BasqueParl: descifrar la huella retórica en el Parlamento Vasco con el procesamiento del lenguaje natural

Julen Orbegozo-Terradillos[1]
Ainara Larrondo-Ureta[2]
Nayla Escribano[3]
Simón Peña-Fernández[4]
Rodrigo Agerri[5]

Recibido: 31/05/2024
Aceptado por pares: 27/07/2024
Enviado a pares: 20/06/2024
Aprobado: 09/08/2024

DOI: 10.5294/pacla.2025.28.1.3

Para citar este artículo / to reference this article / para citar este artigo
Orbegozo-Terradillos, J., Larrondo-Ureta, A. Escribano, N., Peña-Fernández, S. y Agerri, R. (2024). BasqueParl: Descifrar la huella retórica en el Parlamento Vasco con el procesamiento del lenguaje natural. *Palabra Clave*, *28*(1), e2813. https://doi.org/10.5294/pacla.2025.28.1.3

Resumen

Este artículo se centra en el proyecto BasqueParl, que utiliza técnicas de *machine learning* y procesamiento del lenguaje natural (PLN) para analizar la retórica política en el Parlamento Vasco (España). Constituye un buen ejemplo de *code-switching* del lenguaje político en un contexto de bilingüismo y una representación política cuantitativamente equilibrada entre hombres y mujeres. La investigación se basa en el análisis documental de actas parlamentarias (2012 y 2020), comprendiendo 13 872 105 palabras en euskara y español. Se consideran variables como fecha, persona oradora, año de nacimiento, sexo, partido, idioma, lemas y entidades. Para visualizar los

1 ✉ https://orcid.org/0000-0002-2959-4397. Universidad del País Vasco, España. julen.orbegozo@ehu.eus
2 https://orcid.org/0000-0003-2080-3241. Universidad del País Vasco, España. ainara.larrondo@ehu.eus
3 https://orcid.org/0000-0003-3761-7947. Universidad del País Vasco, España. nayla.escribano@ehu.eus
4 https://orcid.org/0000-0003-2080-3241. Universidad del País Vasco, España. simon.pena@ehu.eus
5 https://orcid.org/0000-0002-7303-7598. Universidad del País Vasco, España. rodrigo.agerri@ehu.eus

resultados, se presenta un *dashboard* experimental dividido en cuatro secciones: *Interventions, Tables, LDA* y *Scattertext,* empleando herramientas de lematización y reconocimiento de las entidades nombradas, entre otras. El *dashboard* ofrece una visión gráfica de la actividad parlamentaria, mostrando que el 21 % de la producción retórica es en euskara. Otros hallazgos incluyen que las mujeres intervienen menos y que los partidos minoritarios tienen una presencia discursiva desproporcionada. Además, se observa una distribución de temas políticos en función del reparto tradicional basado en el género de asuntos políticos. Se concluye que en la era del gobierno abierto y el *open data,* estas herramientas son esenciales para promover la transparencia en la Administración pública.

Palabras clave

Gobierno abierto; *open data*; retórica política; procesamiento del lenguaje natural; *machine learning*; indización automatizada; análisis del discurso.

BasqueParl: Cracking the Rhetorical Imprint in the Basque Parliament Using Natural Language Processing

Abstract

This article focuses on the BasqueParl project, which uses machine learning and natural language processing (NLP) techniques to analyze political rhetoric in the Basque Parliament (Spain). It is a good example of political code-switching in the context of bilingualism and a quantitatively balanced political representation between men and women. The investigation builds on the documentary analysis of parliamentary records (2012 and 2020), comprising 13,872,105 words in Basque and Spanish. Variables such as date, speaker, birth year, sex, party, language, slogans, and entities are considered. To visualize the results, we present an experimental dashboard divided into four sections: Interventions, Tables, LDA, and Scattertext, using lemmatization and recognition tools for named entities, among others. The dashboard offers a graphic view of parliamentary activity, showing that 21 % of rhetorical production is in Basque. Other findings include that women speak less and that minority parties have a disproportionate discursive presence. Furthermore, there is a traditional genre-based distribution of political issues. We conclude that in the era of open government and open data, these tools are essential to promote transparency in public administration.

Keywords

Open government; open data; political rhetoric; natural language processing; machine learning; automated indexing; discourse analysis.

BasqueParl: decifrando a marca retórica no Parlamento Basco com processamento de linguagem natural

Resumo

Este artigo se concentra no projeto BasqueParl, que usa técnicas de *machine learning* e processamento de linguagem natural para analisar a retórica política no Parlamento Basco (Espanha). Trata-se de um bom exemplo de *code-switching* da linguagem política em um contexto de bilinguismo e de representação política quantitativamente equilibrada entre homens e mulheres. A pesquisa baseia-se na análise documental de atas parlamentares (2012 e 2020), compreendendo 13.872.105 palavras em *euskara* e espanhol. São consideradas variáveis como data, orador, ano de nascimento, sexo, partido, idioma, slogans e entidades. Para visualizar os resultados, é apresentado um painel experimental dividido em quatro seções: intervenções, tabelas, LDA e *scattertext*, empregando ferramentas de lematização e reconhecimento de entidades nomeadas, entre outras. O painel fornece uma visão geral gráfica da atividade parlamentar, mostrando que 21% da produção retórica está em *euskara*. Outras descobertas incluem que as mulheres intervêm menos e que os partidos minoritários têm uma presença discursiva desproporcional. Além disso, observa-se uma distribuição de tópicos políticos de acordo com a distribuição tradicional baseada no gênero das questões políticas. Conclui-se que, na era do governo aberto e do *open data*, essas ferramentas são essenciais para promover a transparência na administração pública.

Palavras-chave

Governo aberto; open data; retórica política; processamento de linguagem natural; *machine learning*; indexação automatizada; análise de discurso.

Introducción

Esta investigación trabaja con un corpus lingüístico predeterminado (discursos producidos en el Parlamento Vasco, España, durante dos legislaturas, noviembre de 2012 a octubre de 2016 y octubre de 2016 a agosto de 2020), para ofrecer a la comunidad científica un punto de vista innovador en el ámbito del análisis del discurso político. A través de la aplicación del procesamiento del lenguaje natural (PLN), se procesa y analiza una gran cantidad de discursos generados por una población cautiva (preconfigurada y que no requiere una estrategia de muestreo) en un foro institucional. Se trata de una forma novedosa de analizar el lenguaje de la política en el nuevo paradigma del *big data* y el procesamiento de grandes corpus de datos.

Este es un trabajo cuyo interés reside, además, en que constituye un ejemplo de alternancia de código o *code-switching* del lenguaje político en un foro parlamentario y en la realidad sociolingüística bilingüe del País Vasco (España), donde conviven dos idiomas, el español y el euskara. Este trabajo contribuye, así, al corpus de investigaciones que analiza el fenómeno del *code-switching* (proceso de cambio de un código lingüístico a otro, según el contexto social o el marco de la conversación) en contextos políticos en los que los discursos políticos son una herramienta crucial para comunicar, persuadir o informar a la ciudadanía (Heller, 2020).

Estudios sobre la retórica parlamentaria

La retórica parlamentaria ha despertado tradicionalmente un gran interés entre lingüistas, sociólogos y psicólogos (Íñigo-Mora, 2007), y ha sido analizada por diversos autores. Además, el interés académico por este subapartado de la comunicación pública y política ha ido en ascenso con el surgimiento de herramientas informáticas capaces de procesar y sintetizar la gran cantidad de información generada en este tipo de cámaras de representantes.

Ilie (2016) explica, por ejemplo, que una característica distintiva de los parlamentos como instituciones es que el trabajo parlamentario consiste o se limita básicamente a hablar (comunicación monóloga) y a debatir (comunicación dialógica). Según la misma autora, se da por hecho, así,

que los parlamentarios son capaces de actuar e interactuar con el adversario, tanto de manera confrontada (a modo de adversarios, en temas o problemas claramente electoralistas) como de manera colaborativa (cuando se trata de apoyar o llegar a acuerdos sobre determinados temas).

Según Íñigo-Mora (2007, pp. 409 y ss.), el discurso parlamentario es un discurso distintivo, de tipo institucional, que presenta dos aspectos importantes: por un lado, la preferencia por la confrontación de ideas y opiniones entre diputados de distinta ideología y la cooperación entre aquellos otros del mismo partido; por otro, la presencia de una audiencia múltiple. Estas características promueven que los parlamentarios utilicen un lenguaje altamente estratégico con el fin de alcanzar sus fines políticos. Por otra parte, Alcaide et al. (2016), quienes examinan el uso de la emoción en el Parlamento de Andalucía (España), llaman la atención sobre el valor argumentativo y retórico de los discursos parlamentarios expresados mediante el empleo de un lenguaje emocional.

Diversas cámaras de representantes han sido objeto de análisis desde el punto de vista discursivo y argumentativo, caso del Parlamento holandés (Grijzenhou et al., 2010) y el británico (House of Commons) (Íñigo-Mora, 2007; Rheault et al., 2016; Salah, 2014). En España, ha sido objeto de interés el Parlamento de Andalucía, con especial atención a las cuestiones vinculadas al género y a la emocionalidad discursiva (Alcaide et al., 2016; Álvarez-Benito e Íñigo-Mora, 2016) y el Congreso de los Diputados (Íñigo-Mora, 2007).

Por otro lado, los estudios de Rheault y Cochrane (2016), Abercrombie y Batista-Navarro (2018) y Salah (2014) presentan particular interés, al aunar el examen de los discursos parlamentarios y el uso de modernas técnicas de análisis sustentadas en herramientas relacionadas con el *big data* (algoritmos, minería de datos y *machine learning*, respectivamente), mayoritariamente desde el punto de vista de la polaridad sentimental (positivo/negativo). Con anterioridad, estudios como el de Bara et al. (2007) habían puesto su interés en los debates parlamentarios de la House of Commons con el objetivo de comparar y determinar las virtudes o ventajas del Computer-Aided Text Analysis (CATA) a partir de dos aproximaciones analíti-

cas: la técnica de análisis semiautomático (Hamlet) y automático (Alceste). Según explican los autores, "la escritura y el discurso políticos generan una gran cantidad de datos, de hecho, datos tan voluminosos que ningún investigador podría esperar comprenderlos por sí solo o sin ayuda mecánica de algún tipo" (p. 578).

Transformación digital y gobierno abierto

En el ámbito de las instituciones público-políticas y los órganos democráticos que reúnen a electos y electas como parlamentos, diputaciones, etc., la transformación digital se relaciona directamente con el nuevo paradigma del *open government*, o gobierno abierto, así como con el fenómeno del *open data*, o datos abiertos.

El gobierno abierto es definido por Ramírez-Alujas (2012) como el "proceso basado en los principios de transparencia y apertura, participación y colaboración, que intenta responder al agotamiento de los modelos tradicionales de gobernanza en un mundo que está cambiando vertiginosamente" (p. 201). Para Calderón y Lorenzo (2010), la irrupción de internet y los avances tecnológicos han cambiado abruptamente la forma en que la ciudadanía se relaciona con el Gobierno y las instituciones, posibilitando interacciones más horizontales y sin intermediarios.

En ese contexto, Gobiernos e instituciones empezaron un proceso progresivo de facilitar el acceso a la información de carácter público por parte de la ciudadanía. Este fenómeno relacionado con el *open government* es conocido como *open data*, una filosofía y práctica que persigue que determinados datos estén disponibles de forma libre a todo el mundo.

Transparencia, participación e innovación son los tres ejes sobre los que pivota esta filosofía de *open data* (Braunschweig et al., 2012), que, a pesar de su gran aceptación y expansión teórica, no se ha implementado de forma real y efectiva en muchas instituciones públicas. Ejemplo de ello es el trabajo realizado por Pérez-Gordillo y Valle-Cuevas (2017) sobre el portal de transparencia del Parlamento de Andalucía, concluyendo que se pone en manos de la ciudadanía archivos ilegibles o en un formato inapropiado.

Una de las tecnologías fundamentales para favorecer la mencionada transformación digital de organismos e instituciones combina la ciencia computacional y la lingüística: la inteligencia artificial (IA) basada en el PLN. El origen de esta tecnología, relacionada con los sistemas computacionales, se remonta a la década de 1950, periodo en el que Alan Turing publicó un artículo titulado "Computing machinery and intelligence" (Church, 1980). Sin embargo, su desarrollo se ha acentuado especialmente desde las dos últimas décadas del siglo XX hasta nuestros días.

El PLN es un área de conocimiento interdisciplinario entre la informática y la lingüística, que dota a los computadores u ordenadores de la capacidad de entender, interpretar, generar y responder al lenguaje humano (Liddy, 2001). El PLN analiza grandes cantidades de datos o documentos, y extrae información útil para analizar y generar valor (Gonçalves, 2021). Su evolución exponencial se acompaña del incremento de la capacidad de generar datos con nuestras interacciones digitales en el contexto de la web 2.0 (Rodríguez Palchevich, 2008). La disponibilidad de datos en lenguaje natural y de recursos tecnológicos nunca ha sido tan amplia; por ello, el desarrollo de algoritmos que permitan aplicaciones útiles en nuestro día a día con toda esta información es esencial en el PLN (Vázquez Garcia, 2014).

En el contexto de esta investigación, el ámbito de interés es el uso del PLN para el etiquetado morfológico, sintáctico y semántico de determinados mensajes (automatización del análisis de texto). Esta tecnología es capaz de clasificar textos detectando temas automáticamente, reconociendo asuntos en un documento, etc. Este proceso resulta beneficioso, ya que ahorra tiempo al automatizar procedimientos que se realizaban manualmente y permite ofrecer respuestas a preguntas específicas rápidamente, tras haber procesado grandes cantidades de información previamente.

En ese sentido, un parlamento es precisamente un foro donde se generan multitud de documentos creados a partir del lenguaje natural de los parlamentarios, quienes comunican mensajes, transmiten ideas o persuaden sobre diversas premisas a una audiencia determinada. Es, por tanto, un espacio donde existe una estrecha relación entre la lingüística y la gestión

de la información o el archivo de los documentos que reflejan el uso de la retórica parlamentaria.

Justificación y objetivos

Este estudio centra su interés en el Parlamento Vasco, institución pública y legislativa que reúne a 75 electos y electas de las tres provincias de la Comunidad Autónoma del País Vasco, y constituye un paradigma de la descentralización territorial en Estados modernos (Sáiz-Arnaiz, 1985). Este parlamento es el organismo autonómico con mayor capacidad legislativa de España porque el País Vasco ostenta el mayor grado de autonomía en comparación con otras regiones. Además, presenta una considerable pluralidad ideológica y una alta equidad en su composición (53 % de mujeres y 47 % de hombres en la XI legislatura) (Orbegozo Terradillos et al., 2017).

Este trabajo se diseñó desde una doble perspectiva científica. Por un lado, una vertiente que atañe al campo del PLN, con el objetivo de ensayar e implementar una variante novedosa en el campo de las herramientas para el análisis del discurso como metodología científica, capaz de procesar y trabajar con un corpus de millones de palabras. Por otro lado, desde una perspectiva cercana a la comunicación política, esta investigación describe algunas ideas principales de la evolución del discurso político-parlamentario en el foro seleccionado para el estudio.

En este artículo, se presenta un *dashboard* (herramienta informática para visualizar y analizar datos) de carácter experimental y embrionario, que sirve para cruzar numerosa información de la población cautiva de un determinado foro, con una clasificación previa de la población objeto de estudio en función de variables como el género, el partido político, el idioma utilizado, el año, etc.

En resumen, se plantean dos objetivos específicos. Primero, desarrollar una herramienta informática basada en el PLN para el análisis de un vasto corpus lingüístico, que ayude a analizar, desde la perspectiva metodológica del análisis del discurso, la evolución de la retórica parlamentaria en un periodo determinado, poniéndola al servicio de la comunidad cien-

tífica para que pueda ser replicada en contextos similares (O1). Segundo, estructurar y ordenar el corpus de discursos parlamentarios para obtener información, identificando los principales temas en función de variables como la evolución cronológica, la ideología, el género de la persona interviniente o el idioma empleado (O2), demostrando una aplicación práctica del proceso implementado.

Contexto del estudio y diseño metodológico

Obtención de la muestra: Parlamento Vasco y open data

Durante el periodo analizado, las legislaturas X (2012-2016) y XI (2016-2020) del Parlamento Vasco, cuatro partidos políticos repitieron presencia en el hemiciclo: Euzko Alderdi Jeltzalea-Partido Nacionalista Vasco (EAJ-PNV), Euskal Herria Bildu (EH-Bildu), Partido Socialista de Euskadi-Euskadiko Ezkerra (PSE-EE) y Partido Popular (PP). Por su parte, Unión Progreso y Democracia (UPyD) obtuvo un representante en 2012 que no logró revalidar en las siguientes elecciones, y Elkarrekin-Podemos (EP) se incorporó al hemiciclo en 2016 con once diputados. Asimismo, se trata de una cámara de la que también participan otras instituciones, como el Defensor del Pueblo del País Vasco (Ararteko).

EAJ-PNV, de orientación nacionalista y liberal, conformó gobierno, en coalición con el PSE-EE, de orientación socialista, en las dos legislaturas analizadas y de su partido fue el presidente autonómico (*lehendakari*), así como la presidenta del Parlamento Vasco.

Cabe mencionar que durante el estudio se observó una deficiencia estructural a la hora de proporcionar la información en torno a los discursos parlamentarios históricos, pese a contar esta institución con un portal de transparencia (https://www.legebiltzarra.eus/portal/eu/web/eusko-legebiltzarra/transparencia/presentacion) y un apartado específico de *open data*. Así, la persona interesada que pretenda trabajar con la documentación publicada deberá tener conocimientos técnicos o informáticos, porque la información es ofrecida parcialmente y en formato XML o PDF, dificultando su codificación.

Indización automatizada para el análisis de contenido: proyecto BasqueParl

El objetivo principal de esta investigación es presentar un modelo de indización automatizada de discursos parlamentarios para obtener determinada información sobre la evolución de la retórica parlamentaria. Este formato de análisis documental, apoyado en herramientas o *softwares* que posibilitan el tratamiento de datos masivos, constituye una nueva forma de realizar análisis de contenidos, una metodología indirecta basada en el análisis y la interpretación de fuentes documentales existentes, pudiendo ser explotadas en un sentido cuantitativo y cualitativo (Guix Oliver, 2008).

La indización, que consiste en escoger los términos más apropiados para representar el contenido de un documento, es definida como uno de los procesos fundamentales del análisis de contenido. Méndez y Moreiro (1999) afirman que la indización es una técnica que "caracteriza el contenido, tanto del documento como de las consultas de los usuarios, reteniendo las ideas más representativas para vincularlas a unos términos de indización, bien extraídos del lenguaje natural empleado por los autores, o de un vocabulario controlado o lenguaje documental seleccionado *a priori*" (p. 2).

En este caso, se presentan los resultados del proyecto BasqueParl (Escribano et al., 2022), dedicado a la indización automatizada de los discursos plenarios parlamentarios de dos legislaturas (2012-2016 y 2016-2020) en el Parlamento Vasco. La principal característica del corpus es que combina dos idiomas (euskera y español), lo cual añade una categoría o variable interesante para el estudio. BasqueParl cuenta con una web donde se aloja una versión demostrativa (http://legebiltzarra.ixa.eus/bis/), pero la totalidad de los datos y metadatos está disponible para su descarga en un repositorio (https://github.com/ixa-ehu/basqueparl).

BasqueParl cuenta con casi 14 millones de palabras (n = 13 872 105) de transcripciones parlamentarias en bilingüe de sesiones plenarias (intervenciones parlamentarias, n = 41 417), pronunciadas por 165 sujetos que participaron en la actividad parlamentaria. Se procesa el texto en función de las siguientes especificaciones: por un lado, se consideran los párrafos de

los discursos como unidades; por otro, se incluyen metadatos como fecha, nombre de la persona oradora, año de nacimiento, sexo y partido, idioma, lemas y entidades de cada párrafo. El corpus está escrito como un archivo de valores separados por tabuladores (TSV, por sus siglas en inglés) y cada unidad presenta los siguientes campos: "date", "speech_id", "text_id", "speaker", "birth", "gender", "party", "language", "text", "lemmas", "lemmas_stw", "entities", "entities_stw".

Preguntas modelo y categorías de análisis

La realización de un *dashboard* experimental (en referencia a la presentación de los resultados en la versión demostrativa del programa informático) para visualizar los datos procesados permite obtener información en torno a numerosas cuestiones, gracias al cruce de las variables mencionadas. Se trata de una fuente inagotable de información que permite observar la retórica parlamentaria desde diversas perspectivas.

Además de poner el foco investigador en cuestiones de género o de usos lingüísticos, el proyecto BasqueParl sirve para responder, así, a preguntas variadas debido a la múltiple combinación de variables que entran en juego en el estudio. Como muestra del potencial de este tipo de trabajos, en este artículo se seleccionan una serie de preguntas generales y específicas, en función de las categorías mencionadas.

A ese respecto, en el apartado de resultados y a modo de demostración, se responderá a ocho preguntas que coadyuvan a la comprensión de este estudio y a su relevancia política y social, tales como:

- ¿Cuál ha sido el uso parlamentario de ambos idiomas oficiales y el reparto de la palabra en función del género y del partido?

- ¿Cuáles han sido las palabras o entidades más empleadas en función del género? ¿Cuál ha sido el universo semántico de hombres y mujeres?

- ¿Cuáles han sido las palabras o entidades más frecuentes en la retórica parlamentaria de dos partidos opuestos ideológicamente?

- ¿Se pueden detectar *topics* o temas con base en el corpus?

- ¿Cuál es la presencia de un término concreto como "terrorismo" en los diferentes *topics*/temas?

- ¿Cuáles son las palabras o términos que mejor caracterizan el discurso del *lehendakari* (presidente), Íñigo Urkullu, y del consejero de Salud, Jon Darpón?

- Comparando el universo lingüístico de ambos actores, ¿qué términos tienen en común según su retórica parlamentaria?

Para responder a estas preguntas, este estudio toma como referencia el corpus lingüístico completo de las actas parlamentarias. A modo de prueba piloto, se elabora una herramienta para la visualización creada *ad hoc* para este proyecto con una columna vertical que posibilita seleccionar, filtrar y comparar las cinco variables mencionadas, distribuyendo el contenido en cuatro apartados que se explicarán a continuación: *Interventions*, *Tables*, *LDA* y *Scattertext*.

1. *Interventions*: Visualización de una muestra compuesta por 20 intervenciones discursivas del ente que se escoja en cada petición al *software* (20 intervenciones de un parlamentario, 20 intervenciones en un u otro idioma, 20 intervenciones de un determinado partido, etc.). Debido al gran volumen de discursos y microdiscursos, se plasma en este apartado un eventual muestrario (figura 1).

2. Tables: Se emplea el reconocimiento de entidades nombradas (*named-entity recognition* [NER]), que consiste en extraer y clasificar entidades nombradas en textos (Mohit, 2014). Es una forma de clasificar por orden numérico las entidades empleadas en los discursos parlamentarios. Estas entidades son interpretables y sugieren temas, ámbitos, conceptos, asuntos, etc. Los datos de este apartado se visualizan en tres columnas: "entidades", objeto o concepto del mundo real que se identifica en un texto (Baciero Fernández, 2020); *stopwords*, palabras de uso común que se excluyen de las búsquedas para ayudar a indexar y rastrear más rápidamente (Sarica y Luo,

Figura 1. Captura de una selección de discursos en el apartado *Interventions* (*input*: partido/EAJ-PNV).

Fuente: Versión demostrativa BasqueParl.

2021), y *term frequency-inverse document frequency* (TF-IDF),[6] o frecuencia del término por frecuencia inversa del documento, búsqueda del documento más relevante para cierto término en una colección de documentos (Qaiser y Ali, 2018). En la tabla, cada concepto va ilustrado por el número de veces que se repite en el corpus la entidad aludida.

3. *Latent Dirichlet Allocation (LDA)*,[7] o asignación de Dirichlet latente (Sievert y Shirley, 2014), modelo generativo probabilístico, es una de las formas más usadas para la modelación de tópicos que se basa en el concepto de *estructura latente de tópicos* en una gran colección de documentos. Este modelo generativo de aprendizaje automático sigue una hipótesis de "bolsa

6 Mide con qué frecuencia aparece un término o frase en un documento determinado, y lo compara con el número de documentos que mencionan ese término en una colección entera de documentos.

7 La ilustración gráfica de este apartado y los sucesivos podrán observarse en las siguientes secciones de este artículo.

de palabras" donde el orden no importa (una palabra es parte de un tema y no importa en qué parte del documento se encuentra) y permite que conjuntos de observaciones puedan ser explicados por grupos no observados que explican por qué algunas partes de los datos son similares. El algoritmo (de inferencia variacional o muestreo de Gibbs) busca identificar asuntos principales de un conjunto de documentos y considera cada documento como una mezcla de temas (Gelfand, 2020), y cada tema como una distribución de palabras. Así, sirve para encontrar distribuciones de temas y palabras que mejor expliquen el corpus.

4. *Scattertext*: Herramienta que sirve para visualizar las relaciones entre palabras y entidades en un texto (Kessler, 2017). Esta función puede mostrar las palabras que más frecuentemente acompañan a cada una de las entidades para ayudarnos a entender cómo se relacionan las entidades con el lenguaje que las rodea.

Herramientas metodológicas

BasqueParl es un corpus enriquecido con metadatos vinculados al lenguaje de cada fragmento lingüístico, el año de nacimiento del hablante o político parlamentario, el género y el partido político. El desarrollo de este corpus contó con una fase previa de preprocesamiento con la que se buscó preparar el corpus para el análisis de datos mediante diversos procesos o técnicas.

En atención a que los discursos empleados iban variando y saltando de una lengua a otra, se identificó el idioma de cada unidad del corpus a partir de un sistema de detección idiomática nombrado *langdetect*, cuyo funcionamiento se basa en la comparación de los n-grama (subsecuencia de *n* elementos de una secuencia dada) de un texto con otros similares previamente identificados con métricas lingüísticas basadas en una alta probabilidad. La atribución de los idiomas se realizó por párrafos, de manera que las palabras de cada uno de ellos se asignaron a un único idioma, aunque pudieran contener alguna palabra o frase en el otro.

Para la identificación de los parlamentarios y la atribución de sus intervenciones, se realizó también una lematización y un NER para euskera

y español mediante el sistema Flair, sustentado en una arquitectura Bidirectional Long Short-Term Memory (BiLSTM) y un tipo de incrustación de palabras contextual basada en caracteres *word embeddings* (Akbik et al., 2018). El sistema y los *embeddings* han demostrado un alto rendimiento en tareas de etiquetado de secuencias como Part-of-Speech (POS), NER o *semantic role labeling* (SRL).

Asimismo, se aplicaron métodos basados en reglas y técnicas de aprendizaje automático. A este corpus se incorporaron otros metadatos, como el nombre completo de la persona interviniente, su año de nacimiento, el género y el partido político al que representa. Todo ello permitió diferenciar el uso de las lenguas en función del género o del partido de los oradores políticos, así como fomentar una perspectiva analítica comparativa en función del género y según las categorías utilizadas (idioma e ideología, por ejemplo).

Tal y como se ha explicado, para la elaboración de la *demo* o el *dashboard* de BasqueParl, los datos se han ordenado de tal forma que pueden ser elaborados a partir de cuatro categorías ya explicadas: *Tables*, *Interventios*, *LDA*, *Scattertext*.

Resultados

Este apartado comienza con una serie de datos generales obtenidos del corpus o *data set* general y se añade, *a posteriori*, información obtenida de la versión demostrativa de este proyecto.

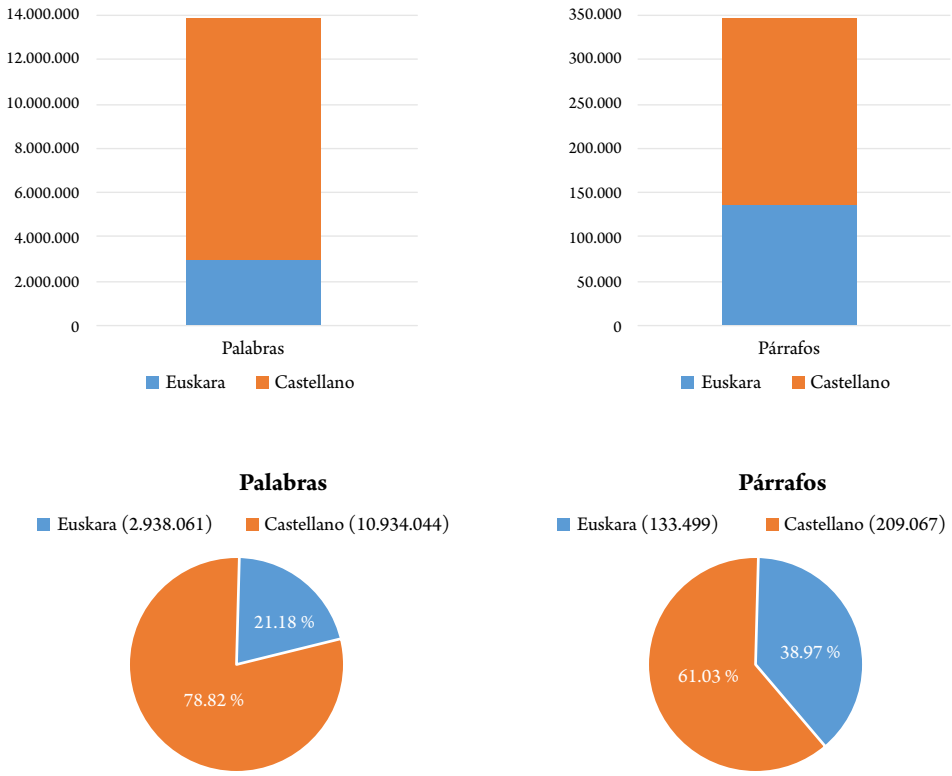
Uso de los idiomas

Los datos evidencian que en los ocho años de discurso parlamentario analizados se empleó en mayor grado el español que el euskera. Además, se sugiere que el euskera se emplea en mayor grado desde el punto de vista de la cortesía o como estrategia política o comunicativa (intervenciones breves o muy breves de introducción, contexto o transmisión del mensaje político).

La presencia ostensiblemente mayor del español concuerda con el uso general de los idiomas por parte de la sociedad vasca. Los sondeos públicos cifran en un 12,6 % la población que emplea el euskera habitualmente en su

vida diaria (Segovia, 2022). En el caso de los parlamentarios y de las parlamentarias, usan el euskera en un 21,2 % de su producción retórica (2 938 061 palabras en euskera, de una muestra total de 13 847 000) (figura 2).

Figura 2. Indicadores de uso del idioma (palabras y párrafos).

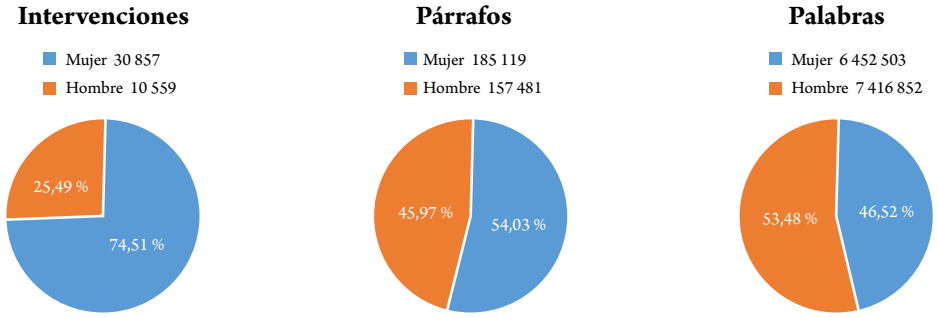


Fuente: Elaboración propia.

Reparto de la palabra en función del género

Las mujeres, en términos absolutos, pese a ser mayoría en la Cámara, han hablado menos que sus homónimos masculinos, tanto desde el punto de vista cuantitativo como del espacio temporal ocupado en la Cámara (figura 3). Sus discursos son más cortos en cuanto a su longitud (más párrafos breves, pero menos palabras en términos totales), lo cual abre dos hipótesis, no excluyentes la una de la otra: podrían explicar con mayor precisión o concisión sus ideas y podrían participar en un menor grado en las intervenciones estratégicas que requieren una mayor duración y profusión retórica.

Figura 3. Indicadores del uso de la palabra en función del género (intervenciones, párrafos y palabras).



Fuente: Elaboración propia.

Cabe explicar que el estudio diferencia entre cinco categorías lingüísticas: *speeches* (discursos), *paragraphs* (párrafos), *words* (palabras), *lemmas* (lemas) y *entities* (entidades). En la interpretación de los resultados, pueden surgir matices en función de qué categoría se seleccione para efectuar comparaciones. En ese sentido, se deben seleccionar las categorías más representativas en cada caso para obtener la información requerida. De hecho, existen circunstancias que pueden distorsionar la muestra y que deberían ser consideradas. Por ejemplo, la presidenta del Parlamento Vasco, Bakartxo Tejeria (EAJ/PNV), quien es mujer y miembro de un determinado partido, ostente el mayor número de intervenciones, porque da y quita verbalmente el turno de palabra a cada parlamentario. Por tanto, para la comparación cuantitativa del uso de la palabra por parte de mujeres y hombres, se prioriza la variable de *words* sobre la de *speeches*, tal y como puede observarse en la figura 3, porque Tejeria realiza miles de breves discursos protocolarios que podrían desfigurar la interpretación de los datos.

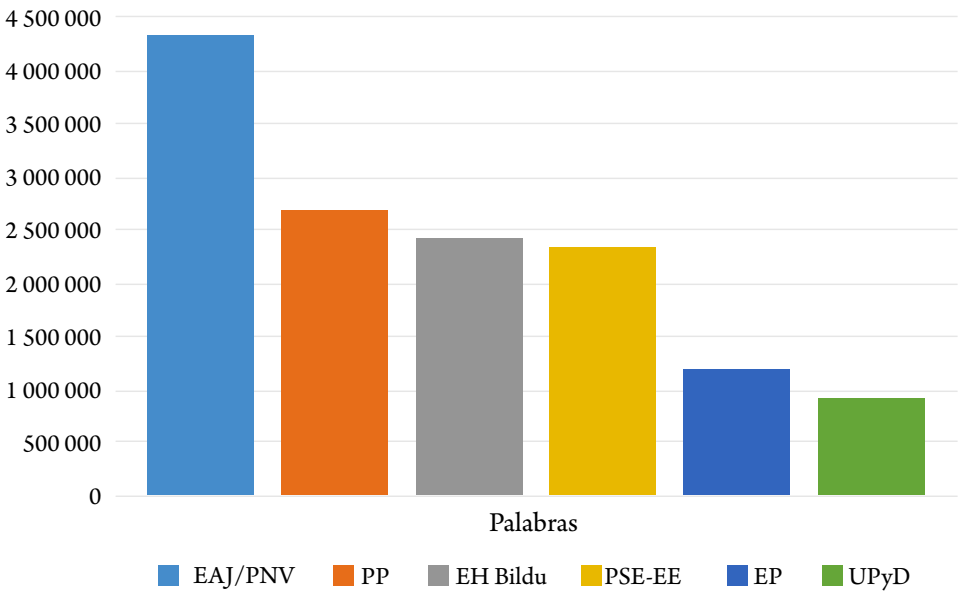
Reparto de la palabra en función del partido

La figura 4 ofrece una visión del uso de la palabra por parte de los partidos políticos con presencia en las dos legislaturas analizadas.

En general, los datos responden a una lógica parlamentaria: el partido con mayor representación ocupa mayor espacio. Sin embargo, hay un detalle que rompe esta tendencia: el Partido Popular, con menos electos y

electas (19 escaños en el total de las legislaturas), ocupa mayor espacio retórico que EH Bildu (39 escaños), principal partido de la oposición. Este dato sugiere que la coalición soberanista emplea sus espacios y sus tiempos parlamentarios con mayor concisión y brevedad, probablemente sin agotar el tiempo reglado para cada intervención. En cambio, los y las electas del Partido Popular realizan intervenciones más largas.

Figura 4. Indicadores del uso de la palabra en función del partido.



Fuente: Elaboración propia.

Apartado *Tables*: indicadores de género y de partido

A través de esta función de la versión demostrativa, puede consultarse el universo semántico de cada parlamentario o parlamentaria, de cada partido, de cada idioma, de cada año de la legislatura, etc. Es decir, la aplicación ofrece las entidades o palabras más frecuentes en cada caso que el equipo investigador interpreta y convierte en temas, tópicos o ámbitos políticos. En otras palabras, esta herramienta sirve para entender sobre qué se ha hablado, en función de múltiples variables.

En ese sentido, las preguntas investigadoras pueden ser variadas para, por ejemplo, conocer el universo retórico de un determinado partido, de un determinado parlamentario, etc. Sin embargo, para este estudio exploratorio se le solicita al *software* que muestre esa "constelación" de conceptos lingüísticos de mujeres y hombres, para comparar los resultados y observar si existen diferencias entre ambos géneros (figura 5).

Figura 5. Palabras o entidades más frecuentes en la retórica parlamentaria (2012-2020) en función del género.

Fuente: Elaboración propia.

Una de las hipótesis que puede efectuarse en este apartado es que la distribución de temas entre parlamentarios y parlamentarias podría responder a un reparto tradicional de roles de género y a un reparto tradicional de temas en política (Orbegozo Terradillos et al., 2021).

A ese respecto, en función de la producción semántica de los ocho años analizados y reflejados en la figura 5, las palabras más ligadas a la gestión administrativa, la geopolítica, las finanzas, la seguridad, la justicia, las cuestiones laborales, los grandes temas estratégicos de país o los grandes problemas políticos, etc., aparecen en la tabla de los hombres parlamentarios en una posición más destacada que en el recuadro de las mujeres. En ese sentido, las palabras más relacionadas con una "agenda" sanitaria, so-

cial, cultural, personal, etc., se ubicarían mejor posicionadas en la tabla femenina (tabla 1).[8]

Tabla 1. Universo semántico de hombres y de mujeres en el discurso parlamentario (palabras preponderantes de cada género)

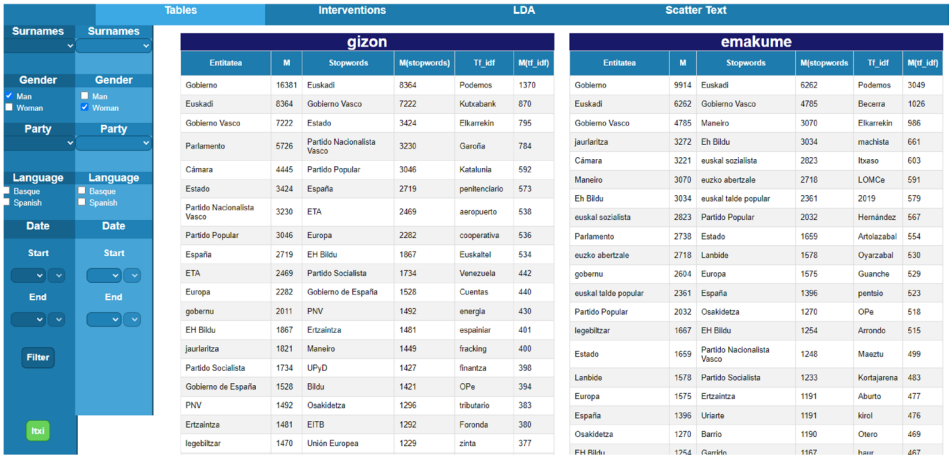
Hombres	Ámbito	Mujeres	Ámbito
ETA	Grandes temas	Lanbide	Derechos sociales
Ertzaintza	Seguridad	Osakidetza	Salud
Unión Europea	Geopolítica	Machista	Igualdad
Europa	Geopolítica	Lomce	Educación
España	Geopolítica	Pentsio (pensión)	Derechos sociales
Venezuela	Geopolítica	OPE	Derechos laborales
Katalunia (Cataluña)	Geopolítica	Kirol (deporte)	Deporte-cultura
Administración	Administración	Haur (niño)	Deporte-cultura
Estado	Administración	Rioja	Gastronomía-cultura
Madrid	Administración	* Entre las mujeres se observan más apelaciones a nombres concretos (Becerra, Itxaso, Hernández, Artolazabal, Meztu, Kortajarena, etc.). Se trata del resultado de las microintervenciones de la presidenta del Parlamento para otorgar el turno de palabra.	
Auzitegi (juzgado)	Justicia		
Penitenciario	Justicia		
Aeropuerto	Ordenación territorial		
Foronda	Ordenación territorial		
Cuentas	Fiscalidad		
Cooperativa	Empresa		
Euskaltel	Empresa		
Kutxabank	Empresa		
Naval	Empresa		
Energía	Estrategia energética		
Fracking	Estrategia energética		

Fuente: Elaboración propia.

8 Para la explicación de los resultados, se obvian en este estudio apelaciones internas a la propia Cámara, al Gobierno Vasco y a partidos políticos.

Tal y como se viene destacando, esta función puede ser empleada para realizar otro tipo de *request*, o peticiones, tomando como base otras variables. A continuación, se incluye, a modo de ejemplo (figura 6), una muestra comparativa del universo retórico de dos fuerzas políticas antagónicas con presencia en el Parlamento Vasco durante el periodo analizado: Unión Progreso y Democracia (UPyD), de ideología liberal y centralista, y Euskal Herria Bildu (EH Bildu), de ideología de izquierdas y soberanista/independentista.

Figura 6. Palabras o entidades más frecuentes en la retórica parlamentaria (2012-2020) en función del partido (UPyD vs. EH Bildu).



Fuente: Versión demostrativa BasqueParl.

La terminología empleada por UPyD se refiere, principalmente, a dos grandes temas de la política en el periodo 2012-2016 (fechas en las que este partido obtuvo representación parlamentaria): el fin de la actividad armada de Euskadi Ta Askatasuna en 2011 y el ciclo de reactivación económica tras la crisis española de 2008-2012. Esta información se deriva de la importancia que tienen las siguientes palabras en el gráfico que trasluce el universo semántico de la formación:

- Terminología relacionada con el fin de la lucha armada de ETA: "ETA", "impunidad", "demócrata", "crimen", "violento", etc.

- Terminología del ámbito económico: "Tribunal Vasco de Cuentas Públicas", "Kutxabank", "cuentas", "privilegio", "tributario", "deshaucio", "multa", "societario", "duplicidad", etc.

Asimismo, en el uso de la retórica parlamentaria por parte del único representante de UPyD en la Cámara, se intuye un uso más subjetivo, afectivo o valorativo de la lengua, con el uso de palabras como "disparate", "superfluo", etc., que, precisamente, no aluden a temas políticos concretos.

Por su parte, en EH Bildu el universo semántico se transforma notoriamente. Resulta relevante observar, de hecho, que el propio uso del lenguaje, de las palabras o de los términos refleja, en gran medida, los marcos o *frames* ideológicos de cada partido.

En este caso, la coalición soberanista se sumerge en otra constelación semántica construyendo un relato político y social paralelo, con palabras como:

- "Gipuzkoa", "Araba" y "Bizkaia" que alucen a la pluralidad y a la diversidad nacional.

- "Euskal Herri", "Estatu", "Madril", "bertako" y "Katalunia" que aluden al sentimiento nacional y a la organización territorial española.

- "Urkullu", "PNV" y "EAJ" que sirven para acentuar el carácter de principal fuerza de oposición.

- "EiTB", "Osakidetza", "Euskaltel", "Kutxabank", "LOMCE", "fracking", "energía", "pentsio", "zaintza" (cuidados, en euskera), etc., que aluden a temas sectoriales estratégicos en ese momento para el país.

La comparación de ambas redes semánticas, pues, ofrece información relevante para interpretar la retórica parlamentaria en función de, en este caso, la variable del partido político o del género del sujeto retórico. Este ejercicio comparativo puede replicarse empleando otras variables como el idioma, el año, el parlamentario, etc.

Apartado *LDA*

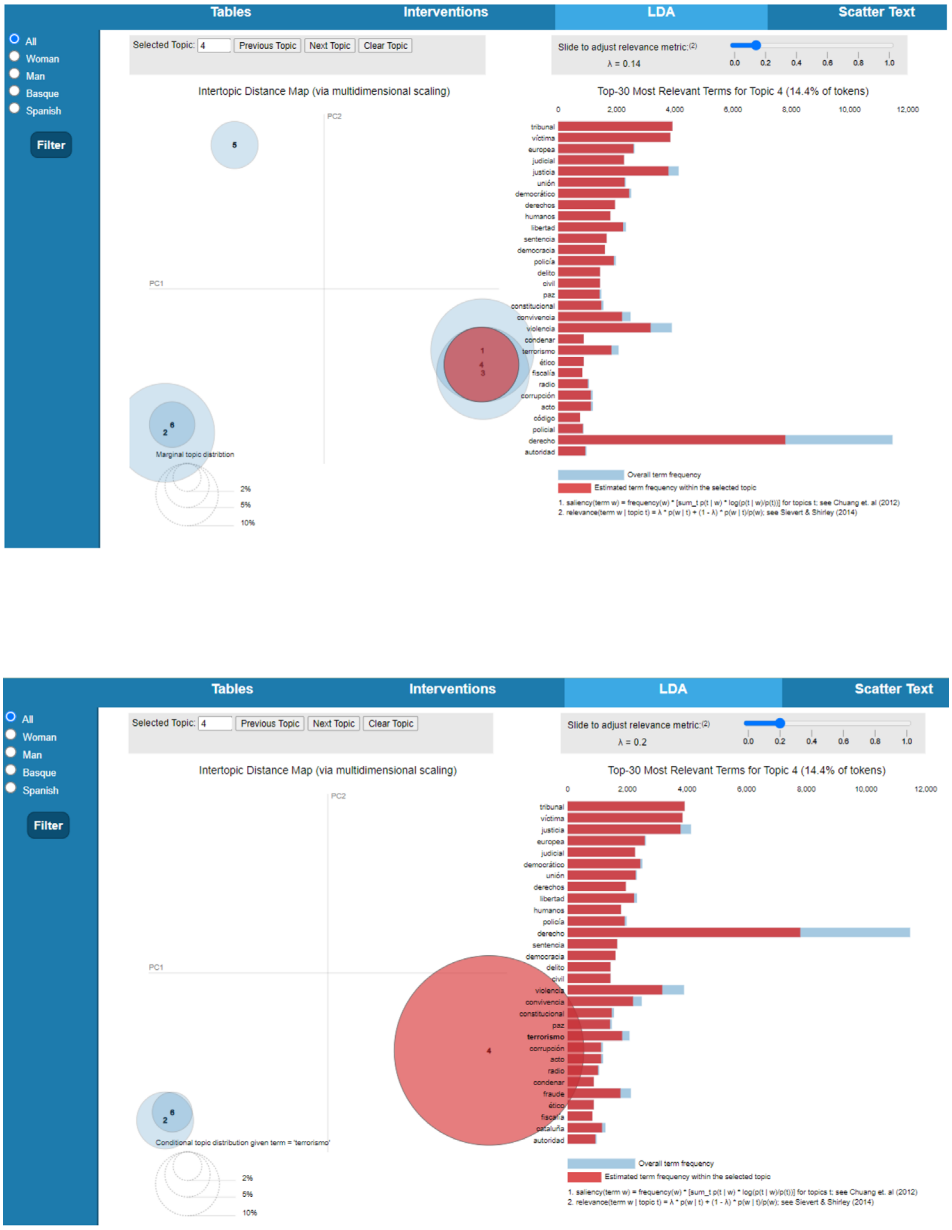
A través de esta función, es posible encontrar las distribuciones de *topics* o temas principales que se tratan en el corpus, con las palabras o términos más frecuentes en cada tema (*Top-30 most salient terms*). El *software* genera "familias" de palabras que sugieren temas latentes derivados del corpus. Es decir, cada palabra es asignada a un "tema", o *topic,* específico basándose en la probabilidad de que esa palabra esté relacionada con ese "tema".

Así, cabe la posibilidad de filtrar el corpus en función del género o del idioma empleado, y analizar si se generan distintos temas o de si los significantes (que, a su vez, sugieren significados concretos) son asignados a los mismos temas en cada caso.

El *software* demostrativo distribuye en 6 *topics* el corpus, cada uno con su lista de palabras más frecuentes. Como se observa en la figura 7, como ejemplo, el tema 4 podría ser titulado "Justicia y terrorismo", ya que alude a un universo semántico con palabras como "tribunal", "víctima", "derecho", "ético", "terrorismo", "condena", etc. Asimismo, cabe la posibilidad de consultar qué importancia tiene una palabra concreta en un *topic* y compararla con su presencia en todo el corpus (*overall term frequency* vs. *estimated term frequency within the selected topic*). Además, se observa cuán probable es que una palabra que obtiene gran relevancia en este *topic* tenga presencia en los otros temas generados por la computadora (véase imagen inferior en la figura 7). Para este estudio, se selecciona la palabra "terrorismo" y el *software* ofrece la información de que principalmente se ubica en el *topic* 4, aunque, en menor medida, también forma parte de la "constelación" inferida de los temas 6 y 2.

La visualización de esta función también incluye un gráfico titulado *Intertopic Distance Map (via multidimensional scaling)* (zona izquierda de las imágenes en la figura 7), donde se representan espacialmente los temas en un mapa, de tal forma que la distancia entre los temas sugiere similitud entre ellos: cuanto más cercanos en el espacio, más similares.

Figura 7. *Topic* 4 de la función LDA (visualización general de tema 4, arriba; y visualización de selección de término "terrorismo", abajo).



Fuente: Versión demostrativa BasqueParl.

Apartado *Scattertext*

El apartado *Scattertext* tiene una orientación o funcionalidad eminentemente comparativa. Posibilita crear gráficos terminológicos en función de la frecuencia de un término concreto en el corpus retórico de un parlamentario o una parlamentaria, de un partido, de hombres y mujeres, etc. La visualización ofrece un listado de palabras con mayor presencia (top 14) según el filtro implementado, pudiéndose consultar, además, la frecuencia de aparición de dicho término por cada 25 000 documentos o una descripción cuantitativa de los documentos considerados en cada eventual submuestra (palabras o documentos analizados).

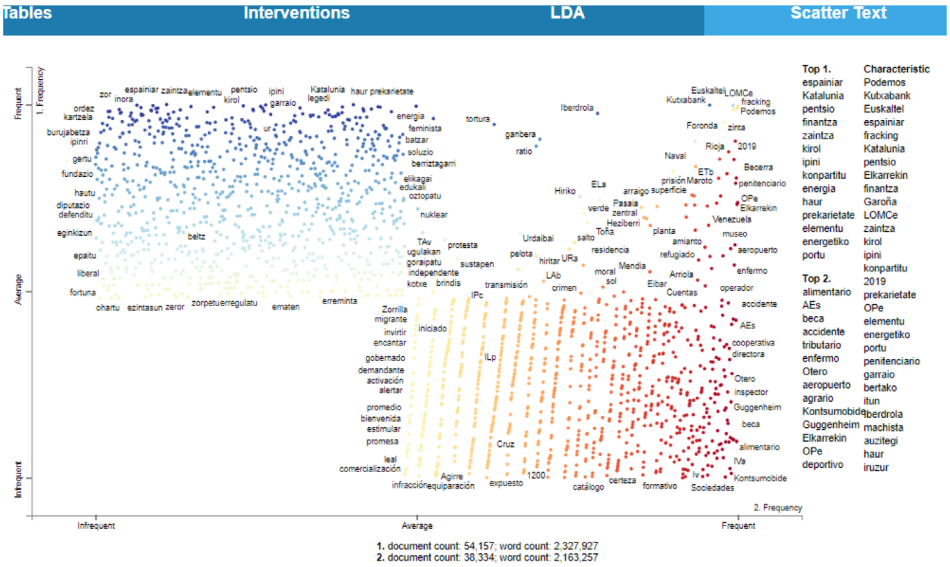
En la figura 8, por la mera distribución espacial de los nodos, puede observarse que el universo semántico de EH Bildu se aproxima y comparte más campo retórico con el del Partido Nacionalista Vasco que con el del Partido Socialista de Euskadi, probablemente porque EH Bildu y EAJ/PNV comparten terminología concreta para referirse a ciertos temas, así como interés por asuntos políticos que ubican en posiciones predilectas en su agenda (autodeterminación, euskera, cuestiones culturales, etc.).

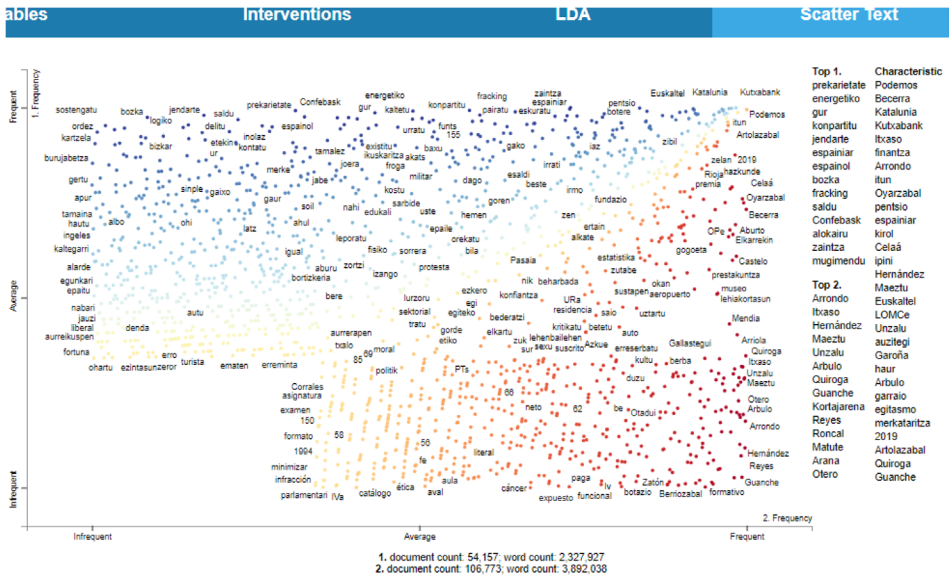
Por otro lado, en la figura 9 puede observarse una comparación entre dos miembros del Gobierno Vasco, el *lehendakari,* Íñigo Urkullu y el consejero de Salud, Jon Darpón. Este es un buen ejemplo para entender que la ubicación y la distancia entre los nodos denotan la proximidad o lejanía conceptual entre los "diccionarios semánticos" de cada sujeto.

En este caso, el consejero tiene unas palabras propias que caracterizan su discurso y que no son empleadas por el presidente: "gripe", "Kontsumobide", "enfermería", "crónico", "clínico", etc. Estos términos aluden, en efecto, a un universo semántico radicalmente sectorial: sanidad, enfermedades, infraestructuras sanitarias, conflictos laborales sectoriales, etc.

Sin embargo, en la figura 9 puede observarse, por ejemplo, que las palabras "Roncal" (en alusión a la parlamentaria Blanca Roncal, del PSE) o "copago" (en alusión a la implementación gubernamental del copago sanitario de medicamentos) son más compartidas por ambos sujetos.

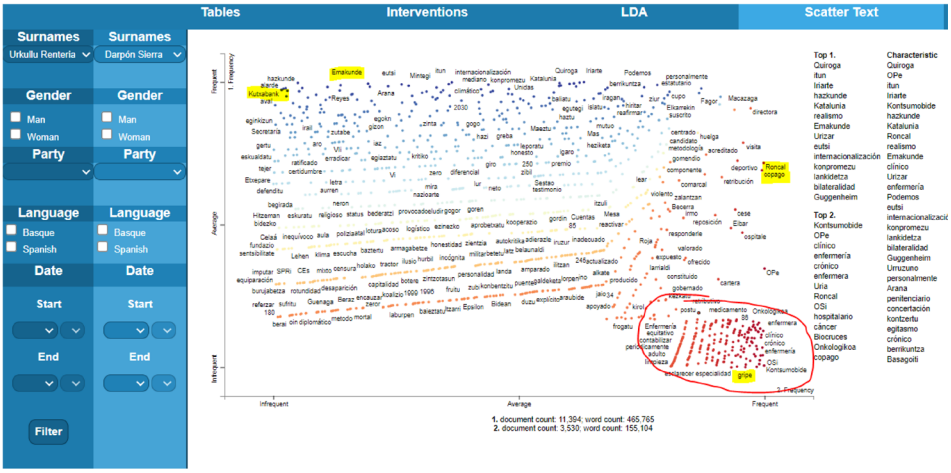
Figura 8. Frecuencia de aparición de palabras (imagen superior: EH Bildu vs. PSE-EE; imagen inferior: EH Bildu vs. EAJ/PNV).





Fuente: Versión demostrativa BasqueParl.

Figura 9. Frecuencia de aparición de palabras (comparativa entre Urkullu, *lehendakari* y Darpón, consejero de Salud).



Fuente: Versión demostrativa BasqueParl.

Por último, las palabras Emakunde (Servicio de Igualdad del Gobierno Vasco) o Kutxabank (banco con participación pública) son ampliamente empleadas por Urkullu, pero obtienen nula presencia en el lexicón de Darpón.

Conclusiones

El proyecto científico BasqueParl parte de una visión de la política como práctica discursiva que busca el entendimiento entre personas, la persuasión entre actores políticos o la aceleración de cambios sociales, todo ello performativamente, a través de la retórica política o del uso de la palabra.

En este proyecto, se estudia la retórica parlamentaria o "huella retórica" de parlamentarios o parlamentarias del Parlamento Vasco durante dos legislaturas. Se obtiene significado de su producción discursiva a través de métodos computacionales relacionados con el PLN. Además, se ensaya la publicación y la visualización de parte de la información obtenida a través de una versión demostrativa que puede ser mejorada o ampliada en el futuro.

El dato es conocimiento, y el lenguaje (como primera forma de conocimiento, tal y como defendía Wittgenstein), o las propias palabras (materia prima de la política y de la democracia), son también formas de conocimiento (Robinson, 2012). A ese respecto, este estudio genera conocimiento del proceso de indización automatizada y el análisis de contenido, favoreciendo, además, el procesamiento y la visualización de la información generada, en aras de la transparencia a la que debería acogerse la Administración pública. Este conocimiento generado viene a reimpulsar una tendencia ya estancada en el seno de la Administración pública, como es el gobierno abierto, o el *open data*.

Se ha demostrado, pues, el gran potencial de introducir estas herramientas en la Administración: se mejora la transparencia, se obtiene conocimiento político-social y científico, etc. En esta investigación, se han respondido, a modo demostrativo, preguntas investigadoras relacionadas con cuestiones del uso de la palabra y el género (las mujeres hablan menos que los hombres), del uso de idiomas oficiales en un foro parlamentario (solo una de cada cinco palabras se pronuncia en euskara) y del peso de cada partido en el reparto de la palabra (hay partidos minoritarios que ocupan mayor espacio que otros con mayor representación).

Por otro lado, se han comparado los universos semánticos que aparecen en diversos actores políticos en función de variables como el género, la ideología, etc., demostrando que esta puede ser una herramienta útil para el análisis del discurso que favorece la comprensión de los procesos sociales desarrollados en el seno de una sociedad.

En el apartado de limitaciones, este proyecto se ha elaborado con información extraída de una versión demostrativa en fase embrionaria, cuyo desarrollo será mejorado desde diversas perspectivas: trabajo de filtrado de datos, mejora de visualizaciones, optimización de la accesibilidad informática, etc. Además, existen variables que han dificultado el procesamiento de datos, favoreciendo la aparición de información sesgada en los resultados, tales como el carácter peculiar de ciertos discursos parlamentarios o el uso de dos idiomas en la misma Cámara.

En el futuro, se antoja pertinente ampliar la muestra temporal de este estudio, mejorar el procesamiento de los datos y la visualización de estos, así como expandir las herramientas y el método empleado a otros foros públicos. Además, el "*big data* lingüístico" puede emplearse para abordar en profundidad otros ámbitos de mayor complejidad computacional y analítica, como el análisis de la evolución de determinados temas o *topics* políticos, comparándolos con el interés social coyuntural, la polarización lingüística de los discursos, la afectividad en el uso del lenguaje, etc.

Referencias

Abercrombie, G. y Batista-Navarro, R. (2018). A sentiment-labelled corpus of Hansard parliamentary debate speeches. En *Proceedings of ParlaCLARIN: Common Language Resources and Technology Infrastructure (CLARIN)* (pp. 280-285). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6241

Akbik, A., Bergmann, T. y Vollgraf, R. (2019). Pooled contextualized embeddings for named entity recognition. En *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (vol. 1, pp. 724-728). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1078

Alcaide, E., Carranza, A. y Fuentes, C. (2016). Emotional argumentation in political discourse. En C. Fuentes y G. Álvarez (cords.), *A gender-based approach to parliamentary discourse* (pp. 129-159). John Benjamins Publishing Company. https://doi.org/10.1075/dapsac.68.08alc

Álvarez-Benito, G. y Íñigo-Mora, I. (2012). Repetición y reiteración en las preguntas orales del Parlamento Andaluz. *Discurso & Sociedad*, *6*(1), 21-48. https://idus.us.es/server/api/core/bitstreams/0c492e74-15ca-4c32-8461-72802253a459/content

Baciero Fernández, J. I. (2020). *Elaboración de un modelo de reconocimiento de "entidades nominales" (NER) para su uso en aplicaciones de procesamiento del lenguaje natural (NLP)* [tesis de grado, Universidad Politécnica de Madrid]. https://oa.upm.es/62858/1/TFG_JOSE_IGNACIO_BACIERO_FERNANDEZ.pdf

Bara, J, Weale, A. y Bicquelet, A. (2007). Analysing parliamentary debate with computer assistance. *Swiss Political Science Review*, *13*(4), 577-605. https://doi.org/10.1002/j.1662-6370.2007.tb00090.x

Braunschweig, K., Eberius, J., Thiele, M. y Lehner, W. (2012). The state of open data. *Limits of Current Open Data Platforms*, *1*, 72-78. http://www2012.wwwconference.org/proceedings/nocompanion/wwwwebsci2012_braunschweig.pdf

Calderón, C. y Lorenzo, S. (coords.) (2010). *Open government: Gobierno abierto*. Algón. https://libros.metabiblioteca.org/server/api/core/bitstreams/240cefc0-76c0-4912-b499-6da26a351de0/content

Church, K. W. (1980). *On memory limitations in natural language processing*. Massachusetts Institute of Technology. https://dspace.mit.edu/bitstream/handle/1721.1/149526/MIT-LCS-TR-245.pdf?sequence=1&isAllowed=y

Escribano, N., González, J. A., Orbegozo-Terradillos, J., Larrondo-Ureta, A., Peña-Fernández, S., Perez-de-Viñaspre, O. y Agerri, R. (2022). Basqueparl: A bilingual corpus of basque parliamentary transcriptions. *arXiv:2205.01506*. https://doi.org/10.48550/arXiv.2205.01506

Gelfand, A. (2000). Gibbs sampling. *Journal of the American statistical Association*, *95*(452), 1300-1304. https://doi.org/10.1080/01621459.2000.10474335

Gonçalvez, T. (2021, 9 de enero). *PLN: ¿Qué es el procesamiento del lenguaje natural?* Alura LATAM. https://www.aluracursos.com/blog/que-es-el-procesamiento-del-lenguaje-natural

Grijzenhout, S., Jijkoun, V. y Marx, M. (2010). Opinion mining in dutch hansards. En *Proceedings of the Workshop From Text to Political Positions* (pp. 1-15). Free University of Amsterdam. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=a31415a5295e795ed30bec05f22d24502f40d967

Guix Oliver, J. (2008). El análisis de contenidos: ¿Qué nos están diciendo? *Revista de Calidad Asistencial, 23*(1), 26-30. https://doi.org/10.1016/S1134-282X(08)70464-0

Heller, M. (2020). Code-switching and the politics of language. En L. Wei (ed.), *The bilingualism reader* (pp. 163-176). Routledge. https://doi.org/10.4324/9781003060406-18

Ilie, C. (2016). Parliamentary discourse and deliberative rhetoric. En P. Ihalainen, C. Ilie y K. Palonen (eds.), *Parliaments and parliamentarism: A comparative history of disputes about a European concept* (pp. 133-145). Berghahn Books. https://doi.org/10.2307/j.ctvgs0b7n.13

Íñigo-Mora, I. (2007). Estrategias del discurso parlamentario. *Discurso & Sociedad, 1*(3), 400-438. https://doi.org/10.14198/dissoc.1.3.2

Kessler, J. (2017). Scattertext: A browser-based tool for visualizing how corpora differ. En *Proceedings of ACL 2017, System Demonstrations* (pp. 85-90). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-4015

Lyddy, E. D. (2001). Natural language processing. En *Encyclopedia of Library and Information Science*. Marcel Decker. https://surface.syr.edu/cgi/viewcontent.cgi?article=1043&context=istpub

Méndez, E. y Moreiro, J. A. (1999). Lenguaje natural e indización automatizada. *Ciencias de la Información, 30*(3), 11-24. http://eprints.rclis.org/12685/1/indizacion99.pdf

Mohit, B. (2014). Named entity recognition. En I. Zitouni (eds.), *Natural language processing of semitic languages* (pp. 221-245). Springer. https://doi.org/10.1007/978-3-642-45358-8_7

Orbegozo Terradillos, J., Iturbe Tolosa, A. y González Abrisketa, M. (2017). Análisis de la nueva estrategia comunicativa de EH Bildu (2016): Hacia una narrativa de la emoción. *Anàlisi: quaderns de comunicació i cultura, 57*, 97-114. https://doi.org/10.5565/rev/analisi.3111

Orbegozo Terradillos, J., Larrondo Ureta, A. y Landaburu Corchete, A. (2021). Emociones y discurso público: Una mirada de género a la retórica política afectiva. *Cultura, Lenguaje y Representación, 26*, 247-266. https://doi.org/10.6035/clr.5838

Pérez-Gordillo, B. y Valle-Cuevas, J. A. (2017). *Portales de transparencia y ciudadanía: Análisis de utilidad y usabilidad del portal de transparencia del parlamento de Andalucía (2016-2017)* [tesis de grado, Universidad de Sevilla]. https://idus.us.es/server/api/core/bitstreams/3833679e-312d-46c5-9fd1-ee6d85837887/content

Qaiser, S. y Ali, R. (2018). Text mining: Use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications, 181*(1), 25-29. https://doi.org/10.5120/ijca2018917395

Ramírez-Alujas, Á. (2012). Gobierno abierto es la respuesta: ¿Cuál era la pregunta? *Más Poder Local, 12*, 14-22. https://pad.undp.org.mx/files/g/820dcf0c1242364677545293.44594fd/banco/archivo/102/0/gobierno-abierto-es-la-respuesta-cual-era-la-pregunta.pdf

Rheault, L. y Cochrane, C. (2020). Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis, 28*(1), 112-133. https://doi.org/10.1017/pan.2019.26

Rheault, L., Beelen, K., Cochrane, C. y Hirst, G. (2016). Measuring emotion in parliamentary debates with automated textual analysis. *PLOS ONE, 11*(12), e0168843. https://doi.org/10.1371/journal.pone.0168843

Robinson, J. (2012). Wittgenstein, sobre el lenguaje. *Estudios, 10*(102), 7-32. https://doi.org/10.5347/01856383.0102.000191959

Rodríguez Palchevich, D. (2008). *Nuevas tecnologías web 2.0: Hacia una real democratización de la información y el conocimiento*. http://biblioteca.udgvirtual.udg.mx/jspui/bitstream/123456789/3564/1/Nuevas_tecnolog%c3%adas_Web_2.0.pdf

Sáiz-Arnaiz, A. (1985). El Parlamento vasco: Relieve constitucional, organización y funcionamiento. *Revista de Estudios Políticos, 46*, 151-182. https://www.cepc.gob.es/sites/default/files/2022-01/16259repne046-047147.pdf

Salah, Z. (2014). *Machine learning and sentiment analysis approaches for the analysis of Parliamentary debates* [tesis de doctorado, University of Liverpool]. https://core.ac.uk/reader/80771780

Sarica, S. y Luo, J. (2021). Stopwords in technical language processing. *PLOS ONE, 16*(8), e0254937. https://doi.org/10.1371/journal.pone.0254937

Segovia, M. (2022, 26 de mayo). El uso del euskera en la calle en Euskadi solo crece un 2% tras 30 años de impulso. *El Independientes*. https://www.elindependiente.com/espana/2022/05/26/el-uso-del-euskera-en-la-calle-en-euskadi-solo-crece-un-2-tras-30-anos-de-impulso/#:~:text=Los%20%C3%BAltimos%20sondeos%20p%C3%BAblicos%20cifran,castellano%20en%20su%20vida%20cotidiana.

Sievert, C. y Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. En J. Chuang, S. Green, M. Hearst, J. Heer y P. Koehn (eds.), *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63-70). https://doi.org/10.3115/v1/W14-3110

Vázquez Garcia, M. (2014). El futuro de las herramientas de procesamiento del lenguaje. *COMeIN, 29*. https://doi.org/10.7238/c.n29.1405