ISSN : 2105-1046 Les Cahiers du GERES

Jasper DEGRAEUWE, Patrick GOETHALS y Pauline VERHOEVE: « Ampliar la caja de herramientas del análisis del discurso asistido por el ordenador: el caso de los cinco sentidos en el discurso turístico »

Universidad de Gante (Bélgica)
Jasper.Degraeuwe@UGent.be
Patrick.Goethals@UGent.be
pauline.verhoeve@gmail.com

Resumen:
En este artículo se presenta un método avanzado para cuantificar el «carácter clave» (*keyness*) de particularidades lingüísticas en los corpus de lenguas de especialidad. Para demostrar su uso, aplicamos la metodología a la dimensión sensorial en el lenguaje turístico. Los resultados muestran que la dimensión sensorial es más de tres veces más clave para el lenguaje turístico de lo que se podía esperar en general, corroborando así la importancia del fenómeno desde una perspectiva cuantitativa. Además, estudiamos la introducción innovadora de un componente de desambiguación semántica automática en la metodología, para lo cual desarrollamos un sistema basado en frases de ejemplo. El sistema consigue una exactitud de entre el 70% y el 80%, y podría convertirse en una herramienta decisiva para escalar la codificación semántica de los corpus de lenguas de especialidad.
Palabras clave:
análisis del discurso asistido por el ordenador, análisis de carácter clave, desambiguación del significado de las palabras, lingüística de corpus, tratamiento automático del lenguaje

Résumé :
Cet article présente une méthode avancée pour quantifier le «caractère clé» (*keyness*) des particularités linguistiques dans les corpus de langues de spécialité. Pour démontrer son utilisation, la méthode s'applique à la dimension sensorielle dans le langage du tourisme. Les résultats montrent que la dimension sensorielle est plus de trois fois plus typique pour le langage touristique que pour le «langage général», corroborant ainsi l'importance du phénomène d'un point de vue quantitatif. De plus, nous étudions l'introduction innovante d'un processus de désambiguïsation lexicale automatique dans la méthodologie, pour laquelle nous développons un système basé sur des phrases exemples. Le système atteint une précision de 70 à 80%, et il pourrait devenir un outil décisif pour le développement de l'annotation sémantique des corpus de langues de spécialité.
Mots-clés : analyse du discours assistée par ordinateur, analyse de caractère clé, désambiguïsation lexicale, linguistique de corpus, traitement automatique de la langue

1. INTRODUCCIÓN

Para la investigación lingüística, el avance tecnológico de poder compilar casi sin limitaciones textos en corpus digitales ha abierto todo un abanico de nuevas oportunidades. Sin embargo, estas oportunidades solo cobran valor si se dispone de los métodos adecuados para procesar los corpus y extraer de ellos los datos deseados, que es el ámbito al cual pretende contribuir el presente trabajo. Nos centramos en la lingüística de corpus para fines investigativos, presentando una metodología que permite estudiar y cuantificar particularidades lingüísticas de lenguas de especialidad y que combina técnicas del análisis de corpus con las más recientes evoluciones en el ámbito del Procesamiento del Lenguaje Natural (NLP por sus siglas en inglés). Para mostrar su uso, aplicamos nuestra metodología a una de las particularidades lingüísticas que dan forma al lenguaje específico del turismo: las referencias textuales a los cinco sentidos.

Empezamos este trabajo por presentar el lenguaje turístico como lengua de especialidad (DANN, 1996), prestando especial atención a la dimensión de los « cinco sentidos », que es uno de los medios lingüísticos/retóricos que se utilizan para describir y promocionar experiencias y destinos turísticos. En la sección 3, profundizamos en la metodología[1] mediante la cual pretendemos desvelar y cuantificar el « carácter clave » (*keyness*) del fenómeno lingüístico en el lenguaje específico. Después, exploramos la integración de un componente de desambiguación semántica automática en la metodología, lo cual debería permitirnos ir aún más allá del tipo de análisis llevado a cabo en la sección 3.

2. ESTUDIO DE CASO: LOS CINCO SENTIDOS EN EL LENGUAJE TURÍSTICO

Para la promoción de destinos y actividades, los textos turísticos se sirven de un discurso propio llamado «el lenguaje turístico» (*the language of tourism*; DANN, 1996). Es una lengua de especialidad que, ofreciendo descripciones detalladas del destino turístico en textos promocionales, pretende convertir a turistas potenciales en turistas reales (SALIM, IBRAHIM y HASSAN, 2012). Para alcanzar este objetivo, el lenguaje turístico se convierte en un instrumento de persuasión que se dirige a su público en términos de sus propias necesidades y motivaciones culturales, transmitiendo así una ideología que determina lo que deberías ver y hacer como turista.

Para transmitir esta « ideología turística », el discurso promocional se apoya en una serie de «tópicos retóricos» (como por ejemplo « viajar es escaparse », « viajar nos cambia », « viajar nos hace salir fuera de lo común » o « viajar nos permite vivir experiencias auténticas »), que se plasman en los textos promocionales a través de una serie de medios lingüísticos y retóricos. Además del uso de adjetivos y metáforas y el empleo de un lenguaje literario (DURÁN-MUÑOZ, 2019; GOETHALS y SEGERS, 2016; JAWORSKA, 2017; MANCA, 2008), una técnica retórica recurrente es la de utilizar palabras y expresiones relacionadas a los cinco sentidos.

Esta dimensión sensorial en los textos turísticos se puede manifestar en la descripción de un destino o una actividad, como queda claro en los siguientes ejemplos:

> Los bosques nos ofrecen setas, bayas y maravilloso aire fresco, así como **olores**, **sonidos** y **sabores** de una de las últimas regiones vírgenes de Europa. (VISITFINLAND.COM, 2020; palabras en negrita marcadas por nosotros)

> Uno de los motivos para ir, es la laguna La Toreadora, que permite al visitante reflejarse en sus aguas, y **sentir** su movimiento. Los visitantes que saben del poder del agua helada, con cuidado se aventuran a **sentir** el frío y agradecer por la energía que brinda. La laguna de Llaviuco es otro **espectáculo** natural tan cercano de Cuenca. De aguas tranquilas y abundante bosque con miles de texturas y colores. Es el único lugar donde el turista podrá **observar** las diminutas “bromelias” y flores de color verde, lila y blanco. (MINISTERIO DE TURISMO DE ECUADOR, 2020; palabras en negrita marcadas por nosotros)

E incluso se puede convertir en el verdadero núcleo de la actividad turística:

> ‘Dining in the Dark’: una cena navideña a oscuras para poner a prueba tus sentidos. Este restaurante te vendará los ojos y no sabrás qué vas a comer: tendrás que confiar en el gusto y el olfato. […] Imagina cómo sería disfrutar de tu plato favorito con los ojos vendados o en completa oscuridad. ¿Te has parado a pensar en el papel que juega la vista cada vez que comes? (MADRID SECRETO, 2020)

[1] El *script* para llevar a cabo la metodología está disponible en github.com/JasperD-UGent/keyness-calculator.

El papel de los sentidos en la toma de decisiones al comprar productos (o servicios, como suele ser el caso en el sector del turismo) es un fenómeno que ya se ha estudiado extensamente (HOLBROOK y HIRSCHMAN, 1982; KIM y FESENMAIER, 2017). Así, por ejemplo, se sabe que etiquetas sensoriales (por ejemplo « naranjas zumosas » en lugar de « naranjas de Florida ») convencen más a los consumidores de comprar alimentos, lo cual ha llevado a la introducción del término « marketing sensorial » (KRISHNA, 2012). En los estudios sobre el lenguaje turístico, la dimensión sensorial es un enfoque de investigación relativamente reciente (MEACCI y LIBERATORE, 2018), pero los resultados confirman que cuanto más la promoción acentúa o destaca el aspecto sensorial del destino, más atractivo el destino se presenta a los turistas potenciales (AGAPITO, MENDES y VALLE, 2013; KRISHNA, 2012).

En resumen, las descripciones sensoriales son un fenómeno lingüístico cuya relevancia se ha evidenciado desde la literatura, con una línea de investigación reciente y de índole cualitativa que se centra específicamente en su empleo en el lenguaje turístico. El siguiente paso de investigación podría consistir en estudiar el fenómeno a gran escala, a fin de describir su distribución en una multitud de textos, contextos y épocas pasadas y futuras. Se trata, pues, de diseñar una metodología que facilite la identificación empírica de los fenómenos que consideramos esenciales en ciertos tipos de discurso. Presentamos una metodología que ofrece una versión avanzada del análisis de carácter clave, un método relativamente común dentro de la lingüística de corpus (sección 3), además de introducir una dimensión que apenas se incluye en estudios de corpus: un componente de desambiguación semántica automática (sección 4). Es importante enfatizar que este tipo de metodología podría contribuir a dar una respuesta a las dos críticas principales a las que se enfrenta, por ejemplo, el análisis crítico del discurso (ACD): el *cherry-picking*, es decir, la selección arbitraria de textos o de ejemplos para demostrar una hipótesis (STUBBS, 1997); y el conjunto limitado de datos que se puede estudiar, debido a las limitaciones de los análisis manuales (JACOBS y TSCHÖTSCHEL, 2019).

3. CUANTIFICACIÓN DEL CARÁCTER CLAVE: UN ANÁLISIS BASADO EN LA PALABRA

3.1. Observaciones preliminares

3.1.1. Compilación y anotación gramatical de los corpus

Nuestro corpus de estudio (CEst) está compuesto por reportajes publicados en El Viajero, el suplemento turístico del periódico español El País. El corpus de referencia (CRef) es una compilación de cinco subcorpus, que reúnen textos periodísticos de diferentes campos y medios (derechos humanos, economía, sanidad), libros de no ficción (historia), y libros de ficción (literatura juvenil). La Tabla 1 ofrece un resumen detallado de los corpus.

Tipo de corpus	Tipo de textos	Detalles	Número de *tokens* (número de palabras de contenido)
CEst	Turismo	Reportajes El Viajero (El País)	7 589 762 (3 210 212)
CRef	Derechos humanos	Noticias de la ONU (un.org/es/)	502 440 (213 406)
	Economía	Artículos periodísticos en línea (Cinco Días)	10 787 219 (4 399 693)
	Historia	Libros de no ficción recientes (> año 2000), varias fuentes	26 744 645 (11 218 195)

	Literatura juvenil	Libros de ficción recientes (> año 2000), varias fuentes	7 528 422 (3 342 247)
	Sanidad	Artículos periodísticos en línea (El País)	9 088 971 (3 819 674)
			54 651 697 (22 993 215)

Tabla 1.
Resumen de los corpus utilizados

Es importante destacar que los corpus utilizados fueron todos anotados mediante un etiquetado morfosintáctico y proceso de lematización uniformes (siguiendo las pautas del proyecto SCAP; GOETHALS, 2018), a fin de garantizar la comparabilidad de los datos. Sobre todo en el caso del español, con su gran variación morfológica, el proceso de lematización es fundamental para poder llevar a cabo un análisis pormenorizado.

3.1.2. Representación de la dimensión sensorial

Operacionalizamos la identificación de la particularidad discursiva de la dimensión sensorial mediante un repertorio de palabras que se refieren explícitamente a los cinco sentidos, tanto en su forma nominal y verbal (« el gusto » – « gustar »; « el oído » – « oír »; « el olfato » – « oler »; « el tacto » – « tocar »; « la vista » – « ver »). La lista no exhaustiva de 82 lemas, presentada en la Tabla 2 con su categoría gramatical (CG) correspondiente bajo paréntesis (S para sustantivos, V para verbos), es el resultado de una selección manual basada en los sinónimos y palabras relacionadas relevantes que se ofrecen en el diccionario en línea WordReference.

Lema (CG)				
acariciar (V)	espectáculo (S)	husmear (V)	olor (S)	ruido (S)
acechar (V)	espiar (V)	magrear (V)	otear (V)	sabor (S)
aroma (S)	estallido (S)	manosear (V)	paisaje (S)	saborear (V)
audición (S)	estampido (S)	miasma (S)	paladar (S)	sobar (V)
avistar (V)	estrépito (S)	mirar (V)	paladear (V)	sonido (S)
catar (V)	estridencia (S)	observar (V)	palpar (V)	tacto (S)
chasquido	estruendo (S)	oído (S)	panorama (S)	tantear (V)
contemplar (V)	fetidez (S)	oír (V)	panorámica (S)	tentar (V)
crujido (S)	fragancia (S)	ojear (V)	percibir (V)	tiento (S)
cuadro (S)	fragor (S)	oler (V)	perfume (S)	tocar (V)
curiosear (V)	gustar (V)	olfacción (S)	perspectiva (S)	toque (S)
degustar (V)	gustillo (S)	olfatear (V)	peste (S)	tufarada (S)
detonación (S)	gusto (S)	olfateo (S)	pestilencia (S)	tufo (S)
distinguir (V)	hediondez (S)	olfato (S)	probar (V)	ver (V)
divisar (V)	hedor (S)	oliscar (V)	regusto (S)	vista (S)
escuchar (V)	horizonte (S)	olisquear (V)	roce (S)	zumbido (S)
esencia (S)				

Tabla 2.
Lista de lemas seleccionados

De esta lista eliminamos las palabras que no aparecen en nuestro corpus de estudio (« hediondez », « magrear », « olfateo », « oliscar » y « tufarada »), resultando en una selección definitiva de 77 lemas. Llama la atención que los lemas eliminados tienen todos una connotación negativa, lo cual muestra que la presencia de la dimensión sensorial en el

discurso turístico no solo depende del dominio semántico, sino también de la polaridad del lema.

3.2. Métricas de cuantificación: Log Ratio y BIC

Para averiguar si una lista de ítems léxicos es típica para el CEst, se debe aplicar una métrica que calcule la diferencia entre las frecuencias en este corpus y las frecuencias en otros corpus (*keyness analysis*). Tradicionalmente, se han utilizado pruebas estadísticas como log-verosimilitud o de chi cuadrado, que permiten deducir si una eventual diferencia es estadísticamente significativa o no. Sin embargo, se ha criticado el uso de estas pruebas porque no indican a qué se debe exactamente la posible significación estadística (GABRIELATOS y MARCHI, 2011; GRIES, 2010): así, por ejemplo, es posible que una diferencia significativa se deba a que se ha analizado una gran cantidad de datos, aunque en realidad la diferencia es bastante limitada. Por consiguiente, se ha propuesto primero medir la magnitud de la diferencia, una medida que averigua si una diferencia o una relación de frecuencia es débil o fuerte, y luego aplicar una prueba estadística que determina si la diferencia de frecuencia es estadísticamente significativa (GABRIELATOS, 2018).

En este estudio, utilizamos Log Ratio (HARDIE, 2014) para calcular la magnitud de la diferencia:

$$\boldsymbol{Log\ Ratio = log_2 \frac{frecuencia\ normalizada\ CEst}{frecuencia\ normalizada\ CRef}}$$

La métrica es fácil de interpretar: un valor de 0 indica que las frecuencias normalizadas son iguales, y un cambio de 1 en la puntuación indica que la diferencia entre las frecuencias se dobla. Así, por ejemplo, un valor de 3 significa que el ítem es 2^3 o 8 veces más frecuente en el CEst que en el CRef (al revés, un valor de -3 significa que el ítem es 8 veces más frecuente en el CRef que en el CEst).

Para determinar la significación estadística de las diferencias de frecuencia, nos basamos en el Bayesian Information Criterion (BIC), una medida estadística que WILSON (2013) aplica al cálculo del *keyness* a fin de permitir un acercamiento más sutil a los valores *p* (en concreto, es una forma de normalizar los valores *p* y de mejorar así su comparabilidad a través de diferentes comparaciones entre corpus de referencia y corpus de estudio). Si se usa para determinar el carácter clave de elementos léxicos, BIC equivale al valor de log-verosimilitud menos el logaritmo natural de la suma de los dos corpus:

BIC = log-verosimilitud – ln(tamaño del CEst + tamaño del CRef)

Los valores de BIC se deben interpretar en términos de « grado de evidencia » (WILSON, 2013) contra la hipótesis nula, es decir, contra la hipótesis de que la diferencia de frecuencia se debe al azar (véase la Tabla 3).

BIC	Grado de evidencia contra H_0
< 0	Ninguna evidencia para rechazar H_0
0 – 2	Evidencia insignificante contra H_0
2 – 6	Evidencia positiva contra H_0
6 – 10	Fuerte evidencia contra H_0
> 10	Muy fuerte evidencia contra H_0

Tabla 3.
Valores del Bayesian Information Criterion (BIC) y su correspondiente grado de evidencia contra la hipótesis nula

3.3. Metodología

Calculamos los valores de Log Ratio de los 77 « lemas sensoriales » seleccionados (Tabla 2) en el corpus de estudio frente a los 6 corpus de referencia (Tabla 1), obteniendo así una cuantificación de su carácter clave en el discurso turístico. Aplicamos los siguientes parámetros:

1) Ya que la particularidad lingüística que estudiamos está relacionada con el contenido (y no con, por ejemplo, construcciones gramaticales), solo consideramos las palabras de contenido en las calculaciones (véase la Tabla 1 para los totales correspondientes), dejando fuera las palabras funcionales como las preposiciones, artículos, modificadores posesivos y demostrativos, etcétera.

2) Para resolver el problema de frecuencias de 0 en el CRef (que resultan en divisiones por 0 en la fórmula de Log Ratio), aplicamos la técnica de «Laplace *smoothing*», que consiste en añadir 1 a todas las frecuencias en los corpus, y añadir el número de *types* al total de palabras. Si bien es cierto que esta transformación aumenta las frecuencias de forma desigual (*100 + 1* implica un aumento del 1%, mientras que *1 + 1* resulta en un aumento del 100%), Laplace *smoothing* produce las estimaciones más adecuadas para frecuencias de 0 (BRYSBAERT y DIEPENDAELE, 2013), contrariamente a, por ejemplo, aproximarlos por un valor infinitesimal (usualmente 1e-18). Consideremos el ejemplo en la Tabla 4, con el corpus económico como CRef: en el caso de la aproximación infinitesimal, « crujido » presenta un carácter clave muchísimo más alto que « acariciar », mientras que esta discrepancia enorme no se evidencia en las frecuencias absolutas. Al aplicar la transformación Laplace, en cambio, los valores de Log Ratio reflejan más intuitivamente las magnitudes de la diferencia (y, por consiguiente, también su carácter clave).

Lema (CG)	Frecuencia en CEst	Frecuencia en CRef	Log Ratio (Laplace)	Log Ratio (aproximación infinitesimal)
acariciar (V)	117	3	5,33	5,74
crujido (S)	14	0	4,35	64,06

Tabla 4.
Ejemplo de la calculación de Log Ratio con Laplace *smoothing*

3) Por último, en los resultados solo incluimos los lemas cuyo valor de BIC es superior a 2 (véase la Tabla 3), eliminando así los casos donde no hay suficiente evidencia para rechazar la hipótesis nula.

3.4. Resultados

Presentamos los resultados del análisis del carácter clave en la Tabla 5. Comparado con el conjunto del CRef (que está compuesto de 5 subcorpus de tipos de textos distintos), las descripciones sensoriales están alrededor de 3 veces más presentes en el discurso turístico de lo que se podía esperar. No obstante, al aplicar la metodología a los subcorpus individuales, los promedios muestran que existen diferencias considerables entre los tipos de textos: la dimensión sensorial es más clave para el discurso turístico en comparación con las noticias sobre derechos humanos (aproximadamente 24 veces más frecuente), economía (alrededor de 12 veces más frecuente), historia (alrededor de 3 veces más frecuente) y sanidad (más de 8 veces más frecuente), pero esta tendencia no se confirma en la literatura juvenil, donde la puntuación media de Log Ratio de -0,26 implica que en este tipo de textos se hace un uso más frecuente de las descripciones sensoriales que en el lenguaje turístico.

Valor	CRef	Subcorpus				
		Derechos humanos	Economía	Historia	Literatura juvenil	Sanidad
Número de lemas con una diferencia de frecuencia significativa (BIC ≥ 2)	46	29	60	47	53	52
Log Ratio (promedio)	1,47	4,49	3,45	1,4	-0,26	3,14

Tabla 5.
Resultados completos del análisis del carácter clave (frecuencias absolutas)

Además de plantearnos la pregunta de lo que puede significar el concepto de « corpus de referencia », esta última observación también nos lleva a una limitación inherente de la lingüística de corpus, que es la influencia que puede ejercer la composición de los corpus en los resultados/valores finales. Parte de esta influencia está relacionada con la « dispersión » de las instancias de un lema en los subcorpus, sobre todo si estos son de tamaño desigual: de hecho, es posible que una frecuencia de ocurrencia alta de un lema se deba a una frecuencia elevada en un subcorpus de tamaño pequeño, implicando que esa frecuencia absoluta alta en realidad no representa la dispersión del lema en el corpus entero. Una serie de subcorpus de tamaño igual resolvería este problema, pero en la práctica entran en juego tantos parámetros (tipo de texto, período de tiempo, variantes geográficas, etcétera) que la compilación de un corpus de referencia completamente equilibrado se vuelve casi imposible.

Para mitigar el efecto de la dispersión, se deben, pues, convertir las frecuencias absolutas en « frecuencias ajustadas », que reflejan más fielmente cómo están repartidas las instancias de un lema en el corpus. La métrica que utilizamos para este propósito es DP_{norm} (GRIES, 2008; LIJFFIJT y GRIES, 2012), una medida sencilla que evita muchos de los problemas que presentan las medidas de dispersión tradicionales (véase GRIES [2008] para una revisión extensa). DP_{norm} se calcula de la siguiente manera (véase la Tabla 6 para el ejemplo del sustantivo « aroma »):

1) Representar las *n* partes del corpus como porcentajes esperados, que corresponden a las proporciones relativas con respecto al tamaño total del corpus.

2) Representar las frecuencias v_{1-n} con las cuales *a* ocurre en las *n* partes del corpus como porcentajes observados, que corresponden a las proporciones relativas con respecto a la frecuencia total de *a*.

3) Computar, en pares, todas las *n* diferencias absolutas de los porcentajes esperados y observados, sumarlas y dividir el resultado por dos.

4) Normalizar este resultado según la siguiente fórmula, donde DP equivale al valor generado en el paso 3), y min(*s*) equivale a la proporción relativa de la parte del corpus más pequeña:

$$DP_{norm} = \frac{DP}{1 - \min(s)}$$

Parte	Frecuencia absoluta	Porcentaje esperado	Porcentaje observado	Diferencia absoluta	DP_{norm}
Derechos humanos	0	0,009	0	0,005	
Economía	19	0,191	0,034	0,079	
Historia	198	0,488	0,35	0,069	$\frac{0,442}{1 - 0,009} = 0,446$
Literatura juvenil	332	0,145	0,588	0,221	
Sanidad	16	0,166	0,028	0,069	
	565	1	1	0,442 (= DP)	

Tabla 6.
Ejemplo de la calculación de DP_{norm} («aroma»)

El resultado final siempre representa un número entre 0 y 1, con valores cercanos a 0 indicando que *a* está repartido por las *n* partes del corpus como cabría esperar dado el tamaño de las *n* partes, y con valores cercanos a 1 indicando que *a* está repartido por las *n* partes del corpus exactamente de la manera opuesta de lo que cabría esperar dado el tamaño de las *n* partes. Por último, para llegar a las frecuencias ajustadas solo hace falta multiplicar las frecuencias absolutas por $(1 - DP_{norm})$, lo cual lleva al siguiente resultado para el ejemplo de «aroma»:

$$\text{frecuencia ajustada}_{aroma} = \text{frecuencia absoluta}_{aroma} * (1 - DP_{norm})_{aroma} = 565 * 0,554 = 312,79$$

La reducción en la frecuencia se debe principalmente al ajuste de la influencia del subcorpus de literatura juvenil, donde *aroma* ocurre mucho más de lo que cabría esperar (un porcentaje observado de 0,588 frente a un porcentaje esperado de solo 0,145).

Finalmente, al volver a calcular los valores de Log Ratio con las frecuencias ajustadas, llegamos a los resultados presentados en la Tabla 7, que corroboran el carácter clave de los cinco sentidos en el lenguaje turístico. Aunque no aumenta el número de lemas clave, el incremento de 0,19 en el promedio sí parece indicar que la dimensión sensorial es aún más clave de lo que sugería el análisis basado en las frecuencias absolutas. Resumiendo, estos resultados destacan la importancia de poder subdividir el CRef en varias partes, en lugar de tomar un solo corpus extenso como único punto de referencia.

Valor	CRef (frecuencias absolutas)	CRef (frecuencias ajustadas)
Número de lemas con una diferencia de frecuencia significativa (BIC ≥ 2)	46	45
Log Ratio (promedio)	1,47	1,66

Tabla 7.
Resumen de los resultados del análisis del carácter clave (frecuencias absolutas versus ajustadas)

4. EL ANÁLISIS DE CORPUS AUTOMATIZADO DE DATOS DESAMBIGUADOS

4.1. Planteamiento del problema

De lo anterior se desprende que una cuantificación que tiene en cuenta la dispersión de los lexemas en el CRef permite cuantificar con mayor precisión el carácter clave de la particularidad discursiva que representan. Sin embargo, el método aún conlleva una limitación severa, pues no va más allá de la frecuencia del lema, sin distinguir entre los diferentes significados de los lemas polisémicos. Así, por ejemplo, « esencia » tiene un significado no sensorial (« fondo »), y un significado sensorial (« extracto aromático »), pero las calculaciones se realizan en base a la suma de las instancias.

Para comprobar en qué medida la polisemia afecta los resultados del análisis del carácter clave, primero dividimos la selección de lemas en dos grupos, presentados en la Tabla 8. Para la distinción de significados, nos basamos en el diccionario Clave (clave.smdiccionarios.com). Ilustramos los cinco tipos de significaciones que se distinguen (subrayados en la Tabla 8) mediante los significados individuales de « acariciar » y « esencia », pero cabe enfatizar que, en su conjunto, ambos lemas pertenecen al grupo « mixto » (ya que tienen tanto significados sensoriales como no sensoriales).

Grupo	Detalles	Número de lemas
Sensorial	Los lemas que tienen únicamente significados sensoriales, que pueden ser referencias explícitas a una acción sensorial (por ejemplo «acariciar» como « hacer caricias, rozar con la mano »), pero también sensaciones u objetos que se pueden percibir a través de uno de los cinco sentidos (por ejemplo « esencia » como « extracto aromático »).	30
Mixto	Los lemas que tienen por lo menos un significado no sensorial, es decir, un significado sensorial que se usa de forma metafórica (por ejemplo « acariciar » como « ambicionar ») o personificada (por ejemplo «acariciar» como está usado en « las olas acarician la orilla »), o un significado que no presenta ninguna relación con la dimensión sensorial (por ejemplo « esencia » como « fondo»).	47

Tabla 8.
Detalles de la división de los lemas seleccionados en un grupo sensorial y un grupo mixto

A continuación, comparamos los valores de Log Ratio de los lemas « mixtos » (MI) con los de los lemas unívocamente sensoriales (SE) en la Tabla 9. Suponiendo que los significados metafóricos, personificados y no sensoriales puedan resultar en valores más bajos para los lemas mixtos, nuestra expectativa es que los ítems sensoriales producirán valores de Log Ratio más altos. Si se confirma esta hipótesis, es un primer indicio de que la polisemia influye en los valores del carácter clave, y que se necesita una operación de desambiguación para, en nuestro caso, poder excluir determinados significados de las calculaciones.

Comprobamos que este es efectivamente el caso: tanto para el CRef en su conjunto como para los subcorpus individuales, los lemas sensoriales presentan valores de Log Ratio considerablemente más elevados (promedio 3) que los lemas mixtos (promedio 0,99). Por consiguiente, se deben interpretar los resultados de los lemas mixtos con mucha cautela, porque la cuantificación basada en las frecuencias de los lemas no permite conocer el peso de cada significado en los valores finales. A fin de superar esta limitación, la metodología necesita de un componente que entre en la semántica, lo cual exploraremos en el apartado siguiente.

Valor	CRef		Subcorpus: Derechos humanos		Economía		Historia		Literatura juvenil		Sanidad	
	SE	MI	SE	MI	SE	MI	SE	MI	SE	MI	SE	MI
Número de lemas estadísticamente significativos (BIC ≥ 2)	15	30	9	20	21	39	16	31	17	36	19	33
Log Ratio (promedio)	3	0,99	5,6	3,98	4,67	2,79	2,45	0,85	0,24	-0,5	4,31	2,46

Tabla 9.
Resultados completos del análisis del carácter clave (frecuencias ajustadas): lemas sensoriales versus mixtos

4.2. Desambiguación del significado de las palabras

En los últimos años, las herramientas y los modelos basados en el Procesamiento del Lenguaje Natural (NLP) se han vuelto cada vez más accesibles para los profesionales de la lengua. En este estudio, aprovechamos las oportunidades que ofrece este enfoque de NLP para llevar a cabo la tarea de la desambiguación del significado de las palabras (*word sense disambiguation* en inglés, WSD por sus siglas), y adaptar la metodología subyacente a los requisitos específicos de nuestro estudio de caso. La tarea de WSD se puede concebir como una tarea de clasificación, con los significados de los ítems léxicos como las clases, y un algoritmo entrenado para asignar cada instancia de los ítems léxicos a una de esas clases. Cabe destacar que el *set* de clases es diferente para cada ítem léxico, por lo cual en realidad WSD comprende *n* tareas de clasificación distintas, donde *n* es el tamaño del léxico (NAVIGLI, 2009).

El primer paso en desarrollar una metodología de desambiguación semántica consiste en definir los significados que se quieren distinguir en un « inventario de significados » (*sense inventory*; apartado 4.2.1). Luego, se necesita un algoritmo que clasifique instancias de palabras ambiguas según este inventario: en este estudio desarrollamos un algoritmo clasificador basado en el modelo de representación de lenguaje muy reciente e innovador de BERT (apartado 4.2.2). Concretamente, exploramos el desarrollo de una metodología en la cual los significados de lemas ambiguos están representados por frases prototípicas que expresan el significado designado. En su forma más sencilla, esta metodología es de carácter

enteramente « supervisado »: el sistema se basa, pues, únicamente en los datos anotados proporcionados por nosotros (es decir, las frases de ejemplo que sirven como punto de referencia) para adivinar el significado de instancias ambiguas. Sin embargo, las frases prototípicas también se pueden integrar en un enfoque semi-supervisado, donde sirven como « frases de semilla » (*seed sentences*) para añadir de forma automática más frases de ejemplo para cada significado, gracias a lo cual el sistema podrá basarse en un número más elevado de datos anotados para realizar las predicciones.

4.2.1. Inventario de significados

Describimos los lemas « mixtos » en un inventario de significados propio[2], orientado hacia las particularidades de nuestro estudio, es decir, la identificación de « significados sensoriales ». Para las distinciones de significados, nos basamos principalmente en la información ofrecida en el diccionario Clave: este diccionario ofrece distinciones semánticas muy sutiles, que usualmente van acompañadas de una frase de ejemplo. Para cada lema, convertimos esta información en una « ficha semántica », distinguiendo entre diferentes « significados principales » y vinculándolos con uno de los tipos de significados descritos en la Tabla 8.

4.2.2. Algoritmo clasificador

Para el desarrollo de un algoritmo clasificador que asigne significados a frases nuevas, utilizamos las frases de ejemplo incluidas en el inventario como representaciones de los diferentes significados, después de lo cual calculamos la similitud que guardan frases nuevas con cada una de estas representaciones. Para llevar a cabo el cálculo de similitud con las frases de ejemplo, utilizamos el modelo de representación de lenguaje Bidirectional Encoder Representations from Transformers (BERT por sus siglas; DEVLIN *et al.*, 2019). En este estudio, usamos el modelo preentrenado BERT-Base Multilingual Cased, que se entrenó en contenido Wikipedia en 104 diferentes idiomas.

En términos sencillos, este modelo convierte el « texto natural » introducido (que puede ser una frase, un párrafo, un par de pregunta-respuesta hasta un documento entero) en representaciones vectoriales (que también se denominan *word embeddings*), lo cual permite la realización de calculaciones matemáticas con el texto. Los vectores se pueden integrar asimismo en modelos de aprendizaje automático, por lo cual modelos como BERT pueden servir en numerosas tareas de NLP, como pueden ser la detección de sentimientos, la clasificación de documentos o la traducción automática. Sin embargo, a pesar de las posibilidades que ofrecen, estas aplicaciones aún no suelen formar parte de la caja de herramientas de la lingüística de corpus.

La ventaja principal que ofrece BERT para la tarea de la desambiguación semántica es que produce representaciones vectoriales que tienen en cuenta no solo el perfil propio del *token*, sino también el perfil de la información contextual. Esto es, mientras que modelos anteriores como word2vec (MIKOLOV *et al.*, 2013) asignaban el mismo valor vectorial a cada instancia de un *token* (en las frases recogidas en la Tabla 10, « palpa » tendría cuatro veces la misma representación vectorial), los vectores generados mediante BERT producen cuatro vectores diferentes de « palpa ». Evidentemente, la información contextual debería contribuir a la desambiguación de los ítems léxicos polisémicos.

En concreto, utilizaremos los vectores de BERT para el cálculo de la similitud coseno, que determina la similitud entre dos vectores dando como resultado un valor entre 0 (ninguna similitud) y 1 (similitud completa). Para las frases de la Tabla 10, significa que podemos clasificar las frases a la izquierda en base a la semejanza que guardan con las dos clases de significados (que representamos, pues, como una frase de ejemplo prototípica). Los valores

[2] Para el inventario completo, véase github.com/JasperD-UGent/sense-inventory-five-senses.

obtenidos indican que BERT efectivamente asigna el significado correcto a ambas frases. En la Tabla 11 y 12, se incluyen dos ejemplos más de la aplicación de BERT, más en concreto para el sustantivo « panorama » y el verbo « saborear ». Destacan claramente el potencial que brinda el modelo de representación del lenguaje para contribuir a la compleja tarea de la desambiguación semántica.

Clases de significados / Frases por clasificar	Clase 1 « tocar con las manos » El médico palpa el vientre de la embarazada.	Clase 2 « experimentar » En el ambiente se palpa un gran nerviosismo.
Se palpa la frente para comprobar si tiene un agujero.	0,85	0,79
En Barcelona, mientras tanto, se palpa una fuerte atmósfera de tensión.	0,64	0,83

Tabla 10.
Ejemplo de la calculación de similitud coseno con vectores de BERT («palpar»; formas de palabra iguales)

Clases de significados / Frases por clasificar	Clase 1 « degustar » Para saborear la comida hay que comer despacio.	Clase 2 « apreciar, experimentar » Saboreas el triunfo por anticipado.
Arqueé las cejas y saboreé el sándwich mixto.	0,68	0,47
Se saboreaba en el clima posmoderno esta muerte de la modernidad.	0,51	0,52

Tabla 11.
Ejemplo de la calculación de similitud coseno con vectores de BERT (« saborear »; formas de palabra diferentes)

Clases de significados / Frases por clasificar	Clase 1 « paisaje, vista » Desde esta montaña se divisa un hermoso panorama.	Clase 2 « situación » En la conferencia se analizó el panorama actual de la literatura española.
Además, el panorama natural que la rodea es espectacular.	0,79	0,56
En esa carta, entre otras cosas, le narra lo que fue la aparición	0,48	0,71

de Heidegger en el panorama filosófico de su época.	

Tabla 12.
Ejemplo de la calculación de similitud coseno con vectores de BERT («panorama»)

4.3. Metodología

Primero, realizamos para cada lema ambiguo una búsqueda de concordancia en los corpus de SCAP, reuniendo todas las frases que contienen el ítem y dividiéndolas después en un *set* de prueba, que sirve para evaluar el desempeño de nuestro sistema de WSD, y un *set* que utilizaremos para el enfoque semi-supervisado. Como ya se ha mencionado en el apartado 4.2, en un sistema semi-supervisado se utilizan las frases de ejemplo originales como frases de semilla a fin de añadir de forma automática más frases de ejemplo a la representación de cada significado. En la práctica, significa que el algoritmo selecciona de ese *set* de « frases nuevas » los mejores candidatos para cada significado y las considera como frases anotadas adicionales. Para determinar qué frases nuevas son los mejores candidatos, nos basamos en la similitud coseno que presentan con las frases de ejemplo originales, y en la diferencia con el valor de similitud coseno segundo más alto.

Para la evaluación de la metodología, calculamos, tal y como está descrito en el apartado 4.2.2, la similitud coseno entre (la representación vectorial de) la instancia ambigua en las frases del *set* de prueba y cada una de las representaciones de los significados en las frases de ejemplo. El significado asignado por el sistema corresponde al significado con el valor de similitud coseno más alto. En el enfoque enteramente supervisado, el algoritmo se basa únicamente en las frases de ejemplo originales como representación de los significados, mientras que en el enfoque semi-supervisado las frases de ejemplo incluyen, además de las frases originales, también las frases nuevas añadidas de forma automática (25 como máximo).

4.4. Resultados

En la Tabla 13 presentamos los resultados de nuestra metodología de WSD (para « gustillo » y « sobar » no había suficientes datos para construir un *set* de prueba), ordenados por su exactitud en el enfoque supervisado. Las dos últimas columnas permiten comparar estos resultados con los porcentajes para el enfoque semi-supervisado y para el significado más frecuente (MFS por sus siglas en inglés), un *baseline* sencillo pero ambicioso que se utiliza frecuentemente en el dominio del NLP para la evaluación de sistemas de WSD. En general, comprobamos que, con un grado de corrección del 66,6%, la metodología funciona razonablemente bien; en el enfoque semi-supervisado este porcentaje incluso se eleva al 71%. Estos resultados se aproximan al *baseline* del MFS, aunque no lo superan. Además, aún se puede mejorar la consistencia del sistema, ya que también en el caso semi-supervisado todavía hay unos diez lemas que no consiguen una exactitud del 50%.

Sin embargo, al considerar en detalle los resultados individuales por lema sí se destaca claramente el potencial que brinda la metodología: en el caso de « palpar », « panorama », « contemplar », « horizonte » y « saborear », por ejemplo, nuestro sistema basado en frases de ejemplo alcanza, con una sola frase de ejemplo por significado como *input*, un grado de corrección que oscila entre el 88,7% y el 97,2% en el enfoque semi-supervisado, correspondiendo a aumentos de entre 27,3 hasta 36,1 puntos frente a la frecuencia del significado más común.

Lema (CG)	Número de significados	Número de frases en *set* de prueba	Supervisado	Semi-supervisado	MFS
gustar (V)	2	99	100,0%	100,0%	100,0%
palpar (V)	2	36	94,4%	97,2%	61,1%
ruido (S)	2	95	93,7%	85,3%	96,8%
oír (V)	2	97	92,8%	83,5%	99,0%
panorama (S)	2	88	90,9%	94,3%	67,0%
mirar (V)	3	96	89,6%	79,2%	92,7%
oído (S)	2	94	89,4%	91,5%	89,4%
escuchar (V)	2	91	89,0%	73,6%	92,3%
contemplar (V)	2	96	84,4%	92,7%	62,5%
horizonte (S)	2	95	83,2%	92,6%	60,0%
vista (S)	4	88	83,0%	79,5%	46,6%
estallido (S)	2	70	82,9%	80,0%	81,4%
sentir (V)	3	96	80,2%	77,1%	78,1%
saborear (V)	2	53	75,5%	88,7%	54,7%
curiosear (V)	2	16	75,0%	56,2%	93,8%
manosear (V)	2	4	75,0%	75,0%	75,0%
tacto (S)	2	44	75,0%	75,0%	77,3%
tiento (S)	2	4	75,0%	100,0%	100,0%
ver (V)	4	97	74,2%	69,1%	68,0%
tantear (V)	2	30	70,0%	80,0%	66,7%
paladear (V)	2	13	69,2%	69,2%	76,9%
gusto (S)	4	97	69,1%	68,0%	56,7%
olfatear (V)	2	6	66,7%	66,7%	100,0%
olfato (S)	2	18	66,7%	88,9%	55,6%
probar (V)	3	88	64,8%	72,7%	39,8%
esencia (S)	2	98	64,3%	74,5%	93,9%
perspectiva (S)	3	95	64,2%	76,8%	48,4%
roce (S)	3	33	63,6%	69,7%	60,6%
observar (V)	3	93	53,8%	45,2%	81,7%
toque (S)	6	89	51,7%	64,0%	59,6%
olisquear (V)	2	4	50,0%	75,0%	75,0%
percibir (V)	3	87	49,4%	75,9%	57,5%
distinguir (V)	4	96	43,8%	43,8%	40,6%
acariciar (V)	3	97	43,3%	74,2%	89,7%
sabor (S)	2	94	42,6%	63,8%	71,3%
tentar (V)	3	30	40,3%	43,3%	83,3%
husmear (V)	2	10	40,0%	50,0%	60,0%
audición (S)	3	9	33,3%	77,8%	44,4%
tufo (S)	3	6	33,3%	16,7%	33,3%
oler (V)	3	90	31,1%	45,6%	58,9%
peste (S)	4	49	28,6%	40,8%	81,6%
cuadro (S)	7	82	24,4%	28,0%	64,6%
tocar (V)	8	90	24,4%	33,3%	38,9%
catar (V)	2	14	21,4%	28,6%	92,9%
paladar (S)	3	32	15,6%	25,0%	84,4%

	2 809	66,6%	71%	71%

Tabla 13.
Resultados completos de desambiguación semántica automática (significados originales)

Ya que este estudio se centra específicamente en la dimensión sensorial, también hemos llevado a cabo un experimento en el cual solo distinguimos dos clases de significados: una clase que incluye todas las significaciones relacionadas explícitamente a los cinco sentidos, y otra clase en la cual se agrupan todos los otros significados, sean metafóricos, personificados o no sensoriales. En el caso de « gusto », por ejemplo, se juntan las significaciones « percepción de sabores » y « sabor en la boca » bajo una « clase sensorial », mientras que las significaciones « placer, voluntad » y « estilo » junto forman la segunda clase. Presentamos los resultados de esta nueva distinción de significados, que por supuesto solo afecta a los lemas con más de dos significados, en la Tabla 14 abajo. Comprobamos que el porcentaje del enfoque supervisado sube del 66,6% al 72,9%, y que la exactitud del enfoque semi-supervisado alcanza el 80% (frente al 71% antes). Además, es importante destacar que este último resultado implica que ahora sí se supera el *baseline* del significado más frecuente.

Lema (CG)	Supervisado	Semi-supervisado	MFS
gustar (V)	100,0%	100,0%	100,0%
palpar (V)	94,4%	97,2%	61,1%
ruido (S)	93,7%	85,3%	96,8%
gusto (S)	92,8%	93,8%	83,5%
oír (V)	92,8%	83,5%	99,0%
panorama (S)	90,9%	94,3%	67,0%
vista (S)	89,8%	94,3%	53,4%
mirar (V)	89,6%	92,7%	92,7%
oído (S)	89,4%	91,5%	89,4%
escuchar (V)	89,0%	73,6%	92,3%
sentir (V)	85,4%	78,1%	82,3%
contemplar (V)	84,4%	92,7%	62,5%
horizonte (S)	83,2%	92,6%	60,0%
estallido (S)	82,9%	80,0%	81,4%
probar (V)	79,5%	89,8%	71,6%
perspectiva (S)	78,9%	84,2%	78,9%
audición (S)	77,8%	100,0%	77,8%
oler (V)	76,7%	72,2%	85,6%
roce (S)	75,8%	87,9%	60,6%
saborear (V)	75,5%	88,7%	54,7%
curiosear (V)	75,0%	56,2%	93,8%
manosear (V)	75,0%	75,0%	75,0%
tacto (S)	75,0%	75,0%	77,3%
tiento (S)	75,0%	100,0%	100,0%
tentar (V)	73,3%	53,3%	100,0%
toque (S)	73,0%	84,3%	79,8%
tantear (V)	70,0%	80,0%	66,7%
tocar (V)	70,0%	72,2%	70,0%
paladear (V)	69,2%	69,2%	76,9%
olfatear (V)	66,7%	66,7%	100,0%

olfato (S)	66,7%	88,9%	55,6%
esencia (S)	64,3%	74,5%	93,9%
distinguir (V)	59,4%	59,4%	59,4%
observar (V)	58,1%	90,3%	81,7%
percibir (V)	50,6%	78,2%	57,5%
olisquear (V)	50,0%	75,0%	75,0%
ver (V)	49,5%	58,8%	68,0%
sabor (S)	42,6%	63,8%	71,3%
husmear (V)	40,0%	50,0%	60,0%
peste (S)	38,8%	44,9%	89,8%
cuadro (S)	37,8%	82,9%	64,6%
tufo (S)	33,3%	50,0%	66,7%
acariciar (V)	30,9%	63,9%	89,7%
catar (V)	21,4%	28,6%	92,9%
paladar (S)	15,6%	31,2%	84,4%
	72,9%	80%	77,9%

Tabla 14.
Resultados completos de desambiguación semántica automática (significados sensoriales vs. otros)

Por último, la Tabla 15 revela el valor añadido de incluir un componente de desambiguación semántica en la cuantificación del carácter clave mediante la metodología elaborada en la sección 3: la recalculación de los valores de Log Ratio de « contemplar » y « horizonte », para los cuales nuestro sistema de WSD alcanzó altos grados de corrección (> 92%), indica que para estos lemas la dimensión sensorial es casi dos veces más clave de lo que sugería el análisis original (un aumento de 0,84 para « contemplar » y 0,92 para « horizonte »). Es importante destacar que, aunque el desempeño general del sistema es más que satisfactorio, el grado de corrección bajo para algunos lemas nos ha llevado a la decisión de no recalcular el carácter clave de la lista de lemas entera, sino de limitarnos a ilustrarlo mediante dos ejemplos concretos de cuyos resultados estamos seguros que son fiables.

Valor	contemplar (1+2)	contemplar (1)	horizonte (1+2)	horizonte (1)
Frecuencia absoluta CEst	1973	1713	657	561
Frecuencia ajustada CEst	1973	1713	657	561
Frecuencia absoluta CRef	5898	3664	1701	951
Frecuencia ajustada CRef	4967,69	2401,5	1483,13	667,59
Log Ratio	1,07	1,91	1,23	2,15
BIC	671,41	1524	273,83	584

Tabla 15.
Recalculación del valor de Log Ratio después de la desambiguación semántica automática para « contemplar » (1: « fijar la vista »; 2: « considerar ») y « horizonte » (1: « línea límite de la superficie terrestre »; 2: « conjunto de perspectivas »)

5. DISCUSIÓN

En lo que precede hemos diseñado una metodología que permite estudiar a gran escala fenómenos lingüísticos en lenguas de especialidad, a fin de complementar los datos de investigación cualitativos con datos empíricos y cuantitativos. En el capítulo 3, hemos explicado en detalle la aplicación de un análisis del carácter clave basado en frecuencias de corpus, que genera valores numéricos fáciles de interpretar. Los resultados muestran que el análisis en su forma básica presenta la limitación de no tener en cuenta la dispersión de los lemas en los corpus, y que al basar las calculaciones en las frecuencias ajustadas se puede llegar a valores de carácter clave más informativos.

No obstante, también esta calculación adaptada es susceptible de mejoras, ya que, sean ajustadas o no, las frecuencias utilizadas siguen estando basadas en la mera suma de las ocurrencias del lema en los corpus, sin distinguir entre los diferentes significados de lemas polisémicos. A fin de superar esta limitación, en el capítulo 4 hemos explorado la incorporación de un componente de desambiguación semántica automática. Los grados de corrección obtenidos destacan el potencial que brinda el sistema de desambiguación automática, pero al mismo tiempo revelan que los resultados son todavía relativamente dispares, y que, para implementar el sistema en futuros proyectos de codificación de corpus, se debería aumentar su consistencia. Se trata de un problema muy específico que requiere más investigación, por ejemplo para desarrollar un sistema de alerta que avisa al investigador de posibles problemas en la codificación automática.

Por último, cabe destacar que la metodología tal y como está diseñada ahora asume que es posible captar tipos de discursos (como por ejemplo « el lenguaje turístico ») en conjuntos de textos cerrados, y que se puede representar el lenguaje general mediante un gran corpus de referencia. En futuros estudios, queremos ahondar en este acercamiento estático de corpus, y estudiar si una conceptualización más flexible y dinámica de las características de corpus de estudio y corpus de referencia podría beneficiar a la metodología y, de esta manera, mitigar aún más el efecto que pueden tener la composición y el tamaño de los corpus utilizados.

6. CONCLUSIÓN

Hemos presentado una discusión pormenorizada de varios instrumentos metodológicos para avanzar en el cómputo del carácter clave (*keyness*) de ítems léxicos, que es una técnica común en el análisis de corpus, pero que, a nuestro parecer, puede perfeccionarse al hacer uso de las últimas evoluciones en el análisis estadístico y en el campo del procesamiento automatizado de la lengua. Al aplicar nuestra metodología a las referencias textuales a los cinco sentidos, una técnica retórica frecuente en el discurso turístico, hemos corroborado la importancia de este fenómeno desde una perspectiva cuantitativa, ya que los resultados del análisis muestran que la dimensión sensorial es más de tres veces más clave (valor de Log Ratio de 1,66) para el lenguaje turístico de lo que se podía esperar en general.

Sin embargo, el aspecto más innovador de este trabajo está en la introducción de un componente de desambiguación semántica automática en la metodología. Hemos explorado el desarrollo de un sistema de WSD basado en frases de ejemplo, que, con unas pocas frases prototípicas como *input*, alcanza un grado de corrección que ronda el 70%. Al ajustar la distinción de significados a la medida de nuestro estudio de caso (es decir, los significados sensoriales frente a los otros significados), la exactitud de la metodología incluso se eleva al 80%. El gran valor añadido de esta operación de desambiguación se ha demostrado mediante la recalculación de los valores Log Ratio de los lemas « contemplar » y « horizonte » (para los cuales el sistema de WSD alcanzó un grado de corrección superior al 92%): en realidad, el carácter clave de su significación sensorial era casi dos veces más alto de lo que indicaba el

análisis original sin desambiguación semántica. No obstante, antes de poder integrar el componente de desambiguación en futuros proyectos se tendrá que estudiar más en detalle el problema específico de las discrepancias entre los resultados por lema.

7. AGRADECIMIENTOS

Para la elaboración del inventario de significados, nos hemos basado en el Diccionario Clave, a cuyos contenidos tenemos acceso gracias a una colaboración de investigación con la Fundación Santa María (SM).

8. BIBLIOGRAFÍA

AGAPITO, D., MENDES, J. & VALLE, P. 2013. « Exploring the conceptualization of the sensory dimension of tourist experiences ». *Journal of Destination Marketing and Management*, *2/2*: 62-73.

BRYSBAERT, M. & DIEPENDAELE, K. 2013. « Dealing with zero word frequencies: A review of the existing rules of thumb and a suggestion for an evidence-based choice ». *Behavior Research Methods*, *45/2*: 422-430.

DANN, G. M. S. 1996. *The Language of Tourism: A Sociolinguistic Perspective*. Wallingford: CAB International, 298.

DEVLIN, J., CHANG, M. W., LEE, K. & TOUTANOVA, K. 2019. « BERT: Pre-training of deep bidirectional transformers for language understanding ». *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1(Mlm)*: 4171-4186.

DURÁN-MUÑOZ, I. 2019. « Adjectives and their keyness: A corpus-based analysis of tourism discourse in English ». *Corpora*, *14/3*: 351-378.

GABRIELATOS, C. 2018. « Keyness Analysis: nature, metrics and techniques ». In TAYLOR, C. & MARCHI, A. (dirs.). *Corpus Approaches to Discourse: A Critical Review*. Oxford: Routledge, 224-258.

GABRIELATOS, C. & MARCHI, A. 2011. « Keyness Matching metrics to definitions ». Corpus Linguistics in the South <https://eprints.lancs.ac.uk/id/eprint/51449/> (consultado el 24/03/2021).

GOETHALS, P. 2018. « Customizing vocabulary learning for advanced learners of Spanish ». In READ, T., SEDANO CUEVAS, B. & MONTANER-VILLALBA, S. (dirs.). *Technological innovation for specialized linguistic domains : languages for digital lives and cultures, proceedings of TISLID"18*. Mauritius: Éditions Universitaires Européennes, 229-240.

GOETHALS, P. & SEGERS, L. 2016. « El uso de los adjetivos en los folletos de turespaña y en la guía de viajes 2.0 minube: un análisis de corpus ». In LÓPEZ SANTIAGO, M. & GIMÉNEZ FOLQUÉ (dirs.). *El léxico del discurso turístico 2.0*. Valencia: Universitat de València, 117-152.

GRIES, S. T. 2008. « Dispersions and adjusted frequencies in corpora *». International Journal of Corpus Linguistics*, *13/4*: 403-437.

GRIES, S. T. 2010. « Useful statistics for corpus linguistics ». In SÁNCHEZ PÉREZ, A. & ALMELA SÁNCHEZ, M. (dirs.). *A Mosaic of Corpus Linguistics: Selected Approaches*. Frankfurt: Peter Lang, 269-291.

HARDIE, A. 2014. « Log Ratio - an informal introduction ». <http://cass.lancs.ac.uk/log-ratio-an-informal-introduction/> (consultado el 25/11/2020).

HOLBROOK, M. B. & HIRSCHMAN, E. C. 1982. « The Experiential Aspects of Consumption: Consumer Fantasies, Feelings, and Fun ». *Journal of Consumer Research*, *8/2*: 132-140.

JACOBS, T. & TSCHÖTSCHEL, R. 2019. « Topic models meet discourse analysis: a quantitative tool for a qualitative approach ». *International Journal of Social Research Methodology*, *22/5*: 469-485.

JAWORSKA, S. 2017. « Metaphors We Travel by: A Corpus-Assisted Study of Metaphors in Promotional Tourism Discourse ». *Metaphor and Symbol*, *32/3*: 161-177.

KIM, J. & FESENMAIER, D. R. 2017. « Measuring Human Senses and the Touristic Experience: Methods and Applications ». In XIANG, Z. & FESENMAIER, D. R. (dirs.). *Analytics in Smart Tourism Design*. Springer, 47-63.

KRISHNA, A. 2012. « An integrative review of sensory marketing: Engaging the senses to affect perception, judgment and behaviour ». *Journal of Consumer Psychology*, *22/3*: 332-351.

LIJFFIJT, J. & GRIES, S. T. 2012. « Review of ((2008)): International Journal of Corpus Linguistics ». *International Journal of Corpus Linguistics*, *17/1*: 147-149.

MADRID SECRETO. 2020. « 'Dining in the Dark': una cena navideña a oscuras para poner a prueba tus sentidos ». <https://madridsecreto.co/dining-in-the-dark-madrid/> (consultado el 07/12/2020).

MANCA, E. 2008. « From phraseology to culture ». *International Journal of Corpus Linguistics*, *13/3*: 368-385.

MEACCI, L. & LIBERATORE, G. 2018. « A senses-based model for experiential tourism ». *Tourism & Management Studies, 14/4*: 7-14.

MIKOLOV, T., CHEN, K., CORRADO, G. & DEAN, J. 2013. « Efficient estimation of word representations in vector space ». *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*: 1-12.

MINISTERIO DE TURISMO DE ECUADOR. 2020. « El Parque Nacional Cajas y su magia natural ». <https://www.turismo.gob.ec/el-parque-nacional-cajas-y-su-magia-natural/%0D> (consultado el 07/12/2020).

NAVIGLI, R. 2009. « Word sense disambiguation: A survey ». *ACM Computing Surveys*, *41/2*.

SALIM, M. A. B., IBRAHIM, N. A. B. & HASSAN, H. 2012. « Language for Tourism: A Review of Literature ». *Procedia - Social and Behavioral Sciences*, *66*: 136-143.

STUBBS, M. 1997. « Whorf's children: Critical comments on critical discourse analysis (CDA) ». *British Studies in Applied Linguistics*, *12*: 100-116.

VISITFINLAND.COM. 2020. « 21 razones para enamorarse de Finlandia ». <https://www.visitfinland.com/es/articulo/grandes-cosas-sobre-finlandia/> (consultado el 07/12/2020).

WILSON, A. (2013). « Embracing Bayes factors for key item analysis in corpus linguistics ». In BIESWANGER, M. & KOLL-STOBBE, A. (dirs.). *New Approaches to the Study of Linguistic Variability*. Peter Lang, 3-11.