Versión final del autor (AM). La versión publicada (VoR) se encuentra en:
Revista Española de Lingüística Aplicada = Spanish Journal of Applied Linguistics, vol. 35, n.º 2, octubre de 2022, pp. 425-48,
https://doi.org/10.1075/resla.20017.dia https://www.jbe-platform.com/content/journals/10.1075/resla.20017.dia

Corpus RLD: el corpus de uso de recursos lingüísticos en el derecho español

Ángel Alonso-Cortés Manteca(*), cortlin@ucm.es,

Juan Manuel Díaz Ayuga(*), https://orcid.org/0000-0003-0470-5537, juadia01@ucm.es,

Ana María Fernández–Pampillón Cesteros(*), https://orcid.org/0000-0002-6606-0159, apampi@filol.ucm.es

(*) Los nombres de los autores aparecen en orden alfabético

Área de Lingüística General. Departamento de Lingüística, Estudios Árabes, Hebreos y de Asia Oriental, Universidad Complutense de Madrid

Resumen

Este artículo explica el corpus RLD, una recopilación de sentencias dictadas por jueces de los Tribunales Supremo y Constitucional, quienes han optado por el uso de diccionarios y la gramática del español como autoridad lingüística en la redacción de dichas sentencias judiciales. Las cuatro fases seguidas para la construcción del corpus: (i) diseño, (ii) implementación, (iii) revisión y (iv) publicación se explican detalladamente. El objetivo del corpus, que incluye 1940 sentencias emitidas entre los años 2012 y 2018, es proporcionar una muestra de datos representativos y empíricamente satisfactorios del empleo de los diccionarios y la gramática como autoridad lingüística en la jurisprudencia española. Este corpus viene a cubrir un vacío en los estudios de Lengua y Derecho: por una parte, la ausencia de grandes corpus de textos jurídicos compartidos, disponibles para cualquier investigador o grupo de investigación interesado, y, por otra, la necesidad de analizar empíricamente el uso del diccionario y la gramática en las sentencias de los Tribunales Supremo y Constitucional.

Palabras Clave: corpus, recursos lingüísticos, sentencias judiciales, jurisprudencia española, Tribunal Supremo, Tribunal Constitucional

1. Introducción

Desde la obra del gran jurista alemán Friedrich von Savigny, *Sistema del Derecho Romano Actual*, publicada entre 1840 y 1849, la lengua del Derecho, la empleada por los jueces y escrita en códigos, ha sido objeto de estudio por la relevancia que tiene en los procesos judiciales. Asimismo, los juristas del siglo XX como Kalinowski (1973), Hernández Gil (1989) y Capella (1968), entre otros, han dedicado obras importantes para destacar la relevancia del lenguaje en el Derecho.

Para la investigación de ese campo de estudio particular se necesita en primer lugar un conjunto de datos suficientemente representativo de la lengua estudiada, y luego habrán de emplearse los instrumentos de la Lingüística teórica y de la Retórica para analizar empíricamente tal muestra. En este sentido, la moderna Lingüística Computacional permite tratar los datos que previamente han sido localizados en los corpus de sentencias. Esto facilita enormemente el objetivo del proyecto "Lengua y Derecho".

El corpus que aquí presentamos forma parte del proyecto "Lengua y Derecho", cuyo objetivo es el análisis lingüístico de los medios léxicos y gramaticales que emplean los jueces en la interpretación de las leyes cuando emiten las sentencias. Los medios lingüísticos fundamentales que utilizan los jueces son los diccionarios y la gramática de la lengua española, tal y como ya han señalado Henríquez Salido et al. (2013) en un interesante estudio acerca del uso de los diccionarios como fuente de autoridad en las sentencias dictadas por los magistrados del Tribunal Supremo. Para estos investigadores, "la gramática y el diccionario se erigen en dos pilares fundamentales del derecho" (Henríquez Salido et al., 2013, p. 121). En particular, el Diccionario de la Real Academia Española (DRAE) es un medio muy utilizado en la interpretación de las leyes y de las pruebas judiciales. Este "ocupa un lugar muy relevante como criterio de *auctoritas*", llegando incluso a ser reconocido por los jueces como "«nuestro Diccionario oficial»". A su vez, las definiciones y acepciones que en él se recogen son consideradas, en ocasiones, como "«definiciones oficiales»" (Henríquez Salido et al., 2013, p. 121).

El proyecto "Lengua y Derecho" tiene el objetivo, primero, de detectar el empleo que hacen los jueces tanto de conceptos lingüísticos (sonido, sílaba, palabra, frase, oración, las relaciones semánticas entre palabras, ambigüedad, vaguedad…) como

las propias interpretaciones de los jueces, incluyendo además cuestiones ortográficas, que con frecuencia afectan a la interpretación.

De acuerdo con lo establecido en el artículo 1.6 del Código Civil, la jurisprudencia realiza una labor complementaria a la del ordenamiento jurídico "al interpretar y aplicar la ley, la costumbre y los principios generales del derecho" (Boletín Oficial del Estado, 2019, p. 2). Para llevar a cabo dicha interpretación de las normas, el propio Código Civil, en su artículo 3.1, establece que deberá atenderse al "sentido propio de sus palabras, en relación con el contexto, los antecedentes históricos y legislativos, y la realidad social del tiempo en que han de ser aplicadas" (2). Asimismo, el artículo 675, relativo a las disposiciones testamentarias, y el 1281, sobre la interpretación de los contratos, establecen, respectivamente, que, si los términos son claros, dichos documentos deberán entenderse "en el sentido literal de sus palabras" (113) o "de sus cláusulas" (185).

De lo anterior se deriva que, tanto para la interpretación de las leyes como de los términos en los que se redacta un contrato o testamento, así como otras pruebas judiciales, el magistrado debe atender "al contenido estrictamente lingüístico", al sentido propio o literal de las palabras, cuando dicta una sentencia, pero sin olvidar aquellos factores que pueden "variar la interpretación original" y "que podríamos denominar contenido extralingüístico" (González Salgado, 2011, p. 59), es decir, lo que el citado artículo 3.1 recoge como "el contexto, los antecedentes históricos y legislativos, y la realidad social." Sin embargo, la jurisprudencia ha manifestado en numerosas ocasiones que "el contenido estrictamente lingüístico prevalece sobre el contenido extralingüístico" (González Salgado, 2011, p. 59). Los jueces se basan frecuentemente en aforismos latinos como *in claris non fit interpretatio* ('en las cosas claras no se hace interpretación'), muy común en las sentencias del Tribunal Supremo (Sánchez Rubio, 2004), o *non ex opinionibus singulorum, sed ex communi usu, nomina exaudiri debent* ('los nombres no se deben entender conforme a las opiniones de cada cual, sino según su uso común') con el fin de demostrar esa prevalencia de lo lingüístico. Para estos, no se ha de "«atribuir un sentido distinto a las palabras cuando su sentido literal es claro», y […] «los nombres no se deben entender según las opiniones de cada cual, sino [según] su uso común»" (Henríquez Salido et al., 2013, p. 97). No es

de extrañar, por lo tanto, que para el jurista Jesús Prieto de Pedro los tres principios fundamentales que deben regir todo lenguaje jurídico sean factores relativos a la comprensión lingüística de los textos: "la claridad, la precisión y la corrección gramatical" (1996, p. 113). Factores que destaca, a su vez, el profesor Bayo Delgado, para quien, si bien "[e]n todos los escritos técnicos la claridad es fundamental", "en las resoluciones judiciales […] están en juego derechos y libertades del individuo. […] Un uso correcto del lenguaje supone garantía de los derechos del inculpado" (1996).

Se explica, así, que los magistrados empleen diversos recursos lingüísticos en la interpretación de las leyes y de las pruebas judiciales. En concreto, es frecuente entre los jueces el uso de los diccionarios, así como del conocimiento que poseen los propios magistrados, especialmente el referido a las categorías morfosintácticas de las palabras. Tal uso les permite pronunciarse, no siempre con acierto, en torno a aspectos léxico-semánticos y morfosintácticos de los escritos que examinan con el fin de emitir su fallo en uno u otro sentido (González Salgado, 2011; Henríquez Salido et al., 2013; Henríquez Salido, 2006). No obstante, un empleo inadecuado de este conocimiento lingüístico puede afectar considerablemente a la interpretación de los jueces y a su decisión en la sentencia.

Podemos encontrar un ejemplo de tal uso en la sentencia nº 96/2004 de la Sección 1ª de la Audiencia Provincial de Pontevedra, del 18 de noviembre de 2004, donde, al emitir su juicio, el magistrado emplea el Diccionario de la Real Academia Española para dotar a la palabra "conocer" de una ambigüedad de la que, claramente, carece. Para ello, el juez establece una distinción tajante entre los usos de "conocer" que emplea un testigo en dos declaraciones diferentes, acudiendo a las acepciones segunda ("Entender, advertir, saber, echar de ver a alguien o algo") y cuarta ("Tener trato y comunicación con alguien") de su definición en el Diccionario de la Real Academia Española (DRAE) con el fin de justificar tal diferencia y la presunta ambigüedad del término. El contexto, no obstante, no permite la ambigüedad apuntada por el magistrado, de la misma forma que no se explica que, en la primera de las declaraciones, acuda a la cuarta acepción en lugar de a la segunda, cuando en el DRAE "las acepciones más frecuentes tienden a aparecer antes que las que lo son menos" (Real

Academia Española, 2014, p. LII). Esto, como vemos, influye considerablemente en su fallo:

> "El recurrente pretende restar credibilidad al testigo, afirmando que incurrió en severas contradicciones. Sin embargo, la atenta lectura de sus manifestaciones no revela contradicción alguna; es más, lo que se intenta hacer pasar por contradicción o retractación (esto es, que, al ser preguntado por los agentes sobre si conocía o mantenía relaciones de algún tipo con la persona que acababa de identificar, el testigo manifestó que "no tiene ningún tipo de relación con ella y que *no la conoce de nada*", cuando después, en el acto del juicio oral, declaró que ante los agentes que "le vio la cara a Manuel, que *le conoce* desde hace 4 años, que *le conoce* de vista, que Manuel *le conoce* y éste a él"), tiene su explicación lógica en *la ambigüedad del término "conocer"*, es decir, la expresión "no tiene ningún tipo de relación" y "no conoce de nada" a una persona no quiere decir forzosamente que no la hubiera visto con anterioridad ni que no supiese de su existencia, sino que "no tenía trato y comunicación con ella" (cuarta de las acepciones del *diccionario de la Real Academia de la Lengua* para la palabra *"conocer"*), lo que, por otra parte, ratificó el testigo en la vista[1]."

Los jueces adoptan tácitamente el principio de interpretación semántica de las palabras de los textos conocido como "lectio facilior", consistente en adoptar la interpretación más fácil y conveniente para interpretar las palabras en los textos. Sin intentar la alternativa o "lectio dificilior."

En otros casos, sin embargo, los magistrados acuden a su propio conocimiento morfosintáctico para intentar demostrar que no existe un sentido unívoco en una construcción gramatical determinada. En la sentencia nº 2466/1993 de la Sala de lo Penal del Tribunal Supremo, del 3 de noviembre de 1993, se menciona la posible ambigüedad en el uso del pronombre "se" en la oración "«... lo que motivó que se entablara entre ellos una discusión, propinándo**se** un puñetazo...»", ya que, de acuerdo con una de las partes, no es posible establecer quién es el objeto del puñetazo:

> "[E]l recurrente entiende que [la oración] no deja claro cuál de las alternativas que puede inferirse del uso del «se» junto al reflexivo y gerundio del verbo «propinar» es la deducible: si hubo un intercambio de puñetazos; si se usa la forma impersonal, quedando indeterminado quién propinó el puñetazo a quien; o si «rizando el rizo», se debe interpretar que cada contendiente se

[1] Los resaltados en cursiva son propios.

> propinó un puñetazo a sí mismo, como cabría deducir del uso del reflexivo «se». En definitiva entiende que queda sin precisar quien inicia la agresión o dio el primer golpe[2].”

Por todo lo arriba expuesto, consideramos que un análisis lingüístico de los medios léxicos y gramaticales que los jueces emplean al emitir sentencias permitiría descubrir, si existen, patrones de incorrecciones en el uso de dicho conocimiento, lo que posibilitaría la elaboración de una guía de buenas prácticas que, sin duda, sería de gran utilidad para los magistrados.

Para llevar a cabo un análisis lingüístico del uso, correcto o incorrecto, de la gramática y los diccionarios que emplean los magistrados, es necesario disponer de una muestra representativa de dicho uso. De la revisión del estado de la cuestión (sección 2), se desprende que no existe tal muestra representativa disponible para la investigación. Por ello, este trabajo que presentamos se ha enfocado en crear un corpus representativo que permita realizar este estudio empírico concreto, pero que, además, pueda reutilizarse o adaptarse para llevar a cabo nuevos estudios lingüísticos sobre Lengua y Derecho. En este sentido, el objetivo que abordamos ahora es la creación de un corpus de la lengua de las sentencias a partir del uso de los diccionarios por los magistrados, que sea suficientemente representativo y, al mismo tiempo, fácil de utilizar y adaptar a nuevas cuestiones de investigación.

A continuación, presentamos el problema que se aborda: la no disponibilidad de corpus en español para un estudio empírico del uso del conocimiento lingüístico como fuente de autoridad en los Tribunales Supremo y Constitucional. En la sección segunda, estado de la cuestión, revisamos los corpus y estudios empíricos sobre Lengua y Derecho realizados con anterioridad. En la tercera sección especificamos la hipótesis y objetivos de la investigación. En la cuarta sección detallamos la metodología de construcción del corpus y los resultados en cada fase. Finalmente, en la quinta sección presentamos las conclusiones y líneas de trabajo futuro.

2. Estado de la cuestión

[2] Tomo los ejemplos anteriores, así como las referencias a ambas sentencias, del trabajo de González Salgado (2011, pp. 62-63, nota al pie 8).

Para la creación del corpus RLD es necesario contar con una muestra representativa de tales sentencias, de forma que el estudio sea lo más riguroso y sistemático posible, y cumpla con dos aspectos imprescindibles en toda investigación lingüística: la representatividad y la autenticidad del material analizado (Sinclair, 2005; Barbera, Corino y Onesti, 2007). En este sentido, la Lingüística de Corpus nos ofrece modelos y procedimientos para construir extensas muestras empíricas que permitan describir tales usos. No es de extrañar, por tanto, que esta disciplina haya revolucionado las diferentes áreas de la Lingüística en las últimas décadas (Tognini-Bonelli, 2001; Wynne, 2005a), y que la utilización de corpus de textos auténticos haya despertado un especial interés en la investigación de discursos de especialidad, pues estos ofrecen a los investigadores en dicho campo un método riguroso con el que compilar, organizar y analizar los discursos y géneros que conforman su objeto de estudio. Dentro de la investigación de lenguas en ámbitos específicos, el examen del lenguaje jurídico se ha beneficiado especialmente de la utilización de una metodología basada en corpus (Taranilla, 2013).

En este sentido, y a pesar de que los corpus que compilan textos exclusivamente jurídicos son muy raros (Pontrandolfo, 2013a), son numerosos los estudios que analizan el discurso jurídico empleado en tribunales y audiencias españolas mediante corpus que funcionan a modo de legitimación empírica de sus investigaciones. Así, entre muchos otros, encontramos ejemplos de proyectos que se basan en corpus para establecer las características del género "sentencia" (Henríquez Salido 2006; Taranilla, 2014, 2015), para analizar el léxico empleado en la jurisprudencia (Henríquez Salido y No Alonso-Misol, 2005; Henríquez Salido, 2007; Polanco Martínez, 2011; Polanco Martínez y Yúfera Gómez, 2013; Pontrandolfo, 2019), para examinar los marcadores discursivos utilizados en las sentencias judiciales (López Samaniego, 2006), para estudiar la cortesía en las sentencias del Tribunal Constitucional (Taranilla, 2009), para analizar la configuración narrativa de los antecedentes de hecho (Taranilla, 2011), para determinar las características del discurso judicial oral (Briz y el Grupo Val.Es.Co., 2012) y escrito (Montolío et al., 2011; Montolío, 2012), para realizar un análisis contrastivo del discurso jurídico en diferentes lenguas (Cabré y Bach, 2004; Gómez Guinovart y Sacau Fontenla, 2007; Ordóñez López, 2008; Álvarez, 2008; Pontrandolfo, 2013a, 2013b; Cunillera-Domènech y Andújar-Moreno, 2017; Godoy Tena, 2017), o, de especial

interés en nuestro caso, para analizar el uso del Diccionario de la Real Academia Española como fuente de autoridad en las resoluciones del Tribunal Supremo (Henríquez Salido et al., 2013).

A pesar de los diferentes corpus y estudios realizados, lo que encontramos en este ámbito concreto es un panorama atomizado, donde, como bien ha sido destacado por Taranilla (2013, p. 313), "los estudios sobre el discurso producido en contextos jurídicos en español […] [parten] de un corpus elaborado específicamente para [una] investigación particular, que, por esa misma razón, acostumbra a tener un tamaño reducido." Asimismo, muchos de estos corpus no han sido publicados como recursos en abierto, estando su uso reservado exclusivamente a sus respectivos compiladores o grupos de investigación. De la detallada descripción de corpus españoles sobre discurso jurídico elaborada por Pontrandolfo (2013a) se colige que los corpus compartidos y disponibles para toda la comunidad científica son, en efecto, una minoría.

Así, entre los corpus cuyo material no se haya publicado en abierto, encontramos: el corpus JUD GENTT, elaborado por el grupo de investigación GENTT (Géneros Textuales para la Traducción), de la Universidad Jaume I de Castellón (Ordóñez López, 2008); el CORPUS del Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra de Barcelona (Cabré y Bach, 2004); el *Corpus de resoluciones judiciales para el Informe Lenguaje escrito*, realizado por el grupo de investigación Estudios de Discurso Académico y Profesional (EDAP) de la Universidad de Barcelona (Montolío et al., 2011), y el Corpus oral de resoluciones judiciales, elaborado por el Grupo Val.Es.Co. (Valencia, Español Coloquial) (2012) de la Universidad de Valencia, ambos parte integrante del *Informe de la Comisión de modernización del lenguaje jurídico*; el corpus JustClar, compilado en el marco del proyecto *Discurso jurídico y claridad comunicativa*, dirigido por Estrella Montolío Durán (Pontrandolfo, 2019); el corpus de sentencias judiciales del Tribunal Supremo, realizado por Henríquez Salido et al. (2013); el corpus elaborado por esta misma autora para el estudio del léxico en la jurisprudencia (Henríquez Salido y No Alonso-Misol, 2005; Henríquez Salido, 2007); el corpus utilizado por López Samaniego para analizar el uso de marcadores discursivos (2006); o el corpus bilingüe anglo-español desarrollado por Álvarez (2008), entre otros. Por su parte, entre los escasos corpus que

han sido publicados en abierto, cabe destacar: el Corpus Lingüístico da Universidade de Vigo (CLUVI) (Gómez Guinovart y Sacau Fontenla, 2007); el *Corpus de Procesos Penales* (CPP), compilado por Taranilla para su tesis doctoral (2011); o el corpus trilingüe español-italiano-inglés COSPE (Corpus de Sentencias Penales), elaborado por Pontrandolfo (2013a) para su trabajo de tesis doctoral.

La atomización del campo de estudio del lenguaje jurídico apuntada por Taranilla (2013) y el hecho de que la mayoría de los corpus existentes en dicho ámbito no se hallen disponibles en abierto implica que no contemos con corpus lo suficientemente extensos y, por lo tanto, representativos, del discurso jurídico español, que puedan ser consultados por cualquier investigador interesado en la materia. La mayoría de los estudios arriba mencionados emplean corpus que rondan un total de diez (Taranilla, 2009), veinte (Henríquez Salido, 2006, 2007; Álvarez, 2008; Cunillera-Domènech y Andújar-Moreno, 2017; Taranilla, 2015), cuarenta (Taranilla, 2014), cincuenta (López Samaniego, 2006) o, a lo sumo, cien documentos (Polanco Martínez y Yúfera Gómez, 2013; Godoy Tena, 2017). Incluso los más extensos no superan los 450 textos jurídicos en español[3]. Así, según el número de documentos, encontramos: el corpus JustClar (Pontrandolfo, 2019), con 241 sentencias del Tribunal Supremo y del Tribunal de Justicia de la Unión Europea, el corpus de la Comisión para la Modernización del Lenguaje Jurídico (Montolío et al., 2011), con 250 sentencias, autos, informes, oficios, decretos de admisión y demanda, citaciones, actas, notas y certificaciones de registradores y notarios, el Corpus de Sentencias Penales (COSPE) (Pontrandolfo, 2013a), con 262 sentencias del Tribunal Supremo y del Tribunal

[3] En el siguiente repaso, obviamos el corpus JUD GENTT (http://www.gentt.uji.es/) y el corpus CLUVI (http://sli.uvigo.es/CLUVI/index.php?lang=gl), por carecer de los datos exactos del número de textos jurídicos en español que incluyen. Se trata de corpus multilingües (JUD GENTT: catalán, español, inglés, alemán y francés; CLUVI: gallego, español, inglés, francés, portugués, catalán, italiano, euskera, alemán, latín, chino) compuestos por textos provenientes de diversos ámbitos especializados (JUD GENTT: textos jurídicos, médicos y técnicos; CLUVI: textos de ficción, informática, divulgación científica, bíblicos, de derecho y administración pública, información de consumo, economía y turismo), de los que solo conocemos el total de textos y palabras en JUD GENTT (900 textos y unos dos millones de palabras) (Ordóñez López, 2008), el total de palabras en CLUVI (49 millones de palabras) y el total de palabras de textos jurídicos en español, gallego y euskera de este último corpus (8.966.468) (Pontrandolfo, 2013a).

Superior de Justicia[4], y el CORPUS del IULA, de la Universitat Pompeu Fabra de Barcelona (Cabré y Bach, 2004), con 412 textos jurídicos de diferente índole[5].

Siguiendo la opinión de Taranilla (2013), consideramos que la superación de esta atomización de la investigación del discurso jurídico en español mediante corpus vendría a través de la elaboración de grandes corpus de textos jurídicos, compartidos en abierto con la comunidad científica, con el objetivo de ofrecer un material lo suficientemente representativo de uno o varios de sus diferentes géneros discursivos, que conformen la base de investigaciones empíricas futuras.

Por otra parte, de los corpus aquí recogidos, únicamente el de Hernández Salido et al. (2013) ofrece un material específico con el que abordar, si bien parcialmente, el problema al que nos referíamos al comienzo de este artículo: el empleo inadecuado del diccionario y la gramática por parte de los magistrados en las sentencias de los Tribunales Supremo y Constitucional. No obstante, consideramos que dicho corpus participa de algunos de los problemas arriba mencionados en relación con otros corpus jurídicos. En primer lugar, a pesar de que se indica que se han seguido tres criterios fundamentales de selección: uno cronológico (sentencias dictadas entre 1 de enero de 2011 y el 30 de septiembre 2012), otro temático (restricción por las palabras clave "Real Academia Española" y "DRAE") y otro referido al origen de las sentencias (Sala Primera de lo Civil, Segunda de lo Penal y Cuarta de lo Social del Tribunal Supremo) (Henríquez Salido et al., 2013), no se especifica el tamaño total del corpus, ya sea en número de textos, palabras o *tokens*, por lo que no podemos evaluar su grado de representatividad. Por otra parte, no hay mención alguna a una compilación electrónica

[4] En realidad, aludimos a uno de sus subcorpus, ya que este corpus multilingüe incluye sentencias en español, italiano e inglés, y su tamaño total es de 782 textos.

[5] El CORPUS del IULA incluye, en concreto, diversos tipos de "textos legislativos (leyes y reglamentos), textos de práctica profesional, textos judiciales (escritos de las partes, diligencias, autos interlocutorios, sentencias), textos teóricos (manuales, monografías, artículos en publicaciones (periódicos, documentación didáctica) y textos instrumentales (diccionarios jurídicos, vocabularios, glosarios)" (Pontrandolfo, 2013a, p. 118). Una vez más, aludimos solo a uno de sus subcorpus, el relativo a textos jurídicos en español, ya que el corpus total abarca textos en cinco lenguas (castellano, catalán, inglés, francés y alemán) y seis áreas de conocimiento (Derecho, Economía, Medio Ambiente, Medicina, Informática y Ciencias del Lenguaje), con un total de 1967 documentos.

del mismo ni tampoco a su publicación como recurso compartido, por lo que es de suponer que este no se encuentra disponible para otros investigadores que, como en nuestro caso, deseen abordar este problema específico de investigación.

3. Hipótesis y objetivos de la investigación

Partiendo, por tanto, de esta doble problemática, una de carácter general: la ausencia de grandes corpus de textos jurídicos compartidos, y una de índole específica: la necesidad de analizar empíricamente el uso del diccionario y la gramática en las sentencias de los Tribunales Supremo y Constitucional, el objetivo que planteamos en el presente artículo es, a su vez, doble.

Por una parte, ofrecer un corpus lo suficientemente representativo en tamaño, cronología y origen del material que abarca, de sentencias judiciales provenientes de dichos tribunales, cuyo acceso público permita ser utilizado por cualquier investigador o grupo interesado.

Por otra, ofrecer un material empírico específico al proyecto "Lengua y Derecho", en el que se incluye la presente investigación, con el que se pueda analizar, en una fase posterior, el empleo (adecuado o inadecuado) del conocimiento lingüístico que puede extraerse de los diccionarios y de las categorías morfosintácticas de las palabras.

4. Metodología

La elaboración del corpus RLD ha comprendido cuatro fases principales: (i) diseño, (ii) implementación, (iii) revisión y (iv) publicación[6].

4.1. Diseño del corpus

4.1.1. Establecimiento de los criterios estructurales

De acuerdo con Sinclair (2005), los criterios estructurales a partir de los que se seleccionan las muestras de lengua de un corpus deben ser exclusivamente externos y

[6] Seguimos, a este respecto, las fases establecidas por Sinclair (2005) para el diseño, creación e implementación de un corpus lingüístico.

no internos, es decir, estos han de derivar de un examen de la función comunicativa del texto y no de aspectos del lenguaje que conforma el texto. No obstante lo anterior, en la selección de la muestra para nuestro corpus se emplearon tanto criterios externos como internos para atender al tipo de análisis que se llevará a cabo en una investigación posterior, como más adelante se verá. Los criterios externos establecidos han sido los propuestos por Sinclair (2005): representatividad, equilibrio y homogeneidad. El único criterio interno que se ha tenido en cuenta es que los textos estén (o puedan estar) relacionados con el uso de los diccionarios y de las categorías morfosintácticas de las palabras como fuentes de autoridad para la interpretación de las resoluciones judiciales.

Para obtener una muestra lo más representativa posible se definieron los subcriterios externos de selección siguientes:

a. tipo de texto: sentencias judiciales,

b. origen: Tribunal Constitucional y Salas Civil, Contencioso-Administrativo, Penal, Militar y Social del Tribunal Supremo,

c. cronología: sentencias dictadas entre el 1 de enero de 2012 y el 31 de diciembre de 2018.

La limitación del corpus a un único género textual dentro de la jurisprudencia española, la sentencia, y a dos tribunales específicos, el Tribunal Supremo y el Constitucional, se fundamenta en dos motivos principales: uno de raíz social y otro de carácter práctico-metodológico.

En primer lugar, desde una perspectiva social, la sentencia es el género de mayor importancia, "el documento más relevante del proceso judicial", puesto que esta "reviste una trascendencia innegable tanto para el ciudadano (sobre cuya vida y patrimonio resuelve) como para la propia jurisprudencia" (Montolío et al., 2011, p. 10). Tal relevancia es aún mayor en las sentencias emitidas por los Tribunales Supremo y Constitucional, pues su repercusión para los ciudadanos es incomparable a la de cualquier otra instancia judicial. Se trata de órganos jurídicos únicos con potestad "en todo el territorio nacional", cuya autoridad está por encima del resto de tribunales en sus respectivos órdenes: en asuntos civiles, penales, contencioso-administrativos y sociales

en el caso del Supremo, y en materia de garantías y derechos constitucionales por lo que se refiere al Constitucional (Consejo General del Poder Judicial 2019). Todo ello hace que cualquier incorrección cometida por sus magistrados en la interpretación lingüística de leyes y pruebas judiciales tenga unas consecuencias mayores a las de cualquier otro órgano jurídico.

En segundo lugar, desde un punto de vista práctico-metodológico, consideramos que el hecho de contar con textos pertenecientes a un mismo género textual y provenientes de un mismo origen nos permitía asegurar el criterio de equilibrio arriba apuntado, al tiempo que el carácter uniforme de los textos, con un tamaño regular, en la mayor parte de ellos, de entre 1000 y 10.000 palabras, consolidaba la homogeneidad del corpus (Figura 1).

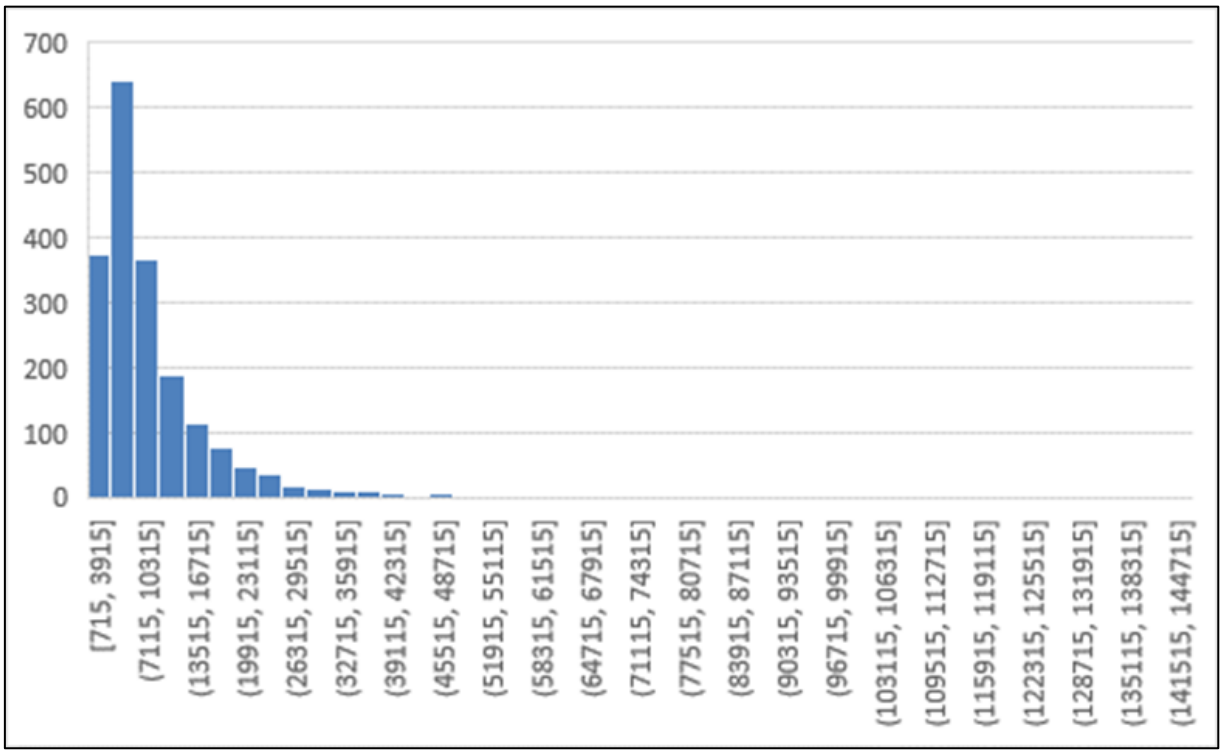

Figura 1. Histograma de sentencias por número de palabras*

* En el histograma, el eje X muestra el número de palabras de las sentencias en intervalos que van desde la sentencia con menor número de palabras (715) hasta la más extensa (144.715). El eje Y, por su parte, muestra la frecuencia absoluta de cada intervalo.

El criterio interno de selección de los textos relacionados con el uso de los diccionarios y de las categorías morfosintácticas de las palabras como fuente de

autoridad se establece de acuerdo con el siguiente criterio temático: la presencia de una o varias de las siguientes palabras clave:

a. aquellas que aluden a los diccionarios y principales diccionarios generales de la lengua española: "diccionario", "Real Academia Española", "RAE", "DRAE", "DUE", "Moliner" (*Diccionario de uso del español* de María Moliner), "Seco" (*Diccionario del Español Actual* de Manuel Seco et al.), "DA" (*Diccionario de Autoridades*) y "DPD" (*Diccionario Panhispánico de Dudas*).

b. Aquellas que se refieren a las categorías morfosintácticas de las palabras: "género", "nombre", "sustantivo", "adjetivo", "verbo", "adverbio", "preposición", "conjunción" y "pronombre".

4.2. Implementación del corpus

4.2.1. Extracción de la muestra

Para la extracción de las sentencias de la muestra de acuerdo con los criterios de selección establecidos en el diseño, se utilizaron la base de datos de *Tirant Online*[7], de la editorial Tirant, y la de la editorial Aranzadi-Thompson[8]. Cada sentencia se extrajo en formato .pdf[9] con su identificador, el número de recurso, como nombre del archivo .pdf, para facilitar su identificación[10]. Asimismo, se elaboró, de forma paralela, una tabla en formato .xlsx para facilitar la búsqueda y análisis de cada componente de la muestra (Figura 2). La tabla contiene la siguiente información de cada sentencia: 1) Nº de recurso; 2) Origen (Tribunal Supremo o Constitucional); 3) Jurisdicción (Sala Civil,

[7] https://www-tirantonline-com.bucm.idm.oclc.org/tol/

[8] http://www.aranzadidigital.es/maf/app/authentication/signon?legacy

[9] Las bases de datos de Tirant y Aranzadi solo permiten exportar las sentencias en formatos pdf. o word.

[10] Los Tribunales Supremo y Constitucional emiten sentencias sobre un recurso (de casación, de inconstitucionalidad, etc.) que se ha interpuesto a partir de una sentencia anterior (de una audiencia provincial o de cualquier otro tribunal dependiente del Supremo y del Constitucional). Por eso, las sentencias siempre se identifican con el número de recurso y, en algunos casos, también con el número de sentencia o resolución. En el corpus se ha utilizado como identificador único el número de recurso, ya que es el que está siempre presente y sirve para identificar cada sentencia concreta.

Sala Contencioso-Administrativo, Sala de lo Penal, Sala de lo Militar y Sala de lo Social); 4) Fecha de emisión; 5) Nombre del ponente.

	A	B	C	D	E
1	Nº de Recurso	Origen	Jurisdicción	Fecha	Ponente
2	931/2008	Tribunal Supremo	Civil	05/01/2012	Rafael Gimeno-Bayón Cobos
3	101/2011	Tribunal Supremo	Militar	10/01/2012	Francisco Javier de Mendoza Fernández
4	894/2009	Tribunal Supremo	Civil	10/01/2012	Juan Antonio Xiol Ríos
5	2120/2009	Tribunal Supremo	Civil	11/01/2012	Francisco Marín Castán
6	16/2010	Tribunal Supremo	Contencioso-Administrativo	12/01/2012	Juan José González Rivas
7	23/2010	Tribunal Supremo	Contencioso-Administrativo	12/01/2012	Rafael Fernández Montalvo
8	1362/2008	Tribunal Supremo	Contencioso-Administrativo	12/01/2012	Eduardo Calvo Rojas
9	1665/2008	Tribunal Supremo	Contencioso-Administrativo	12/01/2012	Eduardo Calvo Rojas
10	1670/2010	Tribunal Supremo	Contencioso-Administrativo	12/01/2012	Ángel Aguallo Avilés
11	1779/2008	Tribunal Supremo	Civil	12/01/2012	Encarnacion Roca Trias
12	2833/2008	Tribunal Supremo	Contencioso-Administrativo	12/01/2012	Manuel Martín Timón
13	3883/2010	Tribunal Supremo	Contencioso-Administrativo	12/01/2012	José Antonio Montero Fernández
14	122/2009	Tribunal Supremo	Contencioso-Administrativo	13/01/2012	Santiago Martínez-Vares García
15	2238/2008	Tribunal Supremo	Civil	13/01/2012	José Ramón Ferrándiz Gabriel
16	6621/2009	Tribunal Supremo	Contencioso-Administrativo	13/01/2012	María Isabel Perelló Domenech
17	820/2009	Tribunal Supremo	Contencioso-Administrativo	16/01/2012	Manuel Campos Sánchez-Bordona
18	1413/2008	Tribunal Supremo	Civil	16/01/2012	Juan Antonio Xiol Ríos
19	4678/2011	Tribunal Supremo	Contencioso-Administrativo	16/01/2012	Agustín Puente Prieto
20	4907/2011	Tribunal Supremo	Contencioso-Administrativo	16/01/2012	Agustín Puente Prieto
21	27/2011	Tribunal Supremo	Social	17/01/2012	Jesús Souto Prieto
22	2142/2010	Tribunal Supremo	Contencioso-Administrativo	17/01/2012	Óscar González González

Figura 2. Tabla con los datos principales de cada sentencia

La extracción se realizó en dos etapas. En la primera se aplicaron los criterios externos de representatividad, equilibrio y homogeneidad. Al aplicar estos criterios, obtuvimos un total de 44.022 sentencias, con un número aproximado de 1000-10.000 palabras por sentencia. La distribución por Tribunales y Salas se muestra en la tabla 1.

Tabla 1. Distribución del total de sentencias de los Tribunales Supremo y Constitucional emitidas entre 2012 y 2018

Tribunal Supremo	Civil	5344
	Contencioso-Administrativo	22517
	Militar	1112
	Penal	6474
	Social	7097
Tribunal Constitucional		1478
Total		**44022**

En la segunda etapa se aplicó el criterio de selección interna quedando la muestra final en 1940 sentencias. Esto significa que 1940 sentencias de los Tribunales Constitucional y Supremo utilizan conocimiento lingüístico (diccionarios o mención a categoría morfosintáctica) de forma explícita, lo que constituye el 4,4% del total de sentencias emitidas por ambos tribunales en los años 2012 a 2018. La tabla 2 muestra la distribución numérica por años, tribunales y salas. Finalmente, la figura 3 establece una comparación entre la distribución por Tribunal y Sala del total de sentencias emitidas en el periodo 2012-2018 y del total de sentencias recogidas en el corpus. Como puede comprobarse, las sentencias del ámbito Contencioso-Administrativo son las más numerosas del corpus (730), suponiendo, aproximadamente, un 37% del total de sentencias recopiladas (1940). Una distribución que se halla en consonancia con los datos totales del periodo considerado, pues las sentencias de este ámbito (22517) constituyen, aproximadamente, el 51% del total (44022).

Tabla 2. Distribución por años, Tribunales y Salas de la muestra final

	Tribunal Supremo					Tribunal Constitucional	**Total**
Año	Civil	Contencioso-Administrativo	Penal	Militar	Social		
2012	77	368	188	38	70	33	774
2013	30	62	86	16	18	7	219
2014	30	46	62	16	33	3	190
2015	18	59	60	16	26	5	184
2016	23	49	76	8	19	13	188
2017	16	60	59	14	12	10	171
2018	10	86	65	11	32	10	214
Total	**204**	**730**	**596**	**119**	**210**	**81**	**1940**

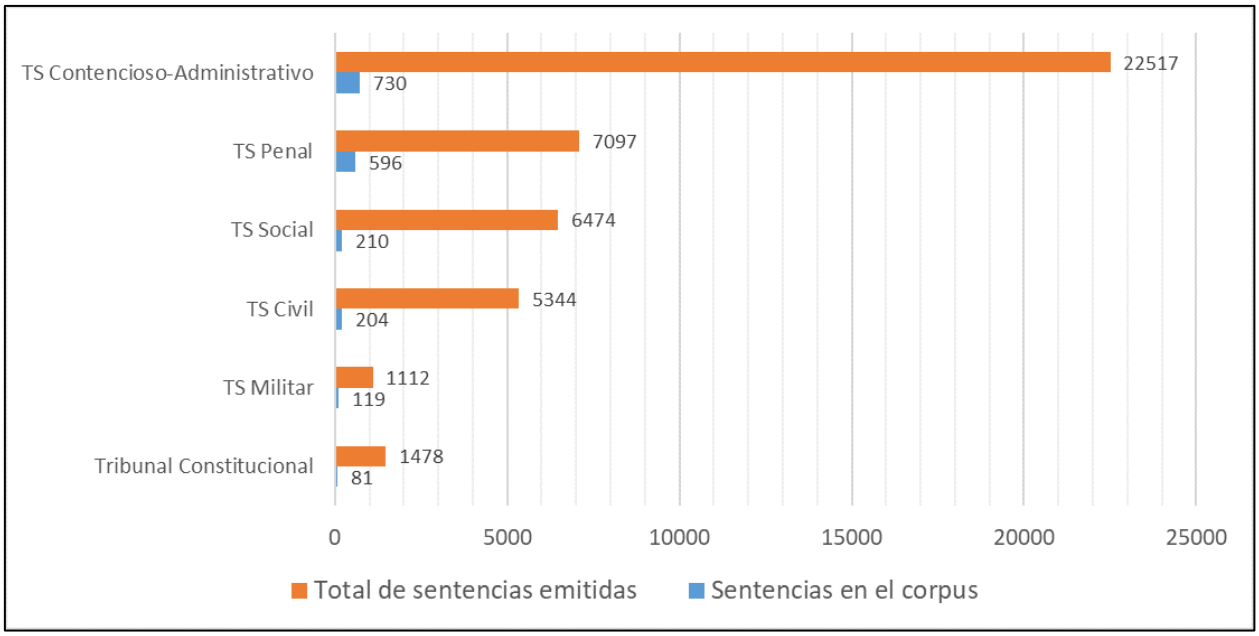

Figura 3. Distribución comparativa del total de sentencias emitidas y sentencias en el corpus por Tribunal y Sala en el periodo 2012-2018

4.2.2. Compilación del corpus

A la hora de compilar el corpus, se utilizó la herramienta de análisis textual en línea *Sketch Engine*[11]. Esta herramienta permite la compilación automática de corpus a partir de archivos de texto en diversos formatos (Arias et al., 2019, p. 16), entre ellos .pdf (Acrobat PDF) y MS .xlsx (MS Excel). A través de ella, se importó el total de 1940 sentencias de la muestra en .pdf, se compilaron y se obtuvo el corpus en formato "texto puro" (con extensión .txt). La compilación en texto puro, un formato que consiste únicamente en secuencias lineales de caracteres (que conforman, típicamente, palabras, números y signos de puntuación) (Sinclair, 2005; Arias Rodríguez et al., 2019, p. 3), asegura que el corpus pueda ser utilizado para el mayor tipo de análisis lingüísticos y con el mayor tipo de software de análisis textual posibles, sin necesidad de llevar a cabo un laborioso proceso previo de conversión de formatos (Sinclair, 2005; Wynne, 2005b; Leech, 2005).

[11] https://auth.sketchengine.eu

El archivo de texto puro resultado de la compilación inicial contiene un total de 19.531.802 palabras y tiene un tamaño de 124 Megabytes (Figura 4).

```
Cabecera: PROYECTO DE PARQUE EÓLICO EL RELUMBRAR: MODIFICACIÓN DE SU
APROBACIÓN Y DECLARACIÓN DE UTILIDAD PÚBLICA DE LOS VIALES DE ACCESO,
DE SERVICIO Y ENTRE AEROGENERADORES, LÍNEAS ELÉCTRICAS SUBTERRÁNEAS Y
PLATAFORMAS DE MONTAJE. IMPUGNACIÓN INDIRECTA DE DECRETO 58/99 DE
CASTILLA-LA MANCHA.
Jurisdicción: Contencioso-Administrativo
Ponente: Eduardo Espín Templado
Origen: Tribunal Supremo
Fecha: 24/02/2015
Tipo resolución: Sentencia
Sala: Tercera
Sección: Tercera
Número Recurso: 5777/2011

ENCABEZAMIENTO:
SENTENCIA
En la Villa de Madrid, a veinticuatro de Febrero de dos mil quince.
VISTOS por la Sala de lo Contencioso-Administrativo del Tribunal Supremo, constituída en su
Sección Tercera por los Magistrados indicados al margen, los recursos de casación tramitados bajo el
número 5.777/2.011, interpuestos por la JUNTA DE COMUNIDADES DE CASTILLA-LA
MANCHA, representada por el Procurador D. Francisco Velasco Muñoz-Cuéllar, y por ENEL
GREEN POWER ESPAÑA, S.A., representada por la Procuradora Dª Pilar Iribarren Cavalle, contra
la sentencia dictada por la Sección Segunda de la Sala de lo Contencioso-Administrativo del Tribunal
Superior de Justicia de Castilla-La Mancha en fecha 8 de julio de 2.011 en el recurso
contencioso-administrativo número 468/2.006 , sobre modificación de la aprobación del proyecto del
parque eólico El Relumbrar y declaración de utilidad pública de los viales de acceso, de servicio y
entre aerogeneradores, de las líneas eléctricas subterráneas y de las plataformas de montaje.
Son partes recurridas VIÑEDOS BALMORAL, S.L., representada por el Procurador D. Juan
Torrecilla Jiménez, la JUNTA DE COMUNIDADES DE CASTILLA-LA MANCHA, representada
por el Procurador D. Francisco Velasco Muñoz-Cuéllar, y ENEL GREEN POWER ESPAÑA, S.A.,
representada por la Procuradora Dª Pilar Iribarren Cavalle.
ANTECEDENTES DE HECHO:
PRIMERO .- En el proceso contencioso-administrativo antes referido, la Sala de lo
Contencioso-Administrativo (Sección Segunda) del Tribunal Superior de Justicia de Castilla-La
Mancha dictó sentencia de fecha 8 de julio de 2.011 , por la que se estimaba parcialmente el recurso
```

Figura 4. Muestra del archivo de texto puro

El archivo del corpus fue documentado con los metadatos XML del *Text Encoding Initiative*, TEI[12] (Figura 5). El modelo de anotación TEI constituye, actualmente, un estándar de facto para el etiquetado de textos académicos. En el caso de los metadatos contiene seis tipos[13]:

1. metadatos administrativos: indican el nombre del corpus <titleStmt /title>, autor del corpus <titleStmt/respStmt>; organismo e institución que financia el proyecto de creación del corpus < titleStmt/funder><titleStmt/sponsor>, investigador principal del

[12] Puede consultarse el repertorio de metadatos TEI en el siguiente enlace: https://tei-c.org/

[13] Nos basamos aquí en la clasificación de metadatos que establece Burnard (2005) a partir de las recomendaciones TEI (*Text Encoding Initiative*).

proyecto <titleStmt/principal>, disponibilidad y restricciones de uso <publicationStmt>, y fecha de publicación <publicationStmt/date>,

2. metadatos editoriales: para indicar la fuente de origen de los textos <sourceDesc>,

3. metadatos de codificación: indican los objetivos del proyecto de investigación <projectDesc>, y los criterios y métodos que se han empleado para seleccionar y compilar las muestras que conforman el corpus<samplingDecl>,

4. metadatos descriptivos: con información sobre las características de los textos <textDesc>, número de textos individuales y palabras dentro del corpus <extent>, idioma <langUsage> y periodo cronológico de producción de los textos <profileDesc/creation/date>,

5. metadatos identificativos (nombre de cada sentencia <filename>[14]) y, finalmente,

6. metadatos de revisión: fecha en la que se realizó cada una de las modificaciones del corpus <revisionDesc>.

```
<?xml version="1.0" encoding="UTF-8"?>
<teiCorpus xmlns="http://www.tei-c.org/ns/1.0">
    <teiHeader xml:lang="es-ESP">
        <fileDesc>
            <titleStmt>
                <title>Corpus Lengua y Derecho</title>
                <respStmt>
                    <resp>compilado y anotado por</resp>
                    <name>Juan Manuel Díaz Ayuga</name>
                </respStmt>
                <funder>Comunidad de Madrid, proyecto "Lengua y Derecho" (PEJD-2018-PRE/HUM-8838).
                    Iniciativa Empleo Juvenil (YEI), cofinanciada en un 91,89% por el Fondo Social Europeo
                </funder>
                <sponsor>Área de Lingüística General. Departamento de Lingüística, Estudios Árabes,
                    Hebreos y de Asia Oriental. Facultad de Filología. Universidad Complutense de
                    Madrid</sponsor>
                <principal>Ángel Alonso-Cortés</principal>
                <principal>Ana Fernández-Pampillón Cesteros</principal>
            </titleStmt>
            <extent>
                <measure unit="MiB" quantity="123">Aproximadamente 123 megabytes</measure>
                <measure unit="doc" quantity="1940">1940 documentos</measure>
                <measure unit="word" quantity="19528210">19.528.210 palabras</measure>
            </extent>
            <publicationStmt>
                <publisher>Universidad Complutense de Madrid</publisher>
                <date>2019</date>
```

[14] Al estar identificado cada documento por su número de recurso, puede rastrearse fácilmente toda su información en la tabla en formato .xlsx (Figura 1).

Figura 5. Extracto de los metadatos TEI del corpus Lengua y Derecho

4.2.3. Anotación del corpus

Para facilitar el análisis lingüístico, el corpus se anotó con la categoría morfosintáctica y el lema de cada *token*. Esta información es imprescindible para poder localizar o distinguir entre palabras con una misma forma, pero pertenecientes a categorías morfosintácticas distintas, para localizar todas las palabras de una misma clase o las formas de un mismo lema. Así, en el marco del proyecto "Lengua y Derecho", en el que se enmarca la creación de este corpus, es necesario localizar únicamente los sustantivos sobre los que el magistrado realiza un comentario metalingüístico.

Para el etiquetado del corpus empleamos, nuevamente, la herramienta *Sketch Engine*. Esta herramienta incluye un etiquetador morfosintáctico, *FreeLing*, para varias lenguas, entre las que se incluye el español. *FreeLing* analiza morfológica y sintácticamente las frases del corpus asignando etiquetas del esquema de anotación propuesto por EAGLES (*Expert Advisory Group on Language Engineering Standards*), con una precisión de hasta un 97% de acierto (Arias Rodríguez et al., 2019). Las reglas de anotación están definidas en un archivo de texto que puede ser modificado para adaptarse al tipo de textos que se desea analizar (Figura 6).

Figura 6. Información del corpus compilado y etiquetado en *Sketch Engine*

El corpus procesado y etiquetado es un archivo de "texto vertical" (con extensión .vert). El texto vertical es texto organizado en columnas que contiene una fila por cada *token* (típicamente palabras) y por cada estructura identificada en el archivo original (típicamente frases y párrafos). El archivo etiquetado en *Sketch Engine*, contiene, en concreto, tres columnas (Sketch Engine, 2019) (Figura 7):

1. en la primera columna se almacenan los *tokens* (uno por cada línea de texto) del corpus,

2. en la segunda columna, las etiquetas mofosintácticas (en inglés *part-of-speech*, POS o clase de palabra) asignadas a cada *token* por el analizardor *FreeLing*, es decir, una anotación de cada *token* con información sobre la clase de palabra a la que pertenece e información morfológica relevante (número, género, tiempo verbal, etc.)[15]; y,

3. en la tercera columna, una lematización del mismo, es decir, una asignación del lema correspondiente a cada forma de palabra del corpus, utilizando etiquetas "lempos", que combinan el lema con la clase de palabra a la que pertenece (p.e., *ir-v*, *casa-n*).

El archivo vertical resultante contiene, además, los metadatos ya mencionados (que se incluyeron en el corpus en formato .txt) junto con nuevos metadatos analíticos que informan sobre el método y tipo de etiquetado aplicado.

[15] Puede consultarse el repertorio de etiquetas POS utilizado por *FreeLing* en este enlace: https://www.sketchengine.eu/spanish-freeling-part-of-speech-tagset/

```
<doc parent_folder="Corpus sentencias judiciales (2).zip"
id="file12107445" filename="Corpus sentencias judiciales
(2)/10001_2016.pdf">
<s>
http://www.tirantonline.comNCFS000
      http://www.tirantonline.com-n
      http://www.tirantonline.comNCFS000
      http://www.tirantonline.com
</s>
<s>
Documento  NCMS000    documento-n      documento  NCMS000
      documento
TOL5.776.076     NP00000    tol5.776.076-n   tol5.776.076
      NP00000    tol5.776.076
</s>
<s>
Jurisprudencia   NCFS000    jurisprudencia-njurisprudencia
      NCFS000    jurisprudencia
Cabecera   NPFS000    cabecera-n cabecera   NPFS000    cabecera
<g/>
:      Fd    :-x   :      Fd    :
-      Fg    --x   -      Fg    -
Robo  NPMS000    robo-n     robo  NPMS000    robo
con   SP    con-i con   SP    con
violencia  NCFS000    violencia-n      violencia  NCFS000
      violencia
e      CC    y-c   y      CC    y
intimidaciÃ³n    NCFS000    intimidaciÃ³n-n intimidaciÃ³n
      NCFS000    intimidaciÃ³n
ejecutado  VMP00SM    ejecutar-v ejecutar   VMP00SM    ejecutar
```

Figura 7. Extracto del corpus compilado como archivo de texto vertical etiquetado

4.3. Revisión del corpus

El cumplimiento de los criterios externos y el criterio interno se verificó consultando de forma semiautomática el corpus utilizando la herramienta de análisis textual *Sketch Engine*. Las comprobaciones que se realizaron fueron:

1. número de palabras por sentencia judicial, Tribunal y Sala: permite comprobar la homogeneidad y equilibrio del corpus,

2. sentencias en las que aparece, como mínimo, una de las palabras clave del criterio de selección interna: permite comprobar si es correcta la selección de las sentencias respecto al criterio interno de selección, es decir, si el corpus es representativo del uso de conocimiento lingüístico en la jurisprudencia de los Tribunales Constitucional y Supremo,

3. concordancias de las palabras clave del criterio de selección interna: permite comprobar que la aparición de las palabras clave es correcta, es decir, realmente indican que se está utilizando conocimiento lingüístico en cada sentencia del corpus. Completa la comprobación de corrección y representatividad del corpus realizada en (2).

Finalmente, es importante señalar que esta verificación de la calidad del corpus fue realizada una sola vez y por un solo investigador.

4.4. Publicación del corpus

Con el fin de que el corpus se halle disponible para futuras investigaciones, ambas versiones del corpus, como texto puro y anotado, así como la tabla Excel para la búsqueda de cada sentencia y un archivo de muestra en formato XML-TEI, se encuentran a disposición de los investigadores en el espacio Google Drive institucional de la Universidad Complutense de Madrid: https://drive.google.com/drive/folders/1CF6sTuhyO3lFI30sxkjsPlXy7wQVe5m9?usp=sharing.

5. Conclusiones y líneas de trabajo futuro

Este trabajo aporta, hasta donde los autores saben, el primer corpus en español en abierto para el estudio del uso del conocimiento lingüístico en el ámbito judicial, concretamente el que puede extraerse de los diccionarios y de las categorías morfosintácticas de las palabras. Además, al tener un tamaño de 1.940 sentencias y 19.531.802 palabras, el corpus constituye uno de los corpus en español más amplios sobre Lengua y Derecho.

El corpus se ha creado atendiendo a los criterios de representatividad, homogeneidad y equilibrio, así como al criterio interno temático de que los textos deben estar relacionados con el uso que realizan los magistrados de la gramática y de los diccionarios como fuente de autoridad para la interpretación jurídica. Finalmente, el corpus se ha creado utilizado formatos y directrices estándares: texto, texto vertical, anotación morfosintáctica conforme el estándar EAGLES y metadatos del *Text Enconding Iniative*. De esta forma se procura que sea lo más portable e interoperable

posible para facilitar al máximo su reutilización y escalado en nuevos trabajos académicos.

Este trabajo permitirá abordar el estudio empírico del uso del conocimiento lingüístico como fuente de autoridad en las sentencias de los Tribunales Supremo y Constitucional. Además, el corpus puede utilizarse en nuevas líneas de investigación, bien en su versión original o bien, aplicando la metodología de construcción presentada, adaptándolo a nuevas cuestiones de investigación.

Financiación

El presente trabajo, enmarcado en el proyecto "Lengua y Derecho" (PEJD-2018-PRE/HUM-8838) ha sido financiado por la Comunidad de Madrid con la Iniciativa Empleo Juvenil (YEI), subvencionada en un 91,89% por el Fondo Social Europeo.

Agradecimientos

Los autores expresan su agradecimiento a Mabel López Medina, bibliotecaria de la Facultad de Derecho de la UCM, su información sobre los distintos corpus de sentencias, y especialmente a las editoriales *Tirant* y *Aranzadi* la autorización para la publicación del corpus en abierto, de forma que pueda ser utilizado por cualquier profesor o investigador.

Bibliografía

Álvarez, Susana (2008). Elementos cohesivos en el lenguaje jurídico: análisis contrastivo de las sentencias judiciales en lengua inglesa y española. En Luis Pegenaute, Janet Ann DeCesaris, Mercedes Tricás Preckler & Elisenda Bernal (Eds.), *La traducción del futuro: mediación lingüística y cultural en el siglo XXI* (pp. 407-418). Barcelona: Promociones y Publicaciones Universitarias, PPU.

Arias Rodríguez, Iván, Ana Fernández-Pampillón, Doaa Samy y Jorge Arús Hita (2019). *Taller sobre herramientas de análisis textual: la herramienta Sketch Engine*. Madrid: Universidad Complutense de Madrid. Disponible en:

https://eprints.ucm.es/13796/9/Taller_de_Sketch_Engine_18_02_2019.pdf [Último acceso: 30/12/2019]

Barbera, Manuel, Elisa Corino y Cristina Onesti (2007). Cosa è un corpus? Per una definizione più rigorosa di corpus, token, markup. En M. Barbera, E. Corino y C. Onesti (Eds.), *Corpora e linguistica in rete* (pp. 25-88). Perugia: Guerra Edizioni.

Bayo Delgado, Joaquín (1996). La formación básica del ciudadano y el mundo del Derecho. Crítica lingüística del lenguaje judicial. *Revista de Llengua i Dret*, 25, pp. 51-72. Disponible en: https://libros-revistas-derecho.vlex.es/vid/formacion-basica-ciudadano-mundo-lenguaje-76844519?_ga=2.8759613.980520530.1580989236-480286434.1580989236 [Último acceso: 06/02/2020].

Boletín Oficial del Estado (2019). *Código Civil y legislación complementaria*. Madrid: Agencia Estatal Boletín Oficial del Estado.

Briz, Antonio y el Grupo Val.Es.Co. (2012). El discurso judicial oral a partir de un análisis de corpus. En Estrella Montolío Durán (Ed.), *Hacia la modernización del discurso jurídico* (pp. 39-64). Barcelona: Edicions i publicacions de la Universitat de Barcelona.

Burnard, Lou (2005). Metadata for Corpus Work. En M. Wynne (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books. Disponible en: http://hdl.handle.net/20.500.12024/2951 [Último acceso: 30/12/2019].

Capella, Juan R. (1968). *El Derecho como lenguaje*, Barcelona: Ariel.

Cabré, Teresa y Carme Bach (2004). El corpus tècnic del IULA: corpus textual especializado plurilingüe. *Panace@*, 16, pp. 173-176.

Consejo General del Poder Judicial (2019). Tribunal Supremo. *Poder Judicial España*. Disponible en: http://www.poderjudicial.es/cgpj/es/Poder-Judicial/Tribunal-Supremo/ [Último acceso: 06/01/2019]

Cunillera-Domènech, Montserrat y Gemma Andújar-Moreno (2017). La expresión lingüística de la valoración en textos jurisprudenciales: Estudio contrastivo francés-español. *Revista signos*, 50.94, pp. 174-194.

Godoy Tena, Francisco (2017). *Análisis macroestructural comparado de un corpus digital bilingüe (inglés-español) de 100 sentencias judiciales británicas y españolas de Primera Instancia y de Instancia Apelativa* (Tesis doctoral). Esther Inmaculada Vázquez y del Árbol (Dir.). Madrid: Universidad Autónoma de Madrid.

Gómez Guinovart, Xavier y Elena Sacau Fontenla (2007). Técnicas de procesamiento lingüístico-computacional de corpus paralelos en "CLUVI" (Corpus Lingüístico da Universidade de Vigo). En Pablo Cano López (Ed.), *Actas del VI Congreso de Lingüística General: Santiago de Compostela, 3-7 de mayo de 2004* (pp. 855-864). Madrid: Arco Libros.

González Salgado, José Antonio (2011). La elección lingüística como fuente de problemas jurídicos. *Revista de Llengua i Dret*, 55, pp. 57-79.

Henríquez Salido, María do Carmo y No Alonso-Misol, Enrique de (2005). *Pautas para el análisis del léxico de la jurisprudencia del Tribunal Supremo*. Madrid: Thomson-Civitas.

Henríquez Salido, Maria do Carmo (2006). Las condiciones de producción y de interpretación de las sentencias de la Sala de lo Penal del Tribunal Supremo. *Revista de Llengua i Dret*, 45, pp. 33-60.

_____ (2007). Los adjetivos calificativos en las sentencias de la Sala de lo Social del Tribunal Supremo. *Revista de Investigación Lingüística*, 10, pp. 101-120.

Henríquez Salido, María do Carmo, Fernando Alañón Olmedo, Josefa Otero Seivane y Pedro F. Rabanal Carbajo (2013). La *auctoritas* de la Real Academia Española en las resoluciones del Tribunal Supremo. *BRAE*, Tomo XCIII, Cuaderno CCCVII, enero-junio, pp. 95-123.

Hernández Gil, Antonio (1989). *Saber jurídico y lenguaje*, Madrid: Espasa.

Kalinowski, George (1973). *Introducción a la lógica jurídica*, Buenos Aires: Olejnik.

Leech, Geoffrey (2005). Adding Linguistic Annotation. En M. Wynne (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books. Disponible en: http://hdl.handle.net/20.500.12024/2951 [Último acceso: 10/02/2020].

López Samaniego, Anna (2006). Los ordenadores del discurso enumerativos en la sentencia judicial ¿estrategia u obstáculo?. *Revista de Llengua i Dret*, 45, pp. 61-88.

Montolío, E., M.ª Á. García Asensio, P. Gras, A. López Samaniego, F. Polanco, R. Taranilla e I. Yúfera (2011). *Estudio de campo. Lenguaje escrito. Comisión para la modernización del lenguaje jurídico*. Madrid: Ministerio de Justicia.

Montolío Durán, E (2012). La situación del discurso jurídico español. Estado de la cuestión y algunas propuestas de mejora. En E. Montolío (Ed.), *Hacia la modernización del discurso jurídico* (pp. 65-94). Barcelona: Edicions i publicacions de la Universitat de Barcelona.

Ordóñez López, Pilar (2008). El Proyecto GENTT. Investigación en traducción: géneros y corpus. *Fòrum de Recerca*, 13, pp. 365-371.

Polanco Martínez, Fernando (2011). Construcción parafrástica y legibilidad en la sentencia judicial: descripción y propuesta de optimización. En María Luisa Carrió, Josefa Contreras, Françoise Olmo, Hanna Skorczynska, Inmaculada Tamarit, Debra Westall (Eds.), L*a investigación y la enseñanza aplicadas a las lenguas de especialidad y a la tecnología* (pp. 109-118). Valencia: Editorial Universitat Politècnica de València.

Polanco Martínez, Fernando e Irene Yúfera Gómez (2013). La construcción parafrástica en las sentencias judiciales. Una propuesta de optimización del discurso. *Revista de educación y derecho. Education and law review*, 7, pp. 1-19.

Pontrandolfo, Gianluca (2013a). *La fraseología en las sentencias penales: un estudio contrastivo español, italiano, inglés basado en corpus* (Tesis doctoral). Helena Lozano Miralles (Dir.). Trieste: Università degli Studi di Trieste.

_____ (2013b). La fraseología como estilema del lenguaje judicial: el caso de las locuciones prepositivas desde una perspectiva contrastiva. En Luisa Chierichetti y Giovanni Garofalo (Eds.), *Discurso Profesional y Lingüística de Corpus: Perspectivas de investigación* (pp. 187-215). Bergamo: Centro di Ricerca sui Linguaggi Specialistici, Università di Bergamo.

_____ (2019). Gerundios 'revelando' normalización en el lenguaje judicial español: consideraciones a partir del corpus JustClar. *Orillas*, 8, pp. 725-749.

Prieto de Pedro, Jesús (1996). La exigencia de un buen lenguaje jurídico y estado de derecho. *Revista de administración pública*, 140, pp. 111-130.

Real Academia Española (RAE) (2014). *Diccionario de la lengua española*. Madrid: Espasa.

Sánchez Rubio, María Aquilina (2004). La interpretación en el derecho: *in claris non fit interpretatio*. *Anuario de la Facultad de Derecho (Universidad de Extremadura)*, 22, pp. 417-435.

Sinclair, John (2005). Corpus and Text - Basic Principles. En M. Wynne (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books. Disponible en: http://hdl.handle.net/20.500.12024/2951 [Último acceso: 30/12/2019].

Sketch Engine (2019). *Sketch Engine User Guide*. Lexical Computing CZ. Disponible en: https://www.sketchengine.eu/guide/ [Último acceso: 30/12/2019]

Taranilla, Raquel (2009). La gestión de la propia imagen en las argumentaciones del Tribunal Constitucional: la función retórica de las estrategias de cortesía. *Revista de Llengua i Dret*, 52, pp. 117-149.

_____ (2011). *La configuración narrativa en el proceso penal. Un análisis discursivo basado en corpus* (Tesis doctoral). Estrella Montolío Durán (Dir.). Barcelona: Universitat de Barcelona.

_____ (2013). Aspectos metodológicos en la confección de un corpus jurídico. Consideraciones a propósito del Corpus de Procesos Penales. *Revista de Investigación Lingüística*, 16, pp. 311-341.

_____ (2014). La variación en la sentencia judicial: hacia una descripción exhaustiva del género. *Comparative Legilinguistics*, 20, pp. 31-52.

_____ (2015). El género de la sentencia judicial: Un análisis contrastivo del relato de hechos probados en el orden civil y en el orden penal. *Ibérica, Revista de la Asociación Europea de Lenguas para Fines Específicos*, 29, pp. 63-82.

Tognini-Bonelli, Elena (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Von Savigny, Friedrich (1840-1849). *System des heutigen römischen Rechts*. Berlin: Bei Veit und Comp.

Wynne, Martin (2005a). Preface. En M. Wynne (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books. Disponible en: http://hdl.handle.net/20.500.12024/2951 [Último acceso: 30/12/2019].

_____ (2005b). Archiving, Distribution and Preservation. En M. Wynne (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books. Disponible en: http://hdl.handle.net/20.500.12024/2951 [Último acceso: 30/12/2019].

RLD Corpus: the corpus of linguistic resources in Spanish jurisprudence

Abstract

This article explains the RLD corpus, a collection of cases where the Spanish High Court and the Constitutional Court Lords or Ladies Justice ("Jueces del Supremo") have turned to Spanish dictionaries and grammar as the linguistic authority to write court judgments. The four main stages followed in the building process of the corpus, (i) design, (ii) implementation, (iii) revision and (iv) publication, are thoroughly explained. The aim of this corpus, which includes 1940 court judgments issued between the years 2012 and 2018, is to provide a representative and empirically adequate sample of data of the use of dictionaries and grammar as linguistic authority in Spanish jurisprudence. This corpus comes to fill a gap in the studies of Language and Law: on the one hand, the lack of large shared corpora of legal texts, available for any researcher or research group interested in this issue, and, on the other hand, the need to analyze empirically the use of dictionaries and grammar in the court judgments of the Spanish High Court and the Constitutional Court.

Keywords: corpora, linguistic resources, judicial sentences, Spanish jurisprudence, Supreme Court, Constitutional Court