Reconocimiento automático de unidades fraseológicas

Jorge Antonio Leoni de León
Université de Genève

1 Introducción

La creación de modelos del lenguaje, ya sea con propósitos académicos o industriales, debe resolver una larga serie de dificultades, todas insoslayables, planteadas por el conjunto de la fenomenología lingüística. Sin embargo, cualquiera que sea la perspectiva adoptada en la investigación, es necesario alcanzar compromisos entre la teoría, que implica una concepción concreta del lenguaje, y la práctica que busca el desarrollo de metodologías para resolver un problema bien determinado. Luego de mucho tiempo de estar en la periferia de las preocupaciones de la lingüística teórica, el léxico ocupa un lugar central de la investigación científica contemporánea. Es difícil, si no imposible, concebir el lenguaje sin alguna forma de entender el léxico; esto se manifiesta dramáticamente en los modelos informáticos del lenguaje, donde es imperativo establecer una estrategia para aprovechar la información léxica en uno u otro sentido (semántico, sintáctico, fonológico, etcétera).

Sin embargo, en los modelos lingüísticos, las limitaciones de los módulos léxicos son rápidamente puestas en evidencia por la fraseología: la frecuencia de las unidades fraseológicas las hace inevitables y su complejidad obliga a una gran fineza en el tratamiento automático, si no se quiere sacrificar una gran parte de la precisión del modelo lingüístico. Por ejemplo, aunque una unidad pluriverbal como "*saco de dormir*" es difícilmente ambigua, podemos interrogarnos sobre las condiciones que hacen surgir la *idiomaticidad* de "*meter la pata*" y que nos conducen a interpretarla, bajo ciertas circunstancias, como una locución o unidad de sentido compleja y, en otras, como una expresión literal.[1] Esta operación es la que llamaremos, de

[1] Sobre la fraseología española recomendamos la lectura de Corpas Pastor (1996).

ahora en adelante, reconocimiento y constituye el tema de este artículo, el cual está basado en una investigación en progreso.

La lexicografía y la lingüística informática tienen en común un problema: la recopilación de unidades de sentido para la construcción de sus diccionarios. Sin embargo, una gran diferencia las separa: en lingüística informática hay una menor necesidad de establecer la definición. Por ejemplo, si el léxico en cuestión está destinado a un analizador sintáctico (o *parser*), lo que interesa de cada entrada es el conjunto de rasgos y relaciones formales que el modelo sintáctico requiera. De esta manera, encontramos datos como género y número (no es lo mismo "un *fondo* azul" que "los *fondos* del fideicomiso"), que también la lexicografía utiliza, y como la posición ("*cierto* texto" contra un "un texto *cierto*") o, incluso, la estructura de casos y funciones temáticas (el verbo *correr* asigna la función temática de *tema*, al objeto, en una oración como "*correr* el programa" y de dirección, al complemento, en "*corrió* hacia el estanque"). Una parte esencial de este trabajo puede ser realizado manualmente, pero los costos son altos y el desarrollo lento. Esto ha llevado a algunos investigadores a proponer la automatización parcial o total del proceso de recolección de datos léxicos.[2] Estas propuestas son extensivas a las unidades fraseológicas; las metodologías se dividen en dos grupos. El primero corresponde a las estrategias estadísticas, que son las más ampliamente utilizadas. De ellas sólo diremos, en general, que están basadas en la probabilidad de que dos o más elementos léxicos (generalmente no más de tres) se produzcan simultáneamente en cierto contexto; es decir, la noción de *frecuencia* es privilegiada. El segundo grupo está conformado por los métodos desarrollados a partir del conocimiento lingüístico. Es nuestro propósito esbozar lo que puede ser una metodología basada en Principios y Parámetros (P&P)[3] para el reconocimiento de unidades fraseológicas; por razones de espacio nos referiremos únicamente a lo que denominamos como *lexemas plurimembres*.

2 Convergencia léxica con unidad de sentido

Antes de definir los elementos que trataremos, señalaremos algunos presupuestos de nuestro modelo léxico (no olvidemos que su objetivo es servir en la recolección – semi-automática de unidades léxicas). El parser sobre el cual nos basamos, deberá reconocer tres niveles de realización léxica:

---

2 El resultado es un aumento de la cantidad de entradas con una disminución de la calidad del léxico, lo que puede ser paliado con una revisión manual del diccionario.

3 Chomsky (1981) sienta las bases de P&P; Haegeman (1994) es una excelente introducción a este modelo generativo.

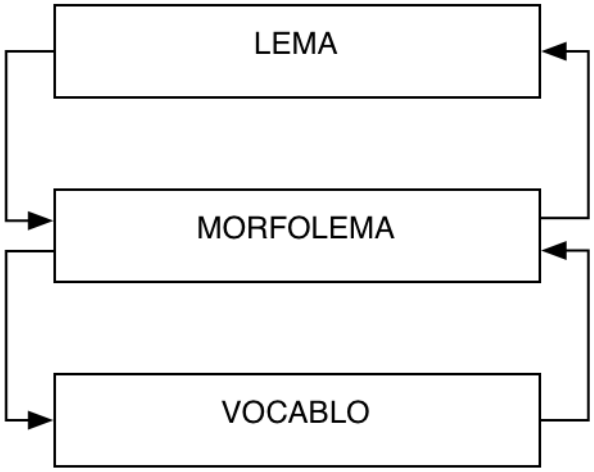

Ilustración 1: Niveles de realización léxica

Las entradas en nuestra base de datos son llamadas lemas, que son formas canónicas de las unidades léxicas, a partir de las cuales se realizan operaciones de derivación (por ejemplo, para *amamos*, *aman* y *amaréis* podemos definir el lema como el infinitivo, *amar*, sin romper la tradición lexicográfica, pero bien podría ser cualquier otra forma que se considere conveniente como la raíz, {*am*-}). Las formas derivadas son los *morfolemas*, los cuales, una vez insertados en la frase, se constituyen en *vocablos*.[4] En la Ilustración 1, las líneas descendentes señalan la derivación; las ascendentes, el reconocimiento. La importancia de estas distinciones radica en la posibilidad de establecer la identidad de dos vocablos correspondientes a un solo lema; esto es más claro en francés:

Ej. 1 Il a toujours vécu ici.

En el Ej. 1, la tercera persona del auxiliar, "*a*", y el participio "*vécu*" son dos vocablos, correspondientes al morfolema "*a vécu*", derivado del lema "*vivre*". De esta manera señalamos la relación entre estos elementos a pesar del adverbio "*toujours*" insertado entre el auxiliar y el participio.

Ahora bien, los lexemas plurimembres, como su nombre lo indica, son lexemas formados por más de un elemento, que a su vez es un lexema también. El resultado es un conjunto lexemático (de varios miembros) portador de una de las siguientes etiquetas

[4] En este modelo, el *lema* y el *morfolema* de un infinitivo coinciden.

gramaticales: *sustantivo*, *adjetivo*, *adverbio* o *preposición*. En esta ocasión nos concentraremos en los sustantivos, por ejemplo: "*saco de dormir*" y "*el pato de la fiesta*".[5] Idealmente, en nuestro sistema, cada lexema plurimembre debe estar representado por un lema plurimembre, o, en su defecto, debe existir la intuición necesaria para reconocerlo,[6] lo que debe ser posibilitado por una síntesis de conocimiento lingüístico (sintáctico en nuestro caso). Estamos, por lo tanto, ante una convergencia, de unidades léxicas, con unidad de sentido, lo que justifica la inclusión de los lexemas plurimembres en cualquier diccionario (como las expresiones francesas "*chemin de fer*" – ferrocarril – y "*pomme de terre*" – papa –).

Esquemáticamente, podemos describir el procesamiento automático del lenguaje como una serie de operaciones de validación de elementos oracionales con la base de datos léxica; conforme estos van siendo identificados, el parser intenta asignarles funciones gramaticales sobre la base de los datos disponibles. El tratamiento llega a su fin cuando la estructura de argumentos del verbo es satisfecha y las posibilidades de combinación están agotadas.[7] No es raro obtener más de una estructura para una misma oración: "*vi al hombre con los binoculares*" puede significar tanto que el sujeto utilizó los binoculares como instrumento, como que el hombre tenía unos binoculares. La importancia del verbo se explica por la riqueza de su estructura de argumentos, la cual incluye informaciones como subcategorización, funciones temáticas, categoría y Caso. Tradicionalmente, estos elementos eran considerados exclusivos del verbo; sin embargo los sustantivos también pueden tener una estructura predicativa compleja que permite identificar objetos directos, indirectos e incluso complementos dentro del sintagma sustantivo:

Ej. 2 Llegada a Ginebra.

Ej. 3 El uso de vehículos motorizados.

En Ej. 2, la subcategorización del sustantivo permite identificar "*Ginebra*" como dirección, destino; similarmente, en Ej. 3, la estructura léxica de "*uso*" impone el empleo de la preposición "*de*" para marcar el objeto directo. La Ilustración 2 representa nuestra clasificación de los sustantivos según el tipo de predicado (o su ausencia) que rijan; dicha ilustración ha sido desarrollado a partir de D'Introno (2002), quien estudia la predicación de los sustantivos y la asignación de Caso al interior de esta categoría.

---

5 El "*pato de la fiesta*" se refiere a una persona ridiculizada en público, un hazmerreír.
6 Wehrli (1997) se ha referido ampliamente al tratamiento de estas unidades.
7 Wehrli (1997) hace una síntesis de las estrategias de procesamiento.

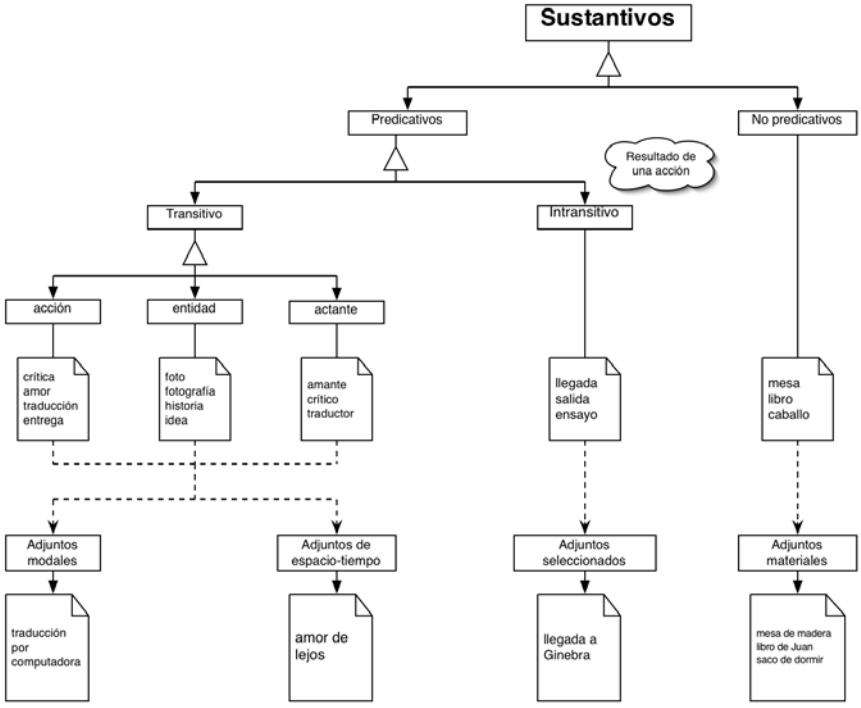

Ilustración 2: Clasificación de sustantivos según su capacidad predicativa.

Los lexemas plurimembres no son formados al azar, sino que la asociación de dos elementos o más, para crear un nuevo sustantivo, debe respetar ciertos principios (algunos de ellos sintácticos, como los que nos ocupan; otros, por ejemplo, semánticos o pragmáticos), que, además, nos pueden guiar en el proceso de reconocimiento. La Ilustración 2 refuerza esta hipótesis a la vez que sugiere otros elementos pertinentes. En primer lugar, los lemas sustantivos de nuestro diccionario deben contar con una descripción adecuada de la subcategorización, para que el sistema esté en la medida de poder estimar una probabilidad de construcción plurimembre. Pongamos por caso la oración "*le dijo de su llegada a Ginebra a Ana*"; un sistema como el que esbozamos debe permitir la correcta desambiguación de esta frase al deducir que el sintagma preposicional "*a Ginebra*" pertenece al sustantivo *"llegada"* y que el otro sintagma preposicional, "*a Ana*", está relacionado con el verbo **decir**. Así, el sistema también podrá proponer "*crítico de Arte*" o "*amante de los animales*" como posibles lemas plurimembres. Pero hay otro aspecto que surge a la luz del esquema de la Ilustración 2: el modelo puede

ser reforzado con una codificación adicional del diccionario; los lemas pueden ser portadores rasgos que indiquen si se trata de materia, lugar o nombre propio. Esto permitiría que sustantivos no predicativos con adjuntos materiales puedan ser considerados como posibles lexemas plurimembres: "*silla de madera*", "*libro de oro*". El conjunto de los rasgos pertinentes queda por ser definido; WordNet (Fellbaum, 1998) puede eventualmente proveer la información de base necesaria para realizar ese trabajo.

3 Comentarios finales

Ha sido nuestra intención remarcar la pertinencia del conocimiento lingüístico en un área de investigación que se encuentra en expansión: la extracción automática de unidades fraseológicas. Las propuestas de formalismos son abundantes y de gran interés, Fontenelle (1997) presenta una síntesis de perspectivas muy interesante. Consideramos muy posible que en el futuro próximo lleguemos a una fusión de métodos y perspectivas que sólo puede enriquecer la lingüística como disciplina científica.

Asimismo esperamos que el problema básico que presenta la fraseología a la lingüística informática haya quedado claro: ¿cómo reconocer una unidad fraseológica? En este momento nos encontramos ante el desafío de responder a esta pregunta sin recurrir a nuestro conocimiento previo. Ya no es posible decir que "*ojo de buey*" es un lexema plurimembre sin poder explicar las razones para que esto sea así: responder correctamente nos pone en la vía de poder modelizar computacionalmente este fascinante fenómeno.

4 Referencias

CHOMSKY, Noam (1981). *Lectures on Government and Binding*. Studies in Generative Grammar. Foris Publications, Dordrecht – Holland/Cinnaminson – Estados Unidos.

CORPAS PASTOR, Gloria (1996). *Manual de fraseología española*. Número 76 de la Biblioteca Románica Hispánica. Gredos, S.A., Madrid.

D'INTRONO, FRANCESCO (2002). *Niveles de complementación nominal*. Sin publicar.

FELLBAUM, CHRISTIANE (1998). *WordNet: An Electronic Lexical Database*. Cambridge, Massachusetts: MIT Press.

FONTENELLE, THIERRY (1997). *Turning a Bilingual Dictionary into a Lexical-Semantic Database*. Lexicographica: Series maior: 79. Editorial Niemeyer, Tübingen.

HAEGEMAN, LILIANE (1994). *Introduction to Government and Binding Theory*. Blackwell, Oxford.

WEHRLI, ÉRIC (1997). *L'analyse syntaxique des langues naturelles: problèmes et méthodes*. Masson: Paris.