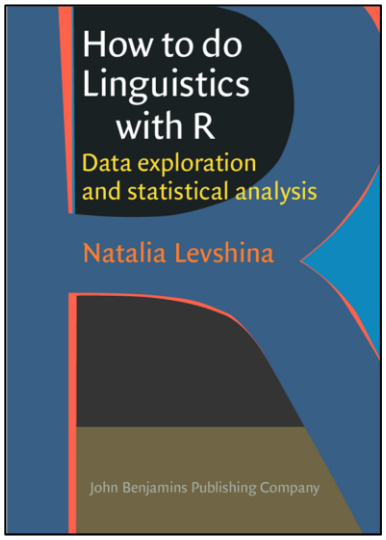
exlibris
Letras

SOBRE *HOW TO DO LINGUISTICS WITH R*, DE NATALIA LEVSHINA

Julia Roberta Carden
Universidad de Buenos Aires
jcarden@filo.uba.ar



∞

How to do Linguistics with R: Data exploration and statistical analysis, de Natalia Levshina; Ámsterdam: John Benjamins, 2015; 443 pp.; ISBN: 978-90-272-1224-5.

La utilización de métodos cuantitativos en el ámbito de la investigación lingüística es un fenómeno cada vez más extendido. Tradicionalmente vinculado a determinadas áreas interdisciplinarias, tales como la psicolingüística, la lingüística computacional, la lingüística aplicada y la neurolingüística, en la actualidad este enfoque está siendo adoptado con creciente frecuencia por otras ramas de la disciplina, que van desde la lingüística diacrónica hasta la sintaxis. En consecuencia, poseer conocimientos estadísticos se ha vuelto necesario para poder evaluar críticamente buena parte de la bibliografía especializada y, en muchos casos, para llevar adelante la propia labor investigativa. *How to do Linguistics with R* se propone ser un manual de estadística dirigido a aquellas personas cuyo interés principal radica en el estudio del lenguaje. Producto de la

experiencia de su autora, Natalia Levshina (Instituto Max Planck de Psicolingüística), en el dictado de cursos de estadística destinados a lingüistas sin formación previa en la materia, *How to do Linguistics with R* es una guía práctica para la ejecución de análisis cuantitativos de datos lingüísticos con el programa informático R. Si bien los conceptos y métodos estadísticos son independientes del campo de estudio al que se aplican, las ventajas de contar con material dedicado específicamente a lingüistas son múltiples. En primer lugar, a diferencia de otros manuales introductorios, este texto describe en detalle métodos no paramétricos y basados en remuestreo, esenciales para el análisis de datos con distribución no normal o escasos, como suelen ser los datos lingüísticos. Asimismo, cada técnica es ilustrada por medio de estudios de distintas áreas de la disciplina, brindando ejemplos claros y útiles de operacionalización de variables y puesta a prueba de hipótesis lingüísticas. Por último, el libro comenta algunas regularidades estadísticas de las lenguas (por ejemplo, la ley de Zipf, el principio de la versatilidad económica y la ley de la abreviación), esclareciendo la naturaleza de variables cruciales, como la frecuencia y la longitud léxicas.

El libro se encuentra estructurado en cuatro partes, la primera de las cuales está conformada por dos capítulos preliminares. El capítulo uno utiliza un estudio sobre diferencias en la tasa de habla entre hablantes de neerlandés de Bélgica y de Holanda para ilustrar conceptos básicos como los de población, muestra, variable, estadística descriptiva, estadística inferencial, prueba de hipótesis, distribución de probabilidad, p-valor, nivel de significación, poder estadístico y grados de libertad. El capítulo dos consiste en una breve introducción a R, programa seleccionado por su popularidad en el ámbito de la investigación científica, su versatilidad y su gratuidad. En este capítulo se brindan instrucciones para su descarga e instalación y se presentan algunos aspectos elementales de su sintaxis. Estar familiarizado con los contenidos de esta primera parte es imprescindible para poder avanzar en la lectura de las tres restantes.

La segunda parte del libro comprende los capítulos tres y cuatro, en los que se aborda la estadística descriptiva de datos cuantitativos y cualitativos, respectivamente. Los datos utilizados en el capítulo tres consisten en la longitud, la latencia media en una tarea de decisión léxica y la frecuencia de una muestra aleatoria de palabras del inglés. En primer lugar, las medidas más típicas de tendencia central y de dispersión son explicadas y calculadas para la variable longitud. A continuación, la distribución de las latencias es explorada mediante diversos métodos gráficos: la creación e interpretación de histogramas, gráficos de densidad, gráficos cuantil-cuantil y diagramas de caja y bigote son estudiadas en detalle. Dado que las latencias parecen poseer una cierta asimetría positiva, se presenta también el test de Shapiro-Wilk como herramienta para evaluar normalidad y se describen distintas estrategias para detectar y tratar valores atípicos. Por último, se examina la frecuencia léxica, variable que posee una distribución extremadamente no normal, tal como predice la ley de Zipf. Las ventajas y desventajas de buscar acercar otros tipos de distribuciones a la gaussiana, así como algunas transformaciones que suelen utilizarse para lograrlo, son analizadas en este punto.

El capítulo cuatro está dedicado a la exploración de variables categóricas. Datos sobre un conjunto de cláusulas clasificadas en función de su transitividad (intransitivas, transitivas o ditransitivas) son utilizados para ilustrar el cálculo de frecuencias, proporciones y porcentajes. Asimismo, se presentan distintas formas de visualización de estas medidas, como los gráficos de torta, de barras y de puntos de Cleveland. En este capítulo también se explica en detalle una medida de dispersión usada en la lingüística de corpus, el desvío de las proporciones. Un estudio

sobre la distribución de los términos básicos para los colores (BCT) en los textos pertenecientes a cuatro registros distintos de un corpus de inglés americano se emplea para esclarecer este concepto.

La tercera parte del libro abarca los capítulos cinco a catorce y se ocupa de la estadística inferencial, es decir, del conjunto de métodos y procedimientos que permiten inferir propiedades de una población a partir de una muestra. El capítulo cinco ofrece herramientas para comparar dos grupos respecto a una característica operacionalizada numéricamente. El trabajo que se presenta como modelo examina si las palabras del inglés de alta y baja frecuencia difieren en cuanto a la cantidad de asociaciones que producen, como sugiere el principio de la versatilidad económica, y a sus puntajes de concretud. La prueba t de Student y la prueba de los rangos con signo de Wilcoxon son descriptas y sus supuestos son explicitados y evaluados.

El capítulo seis introduce el análisis de la relación entre dos variables cuantitativas. ¿Existe un vínculo entre la longitud de una palabra y el tiempo que requiere su reconocimiento? Al intentar dar respuesta a esta pregunta, el texto presenta los coeficientes de correlación de Pearson, Spearman y Kendall, indicando en qué circunstancias debe optarse por cada uno de ellos. Dado que también se explica cómo evaluar la significación estadística de las correlaciones, los conceptos de residuos y homocedasticidad son desarrollados. El capítulo siete, por su parte, está dedicado a la regresión linear, una forma de modelar la relación entre una o más variables predictoras y una variable respuesta numérica. Se retoma el caso trabajado en el capítulo previo, pero buscando dilucidar qué efecto tienen sobre las latencias en el reconocimiento de palabras no solo su longitud, sino también su clase gramatical y su frecuencia. ¿Conocer estas características de una palabra ayuda a predecir cuán difícil será su identificación? A partir de este ejemplo, se estudia cómo ajustar modelos de regresión lineal simple y múltiple. Entre otras cosas, se detalla cómo evaluar las interacciones, verificar el cumplimiento de los supuestos y evitar el sobreajuste. Asimismo, se discuten distintas estrategias de selección de covariables y de identificación de valores atípicos y observaciones influyentes. El capítulo ocho, por último, hace foco en el análisis de la varianza (ANOVA), un caso particular de regresión lineal en el que la o las variables predictoras son categóricas. El texto se estructura alrededor de un trabajo en el que hablantes de una lengua de señas son clasificados de acuerdo a la generación a la que pertenecen (primera, segunda o tercera cohorte) y la edad que tenían cuando fueron expuestos por primera vez a esa lengua (exposición temprana, intermedia o tardía). El objetivo del estudio es establecer si estos factores afectan la cantidad de veces que los hablantes producen una determinada construcción gramatical. En otras palabras, ¿hay diferencias entre los hablantes de distintas cohortes respecto a la frecuencia con que utilizan tal construcción? ¿Y entre los hablantes de la la misma cohorte expuestos temprana y tardíamente a la lengua? Al investigar esta cuestión, el capítulo muestra cómo realizar ANOVAs uni y multifactoriales, de medidas independientes, repetidas y mixtas. Se presentan también alternativas no paramétricas de estas pruebas.

Los capítulos nueve, diez y once están estrechamente vinculados. El primero de ellos se centra en la asociación entre dos variables categóricas. Como modelo se presenta un estudio sobre la relación entre el tipo de uso que se le da a la preposición inglesa *over* (metafórico o no metafórico) y el registro en que aparece (académico o conversacional). ¿Son estas variables independientes? Para explorar los datos, se construyen tablas de contingencia y gráficos de barras apiladas y agrupadas. Distintas medidas de tamaño de efecto, como la razón de probabilidades, la V de Cramer y el coeficiente phi, son calculadas y explicadas. Se presentan también las pruebas chi-

cuadrado de Pearson y exacta de Fisher, que permiten evaluar si la asociación es estadísticamente significativa. El capítulo diez está orientado al estudio de las colocaciones, coligaciones y colostrucciones, por lo que presenta algunas técnicas que permiten medir el grado de atracción entre palabras y otras unidades. Un trabajo sobre la construcción ditransitiva en el ruso y sus colexemas *davat*, *posylat* y *darit* es utilizado para demostrar cómo se calculan las medidas de asociación más típicas, como la atracción, la dependencia, la delta P y el punto de información mutua. El capítulo once, por su parte, está dedicado exclusivamente al análisis distintivo de colexemas, técnica diseñada para comparar o bien dos construcciones casi sinónimas en una misma variedad lingüística en un determinado momento, o bien una misma construcción en momentos históricos distintos o variedades lingüísticas distintas. El ejemplo que se desarrolla en el texto es del último tipo: se comparan las construcciones *quite* + *ADJ* en el inglés británico y el inglés americano por medio de la detección de aquellos colexemas que son atraídos en una variedad y repelidos en la otra. Finalmente, se presenta el análisis distintivo de colexemas múltiple, que es ilustrado mediante la adición de la variedad canadiense al estudio anterior.

Los capítulos doce, trece y catorce explican distintos métodos de clasificación. En el primero de ellos se estudia la regresión logística binaria, análisis que permite modelar la relación entre una o más variables predictoras y una variable respuesta categórica dicotómica. El caso que se propone como modelo examina el uso de los auxiliares *doen* y *laten* en las construcciones causativas del neerlandés. Las variables predictoras son la variedad lingüística (belga u holandesa), el tipo de causalidad (inductiva, volitiva, afectiva o física) y el tipo de predicado efecto (transitivo o intransitivo). ¿Conocer los valores que toman estas variables ayuda a predecir cuál de los dos auxiliares se utilizará? Al abordar este problema, el texto demuestra cómo ajustar y hacer el diagnóstico de un modelo de regresión logística. Brinda asimismo estrategias de selección de covariables y se detiene particularmente en la interpretación de los resultados. En el capítulo trece se explica cómo realizar este mismo tipo de análisis cuando la variable respuesta toma más de dos valores, es decir, se estudia la regresión logística multinomial. Por último, el capítulo catorce presenta métodos de clasificación no paramétricos, útiles en situaciones en las que no es posible ajustar o interpretar un modelo de regresión logística. Un estudio que explora las construcciones causativas del inglés, en el cual la variable respuesta puede tomar tres valores (*make* + *V, have* + *V* y *cause* + *toV),* es utilizado para ilustrar cómo ajustar, evaluar e interpretar árboles de inferencia condicional y cómo calcular el impacto de las distintas variables predictoras a través de la generación de bosques aleatorios.

La cuarta y última parte del libro presenta algunos métodos multivariados, no supervisados y exploratorios, sumamente valiosos a la hora de detectar estructuras en los datos. Los capítulos quince y dieciséis se centran en técnicas propias de la semántica distribucional, es decir, basadas en la noción de que los datos de corpus proporcionan frecuencias distribucionales y que la similitud distribucional refleja similitud semántica o funcional. El primero de ellos introduce el análisis de perfil comportamental a través de un estudio en el que se examina si distintas construcciones causativas con el mismo auxiliar (por ejemplo, *make_V* y *be_made_toV*) tienen distribuciones similares o si, por el contrario, poseen propiedades idiosincrásicas. En primer lugar, el texto explica qué son y cómo pueden crearse vectores con los perfiles comportamentales de cada construcción. Luego, muestra cómo obtener una matriz de distancias entre estos vectores, medidas que representan el grado de similitud entre las construcciones respecto a las variables contextuales incluidas en los perfiles. Por último, discute algunos métodos de segmentación (de clusterización

jerárquica y de partición), útiles para evaluar la presencia y composición de subgrupos de construcciones en los datos. El capítulo dieciséis, por su parte, constituye una introducción general a los espacios vectoriales semánticos. Al igual que el análisis de perfil comportamental, este método realiza una comparación entre los contextos de aparición de los ítems, pero la información distribucional es ahora extraída automáticamente de un corpus anotado. Mediante un ejemplo en el que se analiza la relación entre diversos términos provenientes del campo semántico de la cocina en el inglés, se detalla cómo obtener frecuencias de coocurrencia para cada uno de ellos y calcular su similitud coseno, sobre la que se aplica luego el método de segmentación.

El capítulo diecisiete está destinado al estudio del escalamiento multidimensional, un conjunto de técnicas útiles para la exploración visual de datos multivariados. En el caso que se presenta como ejemplo, setenta y seis variedades del inglés son analizadas en función de veinte características de su gramática. El texto explica cómo obtener una representación gráfica de las semejanzas y diferencias entre esas variedades en dos dimensiones o más. El diagrama resultante permite detectar no solo posibles subgrupos dentro de los datos sino también fuentes interpretables de variabilidad. En este capítulo se indica cómo obtener las distancias de Gower, cómo elegir la cantidad óptima de dimensiones a partir del cálculo del *stress* de Kruskal y cómo evaluar la correlación entre dos matrices mediante el test de Mantel.

El capítulo dieciocho se centra en el análisis de componentes principales y el análisis factorial, métodos que buscan reducir una gran cantidad de variables numéricas correlacionadas a un conjunto más pequeño de dimensiones subyacentes. Para ilustrar su uso, se presenta un análisis multidimensional de la variación de registros en el Corpus Nacional Británico. En este estudio, de la información sobre las frecuencias normalizadas de once elementos gramaticales en sesenta y nueve subcorpus emergen tres dimensiones latentes que explican más de tres cuartas partes de la varianza original. Este capítulo demuestra cómo seleccionar la cantidad óptima de dimensiones, cómo visualizarlas y estudiar su composición y cómo utilizar variables cualitativas suplementarias para facilitar su interpretación. El capítulo diecinueve, por su parte, introduce el análisis de correspondencias, un método similar a los abordados en el capítulo anterior en cuanto a su finalidad, pero diseñado para explorar variables categóricas. Los datos sobre los términos básicos para los colores (BCT) en distintos registros del Corpus de Inglés Americano Contemporáneo, vistos en el capítulo cuatro, son utilizados para explicar cómo representar asociaciones entre variables en la menor cantidad de dimensiones posibles. La selección e interpretación de las dimensiones, la creación de mapas perceptuales y la introducción de elementos suplementarios son estudiadas. Finalmente, el capítulo veinte presenta los gráficos en movimiento como un método útil para la visualización dinámica del cambio lingüístico. En el caso que se toma como modelo, se exploran los cambios en el uso del futuro con *will* y *be going to* en el inglés americano durante el período 1820-2000 por medio de dichos gráficos, cuya creación e interpretación es explicada en detalle.

How to do Linguistics with R nos invita a acercarnos al estudio del lenguaje desde un enfoque cuantitativo. Sin ser abrumador, el texto presenta una enorme variedad de métodos estadísticos, que el lector podrá luego estudiar en mayor profundidad de acuerdo con sus necesidades e intereses. Quizás el mayor mérito de este libro resida en la claridad con la que logra ilustrar la aplicación de tales métodos a la resolución de problemas lingüísticos reales.