ESTADO DE LA NACIÓN

Informe Estado de la Nación en Desarrollo Humano Sostenible 2026

Investigación

Exploración mediante *Machine Learning* no supervisado de los temas y estrategias retóricas del discurso político de años recientes

Investigador:
Elías Chavarría-Mora

San José | 2026

CONSEJO NACIONAL DE RECTORES

UNIVERSIDADES PÚBLICAS POR LA NIÑEZ Y ADOLESCENCIA 2026

PROGRAMA ESTADO DE LA NACIÓN

320.014
C512e

Chavarría-Mora, Elías
Exploración mediante machine learning no supervisado de los temas y estrategias retóricas del discurso político de años recientes / Elías Chavarría-Mora. -- San José, C.R. : PEN, 2026.
1 recurso en línea (43 páginas): archivos de texto PDF, 1100 KB

ISBN 978-9930-654-93-4
Investigación para el Informe Estado de la Nación en Desarrollo Humano Sostenible 2026 (no. 32)

1. CONFERENCIAS DE PRENSA. 2. PODER EJECUTIVO. 3. PRENSA Y POLÍTICA. 4. COSTA RICA. I. Título.

Información de la persona autora:

Elías Chavarría-Mora. https://orcid.org/0000-0001-6424-3915

Esta obra se comparte bajo la licencia
Reconocimiento – No Comercial – Compartir Igual (CC-BY-NC-SA)

Permite usar una obra para crear otra obra o contenido, modificando o no la obra original, siempre que se cite al autor, la obra resultante se comparta bajo el mismo tipo de licencia y no tenga fines comerciales

Indice

Descargo de responsabilidad 4
Resumen ejecutivo 4
Objetivo general 7
Objetivos específicos 7
Preprocesamiento y análisis de diccionario con LIWC 8
Estilo retórico de la comunicación en las conferencias de prensa 9
Metodología: K-medias 9
Clúster 1: Fragmentado 12
Clúster 2: Corporal 13
Clúster 3: Interpelativo 15
Clúster 4: Declaratorio 16
Clúster 5: Informativo 17
Distribución entre clústeres del corpus 19
Temas que dominaron la comunicación en las conferencias de prensa 21
Metodología: Modelación de temas y Asignación Latente de Dirichlet 21
Poderes de la República 22
Acción 24
Consulta 25
Cuantificación 26
Formalismo 27
Poder Ejecutivo 28
País 29
Distribución de temas en el corpus 30
Relación entre lo que se dice, y cómo se dice 32
Bibliografía 36
Anexos 38

Descargo de responsabilidad

- Esta investigación se realizó para el *Informe Estado de la Nación 2026*. El contenido es responsabilidad exclusiva de su autor, y las cifras pueden no coincidir con las consignadas en el capítulo respectivo, debido a revisiones posteriores. En caso de encontrarse diferencia entre ambas fuentes, prevalecen las publicadas en el Informe.

Resumen ejecutivo

- Esta investigación estudia las estrategias retóricas y los temas en el discurso político de la administración Chaves Robles. Puntualmente, se analizan 61 conferencias de prensa entre 18 de mayo de 2022 y 21 de enero de 2024. La investigación sigue una metodología de análisis exploratorio de datos, utilizando diversas técnicas de machine learning no supervisado. Esto para responder la pregunta: ¿Existe una relación entre los temas de los que se habla en el discurso político en años recientes y las estrategias retóricas utilizadas por este?

- El tratamiento de las conferencias de prensa consiste en textos como datos: las conferencias se transforman en una base de datos con 39011 observaciones (cada oración). A estas oraciones se les aplica un análisis de diccionario para obtener 82 variables psicométricas, así como que se utiliza la base para crear una matriz documento-término, donde las columnas corresponden a 46303 términos.

- Las estrategias retóricas son identificadas mediante un modelo de k-medias aplicado a la base de datos con las variables psicométricas, mientras los temas son modelados con una asignación latente de Dirichlet aplicada a la matriz documento-término. La relación entre las categorías resultantes es estudiada mediante una prueba de chi-cuadrado y el estadístico V de Cramér.

- Los resultados indican 5 estrategias retóricas y 7 temas en las conferencias de prensa, a saber:
- Clústeres:
- 1. Fragmentado: oraciones muy cortas y fragmentos de oración
- 2. Corporal: referencias metafóricas a corporalidad o sensorialidad
- 3. Interpelativo: referencias a un interlocutor
- 4. Declaratorio: oraciones cortas y preguntas
- 5. Informativo: oraciones muy largas y con mucha información
- Temas:
- 1. Poderes de la República: Asamblea Legislativa, Ministerios
- 2. Acción: oraciones que indican realizar acciones
- 3. Consulta: preguntas
- 4. Cuantificación: referencias en general a números, o puntualmente a dinero
- 5. Formalismo: formas formales de comunicación, como "don", "presidente", "buenas tardes", perfunctorias
- 6. Poder Ejecutivo: Referencias al poder ejecutivo
- 7. País: referencias al país, Costa Rica, nación, pueblo.
- La estrategia retórica dominante es por mucho la informativa. Los tipos interpelativo y declaratorio son también comunes, y juntos los tres tipos estructuran una lógica conversacional. La estrategia corporal es muy inusual. Con respecto a temas, acción, poderes de la república, país y formalismo son los más comunes. Consulta y cuantificación lo son menos, poder ejecutivo es el tema menos común.

- Con respecto a la pregunta principal respecto a la relación entre estrategias retóricas y temas, las pruebas estadísticas encuentran una relación, aunque moderada. Puntualmente, el tema formalismo se relaciona con un mayor uso de lenguaje fragmentado y un menor uso de lenguaje informativo. Por el contrario, los temas Poderes de la República y Acción se relacionan positivamente con un mayor lenguaje informativo y por el contrario, negativamente con el lenguaje fragmentado. El tema país también presenta un grado de relación con el lenguaje más fragmentado, sugiriendo que no se transmite mucha información sustantiva sobre este tema.

Introducción

- La estrategia comunicativa y el discurso de la administración Chaves Robles, o bien directamente del expresidente, han recibido senda atención en los medios de comunicación, así como algún grado de interés de parte de la academia (por ejemplo ver Arévalo Villalobos 2025; Ochoa Chaves y Martínez Toledo 2025; Tristan-Jiménez, Ochoa Chaves, y Martínez Toledo 2024; Sequeira Rovira 2024). Este trabajo se enmarca dentro de esta veta de investigación, adoptando una estrategia de análisis exploratorio de datos, es decir, inductivamente trabajando con una gran cantidad de datos cuantitativos para explorar dos facetas de la comunicación a partir de los patrones de los datos.

- Puntualmente, se trabajó con un corpus (es decir, una colección de textos) de las conferencias de prensa de Rodrigo Chaves entre el 18 de mayo de 2022 y el 31 de enero de 2024. Este corpus es tratado como datos, es decir, se cuantifica al utilizar las frecuencias de palabras en los textos para crear indicadores y matrices que pueden ser sometidos a técnicas de machine learning no supervisadas de clústering y reducción de dimensionalidad. Esto se realiza para identificar tanto estrategias retóricas, así como temas recurrentes en las conferencias de prensa, y a partir de esto responder la pregunta: ¿Existe una relación entre los temas de los que se habla en el discurso político en años recientes y las estrategias retóricas utilizadas por este?

- Más puntualmente, esta investigación trabaja con la totalidad del discurso que representan las conferencias de prensa, es decir, no solo con lo dicho por el expresidente sino por todas las personas que hablaron en una conferencia particular. El corpus se analiza primeramente mediante k-medias, un modelo de clústering, aplicado a 82 variables psicométricas calculadas a nivel de oración que representan características de cómo se comunica. Esto permite agrupar oraciones similares en "tipos" o estrategias de comunicación. Paralelamente, se analiza también mediante una modelación de temas utilizando una técnica de reducción de dimensionalidad, la asignación latente de Dirichlet (LDA). Esto permite inferir los temas latentes en las conferencias de prensa a partir de las correlaciones entre palabras, párrafos y temas latentes. Finalmente, se analiza la potencial relación entre estrategias retóricas y temas mediante pruebas estadísticas para variables categóricas, como chi-cuadrado y V de Cramér.

Objetivo general

- Estudiar las relaciones entre los temas y las estrategias retóricas del discurso político en años recientes.

Objetivos específicos

- Identificar tipologías de la estructura retórica del discurso político en años recientes. ¿Cómo se habla?

- Identificar los temas latentes principales del discurso político en años recientes. ¿De qué se habla?

- Estudiar las relaciones entre los temas latentes y las estrategias retóricas en el discurso político en años recientes. ¿Existe una relación entre lo que se habla y cómo se habla?

Preprocesamiento y análisis de diccionario con LIWC

- En este trabajo se analizan 611 conferencias de prensa de Rodrigo Chaves, entre 18 de mayo de 2022 y el 31 de enero de 2024. Para realizar los diferentes análisis automatizados, primeramente, se debe realizar un preprocesamiento de los textos, es decir, deben limpiarse las transcripciones de cada conferencia de prensa tal que puedan integrarse al flujo de trabajo. Primeramente, las transcripciones fueron salvadas como documentos de tipo .txt, y en R con ayuda del paquete quanteda (Benoit et al. 2018) se eliminaron las líneas completamente en blanco y se unificaron oraciones que quedaron separadas por saltos de línea (\n). También se automatizó la eliminación de marcas de tiempo (p.ej. 0:06:37) e indicaciones del número de orador (p.ej. “Orador 1” o “Desconocido”), todo esto mediante el uso de expresiones regulares, que son un sistema de escritura en informática que permite identificar patrones. Puntualmente se utilizaron las siguientes:
  - "^\\d{2}:\\d{2}:\\d{2}.* para las marcas de texto.
  - "^Desconocido$" para indicaciones de orador desconocido.
  - "^Orador \\d$" para indicaciones de orador numerado
  - "^Transcribed " para indicaciones de transcripción
- Finalmente, manualmente se eliminaron al inicio de cada transcripción anuncios comerciales no relacionados con la conferencia. Como se mencionó, el objetivo de estos procedimientos es dejar netamente el texto de la transcripción para su análisis, sin metadatos irrelevantes para el este. Posterior a esto, se transformó el corpus en una base de datos, donde las observaciones son 39011 oraciones.

[1] Dos conferencias de prensa puntuales, a saber las del 6 de julio de 2022 y la del 3 de agosto de 2022, no están incluidas en el análisis dado problemas con la transcripción, por lo que se decidió excluirlas para mantener la pureza de los datos. Dicho esto, por la cantidad el corpus, los resultados no deberían de variar enormemente de si se incluyeran, y sería posible a mediano plazo replicar el análisis incluyéndolas, lo cual puede además al compararse con el actual documento puede dar pistas de la robustes de las categorías creadas por los métodos exploratorios.

Estilo retórico de la comunicación en las conferencias de prensa

Metodología: K-medias

- Esta sección busca responder al primer objetivo de investigación planteado, es decir, ¿cómo se comunicó en las conferencias de prensa? Para esto, se utiliza primeramente un análisis de diccionario para crear diferentes métricas sobre el tipo de lenguaje utilizado en las conferencias de prensa, y posteriormente se utiliza una técnica de machine learning no supervisado, puntualmente el k-medias como técnica de clustering, para agrupar las frases utilizadas durante las conferencias en grupos internamente similares, con base en las características retóricas identificadas con el análisis de diccionario.

- Primeramente, un análisis de diccionario se basa en una tabla de datos donde cada palabra que puede estar en el corpus está clasificada en una o más variables. Por ejemplo, esa variable puede ser sentimiento (positivo, negativo o neutro), y un diccionario de ese tipo es la base del análisis de sentimiento. Sin embargo, un análisis de diccionario puede basarse en muchos otros tipos de variables, puntualmente, Linguistic Inquiry Word Count (LIWC) es un proyecto que ha creado un diccionario con más de ochenta variables de basadas en el tipo de palabras utilizadas en los textos y diversos estudios y teorías psicométricas (Boyd et al. 2022). Dado que LIWC es un diccionario propietario, debe utilizarse su software oficial para el análisis, puntualmente utilicé la versión LIWC-22 del software, así como la adaptación del diccionario al idioma castellano de Ramírez-Esparza et al. (2007).

- De esta forma, a la base de datos de 39011 oraciones se le aplica el análisis de diccionario de LIWC el cual crea métricas para 86 variables a partir de los conteos de frecuencias relativas de tipos de palabras por cada oración, una lista de todas las variables está disponible en Pennebaker et al. (2007)2. LIWC también calcula una variable “Dic”, que es el porcentaje de palabras de cada oración capturadas por el análisis, el promedio de esta variable para todo el corpus es 84.48%, un valor relativamente alto, lo cual sugiere que un porcentaje alto de la información en el texto está siendo capturada por el análisis.

[2] Esto porque el diccionario en español es una adaptación del de 2007

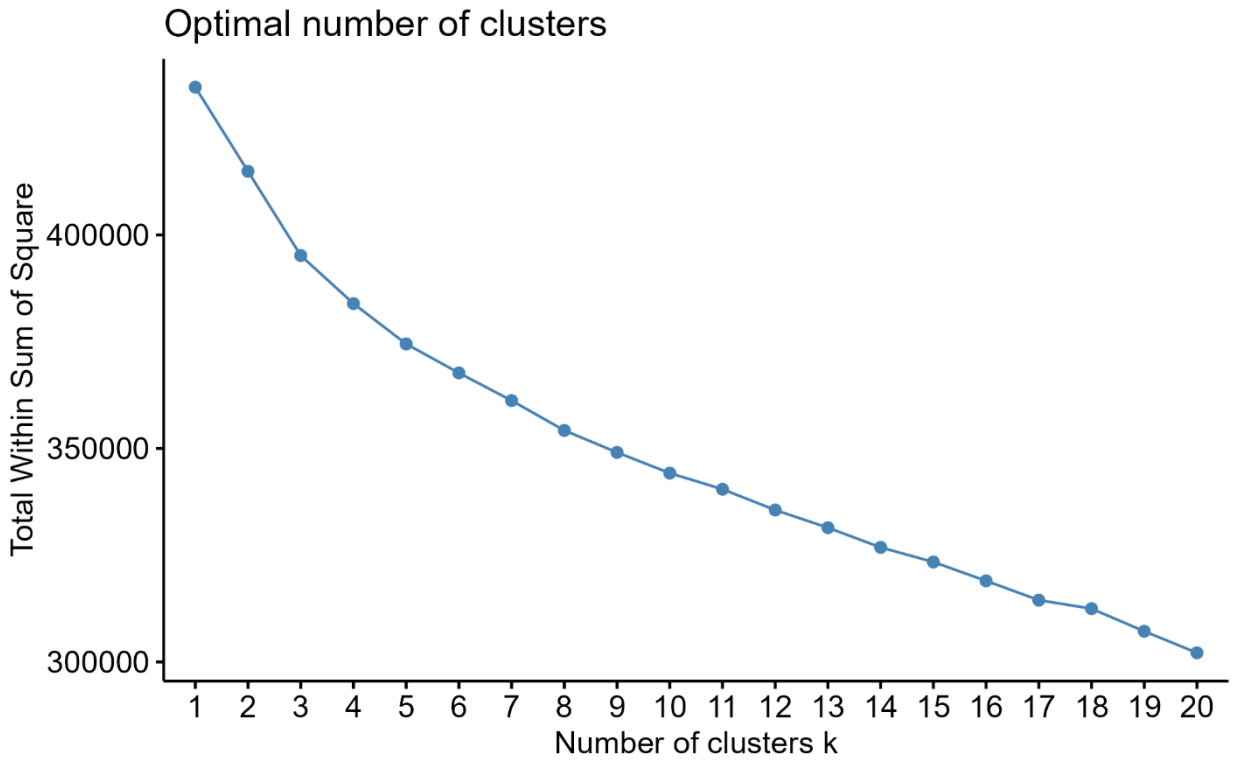
- Una vez obtenida la clasificación de las variables en las categorías de LIWC, se eliminaron unas variables puntuales que no se refieren a información retórica en los textos. Por un lado, la varible "Segment" es constante al igual que "WPS" (promedio de palabras por oración), por tanto se eliminan al no aportar información al modelo a calcularse, "Dic" se elimina al no indicar algo relativo al tipo de comunicación. Con las restantes 82 variables, todas numéricas, estas se estandarizan para que las diferencias en la magnitud no creen distorsiones en los pesos de las variables sobre el algoritmo a utilizarse en el clustering.

- Estas 82 variables se utilizan para calcular un modelo de clustering mediante k-medias. El objetivo de este modelo es agrupar las observaciones (es decir, oraciones) en grupos similares con base en sus características (es decir, las variables lingüísticas calculadas por LIWC). Esto se logra al representar los datos en un espacio n-dimensional, más puntualmente, un espacio de 82 dimensiones donde cada dimensión corresponde a una variable. En este espacio, el algoritmo de k-medias calcula k centroides, es decir, k puntos que están en el centro de una agrupación. En realidad, k-medias inicia localizando los centroides al azar en los datos, y procede de forma iterativa a ajustar su posición hasta que minimiza la suma cuadrada de las distancias al centroide más cercano (Aggarwal 2018). El resultado son grupos de oraciones que presentan atributos lingüísticos similares, es decir, se van a tener k grupos con diferentes formas de comunicarse. Para realizar k-medias se utilizan los paquetes FactoMineR (Le, Josse, y Husson 2008) y factoextra (Kassambara y Mundt 2020).

- Una de las características de k-medias es que el número de grupos (k) debe ser elegido por el investigador, no puede obtenerse automáticamente de los datos. Para seleccionar k, utilizo el método del codo, que consiste en observar un gráfico de la inercia (es decir, la sumatoria de todas las distancias cuadradas entre cada observación y el centroide al que fue asignada) y escoger k basado en un punto donde el gráfico se "aplana", es decir, que aumentar a un clúster más no disminuye mucho la inercia. La figura 1 muestra el codo, para la graficación de este, usé una muestra aleatoria de 5000 oraciones de la base de datos, pues el gráfico calcula inercia para todos los valores de k en un rango, lo cual implicaría realizar k-medias con k=1, luego con k=2, hasta un número predeterminado (utilicé k=20),

lo cual es sumamente demandante computacionalmente si se utilizan las 39011 observaciones con las 82 variables. La muestra aleatoria permite que realizar los cálculos para graficar el codo sea factible, y es razonable esperar que los resultados no van a diferir significativamente de si se usa toda la población

Figura 1
Gráfico de la inercia para escoger k para el modelo k-medias



- El gráfico de codo es visualmente difícil de interpretar en ocasiones, como en este caso. El ángulo parece volverse menos agudo alrededor de k=4 o k=5. De todas formas, la prueba más importante para la calidad del clustering es la interpretabilidad de los grupos; por tanto, se procede a realizar k-medias con k=5, en las siguientes secciones, se interpreta cada clúster formado a partir de las variables más relevantes para diferenciarlo de los demás, así como con una muestra de oraciones por clúster. Como es difícil interpretar los gráficos de la siguiente sección con las 82 variables, se utilizan los gráficos en los que aparecen solo las variables más importantes para perfilar cada clúster, es decir, las 25 con el valor absoluto más alto. Por temas de legibilidad y longitud del documento, todas estas figuras se encuentran en el anexo. Además, para ayudar a la interpretación, se incluyen 10

frases escogidas al azar por cada clúster, que, si bien son pocas, tienen como objetivo dar una idea rápida complementaria de cómo se ve la frase promedio del clúster.

Clúster 1: Fragmentado

- Como los datos fueron estandarizados, cada una de las variables indicadas nos dice qué tan lejos está el centroide en ese valor con respecto al promedio total en desviaciones estándar. El centroide funciona como una especie de "tipo ideal" para las conversaciones de este clúster.

- Los nombres de las variables no son autoexplicativos en todos los casos, por lo cual, en donde sea necesario, incluyo entre paréntesis una vez el significado de la variable. En este primer clúster3, podemos ver valores más altos que el promedio en las variables "Period" (cantidad de puntos), "EmoPos" (emociones positivas), "AlPunc" (Todo tipo de signos de puntuación), "Afect" (palabras que demuestran procesos de afecto o emociones, como "felicidad", "llorar" o "abandono") y en menor grado en "logro"(palabras que indican logros como una meta o impulso psicologico), "Humanos" (referencias a personas como "adulto", "bebé"), "Cuantif"(cuantificadores, como "poco" o "muchos"), "Comma" (comas), "BigWords" (palabras con 7 letras o más, un indicador de complejidad en razonamiento), "Asentir" (palabras que indican estar de acuerdo, como "sí" u "ok") y "Relig" (palabras relacionadas con religión). Por contraparte, el centroide presenta valores más bajos relativamente en "WC" (word count, conteo de palabras), "verbosEL", "Verbos", "TotPron" (total de pronombres), "PronPer" (pronombres personales), "PronImp" (pronombres impersonales), "Present" (verbos en presente), "Prepos" (preposiciones), "MecCog"(procesos cognitivos, como "causa", "conocer"), "Ind", "Funct" (macro-categoría de palabras funcionales que incluyen pronombres, articulos, preposiciones), "El/Ella" (uso de tercera persona singular), "Conjuc"(conjunciones) y "Artículo".

- Con respecto a los que tienen valores más altos que el promedio, estos muestran que las frases en este clúster usan palabras más positivas y emocionales, varios signos de puntuación, y hay pocas palabras por oración, así como poco uso de verbos y varios tipos

[3] La relevancia de las variables para describir al centroide se toma de los gráficos incluidos en el anexo.

de palabras según su función gramatical. Si bien las variables son útiles para identificar clústeres, la descripción del centroide es limitada para entender la estrategia retórica, por lo cual en cada clúster incluyo ejemplos de oraciones tomadas al azar de este.

- Las siguientes son ejemplos de frases en el clúster 1:
  - "Mariela."
  - "Muchas gracias, Presidente."
  - "Sí."
  - "Sí, señor."
  - "Gracias."
  - "Constantemente, todas las semanas..."
  - "Gracias."
  - "Vale, señora."
  - "Harmon Noguera."
  - "¿Cierto?"
- Como se puede ver, más que frases completas, son palabras sueltas o fragmentos cortos en respuestas a preguntas o para referirse a otras personas, por lo tanto, identifico este clúster como uno con un estilo de comunicación fragmentado.

Clúster 2: Corporal

- Para este segundo clúster, tres variables son particularmente relevantes para entender al grupo. "Biolog"(palabras referentes a procesos biológicos como “comer”, “dolor”, “sangre”), "Cuerpo" (partes del cuerpo humano) e "Ingerir" dominan completamente al caracterizar este clúster.

- Además de estas, "Salud", "Sexual" y "Sentir" son relevantes también. En menor grado, aparecen valores más altos que el promedio para "Verbos", "Tentat"(palabras tentativas, como “si”, “o”, “alguno”, “algo”), "Cuantif", "Conjunc" y "Afect", y más bajos para "WC", "Trabajo" (palabras relacionadas con trabajo, escuela y otras responsabilidades de este tipo), "TotPron", "Tiempo"(referencias a momentos como “cuando”, “ahora”, “día”, etc), " Social" (procesos sociales), "Relativ", "PronPer", "PronImp", "Insightde conocimiento y autoconocimiento) , "Ellos", "ElElla" y "BigWords"

- Lo primero a destacar es lo alta que aparece la categoría de palabras referentes a procesos biológicos; las categorías de LIWC son jerárquicas, es decir, categorías como “cuerpo”, “salud”, “sexual” e “ingerir” son todas subcategorías de la categoría biológica. Esto sugiere que todo este clúster de frases está dominado por el sentir; además, indica referencias a procesos de percepción (es decir, táctil). El clúster en general aparece como muy corporal.

- Frases en el clúster 2:

    - "Son cantones como Pocosí, como Matina, como Limón, los que han presentado proyectos."

    - "Generalmente financiamos los principales proyectos de agua, de energía, de salud de todo el país."

    - "Estamos sucediendo esta cohesión, con mucho gusto."

    - "Don Eliezer, mucho gusto saludarlo."

    - "Como bien es sabido, en este país es sumamente difícil y caro para las pymes operar."

    - "¿Un caso como qué?"

    - "Ok. Aníbal, un abrazo."

    - "Promesa suya de campaña, bajar los medicamentos, perdón, los medicamentos en general."

- "Entonces, uno se pone la mano en el corazón y dice, ya van saliendo."
- "Decidí poner todo el peso político de la presidencia en ese ministerio."

- Aquí aparecen oraciones más largas que en el clúster 1. Vemos que las referencias corporales no son tan literales, sino más bien metafóricas, por ejemplo con la figura literaria de “poner la mano en el corazón”, o “poner peso político”, la metafora se refiere a una sensación corporea de presión. Además, se encuentran usos literales de contacto corporal (“abrazo”) y una situación sí muy literal de mensaje sobre temas médicos y de salud: la frase sobre precios de medicamentos. Parecen haber diferentes motivos o formas de hacer referencia a la corporalidad, pero esta se mantiene como un tema que estructura la retórica en estas frases.

Clúster 3: Interpelativo

- Con respecto al tercer clúster, este grupo de frases se caracteriza por el alto uso de pronombres. Puntualmente, aparecen con puntuaciones altas "PronPer", "TotPron", "TuUtd", "Yo", "Vos/Utds", "Ellos", "El/Ella". Aparecen como altos además "Funct", "formal", "Oir", "insight", "Percept" (percepción o atención al ambiente), "Pasado", "Social", "Verbos", "verbYo" y "present". En cuanto a las variables donde el clúster presenta valores más bajos que el promedio del corpus, estas son "WC", "Relativ", "Prepos", "Incl"(palabras de inclusión como “y” o “con”), "Espacio", "EmoPos", "BigWords" y "Afect"

- En este caso, vemos una clara estructura discursiva enfocada alrededor de los sujetos a los cuales se refieren los pronombres. Es decir, son oraciones donde se habla sobre *alguien*, . El tono es además negativo y poco complejo (bajo score en Big Words y word count), con verbos en el pasado. Esto se confirma con las frases incluidas abajo: en todos los casos se comunica sobre una contraparte y además hay claros ataques y lenguaje agresivo, por ejemplo, “a alguna gente en el Congreso le molesta”, “usted avala el uso de trols”. No todas las frases son de conflicto, en algunos casos son más el llamado o referencia a intervención de otros actores en la conversación, por lo tanto, se indica que la estrategia no es necesariamente conflictiva, sino interpelativa.

-

- Frases en el clúster 3:
    - "La última parte, le voy a pedir al señor ministro que venga."
    - "Y eso es lo que a alguna gente en el Congreso le molesta."
    - "Ayer me preguntaban que por qué en temas financieros unos dicen una cosa y otros dicen otra."
    - "Y una condición fundamental es que, como bien decía el señor presidente, en temas de pandemia, muchas personas se quedaron sin trabajo o se les bajó su, este, jornada laboral, este, o fueron simplemente cesados, ¿verdad?, por la situación particular."
    - "La investigación sigue."
    - "Pero se le olvidaron dos cositas, ministro, que a mí me llenan el corazón de alegría."
    - "Usted, don Rodrigo, usted avala el uso de troles."
    - "yo me deciré no me lo está metiendo a mí o a la ministra?"
    - "¿Le parece?"
    - "Lo del transporte está..."

Clúster 4: Declaratorio

- Para el caso del clúster número cuatro, vemos que se distingue por valores más altos que el promedio en "verbYO", "verbosEL", "Verbos", "verbosNOS", "VerbAux", "Tentat", "Subjuntiv", "QMark"(signos de pregunta), "PronImp", "present", "Negacio", "MecCog", "insight", "Excl"(palabras de exclusión como “pero” o “sin”), "Discrep"(palabras como “quiero” o “podría”), "Causa"(¿??), "AllPunc" y "Adverb". Por el contrario, estas frases presentan menos con respecto al promedio en "WC", "PronPer", "Prepos", "Incl", "Ellos", "ElElla" y "artículo".

- De nuevo, se puede notar un patrón lingüístico en estas frases, donde predomina el uso de verbos, así como adverbios que los califican, además de que la predominancia de signos de puntuación sugiere que se están realizando preguntas. Con respecto a los valores negativos, lo que se ve es un menor uso de pronombres, artículos y preposiciones. Los ejemplos de frases sugieren que el estilo es también de oraciones cortas pero más completas que en el clúster 1, como preguntas (“¿Eso cuánto es?”, “¿Por qué no aparece?”) o respuestas directas a estas (“Sin duda”, “¡Vamos!”)

- Frases en el clúster 4:
    - "¿Eso cuánto es?"
    - "Quiero ser muy enfática y muy clara."
    - "Vamos a tomar foto."
    - "Esto sí es algo que es importante."
    - "Entonces es algo que es potestad del Congreso y no es potestad del Poder Ejecutivo."
    - "¿Por qué no aparece?"
    - "Ahora, otra noticia bastante buena."
    - "¿Del tipo de cambio, verdad?"
    - "¡Vamos!"
- "Sin duda."

Clúster 5: Informativo

- El último clúster de frases se identifica por valores altos en "WC" (el único, es decir son las orciones más largas), "Relativ", "Prepos", "Números", "Incl", "Funct", "Espacio", "Dinero" y "artículo"; y valores más bajos que el promedio en "Yo", "verbosEL", "Verbos", "TuUtd",

"Social", "QMark", "PronPer", "Present", "Period", "EmoPos", "Cuantif", "Comma", "Asentir", "AllPunc", "Afect" y "Adverb".

- Son oraciones más largas, con más palabras, incluyen referencias a espacio, a números (no solo la categoria con este nombre, sino también la llamada “Dinero”), y con ver las frases de ejemplo, es claro que en esos casos se realizan frases mucho más largas que atrapan más información.

- Frases en el clúster 5:

    - "Pueblo de Costa Rica, aquí cuando queremos podemos trabajar juntos, sin diferencias de ideología ni partido político."

    - "Entonces, meditando, reflexionando sobre el FES, la importancia que tiene y las discusiones que ocurren todos los años, yo le agradezco a la ministra Anacaterina4 Müller, que logró hacer una negociación de FEDS5 que yo considero muy exitosa para el pueblo de Costa Rica, porque los rectores, haciendo lo correcto, se comprometieron a ciertas medidas para que esos recursos enormes beneficien al pueblo."

    - "Es un proceso que, como uno de mis asesores, don Héctor decía, da un beneficio social porque nos permite reducir tarifas de acuerdo a lo que nos venga a establecer y esto nos permite realmente devolverle al usuario lo que el usuario está esperando y es una reducción en los costos."

    - "Y de finales de 2022, vamos a sumar, a 2024 vamos a sumar 2,215 centros más. en el modelo de conectividad, etcétera."

    - "Sí, bueno, en este caso, según la plataforma Mi Mejor Compra CR, en solo un local en Costa Rica está el precio a la reducción del 27% que señala el ministro, entonces cómo generalizar con un solo precio de un local el comportamiento de todo el producto."

[4] Error en la transcripción, es Anna Katharina Müller, entonces ministra de educación y actual diputada
[5] Error en la transcripción, es FEES

- "Se trata de una evacuación médica donde se requirió utilizar personal policial del grupo de apoyo operacional que estaba apoyando las labores de seguridad ciudadana en la zona."

- "Bueno, ya le avisamos al BCI que gracias, pero no gracias, que esto no va a ocurrir, entonces tenemos que conversar también con la Asamblea Legislativa, pero el proceso de licitación ya no existe, está enterrado."

- "Agradecemos a los jerarcas, medios de comunicación y a las personas que nos acompañaron."

- "Los colegios profesionales son entes públicos no estatales a los cuales se le aplica la regla y esos tenían una serie de problemas relacionados con la ejecución, pero también otras entidades que brindaban servicio a diferentes grupos."

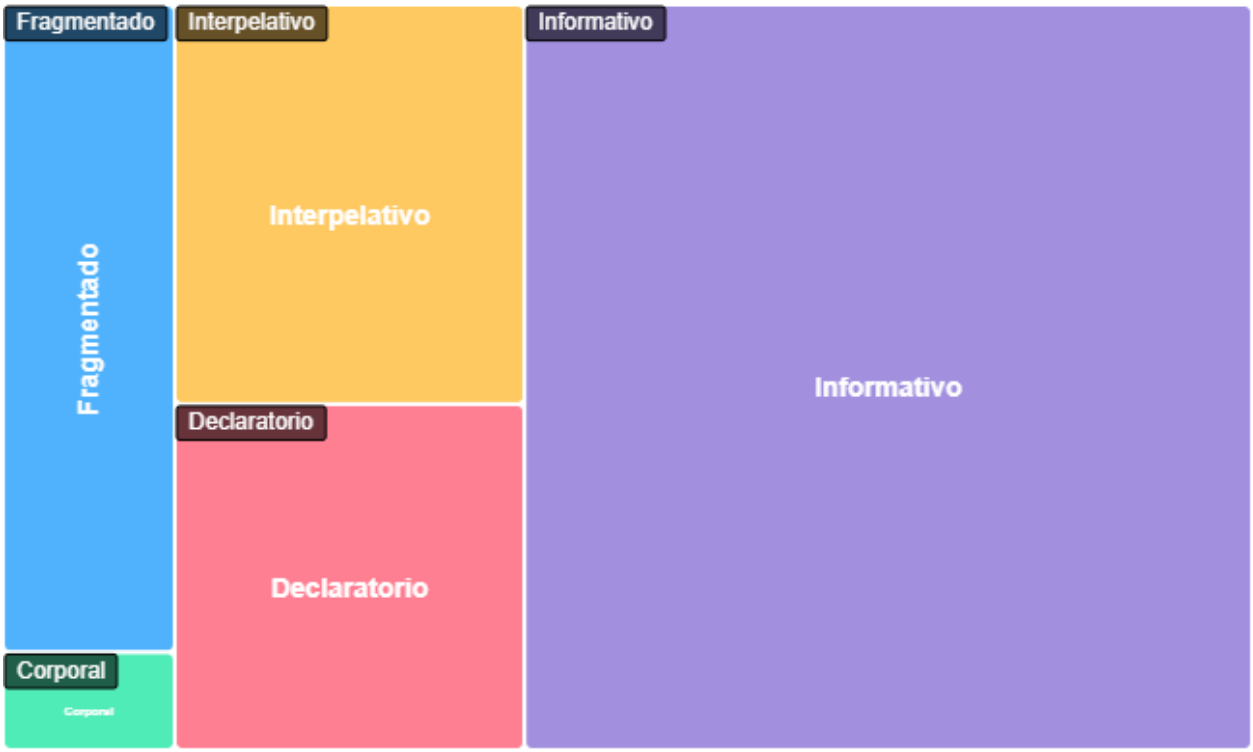
- "Entonces, hemos abierto pensiones de lesividad contra las 162 personas."

- Es inmediatamente notable una diferencia con todos los demás clústeres respecto a la longitud y cantidad de información en este clúster.

Distribución entre clústeres del corpus

- Como se puede ver por el tipo de categorías que surgieron de la interpretación de los clústeres, estos estilos retóricos son complementarios, en el sentido de que en diferentes momentos de una conferencia de prensa se usará uno u otro y en conjunto para crear párrafos completos. Ahora, en las conferencias de prensa, ¿cuánto de lo discutido era informativo? ¿Cuánto tenía una estructura interpelativa? ¿Cuánto declaratorio? En la figura 2 se incluye un gráfico de la frecuencia relativa de cada uno de los estilos retóricos. En el cuadro 3 en el anexo aparecen las frecuencias absolutas por estilo de comunicación.

Figura 2
Frecuencia relativa de estilos retóricos en el corpus



- Como se puede ver, el estilo informativo es totalmente dominante: la mayoría de las oraciones en el corpus son entonces relativamente largas y contienen información verdadera en vez de ser meramente perfunctorias o conectoras. Los tipos interpelativo y declaratorio también son relativamente comunes y tienen la característica de estructurar de alguna forma similar a una conversación: o sea, establecen cotnrapartes con las que se comunica, las interrogan, les responden. Finalmente, el tipo fragmentado es menos común, y el menos común por mucho es el corporal. Esto es curioso pues es una estrategia interesante: se refiere a lenguaje metafórico corporal y sensorial, y puede resultar el tipo de comunicación más emotiva, pero es una estrategia poco utilizada.

Temas que dominaron la comunicación en las conferencias de prensa

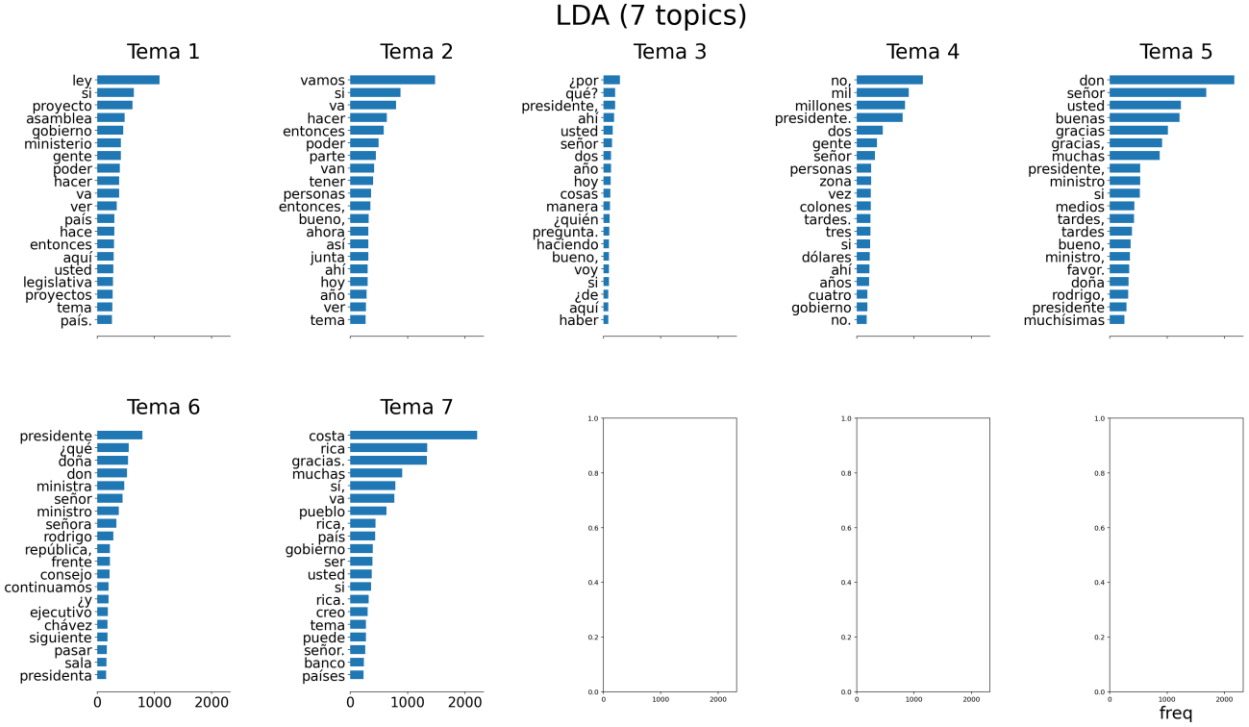
Metodología: Modelación de temas y Asignación Latente de Dirichlet

- Para analizar de qué se habló en las conferencias de prensa, utilizo también un modelo de machine learning no supervisado para la modelación de temas (topic modeling). Puntualmente, utilizo la Asignación Latente de Dirichlet (LDA por sus siglas en inglés), un modelo generativo bayesiano que parte de la siguiente idea: los textos están compuestos de palabras, y tanto las palabras como los textos pertenecen a temas, los cuales son constructos latentes. No podemos observar la pertenencia de palabras o textos a temas, pero sí las frecuencias de las palabras en los textos. Estas frecuencias se pueden representar en una matriz, con lo que el modelo calcula una distribución probabilística de pertenencia tanto de palabras y de textos a los temas (Blei, Ng, y Jordan 2003). En aras de simplificar, es un modelo de reducción de dimensionalidad que, con base en la estructura del corpus, identifica temas latentes.

- Al ser una técnica exploratoria, el objetivo es modelar varias veces con diferentes parámetros y priorizar la interpretabilidad de los resultados. Los modelos los programo en Python utilizando el paquete NLTK (Bird, Loper, y Klein 2009). Para esto se calibran una serie de parámetros, uno es si la matriz es creada por conteos de frecuencias simples en cada texto o mediante conteos de frecuencias ponderados por las frecuencias del término en todo el corpus de acuerdo con la fórmula usual, $tf - idf = tf_{ji} * \log(\frac{n}{n_i})$ (Aggarwal 2018; Bengfort, Bilbro, y Ojeda 2018). Esta ponderación resta relevancia a términos muy comunes en el corpus, que más bien son poco útiles para discriminar entre diferentes temas. También utilizo como parámetro el realizar o no stemming, es decir, recortar o no las palabras a su raíz etimológica lo cual puede ayudar a capturar juntas palabras que son ligeramente diferentes pero tiene el mismo significado; finalmente también uso como parámetros el utilizar todas las palabras, o eliminar aquellas con una frecuencia absoluta mínima de 25 o una relativa máxima de más de 90%, en ambos casos la idea es que las palabras eliminadas podrían no ser tan útiles para discrimiar entre diferentes temas. Esto significa que se probaron 2*2*2=8 parametrizaciones del modelo LDA.

- Para decidir la cantidad de temas, se calculó cada parametrización del modelo con diferente cantidad de temas (k), puntualmente, con 2, 3, 4, 5, 6, 7, 8, 9 y 10 temas. En cada caso, se revisaron las 20 palabras con más peso asignado por el modelo de cada tema, priorizando la interpretabilidad de los temas. Es decir, en total, considerando las diferentes parametrizaciones y los diferentes k tratados, se calcularon 72 modelos de LDA. En última instancia, se escoge la siguiente parametrización como la que produce temas más interpretables: frecuencias absolutas, sin stemming, sin máximos ni mínimos, con k=7. Esta parametrización produce un vocabulario de 46303 términos distintos.

Figura 3
Palabras más importantes por cada tema de acuerdo con el modelo LDA



Poderes de la República

- Uno de los principales motivos para la elección de la parametrización es que crea este tema sumamente interesante: este está estructurado alrededor del poder legislativo, con palabras como “ley”, “proyecto”, “proyectos”, “asamblea”, “legislativa”, “tema”, y otras referencias generales al país y al gobierno, como “gobierno”, “ministerio”, “poder” o “país”, es decir, puede referirse también a las relaciones o interacciones entre poderes.

- Igual que para los clústeres, incluyo 10 oraciones al azar para cada tema. En el caso del tema 1:

    - "Y así como usted, don Carlos, empieza el oscuro financiamiento de su campaña haciendo una afirmación, yo le digo lo que es posible no es probable necesariamente y en este caso estoy seguro que la percepción de la corrupción que habría que ver más con la enorme evasión fiscal que hay en este país, descarada."

    - "Esa es una institucionalidad diseñada para proteger los intereses de grupos muy pequeños y aislarlos de la influencia de la población del país, de la señora que tiene que trabajar en el parque, cerca del parque Kepos6, vendiendo artesanías, para mantener sus monopolios y sus gollerías. la rendición de cuentas es parte de la democracia, el diluir la responsabilidad no debería serlo, eso es caos."

    - "El Ministerio de Salud dice también eso, revisará la suspensión temporal cuando se presente un plan remedial."

    - "Para registrar este momento, tomamos una fotografía oficial, Don Rodrigo."

    - "Sin embargo, se ha utilizado el régimen de refugiados políticos, nuestra legislación, nuestra apertura por parte de grupos que no son refugiados políticos sino que son migrantes económicos."

    - "se levantaban las barreras para entrar al mercado mal asignados."

    - "Los CAIS, hemos logrado bajar la criminalidad dentro del centro penitenciario, la policía acaba, como dijimos hace un rato, detenido a 70 mil personas, reparamos las radiopatrullas, cocinas, dormitorios de policías, etcétera, pero volvamos a lo mismo, la Constitución es clarísima y está bien que haya separación de poderes, pero que cada palo aguante su vela."

[6] Quepos

- "No sé si al Ministro de Transportes, don Luis Amador, o alguien del Ministerio de Transportes directamente pero en el caso de presidencia no hemos recibido esa solicitud como dice el presidente, eso está en la comisión analizándose, tienen consultas audiencias, es un proceso que tiene que seguir la Asamblea Legislativa, hay otros proyectos presentados de plataformas que están en análisis, no convocados pero que están en la corriente legislativa y tiene que seguir el procedimiento, así que esto es nuevo para mí honestamente."

- "¿Tiene otra pregunta?"

- "Se presentó antes del mediodía 11:58 h de la mañana simplemente por falta de voluntad, no se quiso ver en la agenda de la tarde."

Acción

- La categoría presenta verbos de carácter activo o que demuestran la realización de acciones, como “vamos”, “hacer”, “van” o “poder” y otras que pueden señalar esa característica de realizar una acción como “ahora” u “hoy”, es decir como una referencia a hacer en el momento.

- Oraciones ilustrativas:

  - "Ok, bueno, voy a leerle otra vez la pregunta."

  - "La pregunta debería ser, pero ellos nunca lo van a hacer, preguntarse por el amor de Dios cómo pudieron haber obtenido un permiso de operación en estas condiciones por más de una década."

  - "Desde ahora, un llamado a todos los cuerpos policiales a que se sometan a la prueba, por ahora es voluntaria."

  - "Ellos lo lograron, seguimos luchando en la caja, en el mob con licencias."

  - "Señor Presidente, a mí aprovechar para saludar a todas las secretarias de este país."

- "Pero la caja está sujeta a las regulaciones del Ministerio de Salud y del mail sobre las medicinas."
- "Estamos ubicados aquí frente a la entrada al proyecto de la zona industrial HT Evolution, donde se están ejecutando las obras necesarias para garantizar el acceso a dicho proyecto y el tránsito fluido de los usuarios de la Ruta 1."
- "Además, los equipos de la Junta vienen a alivianar un poco también el dolor de aquellas personas que necesitan alivio."
- "Hubo que negociar el préstamo, hubo que negociar con las instituciones."
- "Ya eso es tema de él."

Consulta

- Una categoría con muy poco peso de las palabras que aparecen más, en esta parecen dominar aquellas que forman parte de preguntas, dados los signos de interrogación identificados.
- Oraciones ilustrativas:
  - "Señor Presidente, me hacía un editorial Pero no me hace un editorial esta vez, me hace una pregunta No se preocupe No me preocupo, o sea, con mucho gusto Voy a extender a cuatro, solo para que..."
  - "Dele, señor ministro."
  - "En específico, doña Marlene."
  - "¿Por qué es importante sentarse con ellos?"
  - "Bueno, ya que me hizo dos preguntas, tres preguntas a las primeras dos la respuesta es sí, a la última."

- "No está generando ingresos suficientes para amortizar la deuda, al contrario, está generando pérdidas."
- "Estamos trabajando, decidiendo, mejorando."
- "Esa es la manera elegante de decir, malavanzados, los que no cumplen sus responsabilidades."
- "Justo este lunes iniciamos las negociaciones para un TLC con el Ecuador."
- "Tienen diez días para darlo, yo espero que lo den."

Cuantificación

- Un tema relativamente poco claro, las palabras dominantes en muchos casos son números directamente, o bien referencias al dinero, que también es algo cuantificable. Esa naturaleza de cuantificación se mantiene al ver las oraciones ilustrativas, con referencias a cantidades, precios, cambios en cantidad y similares.
- Oraciones ilustrativas:
  - "El área de salud tiene veintisiete En total, esta doctora y su grupo tienen mapeado que solo en comisiones se van tres mil setecientas quince horas."
  - "Cuesta, la ruta del arroz, todavía hay gente que dice que no ha bajado, porque los intereses son muy grandes y sí, en medicinas, eso es una verdadera..."
  - "De hecho es tan ridícula la posición de la Nación que en el documento que yo le pedí al Banco Centroamericano que resultara ayer en un acto de transparencia sin precedente, hay una carta que le dice al BCIE la encargada de tomar las decisiones o de interactuar con ustedes en esta contratación, en las decisiones que se van a tener que tomar, es la señora Patricia Navarro."
  - "Ah bueno, solo ocho."
  - "Con su Max en el Carmen de Guadalupe, era el diario más barato de..."

- "En este momento se encuentra en construcción los bastiones y el levantamiento de las pilas principales sobre la rotonda."
- "Nos hemos estado reuniendo, don Rodrigo, allá en Punta Arenas, con el sector de importaciones."
- "Don Carlos Mora le dieron 30 mil pesos."
- "Lo felicito."
- "Francisco, tal vez presidente."

Formalismo

- Esta categoría es interesante pues visualmente pudo encontrarse en prácticamente cada una de las 72 parametrizaciones que se probaron. Es una categoría sumamente formalista, con palabras como “don”, “doña”, “señor”, “presidente”, “ministro”, “buenas”, “tardes”, “rodrigo”, “medios”, parece estar compuesta por la formalidad del lenguaje de discurso político. Este es un eje temático interesante, dado el entendimiento popular del discurso de Rodrigo Chaves como poco formal o incluso “chabacano”, pero esto sugeriría que las conferencias de prensa al menos, como un espacio de producción de comunicación relativamente formalizado, apuntan a algún nivel de formalismo y seriedad en el lenguaje utilizado en estas.
- Oraciones ilustrativas:
  - "Si alguien tiene casas, fincas, ganado, caballos caros, algo hay con los caballos y los narcotraficantes, arte, aviones, etc."
  - "Muy bien, tengo al ministro rector del CNP, al señor ministro de Agricultura y Ganadería, don Víctor Julio Carvajal."
  - "Señora ministra, yo creo que usted está en mejor posición para contestar."
  - "Ok. Muchas gracias, señora Ministra."

- "Por favor, don Mario."
- "Perdón, don Rodrigo, disculpe que le interrumpa."
- "Esto es muchísimas gracias, doña Natalia."
- "Uy, colegas, ¿cómo están?"
- "Hasta luego, señor Almaria."
- "Sí, señora Ministra, bienvenida de vuelta, Ministra."

Poder Ejecutivo

- Similar en formalidad a la anterior, pero puntualmente enfocada en el poder ejecutivo, por el uso de “presidente”, “presidenta”, “ministra”, “ministro”, “rodrigo”, “República”, “consejo”, “chávez” (probablemente un error en la transcripción del apellido del presidente, que es homófono) y “continuamos” (frase conocida del presidente alusiva a que el trabajo que realiza el ejecutivo continúa). Nótese además “sala”, tal vez debido a las interacciones entre el poder ejecutivo y la Sala Constitucional.
- Oraciones ilustrativas:
  - "La pregunta sería para la ministra de Salud."
  - "Señor presidente para la segunda denuncia relacionada con el nuevo hospital de Cartago, sumamos a la Ministra de Vivienda y Asentamientos Humanos, Jéssica Martínez Pórez."
  - "En el quinto anuncio hablaremos sobre el cierre del concurso de Ciudad Gobierno."
  - "¿Verdad que curioso?"
  - "También solicitamos la presencia del ministro en ejercicio de planificación y política económica don marlon navarro navarro álvarez la ministra de vivienda y acercamientos humanos doña jessica martínez porras y el ministro de justicia y paz don gerard campos

valverde para iniciarnos con plaza anunciar el proyecto de ciudad gobierno un video correcto Vamos con el video, por favor."

- "Desde el Ministerio de Salud emitió la orden sanitaria en noviembre, como dijo el señor presidente, para proteger la población de Chamorro, Cruzitas, El Roble, Llano Verde, El Jocote y Chorrera."
- "Por favor don."
- "Pero como una salvedad, la Ministra de Hacienda tiene el puesto de Presidenta Ejecutiva del Instituto Nacional de Vivienda y Urbanismo."
- "Continuamos con el siguiente anuncio y trata sobre la implementación de la primera antena 5G en el país."
- "Y también."

País

- Referencias en general al país, con palabras como “costa”, “rica”, “pueblo”, “ustedes” (posiblemente hablando directamente al pueblo de Costa Rica”, inclusive “gracias” podrían ser para el pueblo de Costa Rica.
- Oraciones ilustrativas:
  - "Por ejemplo, nosotros hemos tenido parques nacionales cerrados y vemos que el turismo es una explotación sostenible de la belleza que Dios nos dio y que agradecemos y eso ayuda a generar prosperidad hay países, Noruega tiene el fondo soberano más grande del mundo per cápita producto de haber explotado las reservas de gas y petróleo, creo que tienen ambas verdad señor ministro?"
  - "No le entendí, perdóname, ¿qué es lo que van a recibir?"
  - "Fíjense que hay un condominio residencial que lo construyeron o que lo empezaron a construir."

- "Pero los únicos países que se le ha descubierto con espionaje claro y sencillo a nivel mundial es Estados Unidos y Norteamérica por medio del Prisma hace 10 años, y Israel con el sistema Pegasus hace dos años, que incluso México está pidiendo cuentas sobre esa situación."

- "Muchas cosas."

- "Ahora sí, un aplauso."

- "Puede ser posible que el Consejo de Gobierno anterior que nombró a estas personas haya cometido un acto en violación de la ley con posibles implicaciones penales."

- "Muchas gracias."

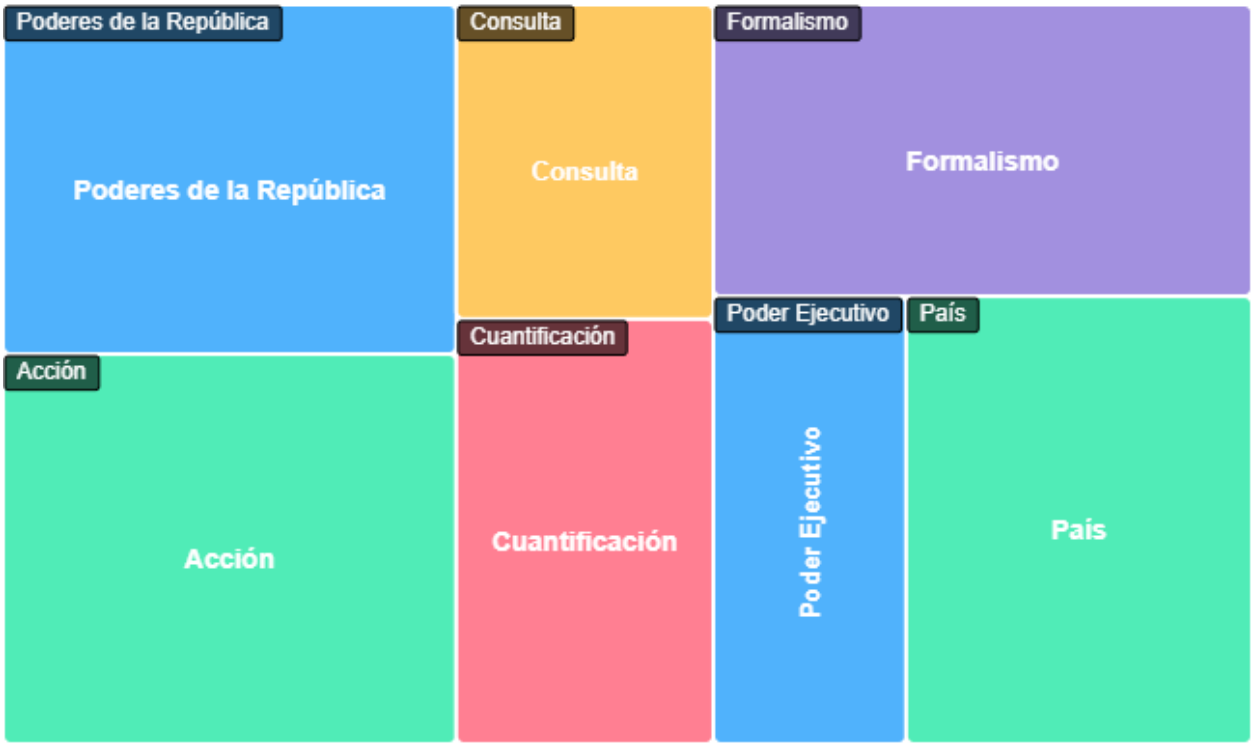
- "Y, claro, cuando yo incorporo o solicito que se realice y se revise los gastos reales de los hospitales nuevos, que su tamaño va siendo el doble, el triple y en algunos hasta cuatro veces el actual, descubrimos efectivamente que los números que se usan para hacer los cálculos actoriales son muy deficientes, son mucho menores que los que realmente va a generar el gasto institucional."

- "Muy buenas tardes al pueblo de Costa Rica."

Distribución de temas en el corpus

- Debe notarse primeramente que el modelo de LDA calcula una probabilidad de pertenencia para cada uno de los siete temas para cada oración en el corpus. Con base en estas probabilidades, asigno pertenencia de cada oración a un solo tema con base en la probabilidad más alta. En el cuadro 4 en el anexo puede verse la distribución de frecuencias por temas utilizada para crear el gráfico.

Figura 4
Frecuencia relativa de estilos retóricos en el corpus



- Como puede verse en la figura 4, hay una distribución relativamente similar entre diferentes ejes temáticos, a diferencia de la dominación del estilo "informativo" en el caso de los clústeres. Acción aparece como el tema más común en las conferencias de prensa, pero las menciones a los otros poderes de la República, las menciones al país o al pueblo, así como el formalismo, todos aparecen con una frecuencia parecida. Los temas de consulta y cuantificación aparecen como algo menos relevantes en relación con la proporción, y curiosamente, el poder ejecutivo propiamente es el tema menos común. Resumidamente, se usa mucho de la comunicación para mantener la fórmula de formalismo, referencias a diferentes poderes y a la nación, así como a comportamiento activo, tal vez esto último es lo que reemplaza las referencias al ejecutivo, en el sentido que este tema de acción puede tratarse de la administración Chaves Robles en realidad.

Relación entre lo que se dice, y cómo se dice

- A partir de los ejercicios anteriores, se encontraron 5 agrupaciones de las frases en los discursos presidenciales según la forma en que se dice, y 7 temas sobre los que se comunica, puntualmente resumidos de nuevo a continuación:

- Clústeres:

- 1. Fragmentado

- 2. Corporal

- 3. Interpelativo

- 4. Declaratorio

- 5. Informativo

- Temas:

- 1. Poderes de la República

- 2. Acción

- 3. Consulta

- 4. Cuantifiación

- 5. Formalismo

- 6. Poder Ejecutivo

- 7. País

- La pregunta que queda pendiente es si existe algún tipo de relación entre el estilo retórico utilizado y los temas tratados. Los cuadros 1 y 2 muestran el cruce entre estas dos variables con respecto al porcentaje del corpus que representa cada combinación. Los

porcentajes diferentes entre columnas y filas en los cuadros sugieren que sí hay una relación tal que algunos tipos de temas coocurren más con algunos tipos de estrategias retóricas.

Cuadro 1
Porcentaje de frecuencia relativa por fila entre estrategia retórica y tema

	Poderes	Acción	Consulta	Cuantificación	Formalismo	Ejecutivo	País
Fragmentado	2,83	4,14	3,47	5,49	45,09	5,21	33,76
Corporal	15,36	19,06	9,67	13,51	18,21	13,51	10,67
Interpelativo	15,09	16,58	12,38	11,49	21,12	10,01	13,32
Declaratorio	12.15	19,12	13,61	12,98	16,02	11,01	15,09
Informativo	21,79	22,88	7,79	12,90	10,38	9,53	14,73

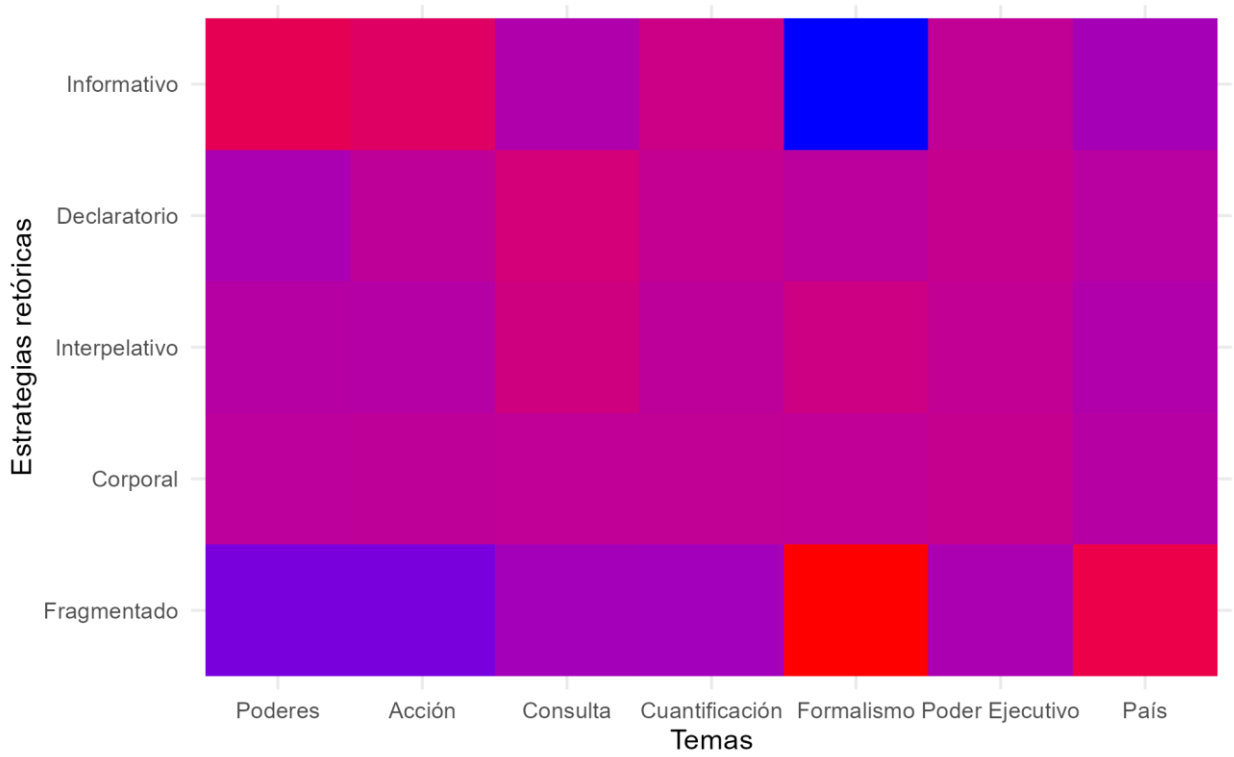
Cuadro 2
Porcentaje de frecuencia relativa por columna entre estrategia retórica y tema

	Poderes	Acción	Consulta	Cuantificación	Formalismo	Ejecutivo	País
Fragmentado	1,97	2,59	4,74	5,55	31,67	6,66	24,07
Corporal	1,61	1,79	1,99	2,06	1,93	2,60	1,15
Interpelativo	13,21	13,00	21,23	14,59	18,64	16,07	11,93
Declaratorio	9,21	12,98	20,20	14,27	12,23	15,30	11,69
Informativo	73,99	69,63	51,83	63,54	35,53	59,36	51,16

- Para corroborar si existe una relación significativa entre el estilo retórico y los temas, realizo una prueba de chi-cuadrado, esto pues ambas variables son cualitativas. Dada la enorme cantidad de observaciones, lo cual tiende a sobrepotenciar las pruebas estadísticas, aumentando la probabilidad de efectos significativos aun con efectos muy sustantivamente pequeños, además de reportar el chi-cuadrado, también indico el valor de V de Cramér, que indica la fuerza sustantiva de la asociación. Puntualmente, la prueba de chi-cuadrado tiene un χ^2=6104,7 con 24 grados de libertad y un p-value de 2,2e-16; es decir la prueba por si sugiere una asociación, aunque como se dijo, puede estar sobrepotenciada. Efectivamente, ese parece ser el caso, pues la V de Cramér es de tan solo 0,1978; un efecto modesto, recordando que este estadístico toma un valor entre 0 y 1.

- Para interpretar la relación entre temas y estrategias puntuales, utilizo la figura 5 la cual es un mapa de calor de los residuos de la prueba de chi-cuadrado. Estos residuos son la diferencia entre la distribución real de los datos y una distribución sintética creada bajo el supuesto de que no existe asociación entre las variables (la hipótesis nula de la prueba de chi-cuadrado). Tonos más fuertes de rojo o azul indican discrepancias grandes entre el valor real y el de la tabla de datos sintética, es decir que existe una asociación fuerte entre las categorías, positiva para rojo y negativa para azul. Tonos morados, en cambio, implican poca diferencia entre los datos reales y los sintéticos.

Figura 5
Mapa de calor de la relación entre estrategias retóricas y temas



- Como puede verse, hay dos relaciones particularmente fuertes: una positiva entre el tema “formalismo” y el lenguaje fragmentado, y una negativa entre “formalismo” y la estrategia informativa. Con una fuerza menor, se pueden ver las relaciones inversas con los temas “Poderes de la República” y “acción” con los estilos informativo y fragmentado. Es decir, aunque la magnitud de las relaciones en general es moderada, como ya se ha dicho, a la hora de enfocar la comunicación en meros formalismos se adopta una estrategia donde las

frases son cortas o incluso incompletas, mientras que los temas de los otros poderes de la república o la toma de acciones sí derivan en un lenguaje más informativo, que, vale la pena recordar, es el predominante en el corpus. Esto demuestra una congruencia donde los temas más sustantivos se comunican en un estilo que permite transmitir más información.

- Por otro lado, el otro cuadrante llamativo es donde se cruza el tema "país" con el estilo "fragmentado". Este tema incluye también nociones como "nación" o "pueblo"; es prematuro en este trabajo exploratorio ligarlo con nociones de populismo, pero sí es llamativo que sea otra categoría que coocurre con el lenguaje más fragmentado, es decir, no se dicen cosas complejas cuando se habla de nación o país, como si se hace con otros temas.

Bibliografía

Aggarwal, Charu C. 2018. *Machine Learning for Text*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-73531-3.

Arévalo Villalobos, Josué. 2025. "Caracterización Del Estilo de Hacer Política a Partir Del Análisis de Los Discursos Del Presidente Rodrigo Chaves Robles Durante La Primera Mitad de Su Mandato Presidencial (2022-2024)." Universidad de Costa Rica.

Bengfort, Benjamin, Rebecca Bilbro, y Tony Ojeda. 2018. *Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning*. O'Reilly Media, Inc.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, y Akitaka Matsuo. 2018. "Quanteda: An R Package for the Quantitative Analysis of Textual Data." *Journal of Open Source Software* 3 (30): 774. https://doi.org/10.21105/joss.00774.

Bird, Steven, Edward Loper, y Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc.

Blei, David M., Andrew Y. Ng, y Michael I. Jordan. 2003. "Latent Dirichlet Allocation:" *Journal of Machine Learning Research* 3: 993–1022.

Boyd, Ryan L., Ashwini Ashokkumar, Sarah Seraj, y James W. Pennebaker. 2022. "The Development and Psychometric Properties of LIWC-22." Austin, TX: University of Texas at Austin. https://www.liwc.app.

Kassambara, Alboukadel, y Fabian Mundt. 2020. "Factoextra: Extract and Visualize the Results of Multivariate Data Analyses." *CRAN: Contributed Packages*. https://doi.org/10.32614/CRAN.package.factoextra.

Le, Sebastien, Julie Josse, y Francois Husson. 2008. "FactoMineR: An R Package for Multivariate Analysis." *Journal of Statistical Software* 25 (1): 1–18. https://doi.org/10.18637/jss.v025.i01.

Ochoa Chaves, Luisa Eugenia, y Yanet Martínez Toledo. 2025. "Rodrigo Chaves y Su Construcción de Un Liderazgo Político." En *Los Yo Narrativos : Relatos de Poder En LATAM-CARIBE*, editado por Omar Rincón, Daiana Bruzzone, y Luisa Uribe, 107–19. Bogotá: Friedrich-Ebert-Stiftung.

Pennebaker, James W., Cindy K. Chung, Molly Ireland, Amy Gonzales, y Roger J. Booth. 2007. "The Development and Psychometric Properties of LIWC2007." Austin, TX. https://www.liwc.app.

Ramírez-Esparza, Nairán, James W Pennebaker, Florencia Andrea García, y Raquel Suriá. 2007. "La Psicología Del Uso de Las Palabras: Un Programa de Computadora Que Analiza Textos En Español." *Revista Mexicana De Psicología* 24 (1): 85–99.

Sequeira Rovira, Paula. 2024. "La Animalización Como Estrategia Política. El Caso Del Presidente de Costa Rica, Rodrigo Chaves Robles." *Tabula Rasa*, 155–79. https://doi.org/10.25058/20112742.n51.07.

Tristan-Jiménez, Larissa, Luisa Ochoa Chaves, y Yanet Martínez Toledo. 2024. "Narrativas Del Gobierno de Rodrigo Cháves Robles: Una Lectura Crítica y Feminista Al Populismo de Derechas En Costa Rica." En *Más Derechos, Menos Derechas : Acerca de La Comunicación y La Democracia En América Latina*, editado por Daiana Bruzzone, Paola Ricaurte, y Omar Rincón, 115–49. Buenos Aires: Consejo Latinoamericano de Ciencias Sociales, Friedrich-Ebert-Stiftung.

Anexos

Gráficos que muestran variable más importante para cada clúster

Figura 6.
25 variables con mayor valor absoluto del clúster 1

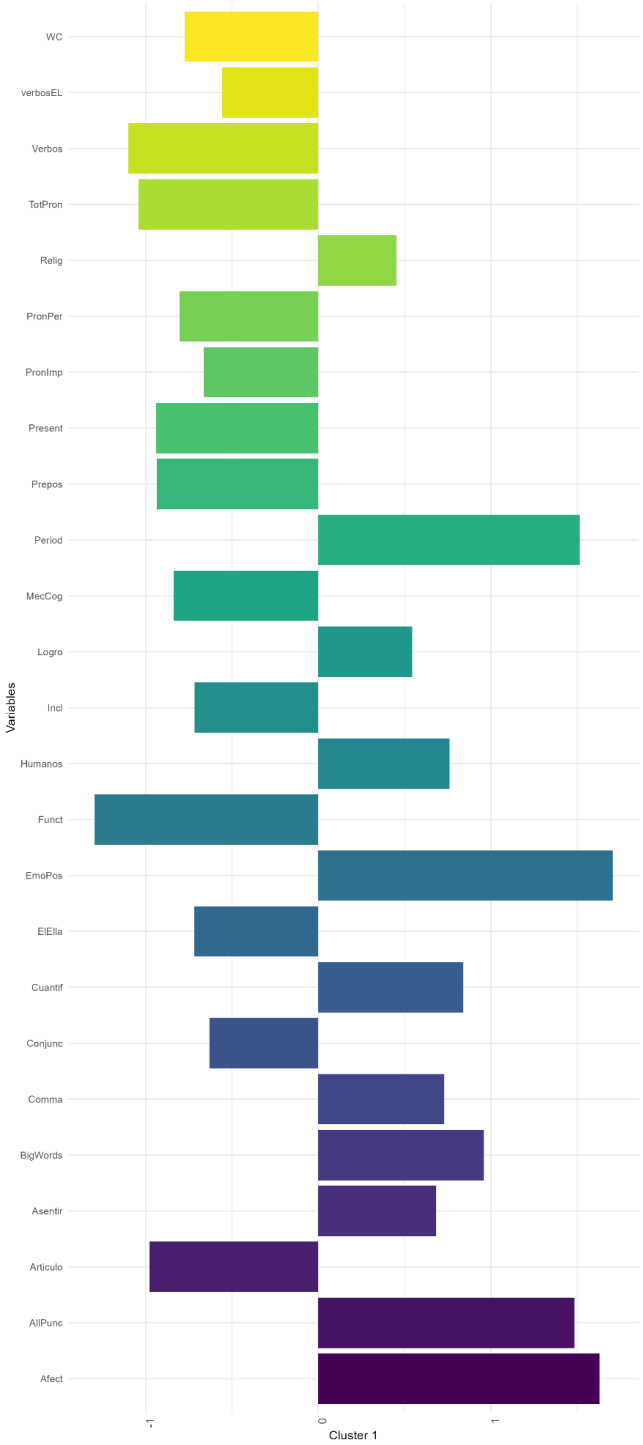


Figura 7
25 variables con mayor valor absoluto del clúster 2

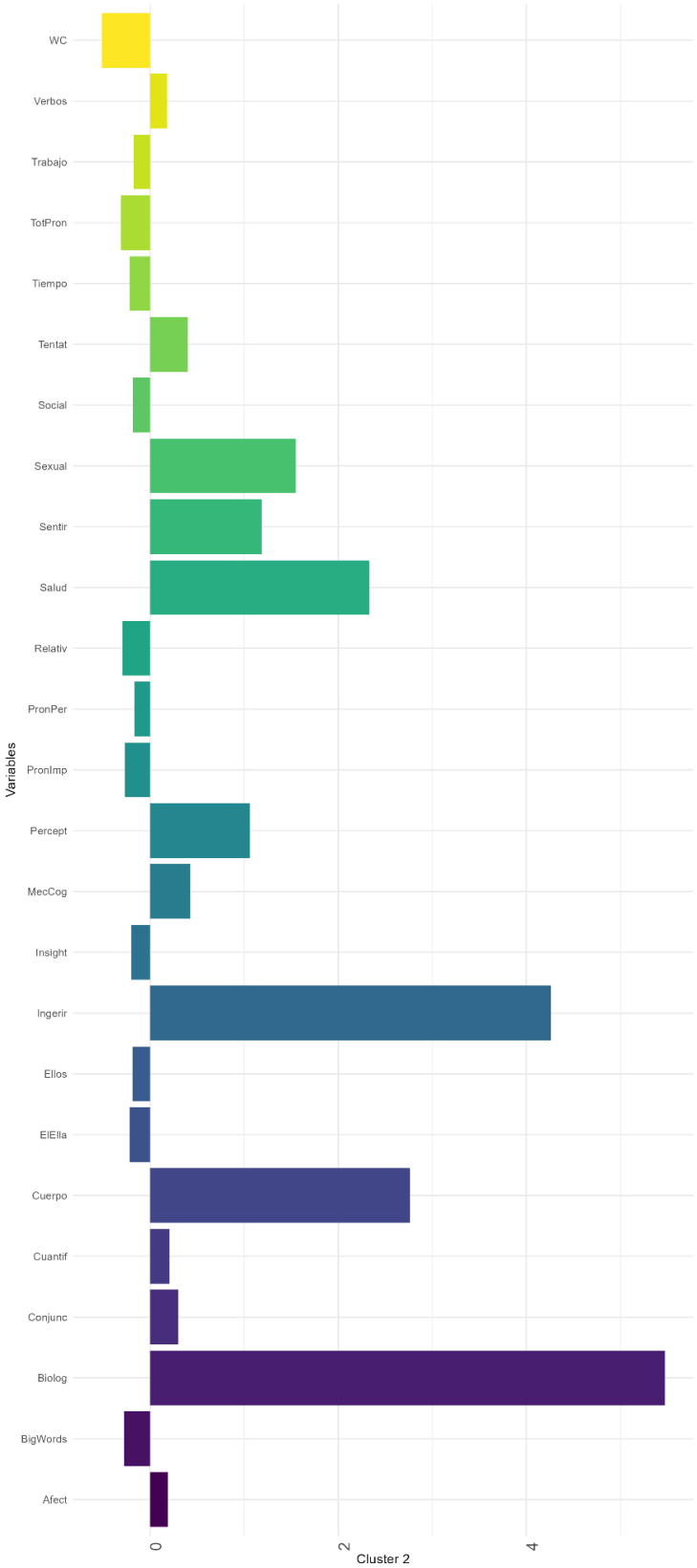


Figura 8
25 variables con mayor valor absoluto del clúster 3

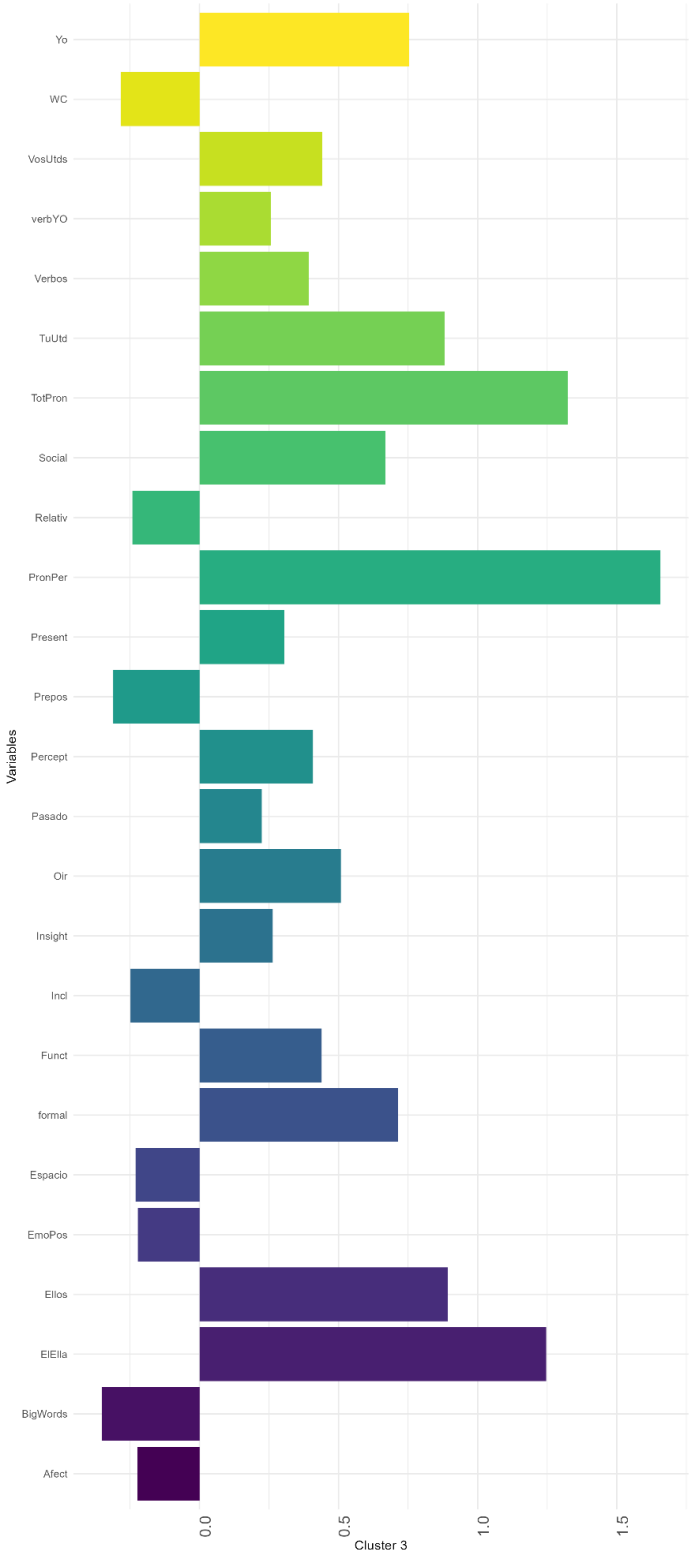


Figura 9
25 variables con mayor valor absoluto del clúster 4

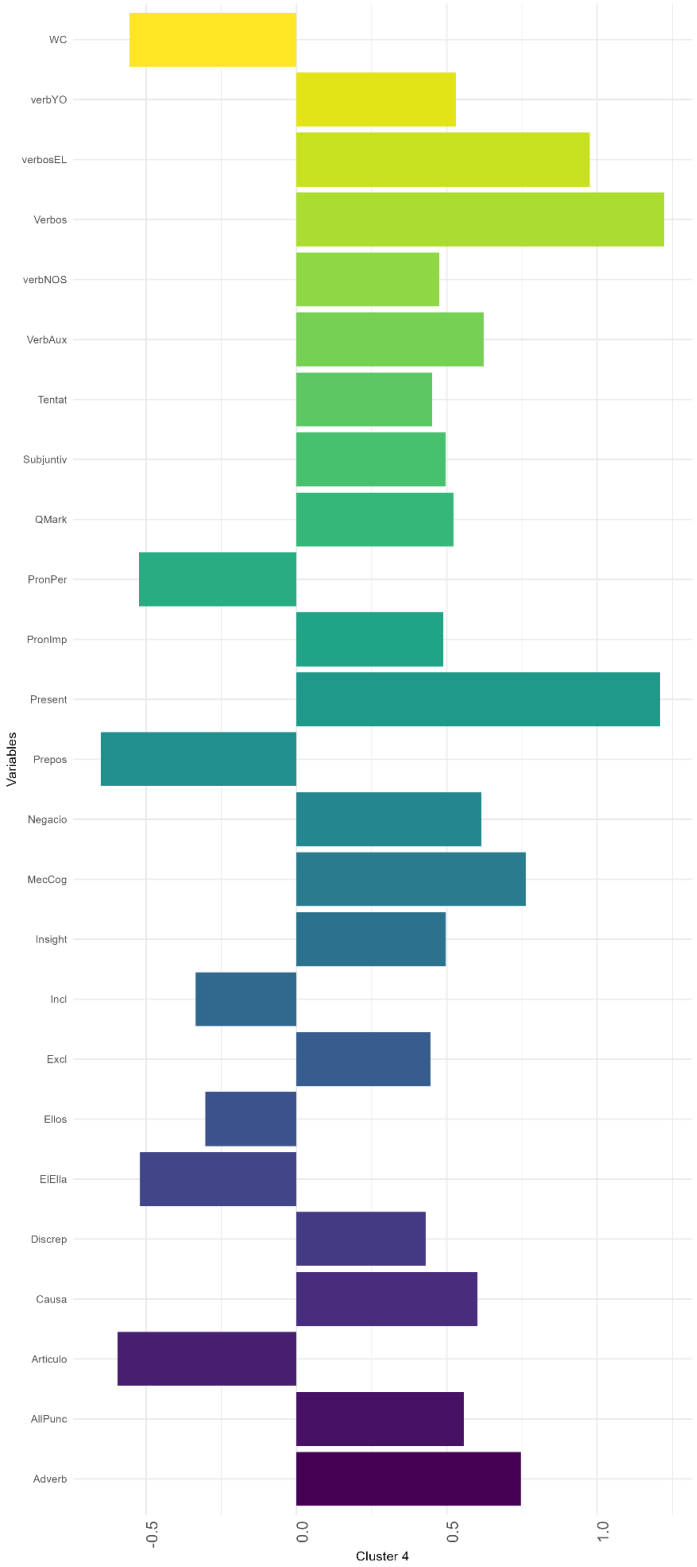
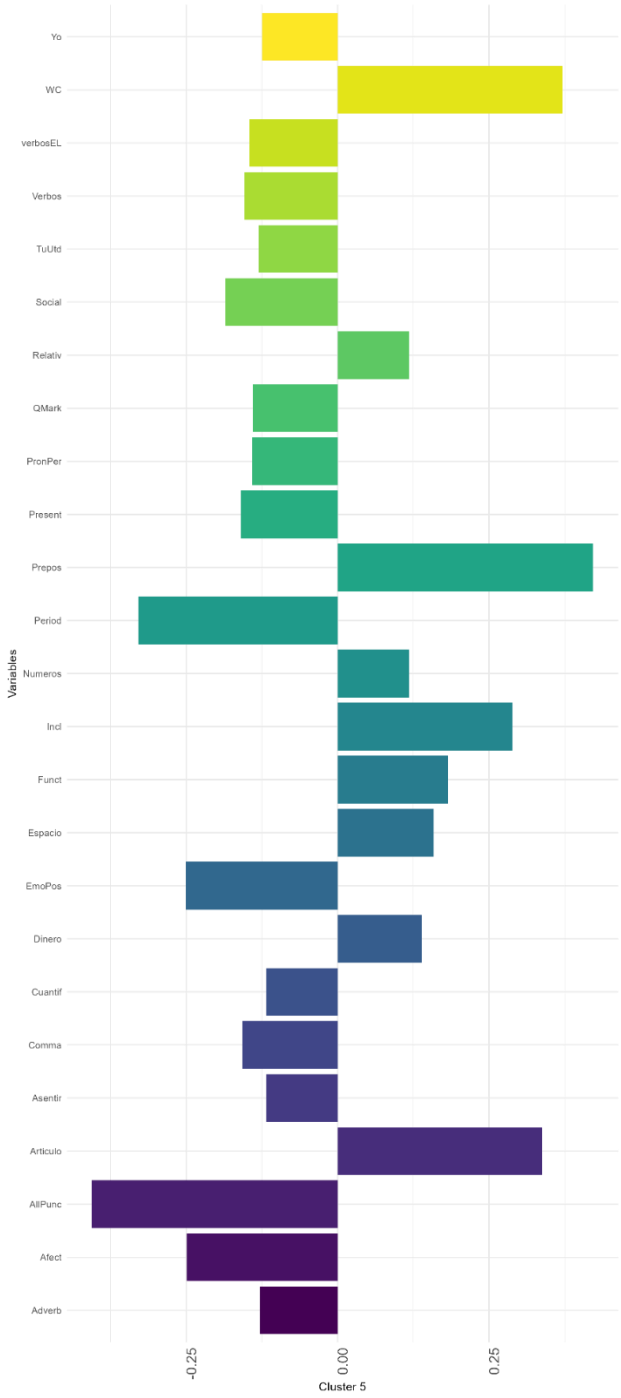


Figura 10
25 variables con mayor valor absoluto del clúster 5



Distribuciones de frecuencias de los clústeres y temas en el corpus

Cuadro 3
Distribución de frecuencias de estrategias retóricas en el corpus

Clúster	Frecuencia absoluta
Fragmentado	4662
Corporal	703
Interpelativo	5856
Declaratorio	5068
Informativo	22722

Cuadro 4
Distribución de frecuencias de temas en el corpus

Tema	Frecuencia absoluta
Poderes de la República	6692
Acción	7466
Consulta	3415
Cuantificación	4614
Formalismo	6637
Poder Ejecutivo	3647
País	6540