajo básico de familiarización con los datos que permite responder la pregunta general "¿qué está pasando acá?" e implica un énfasis en la representación gráfica de los datos, un foco en un proceso iterativo de generación de hipótesis y construcción de modelos tentativos, y una postura flexible respecto a los métodos a adoptar. Para el análisis confirmatorio, en cambio, es necesario especificar a priori las hipótesis y los métodos a emplear. No se trata de una instancia de generación de preguntas sino de respuestas.

Tanto la exploración como la confirmación recurren al mismo conjunto de herramientas; si bien pueden poner énfasis en elementos diferentes, la diferencia crucial entre ambos modos de análisis no pasa por las herramientas que usan, sino que radica en que en el análisis confirmatorio los datos se miran una única vez. En cuanto volvemos a consultarlos, ya estamos realizando una exploración y no podemos interpretar los patrones que encontremos como la confirmación de una hipótesis. Esto implica, necesariamente, que los datos que se emplean para proponer una hipótesis y los que se usan para confirmar jamás pueden ser los mismos.

La exploración y la confirmación se suceden iterativamente, donde los datos que se recolectaron para confirmar una hipótesis son luego explorados para proponer hipótesis diferentes, que luego se ponen a prueba con nuevas observaciones. Es importante considerar que tanto la exploración como la confirmación son importantes, que la exploración va primero y que cualquier estudio puede, y usualmente *debe,* incluir ambas.

Reflexionemos sobre dos ejemplos:

- Queremos estudiar las diferencias de realización del fonema /r/ en hablantes de Argentina. Nuestro conocimiento previo nos sugiere por lo menos dos: una vibrante múltiple dental sonora y una fricativa palatal sonora.

ISSN: 2796-8952

- Tras una búsqueda bibliográfica nos quedamos con estas dos opciones y llegamos a la suposición de que el uso de una u otra tiene relación con la región de origen de le hablante, donde distinguimos entre Cuyo, el Litoral, el Norte, el Centro, la Pampa, la Patagonia y el Río de la Plata.
- También suponemos que hay una influencia del registro, que distinguimos entre formal y coloquial (¿son las únicas posibilidades? ¿cómo determinamos qué registro corresponde a cada caso?).
- Asimismo, consideramos probable que haya una variabilidad importante entre individuos.
- Suponemos que la realización fricativa ocurre con una frecuencia superior en personas del Norte y que es más común en el registro coloquial que en el formal.
- En función de lo anterior, decidimos realizar grabaciones de personas de las seis regiones en un ambiente laboral de oficina, que identificamos como situación formal, y en un ambiente doméstico, que caracterizamos como informal. De las grabaciones extraemos, para cada hablante, la cantidad de veces que produjo cada realización. Luego, realizamos un test para verificar nuestra hipótesis. Al analizar nuestros resultados, nos encontramos con que efectivamente hay una diferencia en la cantidad de realizaciones consistente con nuestra hipótesis anterior, pero nos encontramos con que hay mucha variación entre sujetos que no se explica a través de las variables que consideramos. A partir de eso, agregamos a nuestras observaciones información sobre la edad y el género de cada hablante, y llevamos a cabo un nuevo test para verificar si explican parte de la variación.

- Queremos estudiar las oraciones en las que el objeto directo es un nominal coordinado, del tipo de "los vi a vos y a tu hermano juntos el viernes".
  - Tras hacer un relevamiento bibliográfico llegamos a la propuesta teórica de que, en una tarea de juicios de aceptabilidad, la concordancia del clítico acusativo con el objeto directo –que puede concordar en singular con el primer coordinado o con el segundo, o en plural con ambos– y el modo en el que se interprete el evento descrito por el verbo –un único evento que afecta a ambos OD o dos eventos disjuntos– tienen un efecto en la aceptabilidad percibida. Medimos este efecto en una escala del 0 al 10, donde la concordancia con el primer coordinado tendrá valores altos de aceptabilidad, la concordancia con el último coordinado tendrá bajos valores de aceptabilidad (¿qué serían valores altos y bajos? ¿por qué?) y la concordancia con ambos coordinados tendrá baja aceptabilidad para la lectura de eventos disjuntos, pero alta para un mismo evento.
  - Consideramos que hay variabilidad propia de cada hablante y que el nivel de dominio de otras lenguas y el lugar de origen también son una fuente de variación.
  - Para verificar la hipótesis, diseñamos una encuesta para administrar en línea, con múltiples oraciones que presenten las seis combinaciones posibles de concordancia del clítico y la lectura del evento. En el experimento consultamos a les participantes por su dominio de otras lenguas y su lugar de origen, además de otras preguntas respecto de edad, profesión y carrera. Para testear la relación entre variables, eliminamos de la muestra a aquellas personas que reportaran un dominio avanzado de cualquier lengua además del castellano y a aquellas personas que vinieran de cualquier lugar que no fuera Capital Federal o provincia de Buenos Aires. Tras testear la relación entre las respuestas de los juicios y los valores de lectura y concordancia, encontramos que las diferencias eran coherentes con nuestras hipótesis iniciales, excepto por el caso de concordancia total con lectura de eventos disjuntos, que predijimos con baja aceptabilidad pero mostró valores altos en la muestra. Aparte, encontramos que la lectura por sí sola no explicaba una cantidad significativa de variación, sino sólo en interacción con la concordancia. Cuando exploramos los valores de las demás

ISSN: 2796-8952

preguntas encontramos que cerca de la mitad de les participantes eran estudiantes o graduades de lingüística, por lo que realizamos un test para ver si había una relación significativa entre la profesión y las respuestas de los juicios que no arrojó diferencias significativas entre lingüistas y no lingüistas.

En ambos ejemplos puede verse una estructura común: primero se define un fenómeno de interés que se releva y a partir del cual se identifican variables de interés. Utilizando esas variables, se proponen hipótesis y se diseña un protocolo para recolección de datos que luego se lleva adelante. Esos datos se testean para comprobar las hipótesis iniciales y luego se proponen nuevas hipótesis en función de los hallazgos. En el segundo caso, también se condujo un nuevo test para evaluar la pertinencia de una nueva hipótesis. Considerando esto, puede decirse que entre exploración y confirmación, el análisis cuantitativo supone los siguientes pasos:

1. Se identifica un fenómeno de interés.
2. A partir de datos ya disponibles e investigación previa, se hace un relevamiento sobre el fenómeno. A partir de este relevamiento se identifican las variables que intervienen en el mismo. Es importante la exhaustividad en la identificación de las variables, dado que vamos a condensar el objeto de estudio en el modo en el que estas variables se distribuyen en las observaciones.
3. Se operacionalizan las variables. Esto quiere decir que se definen explícitamente el tipo de variable y el procedimiento para medirla.
4. Se proponen hipótesis sobre el comportamiento de las variables. Es necesario formularlas de manera tal que pueda expresarse una relación cuantitativa entre ellas.
5. Se define el procedimiento de recolección de observaciones, esto es, se define si se toman muestras en un ambiente natural, si se realiza una encuesta o si se conduce un experimento controlado y el protocolo con el que se hace la toma.
6. Si el propósito es conducir un análisis confirmatorio, se determinan también los modelos estadísticos a emplear para testear las hipótesis.
7. Se realiza la toma de datos.
8. Se llevan a cabo los análisis confirmatorios.
9. Los datos obtenidos se exploran para generar nuevas hipótesis.

2.1. Hipótesis

Una hipótesis estadística supone una relación entre dos conjuntos de variables, donde las variables del primer conjunto permiten explicar la variación en las variables del segundo conjunto. Las del primer conjunto se llaman **variables independientes o explicativas**, mientras que las del segundo conjunto son **variables dependientes o de respuesta**. Un estudio empírico impone restricciones sobre la formulación teórica (Gries, 2013). Para ajustarse a ellas, 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.

Asimismo, siempre hay que tener en cuenta que el modo en el que las variables de interés se definan y operacionalicen determina en gran medida la formulación de las hipótesis.

ISSN: 2796-8952

Al definir hipótesis debe definirse también la condición –o conjunto de condiciones– que falsaría cada una. Hay distintos marcos teóricos para la comprobación estadística de una hipótesis, pero la práctica más convencional, NHST (*null hypothesis significance testing*), estipula que para cada hipótesis propia –**hipótesis alternativa H_1**– es necesario formular una hipótesis opuesta –**hipótesis nula H_0**– tal que si H_1 predice la existencia de un fenómeno determinado, H_0 consiste en la inexistencia de ese fenómeno. Aun si no se adhiere al marco teórico de NHST, es importante tener en cuenta que las hipótesis sólo son relevantes en tanto compiten unas con otras. No hay valor en contrastar empíricamente dos hipótesis que realizan las mismas predicciones (ni es realmente posible, estrictamente hablando).

2.2. Diseño del estudio, probabilidades y espacio muestral

Los aspectos formales y materiales de la recolección de datos son también extremadamente importantes. Por ejemplo, para el caso de la realización de la /r/, múltiples maneras de extraer información son posibles: podríamos tomar una única muestra por sujeto, o seleccionar una cantidad fija de participantes y grabarles en condiciones de uso de registro formal e informal. En el caso de las oraciones con OD coordinado, ¿todes les participantes 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? También debemos tener en cuenta constantemente cuestiones como la cantidad de variables independientes y dependientes en consideración, y si hay alguna interacción observable entre ellas. Distintas respuestas a estas preguntas van a suponer el uso de diferentes pruebas. Además, distintos tipos de estudios suponen una diferencia no sólo en la metodología de prueba de hipótesis, sino también en el modo en el que se interpretan las pruebas. Únicamente en el caso de experimentos controlados pueden inferirse relaciones causales entre variables; en todos los demás, pueden establecerse asociaciones significativas, pero no inferirse causalidad.

Estas consideraciones sobre los datos no son sólo relevantes a la hora de pensar en cómo poner a prueba 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 se trabaja (Abney, 2011). Aun si se hallara que los procesos son defectuosos, eso no implica que las hipótesis sean falsas ni que las teorías sean insuficientes. Sin embargo, sí significa que hay valor en intentar reproducir las predicciones de una teoría empleando una metodología más rigurosa que la que se utilizó para construirla. Además, no hay que perder de vista que el protocolo de recolección de datos y su ejecución son los eslabones más débiles en la realización de un estudio. Si los datos están mal recolectados, no hay test de hipótesis que valga, y los errores de toma y medición son extremadamente frecuentes y fáciles de cometer.

El conjunto de todos los posibles resultados de un estudio 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 del 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. Diferentes marcos teóricos estadísticos van a hacer un trabajo distinto sobre el espacio muestral. El marco de NHST supone evaluar la probabilidad de las observaciones de las que disponemos según la hipótesis nula H_0 y, si tienen una probabilidad muy baja, va a sugerir que aceptemos la hipótesis alternativa H_1. Un marco teórico distinto, como el bayesianismo, supone calcular la probabilidad que cada hipótesis le asigna a cada observación en el espacio muestral, y permite así cuantificar y comparar la credibilidad de varias hipótesis en competencia.

ISSN: 2796-8952

2.3. Estadística descriptiva y exploración visual

Cuando se dispone de una muestra sobre la cual proponer o verificar hipótesis, es muy útil agregar la información presente en cada una de las observaciones de nuestra muestra en una métrica que dé cuenta de las propiedades de la muestra en su conjunto (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. Principalmente, los valores medidos en una estadística descriptiva dan cuenta de la **tendencia central** de la muestra (un único valor que representa al conjunto de las observaciones), la variabilidad observada entre distintos valores de una misma variable (**dispersión**) y la **asociación** entre diferentes variables, o sea, cuánta información da una variable sobre otra. La elección de diferentes estadísticos descriptivos, ya sea de tendencia central, dispersión o asociación, depende del modo en el que las variables se hayan operacionalizado.

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 e interpretar de manera rápida e intuitiva dichas medidas. Las visualizaciones a emplear están fuertemente asociadas a los tipos de variables de interés, así como a su cantidad.

2.4. Estadística inferencial

La **estadística inferencial** constituye 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 una muestra tomando en cuenta una hipótesis determinada. En el marco NHST, como ya se dijo, se considera la probabilidad de las observaciones bajo el supuesto de que la hipótesis nula es correcta. De acuerdo con el resultado de la prueba que apliquemos, se acepta H_1 o se la rechaza en favor de H_0. La aceptación o rechazo de una hipótesis depende de la probabilidad de obtener las observaciones de la muestra si la hipótesis nula fuera cierta. Dos tipos de errores son posibles: el de rechazar la hipótesis nula siendo la misma verdadera en la población, conocido como **error de tipo 1 (falso positivo)**, y el de no rechazar la hipótesis nula siendo esta falsa en la población, conocido como **error de tipo 2 (falso negativo)** (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 detalles, las pruebas paramétricas suelen ser más sencillas en términos de procesamiento, pero tienen una mayor cantidad de supuestos sobre la muestra, mientras que las pruebas no paramétricas suelen ser más simples en términos de lo que presuponen, pero son menos sensibles a tendencias en los datos.

Una cuestión crucial sobre la inferencia es que, en su forma más abstracta, supone la capacidad de "predecir el futuro". Al aceptar una hipótesis une se compromete con la idea de que, ante una eventual nueva muestra de la misma población, pueda predecirse la tendencia de las observaciones. Esto quiere decir que no 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.

ISSN: 2796-8952

3. ¿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, las áreas donde el uso de modelos estadísticos puede enriquecer nuestro conocimiento sobre el lenguaje son más de las que podríamos suponer si nos guiáramos por las teorías más difundidas sobre la gramática. En esta sección se pasa revista a algunas de estas áreas.

3.1. 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 comportamentales (p. ej., tiempo de reacción, movimientos sacádicos) o fisiológicas (p. ej., ERP, flujo sanguíneo). En el caso de los 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, etcétera.

3.2. Lingüística de corpus

La lingüística de corpus tiene la particularidad de que 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 (siempre y cuando el alcance de un estudio se limite al corpus y no busque generalizar a otros textos). 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). Cabe aclarar que nada impide, a priori, el uso de un corpus como herramienta para realizar inferencias sobre una población diferente que lo abarque.

3.3. Lingüística computacional y 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, aun si las soluciones propuestas no responden al procesamiento que hace un hablante (Cohen, 2016). Hay una tradición importante de modelos computacionales basados en teorías 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 hasta modelos de las temáticas abordadas en un corpus (Blei, 2012) o modelos de procesamiento acústico de habla.

ISSN: 2796-8952

3.4. Tipología y lingüística histórica

Dunn et al. (2005) muestran 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 emplear variables extralingüísticas como la distancia geográfica, u otras variables relacionadas con el lenguaje, como la similitud entre palabras a través de cognados. De hecho, esta última opción es la prevalente en los métodos filogenéticos actuales.

3.5. Teoría lingüística

Al incorporar herramientas computacionales como parte de las opciones disponibles a le lingüista para desarrollar sus teorías, es posible generar hipótesis cuantitativas en todos los niveles de estudio del lenguaje, ya se trate del estudio de la sintaxis a través de juicios de aceptabilidad (Schütze y Sprouse, 2014), semántica (Goodman y Lassiter, 2015), o básicamente cualquier otro nivel (Bod et al., 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 et al., 2016). Dado que la lógica de los análisis estadísticos se desprende del modo en que 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.

4. Conclusión – ¿Por qué hay que usar métodos cuantitativos para estudiar el lenguaje?

Sin importar la filiación teórica o el nivel de análisis en el que une se especialice, es posible utilizar el conocimiento previo como asidero para introducirse en un paradigma más riguroso formal y empíricamente. Las razones para usar un abordaje cuantitativo tienen que ver, fundamentalmente, con la posibilidad de replicar hallazgos y de comparar teorías entre sí, no con una perspectiva teórica en particular. Uno de los puntos más importantes de esto es que la adopción de métodos cuantitativos permite la integración de diferentes puntos de vista de manera sistemática y la consideración de información de múltiples fuentes distintas bajo un mismo marco metodológico, independientemente de la filiación teórica de las personas que recolectan la información.

Asimismo, 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 une. La ganancia en precisión teórica que se obtiene con el modelado estadístico del lenguaje representa un avance importante en las capacidades de descripción, explicación y predicción de la teoría lingüística.

Finalmente, es importante no perder de vista que el uso de una metodología cuantitativa no elimina la subjetividad en la investigación, pero sí facilita la integración de distintas subjetividades con un objetivo común.

Abney (2011) dice respecto de la lingüística computacional que (trad. mía):

ISSN: 2796-8952

> 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.

El abordaje estadístico está íntimamente ligado al empleo de herramientas computacionales, con lo cual es razonable considerar que la afirmación de arriba es apropiada también para referirse al abordaje cuantitativo. El giro conceptual asociado al uso de métodos cuantitativos tiene que ver con las posibilidades de refinamiento teórico que proporciona una metodología cuantitativa, en consonancia con el resto de la ciencia contemporánea.

5. Referencias

Abney, S. (2011). Data-intensive experimental linguistics. *Linguistic Issues in Language Technology*. https://doi.org/10.33011/lilt.v6i.1235

Arunachalam, S. (2013). Experimental methods for linguists. *Language and Linguistics Compass*, 7(4), 221–232. https://doi.org/10.1111/lnc3.12021

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods, 2*(2), 131–160. https://doi.org/10.1037/1082-989X.2.2.131

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84. https://doi.org/10.1145/2133806.2133826

Bod, R., Hay, J., & Jannedy, S. (2003). *Probabilistic linguistics*. MIT Press.

Casella, G., & Berger, R. L. (2001). *Statistical Inference* (2da ed.). Duxbury Press.

Cohen, S. (2016). *Bayesian Analysis in Natural Language Processing*. Morgan & Claypool Publishers. https://doi.org/10.2200/S00719ED1V01Y201605HLT035

Downey, A. B. (2011). *Think Stats: Probability and Statistics for Programmers* (2da ed.). O'Reilly Media.

Dunn, M., Terrill, A., Reesink, G., Foley, R. A., & Levinson, S. C. (2005). Structural phylogenetics and the reconstruction of ancient language history. *Science*, *309*(5743), 2072–2075. https://doi.org/10.1126/science.1114615

Freedman, D., Pisani, R., & Purves, R. (1998). *Statistics*. W.W. Norton.

Goodman, N. D., & Lassiter, D. (2015). Probabilistic semantics and pragmatics: Uncertainty in language and thought. *The handbook of contemporary semantic theory* (2da ed.). Wiley-Blackwell.

Gries, S. T. (2009). *Quantitative Corpus Linguistics with R: A Practical Introduction*. Routledge.

Gries, S. T. (2013). *Statistics for Linguistics with R: A Practical Introduction* (2da ed.). De Gruyter Mouton.

Johnson, D. E. (2014). Descriptive statistics. En R. J. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (pp. 288–315). Cambridge University Press.

Klavans, J., & Resnik, P. (1996). *The Balancing Act: Combining Symbolic and Statistical Approaches to Language* (1era ed.). The MIT Press.

Kuznetsova, J. (2013). *Linguistic profiles: Correlations between form and meaning*. [Doctoral dissertation, Universitetet i Tromsø].

Levshina, N. (2015). *How to do linguistics with R: Data exploration and statistical analysis*. John Benjamins Publishing Company.

Manning, C., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

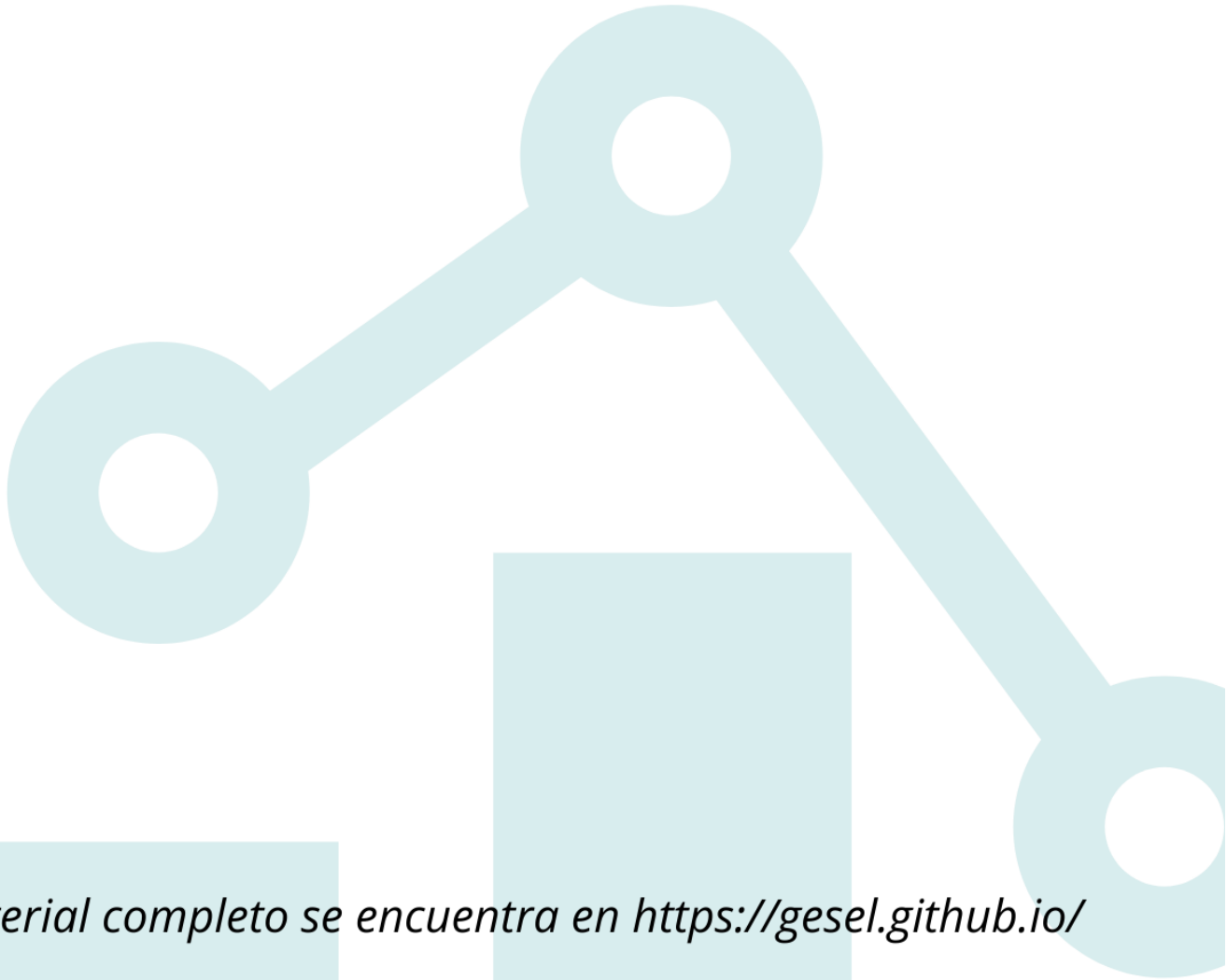
Milin, P., Divjak, D., Dimitrijević, S., & Baayen, R. H. (2016). Towards cognitively plausible data science in language research. *Cognitive Linguistics*, 27(4), 507–526. https://doi.org/10.1515/cog-2016-0055

Schütze, C. T., & Sprouse, J. (2014). Judgment data. En R. J. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (pp. 27–50). Cambridge University Press.

Wikipedia (2018). Statistics — wikipedia, the free encyclopedia. Recuperado el 26 de enero de 2018, de https://en.wikipedia.org/wiki/Statistics

Yang, C. D. (2004). Universal grammar, statistics or both? *Trends in cognitive sciences*, *8*(10), 451–456. https://psycnet.apa.org/doi/10.1016/j.tics.2004.08.006

ISSN: 2796-8952



El material completo se encuentra en https://gesel.github.io/