# Investigando polisemia con corpus: límites y oportunidades

Mariana Montes
Universidad Católica de Lovaina (Bélgica)
mariana.montes@kuleuven.be

**Resumen**

En este capítulo extraemos información distribucional de un corpus para representar las ocurrencias de distintos ítems léxicos, de forma tal que ocurrencias similares aparecen juntas en una nube de puntos. Por ejemplo, el verbo neerlandés *huldigen* puede significar 'rendir homenaje' o 'sostener (una opinión)' dependiendo del objeto directo. La pregunta es: ¿podemos extraer información de un corpus de forma tal que, automáticamente, obtengamos grupos separados de ocurrencias que corresponden a los distintos sentidos de un ítem polisémico? De no ser así, ¿qué información sí puede ser extraída de un corpus, y cómo nos puede ayudar en descripciones semánticas? Ilustrando uno de 32 términos en neerlandés estudiados, mostraremos cómo estos métodos logran capturar patrones colocacionales, pero la medida en que estos caracterizan un sentido en términos definicionales depende del comportamiento distribucional específico de cada palabra.

**Palabras clave:** lexicología, lingüística de corpus, neerlandés, semántica distribucional

## INTRODUCCIÓN

Uno de los principios fundamentales de la Lingüística Cognitiva es la atención por el lenguaje en uso (Langacker, 1988), lo cual fundamenta el creciente interés por métodos empíricos (Geeraerts, 1999, 2005, 2006a, 2016; Glynn, 2014). En particular, la lingüística de corpus y la lingüística computacional ofrecen herramientas esenciales para aprovechar la riqueza de fuentes que continuamente crecen en dimensiones y disponibilidad. Un programa informático puede identificar rápidamente patrones de uso a través de ingentes cantidades de texto, mientras que realizar el mismo procedimiento de manera manual tomaría inmensos recursos en términos de tiempo y esfuerzo. Dentro de este marco, el objetivo del estudio descrito aquí es evaluar y describir la aplicación de modelos distribucionales, una técnica proveniente de la lingüística computacional, al análisis semántico léxico desde la perspectiva de la Lingüística Cognitiva.

La semántica distribucional consiste en el procesamiento de corpus (es decir, grandes compilaciones de textos) para extraer información de frecuencias de ocurrencia; a partir de dichas frecuencias se pueden construir perfiles numéricos para representar distintas palabras. Tales perfiles se prestan a procesamientos matemáticos y estadísticos para llegar a una representación visual como la de la Figura 1. En otras palabras, se inicia con una masa de textos y se finaliza con una representación gráfica en la que cada punto es una ocurrencia de una palabra y la distancia entre los puntos representa la similitud entre los contextos en los

que las respectivas ocurrencias se encuentran: cuanto más cerca se ven en la nube de puntos, más similares los contextos.

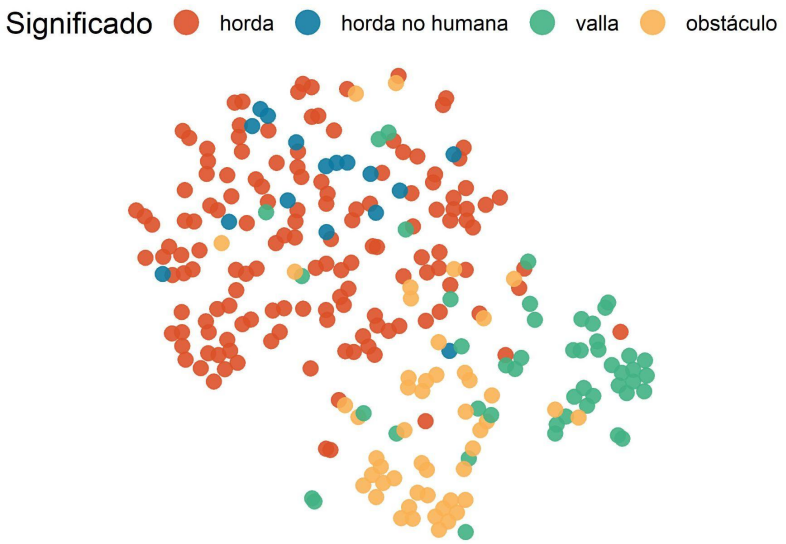


Figura 1. Representación gráfica de ocurrencias de *horde* ‘horda, valla’ en neerlandés, con colores representando anotación semántica manual

Este estudio se enmarca dentro de un proyecto mayor dedicado al desarrollo de herramientas para estudios semánticos con métodos distribucionales (Heylen, Speelman & Geeraerts, 2012; Heylen, Wielfaert, Speelman & Geeraerts, 2015; Lenci, 2018); los objetivos e interrogantes se presentarán en la sección 1. Más concretamente, este trabajo resumirá las conclusiones principales del estudio de 32 sustantivos, verbos y adjetivos en neerlandés (Montes, 2021a); la sección 2 describirá el marco teórico y la metodología. A modo de ejemplo, en la sección 3 se ilustrará el análisis de uno de los sustantivos estudiados, *horde* ‘horda, valla’.

## 1. Proyecto: semántica nefológica

El estudio presentado en estas páginas se enmarca dentro del proyecto “Semántica nefológica” (del griego *nephos* ‘nube’); el nombre hace referencia a la representación de palabras en un espacio semántico multidimensional y establece una analogía entre la exploración de patrones en ese espacio, como por ejemplo la nube de puntos en la Figura 1, y el estudio de las nubes en el campo de la meteorología. El proyecto cubre tres perspectivas principales. La perspectiva semasiológica, dominante en este estudio, se enfoca en la aplicación de métodos distribucionales al estudio de la polisemia (ver también Montes, 2021a). La perspectiva onomasiológica, en cambio, va del significado al significante, y así busca analizar distintas palabras que pueden utilizarse alternativamente para un mismo concepto. Por ejemplo, Montes, Franco y Heylen (2021) identifican los contextos en los que dos verbos neerlandeses que significan ‘destruir’, *vernielen* y *vernietigen*, se alternan o se

especializan. Finalmente, la perspectiva lectométrica busca aprovechar estos métodos cuantitativos a gran escala para comparar variedades lingüísticas (De Pascale, 2019; De Pascale & Zhang, 2021).

Dentro de la perspectiva semasiológica, el proyecto buscaba encontrar configuraciones que nos permitieran identificar significados y/o fenómenos semánticos específicos (por ejemplo, metáfora o metonimia) automáticamente. Como se verá en la sección 2, es posible crear múltiples modelos distribucionales, es decir, múltiples representaciones alternativas del comportamiento de una palabra, según cómo se defina el contexto relevante. Distintas definiciones del contexto resultan en distintas representaciones, distintos modelos; el objetivo era identificar cuál nos permite discriminar significados o fenómenos semánticos automáticamente.

Desafortunadamente, no es posible encontrar una única configuración que resulte en la mejor aproximación a una clasificación manual (ver Montes 2021a, b). Los patrones identificados por un programa informático no son los mismos que los patrones relevantes para un ser humano que hable el idioma en cuestión, si bien ocasionalmente hay acuerdos. El ejemplo que se describirá en la sección 3 es particularmente elegante en su resultado y, al menos parcialmente, puede dar la impresión de que el modelo clasificó las ocurrencias satisfactoriamente. No obstante, la coincidencia entre los resultados automáticos y los resultados manuales depende de características específicas de la palabra analizada, y los mismos parámetros no generan resultados comparables cuando se aplican a otros casos. Los modelos distribucionales identifican patrones textuales, y solo identifican patrones semánticos en la medida en que coinciden con los patrones textuales.

## 2. Marco teórico y metodología

Dado que la semántica distribucional no es tan conocida en el marco de la semántica cognitiva, las subsecciones 2.1 y 2.2 describirán los principios teóricos y técnicos subyacentes. A continuación, la subsección 2.3 enmarcará este enfoque metodológico en el campo de la lingüística cognitiva.

### 2.1. La Hipótesis Distribucional

La metodología central de este estudio se basa en la llamada *Hipótesis Distribucional*, la cual postula una correlación entre distribución y semántica; entre uso y significado (Lenci, 2018). La noción se remonta al menos a Harris (1954)[1] y particularmente a Firth (1957), afamado padre de la lingüística de corpus y autor de una cita altamente popular en introducciones al análisis de corpus y los métodos distribucionales: “You shall know a word by the company it keeps” (Firth, 1957:11), que aplica a las palabras una fórmula del tipo “Dime con quién andas y te diré quién eres”.

[1] Cf. Geeraerts (2017).

En la práctica, la Hipótesis Distribucional sugiere que podemos utilizar información sobre la distribución de una palabra, es decir, su frecuencia de uso en distintos contextos ("con quién andas") para operacionalizar su significado ("quién eres"). Un sistema informático no tiene acceso al significado de un texto de la misma forma que un ser humano, pero sí a números que codifican su frecuencia de uso; si en verdad ambos aspectos están correlacionados, un modelo computacional puede simular o modelar significados lingüísticos.

Los argumentos que apoyan esta noción son variados, pero no siempre están fundamentados en teoría lingüística o resultados empíricos (ver Montes, 2021a:152–154). En efecto, una de las conclusiones principales de la investigación descrita aquí es que no hay una correlación directa entre uso y significado, en el sentido de que no todos los patrones de uso como los identificaría un modelo computacional pueden ser atribuidos a sentidos lexicográficos directamente. No obstante, proponemos que la clasificación de ocurrencias a partir de patrones de uso, o incluso la descripción de términos en función de la claridad y diversidad de los patrones que presenta, siguen siendo un objetivo interesante y útil. Por un lado, el ejemplo de la sección 3 ilustrará que la metodología nos permite al menos describir cómo se suele utilizar una palabra en términos de los contextos formales en los que ocurre y las relaciones entre ellos. Por otro, su mismo agnosticismo semántico, es decir, la falta de una relación directa entre los patrones encontrados y significados lexicográficos, abre las puertas a diversas interpretaciones posibles y un análisis semántico más rico que el mapeo unívoco a definiciones.

### 2.2. Modelos distribucionales

Esta investigación hizo uso, concretamente, de modelos vectoriales de conteo, evitando en lo posible la caja negra de los modelos neurales, aunque el procedimiento en sí sigue siendo complejo e intrincado[2]. Con la explicación en estas líneas intentamos acercar el método a un público de lingüistas sin entrenamiento avanzado en matemática y estadísticas y echar luz sobre las conexiones entre el corpus inicial y la visualización final. La interpretación final es bastante intuitiva, pero puede ser reforzada por una descripción técnica.

Un modelo distribucional o vectorial representa palabras con *vectores*, que son esencialmente listas ordenadas de números. Una lista de colocados y medidas de asociación como pueden resultar de análisis de colocaciones en lingüística de corpus es un ejemplo de un vector: básicamente, representamos una palabra en función de su atracción con otras palabras en el corpus. Por ejemplo, cada fila en la Tabla 1 es un vector que representa una palabra. La primera fila es una representación numérica de *lingüística*; un programa recorrió todo el Corpus del Español (https://www.corpusdelespanol.org/web-dial/), recolectó todas las

[2] El estudio fue llevado a cabo con código abierto desarrollado mayormente dentro del equipo de trabajo como producto del proyecto de investigación macro: ver https://qlvl.github.io/nephosem/tutorials/ y https://montesmariana.github.io/semcloud/. Gran parte del análisis fue llevado a cabo en R (R Core Team, 2021), en particular los paquetes Rtsne, dbscan y ggplot2 (Hahsler & Piekenbrock, 2021; Krijthe, 2018; Wickham, 2016). La visualización puede ser explorada de forma interactiva en https://qlvl.github.io/NephoVis/.

ocurrencias de *lingüística*, registró las palabras que aparecían hasta 4 lugares a su izquierda o derecha y las contó. Ya que hay algunas palabras extremadamente frecuentes y muchísimas palabras poco frecuentes, las frecuencias resultantes fueron transformadas a información mutua positiva (PPMI, ver Church & Hanks, 1989; Gablasova, Brezina & McEnery, 2017), que toma una escala más manejable y representa la *atracción* entre dos palabras. Si bien hay críticas a los resultados de esta transformación por su preferencia por combinaciones infrecuentes, es la medida preferida dentro de la semántica distribucional porque los resultados finales tienden a corresponder a juicios humanos más que en modelos que usen otras medidas de asociación (Kiela & Clark, 2014).

Tabla 1

Matriz de coocurrencias a nivel de tipo, con valores (PPMI) basados en una ventana simétrica de 4 palabras en el Corpus del Español.

| nodo | lenguaje | palabra | español | hablar | comer |
|---|---|---|---|---|---|
| lingüística | 3,55 | 0,53 | 2,09 | 0,00 | 0,00 |
| léxico | 3,51 | 2,74 | 4,27 | 0,04 | 0,00 |
| computacional | 4,27 | 0,00 | 0,00 | 0,00 | 0,00 |
| investigación | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| chocolate | 0,00 | 0,00 | 0,00 | 0,00 | 4,51 |
| mate | 0,00 | 0,00 | 0,00 | 0,00 | 0,46 |

Considerando la Tabla 1, vemos que *lingüística* presenta cierta atracción por *lenguaje* y *español*, menor atracción por *palabra* y ninguna con *hablar* o *comer*; esto indica que, o bien no coocurren en el corpus o que lo hacen muy poco en relación con sus frecuencias individuales. Los valores de *léxico* son similares, con mayor atracción por *palabra* y una ligerísima atracción por *hablar*. Por el contrario, *chocolate* y *mate* no ocurren con las mismas palabras que *lingüística* o *léxico* y en cambio exhiben cierta atracción —alta y baja respectivamente— por *comer*. Un modelo vectorial utilizaría muchísima más información que la presentada en esta tabla, con miles o decenas de miles de columnas en vez de las cinco ilustradas aquí, pero este ejemplo alcanza para explicar el procedimiento. Una vez que se han computado los vectores, podemos comparar las palabras en función de sus valores: vemos que los vectores de *lingüística* y *léxico* son muy similares y diametralmente opuestos a los de *chocolate* y *mate*. De acuerdo a la Hipótesis Distribucional, esto sugiere que *lingüística* y *léxico* son semánticamente similares entre sí y muy distintas de *chocolate* y *mate*, lo cual no es enteramente descabellado.

Ahora bien, la Tabla 1 muestra vectores a nivel de tipo: cada fila combina datos de todas las ocurrencias de una palabra en un corpus, subsumiendo cualquier variación interna en una representación única. Dado que el interés de esta investigación yace precisamente en describir la estructura interna de una palabra, utilizamos representaciones a nivel de caso, como muestra la Tabla 2 para las ocurrencias de *estudiar* reportadas en (1), (2) y (3). No sería factible o útil simplemente identificar las palabras en el contexto inmediato de cada ocurrencia, ya que hay muy baja coincidencia entre estos elementos. En cambio, la estrategia para pasar del nivel de tipo al nivel de caso es utilizar las representaciones a nivel de tipo de las palabras en el contexto inmediato y combinarlas como una representación del caso (Heylen et al., 2015; Schütze, 1998). Por ejemplo, la fila correspondiente a *léxico* en la Tabla 1 es utilizada para representar el ejemplo (1) en la Tabla 2, ya que *léxico* ocurre en ese contexto; la suma de los vectores tipo de *lingüística* y *computacional* (sus respectivas filas en la Tabla 1) resulta en el vector caso del ejemplo (2) en la Tabla 2, y lo mismo para los vectores de *mate* y *chocolate* en la representación de (3).

(1) ¿Te gustaría *estudiar* el léxico del neerlandés?

(2) También *estudian* esto en lingüística computacional.

(3) Cuando *estudio* tomo mate y como chocolate.

Tabla 2

Matriz de coocurrencias a nivel de caso, sumando los vectores de ciertas palabras en el contexto.

| nodo | lenguaje | palabra | español | hablar | comer |
|---|---|---|---|---|---|
| estudiar_1 | 3,55 | 2,74 | 4,27 | 0,04 | 0,00 |
| estudiar_2 | 7,82 | 0,53 | 2,09 | 0,00 | 0,00 |
| estudiar_3 | 0,00 | 0,00 | 0,00 | 0,00 | 4,97 |

Los vectores caso de los ejemplos (1) y (2) resultan entonces muy similares porque los vectores tipo de *lingüística* y *léxico* son similares, mientras que el vector caso de (3) es diametralmente opuesto porque así lo son los vectores tipo de *mate* y *chocolate*. En la práctica, compararíamos muchos más ejemplos simultáneamente y, crucialmente, usaríamos distintos criterios para seleccionar las palabras del contexto inmediato. Los distintos criterios resultan en modelos diferentes que pueden representar distintas relaciones, pero no hay una guía definitiva para seleccionar las palabras más adecuadas para un fin (Montes, 2021b).

A nivel matemático, la comparación se realiza con medidas de distancia como el coseno (Jurafsky & Martin, 2020:105), que tiende hacia 1 cuando los vectores son muy similares entre sí, como entre *léxico* y *lingüística* o entre los ejemplos (1) y (2), y hacia 0 cuando son diametralmente distintos, como entre *lingüística* y *mate* o entre los ejemplos (2) y (3). En base

a estas distancias, podemos procesar los vectores nuevamente para obtener una representación más intuitiva, aunque a costa de cierta información. El algoritmo t-SNE (van der Maaten & Hinton, 2008) intenta reproducir las distancias basadas en una matriz multidimensional (es decir, una tabla con muchas columnas) en un plano en dos dimensiones, que luego se puede graficar con una nube de puntos. La idea es que si dos vectores, como los de los ejemplos (1) y (2), son muy similares, estarían cerca en el gráfico, y la visualización destacaría los grupos de elementos que son similares entre sí. No obstante, las distancias entre puntos lejanos o entre grupos distintos de elementos no son interpretables de la misma manera: la reducción de dimensiones conlleva cierta pérdida de información.

Mientras recurrimos a t-SNE para graficar los modelos, utilizamos el algoritmo de agrupamiento HDBSCAN (Campello, Moulavi & Sander, 2013) para identificar automáticamente los grupos de elementos similares. Este algoritmo busca regiones de puntos cercanos entre sí rodeados de espacios con pocos puntos. Tiene la ventaja de que no intenta agrupar todos los elementos, sino que acepta que algunos pueden no pertenecer a ningún grupo. Además, le atribuye a cada elemento un puntaje en términos de su centralidad o "grado de membresía" al grupo al que fue asignado. Al mapear la clasificación de HDBSCAN como colores en los gráficos, podemos visualizar, identificar, describir y comparar patrones encontrados por una larga cadena de procedimientos automáticos. La práctica concreta se hará más clara en el ejemplo en la sección 3.

### 2.3. Lingüística distribucional y lingüística cognitiva

La atención por el lenguaje en uso, uno de los pilares de la lingüística cognitiva (Geeraerts, 2016; Langacker, 1988), implica alta compatibilidad con una metodología empírica, cuantitativa, que parte de los datos hacia las clasificaciones en lugar de imponerlas *a priori*. Esta es la línea detrás de la semántica cognitiva cuantitativa (Glynn & Fischer, 2010; Glynn & Robinson, 2014; Gries & Stefanowitsch, 2006). Es clave remarcar que la utilización de métodos cuantitativos que maximizan la automatización del análisis de ninguna manera excluye el aspecto cualitativo y el recurso a la introspección e interpretación. La computadora no hace investigación por su cuenta: somos los seres humanos quienes nos hacemos las preguntas, formulamos hipótesis, diseñamos los estudios, interpretamos los resultados y construimos narrativas que constituyen nuestro conocimiento científico (Geeraerts, 2010a).

Un segundo pilar clave de la lingüística cognitiva es la noción de efectos prototípicos (a partir de Rosch, 1978), lo que Geeraerts (2010b) describe como la dimensión cuantitativa en el estudio del significado. La pregunta ya no es qué existe y qué ocurre, sino que se reformula en términos de prominencia, peso, flexibilidad y grados de pertenencia (Geeraerts, 2006b:74).

En este sentido, los modelos distribucionales ofrecen representaciones lingüísticas en línea con esta noción, si bien se limitan a información textual. Los gráficos resultantes de t-SNE representan distancias relativas, sin límites concretos entre grupos de elementos; los

grupos computados por HDBSCAN incluyen información sobre la centralidad y los grados de pertenencia de sus miembros y algunos puntos pueden no pertenecer a ninguno de ellos. Si bien hablaremos de grupos en términos discretos, porque necesitamos pensarlos así para describirlos, la metodología nos permite tratarlos desde una perspectiva más relativa y cuantitativa.

## 3. Ejemplo: hordas y vallas

En esta comunicación nos centraremos en uno de los sustantivos analizados, *horde*. El material de análisis consiste en 278 ocurrencias extraídas de un corpus de periódicos de Flandes (la región de Bélgica donde se habla neerlandés) y los Países Bajos. El corpus en su totalidad, compuesto de textos escritos de género periodístico publicados entre 1999 y 2004, contiene 520 millones de palabras anotadas automáticamente con el lema, la categoría gramatical y relaciones sintácticas (ver Montes, 2021a).

En el caso concreto de *horde*, se extrajeron originalmente 280 ocurrencias que fueron anotadas manualmente por al menos tres individuos distintos, estudiantes de una universidad flamenca. La anotación consistió principalmente en asignar uno de tres significados posibles o "ninguno de las anteriores". Dos de las ocurrencias fueron eliminadas: una pertenecía a un texto en francés y la otra correspondía a instrucciones para un crucigrama.

El sustantivo neerlandés *horde* tiene dos significados no relacionados (homónimos), uno de los cuales se subdivide en una aplicación literal y una metafórica. El ejemplo (4) ilustra el significado 'horda', que cubre 167 ocurrencias de la muestra, mientras que los ejemplos (5) y (6) representan los sentidos literal y metafórico del significado 'valla' (u 'obstáculo'). Estos dos últimos significados cubren 58 y 53 ocurrencias respectivamente.

(4) Hele *horden* van toeristen snellen dan toe op strandressorts, tropische regenwouden en cultuursteden. (*Algemeen Dagblad*, 2003-05-06, Art. 50)

Enteras *hordas* de turistas abarrotan retiros playeros, selvas tropicales y ciudades culturales.

(5) Charles van Commenee (werpen en meerkamp) en Marjan Olyslager (sprint en *horden*) hebben inmiddels een andere werkkring gevonden… (*De Standaard*, 2000-07-13, Art. 97)

Charles van Commenee (lanzamientos y pruebas combinadas) y Marjan Olyslager (carrera de velocidad y *vallas*) ya encontraron un nuevo ambiente de trabajo…

(6) Het internationale ruimtestation-in-aanbouw ISS heeft een belangrijke *horde* genomen. (*Parool*, 2001-01-03, Art. 74)

La estación espacial internacional bajo construcción, EEI, ha sorteado un *obstáculo* importante.

Luego de la anotación, identificamos una categoría más pequeña dentro del significado 'horda' que incluiría los grupos de seres no humanos, desde insectos a vehículos. Esta categoría solo abarca 18 de las ocurrencias de 'horda'. En el caso específico de *horde*,

esperábamos que los modelos automáticos pudieran distinguir fácilmente entre los homónimos y que tuvieran un poco más de dificultad en distinguir entre los usos metafóricos y literales del homónimo. La motivación de esta hipótesis es que los homónimos 'horda' y 'valla, obstáculo' ocurrirían en contextos drásticamente distintos, mientras que los contextos de los usos literales y metafóricos de 'valla' serían más difíciles de discriminar. Para los anotadores la tarea no fue problemática: en 244 de las 280 ocurrencias analizadas (87,14%), todos los anotadores eligieron la misma opción, y en solo 5 casos no hubo ningún acuerdo entre los anotadores.

### 3.1. Resultados: visualización

Como se ha explicado más arriba, la metodología prevé la generación de múltiples modelos distintos, cada uno caracterizado por una cierta definición del contexto. Por medio de un algoritmo de agrupación se seleccionaron 8 modelos representativos para una descripción más profunda y cualitativa, pero a fines de ilustración nos concentraremos en uno solo. En este modelo en particular, cada ocurrencia es representada por los sustantivos, verbos y adjetivos presentes en la misma oración y como máximo a cinco lugares a la izquierda o la derecha del nodo (*horde*). Además, el PPMI entre la palabra y *horde* debe ser mayor que 0; debe haber cierta atracción entre ambas. Estas restricciones eliminaron 19 de las ocurrencias anotadas, porque no tenían ninguna palabra en el cotexto que cumpliera con todos los requisitos. Los vectores que representan los elementos contextuales están definidos por las mismas palabras seleccionadas para el contexto de cada ocurrencia. En otras palabras, si entre las 259 ocurrencias modeladas hay 402 elementos contextuales distintos seleccionados, esos mismos 402 elementos son utilizados como las dimensiones de los vectores de nivel tipo, es decir, las columnas de la Tabla 1.

La Figura 1 representa cada instancia de *horde* como un punto, de forma tal que las ocurrencias con contextos similares aparecen cerca en la nube de puntos. Los colores representan los significados asignados manualmente: el hecho de que forman áreas bastante uniformes y distinguibles indica una fuerte correspondencia entre las representaciones computacionales y los significados lexicográficos. Para mayor claridad, la Figura 2 repite el gráfico resaltando los ejemplos (4) en rojo, (5) en verde y (6) en amarillo. Por un lado, vemos que los tres significados principales cubren áreas relativamente separadas, con poca superposición y mejor distinción entre los homónimos (entre 'horda' y los demás) que entre los usos literal y metafórico del segundo homónimo ('valla' y 'obstáculo'). La categoría de 'horda no humana' es indistinguible del resto de 'horda'.

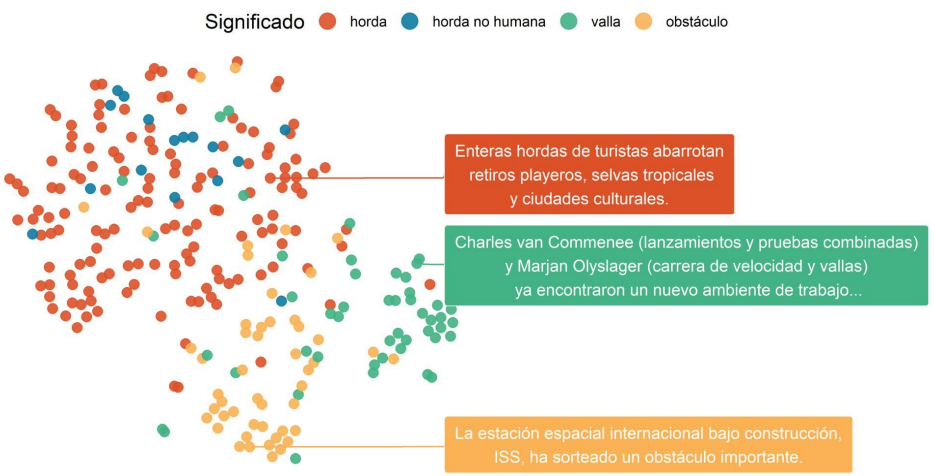


Figura 2. Ocurrencias de *horde* con colores representando los significados y tres ejemplos concretos indicados

Ahora bien, si no tuviéramos anotación manual, una alternativa sería usar HDBSCAN o algún otro algoritmo similar para detectar grupos relevantes. De esta manera obtenemos un gráfico como el de la Figura 3, en el que los colores representan los distintos grupos identificados automáticamente, con gris reservado para los puntos que no fueron asignados a ningún grupo (Grupo NA). Para la descripción de los resultados nos concentraremos en los ejemplos que sí fueron agrupados, ya que representan patrones textuales concretos. No obstante, un análisis más detallado puede incluir descripciones del grado en que los puntos descartados efectivamente pueden ser agrupados con los demás patrones, así como un análisis de los usos de *horde* que presentan más variación que la capturada por los grupos identificados.

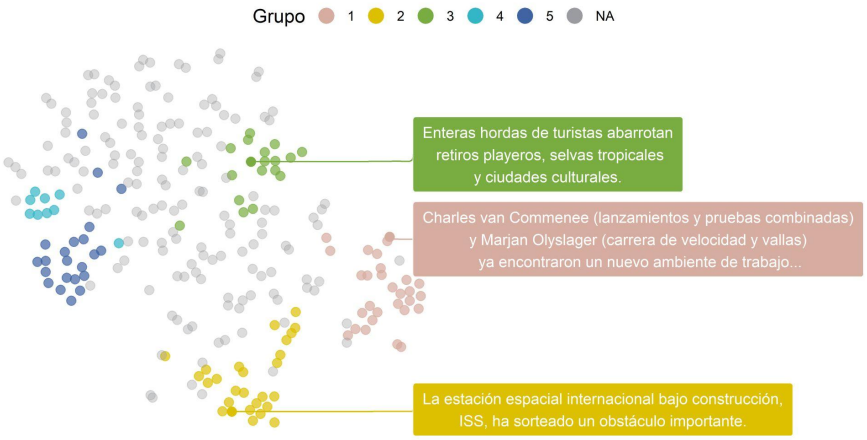


Figura 3. Ocurrencias de *horde* con colores representando grupos automáticos y tres ejemplos concretos indicados

Una forma de describir sucintamente los grupos identificados por el algoritmo es a través de las palabras que coocurren con las instancias de *horde* en un cierto grupo. La Figura 4 indica esta propiedad particular para cada uno de los grupos automáticos, junto con el Valor *F*, que mide la precisión y exhaustividad de una palabra en relación con el grupo que representa. Por ejemplo, el grupo celeste tiene 9 elementos, de los cuales 6 coocurren con el sustantivo *fan* 'fan, fanático'. Dado que todas las ocurrencias de *horde* en la muestra que ocurren con *fan* están incluidas en el grupo celeste, la precisión de *fan* en la definición del grupo es del 100%, o 1 en la escala de 0 a 1. La exhaustividad, en cambio, se refiere a la proporción de elementos del grupo (9) que coocurren con *fan* (6), en este caso dos tercios. El Valor *F* es la media armónica entre los dos valores. En pocas palabras, cuanto más alto es el Valor *F* (más cercano a 1), mayor es la correlación entre el grupo de instancias y la palabra seleccionada. En este caso, seleccionamos la palabra con el Valor *F* más alto para cada grupo. Montes (2021a:105–138) explora la interpretación de distintas relaciones entre los grupos y las palabras que lo representan; tanto grupos caracterizados por una palabra clave como grupos caracterizados por múltiples palabras infrecuentes pero similares o por combinaciones de palabras.

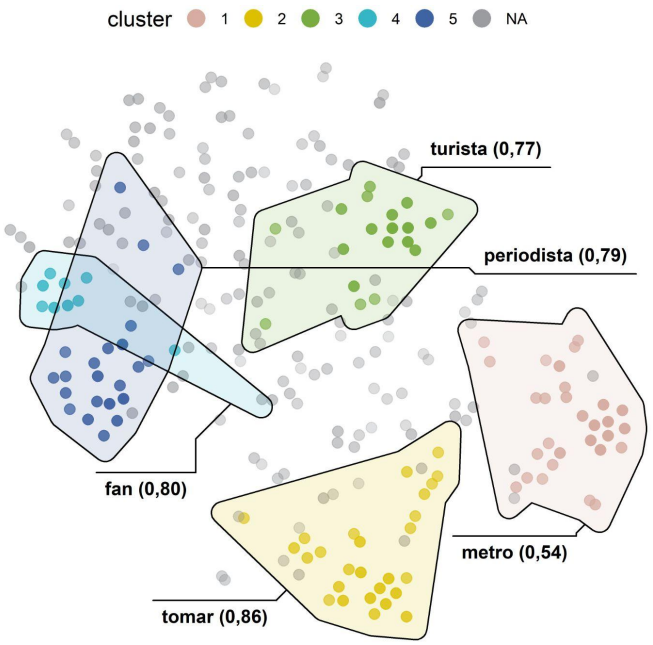


Figura 4. Ocurrencias de *horde* con colores representando grupos automáticos y etiquetas indicando las palabras más representativas de cada grupo

Los grupos identificados por el algoritmo en este modelo de *horde* representan claros patrones textuales que dan lugar a distintas interpretaciones semánticas. Por un lado, los grupos 1 y 2 son respetables representaciones de 'valla' y 'obstáculo'. El primero agrupa instancias que coocurren con *meter* 'metro' y otros términos deportivos, definiendo el contexto de carreras de vallas. El segundo está mayormente caracterizado por la coocurrencia con *nemen* 'tomar', pero también *eerst* 'primero' y *laatst* 'último', porque este significado metafórico ocurre sobre todo en la expresión *de eerste/laatste horde nemen* 'saltar el primer/último obstáculo'. Por otro lado, el significado de 'horda' presenta una gran cantidad de instancias no agrupadas y tres grupos distintos, cada uno caracterizado por un sustantivo distinto: *toerist* 'turista', *fan* 'fan, fanático' y *journalist* 'periodista'. En principio, uno puede atribuir esta distinción a diversos tipos de entidades que pueden formar 'hordas'. En efecto, dos palabras muy frecuentes que no coocurren juntas, como *journalist* y *toerist*, pueden generar sus propios grupos de coocurrencias en una nube de puntos como esta. Sin embargo, un análisis un poco más profundo, en particular comparando con otros modelos, revela que estos grupos de ocurrencias también se diferencian en base a patrones más sutiles, generados por palabras menos frecuentes. Las ocurrencias del grupo 3 no solo tienden a coocurrir con *toerist* sino también con *naar* 'hacia' y distintos verbos relacionados con flujos y movimientos. Las de los grupos 4 y 5, en cambio, así como otras ocurrencias alrededor, tienden a coocurrir con *door* (la preposición del complemento agente) y verbos relacionados con 'rodear' y 'perseguir'. En suma, los grupos no representan simplemente distintos tipos de entidades que pueden constituir "hordas", sino distintas facetas del significado 'horda' que se activan en cada uno de los contextos. La ocurrencia con *toerist* en

el grupo 3 coincide con la predominancia de la faceta de flujo incontrolable de nómadas. En contraste, los grupos 4 y 5 focalizan el aspecto amenazador de las hordas de paparazzi, en cuyo contexto predomina la función gramatical de complemento agente en construcciones pasivas con verbos más violentos.

## CONCLUSIÓN

En esta comunicación presentamos un análisis que utiliza una técnica computacional, modelos distribucionales, para análisis semántico en el marco de la lingüística cognitiva. Primero describimos el marco teórico y metodológico de los modelos distribucionales y los relacionamos con principios fundamentales de la semántica cognitiva. Luego ejemplificamos con un estudio de caso concreto, en el que modelamos 280 ocurrencias del sustantivo neerlandés *horde* 'horda, valla'.

Por medio del análisis mostramos que los espacios vectoriales de nivel caso representan patrones colocacionales/textuales. Como sugiere la Hipótesis Distribucional, es posible identificar significados y fenómenos semánticos automáticamente, pero solo si coinciden con patrones textuales. En el caso de *horde*, las ocurrencias de 'valla' y 'obstáculo' (metafórico) tienden a ocurrir en expresiones relativamente fijas que facilitan la extracción automática. Además, las ocurrencias de 'horda' presentan variación interna, con patrones textuales más sutiles que se corresponden con distintas facetas del significado.

En otras palabras, no es recomendable depender de métodos automáticos ciegamente, pero ciertamente pueden contribuir a una descripción empírica del uso auténtico del lenguaje y de la relación entre patrones textuales y fenómenos semánticos.

## REFERENCIAS BIBLIOGRÁFICAS


Campello, R. J. G. B., Moulavi, D. & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. En: Pei, J., Tseng, V. S., Cao, L., Motoda, H. & Xu, G. (Eds.) *Advances in Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science (pp. 160–172). Berlin, Heidelberg: Springer.

Church, K. W. & Hanks, P. (1989). Word association norms, mutual information, and lexicography. *ACL '89: Proceedings of the 27th annual meeting on Association for Computational Linguistic* (pp. 76–83). Association for Computational Linguistics.

De Pascale, S. (2019). *Token-based vector space models as semantic control in lexical lectometry* (Tesis doctoral). KU Leuven, Leuven.

De Pascale, S. & Zhang, W. (2021). Scoring with Token-based Models. A Distributional Semantic Replication of Socioectometric Analyses in Geeraerts, Grondelaers, and Speelman (1999). En: Kristiansen, G., Franco, K., De Pascale, S., Rosseel, L. & Zhang, W. (Eds.) *Cognitive Sociolinguistics Revisited* (pp. 186–199). De Gruyter Mouton.

Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. En: Firth, J. R. (Ed.), *Studies in Linguistic Analysis*, Special volume of the Philological Society (pp. 1–32). Oxford: Blackwell.

Gablasova, D., Brezina, V. & McEnery, T. (2017). Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence: Collocations in Corpus-Based Language Learning Research. *Language Learning* (online).

Geeraerts, D. (1999). Idealist and empiricist tendencies in Cognitive Linguistics. En: Janssen, T. & Redeker, G. (Eds.), *Cognitive Linguistics: Foundations, Scope, and Methodology* (pp. 163–194). Berlin: Mouton de Gruyter.

Geeraerts, D. (2005). Lectal variation and empirical data in Cognitive Linguistics. In Ruiz de Mendoza Ibáñez, F. J. & Peña Cervel, M. S. (Eds.) *Cognitive linguistics: Internal dynamics and interdisciplinary interaction* (pp. 163–189). Berlin; New York: Mouton de Gruyter.

Geeraerts, D. (2006a). Methodology in Cognitive Linguistics. En: Kristiansen, G., Achard, M., Dirven, R., & Ruiz de Mendoza Ibáñez, F. J. (Eds.) *Cognitive linguistics: Current applications and future perspectives* (pp. 21–49). Berlin: Mouton de Gruyter.

Geeraerts, D. (2006b). *Words and other wonders: Papers on lexical and semantic topics*. Berlin; New York: Mouton de Gruyter.

Geeraerts, D. (2010a). The doctor and the semantician. En: Glynn, D. & Fischer, K. (Eds.) *Quantitative methods in cognitive semantics: Corpus-driven approaches* (pp. 63–78). Berlin; New York: De Gruyter Mouton.

Geeraerts, D. (2010b). *Theories of lexical semantics*. Oxford; New York: Oxford University Press.

Geeraerts, D. (2016). The sociosemiotic commitment. *Cognitive Linguistics*, *27*(4), 527–542.

Geeraerts, D. (2017). Distributionalism, old and new. En: Makarova, A., Dickey, S. M. & Divjak, D. (Eds.) *Each venture a new beginning: Studies in honor of Laura A. Janda* (pp. 29–38). Bloomington, Indiana: Slavica.

Glynn, D. (2014). Polysemy and synonymy: Cognitive theory and corpus method. En: Glynn, D. & Robinson, J. A. (Eds.) *Corpus methods for semantics: Quantitative studies in polysemy and synonymy* (pp. 7–38). Amsterdam; Philadelphia: John Benjamins Publishing Company.

Glynn, D. & Fischer, K. (Eds.). (2010). *Quantitative methods in cognitive semantics: Corpus-driven approaches*. Berlin; New York: De Gruyter Mouton.

Glynn, D. & Robinson, J. A. (Eds.). (2014). *Corpus methods for semantics: Quantitative studies in polysemy and synonymy*. Amsterdam; Philadelphia: John Benjamins Publishing Company.

Gries, S. T. & Stefanowitsch, A. (Eds.). (2006). *Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis*. Berlin; New York: Mouton de Gruyter.

Hahsler, M. & Piekenbrock, M. (2021). *Dbscan: Density based clustering of applications with noise (DBSCAN) and related algorithms*. Disponible en: https://github.com/mhahsler/dbscan

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146–162.

Heylen, K., Speelman, D. & Geeraerts, D. (2012). Looking at word meaning. An interactive visualization of Semantic Vector Spaces for Dutch synsets. *Proceedings of the eacl 2012 Joint Workshop of LINGVIS & UNCLH* (pp. 16–24). Avignon.

Heylen, K., Wielfaert, T., Speelman, D. & Geeraerts, D. (2015). Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis. *Lingua*, *157*, 153–172.

Jurafsky, D. & Martin, J. H. (2020). *Speech and Language Processing.* (Online)

Kiela, D. & Clark, S. (2014). A Systematic Study of Semantic Vector Space Model Parameters. *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality* (pp. 21–30). Gothenburg: ACL.

Krijthe, J. (2018). *Rtsne: T-distributed stochastic neighbor embedding using a barnes-hut implementation*. Disponible en: https://github.com/jkrijthe/Rtsne

Langacker, R. W. (1988). An overview of cognitive grammar. En: Rudzka-Ostyn, B. (Ed.) *Current Issues in Linguistic Theory* (pp. 3–48). Amsterdam: John Benjamins Publishing Company.

Lenci, A. (2018). Distributional Models of Word Meaning. *Annual Review of Linguistics*, *4*(1), 151–171.

Montes, M. (2021a). *Cloudspotting: Visual analytics for distributional semantics* (Tesis doctoral). KU Leuven, Leuven.

Montes, M. (2021b). Modelling meaning granularity of nouns with vector space models. *Papers of the Linguistics Society of Belgium*, *15*.

Montes, M., Franco, K. & Heylen, K. (2021). Indestructible Insights. A Case Study in Distributional Prototype Semantics. En: Kristiansen, G., Franco, K., De Pascale, S., Rosseel, L. & Zhang, W. (Eds.) *Cognitive Sociolinguistics Revisited* (pp. 251–263). De Gruyter Mouton.

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Disponible en: https://www.R-project.org/

Rosch, E. (1978). Principles of categorization. En: Rosch, E. & Lloyd, B. B. (Eds.) *Cognition and Categorization* (pp. 27–48). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Schütze, H. (1998). Automatic Word Sense Discrimination. *Computational Linguistics*, *24*(1), 97–123.

van der Maaten, L. & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Disponible en: https://ggplot2.tidyverse.org