PUBLICACIONES DE LA
ASOCIACIÓN ARGENTINA
DE HUMANIDADES DIGITALES

Volumen 5 2024
e-ISSN 2718-7470
https://doi.org/10.24215/27187470e060
https://revistas.unlp.edu.ar/publicaahd/

El discurso de Navidad del rey de España del año 2022 a través de Analhitza

The 2022 Christmas speech of the King of Spain through Analhitza

Julia Muñoz Moreno de Vega
julia.munoz.morenodevega@gmail.com
Universidad Nacional de Educación a Distancia (UNED)
https://orcid.org/0009-0000-7774-5033

Cita recomendada:

Muñoz Moreno de Vega, J. (2024). El discurso de Navidad del rey de España del año 2022 a través de Analhitza. *Publicaciones de la Asociación Argentina de Humanidades Digitales*, 5, e060. https://doi.org/10.24215/27187470e060

RECIBIDO: 04/10/2023 **ACEPTADO:** 29/10/2024

RESUMEN

El presente estudio examina la plataforma ITXA y su herramienta Analhitza, desarrollada por la Universidad del País Vasco. ITXA proporciona recursos de lingüística computacional para el euskera, el español y el inglés, facilitando el análisis de textos mediante herramientas como Xuxen.
Analhitza permite la extracción automática de información lingüística, siendo útil para áreas como la investigación o la enseñanza. Para poder conocer las capacidades y limitaciones de la herramienta, se ha llevado a cabo un análisis del discurso de Navidad del rey de España del año 2022. En su evaluación, la herramienta demuestra su eficacia en el análisis semántico y morfológico, del mismo modo que se detectan ciertas áreas de mejora.

Palabras clave: lingüística computacional, multilingüismo, procesamiento del lenguaje natural, análisis de textos, Analhitza.

ABSTRACT

This study examines the ITXA platform and one of its tools, Analhitza, developed by the University of the Basque Country. ITXA provides computational linguistics resources for Basque, Spanish and English, facilitating the analysis of texts using tools such as Xuxen.
Analhitza allows the automatic extraction of linguistic information, being useful for areas such as research or teaching. An analysis of the King of Spain's Christmas speech in 2022 has been carried out to understand the tool's capabilities and limitations. In its evaluation, the tool demonstrates its effectiveness in semantic and morphological analysis and detecting certain areas for improvement.

KEYWORDS: Computational Linguistics, Multilingualism, Natural Language Processing, Text Analysis, Analhitza.

1. INTRODUCCIÓN

Las herramientas de procesamiento del lenguaje natural (PLN) y lingüística computacional se utilizan para analizar y comprender el lenguaje humano mediante técnicas de carácter tecnológico e informático. Estos recursos son esenciales para muchas aplicaciones en la era digital, como la traducción automática, la clasificación o generación de texto y el reconocimiento de voz. Asimismo, también son utilizadas dentro del ámbito de la investigación lingüística y en áreas como la inteligencia artificial y la robótica (Ruiz Fabo y Bermúdez Sabel, 2019).

La plataforma ITXA[1] es una iniciativa de la universidad del País Vasco (UPV/EHU) en colaboración con otras instituciones, y su objetivo es ofrecer una batería de recursos y herramientas de lingüística

[1] Disponible en: https://www.ixa.eus/.

computacional en euskera y otras lenguas, como el español o el inglés. Se trata de una plataforma diseñada para desempeñar funciones como el análisis morfológico, sintáctico, semántico y la generación de texto, entre otras, y que proporciona acceso a diversos instrumentos y recursos lingüísticos, como corpus, aplicaciones de procesamiento del lenguaje natural y programas orientados hacia la enseñanza o la investigación. Además, cuenta con un servicio de asesoramiento y consultoría lingüística para empresas y organizaciones que necesiten asistencia en estas áreas. Una de las ventajas más llamativas de este proyecto es su interfaz de usuario fácil de utilizar, que permite a este la configuración y personalización de la herramienta en base a sus necesidades. Gracias a su experiencia y compromiso con la innovación, se ha consolidado como una referencia en el campo de la lingüística computacional, contribuyendo al desarrollo y avance de esta disciplina no sólo en el contexto del euskera, sino también en el del español y el inglés. Entre las plataformas y herramientas que ITXA facilita como recursos para el análisis y procesamiento de lenguaje natural cabe destacar alguna de ellas. Por ejemplo, Xuxen[2] se caracteriza por ser una plataforma de aprendizaje de euskera con herramientas que contribuyen en la evaluación y la generación de ejercicios y recursos didácticos. Otras dos herramientas, también de suma importancia, son Elhuyar-Word[3], un sistema de diccionario integrado en el procesador de textos Word 2000, y EusEduSeg[4], un segmentador automático de discursos en euskera. En el caso de Analhitza, esta sirve para el estudio y visualización de textos y que será el foco principal de este trabajo (Otegi et al., 2017).

Analhitza es una herramienta de procesamiento del lenguaje natural, como ya se ha anticipado, utilizada para la extracción de información lingüística de textos de diferentes idiomas (euskera, español e inglés) de manera automatizada, con el objetivo de facilitar la investigación en el campo de las humanidades. No obstante, no es la

[2] Disponible en: https://xuxen.eus/.
[3] Disponible en: https://www.ixa.eus/node/4465?language=eu.
[4] Disponible en: https://www.ixa.eus/node/4489?language=eu.

única área que abarca, ya que también es útil para la enseñanza de lenguas extranjeras, la industria editorial y la minería de datos.

Esta plataforma cuenta con dos versiones: una en línea[5], que permite a los usuarios subir un archivo de texto, escribirlo directamente en una caja o especificar la URL de un sitio web que contenga el texto a analizar; y una versión interna que tiene una interfaz de línea de comando y permite analizar varios a la vez. De igual forma, ambas versiones generan los resultados en formato de hoja de cálculo Excel.

Analhitza se sirve de diferentes técnicas de PLN para procesar el texto, incluyendo un tokenizador, que sustituye los datos sensibles por símbolos de identificación únicos, que conserven toda la información importante; un lematizador, que reduce las palabras de los textos a su raíz; un etiquetador de partes de la oración, que identifica cada una de ellas; y, finalmente, un reconocedor y clasificador de entidades nombradas. Con esta, es posible estudiar la frecuencia de las palabras, la distribución de las categorías gramaticales y la identificación de patrones de co-ocurrencia, entre otros aspectos.

2. METODOLOGÍA

La propuesta de este trabajo consiste en procesar el discurso de Navidad del rey de España del año 2022[6] y extraer de él información lingüística, a la vez que busca evaluar la herramienta Analhitza. Como bien se ha mencionado, las herramientas de PLN generan una serie de resultados de los que se obtiene información sobre las clases de palabras. En este caso, fueron generadas seis listas con la agrupación de las palabras por categoría gramatical en el siguiente orden: nombres, adjetivos, verbos, adverbios, determinantes, conjunciones y preposiciones. Para cada una de estas, además, se obtuvoun número

[5] Disponible en: https://ixa2.si.ehu.eus/clarink/analhitza.php?lang=es. Un trabajo similar a este en Alonso y Volkens (2012).
[6] Disponible en: https://www.casareal.es/sitios/listasaux/Documents/Mensajenavidad20221224/20221224_mensaje_navidad_esp_rey_felipe.pdf.

concreto de tokens, por ejemplo, para los sustantivos se asocian 347 tokens, es decir, este número representa todas las palabras de esta categoría gramatical sin importar si se repiten o no.

También se realizó un conteo delemas, proporcionados en el apartado siguiente, que representan las formas básicas de las palabras que se utilizan para poder identificar su significado en el diccionario. Por ejemplo, para este texto concreto un lema esencial es *ser* que agrupa este verbo en todas sus variantes. Esta técnica permite recuperar una palabra desde su raíz sin importar sus accidentes morfológicos y así reducir el tamaño de los resultados y por ende, de las matrices de términos con los que se trabaja.

Asimismo, Analhitza analiza también las secuencias de n-grams, que son sucesiones de dos o más palabras consecutivas. Para este caso concreto, un dato que refleje 2-grams sería *Unión Europea*. Este tipo de análisis es muy útil porque identifica patrones y estructuras dentro del propio texto.

A continuación, una vez entendida la finalidad de los diferentes tipos de información que proporciona esta herramienta, se procederá al análisis exhaustivo de ciertos resultados para valorar su efectividad.

2.3. Resultados

Cuando los datos ya han sido obtenidos y descargados, la primera información de valor que proporciona Analhitza para el discurso de Navidad del rey de España es la siguiente:

Nouns	347 tokens	215 types (61.96%)	23.88% of words
Adjectives	115 tokens	82 types (71.30%)	7.91% of words

Verbs	217 tokens	102 types (47.00%)	14.93% of words
Adverbs	64 tokens	27 types (42.19%)	4.40% of words
Determiners	267 tokens	19 types (7.12%)	18.38% of words
Conjunctions	118 tokens	12 types (10.17%)	8.12% of words
Prepositions	215 tokens	13 types (6.05%)	14.80% of words

Tabla 1. Tokens, tipos y porcentajes de palabras del discurso de Navidad del Rey (2022). Fuente: Elaboración propia en Analhitza.

Este informe expresa el número de tokens que hay en el texto en función de las diferentes categorías gramaticales. Como evidencia la tabla, la mayor cantidad está representada por sustantivos que abarcan casi el 70% de palabras del texto. El hecho de que se advierta, por ejemplo, de que de todos estos tokens solo hay 215 tipos diferentes de sustantivos ya implica que muchos de estos se repiten y que, por lo tanto, se encontrarán más adelante lematizados.

Tras esto, se obtuvieron listas desglosadas de palabras agrupadas por categorías, así como también secuencias de *n-grams* que forman estructuras y secuencias lingüísticas típicas. La tabla que se encuentra a continuación muestra un ejemplo de diferentes clases de palabras ordenadas y acompañadas del número de veces que aparecen en el discurso.

Nouns		**Adjectives**		**Verbs**	
11	españa	5	económico	28	ser
9	unión	5	social	19	haber

7	democracia	4	junto	10	estar
7	año	4	cotidiano	9	deber
6	europa	4	nuevo	8	tener
5	institución	3	difícil	8	poder
5	español	3	internacional	6	seguir
5	convivencia	2	político	5	necesitar
4	país	2	ucraniano	4	vivir

Tabla 2. Listado de palabras según frecuencia y categoría gramatical en el discurso de Navidad del Rey (2022). Fuente: Elaboración propia en Analhitza.

Estos datos proporcionan información sobre las palabras más frecuentes enmarcadas dentro de tres categorías gramaticales específicas. Se han seleccionado solamente las 10 más frecuentes de cada grupo, junto con el número de veces que aparecen en el discurso. Por ejemplo, la palabra *España* aparece 11 veces en la categoría sustantivo, *económico* 5 en la de adjetivo y *ser* 28 en la de verbo. Esta información puede ayudar a comprender mejor el contenido y la temática del discurso, ya que la frecuencia indica qué temas se enfatizaron más y cuáles fueron los términos clave utilizados para esto. Además, puede ser útil también para llevar a cabo un análisis más detallado del discurso en términos de su estructura gramatical y uso del lenguaje.

Por otra parte, se facilitan, como ya se ha anticipado, secuencias de n-grams como las adjuntadas a continuación. Estas son bastante útiles para ayudar a identificar patrones y tendencias en el uso del lenguaje, así como para extraer información relevante de un texto. También permiten identificar errores ortográficos y generar modelos de lenguaje para la traducción automática o la generación de texto.

3-grams						
3	económico	G	y	C	social	G
3	por	P	él	Q	,	O
3	,	O	en	P	este	D
3	,	O	que	Q	ser	V
3	el	D	unión	R	europea	R
3	de	P	el	D	unión	R
2	en	P	nuestro	D	país	N

Tabla 3. Secuencia de 3-grams en el discurso de Navidad del Rey (2022). Fuente: Elaboración propia en Analhitza.

Asimismo, también facilita información sobre el número de lemas, palabras y letras recogidos en el texto que pueden servir para establecer, por ejemplo, las redes léxicas y los campos semánticos que abarca este mismo. Para el caso de los lemas, por ejemplo, la herramienta no repite *nuestro* y *nuestra*, mientras que las palabras se recogen por separado. Esto es muy interesante según el tipo de análisis lingüístico que se quiera llevar a cabo.

No obstante, no solo es una información lingüística relevante, sino que de este exhaustivo análisis podemos obtener valiosa información sobre la temática de un texto. En el caso del texto seleccionado, los sustantivos que más ocurrían eran de carácter negativo (*problema, guerra, compromiso, Ucrania, preocupación…*), pero también esperanzador (*confianza, seguridad, compromiso, libertad*…). Esto a primera vista sitúa al analizador de los datos en el contexto socio-político y cultural del texto. En efecto, el 2022 fue un año complejo a nivel global. Lejos de la completa recuperación de la pandemia, tanto España como el resto de los países se vieron también afectados por la guerra de Ucrania. Así, a través de dicha herramienta no solo podemos analizar cuestiones

semánticas, pragmáticas y morfológicas, sino dar cuenta de cómo el contexto influye en el lenguaje y viceversa.

2.4. Alcances y limitaciones de Analhitza

Después de haber analizado el texto con Analhitza, se han podido constatar una serie de alcances y limitaciones. En primer lugar, los aspectos positivos de sus alcances pueden ser resumidos en tres atributos fundamentales: facilidad de uso, lo que la convierte en accesible a usuarios con diferentes niveles de experiencia en procesamiento del lenguaje natural; eficiencia, puesto que es rápida y competente, capaz de procesar grandes cantidades de texto en cuestión de segundos y facilitar datos con agrupaciones de elementos según sus características lingüísticas; y, por último, versatilidad, ya que la gama de funcionalidades es muy amplia y hace que su empleo pueda abarcar diferentes tareas dentro del análisis de un texto.

En contraposición a esto, también se han podido observar una serie de aspectos que podrían tener un rango de mejora. Entre ellos destaca la precisión, puesto que, aunque es una herramienta muy útil, no siempre es capaz de identificar de forma correcta las partes del discurso o las relaciones entre ellas. Por ejemplo, en las secuencias de n-grams se incluyen signos de puntuación que, tal vez, no deberían aparecer o, por otra parte, en cuanto a los *named_entities* se refiere, se ha podido constatar que no hace distinción entre singulares y plurales, de modo que no categoriza estos por lemas.

Asimismo, Analhitza ha sido diseñada para trabajar con tres idiomas (euskera, español e inglés) por lo que quedan restringidos el resto de documentos en otras lenguas. En lo que atañe al contexto, se basa en modelos estadísticos, hecho que implica que su exactitud pueda depender, en gran medida, del contexto en el que se utiliza. Por último, se ha considerado muy importante resaltar que no hay posibilidad de filtrar contenido, como en el caso de las palabras vacías que muchas veces es

importante eliminar de listados de frecuencias donde se pretenda resaltar el contenido semántico de las palabras.

2.5. Otras propuestas de ejercicios

En esta ocasión se ha llevado a cabo un trabajo que analiza el discurso político, para ver la frecuencia de las palabras más utilizadas, las categorías gramaticales predominantes y los n-grams más frecuentes, entre otros aspectos, para comprender mejor la retórica, la temática y el mensaje transmitido. Sin embargo, no es el único ejercicio que se puede realizar[7].

También podría ser interesante comparar diferentes textos literarios de un autor en particular a partir del análisis de cada uno de los textos por esta herramienta. De esta manera, podría obtenerse información muy valiosa acerca de la utilización del vocabulario del escritor de las obras, el estilo y la estructura narrativa.

Por otra parte, asimismo podría enfocarse a una investigación dentro del campo de las redes sociales, donde se pueden analizar los hashtags más utilizados, extraer de ellos las palabras clave que representarán las frecuencias más elevadas, observar los usuarios más influyentes y deducir las tendencias generales de las conversaciones en línea.

Por último, sería muy cautivador a través de esta plataforma examinar las opiniones de los consumidores. ¿De qué manera? Pues bien, utilizando los comentarios que estos mismos escriben en línea. De ellos, podrían extraerse las palabras y categorías gramaticales más utilizadas, lo que puede proporcionar información provechosa sobre las diferentes experiencias de los consumidores con un producto o un servicio en particular.

Estos son solo algunos de los ejemplos de cómo se puede utilizar Analhitza para analizar diferentes tipos de textos. Las posibilidades son

[7] De Fradejas Rueda (2023) hemos aprendido, utilizando además un corpus similar, como realizar estos procesos con R.

casi ilimitadas y dependen de la creatividad y los objetivos de la persona que realiza el estudio.

3. ¿ES COMPATIBLE ANALHITZA CON OTRAS APLICACIONES? ¿CUÁLES?

Analhitza, al ser una herramienta de PLN, se puede integrar con diferentes aplicaciones de visualización de datos para estudiar, interpretar y plasmar los datos de forma visible. Algunas de las que podrían ser compatibles con esta son: Voyant Tools[8], R Studio[9] y Tableau[10].

En el caso de la primera, se trata de una plataforma de visualización de datos lingüísticos en línea que permite explorar y analizar colecciones de textos digitales. Al combinar Voyant Tools con Analhitza se podría utilizar esta última para trabajar con los textos y Voyant Tools para mostrar los resultados. Por ejemplo, si se analizaran no uno, sino todos los discursos de Navidad del Rey de España, podrían cargarse los resultados en Voyant Tools para mostrar las tendencias lingüísticas y los patrones de frecuencia de palabras dentro de la batería de textos, con herramientas disponibles dentro de esta como Cirrus.

R Studio es otra de las alternativas interesantes para combinar con la aplicación analizada en dicha reseña. Es una herramienta de visualización de datos que se utiliza para analizar y visualizar datos en una amplia gama de disciplinas. Al combinar R Studio con Analhitza, se pueden analizar datos lingüísticos utilizando Analhitza y luego visualizar los resultados utilizando los paquetes de visualización de R Studio. Por ejemplo, después de analizar los datos de un chatbot (asistente conversacional) con Analhitza, se pueden cargar los resultados en R Studio y utilizar paquetes como ggplot2 para visualizar las tendencias de conversación y la interacción entre el chatbot y los usuarios.

Como última aplicación destaca Tableau, que se utiliza para crear visualizaciones interactivas y paneles de control para analizar datos en

[8] Accesible desde: https://voyant-tools.org/.
[9] Sitio de RStudio: https://www.r-project.org/.
[10] Accesible desde: https://www.tableau.com/.

tiempo real. Al combinar esta con Analhitza, se pueden ver los resultados de análisis lingüísticos en tiempo real utilizando las capacidades de visualización de dicha aplicación. Por ejemplo, se puede integrar Analhitza en una plataforma de atención al cliente, y utilizar Tableau para visualizar las tendencias de conversación y el rendimiento del equipo de soporte en tiempo real.

4. CONCLUSIONES

En conclusión, Analhitza es una herramienta valiosa para el PLN y el análisis de textos en diversos campos. A través de su capacidad para realizar análisis de frecuencia de palabras, análisis de sentimientos y análisis de temas, los usuarios pueden obtener información desglosada y detallada sobre los textos que desean analizar. La herramienta ofrece una amplia gama de características y opciones de personalización que permiten a los usuarios adaptar su análisis para cumplir con sus objetivos de investigación o negocio específicos.

Además, Analhitza se puede integrar con otras aplicaciones y herramientas de visualización de datos, lo que facilita la creación de gráficos e informes detallados y fáciles de entender sobre los resultados del análisis. También se puede utilizar en una variedad de contextos, incluyendo análisis de redes sociales, análisis de opinión pública, análisis de chatbots y análisis de comentarios de clientes, entre otros.

En cuanto a las desventajas de Analhitza, es importante tener en cuenta que el análisis de texto automatizado no es perfecto y puede haber errores en la interpretación del significado o contexto de las palabras. Además, la precisión del análisis puede verse afectada por la calidad del texto analizado, incluyendo la ortografía, la gramática y el estilo.

Otra posible desventaja es que Analhitza puede ser limitado en términos de idiomas compatibles y conjuntos de datos. Aunque puede ser utilizado en varios idiomas, el análisis y la precisión del análisis pueden variar según el idioma y el conjunto de datos. Además, el análisis de

grandes conjuntos de datos puede requerir mucho tiempo y recursos informáticos, lo que puede limitar su uso en algunos contextos.

Sin embargo, a pesar de estas limitaciones, Analhitza sigue siendo una herramienta útil y valiosa para el análisis de texto y el procesamiento del lenguaje natural. Al aprovechar sus capacidades y limitaciones, los usuarios pueden obtener información valiosa y útil que les ayudará a tomar decisiones más informadas y mejorar sus operaciones en diversas áreas.

REFERENCIAS BIBLIOGRÁFICAS

Agerri, R., Bermudez, J. y Rigau, G. (2014). IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. En N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk y S. Piperidis (Eds.), Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) (pp. 3823-3828). European Language Resources Association (ELRA)

Alonso Sáenz de Oger, S., Volkens, A. y Gómez Fortes, B. (2012). Content-analyzing political texts: A quantitative approach. Centro de Investigaciones Sociológicas (CIS).

Fokkens, A., Etxabe, A. S., Beloki, Z., Ockeloen, C., Rigau, G., van Hage, W. R. y Vossen, P. (2014). NAF and GAF: Linking linguistic annotations [Conferencia]. *10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*. Portorož, Slovenia.

Fradejas Rueda, J.M. (2023). Cuentapalabras. Estilometría y análisis de texto con R para filólogos. Universidad de Valladolid. https://aic.uva.es/cuentapalabras/

Otegi, A., Imaz, O., Díaz de Ilarraza, A., Iruskieta, M. y Uria, L. (2024). *ANALHITZA: A tool to extract linguistic information from large corpora in Humanities research* [Software]. University of the Basque Country. http://ixa.si.ehu.es/node/8862

Ruiz Fabo, P. y Bermúdez Sabel, H. (2019). Navegación de corpus a través de anotaciones lingüísticas automáticas obtenidas por Procesamiento del Lenguaje Natural: De anecdótico a ecdótico. *Revista de Humanidades Digitales*, 4, 136-161. https://doi.org/10.5944/rhd.vol.4.2019.25186