Revista de Humanidades Digitales

Un estudio empírico con Sketch Engine sobre la interfaz sintáctico-pragmática para la identificación de la estructura temática en español

An Empirical Study with Sketch Engine on the Syntactic-Pragmatic Interface for the Identification of Thematic Structure in Spanish

Elena DEL OLMO SUÁREZ
Universidad Complutense de Madrid
elenadelolmo@ucm.es
https://orcid.org/0000-0002-0510-9975

Iván ARIAS RODRÍGUEZ
Universidad Complutense de Madrid
iarias01@ucm.es
https://orcid.org/0000-0002-5385-3643

RESUMEN

En este artículo abordamos, con un estudio de caso, la formalización de la teoría de la progresión temática con el objetivo de explicitar los esquemas de desarrollo conceptual de los textos. Utilizamos la herramienta de análisis textual Sketch Engine, cuyas funcionalidades y limitaciones serán objeto de discusión, sobre un corpus en español de textos periodísticos anotado morfosintácticamente para indagar acerca de los patrones que permiten inferir la estructura temática de las oraciones. Para ello, llevamos a cabo un estudio de los rasgos morfosintácticos y léxicos que caracterizan a los temas y remas de cada oración y, a continuación, los validamos empíricamente en el corpus. Finalmente, reflexionamos sobre la utilidad de la metodología empleada y planteamos futuras líneas de investigación.

PALABRAS CLAVE

Anotación, etiquetado POS, lenguaje, modelización, reconocimiento de patrones.

ABSTRACT

In this article we address, through a case study, the formalisation of the thematic progression theory in order to render explicit the conceptual development patterns within texts. The text analysis tool Sketch Engine, whose features and shortcomings will be addressed and discussed, is applied to analyse a corpus of morphosyntactically annotated Spanish journalistic texts with the aim of exploring which patterns enable the inference of the thematic structure of the sentences. For this purpose, we carried out a study of the morphosyntactic and lexical features which characterise themes and rhemes of every sentence and, subsequently, we empirically validated them in the corpus. Finally, we discuss the applicability of our methodology and identify future lines of research.

KEYWORDS

Annotation, POS-tagging, Language, Modeling, Pattern Recognition.

RHD 6 (2021)

ISSN
2531-1786

Recibido 06/06/2021 – Aceptado 15/08/2021

1. INTRODUCCIÓN

En este artículo presentaremos un estudio de caso que se enmarca en un proyecto más amplio, cuyo objetivo es la verificación empírica de la propuesta de la teoría de la progresión temática (Daneš, 1974) en un corpus de textos periodísticos en español. En el marco de esta teoría, se propone la existencia de patrones discursivos textuales denominados esquemas de progresión temática. Estos patrones muestran explícitamente la secuencia de ideas o conceptos que se desarrollan a lo largo de un texto, una representación que tiene un enorme interés en la Lingüística Computacional, puesto que permitiría formalizar la interpretación o significado intrínseco de los textos y, en consecuencia, abordar tareas como la generación de resúmenes automáticos, extracción de palabras clave, simplificación de textos para la lectura fácil o sistemas de aprendizaje de primeras lenguas –al explicitar aspectos de la redacción de textos que se engloban en el término cohesión textual–, así como sistemas de enseñanza de segundas lenguas, al reflejar en qué medida el texto del aprendiente se acomoda a los patrones de los hablantes nativos.

La teoría lingüística de la progresión temática parte de la clasificación de los conceptos de una frase en tema y rema, siendo el tema la información consabida y el rema la información nueva. La secuencia de temas y remas en un texto o en una conversación o discurso hablado constituye el esquema de progresión temática o esquema implícito de desarrollo conceptual del mensaje. Los esquemas de progresión temática, además, pueden visualizarse, como se muestra en la figura 1, lo que facilita al lingüista el estudio cualitativo de dichos esquemas. En el eje horizontal aparecen representadas las oraciones del texto, mientras que su eje vertical se correspondería con cada uno de los conceptos relevantes en dichas oraciones. El color rojo, azul o amarillo significa que el concepto es tema –color rojo–, rema –color azul–, o si aparece varias veces como tema y una vez como rema –color amarillo.

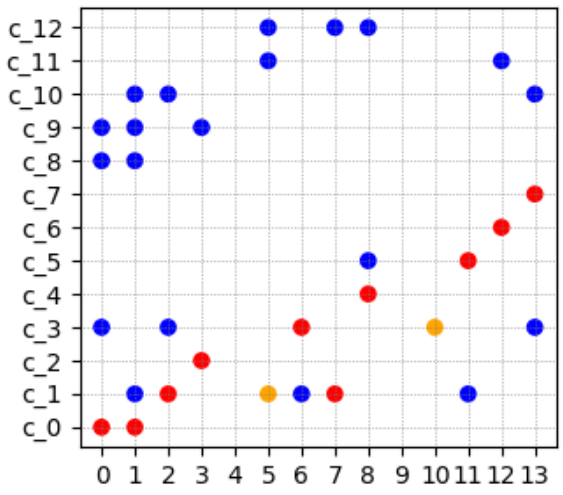


Figura 1. Representación de la progresión temática de un texto. Fuente: elaboración propia.

El problema que abordamos en este artículo es estudiar cómo podrían identificarse de forma automática los temas y remas de cada frase de un texto. Básicamente, la pregunta de investigación es la siguiente: ¿es posible encontrar marcadores sintácticos para la identificación de la es-

tructura temática intraoracional en español? Se trata, pues, de realizar un estudio preliminar cualitativo para identificar estos marcadores con el fin de saber si es viable la identificación automática de los temas y remas de un texto.

Para poder llevar a cabo el estudio, utilizamos una de las mejores herramientas de análisis textual existentes en la actualidad, Sketch Engine[1]. Sketch Engine es una herramienta originalmente desarrollada por el investigador Adam Kilgarriff a partir de sus trabajos sobre lexicografía computacional en los años 90 y principios de 2000 y que continuó su desarrollo en la empresa Lexical Computing, que fundó en 2003. Desde 2018 y hasta 2022 recibe financiación del proyecto europeo ELEXIS, lo que ha permitido a más de 400 universidades y centros de investigación europeos su utilización de forma abierta hasta el 1 de abril de 2022. En la sección 4.2, se presenta un resumen de sus funciones y dónde ampliar esta información o aprender a utilizarla.

La presentación de este estudio empírico sobre los marcadores sintáctico-pragmáticos en español de las estructuras de tema y rema se ha organizado en seis secciones. Esta primera sección ha introducido el objeto de estudio y su motivación. En la segunda sección, se presentan los fundamentos teóricos en los que se basa el trabajo. En la tercera sección, se presentan las hipótesis de trabajo y el objetivo. En la cuarta, las herramientas y metodología utilizadas para el estudio. Las secciones quinta y sexta presentan, respectivamente, los resultados del trabajo y las conclusiones.

2. FUNDAMENTOS TEÓRICOS DEL ESTUDIO

El conocimiento lingüístico que tratamos de formalizar se enmarca en una corriente denominada Lingüística Sistémico-Funcional. El funcionalismo es una corriente lingüística cuyo objeto de estudio, en palabras de Matthiessen (1992, p. 39), es el significado en sí mismo, sin la necesidad de algo externo al mismo, como es el caso de un modelo de mente en la lingüística cognitiva, o un modelo del mundo en la lingüística formal.

El padre de esta corriente, Halliday (1967) propuso delimitar la definición de un concepto tan genérico como el de *cláusula* a partir de los dominios de tres áreas de elección sintáctica. En primer lugar, el dominio de la transitividad vendría dado por el conjunto de opciones para plasmar el contenido cognitivo, es decir, por el conjunto de opciones para la representación lingüística de la experiencia extralingüística, ya venga dada por fenómenos del mundo, sentimientos, pensamientos o percepciones. En segundo lugar, el dominio de la modalidad se identificaría con la organización de los participantes, de tal manera que durante el hecho enunciativo se abrirían varias posibilidades de elección de rol para los hablantes, que pueden realizar diversos actos, tales como preguntar –o ser preguntados–, informar, ordenar, etcétera. Además, la organización de la cláusula obedece a un último dominio de elección obligatorio, el de la dicotomía tema-rema, que constituye la estructura informativa de la misma y cuya categorización binaria agotaría la totalidad de los componentes del mensaje. Este trabajo se centra solo en la localización empírica de los esquemas tema-rema.

[1] Accesible desde: https://www.sketchengine.eu/elexis/.

En las definiciones más clásicas, el tema se define como "aquello de lo que se habla, el punto de partida de una oración", mientras que el rema se define como "lo que se dice del tema" (Halliday, 1967). De acuerdo con ello, el tema constituiría aquella información presentada como conocida, asumida o que sirve de soporte para el rema, que es la información nueva, desconocida o no compartida en el punto en el que se profiere la oración, esto es, el aporte de dicha oración. Para Halliday (2014, p. 66), la delimitación del tema se realizaría localizando el primer sintagma que formara parte de la estructura experiencial de la cláusula. Tal y como señala el autor, el tipo más común de tema es un participante realizado por un sintagma nominal y, en ocasiones, viene anunciado explícitamente con expresiones del tipo *en cuanto a*, *en lo referente a*, etcétera. Sin embargo, como matizaremos más adelante, esta generalización no siempre se cumple. En nuestro trabajo partimos de una descripción de la estructura informativa del enunciado basada en lo que nuestro interlocutor, o bien aquella persona a la que dirigimos nuestro escrito, desconoce, o lo que pensamos que necesita saber (Gutiérrez Ordóñez, 1997, p. 19).

Debido a que la delimitación del tema es eminentemente posicional y lo sitúa en la parte inicial de la oración, ya sea en las caracterizaciones más tradicionales o en las basadas en el concepto de vacío informativo, conviene describir brevemente algunos conceptos que ocupan, necesaria o prototípicamente, esta misma posición en la cadena enunciativa. Estos conceptos son los de tópico y foco, que forman parte de categorías dicotómicas, de la misma manera que las de tema-rema, aunque en distintos niveles de abstracción, tal y como detallaremos a continuación.

La primera de las dicotomías íntimamente ligada a la temática es la de *trasfondo* o *presunción* –que estaría conformado por el conjunto de proposiciones asumidas como verdaderas y compartidas por los participantes en un punto del discurso– y el *foco* o *aserción* –esto es, la información aportada–. Así, por *focalización* se entiende el conjunto de procedimientos por medio de los cuales un constituyente se destaca como foco, es decir, como información nueva o contraria a lo esperado (Cassanova & Franco, 2006), con el fin de realizar una llamada de atención al interlocutor para que advierta la carga semántica de una magnitud (Gutiérrez Ordóñez, 1997, p. 34).

El término, cuyo significado se entremezcla con el del tema, es el de *tópico* o *constituyente topicalizado*. Este concepto se refiere, en sentido general, a la variedad de constituyentes que aparecen en la periferia izquierda, aunque también puede ocupar posiciones en la periferia derecha, e introducen el tema, esto es, la información presupuesta o el enlace con respecto a lo expresado en el resto de la oración (Gallego et al., 2012). El término complementario se conoce como *comentario* y su definición más extendida se limita a describirlo como lo que se enuncia acerca del tópico.

Así, al mismo tiempo que se les atribuye al comienzo de la cadena enunciativa la posición más alta en la jerarquía focalización, dicha posición es también la prototípicamente temática y, asimismo, la que ocupan más frecuentemente los tópicos.

Ya desde la óptica de la interfaz sintáctica, es de esperar una gran frecuencia de sintagmas con la función sintáctica de sujeto como temas oracionales en lenguas con el orden prototípico sujeto-verbo-objeto; tal es el caso del español y del inglés. Que la posición temática venga ocupada por el sujeto antepuesto constituiría, por lo tanto, la opción no marcada, es decir, la opción por

defecto en este tipo de lenguas. Sin embargo, dependiendo del grado de libertad de una lengua en cuanto al orden de palabras, es posible que otros constituyentes precedan al sujeto y ocupen, por tanto, posiciones temáticas. Llamaremos a dichos constituyentes *temas marcados*. Según Halliday (2014, p. 78), los temas marcados siempre implican algún matiz de tipo contrastivo, bien porque contradicen las expectativas asumidas para el lector, bien porque el propósito del autor es remarcar dicho contraste.

Downing (1992, p. 246), por su parte, pone de manifiesto la alta frecuencia con la que los temas marcados se corresponden con información nueva, de tal manera que se considerarían componentes focalizados al comienzo de la oración. Debemos matizar, sin embargo, que en el caso de otros elementos en posición inicial considerados temas marcados por algunos autores, como es el caso de los delimitadores de marco, no existe esta mayor frecuencia de correspondencia con el aporte de la oración. Así, por ejemplo, en la oración *Como destino vacacional, Madrid plantea una serie de inconvenientes que debemos tener en cuenta*, la potencialidad de Madrid como destino turístico puede formar parte de la información nueva en ese punto del texto o bien de la información consabida o asumida como compartida por el autor con los destinatarios del texto, sin que exista en este caso una opción por defecto para su categorización.

Conviene remarcar la también relativa alta frecuencia de aparición de sujetos pospuestos, tanto si estos constituyen la opción marcada –como ocurre en *Escogieron primero los que llegaron antes*– como si constituyen la opción no marcada, como en el caso de los verbos inacusativos o ergativos del español –obsérvese, por ejemplo, la oración *Esta mañana llegó un paquete para ti*, o la oración *En primavera crecieron unas flores preciosas en las macetas que tenía en la fachada oeste*. Nótese, además, que la posposición del sujeto es la opción más natural cuando un constituyente aparece focalizado, esto es, en la periferia izquierda de la oración, como ocurre por ejemplo en *Con un cinco pelado aprobó mi hijo*.

Quizá la pregunta más relevante que quepa plantearse con respecto al objetivo de nuestro trabajo, una vez revisados los fundamentos teóricos en los que se sustenta, tenga que ver con nuestra asunción acerca de la existencia de una interfaz explícita en español entre el nivel pragmático, específicamente el relativo a la estructura temática, y el nivel sintáctico. La cuestión pendiente es determinar si el orden de constituyentes relativamente libre de nuestra lengua sigue verdaderamente un patrón arbitrario, como parece sugerir el término *libertad* incluido en dicha característica tipológica, o bien si se correlaciona con algún aspecto implícito o explícito del hecho comunicativo.

Existe un amplio consenso sobre que el rendimiento del orden de palabras en la significación en cuanto a su capacidad de mostrar funciones representativas es, en español, más bien escaso (Gutiérrez Ordóñez, 1997, p. 29). Desde la óptica de la función informativa del lenguaje, sin embargo, se constata justo lo contrario: la organización de una secuencia se correlaciona efectivamente con el grado de informatividad de cada uno de sus componentes, de tal manera que las oraciones se conforman incluyendo en su primera parte el soporte o información conocida –el tema – que se toma como base para, a continuación, incluir el aporte o información nueva –el rema. Como ya hemos adelantado, el objetivo de nuestro proyecto es justamente demostrar empíricamente

dichas correspondencias. Paralelamente, desde el ámbito de la neurolingüística, aunque sobrepase las pretensiones de este trabajo, diversos autores han demostrado la tendencia a que los argumentos más prominentes precedan a los menos prominentes en términos de orden lineal (Bornkessel-Schlesewsky & Schlesewsky, 2009; Haupt et al., 2008). Así, mediante el análisis de diversos neuromarcadores, se ha constatado que el reanálisis que afecta al orden lineal de las oraciones se procesa de una manera característica, que difiere de la manera de procesar otros tipos de variaciones sintácticas.

3. HIPÓTESIS Y OBJETIVOS

El estudio que se realiza en este trabajo parte de la hipótesis de que es posible, mediante la ayuda de una herramienta de análisis automático de textos, identificar los marcadores morfosintácticos o léxicos de los roles de tema y rema en las frases.

El objetivo general de este estudio de caso es, por lo tanto, estudiar, con la ayuda de la herramienta de análisis textual Sketch Engine, si es posible identificar empíricamente marcadores morfosintácticos o léxicos de los roles de tema y rema, o de estructuras intermedias que posibiliten su posterior identificación, a partir de las oraciones de un corpus de texto en español.

4. HERRAMIENTAS Y METODOLOGÍA

4.1. El corpus Ancora Surface Syntax Dependencies

Para llevar a cabo el estudio se ha escogido el corpus Ancora Surface Syntax Deps (GLiCom-UPF) 1.1, en adelante corpus Ancora, con 17.376 oraciones. El corpus Ancora incluye 225.000 palabras procedentes de textos de la Agencia EFE, 200.000 palabras procedentes del medio escrito *El Periódico* y 75.000 palabras procedentes del corpus LexEsp (Sebastián et al., 2000). Contiene anotaciones correspondientes al nivel morfológico, con las categorías y los rasgos pertinentes para cada palabra, y al nivel sintáctico, con la división de cada oración en constituyentes, la clasificación de estos según su función sintáctica y la transformación automática de dicha estructura de constituyentes al esquema dependencial de núcleo único. Los motivos para escogerlo son su disponibilidad gratuita, su etiquetado de diversos niveles de lengua revisado a mano y el hecho de que es un corpus ampliamente utilizado en el ámbito del procesamiento del lenguaje natural, por lo que puede utilizarse como corpus de referencia en cuanto a los resultados[2].

El formato del corpus es CoNLL-2006 (Buchholz & Marsi, 2006)[3], y sus campos correspon-

[2] Ancora lleva empleándose para la evaluación de tareas de procesamiento del lenguaje natural desde la edición de 2006 de la *Conference on Computational Natural Language Learning* y desde la edición del 2007 del *International Workshop on Semantic Evaluation*, las jornadas científicas más relevantes en dicho ámbito. Además, con él se han entrenado numerosos modelos de análisis lingüístico automático, entre los que destacan por su relevancia actual los de spaCy.

3 El formato CoNLL (cuyo nombre procede de la *Conference on Computational Natural Language Learning*) constituye una convención ampliamente utilizada para la codificación de datos lingüísticos en la que cada palabra aparece representada en una línea con una serie de campos tabulado, en los que se incluye su identificador y sus rasgos distintivos. Los campos y el orden relativo de los mismos varían en función de la version.

dientes al nivel morfológico son la forma, el lema y la categoría morfológica, mientras que sus campos correspondientes al nivel sintáctico son el identificador del núcleo y tipo de dependencia sintáctica que lo relaciona con él. Puede observarse un ejemplo de oración etiquetada en la figura 2.

```
1   Vía Vía np00001 np00001 _   3   nsubj
2   Pública Pública np00001 np00001 _   1   appos
3   ha  haber   vaip3s0 vaip3s0 _   0   root
4   repartido   repartir    vmp00sm vmp00sm _   3   vobj
5   por por sps00   sps00   _   4   prepv
6   el  el  da0ms0  da0ms0  _   7   det
7   momento momento ncms000 ncms000 _   5   pobj
8   siete   siete   dn0cp0  dn0cp0  _   9   det
9   cámaras cámara  ncfp000 ncfp000 _   4   dobj
10  ,   ,   fc  fc  _   3   punct
11  todas   todo    di0fp0  di0fp0  _   12  det
12  ellas   él  pp3fp000    pp3fp000    _   13  pobj
13  en  en  sps00   sps00   _   4   prepv
14  el  el  da0ms0  da0ms0  _   15  det
15  Eixample    Eixample    np00000 np00000 _   13  pobj
16  .   .   fp  fp  _   3   punct
```

Figura 2. Oración anotada del Ancora Surface Syntax Deps. Fuente: elaboración propia.

En cuanto a la tipología textual, como ya hemos adelantado, el Ancora es un corpus de textos periodísticos. La categoría textual es especialmente relevante desde la óptica de la corriente funcionalista en los aspectos relativos a su función del lenguaje predominante (Davies, 1997). En la categoría de textos que componen el corpus escogido, lo esperable es que predomine la función ideacional. Sin embargo, es asimismo esperable la presencia de fragmentos de tipo textual en un porcentaje relativamente alto para mantener la cohesión textual, ya que éste es un objetivo explícito de los textos periodísticos. En cuanto al tipo interactivo, el tipo de texto rige una presencia reducida de los fragmentos en los que se materializa como dominante la función interactiva, por estar los textos periodísticos prototípicamente dirigidos a un lector neutro, cuya relación con el escritor se limita a la del contexto de la informatividad en la prensa escrita.

Este predominio de la función ideacional del lenguaje en el corpus escogido justifica asimismo la decisión tomarlo para este trabajo, ya que las categorías que estudiamos, relativas a la estructura temática, se sitúan al nivel de la estructura informativa del enunciado.

4.2. La herramienta Sketch Engine

La herramienta de análisis textual escogida ha sido la anteriormente mencionada. Sketch Engine. Esta herramienta de exploración de corpus *s* capaz de inferir los patrones y las especificidades de corpus textuales en varias lenguas, entre las que se incluye el español. Cuenta con distintas herramientas de anotación automática y análisis estadístico parametrizables por el usuario, así

como con motores de búsqueda con distintos niveles de abstracción para realizar consultas en todo el corpus, o bien en subconjuntos definidos a partir del mismo.

Entre las múltiples herramientas de que dispone Sketch Engine, cabe destacar especialmente dos de ellas. La primera, denominada Word Sketch, recibe una pieza léxica dada y es capaz de encontrar aquellas otras palabras del corpus de estudio con las que dicha pieza léxica mantiene un determinado tipo de relación sintáctica. Así, es posible encontrar, por ejemplo, aquellos nombres a los que más comúnmente se les modifica un cierto adjetivo, los verbos para los que un cierto nombre actúa frecuentemente de objeto, o sintagmas preposicionales que usualmente aparecen modificando a un cierto nombre. Esta es una herramienta muy funcional, ya que permite hacerse una idea rápida de cómo se trata a una determinada pieza léxica dentro del corpus.

Otra herramienta especialmente interesante para los objetivos de este trabajo es la de Concordancia. Esta herramienta permite encontrar aquellos puntos en el corpus en los que aparece una cierta palabra o expresión. La herramienta es muy flexible, ya que permite utilizar expresiones regulares para definir aquello que se busca, y lo que es aún más útil, permite el uso del llamado lenguaje de consultas de corpus (CQL, *Corpus Query Language*)[4], que admite la búsqueda de estructuras sintácticas muy específicas. En concreto, este lenguaje hace búsquedas por formas concretas, lemas, o incluso permite utilizar restricciones en las etiquetas morfosintácticas con las que Sketch Engine anota automáticamente el texto. Así, es posible referirse a tiempos o modos verbales concretos, así como a números o géneros, o en general a cualquiera de las informaciones que dan las etiquetas utilizadas. Sketch Engine utiliza el etiquetador FreeLing (Padró & Stanilovsky, 2012), que asigna etiquetas basadas en las recomendaciones EAGLES (Leech & Wilson, 1996).

También resulta relevante para este trabajo la herramienta tesauro. Dado un cierto lema, esta herramienta analiza el corpus tratando de encontrar aquellos otros lemas que tienen una mayor similitud semántica con él. Para identificar estos lemas, Sketch Engine procesa el texto del corpus, identificando con qué otros contextos se asocia habitualmente la pieza léxica dada en términos sintácticos, para posteriormente buscar otros contextos asociados usualmente con esas mismas palabras involucradas. Así, los resultados obtenidos son en cierta forma una lista de lemas que resultan intercambiables con el lema dado al aparecer típicamente en las mismas posiciones sintácticas y en contextos equivalentes. Por ejemplo, si utilizamos la herramienta Tesauro con perro, los resultados típicos (que obviamente dependen del corpus que se esté usando) serían gato, mascota o animal, pero también algunos otros que si bien serían normalmente intercambiables con el lema perro en una oración, comienzan a tener una semántica diferente, aunque en muchos aspectos compatible, como bebé, niño, pareja o caballo. En cualquier caso, los lemas aparecen ordenados según su cercanía basándose en una métrica que indica el grado de coincidencia de las palabras con que se relacionan los lemas de la lista en comparación. Allí coaparecen con el lema para el que se ha manejado la herramienta.

Finalmente, merece destacarse que Sketch Engine dispone de múltiples posibilidades de res-

[4] Accesible desde: https://www.sketchengine.eu/documentation/corpus-querying.

tricción de la búsqueda, por ejemplo, buscar palabras que tengan con otras una relación sintáctica del tipo definido en la herramienta Word Sketch, permitiendo hacer búsquedas de patrones de palabras que aparezcan dentro de patrones de palabras mayores, así como obtener los resultados de varias búsquedas en un único resultado combinado, o incluso establecer restricciones según aquellas palabras concretas que se haya encontrado. Por ejemplo, es posible encontrar frases en las que aparezca un verbo en pretérito imperfecto de subjuntivo y en la que posteriormente aparezca ese mismo verbo en condicional de imperativo, y que la persona y el número de ambos sea el mismo. Serán estas funciones las que se utilizarán en este estudio para validar los rasgos morfosintácticos indicativos de los roles de tema y rema[5].

4.3. Metodología de trabajo

Para realizar el estudio se ha seguido una metodología que comprende las tres fases consecutivas siguientes:

- Fase 1. Estudio de los rasgos morfosintácticos y léxicos de determinación de los roles de tema y rema. Para este estudio se escogieron de forma aleatoria diez textos del corpus y se anotó manualmente la información relativa a la estructura temática de las oraciones, se identificaron los posibles rasgos morfosintácticos y léxicos útiles para su identificación y los casos problemáticos o de difícil identificación. Para asegurar su objetividad, se realizó aplicando una anotación y evaluación independiente por pares.
- Fase 2. Validación empírica en el corpus de los rasgos observados en la anterior fase en la totalidad del mismo, y estudio exploratorio de nuevos posibles rasgos con el apoyo de la herramienta Sketch Engine.
- Fase 3. Análisis y síntesis de resultados con el fin de obtener los patrones observados de rasgos morfosintácticos y léxicos de caracterización de los roles de tema y rema en las oraciones.

5. RESULTADOS Y DISCUSIÓN

5.1. Resultados y discusión de la fase 1

En la fase 1 se han anotado manualmente 10 textos del corpus, que contienen un total de 153 oraciones, y se han abstraído una serie de estructuras lingüísticas con el objetivo de clasificar las oraciones en función de la casuística relevante para su anotación temática automática. El relativamente escaso porcentaje de oraciones anotadas para nuestro análisis lo justifica el elevado coste de su anotación manual. Dichas estructuras de rasgos morfosintácticos y léxicos, junto con el porcentaje de las oraciones en que se materializan, son los siguientes:

[5] Se puede consultar Arias et al. (2020) para una descripción detallada de la herramienta o bien la documentación disponible en su página web: https://www.sketchengine.eu/documentation.

Marcadores lingüísticos (en la proposición considerada más informativa)	**Porcentaje de oraciones**
1. Verbos principales de habla y verbos principales de experiencia psicológica complementados por completivas de objeto directo	25.5% (39/153)
2. Sujetos antepuestos al verbo principal	75.2% (115/153)
3. Complementos distintos del sujeto antepuestos al verbo principal	19.6% (30/153)
4. Oraciones sin complementos en posición preverbal y con sujeto no elidido	5.2% (8/153)
5. Verbos principales inacusativos	2.6% (4/153)
5.1. Verbos principales inacusativos con sujetos antepuestos	0.7% (1/153) 25% del total de verbo inacusativos(1/4)
6. Pasivas reflejas	3.9% (6/153)
6.1. Pasivas reflejas con sujetos antepuestos	1.3% (2/153) 33.3% del total de pasivas reflejas (2/6)

Tabla 1. Estructuras relevantes para la anotación automática de la estructura temática. Fuente: elaboración propia.

Incluimos ahora un ejemplo de cada categoría, junto con el identificador del texto del que se ha extraído:

1. El euro, incluso antes de ser creado, protegió a la UE durante el año pasado de la grave crisis financiera internacional y la convirtió en un polo de estabilidad, *señaló* Edlinger. (3_19990102_ssd)
2. *Un español* tiene alrededor del 60% de la renta de un alemán. (3_19990102_ssd)
3. *En Catalunya* se ha experimentado la posibilidad de que las víctimas lleven una alarma en la muñeca para contactar con la policía. (1_20020901_ssd)
4. *Hay* factores como la renta per cápita que pesan y que condicionan las políticas de precios. (3_19990102_ssd)
5. En el plan *participó* un grupo de mujeres de Girona, pero se comprobó que no era efectivo ya que sólo funcionaba en el domicilio y era más fácil que la afectada llevara un teléfono móvil. (1_20020901_ssd)

    5.1. *El más grave y reciente ocurrió* en 1976, cuando un avión de la aerolínea se estrelló durante la maniobra de aterrizaje en el aeropuerto de Bangkok (Tailandia) y fallecieron 72 personas. (3_19991101_ssd)
6. En 13 supuestos *se acordó* el año pasado el establecimiento de un servicio policial de vigilancia y protección personal de la víctima. (1_20020901_ssd)

7. La hija de María del Carmen Costa, que también se llama Carmen, afirmó ayer que la instalación era nueva, y subrayó que *la espita de entrada del gas se cambió* el pasado julio. (1_20020202_ssd)

En cuanto a la relevancia de las categorías propuestas para la delimitación de la estructura temática de las oraciones, hemos observado en primer lugar que, en un porcentaje muy elevado, las oraciones englobaban más de una proposición. Este hecho dificultaba en grado sumo el análisis de la estructura temática de dichas oraciones. Por ello, se delimitó la unidad de análisis para la anotación del tema y el rema oracionales a la proposición más relevante en términos informativos. Obsérvese que, en todas las demás categorías, el alcance se restringe a dicha proposición considerada la de mayor informatividad.

La limitación del análisis a la proposición más informativa favorece el filtrado de la carga informativa de las oraciones. Provoca, sin embargo, la consecuente pérdida de información del resto de proposiciones, cuyos temas y remas podrían estar conectados con los de las proposiciones principales y conformarían diversos patrones de desarrollo, cuya complejidad dejamos fuera del análisis. En todo caso, esta información será objeto de un estudio posterior, una vez resuelto el problema de identificación de los temas y remas principales, mediante la aplicación de un método similar al empleado en este estudio.

En este sentido, se encontró que la estrategia para delimitar la proposición más informativa podría consistir en:

1. Si la relación entre las proposiciones de una oración es paratáctica, esto es, las proposiciones involucradas se sitúan en un mismo plano jerárquico, la oración completa se sustituye por tantas oraciones como proposiciones involucradas en relaciones de dicho tipo contenga dicha oración completa.
2. Si la relación es hipotáctica, esto es, si las proposiciones involucradas pertenecen a distintos planos jerárquicos, se asume que la proposición principal es la más informativa. La viabilidad de esta asunción se discute en la segunda fase de este trabajo.

Durante esta fase, sin embargo, hemos constatado la elevada frecuencia de los verbos propuestos en la primera categoría –los verbos de habla y de experiencia psicológica– en proposiciones principales. Consideramos que, en los casos en que su objeto directo viene saturado por una proposición, es dicha proposición subordinada la que debe seleccionarse como informativamente más relevante. Por ello hemos seleccionado como marcador para la anotación de la estructura temática oracional la presencia de este tipo de verbos acompañados de completiva en las proposiciones principales, ya que constituyen una excepción muy frecuente a nuestra decisión por defecto para los casos de hipotaxis.

La motivación de la elección de las siguientes categorías es bastante más trivial. En el caso de la segunda, recordemos que, en el orden no marcado en español, la posición temática la vienen ocupando, por defecto, los sujetos antepuestos, por lo que es esperable que localizar dicha categoría implique la anotación automática de la estructura temática en un elevado porcentaje de ca-

sos. Conviene, sin embargo, localizar también los casos en los que encontramos complementos distintos al sujeto en posiciones preverbales –tercera categoría–, ya que el estatus informativo que les aporta su posición prototípicamente temática es susceptible de generar patrones de progresión temática característicos que interesa estudiar. De la misma manera, conviene localizar aquellas estructuras temáticas en las que no hay ningún complemento antepuesto al verbo y el sujeto de éste no esté omitido –cuarta categoría–, ya que la estructura temática que presentan se caracteriza por carecer de tema y este patrón de organización informativa.

Por último, las categorías restantes abstraen dos tipos de estructuras en las que el orden no marcado en español no rige la anteposición del sujeto sino, justamente, su posposición. Es el caso de los verbos inacusativos –quinta categoría– y las pasivas reflejas –sexta categoría–. A pesar de su frecuencia relativamente baja en los textos anotados, constituyen estructuras relevantes para la anotación de la estructura temática al presentar un orden no marcado inverso al habitual. Puede comprobarse que, efectivamente, los casos en los que presentan sujetos antepuestos son la minoría –entre un cuarto y un tercio de los casos–, en contraste con lo que ocurre con el corpus total –en el que los sujetos antepuestos aparecen en, aproximadamente, tres cuartas partes de las oraciones. Además, nuevamente, conforman estructuras informativas que merece la pena estudiar, ya que son susceptibles de favorecer ciertos tipos de patrones de progresión temática.

5.2. Resultados y discusión de la fase 2

En esta segunda fase se han diseñado y analizado los marcadores descritos en la fase anterior en la totalidad del corpus gracias a las funcionalidades disponibles en la herramienta Sketch Engine, con el objetivo de constatar la relevancia de las categorías propuestas para el análisis de la estructura temática oracional y, si es posible, localizar nuevas categorías o subcategorías relevantes.

En cuanto al primer marcador, recordemos que con él tratábamos de delimitar los casos de hipotaxis que deben ser contemplados de manera independiente para la aplicación de reglas de anotación automática, por considerarse la subordinada y no la proposición principal como la de mayor relevancia informativa.

Para generalizar esta categoría que caracterizamos en la fase anterior, hemos recurrido en primer lugar a la herramienta Tesauro de Sketch Engine, que permite buscar términos en función de su lema, clasifica el resto de lemas del *corpus* en función de su similitud semántica con nuestra búsqueda en cuestión, y muestra, además, la frecuencia absoluta de éstos. Hemos podido constatar la calidad de la salida para un lema prototípico de los verbos habla, *decir*, y un lema prototípico de los verbos de experiencia psicológica, *creer*. Hemos escogido estos lemas como ejemplos para este trabajo por considerar que presentan una especificidad semántica mínima con respecto a las categorías de las que los consideramos representantes prototípicos.

Mostramos a continuación los 50 lemas más similares semánticamente que devuelve Sketch Engine para los lemas *decir* y *creer*, en la figura 3 y la figura 4, respectivamente.

	Word	Frequency		Word	Frequency		Word	Frequency		Word	Frequency		Word	Frequency
1	explicar	305	11	declarar	151	21	asegurar	364	31	existir	154	41	encontrar	294
2	señalar	242	12	llegar	443	22	ocurrir	117	32	pasar	832	42	recordar	185
3	afirmar	262	13	saber	311	23	esperar	199	33	hablar	190	43	aprobar	125
4	informar	176	14	ir	616	24	ofrecer	132	34	apoyar	95	44	trabajar	131
5	haber	4,888	15	anunciar	237	25	realizar	194	35	participar	130	45	crecer	68
6	ver	449	16	mantener	320	26	contar	222	36	considerar	307	46	cerrar	122
7	llevar	384	17	conocer	222	27	indicar	169	37	acusar	115	47	entrar	128
8	hacer	1,473	18	tener	1,763	28	presentar	324	38	dar	713	48	estar	1,657
9	pedir	284	19	recibir	227	29	reconocer	169	39	llamar	149	49	elegir	76
10	quedar	268	20	salir	194	30	abrir	238	40	poner	383	50	proponer	102

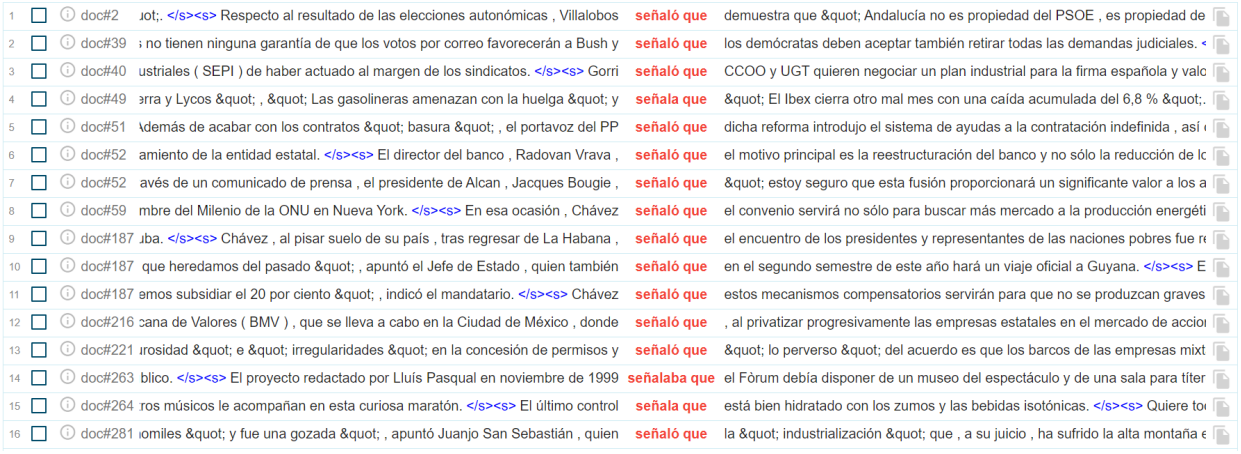
Figura 3. Primeros 50 lemas más similares semánticamente al verbo *decir*. Fuente: elaboración propia.

	Word	Frequency		Word	Frequency		Word	Frequency		Word	Frequency		Word	Frequency
1	admitir	84	11	reclamar	96	21	afectar	122	31	demostrar	108	41	proponer	102
2	sentir	95	12	pedir	284	22	ofrecer	132	32	enseñar	13	42	descender	23
3	defender	129	13	querer	407	23	pagar	95	33	recordar	185	43	temer	31
4	poseer	36	14	parecer	258	24	correr	55	34	necesitar	94	44	confirmar	91
5	esperar	199	15	considerar	307	25	avanzar	59	35	molestar	12	45	recoger	66
6	conocer	222	16	saber	311	26	reivindicar	15	36	marcar	86	46	ver	449
7	reconocer	169	17	usar	57	27	existir	154	37	incluir	141	47	explicar	305
8	crecer	68	18	superar	113	28	confesar	29	38	anunciar	237	48	aprovechar	78
9	aparecer	87	19	formar	114	29	apoyar	95	39	desmentir	23	49	valer	30
10	tardar	26	20	participar	130	30	asegurar	364	40	ocupar	96	50	entender	76

Figura 4. Primeros 50 lemas más similares semánticamente al verbo *creer*. Fuente: elaboración propia.

Como puede observarse, si se analizan ambas listas con detenimiento, es necesario realizar un filtrado posterior, especialmente en el caso de *creer*, ya que estamos interesados en localizar únicamente aquellos verbos en los que lo relevante informativamente sea su completiva, por lo que nos interesa localizar solo las apariciones que contengan una completiva.

Sketch Engine incluye una herramienta que nos facilita el filtrado descrito en el párrafo anterior, la de Concordancias. Dicha herramienta de búsqueda permite el filtrado de secuencias con una extensión propia de Sketch Engine del lenguaje de consultas *Corpus Query Language* (CQL). Así, por ejemplo, en el caso de *señalar*, muy alto en la jerarquía de similitud semántica con *decir*, convendría filtrar la secuencia en la que aparece seguido de la conjunción que, ya que dicha secuencia nos asegura que el sentido que se instancia en las apariciones que recuperemos, será el del verbo de habla y que, además, de él dependerá una cláusula completiva. En la figura 5 pueden observarse algunos de los resultados de la consulta en CQL: [lemma="señalar"][lemma="que"].



1	doc#2	ɹot;. </s><s> Respecto al resultado de las elecciones autonómicas , Villalobos	señaló que	demuestra que " Andalucía no es propiedad del PSOE , es propiedad de
2	doc#39	; no tienen ninguna garantía de que los votos por correo favorecerán a Bush y	señaló que	los demócratas deben aceptar también retirar todas las demandas judiciales. <
3	doc#40	ustriales (SEPI) de haber actuado al margen de los sindicatos. </s><s> Gorri	señaló que	CCOO y UGT quieren negociar un plan industrial para la firma española y valc
4	doc#49	ɔrra y Lycos " , " Las gasolineras amenazan con la huelga " y	señala que	" El Ibex cierra otro mal mes con una caída acumulada del 6,8 % ".
5	doc#51	ʌdemás de acabar con los contratos " basura " , el portavoz del PP	señaló que	dicha reforma introdujo el sistema de ayudas a la contratación indefinida , así
6	doc#52	amiento de la entidad estatal. </s><s> El director del banco , Radovan Vrava ,	señaló que	el motivo principal es la reestructuración del banco y no sólo la reducción de lc
7	doc#52	avés de un comunicado de prensa , el presidente de Alcan , Jacques Bougie ,	señaló que	" estoy seguro que esta fusión proporcionará un significante valor a los a
8	doc#59	mbre del Milenio de la ONU en Nueva York. </s><s> En esa ocasión , Chávez	señaló que	el convenio servirá no sólo para buscar más mercado a la producción energéti
9	doc#187	ɹba. </s><s> Chávez , al pisar suelo de su país , tras regresar de La Habana ,	señaló que	el encuentro de los presidentes y representantes de las naciones pobres fue re
10	doc#187	que heredamos del pasado " , apuntó el Jefe de Estado , quien también	señaló que	en el segundo semestre de este año hará un viaje oficial a Guyana. </s><s> E
11	doc#187	emos subsidiar el 20 por ciento " , indicó el mandatario. </s><s> Chávez	señaló que	estos mecanismos compensatorios servirán para que no se produzcan graves
12	doc#216	ɔana de Valores (BMV) , que se lleva a cabo en la Ciudad de México , donde	señaló que	, al privatizar progresivamente las empresas estatales en el mercado de accio
13	doc#221	ɪrosidad " e " irregularidades " en la concesión de permisos y	señaló que	" lo perverso " del acuerdo es que los barcos de las empresas mixt
14	doc#263	blico. </s><s> El proyecto redactado por Lluís Pasqual en noviembre de 1999	señalaba que	el Fòrum debía disponer de un museo del espectáculo y de una sala para títer
15	doc#264	ɾos músicos le acompañan en esta curiosa maratón. </s><s> El último control	señala que	está bien hidratado con los zumos y las bebidas isotónicas. </s><s> Quiere to
16	doc#281	ɪomiles " y fue una gozada " , apuntó Juanjo San Sebastián , quien	señaló que	la " industrialización " que , a su juicio , ha sufrido la alta montaña e

Figura 5. Concordancias para el lema señalar seguido de la conjunción *que*. Fuente: elaboración propia.

A pesar de que ninguna de las apariciones recuperadas para la búsqueda contiene instancias de la palabra que con una categoría distinta a la de conjunción, convendría incluir restricciones de rasgos múltiples en nuestra consulta, para así evitar la recuperación de pronombres –como ocurriría en una oración del tipo *El producto de los que te señalé que estaba menos accesible era el que realmente quería, pero no me atreví a decírtelo en ese momento*. Dicha búsqueda, que en el *corpus* de estudio devuelve los mismos resultados que la consulta más genérica anterior, sería: [lemma="señalar"][lempos="que-c"].

Aunque el total de los resultados devueltos para la búsqueda anterior (116) se corresponden efectivamente con apariciones de como verbo de habla cuyo objetivo directo presenta forma de completiva, no todos los resultados se corresponden con apariciones en la proposición principal, tal y como se exigiría en el marcador descrito en la fase anterior. Gracias a esta capacidad expresiva, dada por la posibilidad de incluir expresiones regulares en las restricciones para cada palabra en el lenguaje CQL, es posible filtrar las apariciones de los verbos de habla y experiencia psicológica de la lista que generamos anteriormente para recuperar solamente aquellos que son verbos principales, ya que, recordemos, nos interesa seleccionar su completiva solamente cuando ésta depende directamente de un verbo principal –y además no constituye un orden ni está en forma no personal–. La sentencia CQL para lograr esto sería, en el caso del verbo *decir*, [lemma="decir" & tag="VMI.*"], que filtra aquellos verbos principales que no están en forma no personal ni en modo subjuntivo o imperativo, que es el obligatorio para la proposición dependiente en las relaciones hipotácticas. En la figura 6 puede observarse parte de la salida de dicha búsqueda, mientras que en la figura 7 mostramos parte de la salida de su consulta complementaria: [lemma="decir" & tag="VM[^I].*"], que se correspondería con las apariciones del lema decir como verbo no principal de la oración.

Figura 6. Concordancias del lema decir como núcleo de la cláusula principal. Fuente: elaboración propia.

Figura 7. Concordancias del lema decir en subordinadas o en forma no personal. Fuente: elaboración propia.

En cuanto a los siguientes tres indicadores, que se corresponden con los sujetos antepuestos al verbo –segunda categoría–, los complementos antepuestos al verbo distintos del sujeto –tercera categoría– y las oraciones sin complementos antepuestos con sujetos no omitidos, esto es, pospuestos –cuarta categoría–, no hemos podido llevar a cabo su extracción directa con Sketch Engine debido a sus limitaciones para la búsqueda por dependencia sintáctica. Aunque analizaremos este asunto con mayor amplitud en el siguiente apartado, esta carencia en la expresividad de las búsquedas viene dada por la imposibilidad de combinar las restricciones relativas al tipo de relación gramatical con restricciones de precedencia, a causa de que el orden se considera irrelevante en los Word Sketches, que constituyen la única posibilidad de acceder a las dependencias. Sin embargo, sí hemos podido realizar algunas consultas más genéricas para estudiar estas categorías.

En el caso de la tercera categoría, hemos buscado aquellas oraciones que comienzan por preposición, ya que la mayor parte de los complementos distintos del sujeto que aparezcan antepuestos al verbo forzosamente deberán estar precedidos de preposición. La consulta que hemos realizado ha sido <s> [tag="SP"]. Dicha consulta recupera tan solo un 0.49% del corpus total, compuesto por 2.755 apariciones. Hemos analizado manualmente las 100 primeras con el objetivo de descubrir posibles patrones relevantes. En dichas 100 oraciones, hemos constatado la mayor frecuencia –54%– de adjuntos adverbiales –como el marcado con negrita en la oración ***En una in-***

***tervención ante militantes populares**, Villalobos dio las gracias a los militantes de muchos años que hoy tienen una emoción especial*–, seguidos de los delimitadores de marco –como es el caso del marcado con negrita en ***Según cálculos oficiales**, existen alrededor de 18 millones de minas desperdigadas en un único espacio de 288.000 hectáreas de desierto*–, con una frecuencia del 27%. La frecuencia de los elementos con una función textual –como ocurre por ejemplo con el marcado en negrita en ***Por otra parte**, el texto de la patronal contempla la creación de nuevos contratos de trabajo enfocados a la reinserción laboral*– ha resultado del 13%, mientras que los focos –como es el caso del marcado con negrita en ***A todas estas personas** no las vamos a defraudar*, aseguró Villalobos– presentan una frecuencia en los textos revisados del 4%. Por último, se han encontrado algunos casos aislados de errores de anotación -2%.

En el caso de la cuarta categoría, que engloba las oraciones en las que ningún complemento precede al verbo y presentan sujeto pospuesto, en la consulta hemos tenido que obviar la restricción relativa a la precedencia del sujeto. Así, la consulta <s> [tag="V.*"] devuelve 1.402 apariciones, que se corresponden con un 0.25% del corpus total. Con el objetivo de estudiar su casuística y localizar posibles patrones no previstos inicialmente, hemos revisado manualmente las 100 primeras apariciones. La absoluta mayoría se corresponden con verbos con sujeto elidido, cuya referencia es muy accesible en función del contexto precedente, por lo que necesitaríamos revisar un porcentaje mucho mayor de apariciones para estudiar la categoría que nos interesa, o bien utilizar otra herramienta que permita depurar en mayor medida la consulta. Solamente un 7% de los casos presentan sujetos pospuestos. La mayor parte de ellos están conformados por verbos de tipo existencial (existe, hay), aunque también los hay de tipo atributivo (como ocurre en Sigue pendiente una normativa ética para la ficción).

En cuanto a la quinta categoría, referida a aquellas proposiciones cuyo verbo principal es de tipo inacusativo, cuya posición no marcada para el sujeto es la pospuesta al verbo, una opción interesante que proporciona Sketch Engine es la búsqueda a partir de listas de lemas. Al ser la de los verbos inacusativos una categoría que podría considerarse una lista más o menos cerrada, podemos comprobar si los patrones de aparición que hemos observado en nuestra anotación manual se constatan en la totalidad del corpus para una pequeño listado de verbos inacusativos de muestra que se ha comprobado que tienen frecuencias absolutas relativamente altas en el corpus: *llegar*, *faltar*, *quedar*, *aparecer* y *desaparecer*. La consulta CQL que ha permitido esta búsqueda ha sido la siguiente: [lemma="llegar|faltar|quedar|aparecerdesaparecer"]. Parte del resultado obtenido puede observarse en la figura 8.

1	doc#0	8 en la licitación de licencias para construir centrales eléctricas en México y se	quedaron	con dos cada una : Río Bravo y Saltillo para la compañía francesa y Altamira y
2	doc#5	iras a los Juegos Olímpicos , según dijo el propio atleta , que después sólo ha	llegado	a saltar 2,30 metros. </s><s> Ayer , miércoles , sin embargo , Sotomayor afirm
3	doc#6	que en las generales de hace cuatro años , pese a perder casi 30.000 votos y	quedarse	con algo más 1.008.000 y que no sirvió para frenar la derrota global. </s><s> Il
4	doc#6	de los votos al pasar de 547.000 votos a apenas 279.000, mientras que el GIL	aparece	como cuarta fuerza más votada , pese a no obtener escaños con algo más de 3
5	doc#7	inar desde el primer minuto " , dijo el holandés , que recordó que cuando	llega	un nuevo técnico , como ha ocurrido en Tenerife , " siempre hay un impul
6	doc#7	z para exponerle su situación. </s><s> El delantero seguirá en el club , pero si	llega	una buena oferta tendrá las puertas abiertas. </s><s> " Estoy harto de es
7	doc#14	das. </s><s> Una de las cosas que más impresiona al equipo visitante cuando	llega	en autocar es la ubicación del estadio , encajonado en medio de callejuelas est
8	doc#15	t; Será un infierno para el Barça durante los 90 minutos ". </s><s> Rusia	llegó	hoy a la conclusión de que no hay conclusiones definitivas sobre el naufragio d
9	doc#18	hubiese sido un referendo sobre la independencia lo habría perdido , pues no	llega	a la mayoría exigida para dar el paso. </s><s> Pero es que , además , la forma
10	doc#18	erar un incremento del descontento nacionalista en Quebec , pero difícilmente	llegará	a crear dentro de esta legislatura las condiciones ganadoras para que haya un
11	doc#22	ste incumplimiento. </s><s> En Catalunya , el debate sobre tal derecho no ha	aparecido	todavía. </s><s> Creo que debiera iniciarse , porque una condición para que ta
12	doc#23	centenarios censados murieron hace años , aunque su voto por correo pueda	llegar	ahora a su correspondiente colegio electoral. </s><s> Y lo mismo se afirma qu
13	doc#25	conquistado a lo largo de muchos años de lucha. </s><s> Aunque todavía hoy	quedan	lugares y empresas en que el trabajo se realiza en condiciones infrahumanas ,
14	doc#25	nfrahumanas , como regla general el carácter de maldición del trabajo tiende a	desaparecer	. </s><s> El trabajo es la solución para muchas cosas. </s><s> Es la alegría de
15	doc#25	iembro de un equipo , la pasión del intelectual y del científico. </s><s> Aunque	quedan	muchas personas que siguen sintiendo el trabajo como una maldición , el comp
16	doc#28	e de la bolsa de Sao Paulo (Bovespa) cerró ayer con una subida del 8,80% y	quedó	en 8.891 puntos. </s><s> También el mercado de Río de Janeiro , cuyo índice

Figura 8. Concordancias para un listado de verbos inacusativos. Fuente: elaboración propia.

Se ha comprobado en las 100 primeras apariciones que, en consonancia con los resultados obtenidos en la fase 1, el porcentaje de sujetos pospuestos o elididos sobrepasa por mucho al de los antepuestos, que suponen solamente un 27% del total.

Esta herramienta de concordancias es asimismo útil para localizar pasivas reflejas, que constituyen la última de las categorías relevantes de patrones susceptibles de generar estructuras temáticas diferentes de la no marcada en español, según la cual el tema se materializaría en el sujeto expreso y antepuesto. Sin embargo, a pesar de la expresividad de las búsquedas de Sketch Engine, que sí aportan la potencia suficiente, en ocasiones la anotación automática del corpus no es lo suficientemente detallada como para filtrar los resultados como se desearía. Un caso de filtrado deficiente para una de las categorías relevantes para la casuística recién descrita es el de las pasivas reflejas, que incluimos aquí por ser su orden de palabras no marcado distinto al prototípico para la anotación automática de la estructura temática, que, recordemos de nuevo, rige la anteposición del sujeto. Obsérvese cómo, en el caso de las oraciones sexta y séptima, que comparten todos los rasgos morfológicos que codifican las etiquetas *Part of Speech* (*PoS*) tanto del *se* como del verbo, la anotación de los *se* no distingue entre sus apariciones en pasivas reflejas –como es el caso de la séptima oración: Este fin de semana también se disputarán tres eliminatorias– y en verbos pronominales –como es el caso de la sexta oración: *Hewitt se medirá a Meligeni*–.

1	doc#0	rá como asistente en la construcción de Altamira 2 y , posteriormente ,	se	encargará de explotarla como principal accionista . </s><s> EDF y Mitsubis
2	doc#0	en la licitación de licencias para construir centrales eléctricas en México y	se	quedaron con dos cada una : Río Bravo y Saltillo para la co
3	doc#2	ichos años que hoy tienen una emoción especial & quot ; , y	se	confesó también & quot ; especialmente emocionada & quot ; en su
4	doc#4	, salvo el & quot ; todoterreno & quot ; Gustavo Kuerten , no	se	adaptan bien . </s><s> El capitán australiano John Newcombe no pod
5	doc#4	quipo & quot ; aussie & quot ; . </s><s> El doble que	se	enfrentará a la pareja Sandon Stolle-Mark Woodforde será el for
6	doc#4	, abran el fuego en la primera jornada , mientras que Hewitt	se	medirá a Meligeni . </s><s> La otra semifinal enfrentará a España co
7	doc#4	batida del 21 al 23 de julio , pero este fin de semana también	se	disputarán tres eliminatorias para la permanencia en el Grupo Mundia
8	doc#5	, el 15 de agosto en la localidad francesa de Montauban , donde	se	impuso con una marca de 2,28 metros a rivales de segundo orden .
9	doc#6	que no sirvió para frenar la derrota global . </s><s> IU	se	convirtió en la gran derrotada al perder 3 de sus representantes e
10	doc#6	votó al Partido Popular . </s><s> También en Campo Real , donde	se	ubicará el futuro aeropuerto de Madrid casi el 60 por ciento dio
11	doc#7	garle la vida a los blancos . </s><s> Sin Redondo , que no	se	ha recuperado de su esguince en la rodilla derecha , ni Mijato
12	doc#8	correr & quot ; , reconoció Schumacher . </s><s> El bicampeón	se	perderá los grandes premios de Italia (Monza , 12 de septiembre)

Figura 9. Concordancias de la forma *se*. Fuente: elaboración propia.

La documentación sobre las etiquetas morfológicas y la codificación de los rasgos que emplea Sketch Engine para español está disponible en línea, específicamente en el apartado de la documentación correspondiente a la anotación automática en español[6]. Como puede observarse, la codificación de ciertos rasgos morfológicos se lleva a cabo de manera posicional en la propia secuencia de caracteres que conforma la etiqueta *PoS*. Decimos que la capacidad expresiva de Sketch Engine efectivamente permitiría filtrar los verbos en pasiva precedidos por la partícula *se*, o bien directamente estas estructuras a partir de los rasgos de los *se* involucrados en ellas, porque CQL permite la inclusión de expresiones regulares en las cadenas de caracteres que restringen las búsquedas por atributos.

5.3. Resultados y discusión de la fase 3

En la segunda fase hemos utilizado las funciones de Sketch Engine disponibles para el análisis estadístico, el análisis de la similitud semántica y la búsqueda basada en patrones de rasgos. Gracias a ellas, hemos analizado las categorías que detallamos en los resultados de la fase 1. Procedemos ahora a la síntesis el análisis de los patrones observados durante la exploración del corpus de la segunda fase.

Por un lado, verbos principales de habla y verbos principales de experiencia psicológica complementados por completivas de objeto directo. Hemos localizado, gracias a la herramienta de cálculo de similitud semántica de lemas, listados de verbos de los tipos relevantes cuya frecuencia en el corpus es significativa. Además, hemos recuperado aquellos verbos principales que contienen completivas de objeto directo, aprovechando que la completiva debe venir introducida por conjunción, y hemos podido constatar que, efectivamente, en consonancia con nuestros resultados de la primera fase, es su proposición subordinada la que consideramos de mayor relevancia informativa y la que tomaremos, por lo tanto, como unidad de análisis.

Por otro lado, sujetos antepuestos al verbo principal en la proposición considerada más informativa. No hemos podido realizar la consulta correspondiente a este patrón con la especificidad necesaria. La causa es la imposibilidad de combinar restricciones relativas a las dependencias sintácticas y restricciones de precedencia en la herramienta de concordancias. Los rasgos de dependencia sintáctica son únicamente accesibles mediante la sintaxis de los denominados Word Sketches, que, si bien permiten el filtrado por tipos de relación gramatical, entre las que se incluye la de sujeto, no se pueden combinar dichas restricciones con las relativas al orden de palabras. Esta carencia ocasiona que no podamos recuperar por separado los casos marcados y no marcados en cuanto a la estructura informativa, que se rigen por el orden secuencial además de por el tipo de dependencia y que son relevantes para nuestro análisis porque se les supone susceptibles de favorecer ciertos tipos de patrones de progresión temática.

En resumen, no podemos recuperar las dependencias de tipo sujeto que ocurren entre cualquier forma de la categoría nombre y cualquier forma de la categoría verbo, pero no podemos

[6] Accesible desde: https://www.sketchengine.eu/spanish-freeling-part-of-speech-tagset.

distinguir entre aquellas en las que el sujeto precede al verbo –y presentan por tanto la estructura temática prototípica en español– y aquellas en las que el sujeto aparece pospuesto –y presentan por tanto un ordenamiento marcado de los constituyentes en cuanto a su estructura informativa. Obsérvense las siguientes dos consultas, que son equivalentes y recuperan el patrón recién descrito:

- [ws(".+-v",".*subject.*",".+-n")]
- [ws(".+-n",".*subject.*",".+-v")]

Sketch Engine cuenta con la posibilidad de declarar restricciones adicionales relativas al orden de los elementos de la consulta, a través de dos de sus funcionalidades: (i), la identificación de las palabras cuyos rasgos se restringen en la búsqueda en nuestra consulta –precediendo de un entero seguido de dos puntos los corchetes entre los que se encierran los pares clave-valor con los rasgos requeridos–, y, (ii), la función *meet*, que permite establecer condiciones de coaparición entre dos palabras en un cierto contexto precedente –cuya longitud se indica mediante enteros negativos – y siguiente –cuya longitud se indica mediante enteros positivos. Sin embargo, la creación de identificadores está restringida en el caso de los Word Sketches, que constituyen la única posibilidad de acceder a la información dependencial de tipo sintáctico, por lo que no es posible en ningún caso la combinación de rasgos de tipo dependencial sintáctico y rasgos de tipo secuencial.

Asimismo, complementos distintos del sujeto antepuestos al verbo principal en la proposición considerada más informativa. De nuevo, nos encontramos ante un patrón de tipo eminentemente sintáctico que, por la orientación al léxico de la herramienta Sketch Engine, hemos tenido que solventar a partir de su reflejo en el nivel morfológico. En la búsqueda para la exploración del corpus hemos recuperado aquellas oraciones que comienzan por una preposición, ya que los complementos antepuestos distintos del sujeto, mayoritariamente, deben ser introducidos por preposición.

Como resultado del análisis de 100 oraciones de las recuperadas, hemos observado una mayoría de adjuntos adverbiales –54%. Los delimitadores de marco son los siguientes en frecuencia –27%–, seguidos de los ordenadores del discurso –13%– y de los focos –4%. Consideramos todas estas subcategorías relevantes para la descripción de la estructura temática oracional, aunque es especialmente interesante la de los constituyentes focalizados, ya que enfatizan aspectos relativos a la estructura informativa de la cláusula, que son justamente los que tratamos de explicitar.

Los delimitadores de marco, por su parte, constituyen asimismo una categoría distintiva para la estructura temática, ya que restringen el marco de referencias de la cláusula que constituye nuestra unidad de análisis y, con frecuencia, también el de las siguientes oraciones, hecho que conviene explicitar en nuestra anotación. Finalmente, en cuanto a los ordenadores del discurso, aunque se relacionan con la estructura informativa, de momento no los tomaremos en consideración. Sin embargo, como vía de investigación futura, convendría plantearse embeber nuestra anotación de la estructura temática dentro de la estructura retórica del texto y analizar las posibles correlaciones entre los marcadores de ambos niveles abstractos de análisis.

Con respecto a las oraciones sin complementos en posición preverbal y con sujeto no elidido

en la proposición considerada más informativa, nuevamente, hemos tenido que limitar las restricciones del patrón por la imposibilidad de combinación de las restricciones sintácticas y de precedencia. Nuestra búsqueda ha consistido en la recuperación de aquellas oraciones que comienzan con un verbo sin filtrar la presencia de sujetos pospuestos. Por ello, solamente un 7% de los casos que hemos analizado manualmente presentan efectivamente sujetos pospuestos, ya que es más esperable la elisión del sujeto que su aparición pospuesta. Aunque convendría repetir esta búsqueda con la posibilidad de filtrar únicamente los casos relevantes, hemos constatado que la mayoría de las estructuras de este tipo están conformadas por verbos de tipo existencial y de tipo atributivo. La relevancia de este análisis viene dada por la necesidad de validar los casos en los que no se haya extraído ningún tema por su pertenencia a las categorías que se consideren efectivamente atemáticas o remáticas.

En lo que hace a los verbos principales inacusativos en la proposición considerada más informativa, la consulta correspondiente a este patrón se corresponde con la de un listado de lemas procedente de una lista cerrada de verbos. Hemos seleccionado algunos especialmente frecuentes y, en nuestro análisis de 100 de ellos, hemos que, efectivamente, tienden a presentar prototípicamente sujetos pospuestos –en un 73% de los casos. La relevancia de esta categoría estructural para la delimitación de la estructura temática oracional viene dada porque su carácter marcado funciona a la inversa del resto de categorías, por lo que debemos considerar temas marcados los sujetos antepuestos.

Finalmente, con relación a las pasivas reflejas en la proposición considerada más informativa, a pesar de que Sketch Engine cuenta con la potencia suficiente para realizar búsquedas complejas en función de la codificación del estándar de etiquetado, hemos constatado que las categorías contempladas en el etiquetado que ofrece por defecto resultan insuficientes para la búsqueda correspondiente a este patrón, ya que no ofrece una etiqueta distintivas para los verbos ni para el *se* de las pasivas reflejas.

Sin embargo, Sketch Engine contempla la posibilidad de cargar un corpus etiquetado previamente en unos de los estándares más utilizados en Procesamiento del Lenguaje Natural, CoNLL, específicamente su versión del 2006 (Buchholz & Marsi, 2006), tal y como se detalla en su documentación[7]. Esta posibilidad supone, por tanto, una potencial solución al problema relativo a las pasivas reflejas que planteábamos en el apartado anterior, aunque conlleva la revisión de los anotadores automáticos disponibles para español para comprobar si etiquetan con la granularidad requerida, o bien el diseño de un algoritmo que añada rasgos adicionales a las etiquetas morfológicas disponibles, además del posterior procesamiento del corpus para adaptarlo a las particulares del formato que acepta la herramienta.

6. CONCLUSIONES Y TRABAJO FUTURO

Este estudio exploratorio de corpus, guiado por la anotación manual de un pequeño por-

[6] Accesible desde: https://www.sketchengine.eu/documentation/building-sketches-from-parsed-corpora.

centaje del mismo con la información relativa a la estructura temática –esto es, a los temas y los remas de las proposiciones más informativas–, constituye un avance en la validación de la hipótesis del trabajo en el que se enmarca, cuyo objetivo es generar un modelo formal para la anotación automática de la estructura temática.

Sketch Engine constituye una potente herramienta para la exploración de corpus y, gracias a las funcionalidades que implementa, hemos conseguido validar la mayor parte de los patrones estructurales que consideramos relevantes para nuestro objetivo. Las limitaciones que hemos encontrado a la hora de transformar los marcadores en consultas vienen dadas por el enfoque eminentemente léxico de la herramienta. Sin embargo, si bien esto ha supuesto que algunas consultas se han realizado con una genericidad mayor que la deseada, ello ha posibilitado la localización de subcategorías asimismo relevantes para nuestro estudio. Dichas subcategorías merecen ser analizadas con mayor profundidad y, como trabajo futuro, planeamos la revisión de un mayor porcentaje de casos para refinar la delimitación de posibles nuevos marcadores y subcategorías.

Para los siguientes pasos de nuestra investigación, en los que nos centraremos en la creación del modelo formal de reglas de anotación automática de la estructura temática, hemos constatado, gracias a este estudio de caso, la necesidad de buscar un lenguaje de búsqueda con una mayor capacidad expresiva en cuanto a los patrones de dependencia sintáctica. Además, es necesario que la herramienta escogida implemente la posibilidad de reescritura de la estructura de datos del corpus para la creación de las categorías necesarias para la anotación, ya que muchas de ellas son utilizadas por categorías de mayor abstracción –como es el caso, por ejemplo, de la proposición considerada de mayor relevancia informativa, que debe restringir el resto de las categorías aquí analizadas.

Una vez diseñado el algoritmo de anotación automática de la estructura temática, estaremos en disposición de analizar si las estructuras temáticas extraíbles automáticamente presentan efectivamente correlaciones con los patrones de desarrollo conceptual de los textos materializado en su esquema de progresión temática.

REFERENCIAS BIBLIOGRÁFICAS

Arias, I., Fernández-Pampillón, A. M., Samy, D., & Arús, J. (2020). Taller sobre herramientas de análisis textual: La herramienta Sketch Engine. *E-Prints Complutense*. https://eprints.ucm.es/id/eprint/13796/21/Taller de Sketch Engine - Completo - 26-05-2020.pdf

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2009). The Role of Prominence Information in the Real-Time Comprehension of Transitive Constructions: A Cross-Linguistic Approach. *Linguistics and Language Compass*, 3(1), 19-58. https://doi.org/10.1111/j.1749-818X.2008.00099.x

Buchholz, S., & Marsi, E. (2006). CoNLL-X Shared Task on Multilingual Dependency Parsing. En *10th Conference on Computational Natural Language Learning (CoNLL-X)* (pp. 149-164). https://doi.org/10.3115/1596276.1596305

Cassanova, V., & Franco, A. (2006). Tema, rema y focalización: del enunciado al texto. Análisis de títulos y leads de prensa. *Quórum Académico*, 3(2), 54-79. https://dialnet.unirioja.es/

servlet/articulo?codigo=3997660

Daneš, F. (1974). Functional Sentence Perspective &the Organisation of the Text. En František Daneš (Ed.), *Papers on Functional Sentence Perspective* (pp. 106-128). Mouton. https://doi.org/10.1515/9783111676524.106

Davies, F. (1997). Marked Theme as a Heuristic for Analysing Text-Type, Text & Genre. En Jordi Piqué & David Viera (Eds.), *Applied Languages: Theory &Practice in ESP* (pp. 45-80). Universidad de Valencia.

Downing, A. (1992). Organising the Message: Thematic & Information Structures of the Clause. En Downing, A. & Locke, P. (Eds.), *A University Course in English Grammar* (pp. 220-264). Prentice-Hall.

Gallego, Á., & Brucart, J. M. (2012). *El movimiento de constituyentes*. Visor Libros.

Gutiérrez O. , S. (1997). *Temas, remas, focos, topicos y comentarios*. Arco Libros.

Halliday, M. A. K. (1967). Notes on Transitivity & Theme in English Part I. *Journal of Linguistics*, *3*(1), 37-81. https://doi.org/10.1017/S0022226700012949

Halliday, M. A. K., Matthiessen, C. M. I. M., Halliday, M. A. K., & Matthiessen, C. (2014). *An Introduction to Functional Grammar*. Routledge. https://doi.org/10.4324/9780203783771

Haupt, F., Schlesewsky, M., Roehm, D., Friederici, A., & Bornkessel-Schlesewsky, I. (2008). The Status of Subject-Object Reanalyses in the Language Comprehension Architecture. *Journal of Memory & Language*, *59*(1), 54-96. https://doi.org/10.1016/j.jml.2008.02.003

Leech, G. & Wilson, A. (1996). *EAGLES: Recommendations for the Morphosyntactic annotation of Corpora*. https://home.uni-leipzig.de/burr/Verb/htm/LinkedDocuments/annotate.pdf

Matthiessen, C. (1992). Interpreting the Textual Metafunction. En M. Davies & L. Ravelli (Eds.), *Advances in Systemic Linguistics: Recent Theory & Practice* (pp. 37-82). Pinter.

Padró, L., & Stanilovsky, E. (2012). Freeling 3.0: Towards Wider Multilinguality. En N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani & S. Piperidis (Eds.), *Proceedings of the 8th International Conference on Language Resources & Evaluation (LREC'12)* (pp. 2473 -2479). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/430_Paper.pdf

Sebastián, N., Martí, M. A., Carreiras, M. F., & Cuetos, F. (2000). *LEXESP: Léxico Informatizado del Español*. Ediciones de la Universitat de Barcelona.