Artículo de edición especial

Métodos sociológicos e investigación 2022,
vol. 51(4) 1580–1633 © The Author(s)
2022 Pautas para la reutilización
de artículos: sagepub.com/
journals-permissions DOI:
10.1177/00491241221099551
journals.sagepub.com/home/smr

¿Quién hace qué a quién? Cómo hacer que los analizadores de texto funcionen para la investigación sociológica

Oscar Stuhler1

Abstracto

Durante la última década, los sociólogos se han interesado cada vez más en el estudio formal de las relaciones semánticas dentro del texto. La mayoría de los estudios contemporáneos se centran en el mapeo de co-ocurrencias de conceptos o en la medición de asociaciones semánticas a través de incrustaciones de palabras. Aunque conducen a muchos objetivos de investigación, estos enfoques comparten una limitación importante: abstraen lo que se puede llamar la estructura de eventos de los textos, es decir, la acción narrativa que tiene lugar en ellos. Mi objetivo es superar esta limitación mediante la introducción de un nuevo marco para extraer relaciones semánticamente ricas del texto que involucra tres componentes. Primero, una gramática semántica estructurada alrededor de entidades textuales que distingue seis clases de motivos: acciones de una entidad, tratamientos de una entidad, agentes que actúan sobre una entidad, pacientes sobre los que una entidad actúa, caracterizaciones de una entidad y posesiones de una entidad; en segundo lugar, un conjunto integral de reglas de mapeo, que hacen posible recuperar motivos de las predicciones de los analizadores de dependencia; tercero, un paquete R que permite a los investigadores extraer motivos de sus propios textos. El marco de trabajo se demuestra en análisis empíricos sobre la interacción de género en novelas y construcciones de identidad colectiva por parte de los candidatos presidenciales estadounidenses.

[1] Departamento de Sociología, Universidad de Nueva York, Nueva York, NY, EE. UU.

Autor para correspondencia:
Oscar Stuhler, Departamento de Sociología, Universidad de Nueva York, 295 Lafayette Street, 4th Floor, New York, NY
10012-9605.

Correo electrónico: stuhler@nyu.edu

Palabras clave análisis de contenido, análisis de texto, análisis de dependencia, sociología computacional, gramática semántica, narrativa, ciencias sociales computacionales, incrustaciones de palabras, sociología cultural, procesamiento de lenguaje natural

1. Introducción

Durante la última década, los sociólogos se han interesado cada vez más en el estudio formal de las relaciones semánticas dentro del texto. Alentado por la creciente disponibilidad de grandes corpus de texto y recursos computacionales, este programa de investigación comprende una variedad de enfoques. En lo que podría llamarse análisis de redes semánticas, los investigadores derivan y estudian gráficos de co-ocurrencia de conceptos (p. ej., Hoffman et al. 2018; Fuhse et al. 2020; Lee y Martin 2014; Rule, Cointet y Bearman 2015; Padgett et al. 2020 ; Luz 2014). Otros estudios agregan estructuras de coocurrencia para identificar grupos semánticos de términos que concurren con frecuencia, generalmente interpretados como temas (p. ej., Fligstein, Stuart Brundage y Schultz 2017; DiMaggio, Nag y Blei 2013; Karell y Freedman 2019). Más recientemente, los académicos se han movido hacia el uso de incrustaciones de palabras (neuronales) para estudiar distribuciones de similitudes semánticas (p. ej., Kozlowski, Taddy y Evans 2019; Nelson 2021; Stoltz y Taylor 2021).

A pesar de los méritos y la promesa de este floreciente programa de investigación, la mayoría de los enfoques utilizados actualmente por los sociólogos comparten una limitación importante: las relaciones bajo investigación son de co-ocurrencia o, en el caso de incrustaciones de palabras, distancias semánticas derivadas de co-ocurrencia. Estas formas de representación, aunque propicias para muchos objetivos de investigación, están subespecificadas semánticamente de una manera que limita nuestra capacidad para analizar textos. Específicamente, tienden a abstraer lo que se puede llamar la estructura de eventos de los textos (Franzosi 1989: 276; 2009: 16-17), es decir, la acción narrativa que tiene lugar en un texto. Sin embargo, muchas de las preguntas que tradicionalmente se han planteado los sociólogos al estudiar textos giran en torno a los acontecimientos narrativos y la acción. En pocas palabras, en lugar de saber si A y B coexisten o son similares, los sociólogos a menudo quieren saber qué le hace A a B y viceversa. Esto se vuelve especialmente evidente en el trabajo sobre relatos y narrativa (Polletta et al. 2011; Franzosi 1998a; Abell 2004), pero también en los trabajos pioneros de la disciplina que estudiaron formalmente las relaciones en los textos pero las extrajeron mediante codificación manual (Carley 1993, 1994; Carley y Palmquist 1992; Bearman y Stovel 2000; Bearman, Faris y Moody 1999; Tilly 1997; Franzosi 1998b, 1994, 1989, 1990, Duquenne, Mohr y Le Pape 1998; Mohr 1994; Martin 2000; Smith 2007).

En este artículo, demuestro que una forma de superar esta limitación de la mayoría de los métodos contemporáneos es recurrir a analizadores de dependencia, modelos que predicen las relaciones sintácticas entre palabras dentro de oraciones. Los analizadores de dependencia son una gran promesa para la investigación sociológica porque tienen el potencial de extraer relaciones semánticamente ricas de datos textuales. Sin embargo, solo unos pocos sociólogos han utilizado analizadores de dependencia en sus investigaciones (ejemplos recientes son Goldenstein y Poschmann 2019; Stuhler 2021). Esto es sorprendente, dado que los analizadores automáticos han existido desde al menos principios de los años 90, y los modelos que se implementan listos para usar mientras logran altos niveles de precisión recientemente están disponibles en entornos de software populares entre los sociólogos, como R y Pitón.

Sospecho que la razón principal de esto es un desajuste entre el formato de información que predicen los analizadores de dependencia y el tipo de información que quieren los sociólogos. Los analizadores de dependencia generan estructuras gráficas dirigidas bastante complejas en las que los bordes entre las unidades léxicas se etiquetan de acuerdo con su relación de dependencia sintáctica. Las gramáticas de dependencia modernas como el marco Universal Dependencies (Universal-Dependencies 2020) abarcan más de 60 tipos de relaciones sintácticas. Los sociólogos, por otro lado, generalmente buscan información semántica comparativamente menos compleja, como la pregunta de qué le hizo A a B o qué características se les atribuyen.

Para superar este desajuste, propongo un marco que involucra tres componentes. Primero, una gramática semántica simple, centrada en la entidad, que distingue seis clases de motivos: acciones de una entidad, tratamientos de una entidad, agentes que actúan sobre una entidad, pacientes1 sobre los que actúa una entidad, caracterizaciones de una entidad y posesiones de una entidad . entidad. En segundo lugar, un conjunto integral de reglas de traducción derivadas inductivamente que mapean árboles de dependencia complejos en esta gramática. En tercer lugar, un paquete R que permite a los investigadores extraer motivos de sus propios textos según la gramática propuesta.

Demuestro la utilidad de este marco en dos análisis, utilizando corpus de texto de diferentes discursos y de diferentes tamaños. Primero, analizo las relaciones de género en un corpus de novelas estadounidenses (1880-2000). Muestro que las interacciones mujer-mujer son tan frecuentes como las interacciones hombre-hombre en las novelas escritas por mujeres, pero considerablemente menos en las escritas por hombres. También investigo la asimetría en las relaciones entre géneros (por ejemplo, ¿los hombres besan a las mujeres o viceversa?) y sigo su evolución a lo largo del tiempo. Encuentro que el género se ha vuelto menos predecible según el contenido de interacción durante el siglo XX . En segundo lugar, investigo los intentos de los candidatos presidenciales estadour

construir una identidad compartida con su audiencia en discursos de campaña (1950-2020). Específicamente, me enfoco en el uso del pronombre plural en primera persona “nosotros” e identifico diferentes retóricas que se usan para constituir este “nosotros”. Entre otras cosas, muestro que en los últimos años ha habido una tendencia entre los candidatos a aumentar su énfasis en la identidad de campaña a medida que se acerca el día de las elecciones.

2. Enfoques relacionales del texto en sociología

La mayoría de los análisis de contenido cuantitativos tradicionales proceden designando frases clave o códigos conceptuales particulares y luego contando la presencia o ausencia de estos elementos en un conjunto de documentos. Sin embargo, a fines de los años 80 y 90, varios sociólogos comenzaron a adoptar un enfoque relacional para estudiar textos. Más que asignar y contar códigos, estos enfoques buscaban analizar formalmente las relaciones semánticas afirmadas por un texto. A continuación, reviso este programa de investigación, incluidos sus desarrollos más recientes. Debido a que muchos de los enfoques empleados por los sociólogos tienen antecedentes en otras disciplinas, vale la pena enfatizar que mi enfoque se encuentra en el trabajo dentro de la sociología.

Una de las primeras contribuciones a este programa fue la propuesta de Franzosi de estudiar textos desagregándolos en eventos de la forma sujeto-acción-objeto (SAO), construcciones a las que se refirió como “tripletes semánticos” (Franzosi 1989, 1990, 1994). De manera similar, Abell (1987) buscó definir un formalismo general para representar estructuras narrativas. También entre las primeras contribuciones a este programa están los esfuerzos de Roberts (1989, 1997) para diseñar una gramática semántica genérica. Sin embargo, su gramática, formalmente ambiciosa, no solo apuntaba a analizar las cláusulas de evento en sus constituyentes, sino también a clasificarlas de acuerdo con su función discursiva. Finalmente, Carley (1993) desarrolló el análisis MAP, un enfoque que vio en contraste explícito con el análisis de contenido tradicional. La principal innovación del análisis MAP fue representar los datos textuales como redes en las que los conceptos forman nodos y los bordes especifican la relación entre ellos. Qué tipo de relaciones son permisibles y qué conceptos son relevantes se deja al investigador (Carley y Palmquist 1992). En la mayoría de las aplicaciones, este enfoque se usó para generar representaciones en red de modelos mentales a partir de transcripciones de entrevistas (Carley 1988; Carley y Palmquist 1992), pero otras incluyen un estudio sobre imágenes cambiantes de robots en novelas de ciencia ficción (Carley 1994).

Estas contribuciones iniciales fueron seguidas por un conjunto de estudios que buscaban aplicar esta forma de análisis a distintas áreas problemáticas. Algunos de ellos se centraron en la diferenciación entre diferentes tipos de identidades.
Ejemplos notables incluyen las contribuciones de Mohr sobre las categorías de pobreza.

y su asociación con formas particulares de ayuda (Mohr 1994; Mohr y Duquenne 1997; ver también Mohr y Lee 2000) y el trabajo de Martin (2000) sobre animales y sus ocupaciones en la literatura infantil. Otro grupo de académicos se centró en estudiar las estructuras de las redes narrativas, en las que los nodos son eventos y los bordes son vínculos causales o lógicos entre ellos afirmados por el texto (Bearman y Stovel 2000; Bearman et al.
1999; Smith 2007). Mientras tanto, la extracción de relaciones a partir de datos textuales también se hizo popular entre los estudiosos de los movimientos sociales y la política contenciosa. En una de las aplicaciones más impresionantes de una gramática de tripletes semánticos, Tilly (1995, 1997) analiza las reuniones polémicas en la Gran Bretaña de los siglos XVIII y XIX. De un corpus de periódicos, extrajo más de 50.000 informes de acciones relacionales entre clases de actores políticos. Estos fueron luego sometidos a un análisis de modelo de bloque para identificar diferentes facciones políticas. Desde entonces, se han aplicado gramáticas semánticas similares para codificar informes periodísticos sobre eventos de protesta y reivindicaciones políticas (p. ej., Koopmans y Statham 1999; Wada 2004). Finalmente, otra clase de estudios de esta primera fase del análisis de contenido relacional se basa en la dualidad de textos y elementos textuales. Ejemplos de esto incluyen el análisis de Mische y Pattison (2000) de afirmaciones en manifiestos y discursos de organizaciones políticas en Brasil, el trabajo de McLean (1998) sobre elementos retóricos en cartas que buscan favores en la Florencia del Renacimiento, así como el análisis de Ruef (1999) de la salud. actividades relacionadas y formas organizativas en un corpus de publicaciones mediales. Estas obras son conceptualmente distintas de las anteriores, sin embargo, en la medida en que las relaciones entre las entidades textuales surgen de su distribución en los textos, más que de su asociación significativa dentro de un texto.

En la última década, los sociólogos han recurrido cada vez más a la automatización del proceso de extracción de relaciones semánticas del texto. Quizás el enfoque más directo es lo que se podría llamar análisis de redes de palabras, que tiene algo de historia fuera de la sociología (p. ej., Danowski 1993). En este enfoque, el texto se transforma en un gráfico donde los nodos representan palabras o frases y los bordes representan alguna medida de co-ocurrencia dentro de unidades superiores de texto como documentos, párrafos o ventanas en movimiento. En algunos casos, se aplica un algoritmo de detección comunitaria para detectar grupos temáticos dentro de la red. Ejemplos de este enfoque incluyen el análisis de las controversias científicas de Leydesdorff y Hellsten (2006), el estudio de Light (2014) sobre la toma de posesión presidencial de los Estados Unidos, el mapeo de Lee y Martin (2014; véase también Lee y Martin 2014) de los conjuntos de herramientas conceptuales de diferentes pensadores de la Escuela de Frankfurt. , el análisis de Rule et al. (2015) de los cambios en el discurso del Estado de la Unión de EE. UU., la investigación de Fuhse y colegas (2020) sobre lo

significados de “Volk” en la República de Weimar de Alemania, y el análisis del discurso de Padgett et al. (2020) en el consejo florentino Consulte e Pratiche”.

Un enfoque relacionado que se ha vuelto muy popular entre los sociólogos en los últimos años es el modelado de temas, donde las estructuras temáticas se infieren de la co-ocurrencia de palabras en unidades superiores de texto (p. ej., DiMaggio et al. 2013; Fligstein et al. 2017; Karell y Freedman 2019; Mohr y Bogdanov 2013). Los temas se formalizan como distribuciones de probabilidad sobre el vocabulario de funciones. Aparte de su carácter inductivo, sin embargo, este enfoque guarda una mayor semejanza con el análisis de contenido tradicional: en lugar de investigar las relaciones entre las entidades textuales, el objetivo principal suele ser etiquetar los documentos o sus subsecciones con respecto a la prevalencia relativa de diferentes temas.

Un enfoque más directamente relacionado con las relaciones entre unidades textuales es el estudio de la similitud semántica a través de incrustaciones de palabras, también conocida como semántica vectorial. Aquí, las palabras en un corpus se representan como vectores numéricos de igual longitud: están incrustados en un espacio vectorial. Hay diferentes formas de construir estos espacios vectoriales. En principio, se puede considerar que el vector asociado con una palabra en una matriz documento-término es una incrustación. Sin embargo, por lo general, se llevan a cabo pasos de procesamiento adicionales para hacer que las incrustaciones sean relativamente cortas, densas e informativas con respecto al contenido semántico de una palabra, lo que conceptualmente mueve la similitud de las incrustaciones hacia nociones de similitud semántica y lejos de la idea de co-ocurrencia. Por ejemplo, en su análisis de la Biblia protestante y su uso, Hoffman y sus colegas (2018) primero generaron una matriz palabra por palabra en la que una celda captura la cantidad de veces que dos palabras coexisten en una ventana de texto en movimiento (valor ponderado). por distancia en el texto), al igual que en los enfoques basados en la co-ocurrencia discutidos anteriormente. Sin embargo, en lugar de estudiar esta matriz en sí misma como una red, los autores la tratan como un espacio vectorial y calculan las similitudes de coseno entre el vector incrustado de cada palabra.
Luego, los autores pasan a generar y estudiar un gráfico en el que los nodos son palabras y los bordes son similitudes semánticas medidas como similitudes de coseno entre incrustaciones (para un enfoque similar, consulte Puetz, Davis y Kinney 2021).

Hace tiempo que existen técnicas más elaboradas para generar incrustaciones de palabras (p. ej., Deerwester et al. 1990), pero el gran avance de este enfoque en la sociología vino con la capacidad de entrenar eficientemente incrustaciones de palabras de alta calidad a través del ampliamente popular word2vec (Mikolov et al. . 2013) y algoritmos GloVe (Pennington, Socher y Manning 2014) (para revisiones orientadas a aplicaciones de ciencias sociales, véase Arseniev-Koehler 2021; Spirling y Rodríguez 2022). En una aplicación original, Kozlowski y colegas

(2019) utilizan incrustaciones de word2vec entrenadas en Google Ngram Corpus para generar medidas para una serie de dimensiones culturales (p. ej., masculino-femenino, rico-pobre, moral-inmoral). Esto se hace promediando las diferencias de incrustaciones de varios pares de palabras que representan una dimensión (p. ej., "hombre" y "mujer", "masculino" y "femenino", "él" y "ella"). Los vectores así generados se utilizan luego para rastrear la asociación entre diferentes dimensiones culturales a lo largo del tiempo y para proyectar sobre ellas palabras que representan prácticas culturales. Con ligeras variaciones, este enfoque ha sido adoptado en varios artículos de sociología muy recientes. Los ejemplos incluyen estudios sobre género y sus asociaciones cambiantes con estereotipos educativos (Boutyline, Arseniev-Koehler y Cornell 2020) o dominios sociales en general (Jones et al. 2020), conceptos científicos y su evaluación diferencial en culturas de investigación cuantitativa y cualitativa (Kang y Evans 2020), conceptos relacionados con la obesidad y sus asociaciones con diferentes dimensiones culturales (Arseniev-Koehler y Foster 2020), la evolución del discurso de la inmigración estadounidense (Stoltz y Taylor 2021), la interseccionalidad de raza y género en el siglo XIX. Sur de EE. UU. (Nelson 2021), y cambios en el entorno semántico del concepto de “red” (Yung 2021). Ligeramente diferente es el enfoque desarrollado por Stoltz y Taylor, que se centra en estimar la similitud semántica de documentos completos con conceptos focales (2019) o dimensiones culturales (Taylor y Stoltz 2021). En una extensión reciente, los autores muestran cómo se puede usar este enfoque para medir esquemas subyacentes a un texto (Taylor y Stoltz 2020).

La Tabla 1 ofrece una descripción general del trabajo sociológico que se involucra formalmente con las relaciones semánticas en los datos textuales. Enumera los tipos de conceptos textuales estudiados y los tipos de relaciones entre ellos, así como también cómo se extrajeron estos elementos. Los trabajos anteriores que se basaban en la codificación manual tenían un interés definido en relaciones particulares o clases de las mismas: categorías de identidad tratadas (Mohr), actores políticos actuando hacia otros actores políticos (Tilly), animales haciendo trabajos (Martin), eventos siendo causados por otros eventos (Bearman y colegas) y actores, o, más ampliamente, sujetos que actúan hacia objetos (Franzosi, Abell, Carly). En el trabajo contemporáneo, continúa habiendo variación en los temas, pero con el giro hacia los enfoques computacionales, la literatura parece haber convergido en gran medida en dos tipos de relaciones: co-ocurrencia y similitud semántica.

¿Qué debemos hacer con este desarrollo? Sin duda, los recientes avances metodológicos han llevado el estudio formal y relacional del texto a un nuevo nivel de popularidad entre los sociólogos. Lo que antes era un discurso de nicho ahora se puede encontrar en las revistas insignia de la disciplina. No obstante, los primeros estudios que pasaron por la molestia de extraer manualmente semántica

Cuadro 1. Visión general de los enfoques formales y relacionales del texto en sociología.

Publicaciones	Extracción métodos	Clases de textual conceptos	Tipos de relación
(Abel 1987)	Codificación manual	[Agentes] [Eventos]	[Provocando]
(Carley 1988, 1993, 1994; Carley y palmquist 1992)	Mano asistida por computadora codificación	[Conceptos]	[Relación] Nota: Carley lo deja en manos del investigador. para definir clases de relaciones elegibles.
(Franzosi 1989, 1990, 1994)	Codificación manual	[Asignaturas] [Objetos]	[Comportamiento]
(Roberts 1989, 1997) Codificación manual		[Asignaturas] [Objetos] [Comportamiento] [Función discursiva] [Modo] [Tiempo]	Probabilidad Nota: de Roberts esquema completo codifica cláusulas en 17 (1997) variables distintas y luego investiga estadísticas asociaciones entre ellos.
(Mohr 1994; Mohr y Duquenne 1997),	Mano asistida por computadora codificación	[Categorías de pobreza] [Formas de ayuda]	Tratamiento
(Mohr y Lee 2000) Computadora	mano asistida codificación	[Discursos de identidad] [Prácticas de divulgación]	Tratamiento
(Hasta 1995, 1997)	Codificación manual	[actor político]	[Acciones políticas dirigidas]

(continuado)

Tabla 1. Continuación

Publicaciones	Extracción métodos	Clases de textual conceptos	Tipos de relación
(McLean 1998)	Codificación manual	[Palabras clave]	Co-ocurrencia
(Ruef 1999)	Computacionales	[Formas organizacionales] [Relacionados con la salud actividades]	Co-ocurrencia
(Bearman y Stovel 2000; Bearman et al. 1999; Smith 2007)	Codificación manual	[Eventos]	Causal o lógico conexión
(Martín 2000)	Codificación manual	[Animales] [Trabajos]	Haciendo
(Mische y Pattison 2000)	Codificación manual	[Objetivos políticos]	Co-ocurrencia
(Light 2014; Leydesdorff y Hellsten 2006; Fuhse et al. 2020; Lee y Martin 2014; Padget et al. 2020)	computacional	[palabras]	Co-ocurrencia
(DiMaggio et al. 2013; Fligstein et al. 2017; Karell y Liberto 2019; Mohr y Bogdanov 2013)	Computacional	[Temas] [Palabras]	Probabilidad
(Mohr et al. 2013)	Computacional	[Conceptos]	[Comportamiento]
(Regla et al. 2015)	Computacional	[sintagmas nominales]	Co-ocurrencia
(Lee y Martin 2018)	Computacional	[Menciones del autor]	Co-ocurrencia

(continuado)

Tabla 1. Continuación

Publicaciones	Extracción métodos	Clases de textual conceptos	Tipos de relación
(Hoffman et al. 2018)	Computacional [Palabras]		similitud semántica
(Kozlowski et al. 2019)	Computacional [Dimensiones culturales]		similitud semántica
(Stoltz y Taylor 2019)	Computacional [Conceptos focales]	[Documentos]	similitud semántica
(Boutyline et al. 2020)	Computacional [Estereotipos educativos]		similitud semántica
		Género dimensión	
(Jones et al. 2020)	Computacional [Dominios sociales]	Género dimensión	similitud semántica
(Arseniev-Koehler y Fomentar 2020)	Computacional [Dimensiones	culturales] [Palabras relacionadas con la obesidad]	similitud semántica
(Kang y Evans 2020)	Computacional [Palabras]	Dimensión evaluativa	similitud semántica
(Taylor y Stoltz) 2021; taylor y Stoltz 2020)	Computacional [Dimensiones	culturales] [Documentos]	similitud semántica
(Nelson 2021)	Computacional [Dimensiones	culturales] [Instituciones sociales]	similitud semántica
(Stoltz y Taylor 2021)	Computacional [Dimensiones	culturales]	similitud semántica

(continuado)

Tabla 1. Continuación

Publicaciones	Extracción métodos	Clases de textual conceptos	Tipos de relación
		[Palabras relacionadas con la inmigración]	
(Yung 2021)	Concepto de red computacional	[Palabras]	similitud semántica
(Puetz et al. 2021)	Computacional	[sintagmas nominales]	similitud semántica

Nota: Las filas están ordenadas por el primer año de publicación respectivamente. Los corchetes denotan conjuntos en el sentido de que no todos los elementos son idénticos. Esta tabla es el mejor esfuerzo del autor para brindar una descripción general aproximada de una gran cantidad de enfoques y estudios empíricos diferentes. Algunos autores podrían encontrar justificadamente que la complejidad de su enfoque está subrepresentada (p. ej., Roberts 1989, 1997). Otros definen sus esquemas en términos explícitamente flexibles (p. ej., Abell 1989) o realizan análisis variados (p. ej., Kozlowski et al. 2019, Stoltz y Taylor 2021), de modo que su representación en la tabla es de acuerdo con lo que parece ser la aplicación principal .

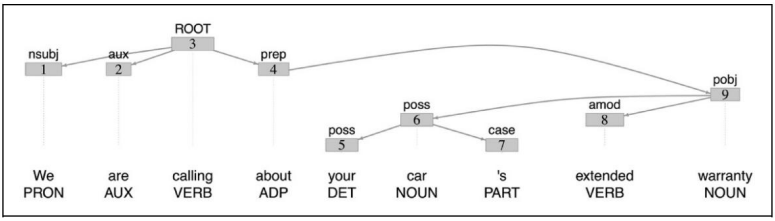


Figura 1. Una oración y el árbol de dependencia correspondiente.
Nota: Las relaciones sintácticas entre tokens se dirigen como lo indican las flechas que apuntan desde la cabeza hasta el token dependiente. La representación de los árboles de dependencia en este documento sigue la convención de colocar la etiqueta de una relación de dependencia sobre el nodo dependiente. Por ejemplo, “Nosotros” es el sujeto nominal (nsubj) de “llamar”.

Las relaciones de los textos son un testimonio del hecho de que muchas preguntas de investigación sociológica van más allá de la co-ocurrencia y la similitud semántica. Se refieren a eventos narrativos (Franzosi 1989: 276; 2009: 16-17) en los que las entidades textuales participan en acciones o son destinatarios de ellas. En particular, esto implica que giran en torno a tipos de relaciones que los enfoques actualmente dominantes no captan. En lo que sigue, describo una forma de superar esta limitación y, por lo tanto, intento reconciliar la literatura contemporánea con sus orígenes.

3. ¿Qué son los analizadores de dependencia?

Una forma de superar las limitaciones descritas de muchos métodos contemporáneos es recurrir a analizadores de dependencia. Los analizadores de dependencia son una gran promesa para la investigación sociológica porque tienen el potencial de extraer relaciones semánticamente ricas de datos textuales. Sin embargo, dentro de la sociología, solo unos pocos han aprovechado esta oportunidad (Stuhler 2021; Goldenstein y Poschmann 2019; Mohr et al. 2013; en ciencias políticas, véase van Atteveldt, Kleinnijenhuis y Ruigrok 2008; van Atteveldt et al. 2017). Mi ambición aquí es primero, identificar y explicar algunas de las complejidades involucradas en hacer que los analizadores de dependencia funcionen para la investigación sociológica; segundo, proporcionar un marco y software asociado que supere estos problemas; y tercero, mostrar el potencial de este marco en los análisis empíricos. Comienzo con una revisión de lo que hacen los analizadores de dependencia (para introducciones más extensas, consulte Jurafsky y Martin 2020; Eisenstein 2019: 243–266).

Los analizadores de dependencia son modelos que predicen las relaciones sintácticas entre elementos léxicos (principalmente palabras) dentro de oraciones. La Figura 1 muestra una oración en inglés que está anotada con un árbol de dependencia. Los árboles de dependencia son gráficos dirigidos que especifican las relaciones entre los elementos léxicos de una oración Lo que cuenta como un árbol de dependencia permisible depende de la gramática de dependencia (se proporciona un glosario de términos potencialmente desconocidos en el material complementario). La mayoría de las gramáticas comparten una serie de propiedades restrictivas: hay un nodo raíz con grado de entrada 0 (generalmente el verbo principal de la oración); todos los demás nodos tienen un grado de entrada de 1 o, dicho de otra manera, cada nodo no raíz depende exactamente de un nodo principal; el gráfico está completamente conectado; no hay ciclos en el gráfico. La justificación y el origen de estos

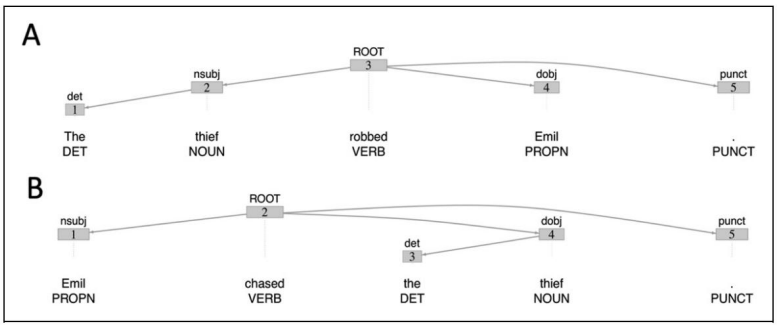


Figura 2. Dos oraciones simples con su correspondiente árbol de dependencia.

las propiedades restrictivas se encuentran más allá del alcance de este documento (para una buena revisión, consulte de Marneffe y Nivre 2019).

Además de estas propiedades estructurales de los árboles de dependencia, las gramáticas de dependencia definen el rango de posibles tipos de dependencia. Estos especifican la función gramatical que desempeña un dependiente con respecto a su cabeza. A lo largo de este artículo, me baso en el esquema ClearNLP (Choi y Palmer 2012), que está estrechamente relacionado con el esquema de dependencia de Stanford (de Marneffe y Manning 2008). En la Figura 1, por ejemplo, “Nosotros” es el sujeto nominal (nsubj) del verbo “llamar” y el límite entre los dos términos está etiquetado en consecuencia. La raíz de la oración, "llamando", tiene dos dependientes más: "son", que sirve como auxiliar (aux) y "sobre", que introduce una frase preposicional (prep).

Los analizadores de dependencias se entrenan y evalúan en conjuntos de datos con relaciones sintácticas anotadas llamadas treebanks. Para el inglés, la mayoría de estos conjuntos de datos se generan mediante la aplicación de reglas de transformación a los treebanks de circunscripciones existentes (p. ej., Choi y Palmer 2012). Los modelos que fueron entrenados en estos datos luego están disponibles como software y pueden implementarse más o menos de manera inmediata en otros textos. En la práctica, los analizadores de dependencia generalmente están integrados en canalizaciones de anotaciones integrales que procesan datos textuales sin procesar mediante la resolución secuencial de un conjunto de tareas, incluida la tokenización de palabras, la división de oraciones, el etiquetado de partes del discurso y la lematización. La información de estas tareas anteriores es la base para el análisis (para obtener información sobre la mecánica de los analizadores de dependencia, consulte Jurafsky y Martin 2020; Eisenstein 2019: 243–266). Existen múltiples canalizaciones de procesamiento populares que incluyen el análisis de dependencias (p. ej., Stanford CoreNLP de Manning et al. 2014; o UDPipe de Straka 2018). Para este documento, utilizo un pipeline2 proporcionado por la biblioteca spaCy (Honnibal et al. 2020), que se implementó recientemente en R (Benoit y Matsuo 2020).

Los árboles de dependencia proporcionan relaciones sintácticas, es decir, relaciones con respecto a la estructura gramatical de una oración. Las relaciones sintácticas pueden no ser de interés per se para la mayoría de los sociólogos. Sin embargo, pueden proporcionar una aproximación de las relaciones semánticas, es decir, relaciones sobre el significado de una oración. Para ilustrar esto, considere la oración en los Paneles A y B de la Figura 2, e imagine que estamos interesados en las acciones en las que está involucrado "Emil". Un análisis de co-ocurrencias nos permitiría inferir que Emil está asociado con "robar" y con "perseguir", pero no está claro si es perseguido y robado o perseguido y robado o una combinación de estos dos. Al construir sobre las relaciones sintácticas, podemos ir más allá de esta información e inferir si él es el agente (hacedor)

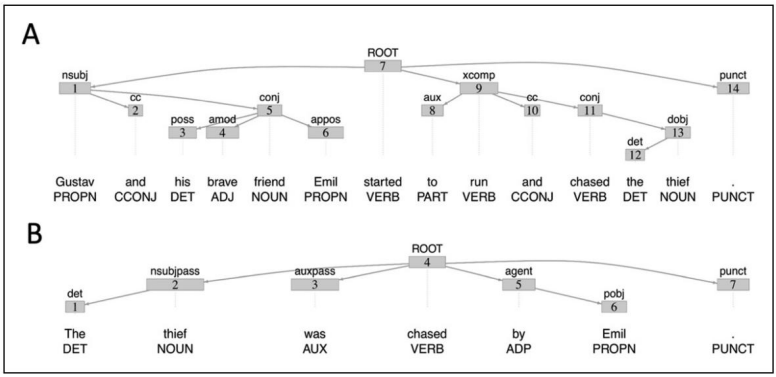


Figura 3. Dos oraciones complejas con su correspondiente árbol de dependencia.

o paciente (receptor) de estas acciones. Para hacer esto, podríamos escribir dos reglas simples:

(a) Encuentre todos los verbos “Emil” es el sujeto nominal de, es decir, todos los nsubj-heads de “Emil”.

(b) Encuentra todos los verbos “Emil” es el objeto acusativo directo de, es decir, todos los dobj cabezas de “Emil”.

Aplicando estas reglas, ahora podríamos inferir que Emil es asaltado y perseguido. Si quisiéramos ir más allá, podríamos expandir nuestras reglas para incluir cualquier dobj-dependiente de los nsubj-heads de Emil, así como los nsubj-dependientes de los dobj-heads de Emil. Esto nos permitiría inferir que Emil fue asaltado por un “ladrón” y está persiguiendo a un “ladrón”; por lo tanto, captaríamos literalmente todo el significado de ambas oraciones.

Si bien iríamos más allá de la simple coocurrencia con nuestras dos reglas, existe un problema: la mayoría de las oraciones que encontramos en los datos textuales reales son sintácticamente más complejas que el ejemplo. Como referencia, la oración promedio en el New York Times contiene 23 palabras (Tauberg 2019). Para ilustrar esto, considere la oración en el Panel A de la Figura 3. Esta oración tiene bastante más que el primer ejemplo. Entre otras cosas, nos enteramos de que Emil se caracteriza por ser “valiente” y que está “corriendo”.

No obstante, una lectura semántica de esta oración aún relativamente corta nos informa que Emil está persiguiendo a un ladrón. Sin embargo, las cosas se complican considerablemente más cuando pasamos a las relaciones sintácticas de la oración.

Primero, “Emil” es un modificador aposicional (appos) del término “amigo”; “amigo” está vinculado a “Gustav” a través de una relación conjunta (conj); “Gustav” es el sujeto nominal (nsubj) de “comenzó” – la raíz de la oración; “iniciado” está vinculado al verbo “ejecutar” a través de una relación de complemento de cláusula abierta (xcomp); “correr” está vinculado a través de otra relación conjunta (conj) con “perseguido”; finalmente, “perseguido” nuevamente tiene el objeto directo (dobj) “ladrón”.
En lugar de una conexión directa y un tipo de relación sintáctica, ahora tenemos una longitud de ruta de cinco y cinco tipos distintos de relación sintáctica entre "Emil" y "chase". La oración en el Panel B de la Figura 3 ilustra otro punto: las oraciones no necesariamente tienen que agregar mucha complejidad para confundirnos con nuestras dos reglas. Las relaciones sintácticas de una cláusula pasiva simple se ven bastante diferentes a las de una activa.

Como Franzosi señaló hace mucho tiempo, los sociólogos generalmente se preocupan por las relaciones semánticas más que sintácticas (Franzosi 1989: 271-272). Las relaciones sintácticas derivables a través del análisis de dependencias pueden proporcionar una base para establecer relaciones semánticas. Sin embargo, son demasiado detallados para ser de mucha utilidad inmediata para los sociólogos, quienes generalmente tienen poca consideración por categorías como "modificador aposicional" o "complemento de oración abierta". Los pocos artículos de sociología que han utilizado analizadores han dejado esta distinción en gran parte sin reconocer (Goldenstein y Poschmann 2019; Mohr et al. 2013; para el mismo punto de crítica, ver Monroe 2019; aunque ver Stuhler 2021). Hacer que los analizadores de dependencia funcionen para la investigación sociológica requiere que reconozcamos y abordemos este desajuste.

4. De la sintaxis a la semántica: proponiendo una gramática semántica centrada en la entidad

¿Cómo podemos superar la brecha entre las relaciones sintácticas de grano fino predichas por los analizadores de dependencia y la información semántica que la mayoría de los sociólogos cuidan?3 En esta sección, propongo una gramática semántica relativamente simple, centrada en entidades, que involucra seis clases de elementos: acciones de una entidad, tratamientos de una entidad, agentes que actúan sobre una entidad, pacientes sobre los que actúa una entidad, caracterizaciones de una entidad y posesiones de una entidad. Argumentaré que podemos mapear las complejas relaciones sintácticas predichas por los analizadores de dependencia en esta gramática semántica a través de un conjunto de reglas de transformación. Más adelante, me refiero a elementos de la gramática semántica como motivos. Uso categorías semánticas como agente, acción y paciente en lugar de categorías sintácticas como sujeto, predicado y objeto para resaltar que los motivos son elementos semánticos.

La gramática propuesta aquí se basa en lo que Franzosi llamó un triplete semántico que, con ligeras variaciones, ha sido un núcleo canónico de gramáticas semánticas para microestructuras textuales (ver, por ejemplo, Franzosi 1989: 273-274; Propp [1928] 1968: 113; van Dijk 1972:287; Todorov y Weinstein 1969:74). La principal diferencia es que mi gramática está centrada en la entidad, por lo que las clases de motivos se distinguen con respecto a la relación que tienen con una entidad central de interés. Esta elección anticipa un caso de uso en el que los investigadores están interesados en extraer relaciones semánticas en torno a un concepto particular o una clase del mismo. En lugar de representar textos completos, el objetivo es extraer y representar las afirmaciones que hace un texto con respecto a estos

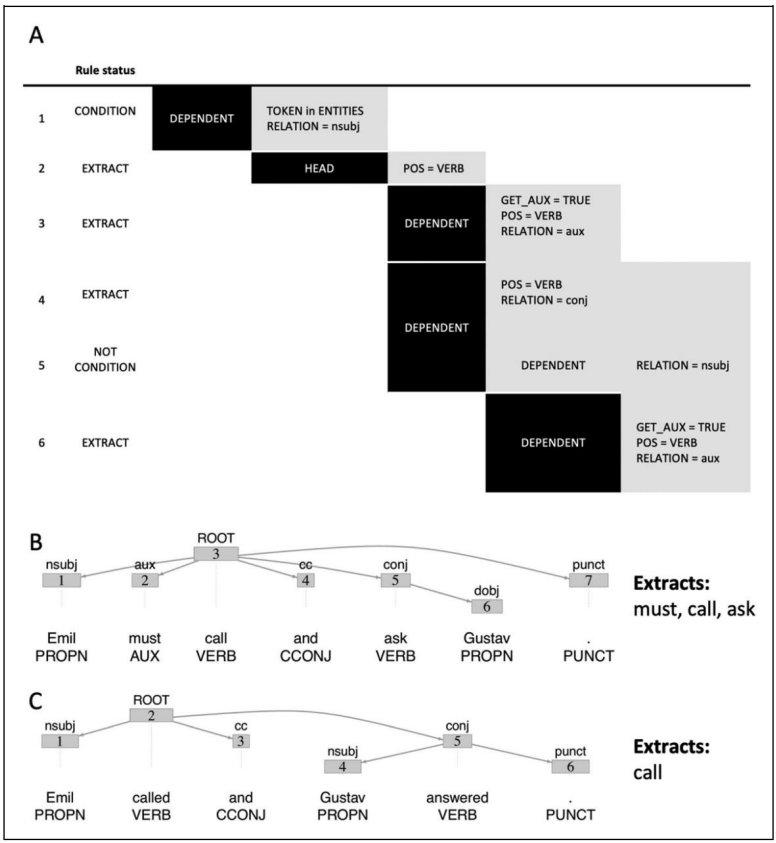


Figura 4. Regla de extracción y dos oraciones ejemplares.

entidades. El trabajo revisado anteriormente corrobora que esta es una preocupación central de la investigación. Además, al incluir caracterizaciones y posesiones, agrego dos motivos de estasis: motivos que se relacionan con estados más que con sucesos (para una discusión de los elementos de estasis en las narrativas, véase Franzosi 2009: 18-19; Chatman 1978: 31-33).

El mapeo de las estructuras sintácticas predichas por el analizador de dependencias en la gramática semántica propuesta se produce a través de un conjunto de reglas de extracción. Estas reglas se crearon en un proceso iterativo e inductivo. Primero, comencé con un conjunto de reglas de extracción simples como las dos formuladas en la Sección 3. Luego, apliqué estas reglas a un conjunto de textos para extraer motivos. Tomé muestras de oraciones y, con base en una lectura semántica del texto, evalué si las reglas dejarían de lado elementos textuales que deberían haber sido extraídos como motivos o extraerían los que no deberían haber sido. A este respecto, la sección anterior ilustró cómo una regla simple (extraer todos los nsubj encabezados de una entidad) logra extraer todas las acciones en un contexto (Oración B en la Figura 2) pero no lo hace en otros (Oraciones A y B en la Figura 2). 3). Estas evaluaciones condujeron, con una probabilidad decreciente a medida que avanzaba el proceso, a la adición de nuevas reglas o al perfeccionamiento de las reglas existentes.

Las reglas especifican criterios con respecto a tokens, etiquetas de parte del discurso y relaciones de dependencia. El código anotado que implementa estas reglas está disponible en forma de un paquete R, semgram, que actualmente está en desarrollo y alojado en Github.4 semgram se basa en funcionalidades de spacyr (Benoit y Matsuo 2020) para análisis de dependencia y rsyntax (Welbers, van Atteveldt y Kleinnijenhuis 2021) para implementar reglas de extracción.

Si bien las reglas en sí mismas son intrincadas y demasiado numerosas para discutirlas aquí, su lógica general se demuestra con un ejemplo. Considere la regla que se muestra en el panel A de la Figura 4, que extrae motivos de acción del texto en torno a una entidad central de interés. Imagina que aplicamos esta regla para extraer acciones de Emil (hiperparámetro: ENTIDADES=[“Emil”]) en las oraciones de los paneles B y C; también especificamos que nos interesa extraer auxiliares como acciones (hiperparámetro: GET_AUX =TRUE). La primera línea busca todos los elementos léxicos cuya representación simbólica se encuentra entre las entidades de interés (FICHAS EN ENTIDADES) que son sujetos dependientes nominales (RELACIÓN =nsubj) de una cabeza. La línea 2 procede a extraer todas las cabezas que son verbos (POS= VERBO). Esto lleva a la extracción de "llamar" en ambas oraciones.

En la línea 3, procedemos a extraer todos los dependientes verbales (POS= VERBO) auxiliares (RELACIÓN =aux) de los elementos identificados en la línea 2. Esto conduce a la extracción de “debe” en la oración del Panel B. Tenga en cuenta que no esto porque especificamos extraer auxiliares como acciones (GET_AUX = TRUE), lo que puede no ser deseable para algunas aplicaciones. En la línea 4, también

Tabla 2. Oraciones ejemplares y extracciones para diferentes clases de motivos.

clase de motivo	Ex.	Oración	Extracto	Comentario
Acción	a.1	Llamadas de ENTIDAD.	una llamada	Cabezas verbales ("llamadas") de las cuales ENTIDAD es el nominal sujeto son las acciones más frecuentes.
	a.2	La ENTIDAD puede llamar.	una_lata, una_llamada	Los verbos auxiliares pueden ser acciones consideradas (hiperparámetro).
	a.3	ENTIDAD llama y pregunta.	una_llamada, una_pregunta	Los verbos en conjunción se consideran acciones.
	a.4	John y ENTITY llamaron.	una llamada	Casos en los que la ENTIDAD es una conjunto dependiente del sujeto nominal ("Juan") se consideran para identificar comportamiento.
	a.5	John fue llamado por ENTIDAD.	una llamada	Las construcciones pasivas con "by" se consideran para identificar acciones.
	a.6	Mi amigo ENTIDAD llamó a John.	una llamada	Instancias en las que ENTIDAD sirve como modificador aposicional de un sujeto nominal (amigo) puede considerarse acciones (hiperparámetro).
	a.7	ENTIDAD quiere llamar.	a_want, a_call	Un verbo en un componente clausal se considera una acción, siempre que no tenga un sujeto nominal dependiente.
	a.8	ENTIDAD quiere que llames.	un querer	
Paciente	P.1	ENTIDAD pregunta John.	P_Juan	Los pacientes suelen ser objetos directos de todos los verbos transitivos. identificadas como acciones.
	P.2	Pedro y ENTIDAD pregunta a John.	P_Juan	
	P.3	Mi amigo ENTIDAD pregunta John.	P_Juan	
	P.4	vino la ENTIDAD y le preguntó a Juan.	P_Juan	
	P.5	ENTITY quiere preguntarle a John.	P_Juan	
	P.6	ENTIDAD llama a John, Jane y Steve.	P_John, P_Jane, P_Steve	Los objetos conjuntos múltiples son considerados simultáneamente como pacientes.
	P.7	ENTIDAD pregunta	P_Juan, P_pregunta	Objetos dativos ("Juan") de

(continuado)

Tabla 2. Continuación

clase de motivo	Ex.	Oración	Extracto	Comentario
		Juan una pregunta.		También se consideran acciones pacientes
	P.8	John es preguntado por ENTIDAD.	P_Juan	Los sujetos pasivos nominales ("John") también se consideran pacientes.
Tratamiento	t.1	John llama ENTIDAD.	t_llamar	Los tratamientos suelen ser verbos de los que ENTITY es el objeto directo o dativo.
	t.2	John le da a la ENTIDAD una manzana.	t_dar	
	t.3	Juan le da a Pedro una entidad.	t_dar	
	t.4	Juan llama a Pedro y ENTIDAD.	t_llamar	Los casos en que ENTIDAD es un conjunto de un objeto directo o dativo, se consideran en los tratamientos de extracción.
	t.5	John le dio a Peter una manzana y ENTIDAD.	t_dar	
	t.6	ENTIDAD fue llamado.	t_llamar	Verbos de los cuales ENTITY es una sujetos pasivos nominales son tratos.
Agente	A.1	John llama ENTIDAD.	A_Juan	Los agentes son a menudo los nominales sujetos de verbos identificados como tratamientos de ENTIDAD.
	A.2	Juan le da a Pedro una ENTIDAD.	A_Juan	
	A.3	Peter y John preguntan ENTIDAD.	A_Peter, A_Juan	Las reglas especificadas con respecto a la relación entre la ENTIDAD y las acciones se aplican a la relación entre el agente y el tratamiento, incluidos los sujetos en conjunción, los verbos en conjunción, las aposiciones o los componentes de la cláusula.
	A.4	Se pregunta ENTIDAD por John.	A_Juan	
	A.5	John vino y preguntó ENTIDAD.	A_Juan	
	A.6	John quiere preguntarle a ENTIDAD.	A_Juan	
	A.7	Mi amigo John le preguntó a su hermano ENTIDAD.	Un amigo, A_Juan	

(continuado)

extraer la conjunción verbal (RELACIÓN =conj) dependientes de los elementos de la línea 2, lo que conduce a la extracción de “preguntar” en la oración del Panel B. Por sí mismo, esto también conduciría a la extracción de “respondió” en la oración del Panel B. panel C, del cual Emil claramente no es el agente. Fue sólo a través de la

Tabla 2. Continuación

clase de motivo	Ex.	Oración	Extracto	Comentario
Caracterización	be.1	ENTIDAD es amable.	ser_amable	La mayoría de las caracterizaciones son dependientes adjetivales de un verbo en cópula (ser, convertirse, permanecer, sentir, mirar y otros).
	be.2	ENTIDAD se ve triste.	estar_triste	
	be.3	ENTIDAD es el ganador.	ser_ganador	Las caracterizaciones también pueden ser dependientes de atributos nominales de un verbo en cópula.
	ser.4	ENTIDAD siguió siendo presidente.	ser_presidente	
	ser.5	ENTIDAD podría ser el presidente.	ser_presidente	Las modificaciones mediante verbos auxiliares (“could”) no afectan el estado de caracterización.
	be.6	ENTIDAD es amable y honesto.	ser_amable, ser_honesto	Los dependientes conjuntos de caracterizaciones se consideran caracterizaciones.
	be.7	ENTITY ganó pero permaneció triste.	estar triste	Al igual que con las acciones, el copular verbo puede tomar diferentes posiciones para que su adjetivo o dependiente nominal sea considerado una caracterización. Los modificadores de adjetivos se consideran caracterizaciones.
	be.8	ENTIDAD va estar triste.	estar triste	
	be.9	ENTIDAD espera permanecer presidente.	ser_presidente	
	be.10	John compró un nuevo y barato ENTIDAD.	ser_barato, ser_nuevo	Los modificadores de adjetivos se consideran caracterizaciones.
	ser.11	El ganador fue ENTIDAD.	ser_ganador	Sujetos nominales de una copular verbo con ENTIDAD como el atributo dependiente puede ser caracterizaciones consideradas (hiperparámetro).
	ser.12	Los ganadores eran Juan y ENTIDAD.	ser_ganador	
	be.13	Mi hermano Ganó la ENTIDAD.	ser_hermano	Cuando la entidad sirve como modificador aposicional, su cabeza se considera una caracterización.
Posesión	H.1	de la ENTIDAD pareja, amigos y los padres se sorprendieron.	H_cónyuge, H_amigo, H_padre	Los sustantivos, junto con sus dependientes en conjunción, que tienen ENTIDAD como modificador de posesión se consideran posesiones.
	H.2	Los descansos y		Cabezas nominales de una preposición

(continuado)

Tabla 2. Continuación

clase de motivo	Ex.	Oración	Extracto	Comentario
		ruedas de la ENTIDAD eran viejos.	H_romper, H_rueda	con “de” que tiene ENTIDAD como objeto dependiente se consideran posesiones.
	H.3	LA ENTIDAD tiene amigos y enemigos.	H_amigo, H_enemigo	Objetos directos del verbo “tener” y sus inflexiones son posesiones consideradas. Nota: “Have” y sus inflexiones no se consideran como motivo de acción. es directo los dependientes de objeto o dativo no se consideran pacientes.
Acción-paciente	aP.1	ENTIDAD pregunta John.	aP_preguntar_a John	Esta es una clase de motivo compuesto, que toma todos los motivos del paciente y los fusiona con la acción respectiva.
	aP.2	ENTIDAD hizo y comió un pastel.	aP_eat_cake	Un sustantivo puede ser objeto de varios verbos transitivos al mismo tiempo (“Hice y me comí el pastel”), pero la gramática sintáctica utilizada aquí no permite inferir la relación entre un segundo verbo y el objeto, de modo que sólo hay una acción por paciente en los motivos acción-paciente (y sólo un tratamiento por agente en los motivos agente-tratamiento).
Agente-tratamiento	At.1	John pregunta ENTIDAD.	En_John_preguntar	Esta es una clase de motivos compuestos, que toma todos los motivos de los agentes y los fusiona con el tratamiento respectivo.

proceso de prueba de reglas sobre datos reales, me di cuenta de que tal regla conduciría a extracciones erróneas cada vez que un segundo sujeto nominal ("Gustav" en este caso) fuera dependiente del verbo en conjunción. Esto me llevó a revisar la regla y agregar la condición negativa en la línea 5, especificando que los elementos de la línea 4 solo se extraerían si no tenían un sujeto dependiente nominal (RELACIÓN =nsubj). Finalmente, en la línea 6, extraemos los dependientes auxiliares verbales de los elementos de la línea 4. Esto no conduce a más extracciones en nuestras dos oraciones.

La Tabla 2 brinda ejemplos para cada clase de motivo y enumera algunos casos más específicos para ilustrar el alcance de las reglas. Estos casos pueden no parecer complejos en su semántica. Sin embargo, tal vez en contra de la intuición, las estructuras sintácticas subyacentes son tales que requieren una consideración especial. La columna de comentarios de la Tabla 2 y la discusión a continuación brindan más detalles al respecto. Tenga en cuenta, además, que la tabla no es exhaustiva de todos los escenarios y reglas posibles. De hecho, la complejidad del conjunto de reglas formales surge principalmente de la necesidad de dar cuenta de la combinación de diferentes escenarios que se muestran en la Tabla 2, digamos una combinación de modificación aposicional, conjunciones de sujeto y verbo y un componente de cláusula abierta, como se ve en la oración en el Panel B de la Figura 3. La Tabla 2 también ilustra el estilo de marcado utilizado para distinguir las diferentes clases de motivos que, como veremos a continuación, pueden ser útiles para procesar y analizar los motivos extraídos. Las letras al comienzo de las palabras indican la clase de motivo respectivo, de modo que a_call, por ejemplo, implica que la entidad de interés participó en la acción de llamar, mientras que t_call se usa cuando llamar es un tratamiento de la entidad.

Los motivos de acción implican que la entidad de interés está haciendo algo. El ejemplo más sencillo de esto es cuando la entidad sirve como sujeto nominal de un verbo (ejemplo a.1). Hay varias construcciones sintácticas, sin embargo, en las que un verbo se considera una acción a pesar de que la entidad no sea su sujeto nominal. Esto incluye instancias en las que la entidad es el conjunto de un sujeto nominal (a.4), hay varios verbos (a.2, a.3, a.7), la entidad sirve como modificador aposicional de un sujeto nominal (a.6), y construcciones pasivas (a.5). Todas las acciones son verbos léxicos o, si se especifican explícitamente, verbos auxiliares.

Los motivos del paciente son cosas hacia las que actúa la entidad de interés. Suelen ser objetos de verbos transitivos que se identificaron como la acción de una entidad.
Estos objetos pueden estar en caso acusativo (P.1-P.6) o en caso dativo si el verbo es ditransitivo (P.7). Cualquier motivo de acción puede conducir a múltiples motivos de Paciente, ya que cualquier verbo transitivo puede tener múltiples objetos conjuntos (P.6).
Más allá de los objetos, los sujetos nominalmente pasivos también son considerados pacientes (P.8).

Los motivos de tratamiento implican que se hace algo a una entidad de interés. Este es el caso cuando la entidad es el objeto de un verbo transitivo. La relación entre los tratamientos y la entidad es análoga a la de las acciones y los pacientes.
La entidad puede funcionar como objeto acusativo (t.1, t.2) o dativo (t.3), como sujeto nominal pasivo (t.6), o como conjunción de cualquiera de estos (t.4, t.5) .

Los motivos de agente son cosas que actúan hacia la entidad de interés a través de un motivo de tratamiento. En la mayoría de los casos, los agentes son el sujeto nominal de un verbo que ha sido identificado como motivo de tratamiento (A.1, A.2). Sin embargo, los agentes no necesitan tomar esa posición y pueden ser conjuntos (A.3) o modificadores aposicionales

(A.7) del sujeto nominal. Generalmente, la relación entre agentes y tratamientos es análoga a la del ente y las acciones, por lo que el verbo transitivo puede tomar posiciones diferentes (A.5, A.6), y construcciones pasivas en las que el ente actúa como sujeto nominal pasivo (A.4) se consideran.

Más allá de estos motivos de proceso, existen dos clases de motivos de estasis. Las caracterizaciones son características adscritas a la entidad de interés. Hay varias maneras en que esto puede suceder. El más común es a través de un verbo copular, que tiene un adjetivo (be.1, be.2, be.6, be.7, be.8) o nominal (be.3, be.4, be.5 , be.9) dependiente del atributo. Sin embargo, los adjetivos también pueden ser dependientes directos de la entidad (be.10) para ser considerados caracterizaciones. Además, se consideran caracterizaciones los sujetos nominales de los verbos copulares con la entidad como atributo dependiente (be.11, be.12) y las cabezas con la entidad como modificador aposicional (be.13).

Las posesiones son cosas que se dice que posee la entidad de interés. El conjunto de reglas explica tres formas en las que esto puede expresarse. Primero, cuando la entidad sirve como modificador de posesión de un sustantivo, dicho sustantivo y sus dependientes en conjunto se consideran posesiones (H.1). En segundo lugar, las construcciones donde la entidad sirve como objeto dependiente de la preposición “de” pueden dar lugar a pos sesiones (H.2). Tercero, si la entidad sirve como sujeto nominal de “have” o una de sus flexiones, su objeto directo y las conjunciones nominales del mismo se consideran posesiones. Tenga en cuenta que "tener" es un verbo transitivo, pero dentro de la gramática, no se considera una acción y, en consecuencia, sus objetos no se consideran pacientes.

Finalmente, hay dos clases de motivos compuestos que vinculan acciones y pacientes (aP.1), así como agentes y tratamientos (At.1). Estos son esencialmente lo que Franzosi llamó tripletes semánticos centrados en la entidad de interés.
Si bien la representación de datos de esta manera puede conducir a altos niveles de escasez, como algunos han señalado (Monroe 2019), demostraré en la sección 5.1 que puede ser un medio poderoso para analizar las representaciones de las relaciones sociales.

5. Aplicación

¿Por qué los sociólogos deberían preocuparse por las gramáticas semánticas y la posibilidad de extraer motivos del texto? En esta sección, utilizo el marco presentado para realizar análisis sobre dos temas sociológicos clásicos: las relaciones de género y la identidad colectiva. Demostraré cómo el marco proporciona una nueva compra analítica sobre estos temas.

5.1 ¿Quién besa a quién? Relaciones de género en la literatura estadounidense (1880-2000)

Gran parte del trabajo reciente revisado en la sección dos estudia la representación del género en los datos textuales. El enfoque actualmente dominante para esto es usar primero palabras clave de género (por ejemplo, "hombre" y "mujer", "niño" y "niña", "masculino" y "femenino") para estimar el género como una dimensión cultural en un espacio de incrustación de palabras (Kozlowski et al. 2019; Boutyline et al. 2020; Jones et al. 2020; Nelson 2021; para una excepción notable a este enfoque, consulte Underwood 2019). En un segundo paso, los investigadores examinan la correlación o el ángulo del coseno entre la dimensión de género y varias otras dimensiones culturales, lo que se interpreta como similitud o asociación semántica. Pero, ¿cuál es el significado de esta medida de asociación semántica? ¿Qué nos dice exactamente, por ejemplo, si encontramos que la feminidad en lugar de la masculinidad está asociada con la riqueza en Google Ngram Corpus (Kozlowski et al. 2019: 922–923)? Según los autores, este patrón se debe a que históricamente las mujeres sirvieron como recipientes para el consumo de los hombres. Esta explicación es plausible. Sin embargo, también señala el hecho de que las relaciones descubiertas por las incrustaciones de palabras pueden subdeterminarse semánticamente y que interpretarlas sigue siendo un desafío.

Las relaciones derivables con una gramática semántica son menos susceptibles a este problema, ya que su nivel de abstracción es considerablemente menor. En lugar de cómo el género se alinea semánticamente con varias dimensiones o conceptos culturales, una gramática semántica nos permite captar cómo las entidades de diferentes géneros se involucran en la acción narrativa y qué atributos y características se les atribuyen explícitamente. Las relaciones de una gramática semántica no solo se mantienen más cercanas al texto original, sino también a las ontologías conceptuales de la mayoría de las teorías sociológicas que, con pocas excepciones, involucran cosas como actores, acciones y atributos. En pocas palabras, en lugar de relaciones semánticas abstractas, podemos capturar representaciones de relaciones sociales.

Para demostrar esto, utilizo el US Novel Corpus (USNC), una colección de 9088 novelas estadounidenses publicadas entre 1880 y 1990 (Chicago-Text-Lab 2021; So, Long and Zhu 2019). Las novelas se seleccionaron en función de la cantidad de existencias de la biblioteca enumeradas en Worldcat.org y, por lo tanto, involucran publicaciones de mercado masivo y altamente canónicas (consulte el material complementario para obtener estadísticas adicionales del corpus). Para identificar entidades de género, utilizo el registro de nombres de la Administración de la Seguridad Social de los EE. UU., que enumera la frecuencia de todos los nombres que se dieron a los niños recién nacidos en los EE. UU. durante cualquier año entre 1880 y 2000 (Wickham 2021). Los uso para construir listas de nombres dados que son indicativos de hombres

de personajes femeninos (ver material complementario para más detalles). Además, considero todas las palabras en mayúsculas precedidas por "Sr.", "Sra.", "Señorita" y "Señora" como entidades de género, así como los pronombres "él", "él", "su", "ella". ," y ella." A continuación, la canalización de lenguaje spaCy se usa para la tokenización de palabras, la división de oraciones y la lematización, así como para anotar el texto con etiquetas de parte del discurso y árboles de dependencia. Finalmente, todas las oraciones se procesan con el paquete semgram R para extraer motivos alrededor de tokens que representan entidades masculinas y femeninas.

Para este análisis, me concentro solo en motivos de acción-paciente, de los cuales hay 7,5 millones para entidades centrales masculinas y 4,6 millones para entidades centrales femeninas. En un paso siguiente, tomo el subconjunto de todos los motivos de acción-paciente que tienen una entidad femenina o masculina en la posición del paciente (1,2 millones). Distinguiéndolos por género de entidad central y género del paciente, genero cuatro conjuntos de motivos: acciones de mujer-hombre, acciones de hombre-mujer, acciones de mujer-mujer y acciones de hombre-hombre. Me refiero a "personajes" masculinos y femeninos a continuación, aunque técnicamente, hay algunos casos en los que las entidades a las que se hace referencia con pronombres o nombres de género pueden no ser técnicamente personajes humanos (por ejemplo, animales, barcos o lugares). La figura 5 proporciona una representación esquemática del flujo de trabajo hasta este punto.

Curiosamente, la mayor parte de la interacción en las novelas estadounidenses parece ocurrir a través de líneas de género (ver Tabla 3). De todos los motivos femeninos de acción-paciente con paciente identificado de género, el 73,7% están dirigidos a personajes masculinos. De todos los motivos de pacientes de acción masculina con paciente identificado por género, el 69,3% están dirigidos a personajes femeninos. En otras palabras, es casi tres veces más probable que una acción de un personaje de género identificado en una novela esté dirigida a un personaje del género opuesto que a un personaje del mismo género. Este es un nivel considerable de heterofilia de género.

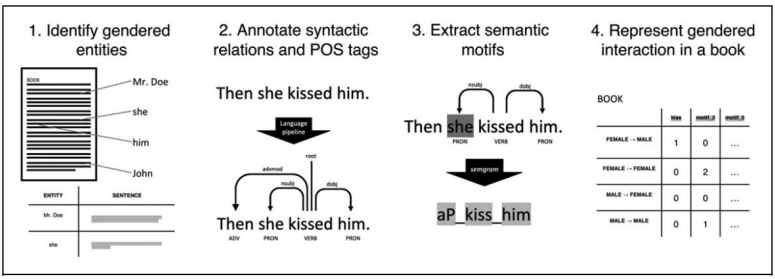


Figura 5. Flujo de trabajo del análisis de la interacción de género en las novelas.

Tabla 3. Frecuencias relativas de acciones del mismo género e intergénero en novelas estadounidenses.

	paciente masculino	paciente femenino	Total
Acciones masculinas	17,3%	39,1%	56,4%
Acciones femeninas	32,2%	11,5%	43,7%
Total	49,5%	50,6%	

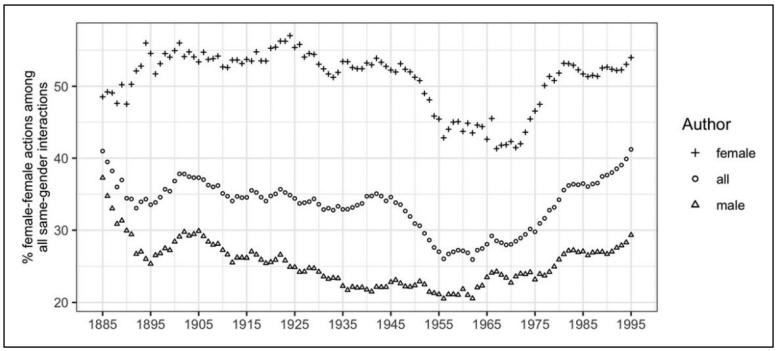


Figura 6. Proporción de acciones mujer-mujer entre interacciones del mismo género.
Nota: Los puntos muestran el porcentaje de acciones mujer-mujer entre todas las interacciones del mismo género promediadas sobre libros escritos dentro de un período de 10 años. Un valor en 1950, por ejemplo, representa todas las publicaciones entre 1946 y 1955.

Un punto de referencia popular para evaluar la representación de las mujeres en las obras de ficción es el Bechdel-Test, que pasa una película o un libro si contiene dos personajes femeninos que hablan entre sí sobre un tema que no sea un hombre. Un número sorprendentemente elevado de obras de ficción no cumplen con este criterio básico, incluida aproximadamente la mitad de las películas que han sido nombradas Mejor Película en los Oscar (BBC 2018). Si bien la gramática propuesta aquí no nos permite evaluar si un libro cumple con el criterio de Bechdel, sí nos permite evaluar las frecuencias relativas de la interacción entre personas del mismo género. En todo el corpus, las acciones mujer-mujer representan solo el 40,0% de todas las acciones del mismo género. La Figura 6 muestra la proporción de acciones mujer-mujer entre todas las interacciones del mismo género promediadas sobre libros y desglosadas por tiempo y género del autor. Hay un declive relativamente constante y acelerado desde 1880 en adelante con interacciones entre personajes masculinos que se vuelven cada vez más frecuentes en comparación con las interacciones mujer-mujer. Inicialmente, esta tendencia está impulsada por autores masculinos, en lugar de femeninos. En las décadas de 1950 y 1960, el predominio de los h

Las interacciones masculinas alcanzan su punto máximo e incluso se pueden encontrar en novelas escritas por mujeres. A partir de ese momento, se produce un fuerte cambio de tendencia en los libros escritos tanto por hombres como por mujeres.

No obstante, un patrón estable a lo largo de todo el período es que en los libros escritos por mujeres las interacciones mujer-mujer y hombre-hombre son igualmente frecuentes. Los hombres, por otro lado, son considerablemente más propensos a escribir sobre interacciones hombre-hombre. A lo largo de las décadas de 1950 y 1960, era de 4 a 5 veces más probable encontrar una interacción hombre-hombre en una novela escrita por un hombre que encontrar una interacción mujer-mujer.

Más allá de la mera frecuencia, la gramática también nos permite investigar la semántica de las relaciones de género. Primero, extraigo todas las acciones para cada una de las cuatro posibles combinaciones de género: masculino-femenino, masculino-masculino, femenino-masculino y femenino-femenino.5 Para encontrar las acciones más características de una combinación, luego calculo la probabilidad relativa de cada acción dentro de una combinación.
Los uso para derivar las probabilidades de una combinación condicional a cada acción, netas de las frecuencias relativas de la combinación. El Panel A de la Figura 7 muestra las acciones más indicativas de combinaciones mujer-hombre y hombre-mujer. Para tener una idea intuitiva de esto, considere la acción "resistir", que está asociada con una probabilidad de .62 para la combinación mujer-hombre. Esto implica que si tuviéramos que observar una instancia de resistencia en una novela con el mismo número de motivos de todas las combinaciones, hay un 62% de posibilidades de que el remitente sea un personaje femenino y el destinatario sea un personaje masculino, en lugar de cualquiera de los otros. tres combinaciones posibles.

Las acciones más indicativas de un personaje masculino actuando hacia uno femenino están relacionadas con el sexo (“penetrar”, “acostarse”, “entrar”), la violencia (“violar”, “aplastar”) y el cortejo (“cortejar”, “ cortejar”, “acariciar”). Los motivos más indicativos de un personaje femenino actuando frente a uno masculino, a su vez, tienen que ver mayoritariamente con la resistencia y el rechazo (“resistir”, “rechazar”, “rechazar”, “eludir”, “evadir”, “rechazar”, ” “divorciarse”, “desafiar”). En general, la relación de roles entre los personajes masculinos y femeninos parece estar definida por el contraste entre la persecución y la resistencia. Por supuesto, solo estamos viendo las acciones que son más características para la combinación de género respectiva, sin perjuicio de que existen modos de interacción predominantes entre personajes masculinos y femeninos que nada tienen que ver con este contraste. El panel B de la figura 7 muestra la probabilidad de que diferentes acciones sean dirigidas desde un personaje femenino hacia uno masculino, en lugar de viceversa. Por ejemplo, si hubo un caso de "amor" entre dos personajes de un género diferente, era igualmente probable que una mujer amara a un hombre (0,52) como que un hombre amara a una mujer. Sin embargo, los personajes femeninos son mucho más propensos a “perdonar” (.64) a los masculinos. Lo mismo se aplica a "mendigar" (.65), "odiar"

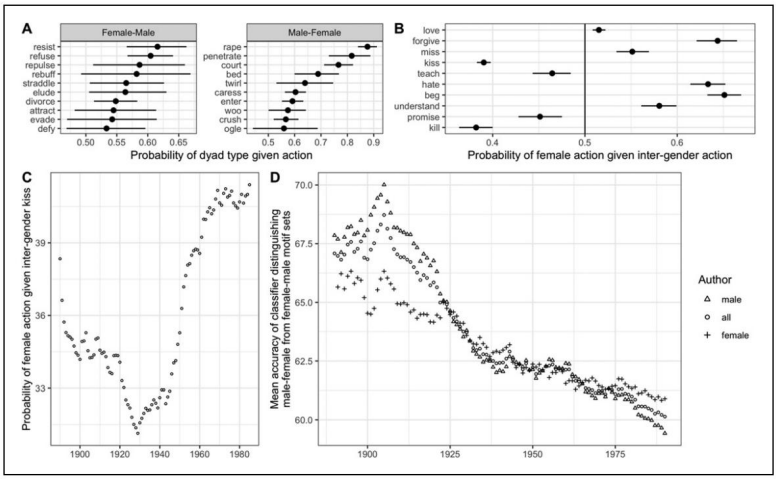


Figura 7. Contenido de la interacción entre géneros en las novelas estadounidenses.
Nota: El panel A muestra las acciones más características de las combinaciones mujer-hombre y hombre-mujer.
El panel B muestra la probabilidad de una combinación mujer-hombre (a diferencia de una combinación hombre-mujer) para un conjunto seleccionado de acciones. El panel C muestra esta estadística a lo largo del tiempo para la acción de besar. Los puntos representan ventanas de tiempo de 25 años.
El panel D muestra la precisión de un clasificador de Bayes ingenuo que distingue conjuntos de motivos masculino-femenino de femenino-masculino, como se analiza con más detalle en el texto.

(.63), “comprensión” (.58) y “falta” (.55). Por otro lado, es más probable que los personajes femeninos sean "enseñados" (0,46), "prometidos" (0,45) o "asesinados" (0,38) por personajes masculinos. También parece haber un desequilibrio de besos bastante considerable: el 61% de todos los besos entre géneros fueron recibidos por personajes femeninos. El panel C investiga este desequilibrio a lo largo del tiempo y revela una tendencia a besar la igualdad durante el siglo XX , aunque esta tendencia no es lineal.

¿Las relaciones de género en general se volvieron más simétricas? La asimetría se puede conceptualizar como el grado en que las acciones son indicativas de una combinación particular de género. Para operacionalizar esto, divido los datos en ventanas móviles de 20 años y considero solo novelas que contienen al menos diez motivos masculinos femeninos y diez masculino-femeninos de acción-paciente, lo que se aplica al 75,1% del corpus. Para cada ventana de tiempo, muestro 196 novelas (95% del número de novelas contenidas en la ventana con la menor cantidad de novelas). Para cada novela, muestro diez motivos de acción masculino-femenino y cinco femenino-masculino. Luego uso los conjuntos de motivos 392 para entrenar a un clasificador bayesiano ingenuo para distinguir conjuntos de motivos masculino-femenino de femenino-masculino. Este clasificador se utiliza para predecir el género.

combinación de todos los conjuntos de motivos retenidos. Repito este procedimiento 1000 veces para cada ventana de tiempo con el fin de obtener estimaciones sólidas de precisión, incluso para las ventanas de tiempo con un número comparativamente pequeño de novelas. El Panel D en la Figura 7 muestra la precisión de clasificación promedio sobre esas ejecuciones para cada ventana. La precisión de la clasificación puede interpretarse como una medida diacrónica de la asimetría en las relaciones de género. Tenga en cuenta que el rango de precisión es una consecuencia del límite artificial de solo 10 motivos por conjunto de motivos. El objetivo aquí no es maximizar el rendimiento de la predicción, sino garantizar la comparabilidad a lo largo del tiempo, de modo que la escala de la precisión informada sea, en cierto sentido, arbitraria.

En general, existe una tendencia inequívoca hacia la simetría. Es decir, la combinación de género de los conjuntos de acción se vuelve menos predecible con el tiempo, cayendo de 68,8 a principios del siglo XX a 60,0% en la última ventana de tiempo entre 1980 y 2000. Curiosamente, el género del autor parece tener poca importancia. diferencia con respecto a la asimetría. Con un sorprendente nivel de concurrencia, tanto los autores masculinos como femeninos siguen la tendencia histórica general de degenerar las interacciones sociales en sus novelas. Una limitación del análisis es que colapsamos los motivos de una combinación de género particular contenida dentro de una novela en lugar de distinguir los lazos entre personajes individuales, lo que podría desdibujar algunas de las dinámicas de género dentro de las novelas. Si bien abordar esto está más allá del alcance de este documento, bien podría valer la pena continuar en futuras investigaciones. La gramática semántica basada en motivos ofrece una nueva oportunidad para estudiar las representaciones de la interacción de género.

5.2 ¿Quiénes somos “nosotros”? La construcción de la identidad colectiva en las campañas presidenciales de EE. UU. (1952–2020)

Cuando los políticos hablan en un mitin de campaña, se enfrentan a un desafío: convencer a la audiencia de que hablan en su nombre. En el contexto de los movimientos sociales, esto se conoce como “problema de correspondencia de identidad” (Snow y McAdam 2000). Al igual que los activistas, los políticos deben intentar construir una identidad compartida en la que se fusionen el orador y la audiencia. Los puntos de referencia semánticos que se prestan a este tipo de trabajo de identidad son muy variados y pueden incluir intereses compartidos, valores, enemigos, historia, destino, pertenencia a una clase, nacionalidad, etnia o identidad regional.

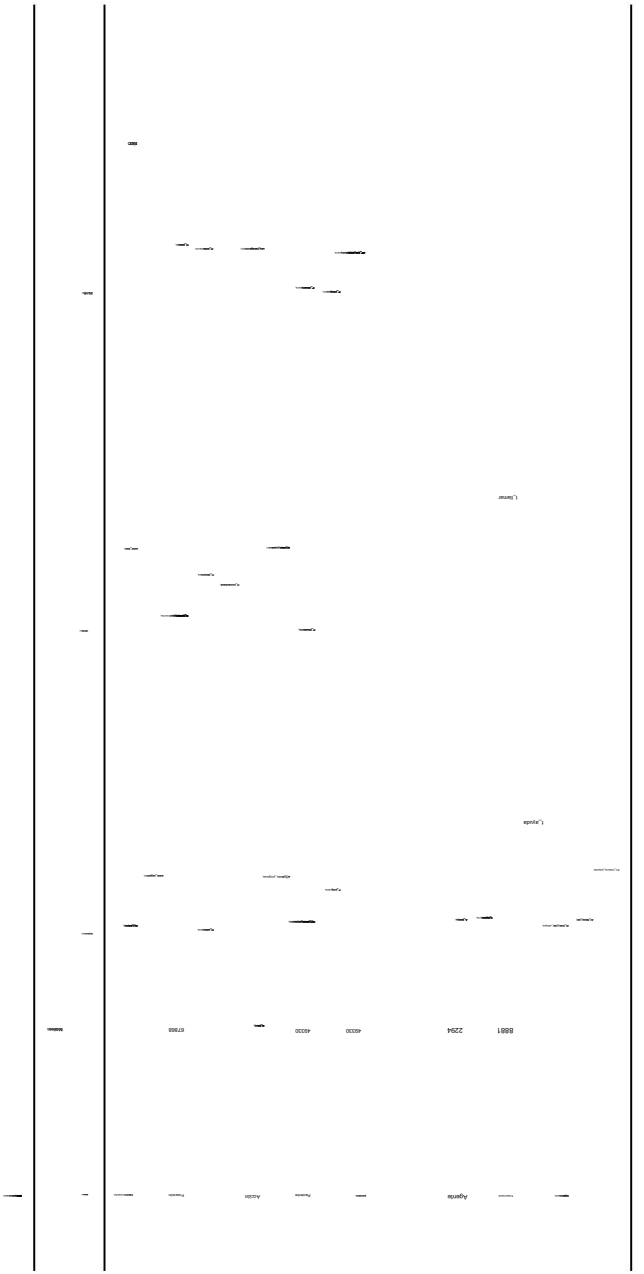
Aquí, me concentro en los intentos de los candidatos presidenciales de EE. UU. de alinear sus identidades con las de la audiencia en los mítines de campaña. Hago esto centrándome en lo que podría decirse que es el ancla léxica más importante en este tipo de identidad.

trabajo: el pronombre plural en primera persona “nosotros”, junto con sus formas objetivas y posesivas “nosotros” y “nuestro”. Los sociólogos políticos han notado durante mucho tiempo que los pronombres juegan un papel clave en la construcción y relación de las identidades sociales (von Wiese 1965; Elias 1978), pero no parecen haber prestado atención explícita a los usos de "nosotros" en el discurso político (aunque ver Wagner-Pacifici 2010:1360-1361). Los lingüistas y los analistas de conversación, sin embargo, han notado su carácter altamente político, ya que los usos de "nosotros" constituyen demarcaciones de fronteras sociales. Son actos de inclusión y exclusión simultáneas (Tyrkko 2016; Bramley 2001; Pennycook 1994; Wodak et al. 2009:45). En el lado inclusivo, el “nosotros” crea una asociación que generalmente involucra al orador y la audiencia, pero que bien puede extenderse más allá de eso. Por el lado exclusivo, los usos de "nosotros" en el ámbito político generalmente implican un "otro" y pueden usarse para enfatizar las dicotomías "nosotros"-"ellos".

Para estudiar las nociones de “nosotros” invocadas por los candidatos presidenciales de EE. UU., utilizo un corpus de texto de discursos de campaña y declaraciones públicas creado por Bonikowski y colegas (2021). Estos datos combinan diferentes fuentes para cubrir las elecciones de 1952 a 2020 y comprenden 2.956 discursos pronunciados por 34 candidatos. Los datos incluyen discursos pronunciados por los nominados de los dos partidos principales entre el 1 de septiembre y el día de las elecciones, así como sus discursos de aceptación de nominaciones.

Extraigo de estos textos todos los motivos en torno a "nosotros", "nuestro" y "nosotros" y los represento en forma lematizada. En total, había 331.422 motivos relacionados con nosotros. Las clases de motivos más frecuentes son acciones, posesiones y pacientes (ver Tabla 4). Los motivos de agente y tratamiento fueron considerablemente menos probables, lo que indica que los candidatos presidenciales tienden a construir un “nosotros” agente en lugar de uno pasivo. Obsérvese, además, que el número de motivos acción-paciente y agente-tratamiento es igual al de los motivos paciente y agente respectivamente, ya que cualquier instancia del primero implica una instancia del segundo (ver aP.2 en la Tabla 2).

La Tabla 4 enumera los 5 motivos más frecuentes para cada clase de motivo. Las principales caracterizaciones en su mayoría atribuyen fuerza y determinación colectivas (be_strong, be_able, be_proud, be_sure). Las posesiones más frecuentes se relacionan con “nuestro” país y su gente, así como con la economía. Los principales motivos de acción son semánticamente ambiguos, mientras que los principales motivos de pacientes también se refieren al país y su gente (P_personas, P_América, P_país), así como a trabajos (P_trabajo). Los motivos de acción-paciente tienden a ser proyectivos y relacionados con objetivos políticos, como la creación de puestos de trabajo (aP_create_job), el progreso (aP_make_progress) o simplemente ganar las elecciones (aP_win_elec
Los motivos más frecuentes de agente, tratamiento y agente-tratamiento se superponen parcialmente, lo que indica que no hay mucha variación en las entidades y



acciones hacia “nosotros”. Dios dándonos (At_God_give), la Biblia diciéndonos (At_Bible_tell), la historia enseñándonos (At_history_teach) y los ataques que nos cuestan (At_attack_cost) son frases populares entre los candidatos.

Para ilustrar cómo los candidatos se basan en diferentes retóricas en la construcción de una identidad colectiva, la Tabla 4 también enumera los 25 motivos más característicos6 relacionados con el “nosotros” para las dos campañas de 2016. Podemos ver, por ejemplo, que el “nosotros” de Hillary Clinton es bien (be_well) y bueno (be_good), mientras que el de Donald Trump está asociado con ser rico (be_rich). Mientras que Clinton parece afiliarse deliberadamente a la actual administración cuando habla de “nuestro” comandante en jefe (H_commander) o Primera Dama (H_Lady), Trump habla de “nuestros” políticos (H_politician) y de drenar (a_drain) el pantano (P_swamp) – construyendo así un “nosotros” que se opone a las élites. Oportunamente, encontramos que en el uso de Trump, el referente de “nosotros” con frecuencia parece ser su propia campaña o la campaña republicana (H_movimiento, P_Casa, aP_ganar_Casa, aP_ganar_estado).

Estos ejemplos conducen luego a una pregunta más general: ¿cuáles son las retóricas dominantes que usan los candidatos presidenciales para construir una identidad compartida de nosotros? Para abordar esta cuestión, construyo una serie de diccionarios de motivos. Los diccionarios de palabras están sujetos a un considerable escepticismo en las ciencias sociales (p. ej., Grimmer y Stewart 2013:274–275). La razón principal de esto es que muchas de las cualidades que uno quisiera medir en un texto no se reducen a la mera presencia o ausencia de palabras particulares. Los diccionarios construidos con motivos en lugar de palabras tienen el potencial de superar este escollo porque contienen información de cláusulas que es constitutiva de muchas cualidades textuales. Considere, por ejemplo, que observamos la palabra "gobierno" en la proximidad del término "nosotros". Esto en sí mismo nos dice relativamente poco sobre el encuadre de "nosotros", ya que la relación semántica entre los dos conceptos permanece sin especificar. Ahora considere, en cambio, que sabemos que la palabra "gobierno" corresponde a una instancia del motivo "A_gobierno". Esto nos dice que el “gobierno” es discutido como un agente que trata a “nosotros”, es decir, el hablante construye una identidad colectiva que se define frente al gobierno. En motivos compuestos como “At_government_tell” o “At_government_treat”, este encuadre se vuelve aún más explícito. Por supuesto, incluso la semántica de los motivos está sujeta a modificaciones por el contexto y hay complejidades textuales que están fuera de su alcance. Por ejemplo, las oraciones “El gobierno nos reprime” y “Mi oponente dice que el gobierno nos reprime” contendrían ambas el motivo “Al_gobierno_reprime”, pero solo en la primera es una afirmación, mientras que en la segunda es una afirmación. observación de segundo orden. Tales deficiencias no deberían ocultar, sin embargo, que los motivos generalmente nos permiten enriquecer los diccionarios con información de cláusulas. sirven como

concentrados de afirmaciones contenidas dentro de un texto, lo que hace que los diccionarios basados en motivos sean potentes medidas de las cualidades textuales.

Para identificar las retóricas dominantes en torno al "nosotros", inspecciono manualmente los 75 motivos más frecuentes de cada clase de motivo: un total de 600 motivos (ver documentación en el Apéndice D). Para evitar el sesgo en contra de las campañas con menos discursos, los selecciono en función de sus probabilidades promedio en todas las campañas. Si bien muchos de los motivos más frecuentes son semánticamente ambiguos (ver Tabla 4), otros parecen caer en registros particulares que se usan alrededor de "nosotros". Identifico cinco retóricas principales en torno al "nosotros" y les asigno motivos individuales, lo que lleva a cinco diccionarios con un promedio de 24 motivos.

La Tabla 5 muestra los motivos asociados a cada retórica. Las asignaciones fueron informadas por la familiaridad de los autores con los discursos a través de proyectos previos, así como la inspección de las instancias en las que ocurrieron motivos específicos.
La prevalencia de estas retóricas se midió luego como el porcentaje de todos los motivos relacionados con "nosotros" pertenecientes al diccionario respectivo. Estos se muestran en la Figura 8.

Una retórica empleada con frecuencia por los candidatos presidenciales postula "nosotros" en oposición al gobierno o la administración actual. A menudo, esto se hace describiendo cómo "nosotros" somos tratados por el gobierno como un agente (A_presidente, A_administración, A_gobierno, A_presidente). La frase más común es que el gobierno nos "dice" cosas, como se ve en el respectivo (At_President_tell, At_government_tell, agent-treatment At_administration_tell). Además, una serie de administración motivos motivos de acción-paciente enfatizan los requisitos para una (aP_necesita_gobierno, aP_necesita_presidente, aP_quiere_presidente), lo que suele ser una forma de enfatizar la incompetencia de la administración actual. Otra forma popular de expresar esto es que no podemos permitirnos más años de la administración actual (aP_afford_year), o que estamos cansados (be_cansado) de las circunstancias actuales.

La retórica de "nosotros" contra la administración es empleada por ambas partes por igual. El uso depende principalmente de quién reta a un presidente o partido en ejercicio (correlación de .60, $p < .001$). Las tres campañas que más se basaron en esa retórica fueron las de Mondale en 1984, Dukakis en 1988 y Obama en 2008, todas las cuales compitieron contra los republicanos después de años de mandato republicano.

También se puede construir un "nosotros" enfatizando una identidad política de seguridad común y evocando una amenaza externa. "Nosotros" somos atacados (t_attack), heridos (t_hurt) o golpeados (t_hit) por "nuestros" enemigos (H_enemy, A_enemy) o adversarios (A_adversary) y manejamos los conflictos (P_peace, P_war) por medio de "nuestras" fuerzas militares (H_tropa, H_militar, H_fuerza).

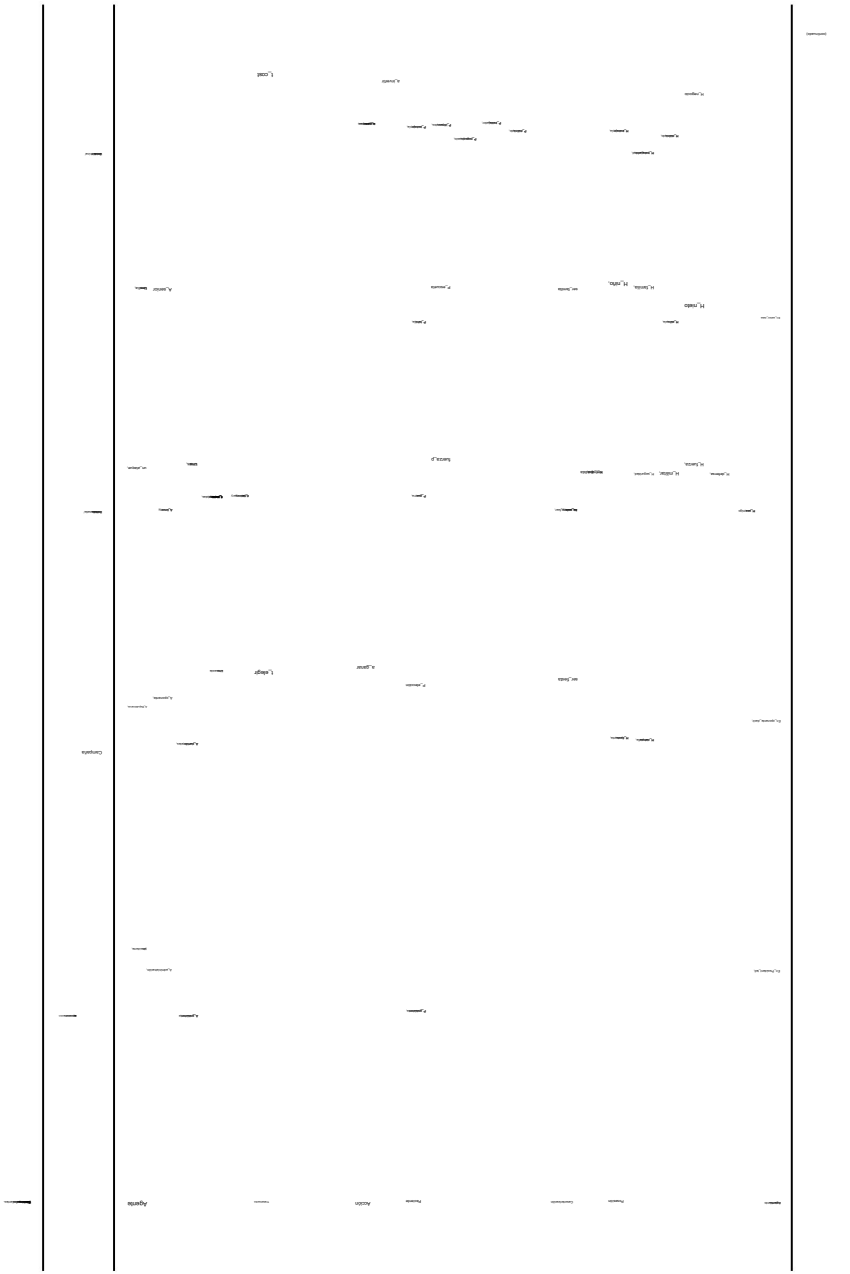
Usar "nosotros" al discutir tales asuntos, en lugar de, digamos, "los Estados Unidos

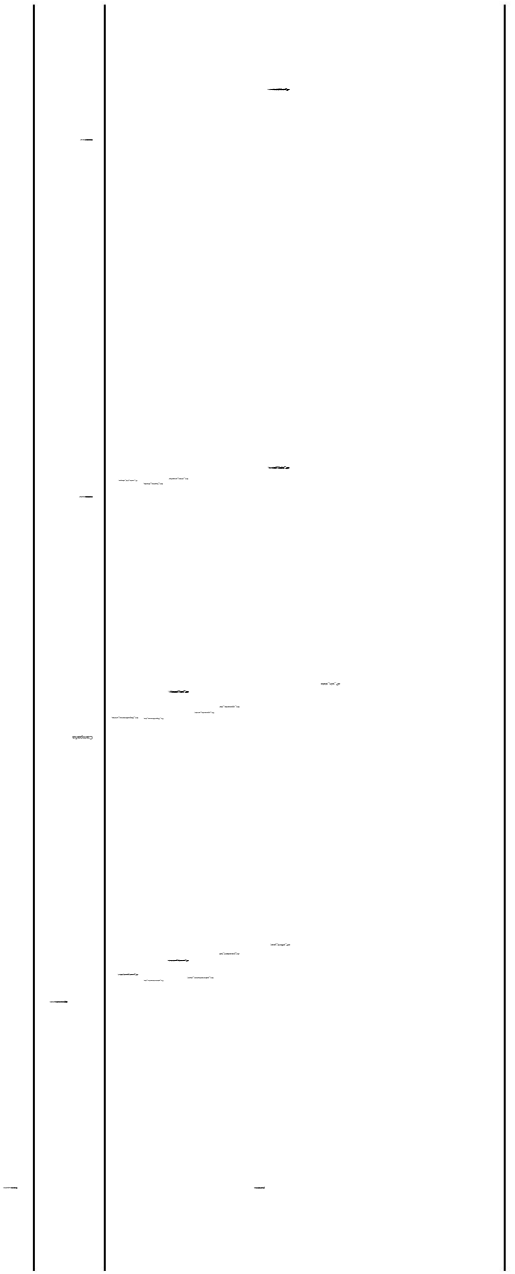
Unidos", fusiona orador y audiencia en una comunidad nacional de destino.
En general, es más probable que los candidatos republicanos empleen tales motivos (correlación de .39, $p < .05$). Sin embargo, el uso de la retórica militar/de amenazas también depende de las circunstancias históricas de una campaña. Con mucho, la prevalencia más alta de esta retórica ocurrió en 2004, la primera elección presidencial después de los ataques del 11 de septiembre y la invasión de Irak.

La retórica más frecuente en torno al "nosotros" se refiere a los Estados Unidos como unidad económica y fiscal. Los candidatos hablan de "nuestra" economía, trabajos y negocios (H_economía, H_trabajo, H_negocio), así como también cómo "nosotros" podemos afectarla (p. ej., aP_crear_trabajo, aP_reducir_impuestos, aP_crecer_economía). El uso de esta retórica se caracteriza por una tendencia histórica interesante: hablar de "nuestra" economía no fue particularmente frecuente hasta principios de los años 80, cuando dicha retórica comenzó a ser cada vez más popular entre los candidatos. Alcanzó su punto máximo en las campañas de 2008 durante la crisis financiera, representando entre el 8 y el 9 % de todos los motivos relacionados con el "nosotros". Desde entonces y con la recuperación de la economía estadounidense, su prevalencia se ha estabilizado nuevamente.

También es común que los candidatos se dirijan a "nosotros" como miembros de la familia. La afirmación aquí no es que el orador y la audiencia compartan una familia. Más bien, hablar de "nuestros" hijos y miembros de la familia (H_niño, H_familia, H_niño, H_nieto) es una forma de enfatizar los roles compartidos de las personas como padres y miembros de familias en general. Con solo dos excepciones recientes (Bush 2004 y Romney 2012), esta retórica tiende a ser más popular entre los candidatos demócratas. Fue, con mucho, el más frecuente en la campaña de 2000 de Gore. Investigaciones posteriores revelan que Gore hizo uso de esta retórica porque al identificar "nosotros" con la familia, pudo hablar sobre las consecuencias aparentemente distantes del cambio climático.

Finalmente, otro uso de "nosotros" es equipararlo con la campaña misma.
Los candidatos hablan sobre "nuestra" campaña, partido u oponente (H_oponente, H_partido, H_campaña), analizan cómo nos trata la otra parte a "nosotros" (A_Republicanos, A_Demócratas, A_oponente, A_oponente_decir), o sobre cómo nos va a "nosotros" en las elecciones (a_gana, aP_gana_elección, aP_toma_el_puesto). Identificar la identidad colectiva con la propia campaña parece ser algo idiosincrásico. No está significativamente asociado con un partido, ni existe una dinámica de retador-titular o una tendencia histórica. En los últimos años, esta forma de retórica del "nosotros" ha sido más frecuente en las campañas de Trump de 2016 y 2020, así como en las campañas de Obama de 2008 y 2012. Los precedentes históricos son la campaña de 1960 de Nixon contra Kennedy y la campaña de 1984 de Mondale contra Nixon. Dicho esto, una mayor investigación de los datos apunta a una dinámica interesante: el grado en que los candidatos identifican "nosotros" con su propia campaña parece cambiar a lo largo de la campaña.





A medida que se acerca el día de las elecciones, la identidad de campaña parece desempeñar un papel cada vez más importante en la retórica de los candidatos. El panel inferior derecho de la Figura 8 muestra la correlación entre la prevalencia de la retórica de "campaña" y el número de días hasta la elección. Hay una asociación negativa para 8 de las 10 campañas desde 2000, aunque no todas estas correlaciones son estadísticamente significativas. Este hallazgo se alinea con la investigación sobre campañas negativas que muestran que los candidatos aumentan su énfasis en los límites de la campaña a medida que se acercan las elecciones (Damore 2002; De Nooy y Kleinnijenhuis 2013). Al comienzo de una campaña, los candidatos tienden a enfatizar el contenido del tema para informar al electorado quiénes son y qué es importante para ellos. A medida que las elecciones se vuelven inminentes y los candidatos han construido con éxito sus perfiles, "nosotros" se vuelve más a menudo un marcador para una campaña de candidatos en yuxtaposición explícita a la del oponente.

En general, este análisis ilustra el potencial que reside en basar los diccionarios en motivos, en lugar de la mera aparición de palabras. Al integrar la información de las cláusulas en nuestra herramienta de medición, las cualidades textuales relativamente complejas se vuelven medibles. Aquí pudimos medir diferentes retóricas al servicio de la construcción de la identidad colectiva, un fenómeno que probablemente eluda muchas otras estrategias de medición. Los análisis también ilustran otra ventaja central del marco importante para los sociólogos: mientras que muchos métodos contemporáneos de aprendizaje automático tienden a operar como cajas negras de facto, una representación a través de motivos semánticos permanece cerca del texto original. Proporciona transparencia que abre la medición para la crítica y el escrutinio.

6. Validación y discusión de las limitaciones

Para evaluar la calidad de la extracción de motivos, extraigo dos muestras de los datos utilizados en los análisis: la muestra 1 consta de 250 oraciones del USNC que contienen entidades de género, estratificadas por década; La muestra 2 consta de 250 oraciones del corpus de discursos de campaña que contienen "nosotros", "nosotros" o "nuestro", estratificados por campaña. Para cada oración, anoto y enumero manualmente todos los motivos alrededor de las entidades centrales. Estas listas sirven como patrón oro y se comparan con los motivos que se extrajeron computacionalmente. Para que este proceso sea lo más transparente posible, los dos conjuntos de datos anotados se proporcionan en el material complementario, junto con comentarios y elaboraciones sobre anotaciones individuales.

Las tablas 6 y 7 muestran los resultados para cada clase de motivo. El valor de recuperación representa la porción de motivos verdaderos que se extrajeron computacionalmente. Por ejemplo, de los 778 motivos que se anotaron en el discurso de campaña

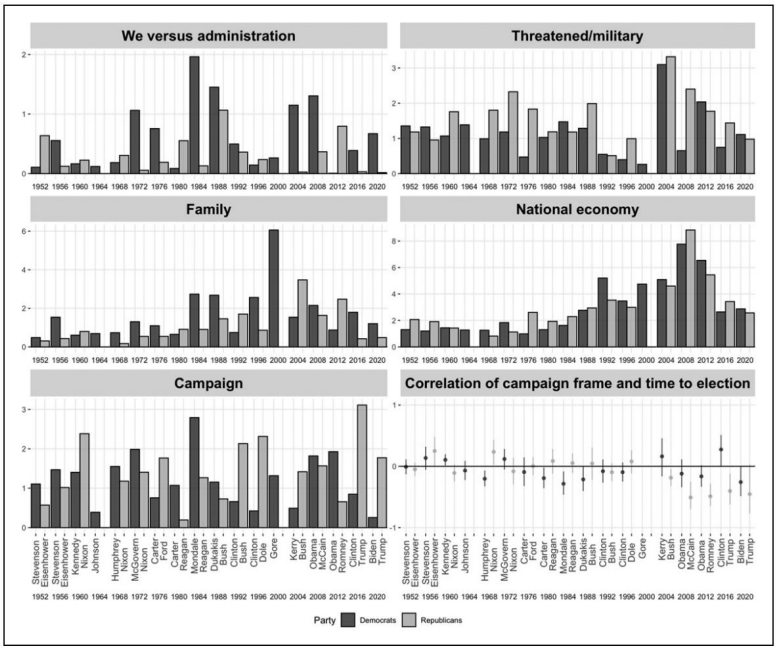


Figura 8. Prevalencia de la retórica del “nosotros” en las campañas presidenciales.
Nota: Los paneles de gráficos de barras muestran la prevalencia de diferentes retóricas como porcentajes de todos los motivos relacionados con "nosotros" que pertenecen al diccionario respectivo. Por ejemplo, el 2,57 % de todos los motivos relacionados con el “nosotros” empleados en la campaña Trump 2020 pertenecen al marco de la economía nacional. El panel inferior derecho muestra la correlación entre la prevalencia de la retórica de "campaña" y el número logaritmizado de días hasta la elección con intervalos de confianza del 90%. La campaña Gore de 2000 se eliminó de este panel porque solo hay 4 discursos en los datos.

muestra, se extrajeron con éxito 652, lo que llevó a un recuerdo general de 0,84. Para la muestra de USNC, el recuerdo alcanza .80. Nuestra capacidad para extraer motivos difiere según la clase de motivo. Mientras que casi todos los motivos de acción (0,90 y 0,82), tratamiento (0,95 y 0,82) y posesión (0,96 y 0,99) se pudieron extraer con éxito, el rendimiento con respecto a los pacientes (0,76 y 0,72), agentes (.47 y .71), y las dos clases de motivos compuestos es considerablemente menor. Esto probablemente se deba al hecho de que estos motivos suelen estar más alejados de las entidades de interés en los gráficos de dependencia. De hecho, la identificación de agentes y pacientes depende de la correcta identificación de acciones y tratamientos, haciendo de los primeros una tarea más compleja. Finalmente, los motivos de caracterización muestran una

Usado	Verdad	Precisión	Recordar		Clasificación
	293	0.96	0.82	Pacientes	
104	130	0.89	0.72		
104	130	0.89	0.72	Tratamientos Acción	
41	40	0.80	0.82		
25	31	0.88	0.71		
25	31	0.88	0.71	Situaciones	
34	51	0.97	0,65		
140	136	0.96	0.99		
723	842	0.93	0.80		

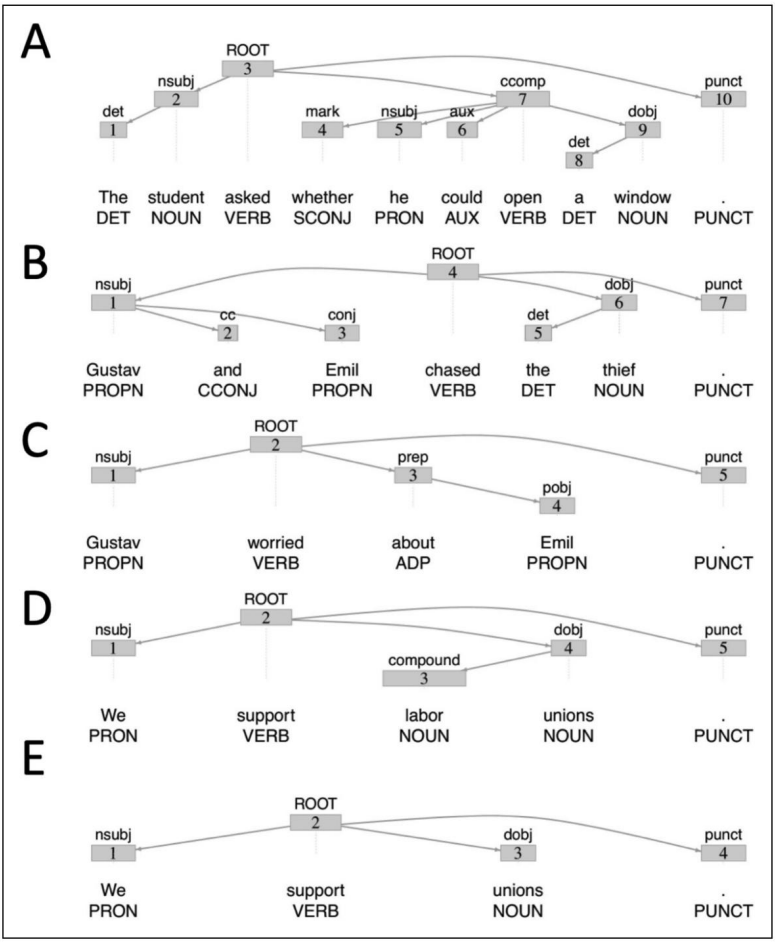


Figura 9. Oraciones ejemplares que ilustran las complejidades en la extracción de motivos a través de reglas sintácticas.

patrón algo distinguido con un recuerdo de .65 en la muestra de USNC y .92 en la muestra de Campaign Speeches. En general, los valores de recuerdo indican que hay margen de mejora en la extracción de motivos y que una cantidad no despreciable de información relevante contenida en los textos se pierde en este proceso.

Sin embargo, una medida quizás más importante con respecto a la validez de los análisis es la precisión, es decir, la porción de motivos extraídos que son verdaderos motivos. La precisión es .90 para el corpus de Campaign Speeches y .93 para el USNC. Nuevamente encontramos diferencias similares entre las clases de motivos, que son, sin embargo, menos pronunciadas que en el recuerdo. En casi todas las clases de motivos, encontramos que alrededor de 9 de cada 10 motivos extraídos son correctos. La única excepción a esto son los motivos de caracterización en la muestra del discurso de la campaña, lo que indica que la calidad de la extracción puede diferir según el estilo del lenguaje.7

Estos números apuntan a un patrón general: nuestra capacidad para extraer motivos deja cierto margen de mejora, especialmente con respecto a las clases de motivos dependientes agente y paciente. Sin embargo, los motivos que se extrajeron son abrumadoramente correctos. Para investigar más a fondo los posibles sesgos en la calidad de la extracción de motivos, realicé una serie de análisis de regresión multinivel sobre covariables a nivel de oración y documento que se proporcionan en el material complementario. Dado que la canalización del lenguaje spaCy utilizada aquí se entrenó en gran medida con datos textuales de fuentes contemporáneas, parecía especialmente importante evaluar si el rendimiento de la extracción cambia con el tiempo. Encuentro alguna evidencia que sugiere que un mayor número de extracciones en una oración aumenta la probabilidad de extracciones erróneas (precisión).
Sin embargo, ni el año de la campaña presidencial ni el año de publicación de una novela son predictores significativos del desempeño de la extracción.
Lo mismo se aplica a la longitud de la oración y al número de motivos verdaderos (es decir, motivos anotados humanos) dentro de una oración. Juntos, estos resultados sugieren que el enfoque propuesto aquí es capaz de extraer motivos válidos en una variedad de aplicaciones. Además, si la heterogeneidad en el rendimiento de la extracción es motivo de preocupación o no, depende de los objetivos del análisis posterior. Si bien parece probable que el rendimiento de la extracción difiera entre tareas y textos muy diferentes, ninguno de los esfuerzos de validación sugiere una vulnerabilidad particular a tales variaciones.

Dicho esto, es importante reconocer que hay información semántica relevante dentro de los datos textuales que simplemente se encuentra más allá de lo que el enfoque propuesto aquí puede extraer en su forma actual. Esto se debe a una serie de limitaciones. El primero, y probablemente el más importante de ellos, se relaciona con el fenómeno de la correferencia; a menudo, varias palabras o frases representan la misma entidad. Para ilustrar esto, considere la oración A en la Figura 9. Una primera lectura de la oración podría sugerir que "estudiante" y "él" se refieren a la misma persona. Suponiendo que estuviéramos interesados en extraer motivos alrededor de entidades de género, idealmente querríamos extraer "a_ask" como motivo de acción, dado que el "estudiante" y "él" son la misma persona. Sin embargo, la entidad actuante asociada con "pedido" no es "él" sino "estudiante", que en sí mismo no es

una entidad de género, por lo que nuestro enfoque no habría podido extraer "a_ask". La correferencia es frecuente en el lenguaje natural y, a menudo, se extiende por los límites de la oración. No hay duda de que, en muchas aplicaciones, nos hace perder información relevante. La oración A, sin embargo, también se presta para ilustrar la complejidad de la correferencia: ¿no es igualmente plausible que la oración se trate de un estudiante de género desconocido que pregunta, por ejemplo, a un maestro si "él" podría abrir una ventana? Sencillamente, no hay forma de saberlo con certeza, especialmente sin basarse en información contextual. La resolución de correferencias, la tarea de identificar todas las palabras y frases que se refieren a la misma entidad, es difícil porque implica interpretaciones complejas. Es un área de investigación activa en el procesamiento del lenguaje natural. Las versiones futuras del enfoque propuesto aquí pueden integrar un componente de resolución de correferencia. Sin embargo, esto inevitablemente tendrá el costo de una precisión reducida, lo que puede no ser deseable para muchas aplicaciones de las ciencias sociales.

En segundo lugar, la gramática propuesta aquí relaciona las entidades centrales con los pacientes sobre los que actúa directamente y con los agentes que actúan directamente sobre ella. Lo hace si, y solo si, los pacientes y los agentes están vinculados a la entidad central a través de un verbo transitivo. Hay, sin embargo, muchas formas en que los textos informan sobre la relación entre entidades que no siguen esta forma. En la oración B de la Figura 9, por ejemplo, Gustav y Emil participan en una acción común, algo que el enfoque propuesto aquí actualmente no captaría. De manera similar, la Oración C nos informa que Gustav estaba preocupado por Emil. Sin embargo, los casos en los que la relación entre entidades está mediada por una preposición ("sobre"), actualmente no se consideran para la gramática. Hacerlo requeriría ampliar la noción de motivos para incluir frases de varias palabras (p. ej., una_preocupación_por) o definir nuevas clases de motivos. El aumento de la complejidad de la gramática de esta manera puede ser el foco del trabajo futuro. Sin embargo, siempre habrá información sobre las relaciones entre entidades que un enfoque basado en la sintaxis no puede capturar.

Esto nos lleva a una tercera limitación: la unidad básica de la gramática es el token. En algunos casos, sin embargo, esto podría llevar al descarte de información relevante. Por ejemplo, si tuviéramos que extraer motivos asociados con "nosotros" en la oración D, extraeríamos "P_union" y "aP_support_union", aunque puede ser relevante qué tipos de uniones se admiten.

Por lo tanto, podría parecer atractivo usar frases como “sindicatos” como motivos, en lugar de símbolos únicos. La razón pragmática para no usar frases de varias palabras como motivos es que esto probablemente introduciría escasez en los datos extraídos que no es útil para la mayoría de los análisis posteriores. Esto se debe a que las frases correspondientes al mismo referente a menudo toman formas ligeramente diferentes, como se ilustra en la oración E. Aquí, el apoyo se dirige a

"sindicatos" solos. El uso de frases, en lugar de tokens, crearía motivos semánticamente congruentes pero léxicamente distintos y, por lo tanto, aumentaría significativamente el tamaño del vocabulario de motivos. Claramente existe una compensación aquí entre la reducción de la escasez y la retención de información potencialmente relevante. Al usar fichas como unidades básicas de motivos, perseguimos los primeros a costa de sacrificar algunos de los segundos.

Finalmente, el enfoque presentado aquí no captura la negación. Aparte del hecho de que integrar la negación complicaría aún más la gramática, también resulta difícil de implementar. La dificultad surge aquí principalmente del hecho de que los patrones sintácticos no siempre parecen proporcionar suficiente información para decidir si una negación se aplica transitivamente a una cadena de verbos o adjetivos. Si bien el desprecio por la negación es la norma entre los enfoques de análisis de texto que se utilizan actualmente en sociología (ver Sección 2), puede valer la pena implementar una aproximación ingenua a la gramática en trabajos futuros. De hecho, ninguna de las diversas limitaciones discutidas aquí debe verse como irresoluble. En cambio, son los sitios de construcción donde los esfuerzos futuros pueden despegar y refinar el marco básico presentado aquí.

7. Conclusión

Los sociólogos han analizado formalmente las relaciones semánticas en los datos textuales desde finales de la década de 1980. Si bien este programa de investigación recientemente ganó atención y encontró su camino hacia la corriente principal de la disciplina, también ha sido dominado por estudios que se enfocan en las co-ocurrencias o en las distancias semánticas entre conceptos. El marco descrito aquí proporciona una alternativa a estos enfoques que tiene al menos tres ventajas.

Primero, el marco hace que la información de las cláusulas sea accesible para la medición. Muchas de las cosas que preocupan a los sociólogos en un texto no pueden reducirse a la presencia o ausencia de palabras particulares. Al integrar la información de las cláusulas en nuestras herramientas de medición, aumentamos nuestra capacidad para capturar cualidades textuales relativamente complejas. Por supuesto, como se ha reconocido, existe un límite a la complejidad argumentativa que pueden representar los motivos. Es por eso que este avance debe describirse de manera realista como gradual: con una gramática semántica, podemos ir más allá de lo que trata un texto y hacia lo que pretende; más allá de la medición de temas y tópicos y hacia la medición de argumentos.

En segundo lugar, la relativa proximidad de su forma de representación al texto original promueve interpretaciones intersubjetivamente válidas. Ningún enfoque formal nos libera de tener que dar sentido a los patrones que descubre. Pero las formas de representación actualmente dominantes tienden a abstraer los eventos y la narrativa.

acción que tiene lugar dentro de los textos, dejando así mucho terreno por recorrer por medio de la interpretación. ¿Qué significa si encontramos que los conceptos coexisten o tienen incrustaciones de palabras similares? Con demasiada frecuencia, parece haber múltiples respuestas plausibles y los criterios para seleccionar entre ellas no son obvios. Si, por otro lado, encontramos que los personajes masculinos de las novelas besan a los personajes femeninos de las novelas con más frecuencia que viceversa, este hallazgo puede sostenerse por sí solo y requiere comparativamente poca interpretación. Para ser claros, esto no quiere decir que dicho resultado no se base en decisiones ex ante, ni que no necesite interpretación, evaluación crítica y preguntas de seguimiento. Sin embargo, sostengo que la brecha entre el patrón formal y la intuición sociológica articulable e intersubjetivamente válida es considerablemente más pequeña.

Esto conduce entonces a lo que probablemente sea la ventaja más importante del enfoque propuesto: su proximidad a la ontología de la teoría sociológica. En este marco, los conceptos que usamos convencionalmente en nuestras descripciones de las estructuras sociales (actores, interacciones, relaciones, roles y atributos, por nombrar algunos) se vuelven medibles en el texto. Con este fin, la sección 5.1 analizó las interacciones entre personajes de género; la sección 5.2 analizó la construcción del "nosotros" como actor en la retórica política; y en otros lugares, recientemente se demostró que se puede usar un enfoque similar para medir roles en datos textuales (Stuhler 2021). No menos importante, este documento es, por lo tanto, un intento de proporcionar herramientas y contornos para lo que podría convertirse en un nuevo programa de investigación. Este programa va más allá del estudio de las asociaciones textuales abstractas y avanza hacia un análisis formal de las representaciones textuales de las estructu

Expresiones de gratitud

Agradezco a Delia Baldassarri, Bart Bonikowski, Adam Braffman, Paul DiMaggio, Jan Fuhse y Patrick Kaminski, así como a los miembros del taller de la Red de Cultura y Acción de la Universidad de Chicago por sus sugerencias y comentarios sobre borradores anteriores. Un agradecimiento especial a Hoyt Long por permitirme el acceso a los EE. UU.
Corpus de novela.

Nota del autor

Se puede acceder a un repositorio que proporciona materiales de replicación y archivos complementarios en el enlace impreso a continuación. En el repositorio también se pueden encontrar descripciones detalladas de los archivos e instrucciones para replicar los análisis. https://osf.io/rq6j7/?view_only=52bccf58ea004439853dff01f464ac50 _

Declaración de Conflicto de Intereses

El autor no declaró ningún conflicto de interés potencial con respecto a la investigación, autoría y/o publicación de este artículo.

Fondos

El autor no recibió apoyo financiero para la investigación, autoría y/o publicación de este artículo.

identificación ORCID

Oscar Stühler https://orcid.org/0000-0001-7391-1743

Material complementario El

material complementario de este artículo está disponible en línea.

notas

1. Si bien esto puede parecer una elección extraña, "paciente" es el término común que se usa en gramáticas semánticas para denotar entidades sobre las que se actúa.
2. Tenga en cuenta que spaCy ofrece varias canalizaciones de procesamiento que proporcionan diferentes compensaciones entre el rendimiento y el costo computacional. Aquí, utilizo el modelo más grande para un rendimiento óptimo. Para obtener información específica sobre la canalización, consulte https://github.com/explosion/spacy-models/releases/tag/
en_core_web_lg-3.1.0 3. Si bien hago una clara distinción aquí entre gramáticas sintácticas y semánticas, vale la pena reconocer que el límite preciso entre estas dos categorías no siempre es obvio y que las gramáticas de dependencia modernas, aunque podría decirse que se centran en la sintaxis, tienen en cuenta algunos criterios semánticos (ver la discusión en de Marneffe y Nivre 2019: 203). Si bien la distinción puede ser, de hecho, más gradual que categórica, es importante resaltar la brecha que existe entre las gramáticas de dependencia detalladas y centradas principalmente en la sintaxis y el tipo de gramática semántica que se propone en este artículo.

4. La versión actual de semgram, así como una demostración, se pueden encontrar en https://github.com/omstuhler/semgram 5.
Los motivos se analizaron en forma lematizada, lo que significa que las características léxicas se redujeron a su forma básica. Esto suaviza la presentación de los resultados pero también ayuda a que los datos sean menos dispersos. Sin embargo, uno bien puede optar por estudiar motivos no lematizados dado que estos pueden contener información relevante para objetivos analíticos distintos a los míos, como, por ejemplo, el tiempo.
6. Los motivos se seleccionaron de acuerdo con una puntuación de exclusividad de frecuencia. La puntuación es la media de un motivo (mi), clasificado por probabilidad bajo un candidato particular (cj), que

es P(mi|cj) y su rango por exclusividad bajo un candidato dado, donde la exclusividad se define como P(cj|mi) después de normalizar cantidades desiguales de datos por candidato.

7. Una inspección más detallada de este resultado en particular no condujo a una explicación clara. Sin embargo, hay un número inusualmente alto de extracciones en la muestra de Discurso de campaña en comparación con la anotación estándar de oro (20% más extracciones que motivos verdaderos). La inspección de los casos de falsos positivos sugiere que esto es una consecuencia de la forma en que los candidatos hablan sobre lo que "somos": a menudo, las oraciones comienzan con "nosotros" y un verbo en cópula (p. ej., "somos"), seguido de múltiples oraciones subordinadas. Los falsos positivos suelen ubicarse en estas oraciones subordinadas.

Referencias

Abel, Pedro. 1987. La sintaxis de la vida social: la teoría y el método de comparación Narrativas. Oxford: Prensa de la Universidad de Oxford.

Abel, Pedro. 2004. "Explicación narrativa: ¿una alternativa a la explicación centrada en variables?" Revisión Anual de Sociología 30(1):287310. doi:10.1146/añorev. soc.29.010202.100113

Arseniev-Koehler, Alina. 2021. "Fundamentos teóricos y límites de las incrustaciones de palabras: ¿qué tipos de significado pueden capturar?. https://arxiv.org/abs/ 2107.10413

Arseniev-Koehler, Alina y Jacob Foster. 2020. "Aprendizaje automático como modelo para el aprendizaje cultural: enseñar a un algoritmo lo que significa ser gordo". https://arxiv. org/abs/2003.12133.

BBC. 2018. "100 mujeres: cómo Hollywood falla a las mujeres en la pantalla". https://www.bbc. com/noticias/mundo-43197774.

Bearman, Peter, Robert Faris y James Moody. 1999. "Bloqueando el futuro: nuevas soluciones para viejos problemas en las ciencias sociales históricas". Historia de las Ciencias Sociales 23(4):50133. doi:10.1017/S0145553200021854

Bearman, Peter y Katherine Stovel. 2000. "Convertirse en nazi: un modelo para redes narrativas". Poética 27(2-3):6090.

Benoit, Kenneth y Akitaka Matsuo. 2020. "spacyr: un envoltorio R para la biblioteca Python spaCy NLP". https://cran.r-project.org/web/packages/spacyr/index.html.

Bonikowski, Bart, Yuchen Luo y Oscar Stuhler. 2021. "¿Política de siempre? Antecedentes de los marcos de la derecha radical en el discurso electoral de los Estados Unidos". https://doi. org/10.31235/osf.io/uhvbp.

Boutyline, Andrei, Alina Arseniev-Koehler y Devin J Cornell. 2020. "Escuela, estudio e inteligencia: estereotipos de género y educación a lo largo de 80 años de medios impresos estadounidenses, 1930-2009". https://doi.org/10.31235/osf.io/bukdg.

Bramley, Nicolette Ruth. 2001. "Pronombres de política: el uso de pronombres en la construcción de 'yo' y 'otro' en entrevistas políticas". Serie Pronombres de política:

el uso de pronombres en la construcción de 'yo' y 'otro' en entrevistas políticas., Edición. Universidad Nacional de Australia: Editor.

Carley, Kathleen. 1988. “Formalizando el Conocimiento del Experto Social”. Métodos sociológicos e investigación 17(2):165232. doi:10.1177/0049124188017002003 Carley, Kathleen.

1993. "Opciones de codificación para análisis textual: una comparación de análisis de contenido y análisis de mapas". Metodología Sociológica 23:75126. doi:10.2307/271007

Carley, Kathleen. 1994. "Extracción de la cultura a través del análisis textual". Poética 22(4):291 312. doi:10.1016/0304-422X(94)90011-6 Carley, Kathleen

y Michael Palmquist. 1992. "Extracción, representación y análisis de modelos mentales". Fuerzas Sociales 70(3):60136. doi:10.2307/2579746 Chatman, Seymour B. 1978. Historia y

discurso. Estructura narrativa en ficción y cine. Ítaca: Prensa de la Universidad de Cornell.

Chicago-Text-Lab. 2021. "Corpus de novelas estadounidenses". https://textual-optics-lab.uchicago. edu/us_novela_corpus.

Choi, Jinho D. y Martha Palmer. 2012. Pautas para el componente de estilo claro para la conversión de dependencia. https://www.mathcs.emory.edu/~choi/doc/cu-2012-choi.pdf

Damore, David F. 2002. "Estrategia del candidato y la decisión de volverse negativo". Investigación Política Trimestral 55(3):66985. doi:10.1177/106591290205500309 Danowski,

James A. 1993. "Análisis de red del contenido del mensaje". Avances en Ciencias de la Comunicación 12:197221.

de Marneffe, Marie-Catherine y Christopher D. Manning. 2008. “Manual de dependencias mecanografiadas en Stanford”. https://downloads.cs.stanford.edu/nlp/software/ dependencies_manual.pdf .

de Marneffe, Marie-Catherine y Joakim Nivre. 2019. "Gramática de la dependencia". Revisión anual de lingüística 5 (1): 197218. doi:10.1146/annurev-lingüística 011718-011842

De Nooy, Wouter y Jan Kleinnijenhuis. 2013. “Polarización en los medios durante una campaña electoral: un modelo de red dinámica que predice el apoyo y el ataque entre los actores políticos”. Comunicación Política 30(1):11738. doi:10.1080/ 10584609.2012.737417

Deerwester, Scott, Susan T. Dumais, George W. Furnas y Thomas K. Landauer. 1990. "Indización por análisis semántico latente". Revista de la Sociedad Estadounidense de Ciencias de la Información 41(6):391407.

DiMaggio, Paul, Manish Nag y David Blei. 2013. "Explotación de afinidades entre el modelado de temas y la perspectiva sociológica de la cultura: aplicación a la cobertura periodística de la financiación de las artes del gobierno de EE. UU.". Poética 41(6):570 606. doi:10.1016/ j.poetic.2013.08.004

Duquenne, Vincent, John W. Mohr y Annick Le Pape. 1998. "Comparación de órdenes duales en el tiempo". Información de Ciencias Sociales 37(2):22753. doi:10.1177/ 053901898037002001

Eisenstein, Jacob. 2019. Introducción al procesamiento del lenguaje natural. Cambridge: Prensa del MIT.

Elías, Norberto. 1978. Qué es la Sociología. Nueva York: Prensa de la Universidad de Columbia.

Fligstein, Neil, Jonah Stuart Brundage y Michael Schultz. 2017. "Ver como la Reserva Federal: cultura, cognición y encuadre en el fracaso para anticipar la crisis financiera de 2008". Revisión Sociológica Americana 82(5):879909. doi:10.1177/ 0003122417728240

Franzosi, Roberto. 1989. "De las palabras a los números: un procedimiento de codificación generalizado y basado en la lingüística para recopilar datos textuales". Metodología Sociológica 19(1):26398. doi:10.2307/270955

Franzosi, Roberto. 1990. "Codificación de datos textuales asistida por computadora". Métodos Sociológicos e Investigación 19(2):22557. doi:10.1177/0049124190019002004

Franzosi, Roberto. 1994. "De las palabras a los números: un marco de teoría de conjuntos para la recopilación, organización y análisis de datos narrativos". Metodología Sociológica 24:10536. doi:10.2307/270980

Franzosi, Roberto. 1998a. "Análisis narrativo: o por qué (y cómo) los sociólogos deberían estar interesados en la narrativa". Revisión Anual de Sociología 24(1):51754. doi:10.1146/ annurev.soc.24.1.517

Franzosi, Roberto. 1998b. "Narrativa como datos: herramientas lingüísticas y estadísticas para el estudio cuantitativo de eventos históricos". Revista Internacional de Historia Social 43(6):81-104. doi:10.1017/S002085900011510X

Franzosi, Roberto. 2009. Análisis narrativo cuantitativo. Mil robles: salvia Publicaciones.

Fuhse, Jan, Oscar Stuhler, Jan Riebling y John Levi Martin. 2020. “Relación de las relaciones sociales y simbólicas en el análisis cuantitativo de textos. Un estudio del discurso parlamentario en la República de Weimar”. Poética 78. doi:10.1016/j.poetic.2019.04.004 Goldenstein, Jan y Philipp

Poschmann. 2019. "Análisis del significado en Big Data: Realización de un análisis de mapas mediante el análisis gramatical y el modelado de temas". Metodología Sociológica 49(1):83131. doi:10.1177/0081175019852762 Grimmer,

Justin y Brandon M. Stewart. 2013. "Texto como datos: la promesa y las trampas de los métodos automáticos de análisis de contenido para textos políticos". Análisis Político 21(3):26797. doi:10.1093/pan/mps028

Hoffman, Mark Anthony, Jean-Philippe Cointet, Philipp Brandt, Newton Key y Peter Bearman. 2018. “La Biblia (protestante), el sermón (impreso) y la(s) palabra(s): la estructura semántica de la Biblia conformista y disidente, 1660–1780”. Poética 68:89-103. doi:10.1016/ j.poetic.2017.11.002

Honnibal, Matthew, Inés Montani, Sofie Van Landeghem y Adriane Boyd. 2020. "spaCy: procesamiento de lenguaje natural de potencia industrial en Python". https://spacy.io .

Jones, Jason, Mohammad Amin, Jessica Kim y Steven Skiena. 2020. "Las asociaciones estereotipadas de género en el lenguaje han disminuido con el tiempo". Ciencias Sociológicas 7:1-35. doi:10.15195/v7.a1

Jurafsky, Dan y James H Martin. 2020. "Capítulo 14. Análisis de dependencia". Procesamiento del habla y el lenguaje. https://web.stanford.edu/ jurafsky/slp3/14.pdf.

Kang, Dong Hyun y James Evans. 2020. "Contra el método: Explosión del límite entre los estudios cualitativos y cuantitativos de la ciencia". Estudios de Ciencias Cuantitativas 1(3):93044. doi:10.1162/qss_a_00056

Karell, Daniel y Michael Freedman. 2019. "Retóricas del radicalismo". Americano Revisión Sociológica 84(4):72653. doi:10.1177/0003122419859519

Koopmans, Ruud y Paul Statham. 1999. “Análisis de reclamos políticos: integración de enfoques de eventos de protesta y discursos políticos”. Movilización 4(2):20321. doi:10.17813/maiq.4.2.d759337060716756

Kozlowski, Austin C., Matt Taddy y James A. Evans. 2019. "La geometría de la cultura: análisis de los significados de clase a través de incrustaciones de palabras". Revisión Sociológica Americana 84(5):90549. doi:10.1177/0003122419877135 Lee,

Mónica y John Levi Martin. 2014. “Codificación, Conteo y Cultura Cartografía." Diario Americano de Sociología Cultural 3(1):133.

Lee, Mónica y John Levi Martin. 2018. "Puerta al Dharma de la Dualidad". Poética 68:1830. doi:10.1016/j.poetic.2018.01.001

Leydesdorff, Loet y Iina Hellsten. 2006. "Medición del significado de las palabras en contextos: un análisis automatizado de controversias sobre 'mariposas monarca', 'Frankenfoods' y 'células madre'". Cienciometría 67(2):23158. doi:10.1007/s11192-006-0096-y

Luz, Ryan. 2014. “De las palabras a las redes y viceversa”. Corrientes Sociales 1(2):111 29. doi:10.1177/2329496514524543

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Berthard y David McClosky. 2014. "El kit de herramientas de procesamiento del lenguaje natural de Stanford CoreNLP". Actas de la 52ª Reunión Anual de la Asociación de Lingüística Computacional: 5560. doi:10.3115/v1/P14-5010 Martín, John Levi.

2000. "¿Qué hacen los animales todo el día?: La división del trabajo, los cuerpos de clase y el pensamiento totémico en la imaginación popular". Poética 27- (2-3):195 231. doi:10.1016/S0304-422X(99)00025-X

McLean, Paul D. 1998. "Un análisis del marco de la búsqueda de favores en el Renacimiento: agencia, redes y cultura política". Diario Americano de Sociología 104(1):5191. doi:10.1086/210002

Mikolov, Tomas, Kai Chen, Greg Corrado y Jeffrey Dean. 2013. "Estimación eficiente de las representaciones de palabras en el espacio vectorial". arXiv:1301.3781.

Mische, Ann y Philippa E. Pattison. 2000. "Componer una arena cívica: públicos, proyectos y entornos sociales". Poética 27(2-3):16394. doi:10.1016/ S0304-422X(99)00024-8

Mohr, John W. 1994. "Soldados, madres, vagabundos y otros: roles discursivos en el Directorio de organizaciones benéficas de la ciudad de Nueva York de 1907". Poética 22(4):32757. doi:10.1016/

0304-422X(94)90013-2 Mohr, John W. y Petko Bogdanov. 2013. "Introducción—Modelos temáticos: qué son y por qué son importantes". Poética 41(6):54569. doi:10.1016/ j.poetic.2013.10. 001

Mohr, John W. y Vincent Duquenne. 1997. "La dualidad de la cultura y la práctica: alivio de la pobreza en la ciudad de Nueva York, 1888-1917". Teoría y Sociedad 26(2/3):305 56. doi:10.1023/A:1006896022092

Mohr, John W. y Helene K Lee. 2000. "De la acción afirmativa a la divulgación: cambios de discurso en la Universidad de California". Poética 28(1):4771. doi:10. 1016/S0304-422X(00)00024-3

Mohr, John W., Robin Wagner-Pacifici, Ronald L. Breiger y Petko Bogdanov. 2013. "Graficación de la gramática de los motivos en las estrategias de seguridad nacional: interpretación cultural, análisis de texto automatizado y el drama de la política global". Poética 41(6):670700. doi:10.1016/j.poetic.2013.08.003

Monroe, Burt L. 2019. "Los significados del "significado" en el análisis de texto científico social". Metodología Sociológica 49(1):1329. doi:10.1177/00811750198 65231

Nelson, Laura K. 2021. "Aprovechamiento de la alineación entre el aprendizaje automático y la interseccionalidad: uso de incrustaciones de palabras para medir las experiencias interseccionales del sur de EE. UU. del siglo XIX". Poética 88. doi:10.1016/j. poético.2021.101539

Padgett, John F., Katalin Prajda, Benjamin Rohr y Jonathan Schoots. 2020. "Discusión y debate político en el tiempo narrativo: la consulta florentina e Pratiche, 1376–1378". Poética 78. doi:10.1016/j.poetic.2019.101377

Pennington, Jeffrey, Richard Socher y Christopher D Manning. 2014. "GloVe: Vectores globales para la representación de palabras". Actas de la Conferencia de 2014 sobre métodos empíricos en el procesamiento del lenguaje natural: 153243. doi:10.3115/ v1/ D14-1162

Penny Cook, Alastair. 1994. "La política de los pronombres". Revista ELT 48(2):1738. doi:10.1093/elt/48.2.173

Polletta, Francesca, Pang Ching Bobby Chen, Beth Gharrity Gardner y Alice Motes. 2011. "La sociología de la narración". Revisión Anual de Sociología 37(1):10930. doi:10.1146/ annurev-soc-081309-150106

Propp, Vladimir. [1928] 1968. Morfología del cuento popular. Austin: Universidad de Prensa de Texas.

Puetz, Kyle, Andrew P. Davis y Alexander B Kinney. 2021. "Estructuras de significado en la política mundial: un análisis de red semántica de la terminología de los derechos humanos en los acuerdos de paz del mundo". Poética. 88. doi:10.1016/j.poetic.2021.101598 Roberts, Carl W. 1989. “Aparte de contar palabras: un enfoque lingüístico para el análisis de contenido”. Fuerzas Sociales 68(1):147 77. doi: 10.2307/2579224 Roberts, Carl W. 1997. "Una gramática semántica genérica para el análisis cuantitativo de texto: aplicaciones al contenido de noticias de radio de Berlín Oriental y Occidental desde 1979". Metodología Sociológica 27(1):89129. doi:10.1111/1467-9531.271020 Ruef, Martín. 1999. "Ontología social y la dinámica de las formas organizacionales: creación de actores de mercado en el campo de la atención médica, 1966-1994". Fuerzas Sociales 77(4):140332. doi:10.2307/3005881 Rule, A., JP Cointet y PS Bearman. 2015. “Cambios léxicos, cambios sustantivos y continuidad en el discurso del estado de la Unión, 1790-2014”. Procedimientos de la Academia Nacional de Ciencias 112 (35): 10837 44. doi:10.1073/pnas.1512221 112. https://www.ncbi.nlm.nih.gov/pubmed/26261302.

Smith, Tammy. 2007. “Límites narrativos y la dinámica del conflicto étnico y la conciliación”. Poética 35(1):2246. doi:10.1016/j.poetic.2006.11.001 Snow, David A. y Doug McAdam. 2000 “Procesos de Trabajo de la Identidad en el Contexto de los Movimientos Sociales: Esclareciendo el Nexo Identidad/Movimiento”. en Self, Identity, and Social Movements, editado por S. Stryker, SJ Owens y RW White. Minnesota: Prensa de la Universidad de Minnesota, págs. 41-67.

Entonces, Richard, Hoyt Long y Yuancheng Zhu. 2019. "Raza, escritura y computación: diferencia racial y la novela estadounidense, 1880-2000". Revista de análisis cultural 3 (2). doi:10.22148/16.031

Spirling, Arthur y Pedro L Rodríguez. 2022. "Incrustaciones de palabras: qué funciona, qué no y cómo distinguir la diferencia para la investigación aplicada". Revista de Política 84(1):10115. doi:10.1086/715162

Stoltz, Dustin S. y Marshall A Taylor. 2019. "Distancia del motor de conceptos: medición de la participación de conceptos a través de incrustaciones de palabras en textos". Revista de Ciencias Sociales Computacionales 2(2):293313. doi:10.1007/s42001-019-00048-6

Stoltz, Dustin S. y Marshall A Taylor. 2021. “Cartografía cultural con incrustaciones de palabras”. Poética88. doi:10.1016/j.poetic.2021.101567 Straka, Milán. 2018. “Prototipo UDPipe 2.0 en CoNLL 2018 UD Shared Task”. Actas de la tarea compartida CoNLL 2018: análisis multilingüe de texto sin procesar a dependencias universales: 197-207

Stuhler, Óscar. 2021. “¿Qué hay en una categoría? Un nuevo enfoque del papel del discurso Análisis." Poética 88. doi:10.1016/j.poetic.2021.101568

Tauberg, Michael. 2019. "¿Qué tan inteligente es su fuente de noticias?". Consultado el 22 de noviembre de 2021 (https://towardsdatascience.com/how-smart-is-your-news-source-1fe0c550 c7d9.

Taylor, Marshall y Dustin Stoltz. 2020. "Análisis de clases de conceptos: un método para identificar esquemas culturales en textos". Ciencias Sociológicas 7:544-69. doi:10. 15195/v7.a23

Taylor, Marshall A. y Dustin S Stoltz. 2021. "Integración de direcciones semánticas con la distancia del motor de conceptos para medir el compromiso de conceptos binarios". Revista de Ciencias Sociales Computacionales 4(1):23142. doi:10.1007/s42001-020-00075-8 Tilly, Carlos. 1995. Contienda popular en Gran Bretaña, 1758-1834. Cambridge:
Prensa de la Universidad de Harvard.

Tilly, Carlos. 1997. "Parlamentarización de la contienda popular en Gran Bretaña, 1758-1834". Teoría y Sociedad 26(2/3):24573. doi:10.1023/A:1006836012345 Todorov, Tzvetan y Arnold Weinstein. 1969. "Análisis estructural de la narrativa".
NOVELA: Un foro de ficción 3(1):706. doi:10.2307/1345003

Tyrkkö, Jukka. 2016. “Buscando umbrales retóricos: frecuencias de pronombres en discursos políticos”. Estudios de variación, contactos y cambios en inglés 17.

Underwood, Ted. 2019. Horizontes lejanos. Evidencia digital y cambio literario. Chicago: Prensa de la Universidad de Chicago.

Universal-Dependencias. 2020. "Marco de dependencias universales". http://www.dependenciasuniversales.org.

van Atteveldt, Wouter, Jan Kleinnijenhuis y Nel Ruigrok. 2008. "Análisis, redes semánticas y autoridad política mediante el análisis sintáctico para extraer relaciones semánticas de artículos de periódicos holandeses". Análisis Político 16(4):42846. doi:10.1093/pan/mpn006 van Atteveldt, Wouter, Tamir Sheafer, Shaul R. Shenhav y Yair Fogel-Dror. 2017. "Análisis de cláusulas: uso de información sintáctica para extraer automáticamente la fuente, el sujeto y el predicado de los textos con una aplicación a la Guerra de Gaza de 2008-2009". Análisis Político 25(2):20722. van Dijk, Teun A. 1972. Algunos aspectos de las gramáticas del texto. Un estudio de lingüística teórica y poética. La Haya/París: Mouton. Von Wiese, Leopoldo. 1965. Die Philosophie der persönlichen Fürwörter. Tubinga: Mohr.

Wada, Takeshi. 2004. “Análisis de eventos de reivindicación en México: ¿Cómo se transforman las protestas sociales en protestas políticas?” Movilización 9(3):24157. doi:10. 17813/maiq.9.3.7wx2pt66130718v3 Wagner-Pacifici, Robin. 2010. "Teorizando la inquietud de los eventos". Americano Revista de Sociología 115(5);1351-1386.

Welbers, Kasper, Wouter van Atteveldt y Jan Kleinnijenhuis. 2021. "Extracción de relaciones semánticas mediante sintaxis". Investigación en comunicación computacional 3(2):116. doi:10.5117/CCR2021.2.003.WELB

Wickham, Hadley. 2021. "nombres de bebé: nombres de bebés de EE. UU. 1880-2017". https://cran.r
project.org/web/packages/babynames/index.html.
Wodak, Ruth, Rudolf de Cillia, Martin Reisigl y Karin Liebhart. 2009. La Construcción Discursiva de la
Identidad Nacional. Edimburgo: Prensa de la Universidad de Edimburgo.

Young, Vicente. 2021. “Un enfoque visual para interpretar la carrera de la red
Metáfora." Poética 88. doi:10.1016/j.poetic.2021.101566

Biografía del autor

Oscar Stuhler es estudiante de doctorado en el Departamento de Sociología de la Universidad de Nueva York. Estudia las representaciones de las estructuras sociales en el discurso.