PUBLICACIONES DE LA
ASOCIACIÓN ARGENTINA
DE HUMANIDADES DIGITALES

Volumen 4 2023
e-ISSN 2718-7470
https://doi.org/10.24215/27187470e048
https://revistas.unlp.edu.ar/publicaahd/

Tomar distancia de los medios. El aporte de Voyant Tools en el análisis discursivo de la representación de los jóvenes del Conurbano Bonaerense y de la Ciudad Autónoma de Buenos Aires en el periódico La Nación

Take distance from the media. The Contribution of Voyant Tools in the Discursive Analysis of the Representation of Youth from Greater Buenos Aires and Autonomous City of Buenos Aires in the Newspaper La Nación

Irene TIMOSZKO
irene.timoszko@gmail.com
Instituto Superior de Formación Docente N 82 Carlos Fuentealba, Argentina.
https://orcid.org/0000-0002-0924-0495

Cita recomendada:

TImoszko, I. (2023). Tomar distancia de los medios. El Aporte de Voyant Tools en el análisis discursivo de la representación de los jóvenes del Conurbano Bonaerense y de CABA en el periódico La Nación. *Publicaciones de la Asociación Argentina de Humanidades Digitales*, *4*, e048. https://doi.org/10.24215/27187470e048

RECIBIDO: 26 de Agosto de 2022 **ACEPTADO:** 10 de Noviembre de 2022

RESUMEN
Este trabajo se propone explorar la herramienta Voyant Tools en el análisis de un corpus de noticias referidas a los jóvenes de dos zonas geográficas distintas (Ciudad Autónoma de Buenos Aires y conurbano bonaerense), publicadas por el diario *La Nación* durante el periodo de un año (abril 2021-abril 2022). El objetivo principal radica en la valoración de las posibilidades de procesamiento de datos a partir de herramientas de lectura distante (Moretti, 2015) en los proyectos de investigación enmarcados en la metodología del Análisis Crítico del Discurso (ACD) de los medios masivos de comunicación (Fairclough, 1992; Fowler et al., 1979). Por otro lado, nos proponemos como objetivo específico realizar un análisis exploratorio sobre las representaciones discursivas de los jóvenes en uno de los periódicos más importantes de la prensa argentina.

PALABRAS CLAVE: Análisis Crítico del Discurso, Voyant Tools, lectura distante, jóvenes, Conurbano Bonaerense, Ciudad Autónoma de Buenos Aires.

ABSTRACT
This work aims to explore Voyant Tools in the analysis of a corpus of news referring to young people from two different geographical areas (Autonomous City of Buenos Aires and Greater Buenos Aires), published by the newspaper *La Nacion* during a period of one year (April 2021-April 2022). The main objective is to assess the possibilities of data processing from distant reading tools (Moretti, 2015) in research projects framed in the methodology of Critical Discourse

Analysis (CDA) of the mass media (Fairclough, 1992; Fowler et al., 1979; Pardo et al., 2018). On the other hand, we propose as a specific objective to carry out an exploratory analysis on the discursive representations of young people in one of the most important newspapers of the Argentine press.

KEYWORDS: Analysis, Voyant Tools, distant reading, young people, Greater Buenos Aires, Autonomous City of Buenos Aires.

1. LA LECTURA DISTANTE COMO HERRAMIENTA METODOLÓGICA DEL ACD

Uno de los principales propósitos del Análisis Crítico del Discurso (ACD) es reconocer determinadas estructuras lingüísticas que –lejos de ser neutrales– enmascaran, sostienen y reproducen discursos de dominación y desigualdad (Van Dijk, 1984, 1995; Fairclough, 1992; Pardo et al., 2018). Estos discursos resultan más poderosos en la medida en que se vuelven hegemónicos (es decir, aceptados y reproducidos) y, paradójicamente, invisibles. Desde el ACD se sostiene que no hay nada más invisible que lo obvio, aquello que se transforma en un sistema de creencias, lo que naturalizamos –aunque no tenga nada de natural– como sentido común (Raiter, 2003).

Este campo disciplinar se orienta a buscar y analizar regularidades en determinadas formaciones discursivas (Foucault, 1959, 1971), las cuales se asientan en representaciones sociales y a la vez las reproducen. Si bien su metodología es cualitativa, la búsqueda de regularidades implica el análisis de corpus extensos para determinar patrones discursivos que resulten representativos.

Precisamente en la búsqueda de regularidades es que consideramos que desde las Humanidades Digitales se ofrece una serie de herramientas metodológicas y conceptuales para pensar el procesamiento de grandes cantidades de información. El concepto de lectura distante (Moretti, 2015) se define como un tipo de lectura a gran escala, lo cual permite el abordaje cuantitativo de una cantidad de datos que sería imposible leer de manera individual, localizada y lineal, y que se realiza con auxilio de programas de procesamiento informático(aunque no de manera exclusiva). El desarrollo de estas herramientas y las investigaciones sobre

su aplicación al campo específico de las Humanidades, revela la necesidad de repensar metodologías tradicionales de análisis:

> Tanto lingüistas computacionales como especialistas de la recuperación de la información han creado y utilizado software para apreciar patrones que no son evidentes en una lectura tradicional o bien para corroborar hipótesis que intuían al leer ciertos textos pero que requerían de trabajos laboriosos, costosos y mecánicos (Gutiérrez de la Torre, 2019).

La pertinencia de este tipo de acercamiento para la investigación sociolingüística reside en la posibilidad de reducir un flujo de texto desestructurado y cualitativo a un conjunto de símbolos manipulables y cuantitativos, lo que permite validar inferencias iniciales. En este artículo exploramos una herramienta que posibilita este tipo de procesamiento: Voyant Tools.

Voyant Tools[1] es una plataforma en línea que ofrece una serie de opciones para el procesamiento de textos. Básicamente, lo que Voyant Tools permite es rastrear un determinado ítem (palabra, frase) de acuerdo a parámetros de presencia, intensidad y frecuencia en un corpus. Estos datos, recabados de manera cuantitativa, a partir de los algoritmos que se aplican en las distintas opciones, nos posibilita a los analistas del discurso confirmar –o no- hipótesis iniciales y descubrir otros datos no visibles al ojo humano. El propósito de este artículo será, entonces, explorar esta herramienta en un corpus de noticias del diario *La Nación* a fin de relevar constantes en la construcción discursiva de los jóvenes de la Ciudad de Buenos Aires.

1.1. Los jóvenes en la prensa argentina

Tanto en el conurbano bonaerense como en la Ciudad de Buenos Aires las condiciones sociales en las que viven los jóvenes no resultan muy diferentes. De acuerdo al Observatorio Sociolaboral de los Jóvenes del Conurbano Bonaerense (2018), el grupo compuesto por personas de

[1] Sitio web de Voyant Tools: www.voyant-tools.org.

entre 15 y 24 años ascendía a 2.165.589, lo que representaba el 16.5 % de la población de esta región, mientras que, de acuerdo a los datos publicados por el Gobierno de la Ciudad de Buenos Aires (Dirección General de Estadísticas y Censos, 2017) la población joven de esta región representa el 19.8%. Si tomamos en cuenta el índice de pobreza, en el conurbano el 42% de los jóvenes son pobres (casi un 10% más del índice de pobreza de la población en general) mientras que en la Ciudad de Buenos Aires el 19.5% de los jóvenes reside en hogares en condiciones de pobreza, un 8% más que la media. Si bien en términos globales el índice de pobreza en la población del conurbano se duplica, también es cierto que en ambas regiones la pobreza afecta en mayor medida al grupo de jóvenes que al resto de la población. El nivel de educación, su permanencia o no en el sistema educativo y sus condiciones laborales tampoco marcan una gran diferencia entre ambas zonas, y mantienen un factor común que es la precariedad y la vulnerabilidad. Ahora bien, si las condiciones socioeconómicas entre ambos grupos son bastante similares, podría pensarse que las representaciones sociales construidas en torno a los jóvenes de un lado y del otro de la General Paz también deberían serlo. Sin embargo, intentaremos demostrar, a partir de un ejercicio de lectura distante de los medios de comunicación, que existen diferencias significativas en el modo en que la prensa argentina los representa.

Para esto, analizamos un corpus de doscientas noticias publicadas en el periodo abril de 2021- abril de 2022 en el diario *La Nación*. Si bien este corpus no resulta representativo del total de la prensa argentina consideramos que sí es útil para llegar a algunas conclusiones provisorias sobre dos aspectos específicos: la viabilidad de las herramientas ofrecidas por Voyant Tools para el procesamiento de un corpus extenso, y más lateralmente, ofrecer un panorama exploratorio de las estrategias discursivas de este diario en la representación de los jóvenes de ambas regiones.

2. METODOLOGÍA

2.1. Selección y recolección del corpus: criterios

Como mencionamos anteriormente, hemos cargado dos corpus en Voyant Tools. Explicaremos brevemente cómo llegamos a seleccionarlos. La primera acción fue escribir los ítems *Jóvenes* y *conurbano*, por un lado, y *Jóvenes* y *CABA*[2], por otro, en el buscador de La Nación, en su edición digital. En la búsqueda se desestimó el uso de mayúsculas y minúsculas y las opciones *ciudad de Buenos Aires* y CABA Luego seleccionamos un rango de noticias del último año, lo cual arrojó los siguientes resultados:

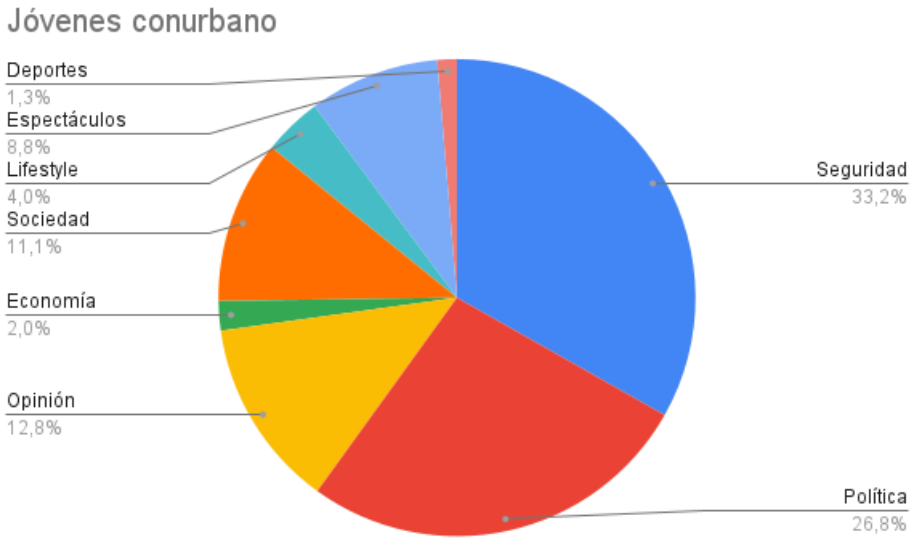


Figura 1. Gráfico de distribución de noticias por sección del corpus *Jóvenes Conurbano*. Fuente: Elaboración propia.

[2] Ciudad Autónoma de Buenos Aires

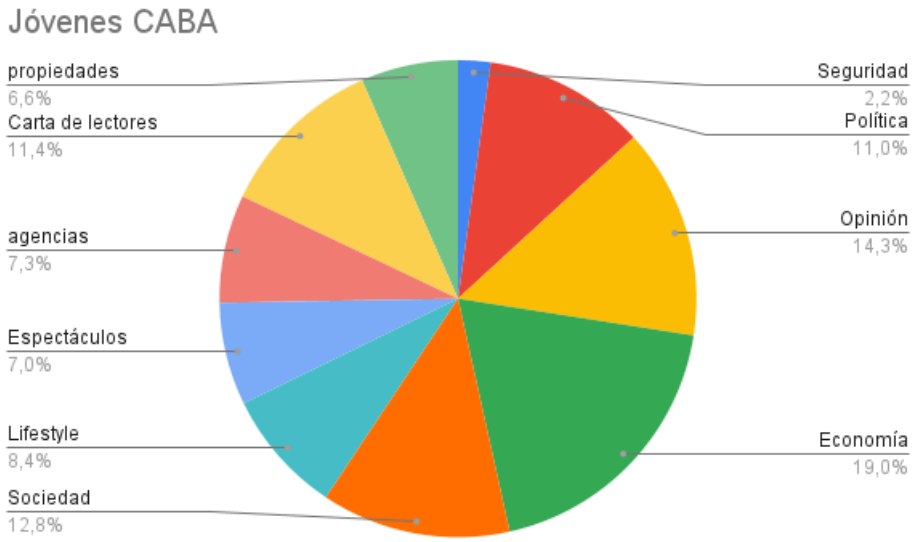


Figura 2. Gráfico de distribución de noticias por sección del corpus *Jóvenes CABA*. Fuente: Elaboración propia.

En esta oportunidad, y teniendo en cuenta que se trata de un trabajo exploratorio de las herramientas Voyant Tools, solo tomaremos el corpus de noticias de la sección más representativa referida a *Jóvenes conurbano*, es decir, las 150 noticias de la sección *Seguridad* y las 52 noticias de la sección *Economía*, la más representativa de la búsqueda *Jóvenes CABA*. No obstante, estos primeros gráficos ya nos aportan un dato interesante en la lectura distante del corpus: la abrumadora preponderancia de noticias en *Seguridad*, en la que los jóvenes del conurbano son los protagonistas, frente a la diversidad de secciones arrojadas en la búsqueda orientada a *Jóvenes CABA*.

Ambos grupos de noticias fueron cargados como documento único cada uno. Es decir, procesamos con Voyant Tools dos documentos. En el próximo subapartado se describirán las herramientas aplicadas y los datos arrojados.

2.2. Explorando (con) Voyant Tools

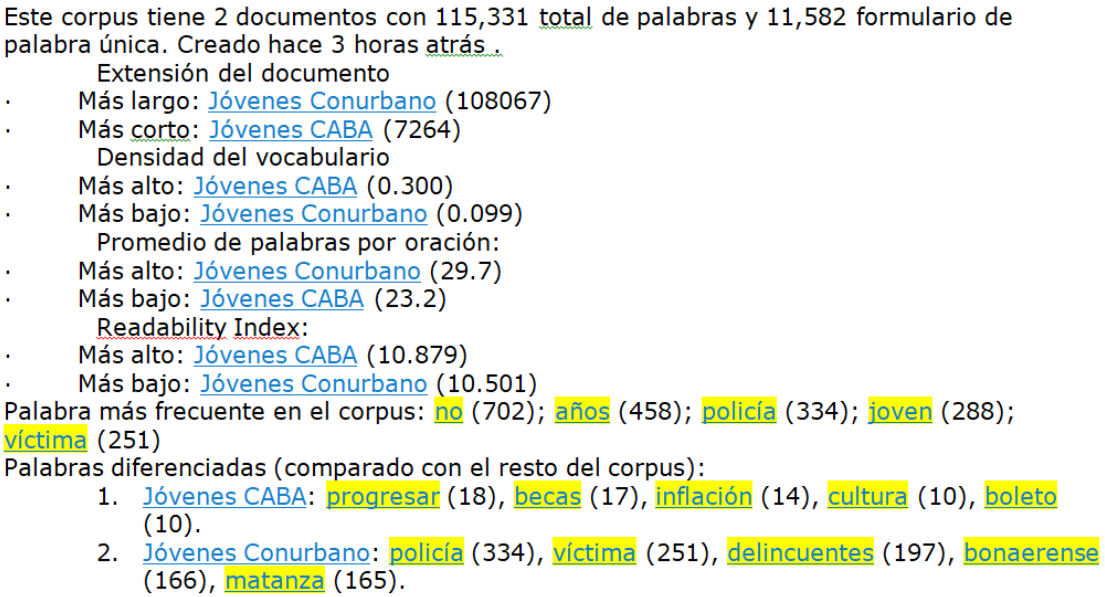
Voyant Tools es una plataforma libre, gratuita y fácil de usar, es intuitiva, se ejecuta en línea y no requiere registro previo. El programa fue iniciado por dos académicos canadienses, Stéfan Sinclair (McGill

University) y Geoffrey Rockwell (University of Alberta), y sus colaboradores. El software permite procesar y estudiar textos propios o ajenos, un documento o una colección de textos (corpus digital), cargados desde el ordenador o desde la red (puede ser un blog, una página web), en formato de texto plano, MS Word, PDF, RTF, HTML o XML. En esta oportunidad, cargamos ambos documentos con formato de texto plano. A continuación, explicaremos los resultados del procesamiento a partir del uso de cinco de las 29 herramientas que ofrece la plataforma.

2.3. Datos arrojados

2.3.1. Primera herramienta: Sumario

La herramienta Sumario nos permite analizar y comparar ambos corpus a partir de cuatro criterios: extensión, densidad de vocabulario, promedio de palabras por oración e índice de legibilidad.



Este corpus tiene 2 documentos con 115,331 total de palabras y 11,582 formulario de palabra única. Creado hace 3 horas atrás .

Extensión del documento
- Más largo: Jóvenes Conurbano (108067)
- Más corto: Jóvenes CABA (7264)

Densidad del vocabulario
- Más alto: Jóvenes CABA (0.300)
- Más bajo: Jóvenes Conurbano (0.099)

Promedio de palabras por oración:
- Más alto: Jóvenes Conurbano (29.7)
- Más bajo: Jóvenes CABA (23.2)

Readability Index:
- Más alto: Jóvenes CABA (10.879)
- Más bajo: Jóvenes Conurbano (10.501)

Palabra más frecuente en el corpus: no (702); años (458); policía (334); joven (288); víctima (251)

Palabras diferenciadas (comparado con el resto del corpus):
1. Jóvenes CABA: progresar (18), becas (17), inflación (14), cultura (10), boleto (10).
2. Jóvenes Conurbano: policía (334), víctima (251), delincuentes (197), bonaerense (166), matanza (165).

Figura 3. Datos arrojados por la herramienta Sumario para el procesamiento de ambos corpus. Fuente: Elaboración propia.

Considerando que la extensión del corpus 1 (*Jóvenes conurbano*) es considerablemente mayor que la del 2 (*Jóvenes CABA*), resulta significativo el dato de densidad de vocabulario. Este se obtiene de la

división del número de palabras únicas entre el número de palabras totales. Cuanto más cercano al valor 1 es el índice de densidad, el vocabulario tiene mayor variedad de palabras, es decir, es más denso. El dato preciso que nos arroja esta herramienta es que, aun cuando la cantidad de palabras del corpus 1 es significativamente mayor, no son tan distintas entre sí. Esto puede deberse, en principio, a la mayor uniformidad temática que implica la sección *Seguridad* frente a una mayor diversidad de la sección *Economía*. Sin embargo, resulta interesante apuntar otra observación: la diversidad (de secciones, de palabras) sigue siendo otra constante en el corpus 2.

La herramienta Sumario también nos permite una primera aproximación al rango de frecuencia de las palabras de cada corpus en común y diferenciadas, pero vamos a visualizar esta frecuencia bruta a partir de la herramienta Cirrus.

2.3.2. Segunda herramienta: Cirrus

Para visualizar los términos con más alta frecuencia bruta (es decir, con más cantidad de apariciones en el corpus) se utilizó la herramienta Cirrus. Si se desplaza el cursor por cada uno de los términos, puede obtenerse el dato exacto de cuántas veces aparece en el documento. Pero lo más interesante es que permite visualizar una nube de palabras que conforma el campo semántico más relevante en cada corpus. Un dato metodológico: en estas herramientas de conteo se pueden descartar palabras vacías (cuyo valor más que semántico es gramatical) a partir de la edición manual de las palabras excluidas. En este caso, se omitieron de la selección adverbios, preposiciones, pronombres, artículos, verbos auxiliares, etc.):

Figura 4. Resultados arrojados por la herramienta Cirrus para el corpus 1 (*Jóvenes Conurbano*). Fuente: Elaboración propia.

Figura 5. Resultados arrojados por la herramienta Cirrus para el corpus 2 (*Jóvenes CABA*). Fuente: Elaboración propia.

2.3.3. Tercera herramienta: Términos y Tendencia

Sin embargo, Cirrus mide la frecuencia bruta de un término en el texto, lo cual puede llevar a una lectura sobredimensionada de un término en un corpus más extenso que otro con menor extensión. Para eso, existe la posibilidad de medir la frecuencia relativa normalizada. Esta se calcula de la siguiente manera: Frecuencia Bruta * 1,000,000 / Número total de palabras. De esta manera, con la herramienta Términos se puede evaluar la tendencia de una palabra en ambos corpus. En este caso, vemos que *policía* tiene una frecuencia relativa 0 para el corpus 2, mientras que presenta una frecuencia muchísimo más alta en el corpus 1:

Figura 6. Frecuencia relativa de *policía* en el corpus 2. Fuente: Elaboración propia.

Figura 7. Frecuencia relativa de *policía* en el corpus 1. Fuente: Elaboración propia.

Al comparar una palabra común a ambos corpus, como *joven*, veremos que, aun cuando en el corpus 2 su frecuencia bruta sea menor, esta diferencia se corrige si tomamos en cuenta la frecuencia relativa, como podemos observar en el siguiente gráfico generado por la herramienta Tendencia. En el mismo, se puede observar que el término *joven* tiene una frecuencia relativa casi similar en ambos corpus, mientras que el término *policía* tiene una frecuencia relativa menor en el corpus *Jóvenes CABA* que en *Jóvenes Conurbano*

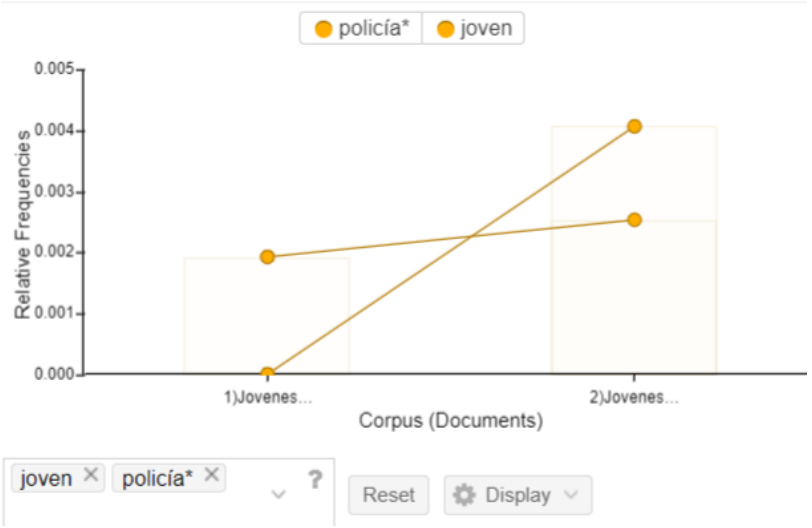


Figura 8. Tendencia: mide la frecuencia relativa de los términos *joven* y *policía* en ambos corpus. En este caso, la primera columna mide el corpus *Jóvenes CABA* y la segunda *Jóvenes Conurbano*. Fuente: Elaboración propia.

2.3.4. Cuarta herramienta: Términos Berry

La herramienta TermsBerry proporciona una forma de explorar términos de alta frecuencia y sus colocaciones (palabras que ocurren en la proximidad). Nos permite visualizar a grandes rasgos el contexto semántico en el que aparece un término en particular. Sin dudas, esta herramienta nos ofrece un análisis visual un poco más preciso que Cirrus, en tanto permite hacer una aproximación al peso de los vínculos entre las palabras-nodos. En este caso, observemos la palabra *joven* en el corpus número 1:

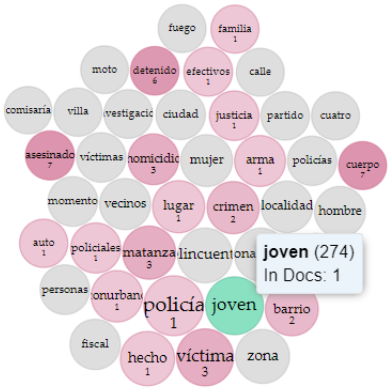


Figura 9. Resultados arrojados por la herramienta Términos Berry para el término *joven* en el corpus 1 (*Jóvenes Conurbano*). Fuente: Elaboración propia.

2.3.5. Quinta herramienta: análisis de contextos

La quinta y última herramienta que analizaremos resulta la más atractiva para profundizar en la metodología del ACD, luego de un primer acercamiento exploratorio. Contextos nos permite recuperar el contenido verbal próximo de un término a la izquierda y a la derecha del mismo. Las opciones permiten ajustar la extensión de ese contexto en cantidad de palabras a un lado y otro, seleccionar un documento o trabajar con varios al mismo tiempo. Otro dato importante es que se puede recuperar la emisión completa con la ubicación del término dentro del corpus:

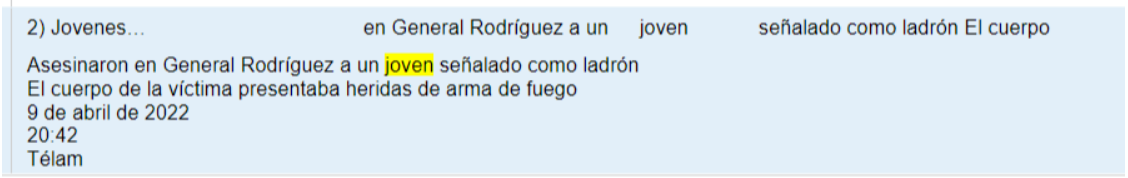


Figura 10. Herramienta Contextos. Ejemplo de emisión completa y ubicación de *joven* en el corpus 1. Fuente: Elaboración propia.

El recurso también permite exportar estos datos en lenguaje HTML, JSON o como documento plano, lo cual es de muchísima utilidad para trabajar con programas de procesamiento semántico y, para el Análisis Crítico del Discurso, implica un abanico de posibilidades de investigación y desarrollo de software que pueda hacer posible, por ejemplo, el análisis de roles temáticos que refieren a posiciones de las entidades discursivas determinadas por los argumentos verbales: agente, experimentador, paciente, beneficiario, etc. En cuanto a los beneficios específicos de Voyant Tools, esta herramienta permite agilizar la identificación de las emisiones gracias la recuperación de los contextos de aparición de un término.

3. CONCLUSIONES

El objetivo principal que planteamos en este artículo es la valoración de la herramienta Voyant Tools para el ACD. De acuerdo a lo expuesto en este estudio exploratorio, podemos ponderar algunas ventajas que ofrece

esta plataforma para el procesamiento de corpus extensos. En primer lugar, Voyant Tools permite identificar patrones discursivos de manera rápida y fiable. Procesar esta cantidad de datos de forma manual requeriría de recursos humanos y de tiempo infinitamente mayor, sin tener en cuenta también un posible margen de error. Sin embargo, no es el ahorro de tiempo y recursos la única ventaja de las tecnologías de procesamiento automático. El abanico de posibilidades de aplicación que las herramientas de software y los bancos de datos estandarizados pueden ofrecer para el ACD en particular, y para la lingüística en general, es innegable. Entre ellos, podemos mencionar proyectos como OntoNotes Release 5.0 (Ralph Weischedel, Martha Palmer, entre otros) desarrollados por Linguistic Data Consortium[3] o ThePropBank[4] (The Proposition Bank) que brindan bancos de datos con información sobre las proposiciones semánticas básicas. En español, podemos mencionar a AncoraNet[5] que, tal como se menciona en su sitio, se trata de un léxico multilingüe catalán, español e inglés que incluye la combinación de en el que se combina información sintáctico-semántica y conceptual procedente de las distintas fuentes integradas Estas herramientas y bancos de datos ofrecen información procesada lingüísticamente, lo cual es un recurso cualitativamente valioso para la investigación.

En segundo lugar, Voyant Tools validó nuestras hipótesis iniciales en relación con la representación discursiva de los jóvenes en la prensa argentina, nuestro objetivo específico, en la medida en que permitió confirmar que:

1. El volumen de palabras destinadas a representar a los jóvenes del Conurbano en contextos de violencia e inseguridad es prácticamente incomparable en cuanto a su dimensión casi absoluta frente a otras secciones del diario.

[3] https://www.ldc.upenn.edu/.
[4] https://propbank.github.io/.
[5] http://clic.ub.edu/corpus/es/ancora.

2. Los jóvenes de CABA, por el contrario, tienen un abanico más amplio de representaciones donde los contextos de inseguridad son casi inexistentes.
3. Si miramos los términos asociados a uno y otro corpus, podremos fácilmente descubrir que, a pesar del gran volumen de palabras que se utiliza para representar a los jóvenes, estas siempre son las mismas: la constancia en roles semánticos asociados a agentes y pacientes de hechos de violencia pareciera demostrar y reafirmar cuáles son sus únicas acciones posibles.

REFERENCIAS BIBLIOGRÁFICAS

Dirección General de Estadísticas y Censos. (2017). *La situación de los jóvenes en la Ciudad de Buenos Aires*. Estadística y Censos de la Ciudad de Buenos Aires. https://www.estadisticaciudad.gob.ar/eyc/publicaciones/situacion_jovenes_caba_2019/index.html#content4-1e

Fairclough, N. (1992). *Discourse and social change*. Polity Press-Blackwell Publishers.

Foucault, M. (1959). *La arqueología del saber.* Siglo Veintiuno.

Foucault, M. (1971). *El orden del discurso*. Tusquets.

Fowler, R. y Kress, G. (1979). *Lenguaje y control*. Fondo de Cultura Económica.

Gutiérrez De la Torre, S. (2019). Análisis de corpus con Voyant Tools. *Programming Historian en español, 3*. https://doi.org/10.46430/phes0043

Moretti, F. (2015). *Lectura distante*. Fondo de Cultura Económica.

Observatorio Sociolaboral de los Jóvenes del Conurbano Bonaerense. (2018). *Vulnerabilidad, precariedad y desafiliación de los jóvenes del conurbano bonaerense en la post crisis. Análisis Sociolaboral II. Actualizado al II trimestre 2018.* Universidad Nacional de San Martín. *https://www.unsam.edu.ar/observatorio-jovenes/analisis18.html*

Pardo, M. L., Marchese, M. C., y Soich, M. (2018). Nuevos aportes desde Latinoamérica para el desarrollo del "Método Sincrónico-Diacrónico de Análisis Lingüístico de Textos". *Chasqui. Revista Latinoamericana de Comunicación*, *139*, 95-114.

Raiter, A. (2003). *Lenguaje y sentido común. Las bases para la formación del discurso dominante.* Biblos.

Van Dijk, T. A. (1984). *Prejudice in discourse*. Bejamin.

Van Dijk, T. A. (1995). Discourse, Semantics and Ideology. *Discourse & Society*, *6*(2), 243-289.