ÑEMITỸRÃ, Vol. 5, Núm. 2 – AGOSTO 2023
Revista Multilingüe de Lengua, Sociedad y Educación

ISSN (en línea): 2707-1642
ISSN (impreso): 2707-1634

Articulo original

El análisis lexicométrico, el corpus y el diccionario previo: caso de la lengua vasca[1]

Lexicometric analysis, the corpus, and the previous dictionary: the case of the Basque language

Juan Abasolo
Universidad del País Vasco (UPV/EHU), España
https://orcid.org/0000-0002-1911-4118

e-mail: juan.abasolo@ehu.eus

Recibido: 18/3/2023
Aprobado: 12/5/2023

RESUMEN
Siguiendo los lineamientos del método propuesto por Reinert (1983, 1990), podemos llevar a cabo una clasificación automática de grandes volúmenes de textos que aborden un determinado contexto semántico, con el fin de identificar los distintos campos semánticos o de sentido en base a las elecciones léxicas realizadas. Esta tarea se encuentra resuelta tanto en el software patrocinado por Max Reinert (1990), conocido como ALCESTE, como en el software de código abierto IRaMuTeQ (Ratinaud & Déjean, 2009) y sus desarrollos posteriores (Barnier, 2022). Como mencionaba Reinert (1990), es fundamental llevar a cabo un pretratamiento lingüístico que se ajuste al sentido semántico del texto, evitando así la variabilidad regida por las normas gramaticales. Para lograr este control sobre la variabilidad, se utiliza una preclasificación con un diccionario que contempla las formas plausibles de aparición en el texto, junto con sus correspondientes valores de significado asociados. En el caso de lenguas aglutinantes como el euskera, esto plantea un desafío particular. En esta comunicación, presentamos el proceso de creación de un diccionario para el uso del software IRaMuTeQ con textos en lengua vasca, su evaluación interna, mediante el análisis de autodescripciones de profesorado universitario, y externa, mediante el análisis de un corpus paralelo multilingüe vasco, castellano, inglés y francés.

Palabras clave: Método Reinert; Iramuteq; léxico; NPL; corpus multilingue.

ABSTRACT
Following the steps of the method described by Reinert (1983, 1990), we can perform an automatic classification of large numbers of texts that address a specific semantic context in order to identify different semantic fields or meanings based on lexical choices. This task is already solved in the software sponsored by Reinert himself, known as ALCESTE, as well as in the open-source software IRaMuTeQ (Ratinaud & Déjean, 2009) and subsequent developments based on it (Barnier, 2022). As Reinert (1990) anticipated, a linguistic pre-processing is necessary to adhere to the proper semantic sense of the text, free from the variability governed by the rules of grammar. This control over variability is based on the use of a dictionary that includes plausible forms of appearance in the text along with their associated meaning values. In agglutinative languages such as Basque, this poses a challenge. In this communication, we present the process of creating a dictionary for the use of the IRaMuTeQ software with texts in the Basque language, along with an analysis of a parallel multilingual corpus.

Keywords: Reinert method; Iramuteq; dictionary; NPL; multilingual corpus.

[1] Este trabajo se ha llevado a cabo con la financiación del proyecto GIU21/016 financiado por la UPV/EHU.

INTRODUCCIÓN

En este estudio, deseamos presentar de manera concisa la adaptación que un grupo de investigadores del equipo EUDIA está llevando a cabo para hacer factible utilizar Iramuteq en lengua vasca. Iramuteq es una herramienta utilizada para el análisis de textos y cuestionarios (Ratinaud & Déjean, 2009).

Nuestra intención es compartir y transparentar el proceso de creación de un léxico adaptado al uso en esta lengua, así como brindar una visión más clara de los aspectos teóricos aplicados en el análisis de textos. Además, también queremos mostrar algunos resultados obtenidos mediante este enfoque de trabajo. Creemos que este estudio puede resultar de interés en otros entornos bilingües, especialmente aquellos con lenguas de naturaleza aglutinante.

LEXICOMETRÍA

Dentro del campo de la lingüística de corpus, el avance en la computación, el desarrollo de técnicas estadísticas más complejas y la capacidad de almacenamiento y procesamiento de grandes volúmenes de datos han permitido un rápido desarrollo de nuevas técnicas, entre las cuales se encuentran la lexicometría y la textometría.

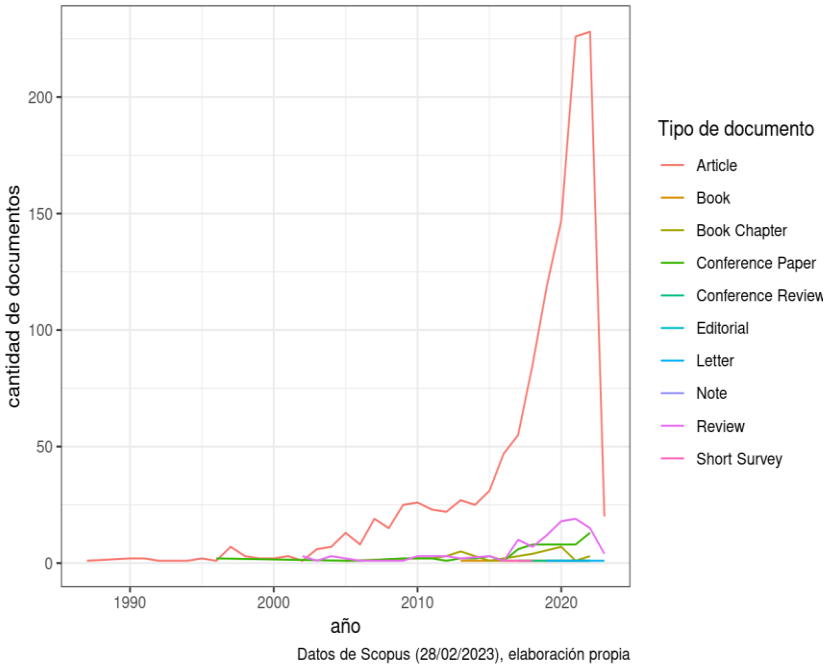
Se podría señalar como el primer paso en las técnicas que hoy se utilizan en la lingüística de corpus fue, en el siglo XIII, la indexación de la Biblia *Sancta Vulgata*, realizada a por Hugo de San Caro y 500 monjes bajo su dirección en su obra *Correctio Biblie*[†] (Hanon, 1991). Sin embargo, probablemente no fue hasta el desarrollo de la semántica interpretativa en los años 80 que se puede hablar de un desarrollo amplio en esta perspectiva. Es posible que el trabajo de Françoys Rastier (1987) sea considerado el punto de partida de este desarrollo, al proporcionar técnicas y herramientas computacionales para la modelación del discurso.

La lexicometría, también conocida como análisis léxico cuantitativo, tiene como objetivo identificar unidades y patrones temáticos en los corpus utilizando enfoques cuantitativos. Una característica distintiva de las metodologías lexicométricas es que, si bien los análisis de corpus se basan en técnicas cuantitativas, los resultados de estos análisis no pueden ser interpretados sin una subsiguiente interpretación cualitativa que les otorgue sentido. Esto se debe a que las técnicas lexicométricas se centran en la búsqueda de patrones y coocurrencias en el corpus analizado. Una vez que se identifica la relevancia estadística de ciertas características, es fundamental que los elementos reconocidos y señalados desde el punto de vista matemático y estadístico sean reinterpretados y dotados de significado.

Entre las diversas técnicas de análisis que permite la lexicometría, posiblemente la perspectiva presentada por Reinert en 1983 (Reinert, 1983) sea la más relevante desde la experiencia académica. Esto se debe, entre otras cosas, al encuentro entre las técnicas que evalúan el grado de significancia estadística de los constructos estudiados y la interpretación cualitativa, lo cual permite identificar y reconstruir significados sociales basados en la práctica textual. Este enfoque ha sido ampliamente valorado por las editoriales científicas, como se refleja en la figura 1, que muestra la publicación de documentos científicos utilizando esta técnica. Es de gran relevancia la amplitud de editores que le dan importancia desde las políticas monetarias (Schonhardt-Bailey & Bailey, 2013) hasta la salud o la educación.

[†] Se puede consultar online la impresión de 1498: *Inkunabeln / Biblia latina: Cum postillis Hugonis de Sancto Caro/: 1. [Basel]. [nach 29. X. 1498 u. nicht nach 1499]*. (1498). http://digital.ub.uni-duesseldorf.de/ink/7856461

Figura 1. Resultados por años de la búsqueda de los descriptores "Alceste", "Reinert" e "Iramuteq" en documentos de la base de datos Scopus. Agregados según tipo de documentos.



Elaboración con base en el trabajo de investigación.

MÉTODO REINERT

A principios de la década de 1980, Reinert (Reinert, 1983) introdujo un algoritmo revolucionario que permitía clasificar los textos según su contexto, mediante un análisis multidimensional. Inicialmente, se presentó en el campo de la psicología social como una herramienta para categorizar las respuestas libres a preguntas abiertas, con el propósito de establecer análisis reproducibles. Tres años después, se presentó la primera implementación informática del algoritmo propuesto, conocida como el programa *ALCESTE - Analyse Lexicale par Contexte d'un Ensemble de Segments de Texte* (Reinert, 1986), junto con su flujo de trabajo, que se explicó en detalle en un artículo posterior en el ámbito de las ciencias sociales (Reinert, 1990).

Aunque no pretendemos abordar todos los aspectos técnicos de esta técnica en este artículo, nuestro objetivo es presentar una reconstrucción de la teoría que sustenta su enfoque técnico. Reinert, en colaboración con Benzecri (1981), propuso el concepto de que el contexto de cada palabra en un texto está formado por el texto completo, excluyendo la palabra en sí misma, tal como se establece originalmente en el concepto "con-texto". Desde una perspectiva teórica, esto permitiría modelar los contextos de las palabras en función de las palabras que las rodean. Para lograr esto, la técnica denominada ALCESTE por Reinert propone describir los contextos mediante la coocurrencia de los términos contextuales.

Una vez que se clasifican las unidades de análisis de contexto, que idealmente deberían coincidir con unidades de sentido, se agrupan utilizando una metodología jerárquica ascendente de conglomerados (cluster). Esta agrupación permite identificar elementos con patrones regulares de aparición y calcular la presencia o ausencia específica de cada uno de los términos en todas las agrupaciones definidas.

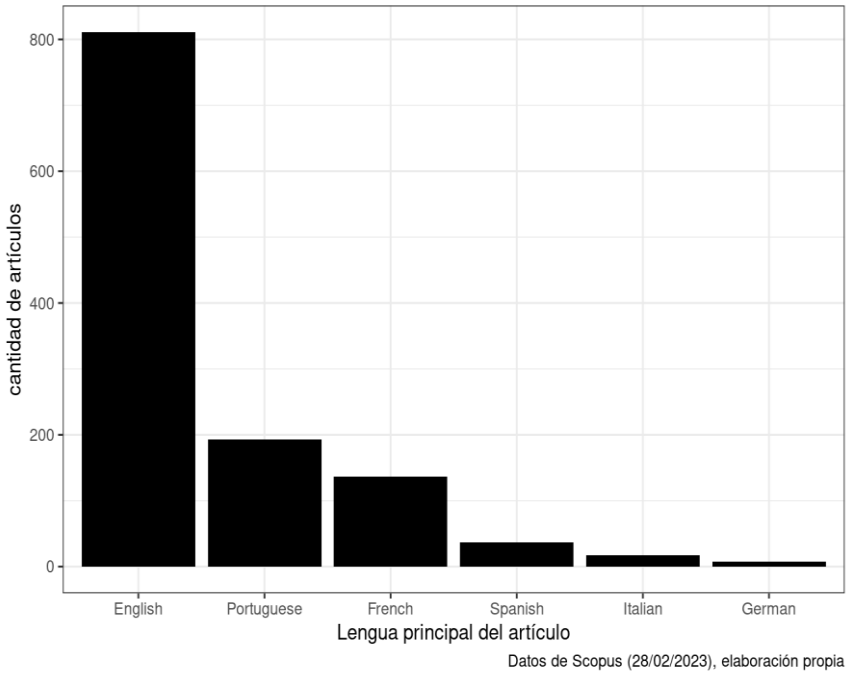
Sin embargo, esta propuesta teórica se enfrenta al desafío del uso real del lenguaje, en el que los elementos conceptuales subyacentes a las formas lingüísticas pueden ser únicos o polimórficos debido a las diversas formas que pueden tomar los lemas. Esta dificultad se agrava especialmente en lenguas aglutinantes, como el euskera.

La solución propuesta por Reinert, que actualmente se está aplicando al euskera, implica contar previamente con una lista ordenada de formas lingüísticas plausibles de aparecer, etiquetadas con sus lemas correspondientes y su valor sintáctico, antes de realizar el análisis. Esto facilita el proceso de asignar sentido a los elementos identificados y señalados mediante las técnicas lexicométricas utilizadas en el análisis.

IRAMUTEQ

Iramuteq (Ratinaud, 2014; Ratinaud & Déjean, 2009) es una adaptación del algoritmo propuesto por Reinert (1983), esta vez utilizando el lenguaje R (Ihaka & Gentleman, 1996), con una interface utilizable por el usuario final. Tanto el código como la aplicación de Iramuteq se distribuyen de forma abierta y gratuita. De hecho, Iramuteq se presenta como una interfaz de R para el análisis multidimensional de textos y cuestionarios (*Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires*), como se indica en el sitio web de descarga‡. Además del análisis multidimensional, Iramuteq también permite realizar otros análisis del corpus tanto multidimensionales como unidimensionales, tales como análisis factorial, nubes de palabras, búsquedas de concordancias y análisis de similitudes. Estos análisis adicionales no serán abordados en detalle en el presente trabajo, pero pueden resultar de interés para los investigadores.

Figura 2. Lenguas de edición de artículos en la base de datos Scopus, que incluyen los descriptores "Iramuteq", "Reinerto" o "Alceste".



Elaboración con base en el trabajo de investigación.

‡ http://www.iramuteq.org

Como se ha destacado en la figura 2, hasta la fecha se ha demostrado la valía de esta herramienta. En la base de datos Scopus se puede observar la relevancia de las publicaciones en varios idiomas, más allá del francés e inglés, que eran los esperados. Al revisar las bases de datos de publicaciones científicas, también se puede observar que en muchos casos las investigaciones mencionan el método Reinert (o Alceste, o Iramuteq), pero el idioma utilizado en la comunicación de la investigación no coincide con el idioma en el que se desarrolló el método. También es importante destacar la notable presencia de publicaciones científicas en Colombia, Brasil o México, además de Estados Unidos, Francia, España o Portugal, que consideran el uso de esta técnica, método y herramienta para describir diferentes representaciones sociales.

LENGUAS

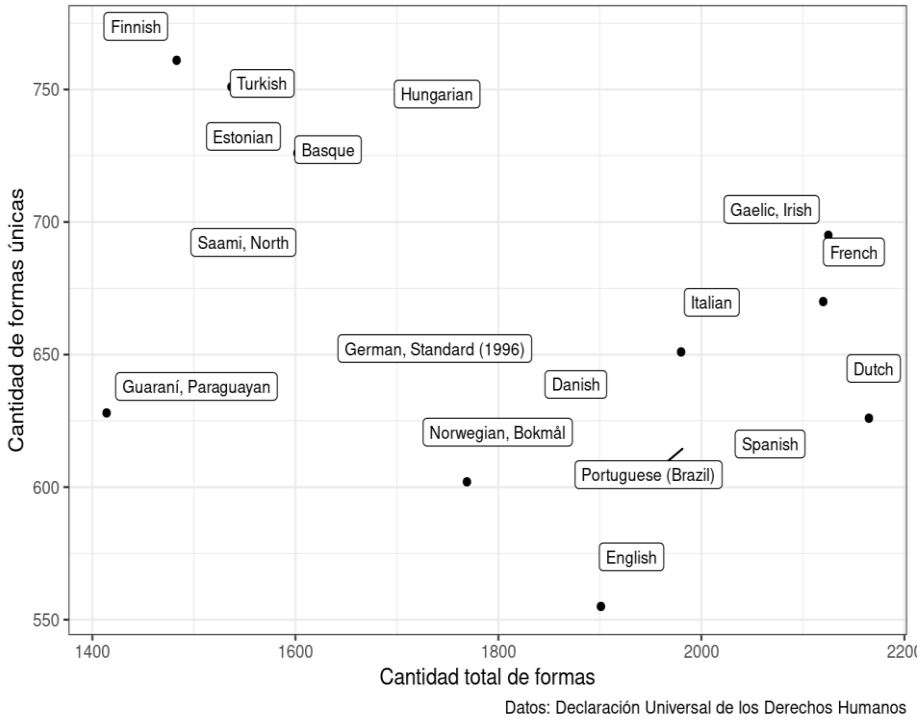
La herramienta Iramuteq, como se mencionó anteriormente, requiere de un léxico ordenado que contenga las formas lingüísticas esperadas en los textos que van a ser analizados. Es relevante destacar en qué idiomas se está utilizando la herramienta y cuáles son sus características diferenciales.

Actualmente, la distribución de Iramuteq viene con algún léxico en alemán, inglés, francés, gallego, griego, italiano, latín, portugués, rumano, español y sueco. De todos los léxicos disponibles, solo algunos se pueden acceder automáticamente desde la interfaz del programa y tienen el tamaño suficiente como para conseguir análisis efectivos. A aquellos léxicos que no ofrecen una opción de selección automática se puede acceder mediante un cuadro de diálogo, que permite asignar un archivo en particular como diccionario.

Esta posibilidad mencionada anteriormente es la que permite y justifica la existencia de numerosos léxicos preparados para su uso en diferentes idiomas y con una mayor cantidad de términos incluidos, además de los léxicos de distribución oficial mencionados anteriormente. Por ejemplo, existe un léxico en lengua castellana distribuido a través de un enlace de Twitter (Ideia [@ideiainova], 2017), o un léxico en alemán distribuido en listas de correo públicas (Loubere, 2023) propias del grupo que desarrolla el programa.

Los léxicos preparados para el método Alceste se encuentran en su mayoría en lenguas indoeuropeas y son principalmente, a excepción del latín, prepositivas. En este trabajo, estamos presentando la adaptación del método a la lengua vasca, que es una lengua no indoeuropea y se caracteriza por ser altamente aglutinante. Esto se traduce en una mayor variedad de formas escritas diferentes y una menor cantidad total de formas para la comunicación de textos equivalentes, como se puede observar en la figura 3, que se representa la cantidad de formas necesarias para la traducción oficial de la Declaración Universal de los Derechos Humanos de la UNESCO, contrastando la cantidad total de formas con la cantidad de formas diferentes.

Figura 3. Relación entre formas diferentes y total de formas empleadas en algunas lenguas.



Elaboración con base en el trabajo de investigación.

EL VASCO

Específicamente, el caso del vasco escrito se destaca como el más divergente en comparación con otras lenguas europeas en cuanto a la cuantificación de la cantidad de formas distintas y totales utilizadas en textos equivalentes. Este fenómeno se atribuye a la presencia de una morfología aglutinante que rige la producción de textos en esta lengua. Para ilustrar esta situación, podemos contrastar alguna de las diversas formas que pueden adoptar los lemas equivalentes "casa" en castellano y "etxe" en vasco:

- casa (casa singular o indefinido)
- casas (casa plural)

En vasco:

- etxe [*casa*]
- etxea [*la casa* (absolutivo)]
- etxeak [*la casa* (ergativo)]
- etxean [*en (la) casa*]
- etxera [*a (la) casa*]
- etxetik [*desde (la) casa*]
- etxerako [*para (la) casa*]
- etxeraino [*hasta (la) casa*]
- etxeko [*de (la) casa*]
- …

Esta lista, que actualmente solo contempla las formas singulares, podría ampliarse para incluir otras formas de uso común, como plurales o indefinidos, además de la amplia variedad de combinaciones de sufijos que son posibles en el idioma, como, por ejemplo.

- etxekoa [*el de (la) casa* (absolutivo)]
- etxekoak [*el de (la) casa* (ergativo)]
- etxekoaren [*del de (la) casa*]
- etxekoarena [*el del de (la) casa* (absolutivo)]
- …

En aras del principio de parsimonia y teniendo en cuenta la capacidad limitada de las herramientas que se utilizarán, no es apropiado trabajar con la lista de todas las formas posibles, sino más bien con aquellas formas que sean plausibles de encontrarse en los diversos textos escritos en vasco.

OBJETIVOS Y MÉTODO

El objetivo de este estudio es desarrollar y evaluar un léxico en lengua vasca para analizar textos escritos en dicho idioma utilizando la metodología propuesta por Reinert. En esta sección, se describen los métodos utilizados para la construcción de los corpus y los procedimientos aplicados posteriormente para su procesamiento.

PRIMER OBJETIVO: CONSTRUCCIÓN DEL LÉXICO EN LENGUA VASCA

El primer objetivo de este estudio fue crear un léxico en lengua vasca. Para lograrlo, se siguieron varios pasos. En primer lugar, se construyó un corpus extenso que reflejara el uso del idioma. Luego, se llevó a cabo un análisis del corpus para identificar las relaciones entre las formas utilizadas, los lemas a los que hacen referencia y las categorías gramaticales en las que se clasifican. Por último, se redujeron estas relaciones encontradas a un conjunto único, basándose en la frecuencia de uso.

Figura 4. Captura de algunas líneas del léxico en español.

Elaboración con base en el trabajo de investigación.

En la imagen de la figura 4 se muestra la estructura requerida para el léxico, con una columna inicial que incluye las formas o tokens, incluyendo las flexiones de los lemas. La segunda columna muestra los lemas correspondientes, mientras que en la tercera columna se proporciona una abreviatura en francés que indica la categoría gramatical asignada.

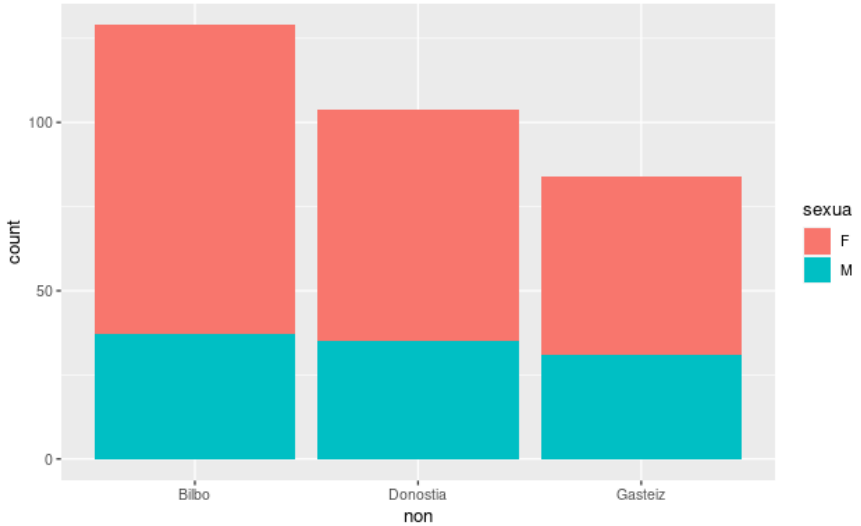
El corpus utilizado en el análisis consistió en artículos y cartas de lectores de una revista de temática general, Argia, en formato digital, que comprendía 1.082.787 frases. Además, se incluyeron 61.567 frases de la selección de artículos en lengua vasca de Wikipedia realizada por el proyecto Common Voice, y 11.379 frases de las publicaciones del año 2020 de la revista en línea Bizkaie escrita en dialecto occidental. En total, se obtuvieron más de un millón de frases en diferentes registros y formas para su análisis.

Para llevar a cabo el análisis y clasificación de las formas lingüísticas, se utilizó el lenguaje de programación R (R Core Team, 2020). Específicamente, se empleó el paquete "udpipe" (Straka et al., 2016) (Wijffels et al., 2020) en su versión 0.8.5, el cual proporciona una interfaz adaptable para aplicar modelos de Universal Dependencies (Nivre et al., 2016). Se accedió al modelo de euskera publicado por Aranzabe y colaboradores (Aranzabe et al., 2015) para llevar a cabo el modelado y clasificación de los lemas y las categorías gramaticales.

SEGUNDO OBJETIVO:
PROBAR LA CONSISTENCIA INTRALINGUAL DEL LÉXICO CONSTRUIDO

Con el fin de evaluar la eficacia del léxico construido, se llevaron a cabo dos enfoques de prueba. En primer lugar, se realizó una evaluación desde la perspectiva de la propia lengua vasca, con el objetivo de comprobar la coherencia de las respuestas obtenidas. Para este propósito, se construyó un corpus de autodescripciones redactadas por profesores de las carreras de magisterio de la UPV/EHU. Estos textos públicos, redactados en lengua vasca, se obtuvieron de las páginas web de las tres facultades donde se imparten las carreras de magisterio en educación infantil o educación primaria. Estos textos fueron procesados con Iramuteq, utilizando el léxico en lengua vasca, para realizar una clasificación jerárquica ascendente de acuerdo con el método ALCESTE propuesto por Reinert.

Figura 5. Distribución del corpus de autodesripciones según campus de pertenencia y género de los autores y autoras.



Elaboración con base en el trabajo de investigación.

El corpus mencionado anteriormente (figura 5), consta de un total de 129 autodescripciones escritas en lengua vasca, de un total de 317 docentes pertenecientes a las tres facultades donde se imparten los grados de magisterio, recopiladas en noviembre de 2021. Esto permitió obtener un total de 1279 pares de unidades de sentido y etiquetas, clasificados según el campus, el departamento al que pertenecen los docentes y su género.

TERCER OBJETIVO:
VERIFICACIÓN DE LA CONSISTENCIA INTERLINGÜÍSTICA DEL LÉXICO CONSTRUIDO

Por otro lado, con el propósito de comprobar la coherencia de los resultados obtenidos en lengua vasca, se construyó un corpus paralelo multilingüe para comparar los resultados en vasco con aquellos en otros idiomas con una mayor tradición en el uso de Iramuteq, como el francés, inglés y castellano. En este caso, el corpus se basa en las cartas paulinas del Nuevo Testamento, segmentadas de manera idéntica, lematizadas utilizando los léxicos proporcionados por Iramuteq en cada lengua y utilizando el léxico específico para los textos en vasco. Estos subcorpus fueron agrupados según el método ALCESTE.

La consistencia de los campos semánticos detectados en diferentes idiomas se verificó en función de las unidades contextuales utilizadas en su construcción. Para ello, se creó una matriz de similitud de los agrupamientos según el método de Jacard y se generó una clasificación jerárquica ascendente.

Para la construcción del corpus paralelo multilingüe, se utilizaron los datos disponibles en el sitio web de la Asociación Bíblica Eslovena[§], que permite la presentación emparejada de diferentes traducciones. En este sentido, se incluyó la única traducción disponible en euskera actual, la edición ecuménica Elizen Arteko Biblia de 1994; la versión en inglés de King James de 1611, la versión francesa de la traducción de Louis Segond de 1910, y la versión en castellano ecuménica de la traducción Dios Habla Hoy de 2002. Estas traducciones se utilizaron como base para el corpus multilingüe y permitieron realizar comparaciones significativas entre las diferentes lenguas.

RESULTADOS

En esta sección, se presentan los resultados obtenidos durante la construcción del léxico, así como los contrastes intra e interlingüísticos del rendimiento del mismo.

CONSTRUCCIÓN DEL LÉXICO

El análisis inicial del corpus reveló un total de 13.065.741 combinaciones diferentes de *forma-lema-categoría gramatical*. No obstante, como se mencionó anteriormente, es necesario reducir esta cantidad a relaciones únicas entre formas y lemas, así como entre lemas y categorías gramaticales. Este proceso de reducción se basa en la frecuencia de uso, priorizando las relaciones más frecuentes. Después de aplicar esta reducción, se logró crear un léxico con 627.479 formas diferentes.

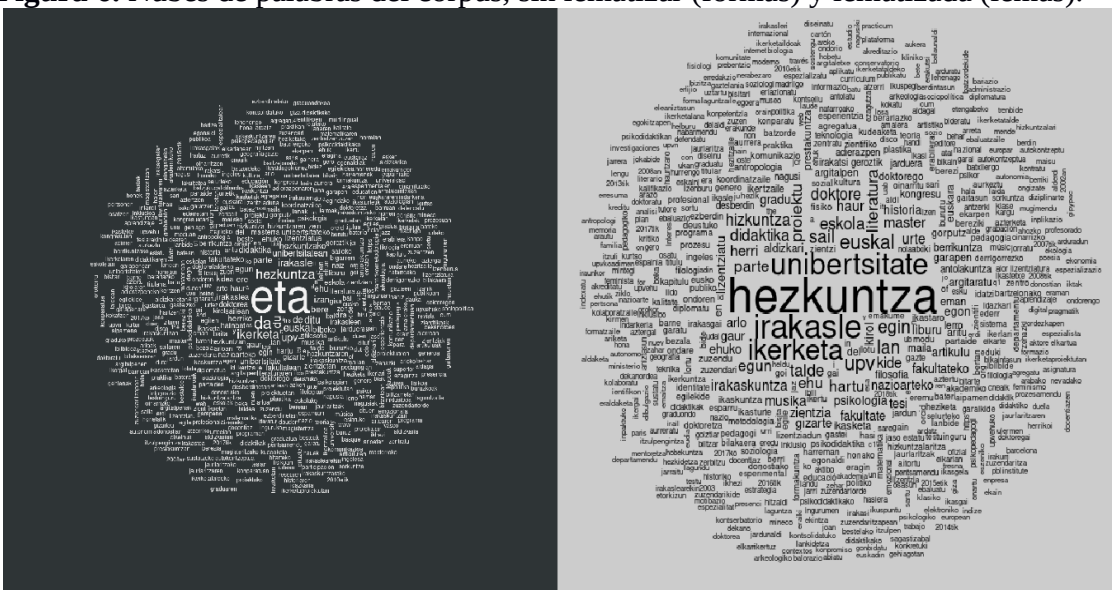
A diferencia de otros léxicos construidos para lenguas prepositivas, debido a las características aglutinantes del vasco, se incluyeron en este léxico un total de 20.093 formas clasificadas como nombres propios. Estas formas fueron etiquetadas como *pro_per*, aunque originalmente esa etiqueta pertenece a la categoría de pronombres personales del léxico francés. En este léxico en lengua vasca, no se realizó una distinción entre estos pronombres y otros, siguiendo las decisiones adoptadas en la construcción de los léxicos para castellano, inglés, portugués e italiano.

§ https://biblija.net

USO EN EUSKERA

Un primer análisis del uso del léxico en lengua vasca confirma que el filtrado por categorías gramaticales y la lematización del corpus generan resultados con evidentes diferencias.

Figura 6. Nubes de palabras del corpus, sin lematizar (formas) y lematizada (lemas).

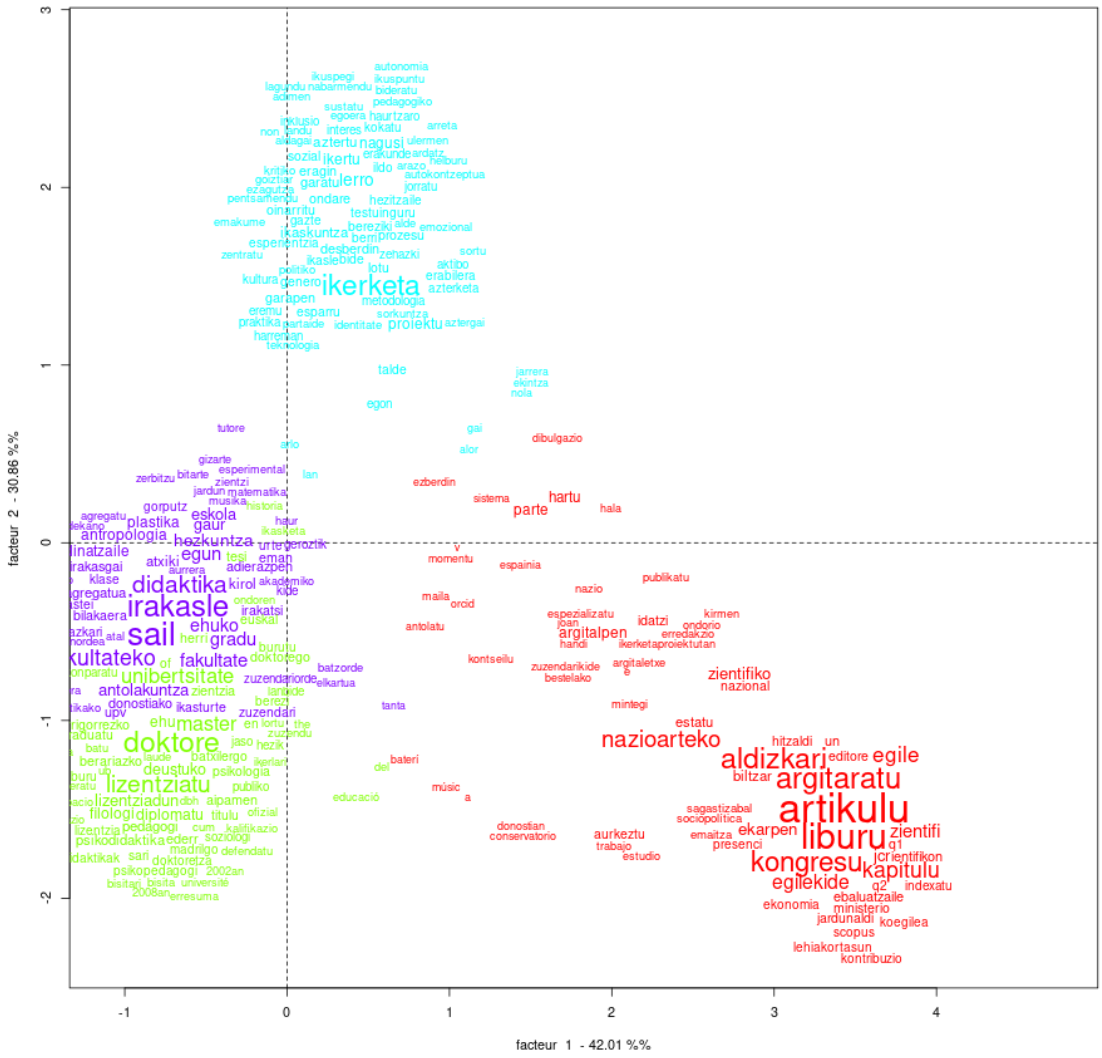


Elaboración con base en el trabajo de investigación.

En la parte izquierda de la figura 6, se muestra la nube de palabras basada en la frecuencia de uso de las formas, mientras que en la parte derecha se presenta la nube de palabras basada en la frecuencia del uso de los lemas, según lo establecido en el léxico. La forma de conjunción “eta” [*y*] destaca en términos de relevancia, en contraposición al lema “hezkuntza” [*educación*]. El lema “hezkuntza” también aparece en la nube de formas y se utiliza en un total de 487 ocasiones, en doce formas diferentes.

Al realizar un Análisis de Clasificación Descendente basado en los contextos léxicos, se puede observar la siguiente proyección (figura 7), donde se identifican cuatro campos semánticos, dos de los cuales están estrechamente relacionados.

Figura 7. Proyección de la dispersión de los lemas de cuatro clases.



Elaboración con base en el trabajo de investigación.

En color rojo se identifica el campo de las publicaciones y las comunicaciones, con especial relevancia de los lemas “artikulu” [*artículo*], “argitaratu” [*publicar*], “liburu” [*libro*], “kongresu” [*congreso*], “aldizkari” [*revista*], entre otros.

El segundo ámbito identificado, marcado en color celeste, abarca la actividad de investigación científica, en la que destacan los lemas “ikerketa” [*investigación*], “metodologia” [*metodología*], “lerro” [*línea*], “proiektu” [*proyecto*], entre otros.

Se observa una interconexión, pero también una clara diferenciación entre los dos últimos ámbitos, relacionados con la autodescripción del profesorado. El ámbito de la actividad docente se representa en color violeta, mientras que el ámbito del recorrido académico se representa en color verde. Los lemas “irakasle” [*maestro*], “didaktika” [*didáctica*], “sail” [*departamento*], “atxiki” [*adjunto*], “eskola” [*escuela*], “hezkuntza” [*educación*], “plastika” [*plástica*], entre otros, se refieren principalmente al ámbito de la acción docente y están relacionados con el recorrido académico, representado por lemas como “doktore” [*doctor*], “lizentziatu” [*licenciado*], “diplomatu” [*diplomado*], “master” y otros.

COMPARACIÓN PARALELA

El análisis de los textos de las cartas paulinas revela un total de cuatro clasificaciones jerárquicas distintas, con diferentes cantidades de campos detectados. Estas cuatro clasificaciones primarias dieron lugar a agrupaciones basadas en similitudes resumidas en la tabla1.

Tabla 1. Resumen de los análisis jerárquicos descendientes de los cuatro subcorpus.

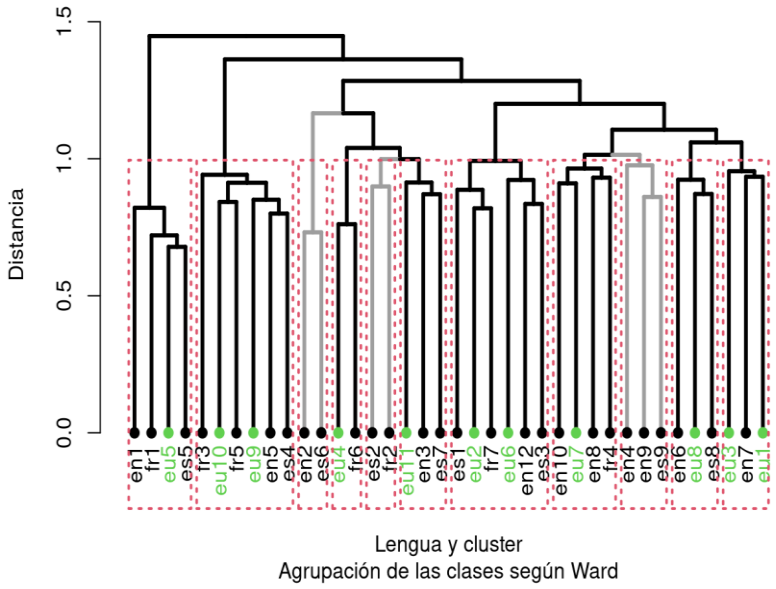
Traducción	Léxico	Ratio clasificado	clases
SEG	francés	0.9751	7
KJB	inglés	0.952	12
DHH	español	0.950	9
EAB	vasco	0.942	11

Elaboración con base en el trabajo de investigación.

El proceso de agrupamiento ascendente se llevó a cabo según la coincidencia en los pasajes asignados a cada clase, siguiendo el algoritmo de Ward. Esta clasificación determina qué grupos se consideran más similares entre los resultados de las clasificaciones. El proceso de agrupamiento basado en la similitud de los componentes de las clases se muestra en el dendrograma de la figura 8, donde se resaltan once clusters, que representan la cantidad de clases resultantes en la clasificación del subcorpus en vasco.

Las agrupaciones de las clases muestran una mayor similitud entre unidades o agrupaciones cuando están más cerca del valor 0 en el eje x. Así, podemos observar que las dos clases que muestran mayor similitud son las etiquetadas como *eu5* y *es5*, correspondientes a la quinta clase de los análisis del subcorpus en vasco y castellano, respectivamente. Estas dos clases forman una unidad con las clases *fr1* y *en1*, correspondientes a las primeras clases de los análisis del subcorpus en francés e inglés.

Figura 8. Dendrograma de agrupación de las clases calculadas en el análisis interlingual.

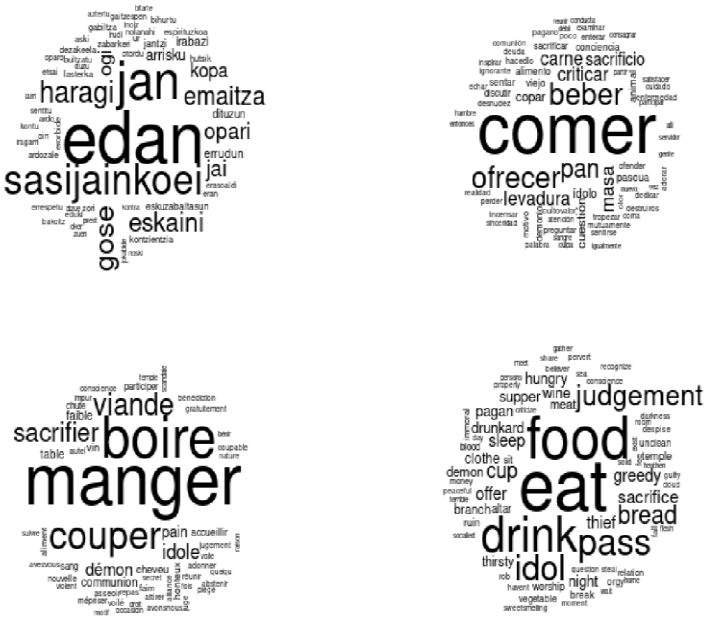


Elaboración con base en el trabajo de investigación.

La primera agrupación que se forma ocurre entre dos clases que utilizan léxicos diferentes (vasco y castellano) para traducciones ecuménicas de finales del siglo pasado. Luego, estas dos clases se agrupan con una clase en francés y otra en inglés. Esta primera agrupación es la primera en formarse y también la última en unirse al resto de las clasificaciones. Esta característica revela que esta agrupación de clases es la más distante de todas las demás agrupaciones calculadas.

A continuación, se presentan las nubes de palabras del primer grupo (figura 9), separadas según las clasificaciones originales de los subcorpus en cada idioma. En estas nubes de palabras se puede apreciar la coincidencia interlingual de muchos de los lemas mencionados, algunos de los cuales se resumen en la tabla 2.

Figura 9. Nubes de palabras de las clases agrupadas en el primer cluster.



Elaboración con base en el trabajo de investigación.

Es importante destacar que, en el caso de este primer cluster, la mayoría de las formas incluidas en los lemas utilizados en las clasificaciones tienen significados equivalentes.

Tabla 2. Resumen de algunas equivalencias en los lemas agrupados en el cluster 1.

eu	es	fr	en
edan	beber	boire	drink
jan	comer	manger	eat
sasijainko(ei)	idolo	idole	idol
haragi	carne	viande	meat
gose	hambre	faim	hungry
ogi	pan	pain	bread
opari	sacrificio	sacrifier	sacrifice
eskaini	ofrecer	-offre (no presente)	offer

Elaboración con base en el trabajo de investigación.

En la clase en lengua vasca, destaca la forma "sasijankoei" [*a los falsos dioses*], que se equipara a *ídolo* (español), *idol* (inglés) e *idole* (francés). La forma en vasco no está lematizada, ya que el equivalente debería haber sido *sasijainko* [falso dios].

A la luz de los resultados presentados en esta sección, se señalan a continuación algunas características sobre las cuales reflexionar.

DISCUSIÓN

En la sección de discusión, se abordan los tres objetivos planteados en el artículo.

Con relación al primer objetivo de construir un léxico para lematizar discursos escritos en lengua vasca, se presenta una primera versión que incluye un total de 473.330 formas con sus relaciones unívocas en tríadas de *forma-lema-categoría gramatical*. Esta versión inicial del léxico incorpora numerosos nombres propios de uso común, actualizados hasta el año 2020 según la prensa. Si bien esto permite detectar de manera precisa y eficiente una multitud de casos, también puede subrepresentar nombres propios relacionados con realidades que no están ampliamente cubiertas por la prensa o actores posteriores a la fecha en la que se recopiló el corpus utilizado para generar el léxico.

En cuanto al segundo objetivo, que se refiere al análisis de los textos y la detección de ámbitos de autodescripción, se considera que los resultados son satisfactorios. El análisis identificó cuatro ámbitos, dos de los cuales mostraron una relación interrelacionada. Esta interrelación entre los ámbitos de acción docente y formación académica en el contexto de las facultades de educación es esperada y coherente en un entorno de formación de docentes.

Respecto al tercer objetivo, que se centra en estudiar la coherencia externa de los resultados obtenidos con cuatro léxicos y cuatro corpus paralelos, se presentan dos aspectos a considerar. En primer lugar, se destacan las características cuantitativas de las clasificaciones, que se muestran en la tabla 1. Luego, se interpreta la comparación de los resultados de las clasificaciones.

En cuanto a las características numéricas de los análisis, se resalta que la clasificación obtenida con el léxico en lengua vasca presenta la tasa más baja de clasificación, con aproximadamente un 6% del corpus sin clasificar, mientras que en francés queda alrededor de un 2.5% sin clasificar. En las otras clasificaciones, hay más unidades sin clasificar que en francés, pero menos que en euskera. En cuanto al número de clases obtenidas, no se observa nada particularmente relevante en la oposición del léxico vasco con respecto a los demás léxicos.

El último punto de reflexión se centra en el proceso de agrupamiento de las clases de los diferentes análisis. Se destaca la relevancia de términos equivalentes en las diferentes clasificaciones de las traducciones utilizando los léxicos correspondientes. Además, se menciona la presencia de formas relevantes, pero no lematizadas, como el caso de la forma "sasijankoei" mencionada anteriormente, lo cual indica la necesidad de revisar detenidamente los resultados del análisis y ajustarlos según sea necesario.

También se observa que, en la clasificación y subdivisión en grupos, en algunos casos no se incluye ninguna clase de alguno de los subcorpus, lo que sugiere que el comportamiento entre las agrupaciones es similar.

En general, la sección de discusión aborda los resultados obtenidos y plantea puntos de reflexión sobre las características del léxico, los ámbitos detectados y el proceso de agrupamiento, resaltando la necesidad de revisiones adicionales y ajustes.

OPORTUNIDADES Y LÍMITES

En esta última sección se destacan varias áreas de enfoque para futuras investigaciones y mejoras en el uso del léxico en textos escritos en lengua vasca.

En primer lugar, es necesario continuar evaluando el desempeño del léxico en diferentes textos escritos en vasco. Esto implica realizar un análisis minucioso para identificar qué elementos lingüísticos se han lematizado correctamente y cuáles no durante el proceso de construcción de las clases. Se requiere un examen detallado para identificar posibles áreas de mejora y perfeccionar el léxico en función de los resultados obtenidos. Para lograr esto, es importante distribuir el léxico entre la comunidad interesada y solicitar su participación y retroalimentación para su uso y mejora continua.

Además, se destaca la necesidad de contar con una herramienta que permita mantener actualizado el léxico, especialmente en lo que respecta a las lematizaciones de los nombres propios. Esto garantizará que el léxico esté al día y pueda adaptarse a los textos a analizar, incluso si se encuentran nombres propios que no estén cubiertos por la prensa o sean actores posteriores a la fecha del corpus utilizado para generar el léxico.

Se subraya que la técnica lexicométrica conocida como ALCESTE o método Reinert ha demostrado ser aplicable con éxito en lenguas aglutinantes, como el vasco. Esto refuerza su utilidad y relevancia en diferentes contextos lingüísticos y culturales, lo que abre oportunidades para utilizar esta técnica en otros idiomas y en diversos campos de investigación.

En resumen, las oportunidades futuras incluyen la evaluación continua y la mejora del léxico en vasco, la participación de la comunidad en su uso y retroalimentación, la actualización constante del léxico y la exploración de la aplicabilidad de técnicas lexicométricas en otros contextos lingüísticos y culturales. Estas acciones contribuirán a fortalecer el léxico y su utilidad en el análisis de textos escritos en lengua vasca.

REFERENCIAS

Aranzabe, M. J., Atutxa, A., Bengoetxea, K., Diaz de Ilarraza, A., Goenaga, I., Gojenola, K., & Uria, L. (2015). Automatic Conversion of the Basque Dependency Treebank to Universal Dependencies. In M. Dickinsons, E. Hinrichs, A. Patejuk, & A. Przepiórkowski (Eds.), Proceedings of the Fourteenth International Workshop on Treebanks an Linguistic Theories (TLT14) (pp. 233–241). Institute of Computer Science of the Polish Academy of Sciences.

Benzécri, J. P. (1981). Pratique de l'analyse des donnees: Linguistique et lexicologie. Dunod.

Hanon, S. (1991). 165. La concordance. Wörterbücher: Ein internationales Handbuch zur Lexikographie, 2, 1562–1567. https://doi.org/10.1515/9783110124200.2

Ideia [@ideiainova]. (2017). Sharing a new version of the Spanish dictionary for #Iramuteq (+500k entries) [Tweet [Link a Archivo]]. In Twitter.

Ihaka, R., & Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics, 5(3), 299–314. https://doi.org/gddc3n

Loubere, L. (2023). Re: [Iramuteq-users] Dictionary in german? | iramuteq.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., & Zeman, D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 1659–1666.

R Core Team. (2020). R: A language and environment for statistical computing [Manual]. R Foundation for Statistical Computing.

Rastier, F. (1987). Représentation Du Contenu Lexical Et Formalismes De L'intelligence Artificielle. Langages, 87, 79–102. https://doi.org/10.3406/lgge.1987.1964

Ratinaud, P. (2014). IRaMuTeQ: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires.

Ratinaud, P., & Déjean, S. (2009). IRaMuTeQ: Implémentation de la méthode ALCESTE d'analyse de texte dans un logiciel libre. Modélisation Appliquée Aux Sciences Humaines Et Sociales MASHS, 8–9.

Reinert, A. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. Les cahiers de l'analyse des données, 8(2), 187–198.

Reinert, M. (1986). Un logiciel d'analyse lexicale. Les Cahiers de l'analyse Des Données, 11(4), 471–481.

Reinert, M. (1990). Alceste une méthodologie d'analyse des données textuelles et une application: Aurelia De Gerard De Nerval. Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique, 26(1), 24–54. https://doi.org/cbhfwp

Schonhardt-Bailey, C., & Bailey, A. (2013). Deliberating American Monetary Policy: A Textual Analysis. The MIT Press. https://www.jstor.org/stable/j.ctt9qf5r7

Straka, M., Hajič, J., & Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 4290–4297.

Wijffels, J., BNOSAC, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague, Republic, C., Straka, M., & Straková, J. (2020). udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the «UDPipe» «NLP» Toolkit (0.8.5). https://CRAN.R-project.org/package=udpipe