Significado, distribución y frecuencia en la categoría preposicional. Una aproximación computacional

Meaning, distribution and frequency of the prepositional category in Spanish. A computational approach

Francesc Reina González
Doctorando de Ciencia Cognitiva y Lenguaje. Departamento de Lingüística de la Universidad de Barcelona **frareina@hotmail.com**

Resumen: Desde las primeras definiciones de la teoría lingüística sobre la preposición se ha considerado una categoría lingüística controvertida. El origen de esa discusión procede de la dificultad para explicar, simultáneamente, su supuesta nuclearidad sintáctica y su valor léxico-semántico. Los utensilios matemáticos que proceden del procesamiento del lenguaje natural, de la lingüística del corpus, así como los algorítmicos del aprendizaje automático están permitiendo acercarse al significado preposicional con resultados muy reveladores, desplazando el debate categorial. Mi investigación sugiere que el significado preposicional pueda ser gradual de manera que su distribución en las secuencias sintácticas será determinante en su caracterización.
Palabras claves: significado preposicional, análisis computacional, semántica distribucional, hipótesis gradual, aprendizaje automático, entropía semántica.

Abstract: Since the first definitions of the linguistic theory of the preposition has been considered a controversial linguistic category. The origin of this discussion comes from the difficulty to explain, simultaneously, its supposed syntactic nuclearity and its lexical-semantic value. The mathematical tools that come from the processing of natural language, corpus linguistics, as well as the algorithms of machine learning are allowing us to approach the prepositional meaning with very revealing results, displacing the categorical debate. My research suggests that prepositional meaning can be gradual so that its distribution in the syntactic sequences will be decisive in its characterization.
Keywords: prepositional meaning, computational analysis, distributional semantics, gradual hypothesis, automatic learning, semantic entropy.

1. La naturaleza de las preposiciones: límites y dificultades

La preposición ha sido, tradicionalmente, una categoría lingüística controvertida desde su conducta sintáctica hasta su valor léxico-semántico. Estos conflictos descriptivos han estimulado aproximaciones teóricas muy apriorísticas en la explicación de su comportamiento dentro del español y en lenguas de otras familias tipológicas (con presencia de caso morfológico y preposiciones, o con posposiciones). Ambas circunstancias, cierta imposición empírico-teórica y las dificultades descriptivas propician, según creo, el desplazamiento del debate categorial y la búsqueda desde otras perspectivas más esclarecedoras en sus predicciones empíricas y teóricas.

2. De la teoría lingüística a las aproximaciones computacionales

Lloret, E.; Saquete, E.; Martínez-Barco, P.; Moreno, I. (eds.) *Proceedings of the Doctoral Symposium of the XXXIV International Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*, p. 1–5 Sevilla, Spain, September 19th 2018.

Hace ya más de medio siglo que desde diferentes perspectivas de la teoría lingüística, la gramática generativista o la gramática cognitiva, la preposición ha supuesto muchos retos de distinta condición empírica y estilo analítico. Los trabajos categoriales pioneros de (Chomsky 1970, 1981), (Jackendoff 1977) o (Van Riemsdijk 1978) incluyen la preposición entre las categorías léxicas mayores con los rasgos [-N -V]. Además, la distinción empírico-conceptual, entre categorías léxicas y funcionales de finales de la década de los años 80 del siglo XX, permitió abrir una línea de investigación en la cartografía sintáctica y los grados de interpretación en piezas lingüísticas como la preposición, ya que sabemos que comparte rasgos de ambas.

Sendas caracterizaciones binarias de la preposición [-N –V] o [± F ± L] han estado sometida, igualmente, a discusiones de diversa consideración. Y, de hecho, todavía sigue viva. Quizás una de las propuestas más radicales, al negarle su lugar entre las categorías léxicas, fue la de (Mark C. Baker 2003).

Mientras tanto, el progreso en los procedimientos y las heurísticas computacionales (estadísticos o algorítmicos) han permitido investigar el papel de esta clase de partículas con otros propósitos y con otras consecuencias explicativas. Se pueden leer una multitud de monografías sobre la conformación de esos recursos, desde (Grishman 1986) hasta (Clark, Fox y Lappin 2013), pasando por los clásicos de (Charnik 1993) y el didáctico compendio de (Manning y Schütze (1999). Además se han desarrollado proyectos aplicados globales en el tratamiento computacional de esta categoría. Una de la más exhaustivas para el inglés fue *The preposition Project* de Litowski (http://www.clres.com/prepositions.html), cuyo objetivo fue la desambiguación de la semántica preposicional (un problema esencial en la traducción automática, por ejemplo).

Asimismo, desde los años noventa del siglo XX, se han acumulado evidencias descriptivas nuevas, no solo en la teoría lingüística más formalista, sino en otras que se asientan en cuestiones y hechos semánticos. Sería el caso de (Zelinsky-Wibblet 1993), (Saint-Dizier 2006), o las compilaciones de (Kurzon & Adler 2008) y (Hagège 2010) para cualquier tipo de adposición (pre- o post- posicional). Una de las mejores síntesis del estado del arte en el ámbito computacional fue el monográfico de (Baldwin, Timothy; Kordoni; Valia y Villavicencio, Aline 2009) recogido en la influyente revista *Computational Linguistics*. Allí se despliega un abanico de cuestiones sobre cómo procesa esta clase de partículas, sobre el especial significado que aporta, sobre el valor de su frecuencia en los corpus y sobre la diversidad de aplicaciones que se ven afectadas por la comprensión de esta clase de palabras.

Por último, con la progresiva construcción de corpus lingüísticos en diferentes lenguas, incluso antes del actual desarrollo de la minería de datos o Big Data, se han venido multiplicando las generalizaciones sobre la distribución sintáctica preposicional así como de los valores semánticos de las preposiciones en multitud de lenguas. En ese sentido, tanto la nueva semántica formal y distribucional, véase (Boleda, G. and A. Herbelot 2016), como las herramientas de aprendizaje automático (conocidas como Machine Learning), véase (Mikolov 2014), están favoreciendo un acercamiento más afinado a la multitud de datos heterogéneos que suelen conformar la preposición.

3. Frecuencia, distribución y significado de las preposiciones. La hipótesis de la gradualidad semántica

La perspectiva metodológica se enmarca en una visión empírica y computacional respecto de los recursos de observación, predicción, medición y generalización de los hechos lingüísticos. Así, y a partir de

tres conceptos recurrentes en el procesamiento del lenguaje natural, como la frecuencia de piezas léxicas, su distribución y el significado propongo una hipótesis general para la semántica que subyace a las preposiciones en español.

El enunciado hipotético es el siguiente: *los valores semánticos de las preposiciones del español pueden ordenarse de manera gradual que alcanza desde la funcionalidad completa hasta la lexicidad. En esta secuencia, y según la diversidad de construcciones, sintagmas y contextos, podemos identificar y medir ese valor a partir de la distribución ontosemántica de los SSNN u otras clases de sintagmas que coocurran.*

La gradualidad oscila entre los valores y los usos más funcionales o vacíos como las preposiciones que conforman las locuciones prepositivas, la *a* del CD/CI o el *por* del complemento agente en las oraciones en voz pasiva, y los valores y los usos locativos *de a, de, en, por, hacia, sobre o hasta*, los temporales de *a, entre, bajo, desde, durante o tras*, o los nocionales de *mediante, por* y *para*, considerados léxicos o plenos.

El desarrollo de esta hipótesis se articula en torno a seis objetivos que se relacionan a continuación.

a). Presentar los elementos historiográficos, descriptivos y teóricos de la categoría gramatical a lo largo de los hitos más valiosos de la tradición y de los modelos de investigación lingüística más recientes.

b). Caracterizar las limitaciones de las propuestas que se han ido aportando en el análisis de casos concretos para la preposición en español y en otras lenguas.

c). Medir la frecuencia de similitudes semánticas de las piezas preposicionales, en algunas estructuras sintáctico-argumentales con el uso de herramientas de algoritmos probabilísticos del campo del aprendizaje automático para el procesamiento del lenguaje natural en corpus lingüísticos (al estilo de redes neurales, como Word2vec o Gcluto). (Experimentos 1 y 3).

d). Observar, identificar y analizar posibles correlaciones semánticas a partir de valores como la entropía en corpus del español y a través de clasificaciones ontosemánticas estandarizadas procedentes de la red ontosemántica de Wordnet (experimento 2). Para su funcionamiento y organización se puede leer (Fellbaum 2006).

e). Proponer, a la luz de los resultados anteriores, un eje-espacio gradual de valores-piezas preposicionales desde la funcionalidad absoluta (marcadoras de caso, por ejemplo) hasta el significado más cuantificable (valores locativo-espaciales, temporales o nocionales). Ese eje tendría tres zonas continuas de significado: funcional, semifuncional y léxico.

f). Explorar las posibilidades interlingüísticas de la propuesta (con lenguas que admitan caso, preposiciones o posposiciones) y comprobar su capacidad generalizadora.

4. Trabajo metodológico y experimentos

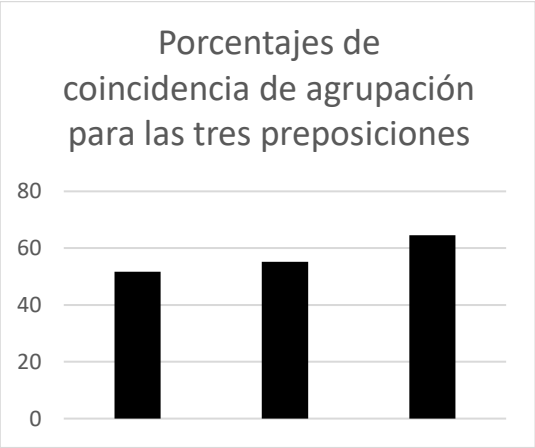
Se están realizando tres experimentos diferentes con objeto de verificar empíricamente las diferentes fases y/o grados semánticos de las preposiciones, tal y como se estipula en la hipótesis.

El primero de ellos ha analizado la gradualidad en el caso de tres preposiciones: *a, hacia* y *hasta* en los contextos sintácticos para 90 verbos de movimiento del español. A través de la herramienta CLUTO (agrupación por Word Embedding), véase (Karypis 2003) para su funcionamiento, hemos procedido a realizar agrupaciones (clustering) de 71.000 sintagmas preposicionales (SSPP) disponibles en los corpus WikiCorpus, Ancora y Semsem, en 3, 4 y 5 grupos. Estos recursos de "agrupación" se fundan en el "Word embedding". El objetivo de este procedimiento es cuantificar y categorizar propiedades semánticas entre elementos lingüísticos a partir de los contextos donde coocurren y que se representarán en vectores. Estos modelos de espacio

vectorial representan ("embed", incrustan) palabras en un espacio vectorial continuo en el que palabras semánticamente similares se asignan a puntos cercanos.

Los resultados han sido muy satisfactorios en la medida en que porcentualmente se verifican y se confirman las agrupaciones realizadas por dos anotadores humanos. El aprendizaje automático acredita la predicción humana en la asignación de significados. Nos encontramos con un 51,65 % para la preposición *a*, un 55,17 para la preposición *hacia* y un 64,6 para la preposición *hasta*.

Gráfico 1



A, HACIA y HASTA

En el segundo experimento hemos utilizado el concepto de entropía (H) de (Claude E. Shannon 1948), procedente de la teoría de la información, para medir la clase de significado que contienen los SSPP en los verbos de régimen en comparación con otras clases de verbos. Hemos seleccionado un total de 140 verbos, de los cuales 48 poseen SSPP considerados de régimen verbal (obligatorios argumentalmente con una determinada preposición). Hemos realizado una búsqueda de SSPP en el Corpes anotado de la RAE que se ha cruzado con el Wikicorpus, es decir, se ha comprobado la coincidencia entre ambos repertorios. Y por último, los hemos clasificado en tres tipos: (i) los funcionales, cuya preposición es la *a* (ii) los semifuncionales, complementos de régimen verbal con *de*, *en* o *con*, y (iii) los léxicos, complementos adjuntos o circunstanciales, con preposiciones como *de*, *en*, *con*, *sin*, *a*.

Una vez obtenidos los ficheros con los SSPP coincidentes para cada verbo se procede a su asignación semántica según las siguientes clases: humano, entidad abstracta, locativo, temporal, evento o actividad, objeto o artefacto, y modalidad (elegidas, por su importancia representativa, de la red ontosemántica de Wordnet). Los datos obtenidos se introducen en una tabla con la función logarítmica de la entropía que se aplica a los tres grupos, funcionales (F), semifuncionales (SF) y léxicos (L).

Las cifras indican que la entropía, es decir, el grado de azar o desorden de esos subgrupos es el más bajo en los *funcionales* (1,568), el más alto en los *léxicos* (2,512), e intermedio en los *semifuncionales* (2,321), donde se encuentran los verbos de régimen. La predicción coincide con la descripción gramatical: cuánta más restricción argumental tiene el verbo más previsible es el significado del SN seleccionado. La preposición, por tanto, sigue la gradación semántica y la medida entrópica está justificada como recurso explicativo.

Tabla 1

TIPOS DE SSPP	**F**	**SF**	**L**
Humanos	287	86	22
Abstractos	33	273	58
Locativos	42	35	128
Temporales	12	23	33
Eventos	7	50	47
Objetos	5	103	20
Modales	26	41	102
TOTAL	**412**	**611**	**410**
Entropía (H)	**1,568**	**2,321**	**2,512**

El tercer experimento está en curso y en fase explorativa. El propósito es intentar que parte del reconocimiento entrópico del segundo experimento se pueda producir de manera automática gracias a un parset, es decir, que los tres grupos de SSPP emerjan

sin la necesidad de la intervención humana, o con un grado de reconocimiento representativo.

Al final de la investigación, procederemos a elegir algunas lenguas con caso y preposición (alemán, ruso o polaco), posposiciones (vasco, chino, húngaro o hindi), o con muy pocas preposiciones (igbo, hablada en Nigeria) y someterlas al contraste experimental que hemos realizado con el español.

Agradecimientos

A mis directores de tesis la Dra. Irene Castellón y el Dr. Lluís Padró por sus ideas, sugerencias y orientaciones; y al Dr. Horacio Rodríguez por sus comentarios y su generosidad durante el Simposio.

Bibliografía

Baker, Mark C. 2003. *Lexical Categories. Verbs, nouns and adjectives*, Cambridge University Press, Cambridge.

Baldwin, Timothy; Kordoni, Valia y Villavicencio, Aline 2009. *Prepositions in Applications: A Survey and Introduction to the Special Issue*, Computational Linguistics, 35(2), páginas 119-149.

Boleda, Gemma y Herbelot, Aurélie 2017. *Formal Distributional Semantics: Introduction to Special Issue*, Computational Linguistics, 42(4), páginas 619-635.

Charniak, Eugene 1993. *Statistical Language Learning*, The MIT Press, Cambridge.

Chomsky, Noam. 1970. *Remarks on Nominalization*. En Jacobs, Roderick A. and Rosenbaum, Peter S. (eds.), *Readings in English Transformational Grammar*, páginas184-221. Ginn, Boston.

Chomsky, Noam 1981. *Lectures on Government and Binding*, Foris, Dordrecht.

Clark, Alexander; Fox, Chris y Lappin, Shalom 2013. *The Handbook of Computational Linguistics and Natural Language Processing*, Wiley-Blackwell, Oxford.

Fellbaum, C. 2006. *WordNet(s.* En Keith Brown, (Ed.) *Encyclopedia of Language & Linguistics*, 2da. edición, Vol. 13, páginas 665-670, Elsevier, Oxford.

Grishman, Ralhp 1986. *Computational Linguistics. An introduction*, Cambridge University Press, Cambridge.

Hagège, Claude 2010. *Adpositions*, Oxford University Press, Oxford.

Jackendoff, Ray 1977. *X' Syntax: A Study of Phrase Structure*, The MIT Press, Cambridge.

Karypis, George 2003. *CLUTO. A clustering toolkit*, University of Minnesota, Technical Report, 02-017.

Kurzon, Dennis & Adler, Silvia 2008. *Adpositions. Pragmatic, semantic and syntactic perspectives*, John Benjamins, Amsterdam.

Litkowski, Ken y O. Hargraves, O. 2005. *The Preposition Project*. En ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications, páginas 171–179, Colchester.

Mannig, Christopher D. y Schütze, Hinrich 1999. *Foundations on statistical natural languages processing*, The MIT Press, Londres.

Mikolov, Tomas; Le, Quoc 2014. *Distributed Representations of Sentences and Documents*, Proceedings of the 31th Internacional Conference on Machine Learning, vol. 32(2), pàgines 1118-1196, Beijing.

Riemsdijk, Van, Henk 1978. *A Case Study in Syntactic Markedness: the Binding Nature of Prepositional Phrases*, Foris, Dordrecht.

Shannon, Claude E. 1948. *A Mathematical Theory of Communication*, The Bell System Technical Journal, Vol. 27, páginas 379–423, 623–656, julio y octubre.

Saint-Dizier, Patrick 2006. *Syntax and Semantics of prepositions*, Springer, Dordrecht.

Zelinsky-Wibblet, Cornelia (ed.) 1993. *The semantics of Prepositions*, Mouton de Gruyter, Berlín.