



“Sesgo de Género en los Medios Virtuales al reportar Diputados y Diputadas Chilenos”

**TESIS PARA OPTAR AL GRADO DE
Magister en Análisis Económico**

**Alumno: Juan Eduardo Concha Olivares
Profesor Guía: Francisco Pino, Valentina Paredes**

Santiago, diciembre 2020

Sesgo de Género en los Medios Virtuales al reportar Diputados y Diputadas Chilenos

Juan E. Concha

Profesores Guía: Francisco Pino y Valentina Paredes

23 de septiembre, 2019

Abstract

Are there any differences in the treatment male and female deputies receive on virtual media, that might lead to think there's media bias? This paper analyses differences in number of news pieces, number of words and use of words in reports of deputies from the 2014-2018 period. Using a novel database that consists of 6,079 news pieces gathered from virtual media, it is found that there's a tendency to find significant differences in the report of female deputies: between 15.5% and 27.8% less words and a negative, but non-significant effect on number of news pieces. Also, there's word use differentiation, with a strong tendency to associate male deputies with agentic topics, while female deputies are associated to communal topics. This opens several investigation routes that could lead to untangle the causality of these results

Resumen

¿Existen diferencias en el trato que reciben diputados y diputadas en los medios virtuales que puedan llevar a pensar que existe un sesgo en los medios? Este trabajo analiza las diferencias en la frecuencia de noticias, número de palabras y uso de palabras en el reporte de los diputados del período 2014-2018. Utilizando un set de datos novedoso que consiste de 6.079 noticias recogidas de medios virtuales se encuentra que tienden a existir diferencias significativas en reportes a las diputadas: entre un 15,5% y un 27,8% menos palabras y un efecto negativo, pero no significativo en el número de noticias. Además, hay una diferenciación en el uso de las palabras, con una fuerte tendencia a asociar diputados a temas agénticos y a las diputadas con temas comunales. Esto abre varias rutas para continuar investigando y así comenzar a desentrañar las causalidades de estos resultados.

Keywords: Gender, Political Economy, Elections, Media

JEL codes: L82, J16, P16, C88

Highlights

- Se encuentra que diputadas reciben entre un 15,5% y un 27,8% menos de palabras por noticia.
- Se encuentra efecto negativo, pero no significativo en el número de noticias por diputada.
- Se encuentra un uso diferenciado de las palabras en noticias de diputados y de diputadas.
- Se proponen canales para la causalidad de los resultados.

I. Introducción

Los medios y su efecto en el proceso de decisiones de las personas han sido sujeto de estudio de la Economía hace ya bastante tiempo. Entre los casos que han sido estudiados por la disciplina se encuentra el impacto de la campaña televisiva del No en el plebiscito de 1988 en Chile (González & Prem (2018)), el efecto de la entrada de periódicos en la participación electoral (Gentzkow et al (2011), Gentzkow (2006)), el efecto de la radio en el gasto público (Strömberg (2004)) e incluso el efecto que tuvo mostrar núcleos familiares pequeños de las telenovelas en las tasas de fertilidad de sectores de Brasil (La Ferrara et al (2012)).

Así, si los medios tienen la capacidad de alterar las decisiones, también nace el riesgo de que intenten manipularlas. Este es el caso de los medios sesgados, los que, además de su misión básica de informar, también suman la de persuadir. Este sesgo puede ser explícito, en el caso de un medio que busque maximizar una renta política, además de la renta económica (Kaplan & DellaVigna, 2007), o bien puede ser implícito y provenir, por ejemplo, de la subjetividad de los periodistas (Baron, 2006).

Considerando que la mayor parte de la literatura se ha enfocado en el posible sesgo político en los medios, existe espacio disponible para analizar una segunda posibilidad: la existencia de sesgo de género en los medios. Así, este trabajo busca explorar las diferencias de género que se pueden encontrar en los medios virtuales en el reporte de los diputados del período legislativo LIV de Chile (2014-2018), medido a través del número de noticias por diputado, el número de palabras por noticia, el uso de palabras y las temáticas asociadas a estas.

La hipótesis subyacente es que, si existen diferencias en estas variables, esto podría estar ligado a un sesgo implícito en los medios. Particularmente, el sesgo correspondería a diferencias entre la productividad de un diputado/a y el reporte efectivo que este recibe. De esta manera, si existe una diferencia sistemática entre la “productividad” de diputados de uno u otro género y el nivel de reportaje asociado, se estaría hablando de sesgo. Como medida base de productividad, se utilizaron los minutos hablados en intervenciones en Sala, además del número de intervenciones por diputado/a. Al analizar estos valores, se encuentra que existe una diferencia positiva, pero no significativas tanto en el tiempo de intervención per cápita de las diputadas con respecto a los diputados, como en el número de intervenciones per cápita (Anexo 1).

Para esta aproximación inicial se seleccionan noticias de Emol y La Tercera, dado que son los dos medios virtuales más grandes que pertenecen a dueños distintos (Medios Digitales (2015)). Junto a esto, se decide limitar inicialmente el análisis sólo a diputados debido a que son un grupo grande (120 en total) con tareas relativamente homogéneas relativo a otros grupos como los

ministros que tienen niveles muy distintos de relevancia mediática, por la naturaleza de sus actividades.

Utilizando una base de datos novedosa y aplicando metodologías de Procesamiento de Lenguaje Natural se logra realizar tanto un análisis de número de noticias como de palabras por noticia para cada diputado/a, encontrando diferencias sistemáticas, particularmente en el número de palabras por noticias. Además, el texto de las noticias se vectoriza para ajustar un modelo Logit-Ridge que permite identificar las palabras que tienen mayor poder predictivo para el género del diputado/a asociado a cada noticia. Finalmente, se analizan las diferencias en las palabras que son más predictivas para cada género.

Lo que queda del trabajo se estructura de la siguiente manera: en la sección II se verá la Literatura previa tanto de otras Ciencias Sociales, como el principal trabajo económico que motiva este paper. En la sección III se presentan los Datos. En la sección IV se repasa la metodología. En la sección V se presentan los resultados y, finalmente, se dan las conclusiones en la sección VI.

II. Literatura Previa

Para esta sección, inicialmente se repasarán trabajos que específicamente se refieran al sesgo de género en los medios desde las ciencias sociales, dado que la Economía no ha desarrollado este tema extensivamente. Luego se realizará un breve repaso de la literatura económica del sesgo en los medios y finalmente se revisará el trabajo económico que más inspira este trabajo.

Primero, desde la Psicología, Sendén et al (2015), hacen análisis de sentimiento sobre 400.000 noticias de Reuters, donde se busca ver la diferencia entre cómo se utilizan los pronombres “Él” y “Ella” en los textos. Los autores encuentran inicialmente que el pronombre ‘Él’ se utiliza 9 veces más seguido y en una segunda etapa utilizan una base llamada Affective Norms for English Words (ANEW), que rankea del 1 al 9 un listado de 1.034 palabras según la positividad del sentimiento que expresan. Luego, los autores vectorizan las palabras de Reuters en un espacio semántico multidimensional, lo que les permite calcular distancias entre palabras. Habiendo hecho esto, predicen su nivel de sentimiento usándolas como regresores de las palabras de ANEW. Finalmente, computan el nivel de sentimiento predicho de las 15 palabras anteriores y

posteriores a “Él” y a “Ella”. En base a esto, los autores encuentran que “Él” se usa más positivamente que “Ella” de manera estadísticamente significativa.

Desde la Ciencia de la Comunicación, Heldman, C., Carroll, S. J., & Olson, S. (2005) realizan un estudio cualitativo sobre el reporte de los medios respecto a la campaña de Elizabeth Dole, una de las contendientes a ser la candidata presidencial Republicana para las elecciones de 1999. Los autores encuentran que, a pesar de ser la segunda favorita en las encuestas, el nivel de cobertura recibido no fue consistente con esto. Además, se encuentra que la cobertura fue diferente a sus pares de género masculino, dando más atención a su personalidad y apariencia física.

Luego, Devitt (1999), desde la Ciencia de la Comunicación y Kahn (1996) desde la Ciencia Política, realizan trabajos más exhaustivos, donde se analiza el nivel y tipo de cobertura que reciben candidatas a Diputadas y Senadoras. Las conclusiones a las que llegan son similares: aun controlando por nivel de competencia e incumbencia, las candidatas reciben menor cobertura (no así las candidatas a alcaldesa y gobernadora) y hay un mayor foco en las responsabilidades familiares de las candidatas.

En lo que respecta a la literatura económica del sesgo en los medios, esta es abundante y considera diversas definiciones de sesgo. Anderson et al (2016) sistematizan esta literatura de la siguiente manera: una primera separación es según si el sesgo es explícito o implícito.

El primero es el más directo, pues tiene que ver con la postura que mantienen abiertamente ciertos medios. Dada la naturaleza de esta investigación, no se hará mayor énfasis en este tipo de sesgo, pues no existe medio que se declare abiertamente a favor o en contra de uno u otro género.

El sesgo implícito, en cambio, es uno que se puede inferir de la conducta de los medios de prensa, a pesar de que éstos no estén buscando ser abiertamente sesgados y se subdivide en tres tipos: comparación, intensidad de temas y tono. El primero se refiere a cómo se compara el léxico de los medios con otro *benchmark* político, el que usualmente es el vocabulario de los políticos o bien el de revistas partisanas. El segundo se refiere a la capacidad de definir la agenda de los medios, y se basa en la idea de que la cantidad de cobertura que los medios le dan a un tema puede influenciar la importancia que los lectores le dan a ese tópico. Finalmente, el último se basa en la idea de que un medio está sesgado a favor de un partido o posición política si sistemáticamente muestra a este partido o posición de manera favorable y/o muestra al partido o posición opositor de manera negativa.

El tipo de sesgo más relevante para este trabajo será, el de intensidad de temas, puesto que el efecto más probable es que los medios retraten a diputados y diputadas de maneras diferentes,

ya sea por un número distinto de palabras o noticias o bien por un uso distinto del lenguaje para cada género.

Finalmente, desde la Economía, existe amplia literatura de sesgo político en los medios y del efecto del género en la política. El principal motivador de este trabajo es Kaplan & DellaVigna (2007), donde estiman el impacto de la entrada de Fox News, un medio abiertamente Republicano, sobre el porcentaje de votos a favor del candidato Republicano de 2004: George W. Bush. Los autores encuentran un efecto positivo, significativo y robusto a una serie de tests, que, de acuerdo con sus estimaciones, generó veinte veces la cantidad de votos necesaria para cambiar los resultados de la elección. Esto debido a que Florida fue un *swing state* y la votación se decidió por un margen de 537 votos. Esto demuestra que los medios son efectivamente capaces de alterar opiniones y, en consecuencia, alterar elecciones, por lo que es de suma relevancia explorar si los medios pudieran estar afectando la percepción que las personas tienen de los diputados debido a diferencias en el reporte asociadas al género de estos.

Desde Chile, Navia et al (2013) evalúan si existe tanto sesgo de género en el reporte de los primeros 100 días de gobierno de Michelle Bachelet (2006-2010) y si existe sesgo político para los primeros 100 días del gobierno de Sebastián Piñera (2010-2014) de parte los medios de prensa escrita *El Mercurio* y *La Tercera*. Los autores finalmente encuentran sesgo político a favor de Piñera, pero no encuentran sesgo de género en contra de Bachelet.

En términos del efecto del género en las políticas públicas, existe una amplia literatura económica que comienza con Chattopadhyay y Duflo (2004), donde estudian el impacto del uso de cuotas de género para el Concejo del Pueblo: desde 1993, aleatoriamente, un tercio de los Concejos deben ser presididos por mujeres, experimento natural que aprovechan los investigadores. Finalmente, encuentran que efectivamente los líderes del Concejo del Pueblo tienden a desarrollar políticas más asociadas a las necesidades del género de quién lo preside. Además, trabajos como el de Brollo & Troiano (2016) abordan y encuentran para Brasil un efecto positivo de la elección de una mujer como alcaldesa, disminuyendo la corrupción, aumentando los resultados de salud y recibiendo más transferencias federales. En cambio, Ferreira & Gyourko (2014) no logran encontrar diferencias significativas entre alcaldes hombres y mujeres en una serie de medidas: tamaño del gobierno, tasas de criminalidad y composición del gasto municipal. Cabe mencionar, además, que estos autores encuentran resultados opuestos, a pesar de utilizar la misma estrategia de identificación: hacer regresión discontinua sobre elecciones reñidas entre hombres y mujeres, lo que permite asumir que en el margen los candidatos son iguales, a excepción de su género.

Como se vio en los párrafos anteriores, si bien la evidencia económica respecto al impacto del género de los políticos sobre los resultados de política pública no es concluyente, sí existe importante evidencia del impacto que puede tener un medio sesgado en el comportamiento de los votantes. Así, si en la dimensión política izquierda-derecha se puede generar cierta manipulación, cabe preguntar si en otra dimensión los medios pueden impactar el resultado de las votaciones. Es por esto que nace la idea de investigar cómo se diferencia el reporte de hombres y mujeres del mundo de la política en los medios virtuales, de manera de tener un acercamiento inicial a posibles investigaciones respecto a esta temática en la prensa chilena.

III. Datos

Los datos utilizados en este trabajo se obtienen a partir de un proceso de varias etapas. En un paso inicial, un *web scraper* (algoritmo diseñado para navegar por internet) busca el nombre de cada uno de los 120 diputados del período 2014-2018 (101 hombres y 19 mujeres) en los portales de noticias de Emol y La Tercera y obtiene el link de todas las noticias arrojadas por estos.

Luego, entra a cada uno de los links y en una base de datos se guardan el titular, bajada de título, cuerpo, fecha y medio de origen de cada una de las 14.064 noticias. Junto a esto, se limitan las noticias a aquellas que hayan sido publicadas entre enero de 2014 y marzo de 2018, acaparando una parte importante de la campaña y el período completo de ejercicio de los diputados. Además, se eliminan casos inesperados como noticias sin cuerpo. En este punto, la muestra se reduce a 6.079 noticias y se le llamará “base bruta” de ahora en adelante.

No obstante, esta técnica de recolección de datos tiene dos principales limitantes: dadas las funcionalidades del buscador de noticias de *La Tercera*, sólo es posible obtener hasta 9 páginas de noticias para cada diputado y, junto a esto, no tiene posibilidad de seleccionar fechas de búsqueda. Estos dos problemas, llevan a que para diputados que han sido mediáticos entre abril de 2018 y el presente, la gran mayoría de las noticias recolectadas pertenezcan a este período. Al poder recolectar un número limitado de noticias, esto lleva a que queden subrepresentados o derechamente desaparezcan de la muestra. Ejemplos de este problema son Ignacio Urrutia, Karla Rubilar y Maya Fernández, donde se pierden todas sus noticias y Fuad Chahin, Marcela Sabat, Leonardo Soto, Nicolás Monckeberg, Karol Cariola, Daniel Núñez Vlado Mirosevic, Javiera Macaya, Jaime Bellolio y Paulina Núñez, donde hay pérdidas superiores al 80%.

Al no contar con la población completa de noticias del período para algunos diputados en La Tercera y considerando que las noticias se presentan en orden de relevancia o, en otras palabras,

de manera no aleatoria, podría existir algo de sesgo de selección en las noticias de estos diputados/as, en la medida en que la relevancia tuviera relación con las variables dependientes o con el género. No obstante, no es evidente un canal sobre el que la relevancia, medida a través del número de clicks en la noticia, pudiera correlacionar con las variables de interés. Además, considerando que este problema se da para un subconjunto de los diputados y en un solo periódico, en caso de existir un problema de sesgo de selección, este debiese ser menor.

Además, al analizar con mayor detalle los resultados de las regresiones obtenidas utilizando datos de cada periódico por separado, se obtienen resultados altamente similares (Tablas A1, A2, A3 y A4), lo que lleva a inducir que el truncamiento de la selección de noticias de algunos diputados/as no está generando un problema mayor.

Por último, existen dos niveles de filtro extra, que se aplican según el análisis a realizar. El primero, limita la muestra a noticias que mencionen a diputados de un solo género, lo que elimina el ruido que introducen las noticias que mencionan a diputados de ambos géneros, dificultando un análisis más preciso sobre el trato del género de cada diputado. El segundo nivel, además limita de manera que se mencione exclusivamente a un diputado por noticia, mejorando el análisis respecto al número de noticias por diputado, ya que el nivel de filtro anterior sesga los resultados en contra de las diputadas, debido a que, dado que son una proporción bastante menor a los diputados, por construcción es menos probable que una noticia que mencione a más de un diputado mencione exclusivamente a mujeres.

Preprocesamiento

Teniendo la muestra definida, un paso importante es preprocesar el texto. Considerando la naturaleza de alta dimensionalidad de los datos de texto, se han desarrollado una serie de prácticas estándar que permiten reducirla de manera importante (Gentzkow et al (2017)). Los pasos de procesamiento aplicados son:

- Eliminación de la puntuación, dígitos y tildes.
- Eliminación de palabras de 3 o menos letras.
- Eliminación de *stopwords*: palabras muy comunes que no entregan información, como “cuando” o “porque”.
- Eliminación de las menciones a nombres de diputados: revelan a qué género corresponde la noticia.

- *Stemming* de las palabras: son reducidas a su raíz, de modo que, por ejemplo, “diputados”, “diputadas”, “diputado” y “diputadas” son todas transformadas a “diput”. Esto cumple una doble función: disminuir la dimensionalidad del problema al reducir el número de palabras diferentes y, además, vuelve el vocabulario neutro en términos de género, de manera que las noticias no sean trivialmente fáciles de predecir con la metodología que se detallará en la siguiente sección.

Habiendo hecho esto, la dimensionalidad de la base bruta se reduce de un total de 2.760.533 palabras, con 57.005 palabras diferentes a 1.052.810 palabras totales con 18.648 palabras diferentes.

Finalmente, la base con datos de las noticias se une a través del nombre del diputado con otra base que contiene más información de estos: partido político, tendencia política, coalición, género, número de períodos en el Congreso, edad y una variable *dummy* para Presidencia de la Cámara de Diputados.

Por último, es relevante mostrar información básica respecto a los distintos niveles de bases de datos:

Tabla 1: Estadística descriptiva básica de los tres niveles de base de datos

Base de datos	Número de noticias	Número de diputados
Bruta	6093	120
1 género por noticia	4290	120
1 diputado por noticia	2418	115

Para el nivel más exigente de filtro, no existen noticias para Clemira Pacheco, Juan Morano, Pedro Pablo Álvarez, René Manuel García y Roberto Poblete. Además, el paso de la base bruta a la base de 1 género por noticia representa una disminución del 29,3% en el número de noticias, mientras que el paso a 1 diputado por noticia representa una disminución del 61,8%.

Tabla 2a: Número medio de palabras y número de noticias por diario

Diario	Media de n° de palabras por noticia	N° de noticias
Emol	344.34	1577
La Tercera	485.77	848

También resulta relevante el comportamiento de otras variables como el periódico de origen, el partido político o el año, lo que es posible ver en las Tablas 2a, 2b y 2c.

Para los diarios, en la Tabla 2a se observa que las noticias en el diario La Tercera tienden a ser más largas, pero representan una proporción menor del total.

Tabla 2b: Número medio de palabras y número de noticias por año

Año	Nº de palabras	Nº de noticias
2014	349.17	329
2015	345.57	332
2016	386.69	463
2017	414.69	902
2018	431.75	399

En lo que respecta a los años, en la Tabla 2b se observa que una proporción muy importante de las noticias se concentra en 2017 y 2018 (considerando que sólo se obtuvieron noticias hasta marzo). Esto tendría una explicación doble: por un lado, a medida que avanza la campaña para el siguiente período, efectivamente es de esperar un aumento significativo en el flujo de noticias. No obstante, la limitación que presentó el portal de La Tercera a la hora de obtener las noticias también genera una tendencia hacia el presente, como se comprueba en el Anexo 2. Además, se observa que el largo de las noticias tiende a aumentar a medida que pasa el tiempo, aunque se esté mencionando a sólo un diputado/a por noticia.

Finalmente, en la Tabla 2c se compara el porcentaje de noticias dedicadas a diputados de cada partido y el porcentaje de los diputados que pertenecen a cada uno de estos. En términos generales, no hay grandes desviaciones entre el porcentaje de noticias y diputados para cada partido, posiblemente con la excepción de Evópoli, representado sólo por Felipe Kast. Además, llama la atención que Frente Amplio es el partido al que se le dedican menos palabras, mientras que el Partido Socialista y el Partido Por la Democracia son los partidos con más de un diputado que se les asigna más.

Tabla 2c: Número medio de palabras y porcentaje de noticias y diputados de cada partido

Partido	N° de palabras	% de noticias	% de diputados
UDI	367.98	27.2%	24.3%
RN	384.13	13.3%	12.2%
PDC	412.34	13.0%	16.5%
PC	354.82	9.9%	5.2%
PPD	419.92	9.3%	13.0%
PS	424.72	9.0%	13.0%
IND	420.20	7.3%	4.3%
PRSD	390.16	4.1%	5.2%
Evópoli	466.29	3.1%	0.9%
AMP	337.56	1.6%	1.7%
IND-PDC	499.38	0.9%	0.9%
IC	385.76	0.7%	0.9%
PL	468.60	0.4%	0.9%
MIRAS	476.00	0.2%	0.9%

3.1 Estadística Descriptiva

3.2

Tabla 3: Estadística descriptiva básica para la base de datos bruta.

Estadístico	Hombre	Mujer	Diferencia
Número de noticias	5065	892	4173*** (0.00)
Noticias per capita	50.15	46.95	3.20 (0.73)
Media de palabras por noticia	482.45	457.93	24.52 (0.06)
Std de palabras por noticia	369.34	330.46	38.88 (0.75)
Mínimo de palabras por noticia	61	58	3
Percentil 25	273	258	15
Percentil 50	386	362.5	23.5
Percentil 75	567	542.25	24.75
Máximo	4411	3143	1268

Nota: estadísticos fueron calculados utilizando la base de datos bruta. Valores p entre paréntesis. *** p<0.001 ; ** p<0.01; * p<0.05

Resulta interesante tener una idea de qué es posible encontrar en los datos. En primera instancia, al analizar la base bruta, que incluye noticias en que son mencionados diputados de ambos géneros, en la Tabla 2 es posible notar que la magnitud de las diferencias es relativamente baja. De hecho, al hacer test-t sobre la media de palabras por noticia y las noticias per cápita, no se encuentran diferencias significativas. Las diferencias en número de noticias totales son mecánicas y responden a la diferencia en la proporción de diputados y diputadas en el Congreso. No obstante, las noticias que mencionan hombres en todas las medidas tienen valores mayores a las que mencionan mujeres.

Sin embargo, al refinar el nivel de filtro a noticias que mencionan diputados/as de un solo género, las diferencias se vuelven más notorias.

En la Tabla 3, se ve que el número de noticias per cápita es más del doble para los hombres y, en promedio, se les dedica 17% más de palabras por noticia. Ambas diferencias son significativas.

Tabla 4: Estadística descriptiva básica para la base de datos con noticias que mencionan diputados de un solo género.

Estadístico	Hombre	Mujer	Diferencia
Número de noticias	3939	358	3581*** (0.00)
Noticias per capita	39.00	18.84	20.16* (0.011)
Media de palabras por noticia	449.16	371.23	77.93*** (0.00)
Std de palabras por noticia	341.38	267.69	73.69 (0.35)
Mínimo de palabras por noticia	61	67	-6
Percentil 25	261	214.25	46.75
Percentil 50	364	309.5	54.5
Percentil 75	523	427	96
Máximo	4411	3143	1268

Nota: estadísticos fueron calculados utilizando la base de datos con un género por noticia. Valores p entre paréntesis. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Un punto que considerar de estas estadísticas es que, dado que existen 5.3 veces más diputados que diputadas, aunque los diputados a mencionar en una noticia fueran seleccionados aleatoriamente, la probabilidad de que la noticia mencione exclusivamente hombres es bastante elevada, aunque no existiera ningún sesgo. Por lo mismo, se hace un último análisis para noticias que mencionan sólo un diputado.

En este caso, si bien los hombres siguen teniendo más noticias per cápita y palabras asignadas por noticia a lo largo de la distribución, se observa que para las noticias per cápita la diferencia no es significativa. En todos los casos analizados, además, el número de palabras de mujeres está sujeto a una varianza menor, a pesar de contar con menos observaciones.

Tabla 5: Estadística descriptiva básica para la base de datos de noticias que mencionan un solo diputado por noticia.

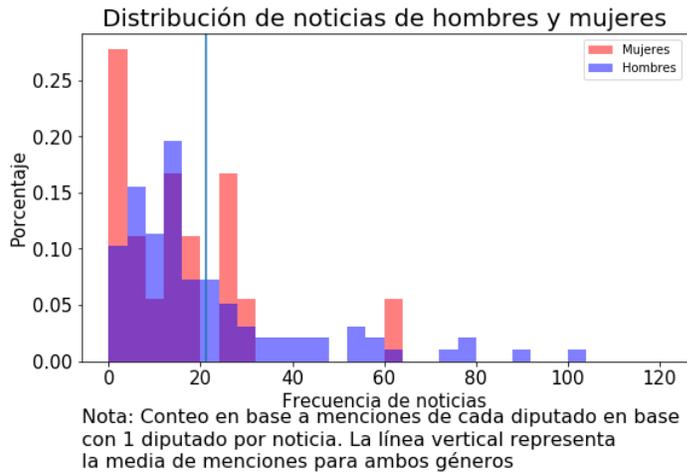
Estadístico	Hombre	Mujer	Diferencia
Número de noticias	2158	267	1891 (0.00)
Noticias per capita	21.37	14.05	7.31 (0.22)
Media de palabras por noticia	398.66	354.52	44.13* (0.02)
Std de palabras por noticia	294.71	224.56	70.15 (0.84)
Mínimo de palabras por noticia	61	67	-6
Percentil 25	235	208.5	26.5
Percentil 50	323	301	22
Percentil 75	461.75	424	37.75
Máximo	3414	1730	1684

Nota: estadísticos fueron calculados utilizando la base de datos con un diputado por noticia. Valores p entre paréntesis. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Considerando lo anterior, cabe entrar al detalle de la distribución de las noticias de cada diputado/a. En este caso, se limitará el análisis para la base que considera noticias que mencionan a un solo diputado/a.

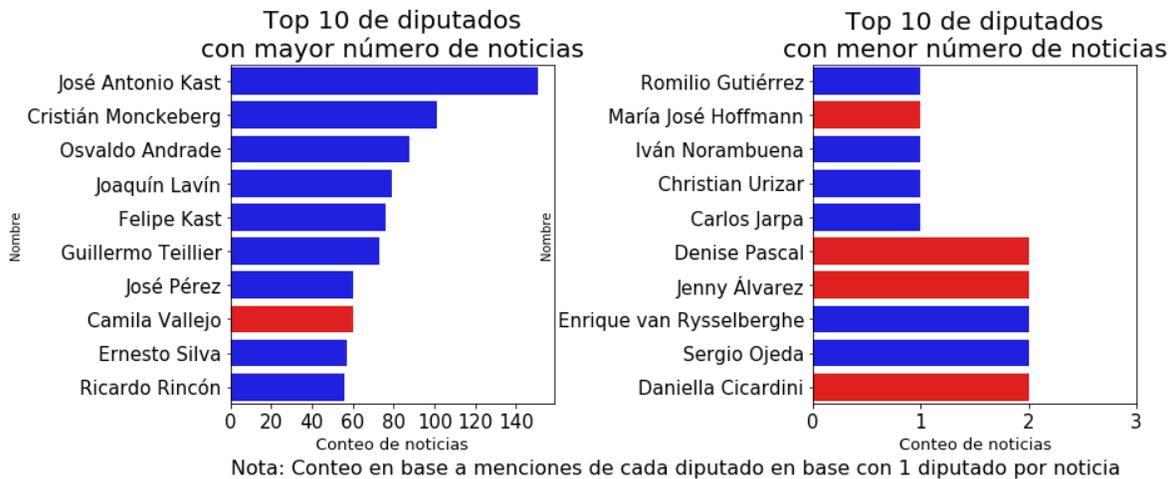
La Figura 1 muestra la distribución del número de noticias dedicado a cada diputado/a en la base con 1 diputado/a por noticia, separado por género. Se aprecia que una porción muy importante de las mujeres recibe un número de noticias menor a la media (línea vertical) y que, además, estas están bastante más concentradas.

Figura 1: Distribución del número de noticias de hombres y mujeres.



Luego, en la Figura 2 es posible ver los 10 diputados/as con mayor y menor número de noticias. Efectivamente se comprueba que el top 10 está dominado por hombres, con la sola excepción de Camila Vallejo.

Figura 2: 10 diputados/as con mayor y menor número de noticias



IV. Metodología

En este trabajo, la metodología se divide en dos: en una primera etapa, el análisis será extensivo, con la idea de testear si existen diferencias tanto en el número de palabras que se le dedica a cada diputado/a como en el número de noticias para tener una idea de si existen diferencias en

el interés periodístico que generan los diputados vs las diputadas a nivel particular y agregado, respectivamente.

En segunda instancia, se realiza un análisis en un cuanto al uso de palabras: se busca encontrar qué palabras son más predictivas del género del diputado al que se refiere una noticia. Esto se realiza con la intención de recolectar evidencia que pudiera sustentar la hipótesis de que las diputadas y los diputados podrían tener un trato distinto por la prensa y si es así, poder plantear distintas explicaciones para estas diferencias.

Número de palabras y noticias

Para identificar diferencias en número de palabras y noticias, la estrategia es sencilla. Se plantean las siguientes regresiones:

$$y_{n,d,t,p} = \beta_M \text{Mujer} + \beta_1 X_{d,t,p} + \epsilon_{n,d,t,p}$$

$$y_{n,d,t,p} = \beta_{M_1} \text{Mujer} + \beta_{M_2} \text{Mujer} * \text{LaTercera} + \beta_1 X_{d,t,p} + \epsilon_{n,d,t,p}$$

Donde $y_{n,d}$ representa el número de palabras en la noticia n , para el diputado d , en el año t y el periódico p , Mujer es una dummy que toma valor 1 si el género del diputado/a mencionado en la noticia es mujer, La Tercera es una dummy que toma valor 1 si el periódico es La Tercera y 0 si es Emol y $X_{d,t,p}$ son los controles, que consideran: el número de minutos que el diputado/a intervino en las sesiones de la Cámara en el último mes (*proxy* de productividad del diputado/a), el número de períodos en la Cámara del diputado d , si es Presidente/a de la Cámara, el partido del diputado, número de comisiones en las que participa, edad y, por último, el periódico de origen de la noticia.

Para el número de noticias por diputado, la estrategia es prácticamente la misma:

$$y_{d,t,p} = \beta_M \text{Mujer} + \beta_1 X_{d,t,p} + \epsilon_{d,t,p}$$

$$y_{d,t,p} = \beta_{M_1} \text{Mujer} + \beta_{M_2} \text{Mujer} * \text{LaTercera} + \beta_1 X_{d,t,p} + \epsilon_{d,t,p}$$

Siendo la única diferencia que ahora $y_{d,t,p}$ representa el número de noticias que recibió el diputado/a d , en el año t , en el periódico p .

Uso de palabras

Para explorar qué palabras son más predictivas del género del diputado mencionado en una noticia, la metodología será distinta. En este caso se aplica Procesamiento de Lenguaje Natural al texto de cada una de las noticias, basado en la metodología propuesta por Gentzkow, Kelly & Taddy (2017), para predecir en base a la distribución de las palabras.

La idea es la siguiente: posterior al preprocesamiento del texto, se mapea un *corpus* (conjunto de documentos) D , compuesto por n noticias $\{D_i\}$ en una matriz numérica C , para predecir una variable \hat{V} , que en este caso corresponde al género del diputado asociado a cada noticia. Cabe mencionar que, para este ejercicio se trabaja con la base filtrada tal que se mencionen diputados de un solo género. La matriz está compuesta de n filas c_i , y p columnas, representando cada una un *token*.

Los *tokens* corresponden a la unidad de análisis elegida por el investigador, tales como caracteres, palabras, oraciones, etc. En este caso, los *tokens* serán tanto las palabras como los bigramas. En otras palabras, la matriz C tendrá una columna por cada palabra única utilizada y por cada combinación de dos palabras consecutivas que existan a lo largo de todos los documentos.

Por último, cada celda de la matriz C , está dada por una función $f(c_i)$, que en este caso corresponde simplemente al número de apariciones de un token en la noticia $\{D_i\}$. A continuación, un ejemplo que permite aterrizar el concepto y comprender el impacto del preproceso del texto.

Se tiene un corpus D , que contiene dos noticias: $\{D_1\}$ y $\{D_2\}$, donde la primera consiste en el texto “La diputada socialista” y la segunda en “El diputado de Renovación Nacional”. Al vectorizar ambas noticias, se genera la siguiente matriz C_1 de dimensiones 2×8 :

Tabla 6: Ejemplo de matriz C antes de preprocesar el texto

Noticia	<u>La</u>	<u>diputada</u>	<u>socialista</u>	<u>El</u>	<u>diputado</u>	<u>de</u>	<u>Renovación</u>	<u>Nacional</u>
$\{D_1\}$	1	1	1	0	0	0	0	0
$\{D_2\}$	0	0	0	1	1	1	1	1

Evidentemente, considerar todas las palabras de todas las noticias llevaría a una matriz C con un número muy elevado de columnas. Es por esto mismo que el preprocesamiento es clave: la dimensionalidad del problema se reduce considerablemente y en este caso los textos pasarían a

ser $\{D_1\}$ = “diput socialist” y $\{D_2\}$ = “diput renov nacion”, lo que lleva a la siguiente matriz C_2 de dimensiones 2×4 :

Tabla 7: Ejemplo de matriz C luego de preprocesar el texto.

Noticia	diput	socialist	renov	nacion
$\{D_1\}$	1	1	0	0
$\{D_2\}$	1	0	1	1

Logrando una reducción efectiva del 50% en el número de celdas. En la práctica, al preprocesamiento se le suman dos restricciones: cada *token* debe aparecer al menos 3 veces a lo largo de todos los documentos y puede aparecer máximo en un 98% de los documentos. Habiendo hecho esto, la matriz C se reduce de 416.170 columnas a 47.481, lo que representa una mejora del 88,6%.

Una vez preparadas la matriz C y el vector V a predecir, se realiza un proceso de *cross-validation*, para poder predecir el género del diputado asociado a la noticia. En este caso, el proceso consiste en dividir el set de datos aleatoriamente, de manera de ajustar el modelo con un 85% de los datos y luego predecir sobre el 15% restante.

Esta metodología se usa debido a que los datos son de alta dimensionalidad, sobre todo considerando que $p > n$, se genera un sobreajuste del modelo, lo que implica que el modelo tiende a predecir muy bien dentro de la muestra, pero a cambio de una caída importante en el rendimiento para predecir fuera de muestra.

El modelo seleccionado para predecir es un Logit-Ridge, que se caracteriza por utilizar penalización ℓ_2 (Hoerl & Kennard (1970)). De esta manera, el modelo será:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \eta_i = \alpha + x_i'\beta$$

Donde p_i es la probabilidad de que una noticia esté asociada a una diputada y donde x_i es una transformación conocida de c_i , que en este caso es $x_i = c_i$. La diferencia con un modelo Logit tradicional, se encuentra en función de pérdida a minimizar:

$$\operatorname{argmin}_{\beta_0} \left\{ l(\alpha, \beta) + n\lambda \sum_{j=1}^p \beta_j^2 \right\}$$

Donde $l(\alpha, \beta)$ es la función de pérdida clásica de Logit y $n\lambda \sum_{j=1}^p \beta_j^2$ es el término de regularización que penaliza por desviaciones del cero de β_j . Esta penalización permite encontrar valores para los parámetros β_j a pesar de que son de mayor dimensión que el número de observaciones. De

esta manera, se elimina el problema de la econometría tradicional, donde, si el número de parámetros p es mayor al número de observaciones n , se obtienen $(p-n)$ parámetros que son combinaciones lineales de los primeros p parámetros, lo que a su vez lleva a un ajuste perfecto (implicando una muy mala generalización fuera de muestra).

V. Resultados

En la siguiente sección se presentarán los resultados de las regresiones descritas en la sección de Metodología, tanto para para la frecuencia de noticias y palabras, como para el uso de palabras.

Número de palabras y noticias

En la Tabla 8 se examina el efecto del género del diputado asociado a una noticia en el número de palabras de dicha noticia. En la columna 1, que considera noticias donde se menciona a diputados de un solo género, se observa que el hecho de que la noticia esté referida a diputadas se asocia a una disminución de 68.66 palabras, aun controlando por la productividad del diputado/a.

En la columna 2, se agrega la interacción entre género y diario, para aislar si el efecto está centrado en uno de los dos diarios analizados. Efectivamente se encuentra que una noticia referida a diputadas en La Tercera estaría asociada a una disminución adicional de 91.95 palabras, mientras que el coeficiente de Género disminuye su magnitud más de un 50% a -31.19 palabras por noticia.

No obstante, dado que existen 5.3 veces más diputados que diputadas, es posible que las noticias que mencionan sólo diputados contengan más personas que las que mencionan sólo diputadas y, por lo tanto, más palabras. Por lo mismo, en las columnas 3 y 4 se restringe la muestra sólo a noticias que mencionan exclusivamente un diputado o diputada. En el caso de la columna 3, aún se encuentra un efecto negativo y significativo de -50 palabras, no así en la columna 4, donde, si bien se mantiene el signo del efecto, este deja de ser significativo al 5%. Además, la interacción también es negativa y no significativa.

Tabla 8: Efecto del género y diario sobre el número de palabras en una noticia

	Un género por noticia		Un diputado por noticia	
	(1)	(2)	(1)	(2)
Número de períodos	2.535 (0.556)	2.26 (0.6)	0.948 (0.855)	0.91 (0.861)
Edad	-0.742 (0.168)	-0.674 (0.21)	-0.781 (0.285)	-0.723 (0.323)
Número de comisiones	-4.619 (0.347)	-4.392 (0.371)	1.037 (0.862)	1.205 (0.84)
Presidente_Camara	-8.974 (0.711)	-6.451 (0.79)	15.854 (0.636)	17.13 (0.609)
Minutos de intervención	0.0 (0.638)	0.0** (0.015)	-0.0 (0.129)	-0.0 (0.189)
Mujer	-68.658*** (0.0)	-31.191** (0.041)	-49.991*** (0.007)	-32.562* (0.076)
La Tercera	188.285*** (0.0)	195.795*** (0.0)	136.045*** (0.0)	141.25*** (0.0)
Mujer*LaTercera		-91.947** (0.01)		-47.439 (0.21)
Dummy Año	Sí	Sí	Sí	Sí
Dummy Partido	Sí	Sí	Sí	Sí
Observaciones	4298	4298	2430	2430
R ²	0.088	0.089	0.068	0.069

Notas: La tabla reporta estimaciones por OLS del efecto del género del diputado mencionado en una noticia y el periódico de origen de est sobre el número de palabras utilizadas en una noticia. *Número de períodos* corresponde a cuantos períodos ha estado en la Cámara el diputado previo a la elección de 2014, *Número de comisiones* corresponde al número de comisiones a las que pertenece el diputado. *Presidente_Camara* es una dummy que toma valor 1 si el diputado era Presidente de la Cámara en el año en que fue publicada la noticia. *Diario_La Tercera* es una dummy que toma valor 1 si la noticia proviene de La Tercera y 0 si proviene de Emol. La diferencia entre la especificación (1) y (2) es que la primera no considera el efecto interactivo entre diario y género. Se utilizan errores estándar robustos. Se reportan valores p entre paréntesis,

El efecto encontrado sería bastante considerable, teniendo en cuenta que el número promedio de palabras de una noticia en las muestras de las primeras dos columnas es de 443.22, y que la disminución en el número de palabras debido al género estaría entre -68.66 (columna 1) y -123.13 (columna 2, noticias sobre diputadas en La Tercera). Esto implicaría una disminución de palabras que representaría entre un 15,5% y un 27,8% de la media. Para el caso de las columnas

3 y 4, el número de palabras promedio desciende a 394.3, sin embargo, resulta más difícil llegar a conclusiones, debido a la falta de significancia de los resultados.

Tabla 9: Efecto del género y diario sobre el número de noticias que recibe un diputado

	Un género por noticia		Un diputado por noticia	
	(1)	(2)	(1)	(2)
Número de períodos	0.13 (0.529)	0.13 (0.527)	0.294 (0.2)	0.296 (0.199)
Edad	-0.032 (0.159)	-0.033 (0.157)	-0.013 (0.513)	-0.013 (0.509)
Número de comisiones	-0.792** (0.014)	-0.791** (0.014)	-0.849*** (0.001)	-0.848*** (0.001)
Presidente_Camara	10.142** (0.032)	10.136** (0.032)	7.572** (0.016)	7.559** (0.016)
Minutos de intervención	-0.0 (0.283)	0.0* (0.096)	-0.0 (0.151)	-0.0 (0.75)
Mujer	-2.726*** (0.0)	-3.177*** (0.001)	-0.805 (0.196)	-1.006 (0.242)
La Tercera	-2.197*** (0.003)	-2.315*** (0.003)	-2.256*** (0.0)	-2.312*** (0.001)
Mujer*LaTercera		1.01 (0.429)		0.459 (0.676)
Dummy Año	Sí	Sí	Sí	Sí
Dummy Partido	Sí	Sí	Sí	Sí
Observaciones	651	649	540	540
R ²	0.148	0.147	0.151	0.151

Notas: La tabla reporta estimaciones por OLS del efecto del género del diputado mencionado en una noticia y el periódico de origen de esta sobre el noticias publicadas en un año y diario dados. *Número de períodos* corresponde a cuantos períodos ha estado en la Cámara el diputado previo a la elección de 2014, *Número de comisiones* corresponde al número de comisiones a las que pertenece el diputado. *Presidente_Camara* es una dummy que toma valor 1 si el diputado era Presidente de la Cámara en el año en que fue publicada la noticia. *Diario_La Tercera* es una dummy que toma valor 1 si la noticia proviene de La Tercera y 0 si proviene de Emol. La diferencia entre la especificación (1) y (2) es que la primera no considera el efecto interactivo entre diario y género. Se utilizan errores estándar robustos. Se reportan valores p entre paréntesis,

En la Tabla 9, en cambio, se analiza el efecto del género de un diputado sobre el número de noticias escritas asociadas a este en un año y un periódico dados. En las noticias donde se menciona a uno o más diputados de un mismo género (columnas 1 y 2) se encuentra un efecto negativo y significativo en el número de palabras de entre -2.73 (columna 1) y -3.18 (columna 2),

lo que representaría una disminución de entre un 41.4% y un 48.3% con respecto a la media de 6.58 noticias por diputado al año, por diario. En cambio, en las columnas 3 y 4, si bien los efectos encontrados son negativos, esto no son significativos, lo que podría estar implicando que el efecto de las columnas 1 y 2 está dado principalmente por el hecho de que es más probable encontrar noticias que mencionen exclusivamente a hombres, por el hecho de que son una mayor proporción en el Congreso. Otro punto a considerar es que para las columnas 3 y 4 la muestra fue agrupada de manera de obtener el número de noticias por diputado, lo que lleva a una disminución de aproximadamente el 85% en el tamaño muestral y, por lo tanto, una pérdida considerable del poder de los test.

Tabla 10: Efecto del género y diario sobre el número de palabras en una noticia, errores estándar clusterizados a nivel diputado.

	Un género por noticia		Un diputado por noticia	
	(1)	(2)	(1)	(2)
Número de períodos	2.535 (0.644)	2.26 (0.684)	0.948 (0.878)	0.91 (0.883)
Edad	-0.742 (0.248)	-0.674 (0.296)	-0.781 (0.268)	-0.723 (0.311)
Número de comisiones	-4.619 (0.496)	-4.392 (0.524)	1.037 (0.883)	1.205 (0.865)
Presidente_Camara	-8.974 (0.691)	-6.451 (0.782)	15.854 (0.447)	17.13 (0.413)
Minutos de intervención	0.0 (0.47)	0.0*** (0.0)	-0.0*** (0.007)	-0.0** (0.02)
Mujer	-68.66*** (0.003)	-31.19 (0.114)	-49.99** (0.033)	-32.56 (0.159)
La Tercera	188.285*** (0.0)	195.795*** (0.0)	136.045*** (0.0)	141.25*** (0.0)
Mujer*LaTercera		-91.947** (0.016)		-47.439 (0.257)
Dummy Año	Sí	Sí	Sí	Sí
Dummy Partido	Sí	Sí	Sí	Sí
Observaciones	4298	4298	2430	2430
R ²	0.088	0.089	0.068	0.069

Notas: La tabla reporta estimaciones por OLS del efecto del género del diputado mencionado en una noticia y el periódico de origen de est sobre el número de palabras utilizadas en una noticia. *Número de períodos* corresponde a cuantos períodos ha estado en la Cámara el diputado previo a la elección de 2014, *Número de comisiones* corresponde al número de comisiones a las que pertenece el diputado. *Presidente_Camara* es una dummy que toma valor 1 si el diputado era Presidente de la Cámara en el año en que fue publicada la noticia. *LaTercera* es una dummy que toma valor 1 si la noticia proviene de La Tercera y 0 si proviene de Emol. La diferencia entre la especificación (1) y (2) es que la primera no considera el efecto interactivo entre diario y género. Se utilizan errores estándar clusterizados a nivel diputado. Se reportan valores p entre paréntesis.

Las Tablas 10 y 11 reportan los mismos resultados de las Tablas 8 y 9, respectivamente, pero utilizando errores estándar con clustering a nivel de diputado. Se observa que los resultados son muy similares, sin embargo, la variable Mujer, en la especificación (2), pierde su significancia. En la Tabla 11 no se encuentran mayores diferencias.

Tabla 11: Efecto del género y diario sobre el número de noticias que recibe un diputado, errores estándar clusterizados a nivel diputado

	Un género por noticia		Un diputado por noticia	
	(1)	(2)	(1)	(2)
Número de períodos	0.13 (0.661)	0.13 (0.659)	0.294 (0.351)	0.296 (0.35)
Edad	-0.032 (0.366)	-0.033 (0.363)	-0.013 (0.609)	-0.013 (0.606)
Número de comisiones	-0.792* (0.063)	-0.791* (0.063)	-0.849** (0.012)	-0.848** (0.012)
Presidente_Camara	10.142*** (0.001)	10.136*** (0.001)	7.572** (0.011)	7.559** (0.011)
Minutos de intervención	-0.0*** (0.0)	0.0*** (0.0)	-0.0*** (0.0)	-0.0 (0.33)
Mujer	-2.726*** (0.002)	-3.177*** (0.007)	-0.805 (0.242)	-1.006 (0.315)
LaTercera	-2.197** (0.015)	-2.315** (0.018)	-2.256*** (0.004)	-2.312*** (0.006)
DiarioGenero		1.01 (0.454)		0.459 (0.697)
Dummy Año	Sí	Sí	Sí	Sí
Dummy Partido	Sí	Sí	Sí	Sí
Observaciones	651	649	540	540
R ²	0.148	0.147	0.151	0.151

Notas: La tabla reporta estimaciones por OLS del efecto del género del diputado mencionado en una noticia y el periódico de origen de esta sobre el noticias publicadas en un año y diario dados. *Número de períodos* corresponde a cuantos períodos ha estado en la Cámara el diputado previo a la elección de 2014, *Número de comisiones* corresponde al número de comisiones a las que pertenece el diputado. *Presidente_Camara* es una dummy que toma valor 1 si el diputado era Presidente de la Cámara en el año en que fue publicada la noticia. *LaTercera* es una dummy que toma valor 1 si la noticia proviene de La Tercera y 0 si proviene de Emol. La diferencia entre la especificación (1) y (2) es que la primera no considera el efecto interactivo entre diario y género. Se utilizan errores estándar clusterizados a nivel diputado. Se reportan valores p entre paréntesis,

Uso de palabras

Para el uso de palabras se corrió el modelo Logit-Ridge sobre la muestra que considera noticias que mencionan diputados de un solo género. Las variables explicativas son el conteo del número de veces que aparece cada una de las 47.481 palabras y bigramas, obtenidas de acuerdo con la metodología presentada en la sección IV.

A partir de los resultados, en la Tabla 12 se muestra la matriz de confusión que permite medir la capacidad predictiva del modelo.

Tabla 12: Matriz de confusión del modelo Logit-Ridge

		Valores predichos		
		Hombre	Mujer	Total
Valores reales	Hombre	857	8	865
	Mujer	46	33	79
Total		903	41	944
Accuracy	94.3%	Prevalencia	8.4%	
Recall	41.8%	Precision	80.5%	

Es posible ver que el modelo predice bien el género asociado a una noticia el 94.3% de las veces. De estas, predijo 41 noticias sobre mujeres, de las cuales 33 eran efectivamente de mujeres (Precision de 80.5%). Sin embargo, 46 noticias que eran de mujeres son clasificadas como noticias de hombres (Recall de 41.8%).

En base a los resultados y a la experiencia empírica, se hipotetiza que, dentro del conjunto de noticias sobre mujeres, existe un subconjunto que es diferenciable en base al vocabulario de las noticias de hombres y que el resto de las noticias usa un lenguaje indistinto al que se usa en noticias sobre hombres, lo que dificulta su predicción.

Luego, para ahondar en esta hipótesis, se presentan en la Tabla 11 los parámetros con mayor capacidad de predecir el género masculino y el género femenino. A priori, resulta difícil encontrar relaciones evidentes de los predictores, por lo que se proponen una serie de categorías para clasificarlas, basadas en los contextos en que estas palabras se encuentran: comisión, declarativa, jerárquica-organizacional, agéntica, comunal, específica y otra.

Comisión se refiere a palabras que estén asociadas a comisiones con una proporción de hombres o mujeres particularmente alta. En las Tablas A3 y A4 del Anexo se presentan las comisiones con mayor participación de cada género.

Declarativa se refiere a palabras utilizadas previo o luego a una cita a un diputado. En esta categoría caen palabras como “señalar”, “declarar”, “plantear” y “recalcar”.

Jerárquica-organizacional se refiere a palabras que infieren una jerarquía u organización. En este grupo se consideran palabras como “Corte”, “presidente”, “Director”, “Presidencial”, “representante”, “vocero”.

Tabla 13: 20 mayores predictores de noticias sobre hombres y mujeres

Hombres		Mujeres	
Predictor	Efecto Marginal	Predictor	Efecto Marginal
Corte	-0.64	Mujer	0.85
Las Condes	-0.56	Antofagasta	0.77
Queda	-0.55	Respuesta	0.55
Presidente	-0.49	Vocero/a	0.51
Director(a) / Directo	-0.48	Partido comunista	0.50
Señalar	-0.47	Guillier	0.50
Juicio	-0.46	Diputado/a comunista	0.48
Últimatum/ último	-0.45	Política pública	0.47
Operar / Operativo	-0.45	Jornada	0.46
Presidencial	-0.44	Representante/ representar	0.43
Noviembre	-0.42	Concentración	0.41
Gremial	-0.41	Valor	0.40
Consecuente	-0.40	Ñuñoa	0.40
Declarar/declaró	-0.40	Fuerte	0.40
Timonel	-0.40	Asalto	0.39
Plantear	-0.39	Lily	0.39
Enviar	-0.39	Televisión	0.39
Deporte	-0.39	Ministro/a educación	0.39
Aumento / Aumentar	-0.36	Recalcar	0.37
Institución/ Instituto	-0.36	Aceptar renuncia	0.37

Notas: el efecto marginal de un token t es el cambio en la probabilidad de que una noticia sea clasificada como sobre mujeres, por cada vez que aparece t en la noticia.

Agéntica y comunal se refieren a la teoría de las características agénticas y comunales (Carlson (1971), Trapnell & Paulhus (2012)). Esta teoría, plantea que a los hombres usualmente se les asocia con características que Carlson llama agénticas: ascender en la jerarquía social, uso de la fuerza, competitividad, independencia y asertividad, entre otros. Las mujeres, en cambio, se han asociado a características comunales: preocupación por la comunidad, expresividad, desinterés. Ejemplos de palabras agénticas serían “presidente”, “Corte” o “fuerte” y para comunal se

encontrarían palabras como “vocero/a”, “política pública” o “representante”, pues evocan un sentido de comunidad.

Específica se refiere a palabras asociadas a ciclos específicos de noticias que tuvieron una participación especialmente alta de uno u otro género, pero que con alta probabilidad no tienen valor explicativo para algún fenómeno que determine el género asociado a una noticia en otro período de tiempo o bien están claramente asociadas a diputados/as específicos. En otras palabras, son palabras que fuera de este período del Congreso es muy probable que no tengan poder predictivo. En esta categoría se consideran palabras como “Las Condes”, “Noviembre”, “Antofagasta”, “Partido comunista”, “Diputado comunista”, “Guillier”, “Ñuñoa”, “Lily”, “Ministro/a educación”, “aceptar renuncia”.

Finalmente, Otra se refiere a palabras que no caen en ninguna de estas categorías o pueden tener una interpretación fuera de ellas, pero que podrían tener poder predictivo generalizable fuera de la muestra.

Para el análisis siguiente, con el fin de aterrizar lo mencionado, se mostrará entre paréntesis el ratio entre el porcentaje que representa la palabra del total de palabras usadas entre mujeres y hombres, es decir:

$$ratio_p = \frac{\left(\frac{n_{p,M}}{n_M}\right)}{\left(\frac{n_{p,H}}{n_H}\right)}$$

Donde $n_{p,G}$ representa el número de veces que se repite la palabra p en el subconjunto de noticias asociadas al género G y n_G representa el número total de palabras existentes en el *corpus* asociado al género G .

En las Tablas 14a y 14b se presentan las categorías asociadas a los predictores de mujeres y hombres, respectivamente. Llama la atención que, más allá de palabras específicas, existe una mayor incidencia de palabras asociadas a la categoría jerárquica-organizacional en hombres que mujeres: “Corte” (0.58), “presidente” (0.53), “Juicio” (0.61), “presidencial” (0.57), “timonel” (0.18), “Institución/Instituto” (0.42/0.64) para los primeros vs “Representante” (1.9), “Vocero/a” (2.07) en el caso de las últimas.

Tabla 14a: Categorías asociadas a mayores predictores de mujeres

Mujeres	
Predictor	Categoría
Mujer	Comisión (Familia, Desarrollo Social) / Comunal
Antofagasta	Específica / Comisión (Minería)
Respuesta	Otra
Vocero/a	Jerárquica / Comunal
Partido comunista	Específica
Guillier	Específica
Diputado/a comunista	Específica
Política pública	Otra / Comunal
Jornada	Otra
Representante/ representar	Jerárquica/Comunal
Concentrarse/concentración	Específica
Valorar	Otra
Ñuñoa	Específica
Fuerte	Otra / Agéntica
Asalto	Comisión (Seguridad Ciudadana)
Lily	Específica
Televisión	Otra
Ministro/a educación	Específica
Recalcar	Declarativa
Aceptar renuncia	Específica

Además, cabe notar que las palabras de los hombres se asocian a lo agéntico (alta jerarquía) y las de las mujeres a lo comunal (representación de una comunidad). Profundizando en la teoría agéntico-comunal destaca que los diputados no tienen ningún predictor comunal y las mujeres sólo uno agéntico: “fuerte” (1.86).

Luego, una segunda diferencia se da en la categoría declarativa: los hombres consideran “señalar” (0.87), “declarar” (0.8) y “plantear” (0.51), a diferencia de las mujeres donde sólo se encuentra “recalcar” (1.43).

Por último, de la categoría Otra, en el caso de las diputadas existe una mayor incidencia. Si bien, debido a la diversidad de las palabras el análisis no es claro, llaman la atención para mayor estudio palabras como “fuerte” (1.86), “televisión” (5.56) y “valorar” (1.34). En el caso de los diputados tanto “operar/operación” (0.47) como “consecuente” (0.42) llaman a mayor investigación.

Tabla 14b: Categorías asociadas a mayores predictores de hombres

Hombres	
Predictor	Categoría
Corte	Jerárquica / Agéntica
Las Condes	Específica
Quedar	Otra
Presidente	Jerárquica / Agéntica
Director(a) / Directo	Jerárquica / Agéntica
Señalar	Declarativa
Juicio	Jerárquica / Comisión (Constitución y Justicia)
Último	Específica
Operar / Operación	Otra
Presidencial	Jerárquica / Agéntica
Noviembre	Específica
Gremial	Específica/Otra
Consecuente	Otra
Declarar	Declarativa
Timonel	Jerárquica / Agéntica
Plantear	Declarativa
Enviar	Específica
Deporte	Comisión (Deportes)
Aumento / Aumentar	Específica/Otra
Institución/ Instituto	Jerárquica

Otros puntos que considerar son que en ambos casos las comisiones no son particularmente determinantes de los resultados. Junto a esto, hay un mayor número de palabras específicas en el caso de las mujeres, pero esto podría darse debido a que la muestra es significativamente más pequeña, lo que dificulta abstraer conceptos, por lo que hay una mayor dominancia de ciclos de noticias en que se mencionan mujeres y se repiten palabras similares, como el caso de “partido comunista” y “diputado/a comunista” que hacen alusión a Camila Vallejo y Karol Cariola, ambas diputadas relativamente mediáticas y que tienden a compartir noticias.

Discusión

Los resultados presentados en las dos secciones anteriores más que causales pretenden ser introductorios, abriendo la puerta para mayor investigación tanto desde la Economía, como otras ciencias sociales. No obstante lo anterior, se pueden esgrimir algunas hipótesis en base a lo encontrado.

Primero, en cuanto al análisis de frecuencias se encuentra un menor número de palabras y de noticias para mujeres de manera robusta, incluso controlando por la productividad de los diputados/as. Esto induce a pensar que, para los medios las diputadas resultarían menos noticiosas, tanto en términos de contenido de la noticia (número de palabras) como de frecuencia (número de noticias), especialmente en la primera. Entre las posibles explicaciones de este resultado, se encontrarían:

- a) Existe efectivamente un sesgo en contra de las diputadas de parte de los medios.
- b) Las diputadas efectivamente son menos mediáticas, teniendo menor contacto con la prensa que su contraparte masculina.

La hipótesis a) implicaría que consciente o inconscientemente los medios virtuales estarían perjudicando a las diputadas con un reportaje más escaso, potencialmente llevando a menores probabilidades de re-elección al haber menor exposición mediática. Existiría, por lo tanto, un costo social asociado a las decisiones de los medios. En cambio, si se diera la hipótesis b), la menor presencia mediática sería endógena a las diputadas y, por lo tanto, el costo potencial no debiese asociarse a los medios. Dilucidar cuál hipótesis prima es una pregunta abierta y un trabajo complejo, debido a la dificultad de obtener una estrategia de identificación que permita establecer causalidad.

Luego, en cuanto al uso de palabras, parecen particularmente merecedores de mayor investigación la diferenciación entre lo agéntico y lo comunal para diputados y diputadas, respectivamente, y el mayor número de palabras declarativas para los hombres.

En el primer caso, dos hipótesis serían:

- a) Los medios potencian estos temas al haber una asociación más fuerte entre el género del diputado y el tema a discutir en la noticia.
- b) Las diferencias se darían porque efectivamente los diputados y diputadas se enmarcan endógenamente en temas agénticos y comunales.
- c) Una combinación de ambas.

Que se cumpliera total o parcialmente la hipótesis a) sería el caso más pernicioso, pues potenciaría potenciales estereotipos y, sobre todo, teniendo en cuenta que los políticos usualmente están asociados a características más agénticas, podría tener un efecto negativo sobre las votaciones que reciben las mujeres, lo que a su vez podría repercutir en resultados de política pública y así en el bienestar social, como se vio en la sección II.

Finalmente, merece estudio el hecho de que se encuentren más palabras declarativas como predictores en el caso de los hombres. La primera pregunta por responder sería si es que a los hombres se les cita más que a las mujeres; luego, en caso de ser positiva la respuesta, se podría ahondar en la causa de este comportamiento.

VI. Conclusiones

¿Existe un trato diferenciado a las diputadas del Congreso 2014-2018 en los medios virtuales? La evidencia encontrada en este trabajo muestra que existen diferencias significativas, tanto en la frecuencia de noticias y palabras, como el uso de estas.

Esta pregunta se aborda a partir del análisis del texto de noticias *scrapeadas* de los portales virtuales de La Tercera y Emol para el mismo período del Congreso y se encuentra, bajo distintos niveles de significancia, que el número tanto de noticias por diputada, como de palabras por noticia para cada diputada tiende a ser menor. Además, desde el uso de palabras se encuentran palabras predictoras del género de una noticia que posiblemente no se relacionen con ámbitos estrictamente profesionales. No obstante, debido a la naturaleza de este problema, esgrimir canales causales concluyentes requerirá de un trabajo más profundo.

En términos de posibles canales para estas diferencias, se postulan algunas hipótesis. Para las diferencias en número de noticias y palabras por noticia, las diferencias podrían ser inherentes a los medios, por un sesgo que venga desde algún componente de la cadena de producción de noticias o bien podría ser inherente a las diputadas, buscando menor contacto con la prensa o un efecto mixto de ambas. Por el lado del uso de palabras, las fuentes de variación son múltiples, pero son dos los fenómenos que más llaman la atención: el uso de palabras agénticas y comunales para hombre y mujeres, respectivamente y la asociación de palabras declarativas a los hombres. Nuevamente, la pregunta es si las diferencias son endógenas a los diputados y diputadas o a los medios, o bien una mezcla de ambas.

Este documento, hasta donde pudo investigar el autor, es el primer trabajo que utiliza esta metodología y, por lo tanto, pretende abrir una puerta para mayor investigación de estos campos, no sólo para la Economía, sino también para otras ciencias sociales, aportando con la evidencia que pudo ser extraída de estos datos.

VII. Bibliografia

Adena, M., Enikolopov, R., Petrova, M., Santarosa, V., & Zhuravskaya, E. (2015). Radio and the Rise of the Nazis in Prewar Germany. *The Quarterly Journal of Economics*, 130(4), 1885-1939.

Anderson, S. P., Waldfogel, J., & Stromberg, D. (Eds.). (2016). *Handbook of Media Economics, vol 1A*. Elsevier.

Banerji, A. V., Imbert, C., & Pande, R. (2013). Entry, exit and candidate selection: Evidence from India.

Baron, D. P. (2006). Persistent media bias. *Journal of Public Economics*, 90(1-2), 1-36.

Brollo, F., & Troiano, U. (2016). What happens when a woman wins an election? Evidence from close races in Brazil. *Journal of Development Economics*, 122, 28-45.

Carlson, R. (1971). Sex differences in ego functioning: Exploratory studies of agency and communion.

Chattopadhyay, R., & Duflo, E. (2004). Women as policy makers: Evidence from a randomized policy experiment in India. *Econometrica*, 72(5), 1409-1443.

DellaVigna, S., & Kaplan, E. (2007). The Fox News effect: Media bias and voting. *The Quarterly Journal of Economics*, 122(3), 1187-1234.

Devitt, J. (1999). Framing gender on the campaign trail: Women's executive leadership and the press. Washington, DC: Women's Leadership Fund.

Durante, R., & Knight, B. (2012). Partisan control, media bias, and viewer responses: Evidence from Berlusconi's Italy. *Journal of the European Economic Association*, 10(3), 451-481.-

Ferreira, F., & Gyourko, J. (2014). Does gender matter for political leadership? The case of US mayors. *Journal of Public Economics*, 112, 24-39.

Gentzkow, M. (2006). Television and voter turnout. *The Quarterly Journal of Economics*, 121(3), 931-972.

Gentzkow, M., Kelly, B. T., & Taddy, M. (2017). Text as data (No. w23276). National Bureau of Economic Research..

Gentzkow, M., Shapiro, J. M., & Sinkinson, M. (2011). The effect of newspaper entry and exit on electoral politics. *American Economic Review*, *101*(7), 2980-3018.

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American economic review*, *90*(4), 715-741.

González, F., & Prem, M. (2018). Can television bring down a dictator? Evidence from Chile's "No" campaign. *Journal of Comparative Economics*, *46*(1), 349-361.

Heldman, C., Carroll, S. J., & Olson, S. (2005). "She brought only a skirt": print media coverage of Elizabeth Dole's bid for the Republican presidential nomination. *Political communication*, *22*(3), 315-335.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67.

Kahn, K. F. (1996). *The political consequences of being a woman: How stereotypes influence the conduct and consequences of political campaigns*. Columbia University Press.

La Ferrara, E., Chong, A., & Duryea, S. (2012). Soap operas and fertility: Evidence from Brazil. *American Economic Journal: Applied Economics*, *4*(4), 1-31.

Medios Digitales (2015). Ranking de medios digitales en Chile. Recuperado de <https://mediosdigitales.info/2015/02/05/ranking-de-medios-digitales-en-chile/>.

Navia, P., Osorio, R., & Valenzuela, F. (2013). Sesgo político en las lunas de miel presidenciales: El Mercurio y La Tercera, 1994-2010. *Intermedios: Medios de comunicación y democracia en Chile*, *35-57*.

Sendén, M. G., Sikström, S., & Lindholm, T. (2015). "She" and "He" in news media messages: pronoun use reflects gender biases in semantic contexts. *Sex Roles*, *72*(1-2), 40-49.

Strömberg, D. (2004a). Radio's impact on public spending. *The Quarterly Journal of Economics*, *119*(1), 189-221.

Trapnell, P. D., & Paulhus, D. L. (2012). Agentic and communal values: Their scope and measurement. *Journal of personality assessment*, *94*(1), 39-52.

VIII. Anexos

Tabla A1: Número y minutos de intervenciones per cápita.

Estadístico	Hombre	Mujer	Diferencia
Minutos de intervención per cápita	537.7	566.56	-28.86 (0.74)
Intervenciones per cápita	118.39	132.1	3.2 (0.46)

Tabla A2: Efecto del género y diario sobre el número de palabras en una noticia. Sólo La Tercera

	Un género por noticia		Un diputado por noticia	
	(1)	(2)	(1)	(2)
Número de períodos	0.591 (0.946)	-11.455 (0.316)	0.591 (0.937)	-11.455 (0.275)
Edad	-0.496 (0.624)	-0.435 (0.768)	-0.496 (0.617)	-0.435 (0.762)
Número de comisiones	-17.99 (0.198)	-10.674 (0.539)	-17.99* (0.085)	-10.674 (0.487)
Presidente_Camara	123.47*** (0.002)	-24.608 (0.64)	-123.47 (0.179)	-24.608 (0.854)
Minutos de intervención	0.0 (0.617)	-0.0* (0.064)	0.0 (0.723)	-0.0 (0.232)
Mujer	111.202** (0.011)	69.515* (0.067)	111.202*** (0.003)	69.515* (0.078)
Dummy Año	Sí	Sí	Sí	Sí
Dummy Partido	Sí	Sí	Sí	Sí
Observaciones	4298	4298	2430	2430
R ²	0.088	0.089	0.068	0.069

Notas: La tabla reporta estimaciones por OLS del efecto del género del diputado mencionado en una noticia y el periódico de origen de est sobre el número de palabras utilizadas en una noticia. *Número de períodos* corresponde a cuantos períodos ha estado en la Cámara el diputado previo a la elección de 2014, *Número de comisiones* corresponde al número de comisiones a las que pertenece el diputado. *Presidente_Camara* es una dummy que toma valor 1 si el diputado era Presidente de la Cámara en el año en que fue publicada la noticia. *Diario_La Tercera* es una dummy que toma valor 1 si la noticia proviene de La Tercera y 0 si proviene de Emol. La diferencia entre la especificación (1) y (2) es que la primera no considera el efecto interactivo entre diario y género. Se utilizan errores estándar robustos. Se reportan valores p entre paréntesis,

Tabla A3: Efecto del género y diario sobre el número de palabras en una noticia. Sólo Emol.

	Un género por noticia		Un diputado por noticia	
	(1)	(2)	(1)	(2)
Número de períodos	3.154 (0.604)	5.724 (0.364)	3.154 (0.342)	5.724 (0.181)
Edad	-0.863 (0.113)	-0.497 (0.404)	-0.863** (0.014)	-0.497 (0.301)
Número de comisiones	8.067** (0.048)	9.362* (0.051)	8.067** (0.011)	9.362** (0.029)
Presidente_Camara	37.875*** (0.003)	50.545*** (0.001)	37.875** (0.032)	50.545** (0.03)
Minutos de intervención	-0.0* (0.077)	-0.0*** (0.001)	-0.0 (0.314)	-0.0 (0.11)
Mujer	-40.702** (0.012)	-19.236 (0.348)	-40.702*** (0.003)	-19.236 (0.261)
Dummy Año	Sí	Sí	Sí	Sí
Dummy Partido	Sí	Sí	Sí	Sí
Observaciones	4298	4298	2430	2430
R ²	0.088	0.089	0.068	0.069

Notas: La tabla reporta estimaciones por OLS del efecto del género del diputado mencionado en una noticia y el periódico de origen de est sobre el número de palabras utilizadas en una noticia. *Número de períodos* corresponde a cuantos períodos ha estado en la Cámara el diputado previo a la elección de 2014, *Número de comisiones* corresponde al número de comisiones a las que pertenece el diputado. *Presidente_Camara* es una dummy que toma valor 1 si el diputado era Presidente de la Cámara en el año en que fue publicada la noticia. *Diario_La Tercera* es una dummy que toma valor 1 si la noticia proviene de La Tercera y 0 si proviene de Emol. La diferencia entre la especificación (1) y (2) es que la primera no considera el efecto interactivo entre diario y género. Se utilizan errores estándar robustos. Se reportan valores p entre paréntesis,

Tabla A4: Efecto del género y diario sobre el número de noticias que recibe un diputado. Sólo La Tercera.

	Un género por noticia		Un diputado por noticia	
	(1)	(2)	(1)	(2)
Número de períodos	-0.08 (0.651)	0.03 (0.825)	-0.08 (0.678)	0.03 (0.838)
Edad	-0.001 (0.972)	0.013 (0.41)	-0.001 (0.972)	0.013 (0.395)
Número de comisiones	0.099 (0.756)	-0.34* (0.075)	0.099 (0.765)	-0.34* (0.09)
Presidente_Camara	3.808 (0.383)	8.395*** (0.0)	3.808 (0.361)	8.395*** (0.0)
Minutos de intervención	-0.0 (0.421)	-0.0 (0.285)	-0.0 (0.518)	-0.0 (0.462)
Genero_M	-2.082** (0.036)	-0.267 (0.685)	-2.082** (0.034)	-0.267 (0.717)
Dummy Año	Sí	Sí	Sí	Sí
Dummy Partido	Sí	Sí	Sí	Sí
Observaciones	319	319	252	252
R ²	0.20	0.2	0.184	0.184

Notas: La tabla reporta estimaciones por OLS del efecto del género del diputado mencionado en una noticia y el periódico de origen de esta sobre el noticias publicadas en un año y diario dados. *Número de períodos* corresponde a cuantos períodos ha estado en la Cámara el diputado previo a la elección de 2014, *Número de comisiones* corresponde al número de comisiones a las que pertenece el diputado. *Presidente_Camara* es una dummy que toma valor 1 si el diputado era Presidente de la Cámara en el año en que fue publicada la noticia. La diferencia entre la especificación (1) y (2) es que la primera utiliza errores estándar robustos, mientras que la segunda utiliza errores estándar clusterizados a nivel diputado. . Se reportan valores p entre paréntesis,

Tabla A5: Efecto del género y diario sobre el número de noticias que recibe un diputado.
Sólo Emol

	Un género por noticia		Un diputado por noticia	
	(1)	(2)	(1)	(2)
Número de períodos	0.32 (0.604)	0.608 (0.356)	0.32 (0.44)	0.608 (0.22)
Edad	-0.096 (0.188)	-0.057 (0.334)	-0.096** (0.04)	-0.057 (0.212)
Número de comisiones	-1.741** (0.012)	-1.338** (0.014)	-1.741*** (0.001)	-1.338*** (0.003)
Presidente_Camara	13.381** (0.015)	7.314** (0.035)	13.381** (0.019)	7.314* (0.055)
Minutos de intervención	0.0** (0.026)	-0.0 (0.317)	0.0 (0.22)	-0.0 (0.746)
Mujer	-3.309** (0.011)	-1.186 (0.305)	-3.309*** (0.001)	-1.186 (0.208)
Dummy Año	Sí	Sí	Sí	Sí
Dummy Partido	Sí	Sí	Sí	Sí
Errores estándar clusterizados	No	Sí	No	Sí
Observaciones	330	330	288	288
R ²	0.187	0.187	0.18	0.18

Notas: La tabla reporta estimaciones por OLS del efecto del género del diputado mencionado en una noticia y el periódico de origen de esta sobre el noticias publicadas en un año y diario dados. *Número de períodos* corresponde a cuantos períodos ha estado en la Cámara el diputado previo a la elección de 2014, *Número de comisiones* corresponde al número de comisiones a las que pertenece el diputado. *Presidente_Camara* es una dummy que toma valor 1 si el diputado era Presidente de la Cámara en el año en que fue publicada la noticia. La diferencia entre la especificación (1) y (2) es que la primera utiliza errores estándar robustos, mientras que la segunda utiliza errores estándar clusterizados a nivel diputado. . Se reportan valores p entre paréntesis,

Tabla A6: Porcentaje del total de noticias de cada diario para cada año

Año	Emol	La Tercera
2014	19%	4%
2015	16%	9%
2016	21%	15%
2017	32%	48%
2018	12%	25%

Tabla A7: 10 Comisiones con mayor participación de hombres

Comisión	Hombres	Mujeres	n_personas
Control del Sistema de Inteligencia del Estado	100.0%	0.0%	7
DDHH	100.0%	0.0%	13
Deportes	100.0%	0.0%	13
Hacienda	100.0%	0.0%	13
Defensa Nacional	92.3%	7.7%	13
Economía, Consumidores y PyMEs	92.3%	7.7%	13
Trabajo	92.3%	7.7%	13
Constitución y Justicia	91.7%	8.3%	12
Regimiento Interno	90.9%	9.1%	11
Ética	88.9%	11.1%	9
Revisora de cuentas	87.5%	12.5%	8

Tabla A8: 10 comisiones con mayor participación de mujeres

Comisión	Hombres	Mujeres	n_personas
Familia	53.8%	46.2%	13
Desarrollo Social	69.2%	30.8%	13
Recursos Hídricos	69.2%	30.8%	13
Educación	70.0%	30.0%	10
Bomberos	76.9%	23.1%	13
Ciencia y Tecnología	76.9%	23.1%	13
Minería y Energía	76.9%	23.1%	13
Obras Públicas y Transportes	76.9%	23.1%	13
Salud	76.9%	23.1%	13
Seguridad Ciudadana	76.9%	23.1%	13
Vivienda y Bienes Nacionales	76.9%	23.1%	13
Zonas Extremas	76.9%	23.1%	13

