Navegación de corpus a través de anotaciones lingüísticas automáticas obtenidas por Procesamiento del Lenguaje Natural: de anecdótico a ecdótico

Pablo Ruiz, Helena Bermúdez Sabel

Revista de Humanidades Digitales

Navegación de corpus a través de anotaciones lingüísticas automáticas obtenidas por Procesamiento del Lenguaje Natural: de anecdótico a ecdótico

Corpus Navigation through Automatic Linguistic Annotations Obtained by Natural Language Processing: From Anecdotic to Ecdotic

Pablo RUIZ FABO
Universidad de Estrasburgo
ruizfabo@unistra.fr
https://orcid.org/0000-0002-4349-4835

Helena BERMÚDEZ SABEL
Universidad de Lausanne
helena.bermudezsabel@unil.ch
https://orcid.org/0000-0002-8627-1367

ABSTRACT

The article presents two case studies on the application of Natural Language Processing (NLP) technologies or Computational Linguistics to create corpus navigation interfaces. These interfaces help access relevant information for specific research questions in Social Sciences or Humanities. The paper also focuses on how these technologies for automatic text analysis can allow us to enrich scholarly digital editions. The theoretical framework that connects the aforementioned technologies with academic digital edition is described, and a reflection is made on the application of such technologies and interfaces to digital scholarly editing.

KEYWORDS

Natural Language Processing (NLP), Digital Edition, Automatic Text Analysis.

RESUMEN

En el presente artículo se describen dos estudios de caso de aplicación de tecnologías de Procesamiento del Lenguaje Natural (PLN) o lingüística computacional para crear interfaces de navegación de corpus, las cuales facilitan el acceso a información de relevancia particular para cuestiones de investigación en ciencias sociales o humanidades. Se presta también atención a cómo estas tecnologías de análisis automático de texto pueden ayudar a enriquecer ediciones digitales académicas. Para ello se describe igualmente el marco teórico que conecta las tecnologías citadas con la edición digital académica y se reflexiona sobre aplicaciones de las tecnologías e interfaces descritas a la edición digital académica.

PALABRAS CLAVE

Procesamiento del Lenguaje Natural (PLN), edición digital, análisis automático de texto.

RHD 4 (2019)

ISSN
2531-1786

Recibido 01/07/2019 – Aceptado 10/07/2019

1. INTRODUCCIÓN

La disponibilidad masiva de textos en formato digital sobre todo tipo de temas de Humanidades y Ciencias Sociales está cambiando la forma de hacer investigación en estas áreas (Venturini, Cardon y Cointet, 2014). Los estudios basados en un gran volumen de datos textuales requieren herramientas de análisis textual automático que permitan acceder a contenido relevante para las preguntas de investigación en cada caso. Consideramos que los productos resultantes de la aplicación de estas herramientas no constituyen ediciones digitales de los textos, entre otras razones porque no se analiza el texto en su conjunto, sino que se efectúa un acceso parcial del mismo mediado por herramientas de automatización. Sin embargo, estas herramientas pueden servir para enriquecer ediciones digitales académicas y asistir en su proceso de creación. Además, existen elementos comunes entre el proceso de aplicación de herramientas informáticas de análisis textual automático y el proceso de edición digital académica, por lo que ambos procesos pueden complementarse mutuamente.

La estructura del trabajo es la siguiente: tras una descripción del marco teórico que conecta las tecnologías citadas con la edición digital académica en el apartado 2, en el 3 se describen los corpus de nuestros estudios de caso. El que sigue, cubre las tecnologías de Procesamiento de Lenguajes Naturales (PLN) aplicadas, y el apartado 5 las interfaces de navegación de corpus que las explotan. Finalmente, el apartado 6 reflexiona sobre aplicaciones de las tecnologías e interfaces descritas a la edición digital académica[1].

2. MARCO TEÓRICO

El proceso editorial ha sufrido grandes cambios desde el punto de vista teórico y metodológico gracias a la mediación tecnológica. En un primer momento de transición, el principal objetivo de la computación era el de auxiliar en las tareas de edición menos intelectuales (Tanselle, 2006) así como favorecer la difusión de resultados. Esta limitación del papel de la tecnología en el proceso ecdótico está claramente superada, y en la actualidad se demanda un mayor debate teórico centrado en el diseño de nuevos modelos editoriales que aprovechen las numerosas ventajas del soporte digital (Sahle, 2013; Pierazzo, 2015).

Las tendencias en el campo de la edición digital académica están íntimamente relacionadas con las escuelas filológicas con mayor tradición en cada región, como el neolachmannianismo ibérico e italiano o la crítica genética francesa. Sin embargo, estos modelos

[1] Este trabajo fue realizado con el apoyo de una beca doctoral Région Île-de-France para Pablo Ruiz Fabo y el proyecto *Paleografía, Lingüística y Filología. Laboratorio on-line de la lírica gallego-portuguesa* (FFI2015-68451-P MINECO) para Helena Bermúdez Sabel.

tradicionales están siendo abandonados en favor de formatos que solo pueden ser concebibles en soporte digital, sin que el cambio conlleve una pérdida del rigor filológico.

No cabe duda de que la publicación en la web ofrece la posibilidad de incorporar múltiples recursos que enriquecen el texto con información de distinto cariz. El objetivo de incorporar información adicional no debería ser el de crear meros repositorios en los que se almacenan datos relacionados con un artefacto textual determinado: toda esta información debe ser curada, gestionada, analizada e interpretada, es decir, debe ser mediatizada por quien edita. Por tanto, el proceso de edición incluye la evaluación de cómo estos recursos van a permitir una exploración exhaustiva y eficiente del texto. Una vez que el equipo editorial ha decidido cuál es la información que será publicada, deben implementarse toda una serie de tareas relacionadas con el diseño de interfaz (Pierazzo, 2011).

En los próximos apartados se presentarán diversas herramientas y su implementación concreta a través de dos estudios de caso. El objetivo de este trabajo es mostrar cómo se puede crear una experiencia de lectura única gracias a la utilización de interfaces, y cómo estas pueden proporcionar una dimensión multifuncional a la propia presentación del texto.

3. CORPUS ANALIZADOS

Nuestro estudio de caso se centra en dos corpus de contenido político. La lengua de ambos corpus es el inglés. En estos corpus hay varios tipos de información relevantes para preguntas de investigación en ciencias sociales y políticas: los temas mencionados en el corpus, los actores que intervienen, así como qué actores abordan qué temas. Las tecnologías de PLN utilizadas, que serán descritas en el apartado 4, permiten identificar esos tipos de información automáticamente. A continuación, damos más detalles sobre los corpus y las transformaciones de formato que fueron necesarias para su tratamiento computacional.

3.1. Descripción

El primer corpus es PoliInformatics[2] (Smith, Cardie, Washington y Wilkerson, 2014), que alberga materiales heterogéneos sobre la crisis financiera norteamericana de 2007-2008. El corpus completo, del que hemos extraído una muestra, contiene legislación, informes del Congreso estadounidense sobre la crisis, transcripciones de reuniones sobre política monetaria en la Reserva Federal, y transcripciones de las sesiones de investigación del Congreso sobre el impacto de la legislación y políticas de respuesta a la crisis. El corpus también contiene una transcripción de la primera audiencia pública de la comisión FCIC (Federal Crisis Inquiry Commission) que, nombrada por el Congreso, se ocupó de investigar las causas de la crisis. Nuestra muestra contiene aproximadamente 400.000 palabras y dos tipos de documentos:

[2] Accesible desde: http://poliinformatics.org/.

primero, el informe Wall Street and the Financial Crisis: Anatomy of a Financial Collapse[3], un informe oficial del Congreso sobre las causas de la crisis, basado en procesos judiciales y entrevistas con expertos, con unas 318.000 palabras; en segundo lugar, la muestra incluye las transcripciones de la primera audiencia de la ya mencionada comisión FCIC, con unas 82.000 palabras que recogen 859 intervenciones en la audiencia.

El segundo corpus es una parte del volumen 16 del Earth Negotiations Bulletin[4] (ENB-Boletín de negociaciones de la Tierra). La parte del volumen 16 que hemos analizado consiste en resúmenes diarios de las intervenciones de los participantes en cumbres internacionales de política climática. Acuerdos como el Protocolo de Kioto de 1997 o el Acuerdo de París de 2015 se negocian en estas cumbres. Los participantes son países, grupos de países y organizaciones no gubernamentales. Nuestra muestra del corpus cubre negociaciones desde 1995 hasta 2016, con un total de aproximadamente 505.000 palabras.

3.2. Curación de los materiales fuente

Si bien los corpus estaban nativamente en formato digital, era necesario preprocesarlos para que su formato fuera apropiado para su tratamiento con herramientas de PLN y para hacerlos disponibles a través de interfaces de navegación de corpus.

En el corpus PoliInformatics, los documentos de nuestra muestra (ver apartado 3.1) habían sido convertidos a texto simple por los proveedores de los datos (Smith et al., 2014) a partir de originales en PDF. En esta versión inicial en texto plano la identificación de párrafos presentaba algunas inconsistencias, las cuales pudieron ser corregidas sobre la base de determinados rasgos formales. A través de estas operaciones de preproceso se crearon dos versiones del corpus: una versión en texto simple, que se utilizó para los análisis PLN; y una versión en XML respetando un formato compatible con el servidor de búsqueda Solr[5], en el que se indexaron los XML para permitir la navegación a través de búsqueda de texto libre.

El preproceso requerido para el corpus ENB suscita un mayor interés ya que, cuando nuestro trabajo se desarrolló, los documentos más antiguos del corpus (que se extiende de 1995 a 2016) estaban disponibles en versiones obsoletas de HTML que podían presentar problemas para su transformación a texto simple o XML. Asimismo, el corpus había usado diferentes plantillas a lo largo de sus más de 20 años de historia, de modo que los detalles del preproceso (los elementos HTML a considerar como títulos o cuerpo del documento) diferían según la fecha del texto. Aparte de estos factores, el preproceso fue similar al seguido para el corpus PoliInformatics: se eliminó el marcado HTML y se sistematizó la segmentación en párrafos. Se

[3] Wall Street y la crisis financiera: Anatomía de un colapso financiero (traducción propia). Accesible desde: https://bit.ly/2nd1mNL.
[4] Accesible desde: http://enb.iisd.org/enb/vol12/.
[5] Accesible desde: https://lucene.apache.org/solr/.

creó una versión en texto simple y una versión en XML compatible con Solr. El preproceso de este corpus presenta ciertos desafíos por lo que, para facilitar la investigación sobre el mismo, se publicó la versión preprocesada en texto simple y XML[6].

4. TECNOLOGÍAS DE PROCESAMIENTO DE LENGUAJE NATURAL APLICADAS

Hay varios tipos útiles de información básica para la investigación en Ciencias Sociales y Políticas sobre corpus como los ya descritos, como los temas mencionados en el corpus y los actores (como personas u organizaciones) que intervienen en él, o la relación entre actores y temas –es decir, qué actores abordan qué temas–. Se han aplicado varias tecnologías para identificar automáticamente esta información. Hemos considerado que los sintagmas que aparecen en el corpus cumpliendo ciertas características estadísticas y léxicas, que se detallarán más abajo, son una indicación de los temas contenidos en él.

Para encontrar estos sintagmas se han aplicado dos métodos: el enlazado de entidades (*entity linking*) (Rao, McNamee y Dredze, 2013) y la extracción de palabras clave (*keyphrase extraction*) (Kim, Medelyan, Kan y Baldwin, 2010). El enlazado de entidades encuentra en una base de conocimiento –como Wikipedia o su versión web semántica, DBpedia (Auer, Bizer, Kobilarov, Lehmann, Cyganiak e Ives, 2007)– conceptos a los que una expresión del texto se refiere. Esto es, se proyecta un esquema de conocimiento externo al texto para etiquetar expresiones contenidas en él. Al mismo tiempo, se da al texto un punto de entrada a los datos abiertos enlazados (Heath y Bizer, 2011), ya que este se etiqueta con conceptos disponibles en la web de datos abiertos.

En cuanto a la identificación automática de la relación entre actores y temas, este tipo de anotación es interesante –particularmente en el caso del corpus ENB–, ya que nos ocupamos de negociaciones políticas en las que unos actores se oponen a otros en diferentes aspectos. Identificar la posición de un actor con respecto a un ítem de la negociación puede ayudar a encontrar patrones de apoyo y oposición entre los actores. Este tipo de análisis va más allá de identificar expresiones que *coocurren* en el corpus, un método útil y ampliamente utilizado (Venturini y Guido, 2012; Poibeau y Ruiz Fabo, 2015), pero que no permite la identificación explícita de la naturaleza de la relación entre los términos en coocurrencia. La extracción de relaciones se ha aplicado a textos de ciencias políticas en estudios como Van Atteveldt (2008), Van Atteveldt, Sheafer, Shenhav y Fogel-Dror (2017), además de Diesner (2013) y referencias ahí citadas. Sin embargo, nuestro trabajo se concentra en predicados de relación tanto verbales como nominales, proporcionándose además una interfaz de usuario para explorar los resultados (ver apartado 5).

[6] Accesible desde: https://bitbucket.org/pruizf/enb/src.

Este apartado describe las tecnologías de PLN aplicadas a nuestros corpus para enriquecerlos con anotaciones lingüísticas (conceptos, palabras clave, relaciones) y facilitar una navegación de corpus que explore con precisión dichas anotaciones dando respuesta a cuestiones concretas de investigación. El apartado 4.1 describe el enlazado de entidades, el siguiente se concentra en la extracción de palabras clave, y el 4.3 discute la extracción de relaciones.

4.1. Enlazado de entidades

La primera tecnología aplicada al corpus es el enlazado de entidades. Se describen a continuación la tarea y lo que aporta al análisis del corpus, cómo se ejecuta la tarea así como los resultados que proporciona y cómo se evalúa.

4.1.1. Definición de la tarea y utilidad

El enlazado de entidades (*entity linking*) (Rao et al., 2013; Cornolti, Ferragina y Ciaramita, 2013) consiste en la identificación de expresiones (conocidas como menciones) dentro de un texto que tienen una referencia dentro de una base de conocimiento (BC) como DBpedia o similares[7], determinando también la referencia correcta. Hay dos fenómenos que dificultan esta tarea. Primero, hay varias maneras de referirse al mismo concepto de la BC (sinonimia parcial entre las menciones). Por ejemplo, la persona que ganó el premio Nobel de física de 1903 corresponde al concepto Marie_Curie en DBpedia[8]. Este concepto puede estar expresado en un texto a través de menciones como Marie Curie, Marie Skłodowska Curie o Mme Curie, entre otras. Se da también el fenómeno opuesto: una mención puede referirse a varios conceptos de la BC –podríamos decir que se trata de homonimia–. Por ejemplo, la mención Curie puede referirse al concepto Marie_Curie ya citado, pero también a Pierre_Curie[9] y a la unidad de radioactividad Curie[10], entre otros conceptos.

El enlazado de entidades tiene dos utilidades inmediatas: ayuda a encontrar documentos o pasajes que se refieren a los mismos temas –vistos como conceptos de una base de conocimiento–, haciendo abstracción de la variedad de expresiones con la que se hace referencia a esos temas; asimismo, ya que proporciona anotaciones de bases de conocimientos pertenecientes al ecosistema de datos abiertos enlazados (*Linked Open Data*), su aplicación constituye una primera puerta de entrada a publicar textos como datos abiertos enlazados.

[7] Por ejemplo YAGO (Suchanek, Kasneci y Weikum, 2007) o BabelNet (Navigli y Ponzetto, 2012).
[8] Accesible desde: http://dbpedia.org/page/Marie_Curie.
[9] Accesible desde: http://dbpedia.org/page/Pierre_Curie.
[10] Accesible desde: http://dbpedia.org/page/Curie.

4.1.2. Proceso de enlazado

Un procedimiento genérico de enlazado de entidades tiene lugar según tres etapas: (1) la detección de menciones, (2) la generación de conceptos candidatos y (3) la desambiguación de menciones.

> 1. Detección de menciones: Consiste en la identificación de expresiones que podrían referirse a conceptos de la BC. Puede basarse en un diccionario de posibles menciones: uno que contenga el texto de todos los links en los artículos de Wikipedia –ya que Wikipedia es el origen de DBpedia, por lo que los links de Wikipedia son posibles menciones a conceptos de DBpedia–. Por ejemplo, en Wikipedia, tanto los links Marie Curie como Maria Skłodowska enlazan con la página Marie_Curie, por lo que estas secuencias quedarían incluidas en un diccionario de menciones. La detección puede basarse también en la aplicación de tecnologías como el reconocimiento de entidades nombradas (Tjong Kim San y De Meulder, 2003), que identifican de forma productiva (sin restringirse a un diccionario previo) tipos de expresiones de distintas categorías, como personas, organizaciones, lugares y productos, entre otras.
>
> 2. Generación de candidatos: El objetivo de esta etapa es obtener un conjunto de conceptos de la BC que sean susceptibles de ser buenos candidatos para una mención dada. Para ello se pueden aplicar medidas de similitud entre la mención y las etiquetas de los conceptos candidatos (la etiqueta de un concepto es el nombre canónico asignado a este en la BC). Se suelen aplicar también otras operaciones de transformación de cadenas, como generación de acrónimos o iniciales. Por ejemplo, dada la mención M. Curie, tanto el concepto Marie_Curie como Pierre_Curie podrían ser considerados candidatos, en el sentido de que M. Curie contiene la inicial del nombre de pila Marie, pero también es la abreviatura de Monsieur, para Monsieur Curie.
>
> 3. Desambiguación de menciones: Esta etapa suele basarse en varias fuentes de información. Por un lado, se establece la probabilidad a priori de que una mención esté asociada a un concepto. Si la mención Curie aparece como link en Wikipedia para la página de la unidad de radioactividad en mayor proporción que para la página de Marie Curie, la probabilidad a priori del concepto candidato Curie será mayor que la del concepto candidato Marie_Curie. Otro factor importante para la desambiguación de una mención es su contexto, el cual será comparado con el contexto de esa mención en textos relacionados con cada candidato en la BC. Aquellos candidatos cuyo texto relacionado en la BC

presenten mayor coincidencia con el contexto de la mención en el documento serán prioritarios según este criterio. El grado de coincidencia entre los contextos puede medirse según el número de palabras en común, ignorando palabras vacías o *stopwords* (como artículos o preposiciones). Por ejemplo, si la mención Curie se refiere a una unidad de radioactividad, es razonable esperar que el léxico alrededor de la mención en un documento coincida más con el léxico de la definición de esa unidad en la BC que con el léxico de los textos relacionados con Marie o Pierre Curie. Finalmente, otro factor para determinar la adecuación de cada candidato para una mención es la coherencia de ese candidato con respecto a otros propuestos para menciones próximas. La coherencia se define de formas diferentes según la herramienta. La noción se basa en la aplicación de la teoría de grafos sobre el grafo de links de la base de conocimiento (Milne y Witten, 2008a)[11]. La intuición detrás de la coherencia es que, si en un texto hay varias menciones no ambiguas a unidades de medida, una mención como Curie, que es ambigua entre una unidad de medida y otros tipos de conceptos, será más coherentemente desambiguada en ese texto como la unidad de medida, dado que los mejores conceptos candidatos para las menciones que la rodean son también unidades de medida.

Las herramientas de enlazado de entidades que hemos aplicado en la interfaz descrita más abajo son: Wikipedia Miner (Milne y Witten, 2008b), TagMe2[12] (Ferragina y Scaiella, 2010; Cornolti et al., 2013) y DBpedia Spotlight[13] (Mendes, García-Silva y Bizer, 2011; Daiber, Jakob, Hokamp y Mendes, 2013). Las dos primeras usan tanto el contexto como la noción de coherencia para la desambiguación, mientras que el tercero usa sólo el contexto.

4.1.3. Resultados del proceso

El resultado del proceso de enlazado descrito es una lista de conceptos candidatos para cada mención, donde el orden de los candidatos lo determinan diferentes ponderaciones de los factores de desambiguación mencionados en el subapartado 4.1.2. Además de los candidatos en sí, el enlazado de entidades suele cuantificar los candidatos con una puntuación de confianza y una puntuación de coherencia. La confianza es una estimación por parte del algoritmo de desambiguación de la calidad de los resultados obtenidos. Por ejemplo, si había muchos candidatos en competición, o poca evidencia para estos, la confianza puede descender. La coherencia refleja en el criterio mencionado anteriormente, una especie de cohesión temática de

[11] Usando Wikipedia como BC.
[12] Véase demostración de la herramienta en: https://tagme.d4science.org/tagme/.
[13] Véase demostración de la herramienta en: http://demo.dbpedia-spotlight.org/.

un candidato con otros candidatos propuestos para menciones en el entorno de la mención de dicho candidato. Estas puntuaciones pueden usarse para filtrar resultados y quedarse sólo con los resultados susceptibles de tener mayor calidad. Unos indicadores de calidad bajos pueden también indicar a un usuario qué resultados deberían ser priorizados en una posible etapa de verificación manual. Los indicadores de calidad mencionados han sido explotados en la interfaz de navegación para el corpus PoliInformatics, que serán descrita posteriormente, para dar a los usuarios un modo de evaluar las anotaciones automáticas presentes en la interfaz.

4.1.4. La evaluación en enlazado de entidades

La evaluación del enlazado de entidades ha sido descrita en detalle por Cornolti et al. (2013). Esta implica comparar las anotaciones automáticas con anotaciones de referencia creadas manualmente por personas expertas. La comparación se puede hacer de forma más o menos estricta: además de la coincidencia del concepto para una mención, se puede exigir una coincidencia exacta entre las menciones detectadas automáticamente y las menciones de referencia, o se puede tolerar una coincidencia parcial en la mención.

Los resultados cuantitativos obtenidos para esta tecnología con ese tipo de evaluación sobre diversos corpus estándar pueden parecer modestos si nos fijamos puramente en las cifras obtenidas. Dependiendo del corpus de referencia y del método de evaluación, los resultados para las herramientas aplicadas en nuestro trabajo varían entre 44.5 puntos y 79.7 puntos de F1[14] (Ruiz Fabo y Poibeau, 2015). Sin embargo, las cifras obtenidas en una tarea cuantitativa estándar no siempre reflejan la utilidad de los resultados para especialistas de un dominio, en una tarea aplicada, como se ha argumentado en Ruiz Fabo (2017, pp. 68-70). Una forma de mejorar los resultados de enlazado de entidades es combinar las salidas de varios sistemas de forma que los sistemas se complementen, como se ha hecho en la interfaz descrita más abajo.

4.2. Extracción de palabras clave

La segunda tecnología aplicada a la anotación automática del corpus es la extracción de palabras clave. Se describe a continuación la tarea que esa tecnología aborda y aspectos de su evaluación.

4.2.1. Definición de la tarea y utilidad

La extracción de palabras clave (*keyphrase extraction*) (Kim et al., 2010) identifica las expresiones más importantes en un texto, según criterios léxicos –filtrándose las palabras vacías,

[14] La medida F1 es la media armónica de la precisión (P) y la exhaustividad (R). Las definiciones son: P = resultados correctos sistema / resultados sistema ; R = resultados correctos sistema / resultados de referencia ; F1 = 2PR / (P + R).

entre otros, –morfológicos– según la secuencia de categorías gramaticales de cada expresión– y estadísticos –se puede exigir una frecuencia mínima o ignorar expresiones que aparecen en demasiados documentos, al no ser estas discriminantes dentro del corpus, pues no sirven para aislar un subconjunto dentro de este–. Se suele usar para dar una visión general del contenido de un corpus, como en Moretti et al. (2014) o Rayson (2008); este es también el uso que hemos hecho de esta tecnología en nuestro trabajo. La extracción de palabras clave se ha limitado al corpus ENB.

Esta tarea complementa a los resultados del enlazado de entidades. Contrariamente al enlazado de entidades, esta extracción no requiere la disponibilidad, en una base de conocimiento (BC) externa al corpus, de conceptos relevantes como referencia de una expresión. Por tanto, se pueden extraer expresiones sobre aspectos del texto ausentes de la base de conocimientos, por ejemplo, porque sean demasiado especializados para una BC genérica como DBpedia.

4.2.2. Proceso de extracción

La anotación se llevó a cabo con Yatea (Aubin y Hamon, 2006)[15], un extractor de palabras clave basado en reglas. Toma como entrada un texto etiquetado con categorías gramaticales y, basándose en estas etiquetas, primero identifica frases nominales de acuerdo con una serie configurable de patrones de categorías gramaticales indicativas de estas. Los sintagmas nominales resultantes son después filtrados para eliminar candidatos que contienen secuencias no informativas –así, un candidato que contenga la secuencia *of course* será eliminado–. La herramienta fue configurada para extraer tanto expresiones de una sola palabra como de varias.

4.2.3. La evaluación en la extracción de palabras clave

La extracción de palabras clave es un procedimiento básico de minería de textos usado rutinariamente en recuperación de la información y motores de búsqueda (Rose, Engel, Cramer y Cowley, 2010). Una de sus aplicaciones industriales es la indexación y minería de literatura científica, y su utilidad para el análisis de textos nos parece indiscutible.

La evaluación cuantitativa de esta tecnología suele hacerse por comparación con palabras clave anotadas manualmente por expertos que son usadas como referencia. A pesar de la manifiesta utilidad de esta tecnología, las evaluaciones cuantitativas comparando con anotaciones de referencia arrojan cifras muy bajas de F1. En la tarea del workshop SemEval coordinada por Kim et al. (2010), los mejores sistemas obtuvieron una F1 de entre 25 y 30 puntos.

[15] Accesible desde: https://metacpan.org/pod/distribution/Lingua-YaTeA/lib/Lingua/YaTeA.pm.

El caso de la evaluación cuantitativa de la extracción de palabras clave sugiere, como ya se ha mencionado para el enlazado de entidades, que las métricas cuantitativas utilizadas en PLN para evaluar una tecnología no siempre son un buen correlato de la utilidad de esa tecnología en escenarios de aplicación concretos (Ruiz Fabo, 2017). Por eso es conveniente complementar esas métricas cuantitativas con evaluaciones cualitativas en una tarea de aplicación, como hemos hecho en la evaluación de las interfaces presentadas en este trabajo, particularmente la interfaz para el corpus ENB (ver subapartado 5.2.3).

4.3. Extracción de relaciones

Esta tecnología se aplicó solamente al corpus ENB. Describimos la tarea de extracción, sus etapas y evaluación. El código desarrollado está disponible públicamente[16].

4.3.1. Definición de la tarea

Particularmente en el caso del corpus ENB (ver sección 3.1), que contiene resúmenes de negociaciones diplomáticas sobre el cambio climático, es importante saber no solo qué temas se discuten en el corpus, sino también qué actores hablan de cada tema, y con qué actitud. Con el fin de anotar automáticamente esa información en el corpus ENB, hemos creado una *pipeline* que identifica los actores que emiten un enunciado en el corpus, así como el ítem de la negociación mencionado por el actor y el predicado (verbal o nominal) a través del cual el actor se expresa. Esta información se formaliza mediante proposiciones, es decir, triples con la forma (actor, predicado, ítem de negociación). Los actores son participantes en la negociación, como países y otras organizaciones. Los predicados son verbos o sustantivos de enunciación, como *state* (afirmar) o *statement* (declaración).

La figura 1 muestra un ejemplo de una oración del corpus y las proposiciones extraídas a partir de ella. Como se ve en la figura, varios actores (Unión Europea y Nueva Zelanda) apoyan el ítem considerado, mientras que China, Malasia y Bután están en contra. Esto nos permite inferir que, en relación con el ítem de negociación considerado, el primer grupo de actores se opone al segundo. Esto es, la extracción de proposiciones sirve también como primer paso para encontrar patrones de apoyo y oposición entre los actores de la negociación.

[16] Accesible desde: https://github.com/pruizf/pasrl.

1 - Múltiples predicados verbales

The EU, with NEW ZEALAND and opposed by CHINA, MALAYSIA and BHUTAN, supported including the promotion of natural regeneration within the definitions of "afforestation" and "reforestation."

	Proposiciones			Tipo de predicado
	Actor	*Predicado*	*Ítem de la negociación*	
1	European_Union	supported	including the promotion of natural regeneration within the definitions of "afforestation" and "reforestation."	support
2	New_Zealand			support
3	China	~supported	including the promotion of natural regeneration within the definitions of "afforestation" and "reforestation."	opposition
4	Malaysia			opposition
5	Bhutan			opposition

2 - Predicado nominal

Much of the discussion was on a proposal by the G-77/China to include research and development in the transport and energy sectors in the priority areas to be financed by the SCCF.

	Proposiciones			Tipo de predicado
	Actor	*Predicado*	*Ítem de la negociación*	
1	Group_of_77/China	proposal	to include research and development in the transport and energy sectors in the priority areas to be financed by the SCCF.	support

Figura 1. Se muestra el análisis de dos oraciones en las proposiciones que las componen. En la primera se extraen cinco proposiciones. En la segunda, con un predicado nominal, hay una sola proposición. Una tilde en el predicado de una proposición indica que sus actores han mostrado oposición al ítem de la negociación considerado.

4.3.2. Proceso de extracción de proposiciones

El proceso que hemos definido se basa en una cadena de tareas de PLN, así como en reglas de extracción que se aplican al resultado de estas tareas. La librería de PLN que hemos utilizado es IXA Pipes (Agerri, Rigau y Bermúdez, 2014). El proceso utiliza también una base de posibles actores y posibles predicados de enunciación a considerar; la lista completa se encuentra en Ruiz Fabo (2017)[17]. Si bien la lista de predicados es imprescindible, la cadena puede funcionar también sin una lista predefinida de actores: se considerarán en ese caso como posibles actores todos los sujetos de predicados de enunciación. Para cada oración que contiene predicados de enunciación según la lista mencionada, las informaciones proporcionadas por la cadena de PLN y que son explotadas para la extracción de proposiciones son las siguientes[18]:

- Etiquetado de roles semánticos: Esta tecnología identifica, dado un predicado nominal o verbal, qué argumentos intervienen en la acción o estado denotado por el predicado. Determina roles como agente, tema, u otros que expresan circunstantes o nociones adverbiales (por ejemplo, una expresión de tiempo). El esquema de roles utilizado es el de la base léxica PropBank (Palmer, Gildea y Kingsbury, 2005). Los roles utilizados como base de las proposiciones identificadas fueron agente y tema, si bien también se usaron otros roles para casos menos frecuentes.

[17] También disponible en: https://sites.google.com/site/nlp4climate/domain-model.
[18] Para más detalle véase Ruiz Fabo (2017).

- Análisis de dependencias: Esta tecnología identifica funciones sintácticas como sujeto u objeto directo. Se ha usado para complementar las salidas del etiquetado de roles semánticos (en caso de no identificación por este de roles relevantes) y la de la resolución de anáforas pronominales.
- Resolución de anáforas pronominales: Esta tarea consiste en encontrar el sintagma al que un pronombre personal se refiere (el antecedente). La tarea es compleja y en este estudio nos hemos concentrado en un pequeño subconjunto de los casos posibles a resolver. La librería IXA Pipes lleva a cabo la resolución de correferencias, y sobre las salidas de esta hemos resuelto los casos más básicos de anáfora pronominal gracias a la aplicación de reglas.
- Resolución de la negación: Esta tarea es compleja y tratamos solamente los casos más claros. Los predicados negados fueron identificados según la presencia de roles PropBank que indican negación (AM-NEG), y según la presencia de indicadores de polaridad negativa (*not*, *no*) en una ventana alrededor del predicado.

Las triples identificadas reciben una puntuación de confianza de 0 a 5, según el valor informativo que cabe esperar de la proposición y según una estimación de hasta qué punto es posible que la proposición haya sido correctamente identificada (Ruiz Fabo, 2017, p. 84). Por ejemplo, si la resolución de anáforas pronominales se ha aplicado, la confianza disminuye, ya que esa etapa puede haber introducido un error. La puntuación de confianza puede ser aplicada por los usuarios en la interfaz para analizar un conjunto de proposiciones más amplio (potencialmente con más errores) o más restringido pero más fiable.

4.3.3. Evaluación

La extracción de proposiciones fue evaluada cuantitativamente con un conjunto de 313 proposiciones identificadas manualmente en 100 oraciones del corpus, que se compararon con las extracciones automáticas. Se buscaron oraciones que contuvieran los desafíos típicos del corpus: presencia en la misma oración de múltiples actores y predicados, y negación[19]. El criterio para considerar que una extracción fuera correcta fue la coincidencia exacta para los tres elementos de la proposición (actor, predicado e ítem de la negociación). Usando este criterio, la puntuación F1 fue de 69 puntos. Es un criterio de evaluación estricto y el resultado cuantitativo nos parece aceptable. Como se ha mencionado, es conveniente complementar la evaluación cuantitativa con evaluaciones cualitativas realizadas por especialistas del dominio que se está

[19] El juego de test se encuentra accesible desde: https://sites.google.com/site/thesisrf/proposition-extraction-test-set.

modelizando (en este caso la política sobre el cambio climático). Se llevó a cabo una evaluación con expertos de la interfaz de navegación de corpus que explota la extracción de proposiciones (ver apartado 5.2). Estos consideraron que la calidad de la extracción de proposiciones era suficiente para sus actividades de exploración del corpus.

4.3.4. Enriquecimiento de las proposiciones con entidades y palabras clave

Para aprovechar estos metadatos en la interfaz, se efectuó un enlazado de entidades con DBpedia, además de una extracción de palabras clave, limitando las salidas disponibles en la interfaz a aquellos metadatos que provenían de menciones contenidas en ítems de la negociación: así nos aseguramos de que las palabras clave o entidades eran efectivamente parte de temas discutidos en la negociación, y no parte de otros constituyentes de las oraciones. Además de estas anotaciones, también se anotaron menciones a conceptos del tesauro *Climate Thesaurus* (Bauer, Recheis y Kaltenböck, 2011) que cubre términos de política climática. Usamos la API Climate Tagger[20] para identificar estas menciones. El uso de un tesauro nos da una referencia a conocimiento externo, como DBpedia, pero supone una base de conocimiento específica al dominio y que por tanto puede contener conceptos que describen de forma más precisa el contenido del corpus.

5. INTERFACES DE NAVEGACIÓN DE CORPUS

En general, los objetivos de las interfaces desarrolladas son: ayudar a los investigadores a encontrar información relevante sobre cuestiones de investigación específicas de su dominio; dar información que no sería fácilmente accesible sin usar estas interfaces; y ofrecer anotaciones lingüísticas obtenidas por PLN a las que las interfaces dan acceso. Estas informaciones se basan en constructos lingüísticos no disponibles a través de una búsqueda por palabras en el texto integral.

Esta sección describe las interfaces desarrolladas para examinar los corpus PoliInformatics y ENB, sus funciones, los tipos de navegación que permiten, y la evaluación llevada a cabo, que en el caso del corpus ENB tuvo lugar con especialistas de dominio[21].

5.1. PoliInformatics

El corpus PoliInformatics (Smith et al., 2014) contiene materiales heterogéneos (informes, entrevistas) sobre la crisis financiera americana de 2007-2008; nuestra muestra fue descrita en el apartado 3.2. La interfaz[22] combina el acceso al texto mediante búsqueda por facetas (con

[20] Accesible desde: http://api.climatetagger.net.
[21] Las interfaces están accesibles desde: http://apps.lattice.cnrs.fr/prf/.
[22] Accesible desde: http://apps.lattice.cnrs.fr/nav/gui/.

entidades enlazadas como facetas) y la búsqueda de texto libre. Describimos a continuación los objetivos de esta interfaz, sus funciones y cómo se ha evaluado.

5.1.1. Objetivos

El objetivo general es ayudar a los investigadores a seleccionar un conjunto de entidades con las que modelar el corpus, a partir de los resultados del enlazado de entidades descrito en el apartado 4.1. La interfaz responde a las necesidades identificadas por investigadores en Ciencias Políticas con quienes habíamos colaborado, y además palia problemas inherentes a la aplicación del enlazado de entidades. Venturini y Guido (2012, pp. 7-15) mencionan que sería conveniente contar con métricas de la calidad de los resultados para guiar el filtrado manual de entidades por parte de los investigadores. El filtrado manual complementa a un posible filtrado automático, tanto para eliminar entidades incorrectas como para recuperar entidades eliminadas automáticamente que son correctas. Teniendo en cuenta esto, la interfaz muestra indicadores de la calidad de los resultados.

Por otro lado, la calidad de los resultados de enlazado de entidades devueltos por cada herramienta varía en función del corpus. Para conseguir una calidad más homogénea independientemente del corpus, hemos combinado los resultados de varias herramientas. La interfaz muestra los resultados de cada una, además de dar la opción de llevar a cabo una selección automática entre estos resultados.

5.1.2. Funcionalidades

Comenzamos con una descripción general de la interfaz, para discutir después cómo la interfaz aborda las necesidades descritas más arriba. Otros aspectos de la interfaz se describen en Ruiz Fabo et al. (2015) y Ruiz Fabo (2017, p.144).

5.1.2.1. Funciones de búsqueda

La interfaz se divide en dos paneles (figura 2). El panel de entidades (izquierda) contiene dos áreas de búsqueda: *Search Text* (buscar texto [1] en la figura 3) y *Search Entities* (buscar entidades [5] en la figura 3). A la derecha encontramos el panel de documentos.

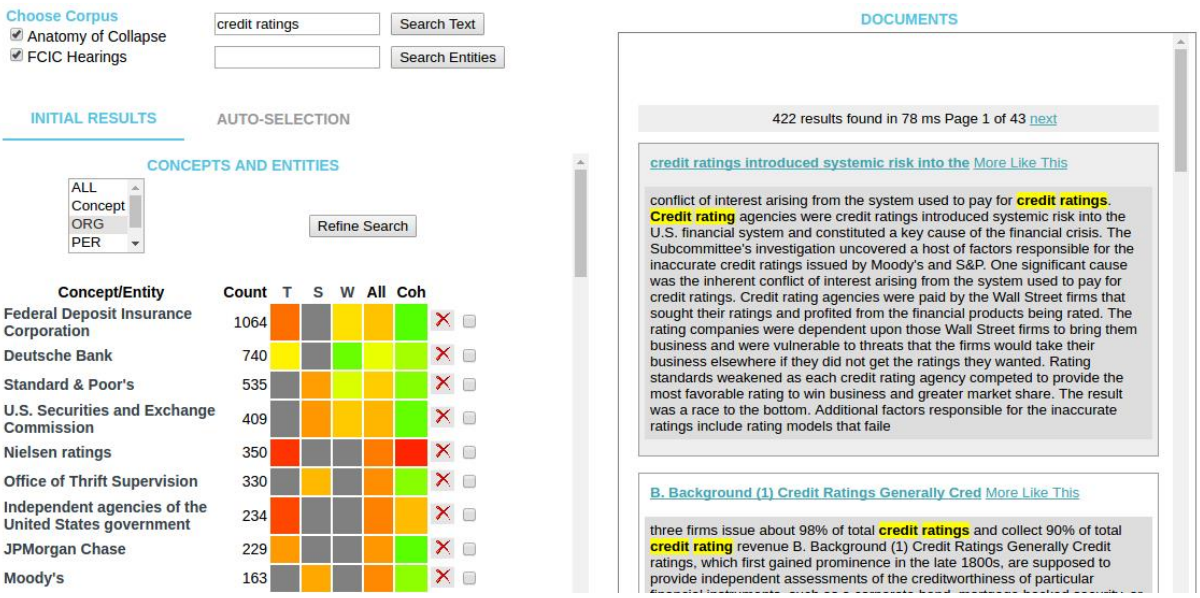


Figura 2. La interfaz PoliInformatics, con el panel de entidades a la izquierda y el panel de documentos a la derecha. Se muestran los resultados de la consulta de texto libre *credit ratings* (evaluación de crédito).

Introduciendo una expresión en *Search Entities* (buscar entidades), el panel de entidades mostrará las entidades del corpus que corresponden a la consulta, y el panel derecho mostrará los documentos en los que esas entidades han sido anotadas. Introduciendo una expresión en *Search Text* (buscar texto), el panel de documentos mostrará los documentos que corresponden a la consulta, y el panel de entidades mostrará las entidades que han sido anotadas en esos documentos.

Los resultados de las búsquedas se pueden filtrar por tipo de entidad: los tipos a mostrar se eligen en el menú desplegable de la izquierda que contiene los tipos ORG (Organización) y PER (Persona) entre otros tipos, y el botón *Refine Search* (filtrar resultados [6] en la figura 3) activa el filtrado. Los resultados se pueden restringir también a un conjunto de entidades activando las casillas en la fila correspondiente a cada entidad. Cuando los resultados son filtrados, en el panel de documentos se muestran únicamente los documentos correspondientes a las entidades filtradas.

La búsqueda de texto libre se lleva a cabo a través de un servidor Solr[3] donde se indexó el corpus. La búsqueda por entidades tiene lugar en una base MySQL en la que se indexaron los metadatos. El uso de identificadores compartidos entre el índice Solr y la base de datos posibilita la unión de los resultados de ambas fuentes.

5.1.2.2. Indicadores de la calidad de los resultados

En el panel de entidades hay varias columnas coloreadas para cada entidad. T, S y W ([2] en la figura 3, *vid. infra*) reflejan los resultados de distintas herramientas de enlazado de entidades: T contiene los resultados de TagMe2, S muestra los resultados de DBpedia Spotlight, y W los de Wikipedia Miner. Los colores de T, S y W indican la puntuación de confianza para

la anotación (apartado 4.1.3). Se trata de una escala de 0.0 (rojo) a 1.0 (verde), con amarillo como punto medio. El color da una indicación visual de la calidad del resultado, y la puntuación numérica se puede leer pasando el ratón sobre la celda de la tabla. Una celda gris para un anotador quiere decir que este no ha encontrado dicha entidad –lo cual también indica que podemos confiar menos en ella, ya que no todos los anotadores la han encontrado–. La columna *All* ([3] en la figura 3) es una media de las puntuaciones de confianza para los tres anotadores.

La columna *Coh* muestra la puntuación de coherencia entre cada entidad y un conjunto representativo de entidades del corpus. La noción de coherencia, que se explicó anteriormente, y los detalles sobre su computación se encuentran en Ruiz Fabo (2017, pp. 142-145). En términos informales, una puntuación de coherencia baja indica que la entidad no es consistente con los temas del corpus, y por tanto cabe esperar que la anotación sea incorrecta. Por ejemplo, la anotación marcada con [4] en la figura 3 tiene una coherencia baja con el corpus (la entidad *Nielsen_ratings* es una agencia de evaluación de audiencia, no de evaluación de crédito, que sería más coherente con los temas del corpus).

Además de las indicaciones visuales de la calidad de las anotaciones, ya mencionadas, su contexto, disponible en los documentos del panel de la derecha, puede usarse para decidir si una anotación es correcta o no. Aparte de la selección manual de entidades según las indicaciones de calidad, la interfaz puede usarse para seleccionar automáticamente un subconjunto de entidades de entre las que presentan mejor calidad y sobre las que los anotadores tienen mayor acuerdo ([7] en la figura 3). El método de selección se describe en Ruiz Fabo y Poibeau (2015).

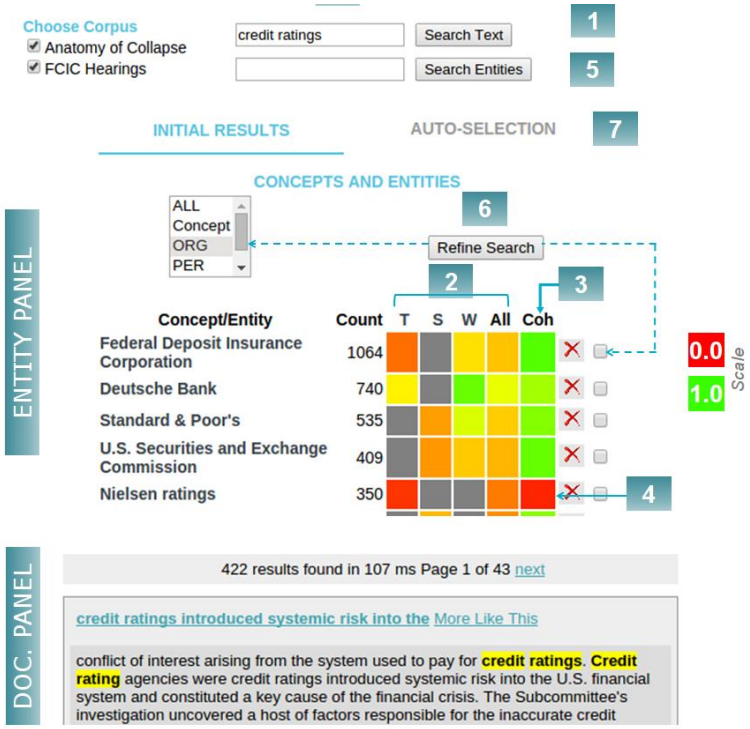
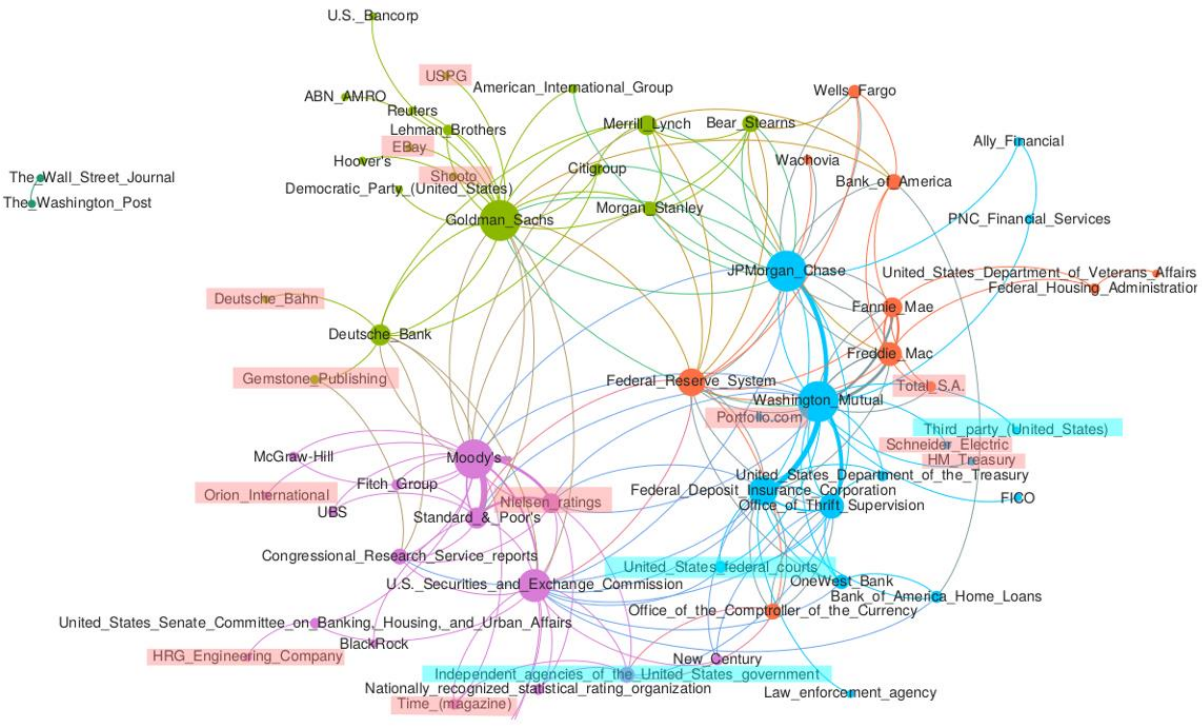


Figura 3. Funciones de búsqueda e indicadores de la calidad de los resultados en la interfaz PoliInformatics.

5.1.3. Evaluación

La interfaz muestra métricas que indican la calidad de las anotaciones. La validez de estas métricas se evaluó en los trabajos que las describen en detalle (ver apartado 4.1.3), así como los resultados de las herramientas de anotación de entidades y su procedimiento de combinación. A diferencia de la interfaz ENB (apartado 5.2), que fue evaluada por expertos de dominio, hemos evaluado la interfaz PoliInformatics nosotros mismos; una mejora de la evaluación sería repetir la tarea con expertos.

Como evaluación cualitativa a través de una tarea de aplicación, llevamos a cabo la siguiente experiencia. En respuesta a la pregunta ¿qué organizaciones son relevantes en la crisis financiera?, creamos redes de coocurrencia entre las organizaciones identificadas por el enlazado de entidades en el corpus. Las redes representan las entidades que aparecen en la misma oración; las entidades que coocurren con mayor frecuencia están más próximas en la red. Después, la interfaz se utilizó para consultar las indicaciones de calidad de las entidades de la red. Esto reveló varias entidades cuyas medidas de coherencia o confianza eran bajas (indicadas en rojo o azul en la parte superior de la figura 4). Para las entidades indicadas en azul, fue necesario verificar el contexto en los documentos, ya que eran temáticamente coherentes con el contenido del corpus. Si bien esta evaluación fue una tarea de validación informal, sugiere la utilidad de la información proporcionada por la interfaz para recopilar evidencia relevante para una pregunta de investigación.



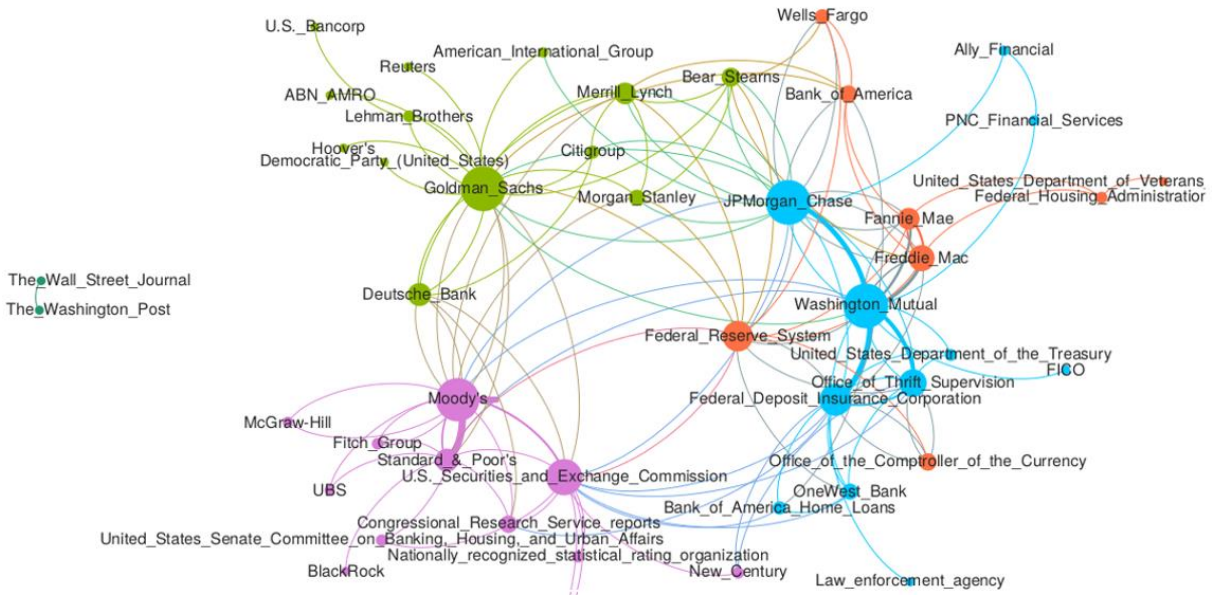


Figura 4. Red de co*ocurrencias* entre organizaciones de corpus PoliInformatics. La red de arriba contiene entidades erróneas cuyos indicadores de calidad eran bajos en la interfaz. Fueron después eliminadas creando una red con más consistencia entre los miembros de cada grupo.

5.2. Earth Negotiations Bulletin

El corpus contiene resúmenes de las intervenciones de los participantes en cumbres internacionales de política climática. La interfaz[23] permite navegar estas intervenciones a través de diversas búsquedas estructuradas, restringidas a componentes lingüísticos como los emisores de cada intervención y los ítems de la negociación sobre los que hablan. Una búsqueda de texto libre también es posible. Metadatos extraídos de los mensajes (palabras clave y otros) pueden usarse para filtrar los resultados.

5.2.1. Objetivos

El objetivo de esta interfaz es permitir la comparación del comportamiento de diferentes actores en la negociación. De manera más genérica, se pretendía evaluar hasta qué punto la interfaz permite obtener información sobre el corpus que no habría estado disponible para la comunidad científica sin su uso, o si la interfaz puede servir para arrojar nuevos datos sobre el corpus.

5.2.2. Funcionalidades

Se describió más arriba (apartado 4.3) la *pipeline* para la extracción de proposiciones o triples (actor, predicado, ítem de la negociación) en las que se basan las funciones de búsqueda de la interfaz. La interfaz permite una búsqueda restringida a proposiciones que contengan ciertos actores, ciertos predicados o tipos de predicados (apoyo, oposición,

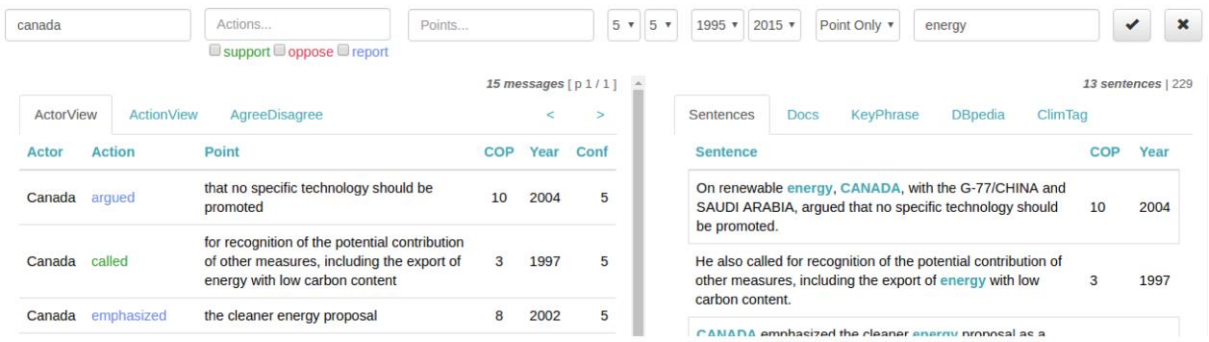
[23] Para instancia véase: http://apps.lattice.cnrs.fr/ie/uidev/, y para código véase: https://github.com/pruizf/pasrl.

enunciación neutra) o cuyos ítems de negociación contengan la expresión buscada. Aparte de esto, también se ha habilitado una búsqueda de texto libre. Además, se pueden limitar las búsquedas a un rango de fechas y a un rango de puntuaciones de confianza para las proposiciones.

Para cualquier búsqueda, es posible leer las oraciones que contienen las proposiciones devueltas o sus documentos (con la oración resaltada), o bien leer compilaciones de metadatos que proporcionan una visión de conjunto del contenido de los ítems de negociación presentes en los resultados. Estos metadatos son palabras clave, entidades de DBpedia, o conceptos del Climate Thesaurus (Bauer et al., 2011), especializado en política climática. Las proposiciones se muestran en el panel de la izquierda, y sus contextos (oración, documento) o metadatos en el panel de la derecha (figura 5, *vid. infra*).

Además, con la pestaña *AgreeDisagree*, se puede elegir un par de actores y mostrar las oraciones –y metadatos extraídos de sus proposiciones– en las que esos actores han mostrado un acuerdo o desacuerdo, pudiendo filtrarse los resultados según los metadatos mencionados en las proposiciones (palabras clave, conceptos de Climate Thesaurus, o entidades de DBpedia).

Por ejemplo, la figura 5 muestra las proposiciones en que Canadá y un grupo de estados insulares llamado AOSIS hablan de energía. Las palabras clave extraídas de los ítems de negociación abordados nos muestran que la exportación de recursos energéticos (*energy exports*) es un tema abordado por Canadá –quizás porque pueden exportar energía– pero no por AOSIS, que se expresa en cambio sobre la eficiencia energética (*energy efficiency*). Ambos actores hablan de energía limpia o renovable, con expresiones como *cleaner energy* o *renewable energy*.



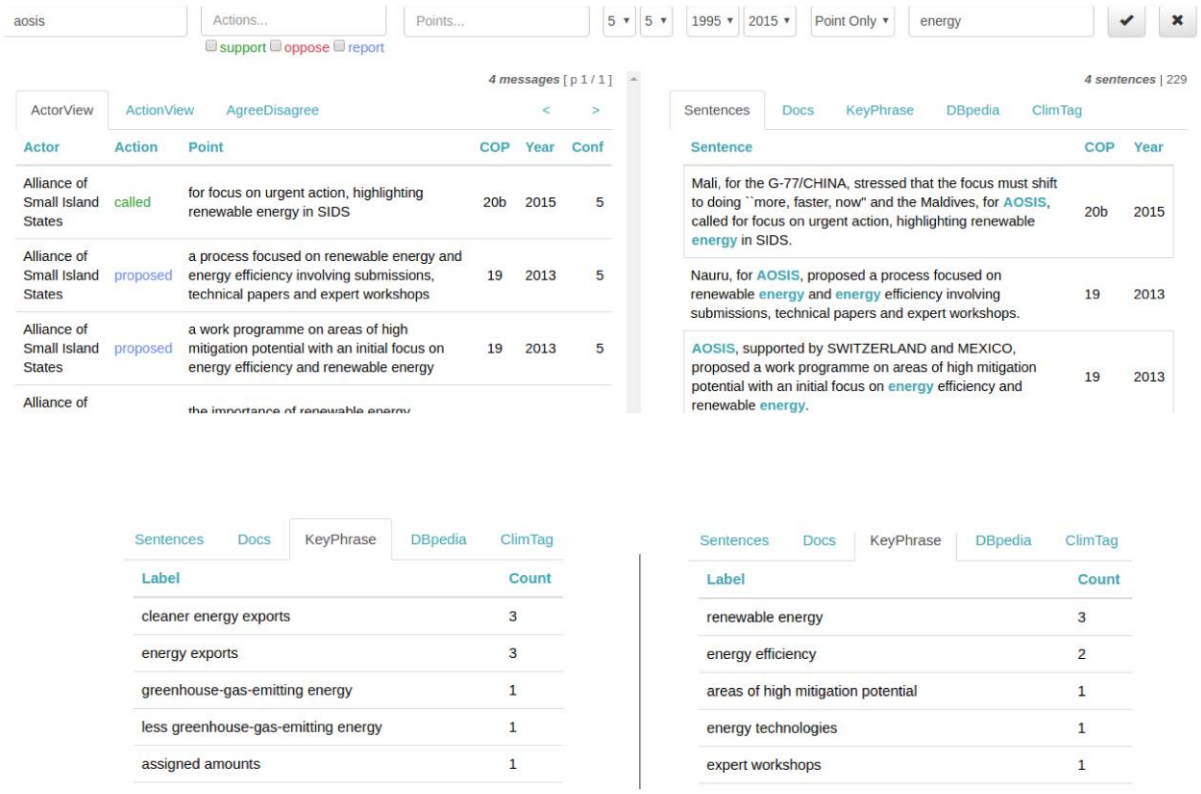


Figura 5. Consultas para proposiciones con Canadá (arriba) o AOSIS (centro) como actores y *energy* energía en el ítem de la negociación. Las imágenes inferiores muestran las palabras clave extraídas de los resultados para Canadá (izquierda) y AOSIS (derecha).

5.2.3. *Evaluación*

La evaluación tuvo lugar con tres expertos del dominio que habían producido publicaciones científicas sobre el corpus, y usaron la interfaz para efectuar búsquedas relacionadas con sus preguntas de investigación. La actividad tuvo una duración de una hora por experto; se registraron las búsquedas realizadas por los expertos, así como los comentarios que hicieron sobre los resultados proporcionados por la interfaz.

Los resultados de la evaluación sugieren que este tipo de interfaz y la información que ofrece a los investigadores puede llevar a descubrir nuevas vías de investigación o datos no previamente conocidos sobre el corpus. Concretamente, gracias a la disponibilidad sistemática de triples (actor, predicado, ítem de negociación), una de las expertas observó que ciertos actores se comportan de forma legalista dentro de la negociación, insistiendo en aspectos formales y legales, mientras que otros actores insisten en verdaderas acciones de gestión del cambio climático. Esta podría ser una dimensión para comparar los actores en la que la investigadora no había pensado previamente. Dos de los expertos mencionaron que, ya que la interfaz no exige una lista de actores previa y muestra resultados para actores que quedan fuera de la lista de países participantes en la negociación habitualmente estudiada, se ofrecen resultados sobre actores poco estudiados, como ciertas organizaciones no gubernamentales, grupos de pueblos indígenas o grupos de interés en el género y cambio climático. En este

sentido, consideramos que los objetivos de la interfaz se ven cumplidos, según los comentarios obtenidos en la evaluación con expertos del dominio.

6. VALORACIÓN FINAL: APLICACIONES EN EDICIÓN DIGITAL ACADÉMICA

En palabras del célebre filólogo italiano Gianfranco Contini, es una *osservazione elementare* que la constitución textual y la información lingüística se condicionan recíprocamente (Contini, 1986, p. 149). Aunque el ensayo de Contini estaba especialmente centrado en la problemática de las variantes lingüísticas formales durante la fijación del texto crítico, es evidente que el estudio de la lengua no solo es imprescindible durante el proceso de presentación del texto editado, sino que también es determinante en la vertiente más interpretativa del proceso ecdótico. Uno de los objetivos de la labor filológica es el de ofrecer al público un producto textual proporcionando todos los elementos necesarios para que el significado original del texto, así como su evolución a lo largo de la historia, sean accesibles al auditorio deseado. En este sentido, interfaces como las aquí presentadas ayudan al editor en esta tarea.

Una de las ventajas del soporte digital es la oportunidad de introducir información complementaria al texto de una manera fácilmente accesible a quien realiza la lectura sin que esta información incomode o interrumpa el proceso de lectura. En este sentido, el enlazado de entidades ofrece la posibilidad de enriquecer la edición con contenido contextual gracias al cual se identifican personajes, lugares y otro tipo de referencias –tanto aquellas propias de la ficción como las entidades con un referente real–. Una edición crítica en papel, por ejemplo, con su aparato y sus notas interpretativas, también contiene este tipo de información, exigiendo una ruptura de la lectura secuencial. Las anotaciones que acompañan una edición en papel son incorporadas porque se considera que son referencias que el lector debe conocer para poder comprender el significado del texto original. En el soporte digital, gracias a la implementación de interfaces bien diseñadas, se habilita una lectura radial realmente eficiente (Tanselle, 2006, p. 5), es decir, se puede forzar al usuario a colocar el texto en diferentes contextos gracias a la exploración externa, siendo tarea del equipo editorial la de proporcionar las herramientas necesarias para tal exploración. Las interfaces son la herramienta que permite que el lector pueda llevar a cabo la experiencia lectora que más se adecúe a sus intereses y necesidades.

Los resultados de procesos de extracción de palabras clave y proposiciones pueden ser explorados durante el análisis literario-cultural que acostumbra a acompañar una edición. Las posibilidades de implementación son muy numerosas, pero una aplicación inmediata desde una perspectiva comparatista es la de contextualizar una obra dentro de la producción de ese autor o autora o, incluso, abrir el marco de comparación e introducir la producción literaria de sus coetáneos. Este tipo de estudios permite responder preguntas sobre las temáticas y posicionamientos privilegiados dentro y fuera del canon literario. En la actualidad, la comunidad

académica tiene acceso a multitud de textos en formato digital que son propensos a ser utilizados para este tipo de análisis que demanda el uso de grandes corpus. En ocasiones, la calidad científica de los textos no siempre es la deseada: de ahí la necesidad de ediciones filológicamente rigurosas, especialmente para estudios que requieran una lectura minuciosa, detallada, gracias a la cual poder describir los pormenores del texto. Sin embargo, para el desarrollo de macroanálisis (Jockers, 2013), el análisis basado en métodos computacionales y cuyo objeto de estudio acostumbran a ser colecciones de documentos, los errores que pueda haber en las ediciones disponibles se diluyen; estos estudios se concentran en regularidades que aparecen de forma robusta en grandes colecciones, de forma que los casos excepcionales o errores esporádicos tienen un impacto menor en los resultados.

Como ya ha sido apuntado, una edición añade información que supera la del contenido del texto propiamente dicho. Es parte del trabajo editorial decidir cómo es presentado el texto, establecer los mecanismos a través de los cuales se pueden interrogar los materiales, y habilitar diferentes puntos de acceso a la información analítica e interpretativa, jerarquizando todos esos datos para evitar la sobrecarga informativa. Esto significa que el proceso de edición en un entorno digital demanda conocimiento que va más allá del ámbito lingüístico y literario. Consecuentemente, existen ciertas reticencias dentro de la comunidad filológica sobre la inclusión de todos estos elementos adicionales como parte de la labor ecdótica. En soportes analógicos, un editor era capaz de realizar todas las tareas relacionadas con la edición del texto objeto de estudio. Actualmente se le demanda que en muchas ocasiones dependa de personal con formación tecnológica para el desarrollo y publicación de su trabajo filológico. Con esta contribución queremos mostrar algunas herramientas que podrían ser adaptadas para su uso en entornos de edición intuitivos que fomenten la aplicación de métodos computacionales en el proceso de edición académica, favoreciendo que el editor mantenga su autonomía sin renunciar a las ventajas proporcionadas por los avances tecnológicos.

REFERENCIAS BIBLIOGRÁFICAS

Agerri, R., Rigau, G. y Bermúdez, J. (2014). IXA Pipeline: Efficient and Ready to Use Multilingual NLP Tools. En N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk y S. Piperidis (Eds.), *LREC 2014, Ninth International Conference on Language Resources and Evaluation*. Reykjavik: European Languages Resources Association. Recuperado de https://bit.ly/2Z2EX2I el 13/08/2019.

Aubin, S. y Hamon, T. (2006). Improving Term Extraction with Terminological Resources. En T. Salakoski, F. Ginter, S. Pyysalo y T. Pahikkala (Eds.), *Advances in Natural Language Processing. FinTAL 2006* (pp. 380-387). Berlin: Springer. Recuperado de https://hal.archives-ouvertes.fr/hal-00091444 el 13/08/2019.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak y Ives, Z. (2007). DBpedia: A Nucleus for a Web of Open Data. En K. Aberer, K. Choi, N. Noy, D. Allemang, K. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber y P. Cudré-Mauroux (Eds.), *The Semantic Web. ISWC 2007, ASWC 2007* (pp. 722-735). Berlin: Springer. Recuperado de https://bit.ly/2YMOI9x el 13/08/2019.

Bauer, F., D. Recheis y M. Kaltenböck (2011). data.reegle.info – A New Key Portal for Open Energy Data. En J. Hřebíček, G. Schimak y R. Denzer (Eds.), *Environmental Software Systems. Frameworks of eEnvironment. ISESS 2011* (pp. 189-194). Berlin: Springer.

Contini, G. (1986). *Breviario di ecdotica*. Milán: R. Ricciardi.

Cornolti, M., Ferragina, P. y Ciaramita, M. (2013). A Framework for Benchmarking Entity-Annotation Systems. En D. Schwabe, V. Almeida y H. Glaser (Coords.), *Proceedings of the 22nd International Conference on World Wide Web* (pp. 249-260). New York: ACM. Recuperado de https://bit.ly/2C3I4Pu el 18/08/2019.

Daiber, J., Jakob, M., Hokamp, C. y Mendes, P. N. (2013). Improving Efficiency and Accuracy in Multilingual Entity Extraction. En M. Sabou, E. Blomqvist, T. Di Noia, H. Sack y T. Pellegrini, *Proceedings of the 9th International Conference on Semantic Systems* (pp. 121-124). New York: ACM. Recuperado de https://bit.ly/2MqeUjV el 18/08/2019.

Diesner, J. (2013). *From Texts to Networks: Detecting and Managing the Impact of Methodological Choices for Extracting Network Data from Text Data* (Tesis doctoral). Universidad Carnegie Mellon, Pittsburgh. doi:10.1007/s13218-012-0225-0

Ferragina, P., y Scaiella, U. (2010). Tagme: On-the-fly Annotation of Short Text Fragments (by Wikipedia Entities). *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 1625-1628. doi:10.1145/1871437.1871689

Heath, T. y Bizer, C. (2011). Linked Data: Evolving the Web into a Global Data Space. *Synthesis Lectures on the Semantic Web: Theory and Technology, 1*(1), 1-136. Recuperado de http://linkeddatabook.com/editions/1.0/ el 18/08/2019.

Jockers, M. L. (2013). *Macroanalysis: Digital Methods & Literary History*. Illinois: University of Illinois Press.

Kim, S. N., Medelyan, O., Kan, M. Y. y Baldwin, T. (2010). Semeval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles. En K. Erk y C. Strapparava (Eds.), *Proceedings of the 5th International Workshop on Semantic Evaluation* (pp. 21-26). Stroudsburg: ACL. Recuperado de http://dl.acm.org/citation.cfm?id=1859668 el 18/08/2019.

Mendes, P. N., Jakob, M., García-Silva, A., y Bizer, C. (2011). DBpedia Spotlight: Shedding Light on the Web of Documents. En C. Ghidini, A. C. Ngonga Ngomo, S. Lindstaedt y T. Pellegrini (Eds.), *Proceedings of the 7th International Conference on Semantic Systems* (pp. 1-8). Recuperado de http://dl.acm.org/citation.cfm?id=2063519 el 18/08/2019.

Milne, D. y Witten, I. H. (2008a). Learning to Link with Wikipedia. En J. G. Shanahan (Coord.), *Proceedings of the 17th ACM Conference on Information and Knowledge Management* (pp. 509-518). Recuperado de http://dl.acm.org/citation.cfm?id=1458150 el 18/08/2019.

_____ (2008b). An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. En D. M. Hamilton (Dir.), *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy* (pp. 25-30). Recuperado de https://bit.ly/2P5FZei el 18/08/2019.

Moretti, G., Tonelli, S., Menini, S., Sprugnoli, R., Basili, R., Lenci, A. y Magnini, B. (2014). ALCIDE: An online platform for the Analysis of Language and Content in a Digital Environment. En R. Basili, A. Lenci y B. Magnini (Eds.), *Proceedings of the First Italian Conference on Computational Linguistics* (pp. 270-274). Pisa: Università di Pisa. Recuperado de https://bit.ly/2L0Rs9J el 18/08/2019.

Navigli, R. y Ponzetto, S. P. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence, 193*, 217-250. doi:10.1016/j.artint.2012.07.001.

Palmer, M., Gildea, D. y Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational linguistics*, *31*(1), 71-106. doi:10.1162/0891201053630264.

Pierazzo, E. (2011). A Rationale of Digital Documentary Editions. *Literary and Linguistic Computing, 26*(4), 463-477. doi:10.1093/llc/fqr033.

_____ (2015). *Digital Scholarly Editing: Theories, Models and Methods*. Recuperado de http://hal.univ-grenoble-alpes.fr/hal-01182162 el 20/08/2019.

Poibeau, T. y Ruiz Fabo, P. (2015). Generating Navigable Semantic Maps from Social Science Corpora. *Digital Humanities Conference 2015*. Sydney: ADHO. Recuperado de https://bit.ly/2za1e3A el 18/08/2019.

Rao, D., McNamee, P. y Dredze, M. (2013). Entity Linking: Finding Extracted Entities in a Knowledge Base. En T. Poibeau, H. Saggion, J. Piskorski y R. Yangarber (Eds.), *Multi-Source, Multilingual Information Extraction and Summarization* (pp. 93-115). Berlin: Springer. Recuperado de https://bit.ly/30kZqRs el 18/08/2019.

Rayson, P. (2008). From Key Words to Key Semantic Domains. *International Journal of Corpus Linguistics, 13*(4), 519-549. doi:10.1075/ijcl.13.4.06ray

Rose, S., Engel, D., Cramer, N. y Cowley, W. (2010). Automatic Keyword Extraction from Individual Documents. En M. W. Berry y J. Kogan (Eds.), *Text Mining: Applications and Theory* (pp. 1-20). West Sussex: Wiley.

Ruiz Fabo, P. (2017). *Concept-Based and Relation-Based Corpus Navigation: Applications of Natural Language Processing in Digital Humanities* (Tesis doctoral). École normale supérieure, Science et Lettres Research University, París. Recuperado de https://hal.archives-ouvertes.fr/tel-01575167 el 18/08/2019.

Ruiz Fabo, P. y Poibeau, T. (2015). Combining Open Source Annotators for Entity Linking through Weighted Voting. *Joint Conference on Lexical and Computational Semantics (*SEM 2015)* (pp. 211-215). Recuperado de https://hal.archives-ouvertes.fr/hal-01173967/ el 18/08/2019.

Sahle, P. (2013). *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 2: Befunde, Theorie und Methodik*. Norderstedt: BoD. Recuperado de https://bit.ly/2TLYIQh el 20/08/2019.

Smith, N. A., Cardie, C., Washington, A. y Wilkerson, J. (2014). Overview of the 2014 NLP Unshared Task in PoliInformatics. En C. Danescu-Niculescu-Mizil, J. Eisenstein, K. McKeown y N. A. Smith (Eds.), *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science* (pp. 5-7). Recuperado de https://bit.ly/2Nl75vx el 18/08/2019.

Suchanek, F. M., Kasneci, G. y Weikum, G. (2007). YAGO: A Core of Semantic Knowledge. En C. Williamson y M. E. Zurko, *Proceedings of the 16th International Conference on World Wide Web* (pp. 697-706).

Tanselle, G. T. (2006). Foreword. En L. Burnard, K. O'Brien O'Keeffe y J. Unsworth (Eds.), *Electronic Textual Editing* (pp. 3-7). New York: Modern Language Association of America.

Tjong Kim Sang, E. F. y De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. En W. Daelemans y M. Osborne (Eds.), *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL, 4* (pp. 142-147). Edmonton: ACL. Recuperado de https://bit.ly/2Z67xF3 el 18/08/2019.

Van Atteveldt, W. (2008). *Semantic Network Analysis: Techniques for Extracting, Representing and Querying Media Content*. Charleston, SC: BookSurge. Recuperado de http://dare.ubvu.vu.nl/bitstream/handle/1871/15964/complete%20dissertation.pdf?sequence=5 el 18/08/2019.

Van Atteveldt, W., Sheafer, T., Shenhav, S. R. y Fogel-Dror, Y. (2017). Clause Analysis: Using Syntactic Information to Automatically Extract Source, Subject, and Predicate from Texts with an Application to the 2008-2009 Gaza War. *Political Analysis, 25*(2), 207-222.

Venturini, T., Cardon, D. y Cointet, J.-P. (2014). *Réseaux, 6. Méthodes Digitales. Approches quali/quanti des données numériques*. doi:10.3917/res.188.0009

Venturini, T. y Guido, D. (2012). Once Upon a Text: an ANT Tale in Text Analysis. *Sociologica, 6*(3). doi:10.2383/72700.