UNIVERSIDAD DE MÁLAGA | uma.es

CREACIÓN DE UN CORPUS DE NOTICIAS DE GRAN TAMAÑO EN INGLÉS, ESPAÑOL Y CATALÁN PARA EL ANÁLISIS DEL DISCURSO DE LA VIOLENCIA DE GÉNERO

Carla Fernández Melendres

XXXVI Congreso Internacional de la Asociación de Jóvenes Lingüistas

21 de septiembre de 2022

ÍNDICE

1. Proyecto NEWSGEN
2. Contexto
3. Marco teórico
4. Objetivos
5. Metodología
6. Presentación del corpus
7. Aplicaciones
8. Perspectivas de futuro

1. PROYECTO NEWSGEN

"Valores informativos e ideología: La construcción discursiva intercultural de género y desigualdades sociales en la prensa (digital) a través de la lingüística de corpus" **[NEWSGEN]**

Objetivo principal

Estudio y análisis de los discursos públicos en torno a **género y desigualdad social** en la prensa digital

- Análisis diacrónico
- Impacto político, cultural, social e ideológico

2. CONTEXTO

- La prensa desempeña un papel social fundamental en la formación de la opinión pública, reproduciendo o resistiendo los discursos de desigualdad (van Dijk, 1991)
- Aunque el análisis de las ideologías en el discurso de noticas goza de una larga tradición, solo recientemente los lingüistas han comenzado a utilizar corpus de gran tamaño y metodologías de lingüística de corpus para el estudio de noticias:
  - Baker et al. (2008, 2013) para la representación del Islam y los musulmanes en el Reino Unido
  - Potts et al. (2015) sobre el huracán Katrina
  - Fuster-Márquez y Gregori-Signes (2017) sobre el turismo en la prensa española
  - Maruenda-Bataller (2021); Santaemilia (2021) sobre la violencia de género en la prensa española

3. MARCO TEÓRICO

Análisis Crítico del Discurso

- Una de las principales premisas es que la prensa desempeña un papel social crucial en la formación de la opinión pública en, por ejemplo, dos direcciones opuestas:
  (1) reproduciendo (normas, valores y creencias comunes) o
  (2) resistiendo los discursos (es decir, desafiando estas normas sociales, valores, creencias) (Van Dijk 1991).
- El enfoque principal es el análisis de las ideologías y las desigualdades en los discursos de las noticias, que goza de una larga tradición en los estudios críticos del discurso.

4. OBJETIVOS

Diseño, compilación y anotación de un corpus periodístico (noticias y editoriales) del Reino Unido, Estados Unidos, España y Cataluña en torno al tema de la violencia de género.

Corpus digital de noticias diacrónico y multilingüe sobre la violencia de género

5. METODOLOGÍA

Diseño del corpus

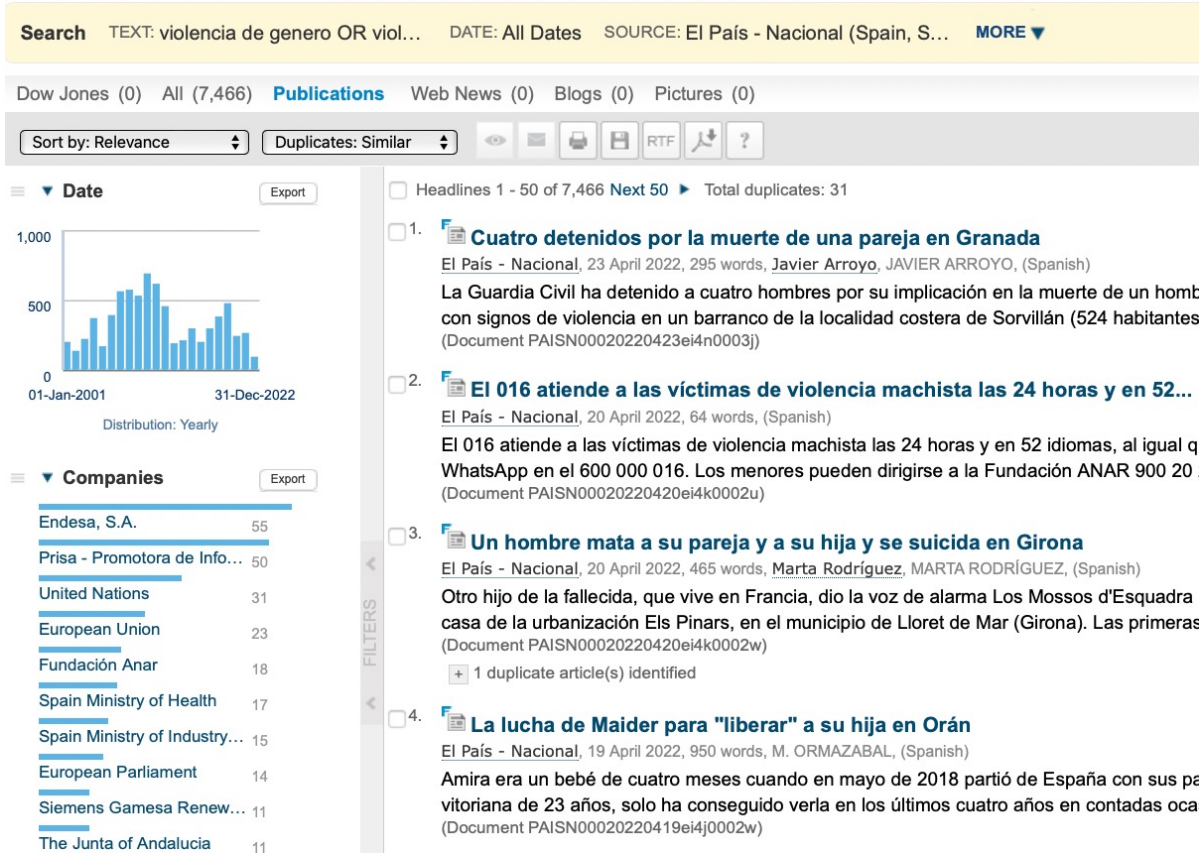
- Palabras clave en torno al tema de la violencia de género
- Según Xiao (2010: 148-153), los corpus deben definirse en términos de **tamaño**, **representatividad** y **equilibrio**.

1. **Tamaño**. Bowker y Pearson (2002: 49) consideran que no existe un número ideal de palabras preestablecido, ya que depende del propósito del estudio. Sinclair (1991: 18) cree que "un corpus debe ser lo más grande posible y debe seguir creciendo".
2. **Representatividad**. Definida por Biber (1993: 243) como "el grado en que una muestra incluye toda la gama de variabilidad de una población"
3. **Equilibrio**. Sinclair (2005) "las proporciones de los diferentes tipos de texto en un corpus deben corresponder a juicios informados e intuitivos". Douglas (2003: 34) considera que el equilibrio de un corpus es secundario para una buena práctica.

5. METODOLOGÍA

Compilación del corpus (I)

Extracción de artículos de prensa de la base de datos **Factiva**



5. METODOLOGÍA

Compilación del corpus (II)

Limpieza de textos y eliminación de duplicados

EL PAÍS

España

Cuatro detenidos por la muerte de una pareja en Granada

Javier Arroyo

JAVIER ARROYO

295 words
23 April 2022
El País - Nacional
PAISN
N (Nacional)
21
Spanish

La Guardia Civil ha detenido a cuatro hombres por su implicación en la muerte de un hombre y una mujer cuyos cuerpos fueron hallados anteayer con signos de violencia en un barranco de la localidad costera de Sorvillán (524 habitantes, Granada), según el delegado del Gobierno en Andalucía, Pedro Fernández. La investigación sigue su curso y no se descartan nuevas detenciones, según la misma fuente. Los fallecidos son un hombre de 61 años y su pareja, una mujer de 47, que fue concejal del PP en la localidad vecina de Gualchos-Castell de Ferro entre 2015 y 2019.

La jueza Irene Navarrete Cánova, titular del juzgado de Instrucción 1 de Motril, decretó ayer el secreto de sumario. El barranco de Los Yesos, en el que se encontraron los cuerpos, está cerca del cortijo en el que ambos residían. Fuentes de la investigación habían descartado desde el inicio que el caso tenga relación con la **violencia machista**.

La tarde del jueves, día en el que fueron hallados los cadáveres, la línea de investigación prioritaria vinculaba las muertes con un ajuste de cuentas por drogas. El fallecido ha sido identificado como J. A. V., apodado el Rey León, a quien se relaciona con el tráfico y plantación de cáñamo, según informa Efe.

La investigación se inició a primeras horas de la mañana del jueves, cuando la Guardia Civil recibió la llamada de la hija del hombre fallecido informando de la desaparición de la pareja y del hallazgo del vehículo de su padre en la salida de Castell de Ferro de la autovía A-7, cercana al domicilio, con restos de sangre. Pocas horas después, los agentes encontraron los dos cadáveres, semidesnudos, en el barranco.

Diario AS. SL

Document PAISN00020220423ei4n0003j

5. METODOLOGÍA

Compilación del corpus (III)

Clasificación, etiquetado y procesado de archivos de corpus

- Año de publicación: codificado con cuatro dígitos (AAAA)
- Nombre del país o región
- Nombre del periódico
- Un código de identificación numérico para identificar cada artículo, de modo que cada uno de los archivos reciba un código de identificación único
- Ejemplo: 2020_SP_ABC_0145.txt
  (año_país/región_ periódico_código)

5. METODOLOGÍA

Anotación del corpus

- Leech (1997:2) define la anotación de corpus como "la práctica de añadir información **lingüística interpretativa** a un corpus electrónico de datos lingüísticos hablados y/o escritos".
- En TXT y XML, listo para usar en Sketch Engine
- Metadatos:
  - <país>
  - <periódico>
  - <fecha>
  - <autor>
  - <sección_de_noticias>
  - <línea editorial>
  - <contador de palabras>
  - <idioma>
  - <titular>
  - <cuerpo>

```
<xml>
<article country="UK" newspaper="The Guardian" date="2000-01-03" year="2000" month="January" day="03" author="Diane Taylor" news_section="No section" editorial="left-centre" wordcount="805 words" language="English">
<section section_type="headline">
Go it alone.
</section>
<section section_type="body">
Ann Hudson finally found the courage to leave her abusive husband after 26 years of marriage […]
</section>
</article>
</xml>
```

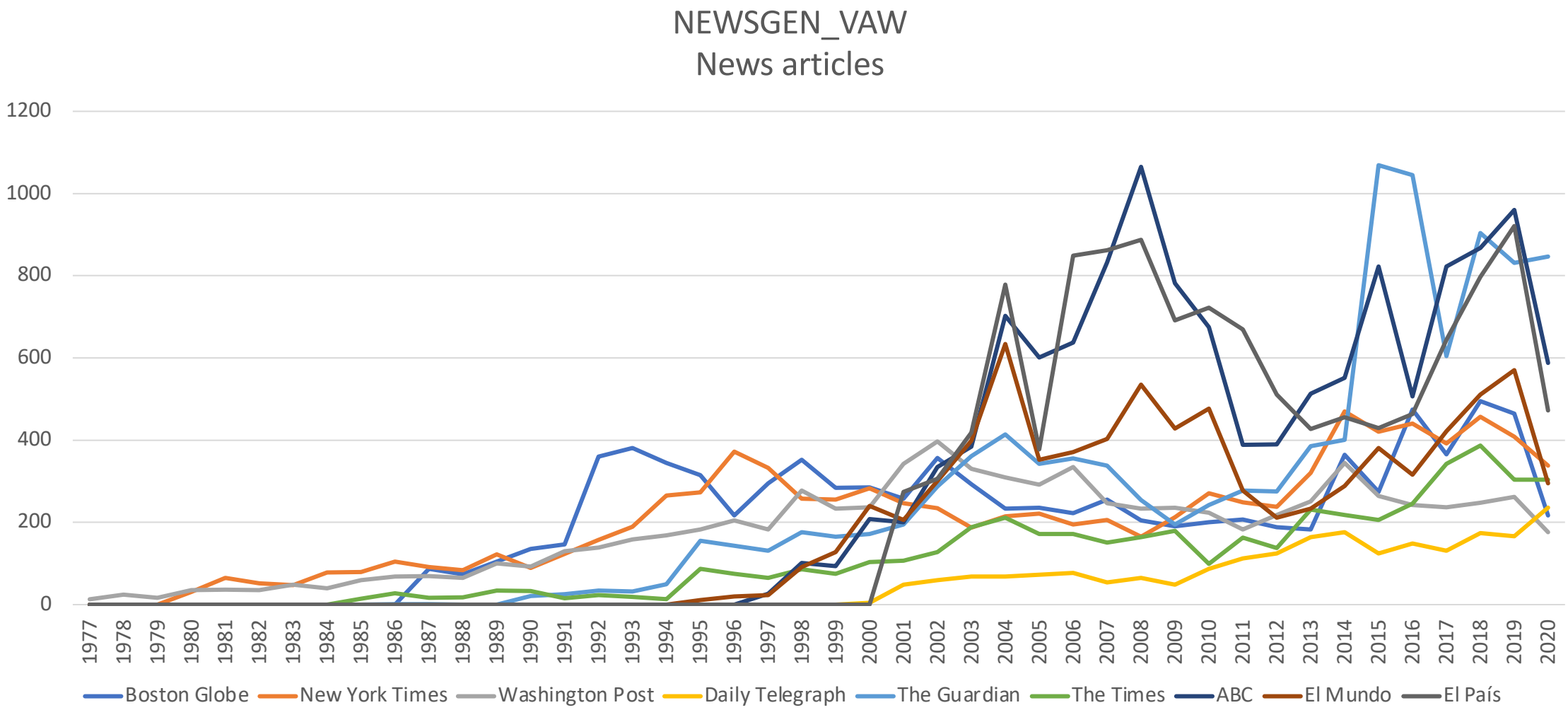
6. PRESENTACIÓN DEL CORPUS

UNIVERSIDAD DE MÁLAGA | uma.es

País	Periódico	Año	Textos	Tokens	Palabras	%
USA	Boston Globe	1987-2020	9.065	9.421.687	7.883.912	11,9
	New York Times	1980-2020	9.243	11.544.100	9.659.914	14,6
	Washington Post	1977-2020	7.992	10.044.341	8.404.935	12,7
Corpus USA			**26.300**	**31.010.128**	**25.948.761**	**39,2**
UK	Daily Telegraph	2000-2020	2.209	1.686.959	1.431.972	2,1
	The Guardian	1986-2020	10.729	15.287.037	12.976.453	19,3
	The Times	1985-2020	4.817	4.485.931	3.807.884	5,7
Corpus UK			**17.755**	**21.459.927**	**18.216.309**	**27,1**
SP	ABC	1997-2020	13.056	7.288.477	6.312.549	9,2
	El Mundo	1995-2020	8.126	6.420.942	5.561.175	8,1
	El País	2001-2020	11.955	7.457.408	6.458.865	9,4
Corpus SP			**33.137**	**21.166.827**	**18.332.589**	**26,8**
CA	Diari Ara	2013-2020	2.580	2.051.241	1.756.476	2,6
	El Periódico	2014-2020	2.681	1.801.817	1.542.895	2,3
	La Vanguardia	2013-2020	2.584	1.553.528	1.330.284	2,0
Corpus CA			**7.845**	**5.406.586**	**4.629.655**	**6,9**
TOTAL			**85.037**	**79.043.468**	**67.127.314**	

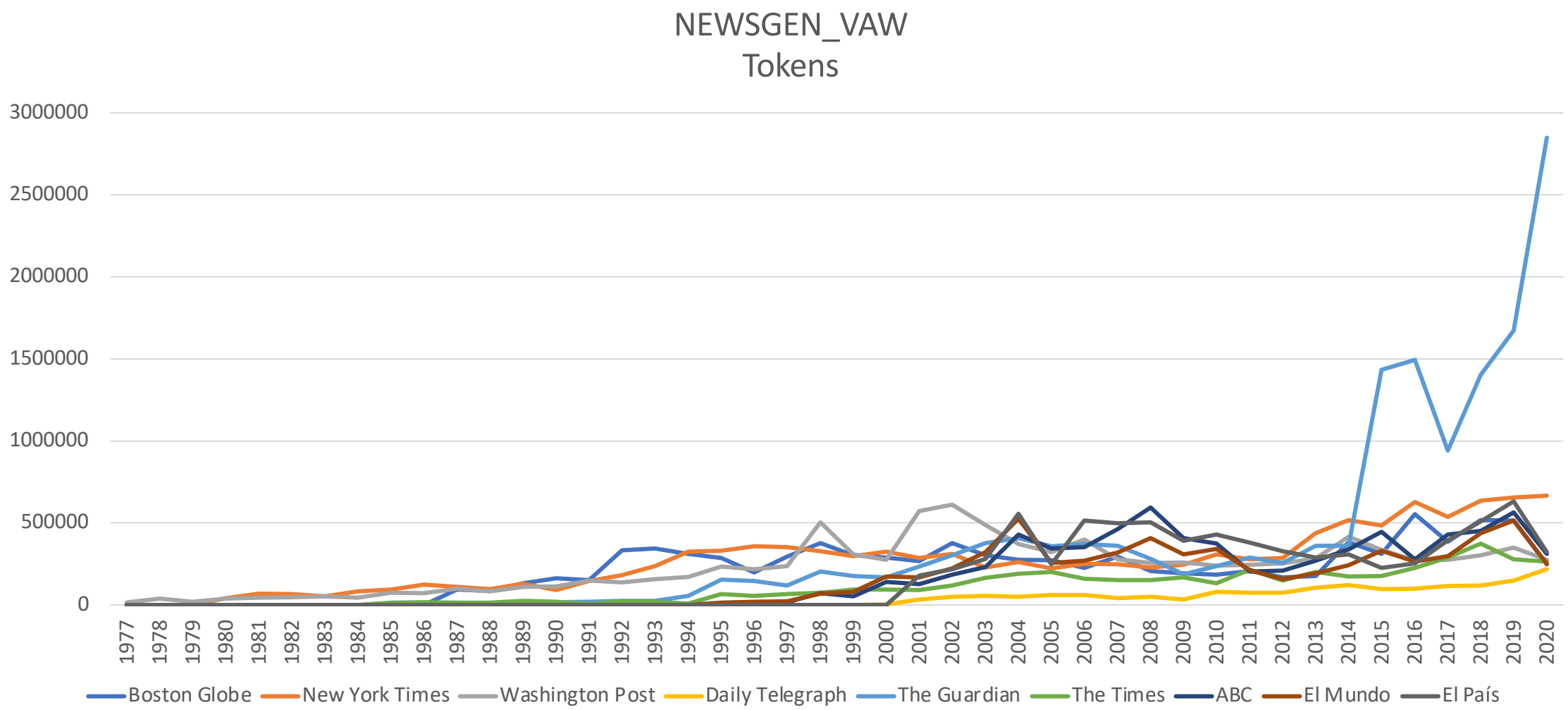
6. PRESENTACIÓN DEL CORPUS

EEUU
Reino Unido
España



6. PRESENTACIÓN DEL CORPUS

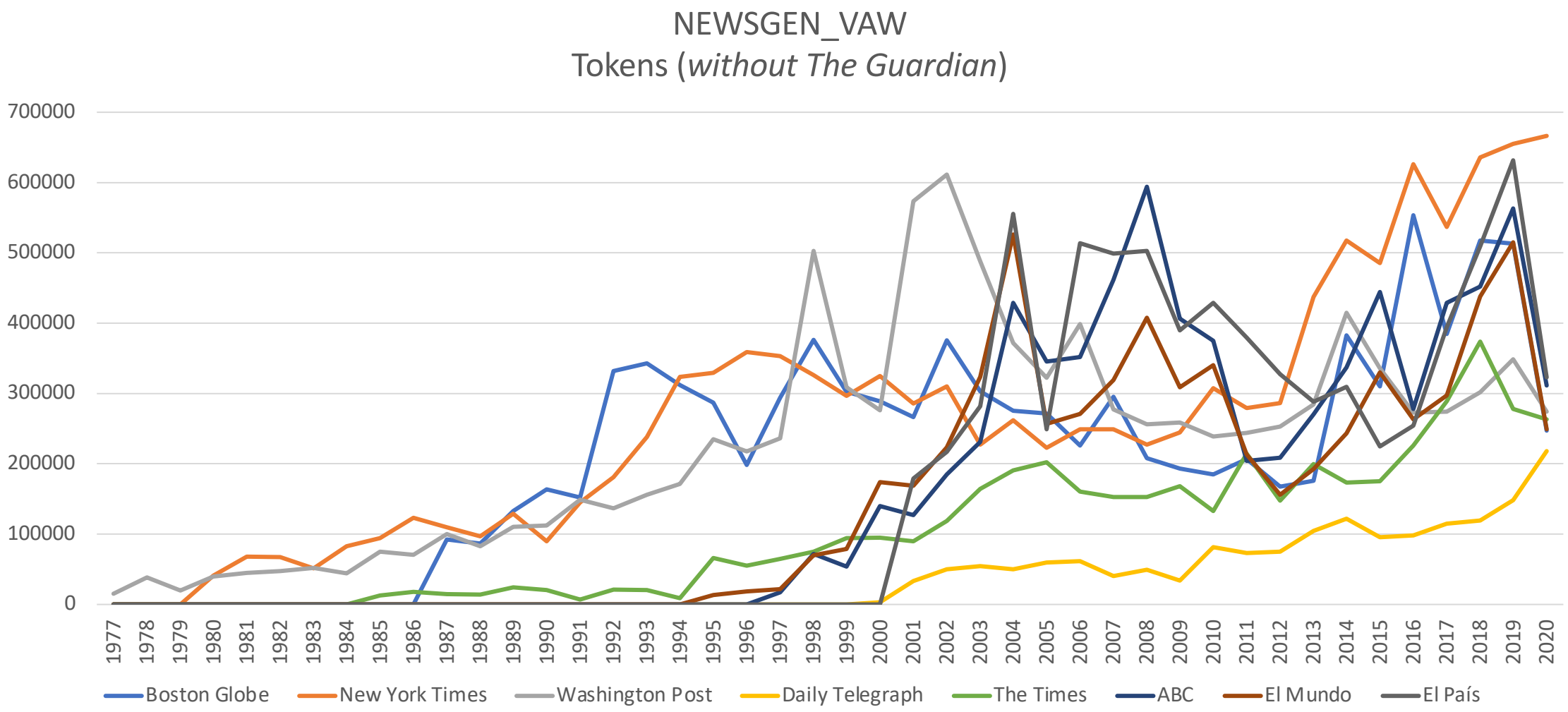
EEUU
Reino Unido
España



6. PRESENTACIÓN DEL CORPUS

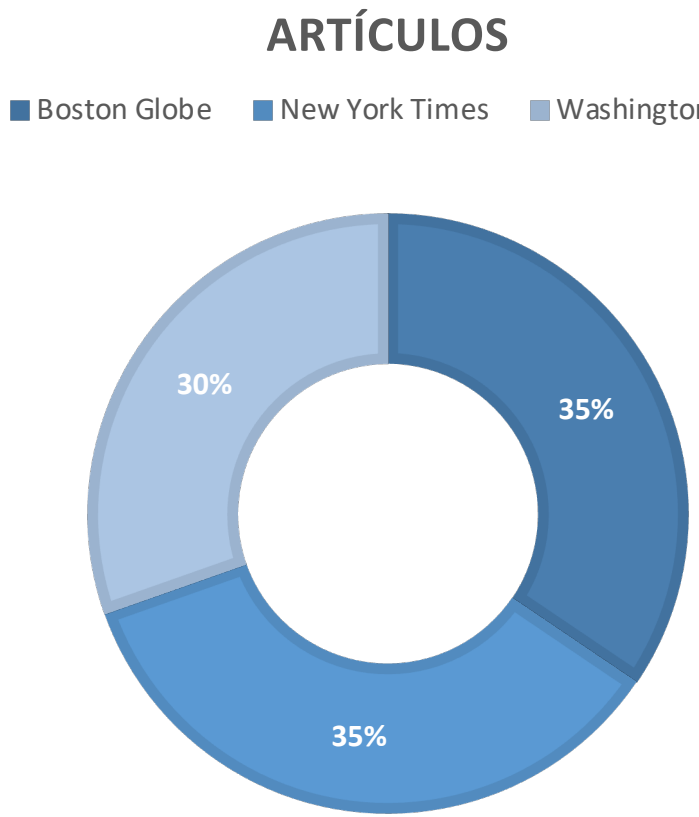
EEUU
Reino Unido
España

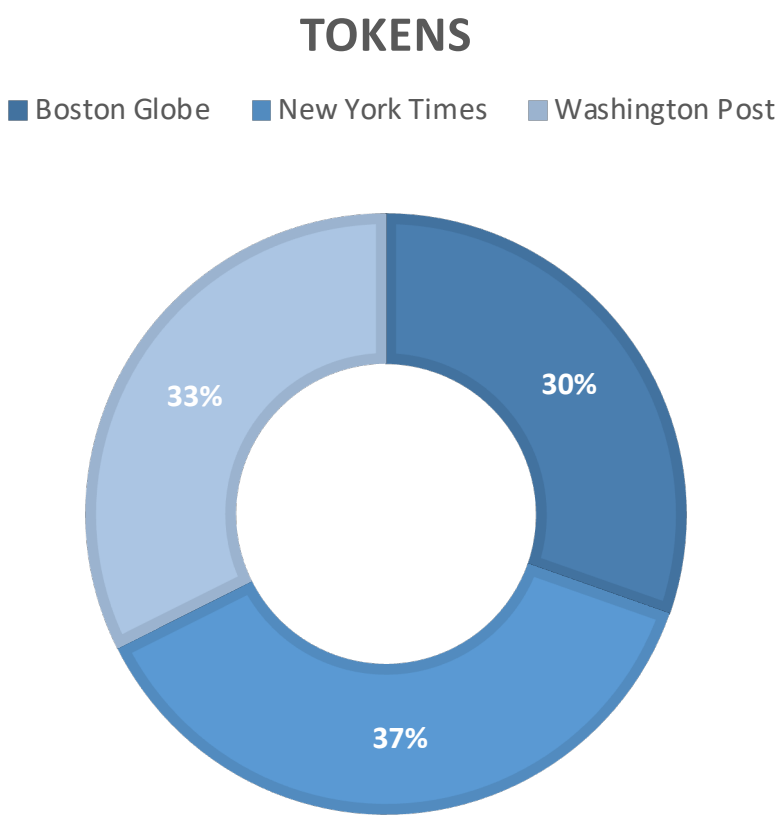
UNIVERSIDAD DE MÁLAGA | uma.es



6. PRESENTACIÓN DEL CORPUS

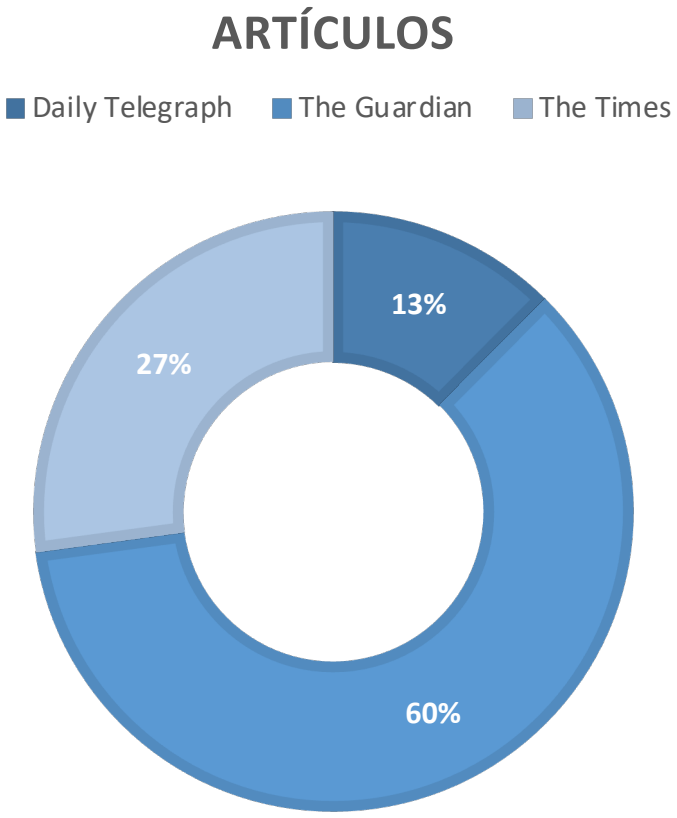
EEUU

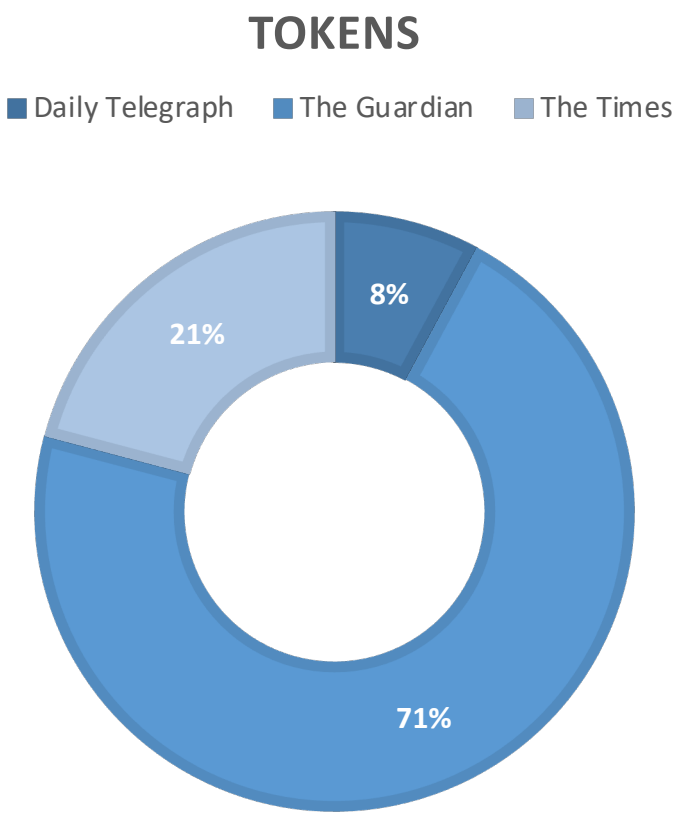




6. PRESENTACIÓN DEL CORPUS

Reino Unido

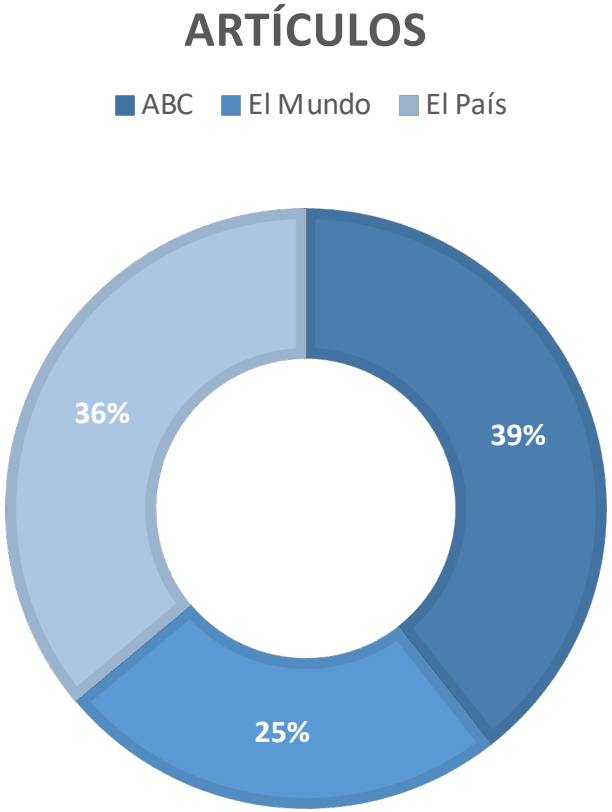


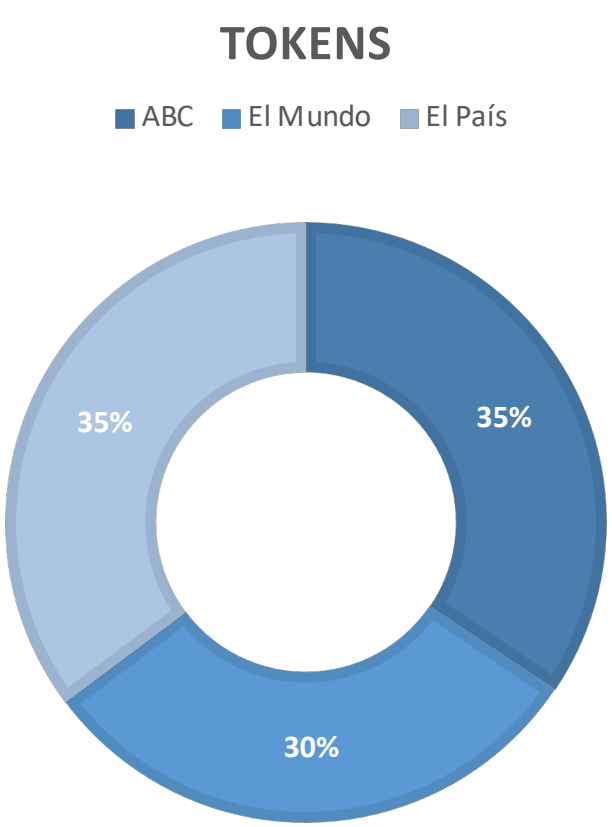


6. PRESENTACIÓN DEL CORPUS

UNIVERSIDAD DE MÁLAGA | uma.es

España





7. APLICACIONES

Este corpus está diseñado para examinar aspectos de interés periodístico, pero permite otros tipos de análisis del discurso (cuantitativos y/o cualitativos):

- Análisis diacrónico
- Impacto político, cultural, social e ideológico
- Posibles búsquedas seleccionando entre los metadatos: titulares, cuerpo, fecha, país/región, periódico, autor, sección, editorial, longitud, variedades lingüísticas (inglés británico vs. inglés americano), etc.

7. APLICACIONES

Trends: cambios en el tiempo / Daily Telegraph (2000-2020)

	Word	Trend ↓		Frequency	Sample
1	abusers	↗	2.25	80	
2	online	↗	1.96	246	
3	opinion	↗	1.96	82	
4	era	↗	1.88	69	
5	via	↗	1.66	111	
6	survivors	↗	1.60	110	
7	battered	↘	-1.60	82	
8	campaigners	↗	1.54	86	
9	murders	↘	-1.54	143	
10	brilliant	↗	1.48	77	
11	killings	↘	-1.48	117	
12	intended	↘	-1.43	89	
13	murder	↘	-1.43	681	
14	overseas	↗	1.43	99	
15	conflict	↗	1.38	123	
16	amid	↗	1.33	106	
17	economy	↗	1.33	83	

	Word	Trend ↓		Frequency	Sample
18	cost	↘	-1.28	137	
19	experiences	↗	1.23	108	
20	killing	↘	-1.23	228	
21	round	↘	-1.19	134	
22	sport	↗	1.19	177	
23	warned	↗	1.19	273	
24	sexist	↗	1.19	75	
25	near	↘	-1.15	200	
26	supporting	↗	1.15	111	
27	decide	↘	-1.11	77	
28	students	↗	1.07	154	
29	passion	↘	-1.00	137	
30	beaten	↘	-1.00	95	
31	bought	↘	-1.00	82	
32	wives	↘	-1.00	120	
33	charity	↗	0.97	446	
34	chief	↗	0.97	462	

	Word	Trend ↓		Frequency	Sample
35	support	↗	0.93	646	
36	area	↘	-0.90	192	
37	abuse	↗	0.90	1,495	
38	either	↘	-0.90	197	
39	red	↘	-0.90	125	
40	girl	↘	-0.87	352	
41	process	↗	0.87	171	
42	due	↗	0.87	237	
43	trial	↘	-0.87	398	
44	common	↘	-0.87	190	
45	yesterday	↘	-0.84	753	
46	large	↘	-0.81	181	
47	parents	↘	-0.78	527	
48	change	↗	0.75	399	
49	include	↘	-0.75	246	
50	prepared	↘	-0.75	106	

7. APLICACIONES

“Análisis de sentimiento de base lingüística con parsing retórico-discursivo (DisParSa)”

- El **análisis de sentimiento** extrae información relativa a la **polaridad** y la **intensidad** de las **emociones** expresadas en un texto (Liu, 2015).
- Sin embargo se limita a extraer información a nivel de aspecto, oración o documento (Socher et al., 2013).
- La extracción de opiniones implica necesariamente el análisis de la estructura discursiva
- **Teoría de la Estructura Retórica (RST)**: modelo desarrollado por Mann y Thompson (1988) y posteriormente implementado como analizador sintáctico o parser automático por Marcu (1997, 1999) y otros.

7. APLICACIONES

Editoriales

- Tipo de texto persuasivo con gran complejidad estructural a nivel discursivo
- Opinión que transmite una postura sobre un tema controvertido y de interés público
- Es habitual que se describan y contrasten ideas y posturas contrarias
- Propagan ideologías particulares o recomiendan ciertas actitudes (van Dijk, 1992)
- Varios autores han investigado las estrategias argumentativas en editoriales mediante el análisis sintáctico del discurso (Al-Khatib et al., 2016; Bal, 2014; Breeze, 2016; Le, 2003; Scheffer y Stede, 2016)
- Pero los corpus anotados existentes son pequeños, lo que limita su utilidad en tareas de análisis automatizado del lenguaje

8. PERSPECTIVAS DE FUTURO

- Siguiente paso en la tesis:
  Anotar una muestra significativa del corpus (**editoriales**) con información del discurso que se utilizará de entrenamiento y evaluación

REFERENCIAS

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyzanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse and Society*, 19(3), 273–306.

Baker, P., T. McEnery & C. Gabrielatos (2013) Sketching Muslims: A Corpus Driven Analysis of Representations Around the Word 'Muslim' in the British Press 1998–2009. *Applied Linguistics* 34(3): 255-278.

Bednarek, M. & H. Caple (2017) *The Discourse of News Values: How News organizations create newsworthiness*. OUP.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing* 8/4: 243–257.

Bowker, L. & Pearson, J. (2002). *Working with Specialized Language: A Practical Guide to Using Corpora*. London: Routledge.

Butler, J. (1990) *Gender trouble: feminist and the subversion of Identity.* Routledge.

Ehrlich, S., M. Meyerhoff & J. Holmes (2014) *The Handbook of Language, Gender & Sexuality*. Chichester: Wiley.

Factiva [database]. Retrieved from www.factiva.com.

Fuster-Márquez, M. & C. Gregori-Signes (2019) La construcción discursiva del turismo en la prensa española (verano de 2017). *Discurso & sociedad* Vol 13(2): 195-224.

Fuster-Márquez, M., J. Santaemilia, C. Gregori-Signes & P. Rodríguez Abruñeiras (2021) Insights from corpus-assisted discourse analysis: Unveiling social attitudes and values. In Fuster-Márquez, M., J. Santaemilia, C. Gregori-Signes & P. Rodríguez Abruñeiras (eds.) *Exploring discourse and ideology through corpora. In Linguistic Insights*. Bern: Peter Lang, 7-16.

Kilgarriff, A., V. Baisa, J. Busta, & M. Jakubicek (2014) The Sketch Engine: Ten Years On. *Lexicography* 1 (1): 7-36.

Lazar, M. (2005) *Feminist Critical Discourse Analysis: Gender, Power and Ideology*. Palgrave-Macmillan.

Leech, G. N. (1997). Introducing corpus annotation. In: Roger Garside, Geoffrey Leech and Tony McEnery (eds.), *Corpus Annotation*, (pp, 1–18). London/New York: Longman.

Maruenda-Bataller, S. (2021) The role of news values in the discursive construction of the female victim in media outlets: A comparative study. In Fuster-Márquez, M., J. Santaemilia, C. Gregori-Signes & P. Rodríguez Abruñeiras (eds.) *Exploring discourse and ideology through corpora*. In Linguistic Insights. Bern: Peter Lang, 141-166.

Partington, A., A. Duguid & C. Taylor (2013) Patterns and Meanings in Discourse : Theory and Practice in Corpus-assisted Discourse Studies (CADS). John Benjamins Publishing.

Potts, Amanda, Bednarek, Monika & Caple, Helen (2015) How can computer-based methods help researchers to investigate news values in large datasets? A corpus linguistic study of the construction of newsworthiness in the reporting on Hurricane Katrina. *Discourse and Communication* 9 (2): 149-172.

Santaemilia, José & Sergio Maruenda (2011) Building a comparable corpus (English-Spanish) of newspaper articles on gender and sexual (in)equality (GENTEXT-N): Present and future applications in the analysis of socio-ideological discourses. In María Luisa Carrió Pastor & Miguel Ángel Candel Mora (eds.) *Actas del III Congreso Internacional de Lingüística de Corpus - Las Tecnologías de la Información y las Comunicaciones: Presente y futuro en el análisis de corpus. Valencia*. Editorial Universitat Politècnica de València. 197-204

Santaemilia, José & Sergio Maruenda (2013) "Naming practices and negotiation of meaning: A corpus-based analysis of Spanish and English newspaper discourse." In Istvan Kecskes & Jesús Romero Trillo (eds.) *Research Trends in Intercultural Pragmatics*. Berlin: De Gruyter Mouton. 439-457.

Santaemilia, José & Sergio Maruenda (2014) The linguistic representation of gender violence in (written) media discourse: The term 'woman' in Spanish contemporary newspapers *Journal of Language Aggression and Conflict* 2(2): 249-273.

Santaemilia, José (2021) News values as evaluation. Main naming practices in Violence Against Women news stories in contemporary Spanish newspapers: *El País* vs. *El Mundo* (2005-2010). RiCL 9 (2): 90-113.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. (2005). Corpus and text – Basic principles. In Martin Wynne ed. *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books. http://users.ox.ac.uk/~martinw/dlc/index.htm (7 May, 2020.)

Van Dijk, T. (2001). Critical discourse analysis. In D. Tannen, D. Schiffrin, & H. Hamilton (Eds.), *Handbook of discourse analysis* (pp. 352–371). Oxford: Blackwell.

Xiao, R. (2010). Corpus creation. In Nitin Indurkhya and Frederick J. Damerau eds. *Handbook of Natural Language Processing.* (pp. 147–165). Boca Raton: Chapman & Hall/CRC. ,5

UNIVERSIDAD DE MÁLAGA | uma.es

MUCHAS GRACIAS

Carla Fernández Melendres
cfdz@uma.es