

MERCÈ LORENTE CASAFONT

Universitat Pompeu Fabra. Barcelona

CONSTRUCCIONES VERBALES EN EL DISCURSO DE LA GENÓMICA. TIPOLOGÍA VERBAL Y DISCURSO CIENTÍFICO*

Abstract. Lorente Casafont Mercè, *Construcciones verbales en el discurso de la genómica. Tipología verbal y discurso científico* [Verbal constructions in the discourse about genomics. Verbal typology and scientific discourse]. *Studia Romanica Posnaniensia*, Adam Mickiewicz University Press, Poznań, vol. XXXI: 2004, pp. 353-359. ISBN 83-232-1353-4, ISSN 0137-2475.

Based on the linguistic approach of the Communicative Theory of Terminology, this paper, focuses on the semantic classification of verbs taken as a sample of the Human Genome discourse in the Spanish language, in order to describe specialized lexical units and codify verbal lemmas on computational dictionaries for specialized corpora processing and information retrieval. The semantic classification proposals of verbs come from Levin (1993) and the CLIPS project (ILC-CNR, Pisa).

1. INTRODUCCIÓN

La descripción de las unidades léxicas de categoría verbal ocupa parte importante de la bibliografía lingüística, tanto desde la sintaxis como desde la semántica. Los modelos lexicalistas han reforzado esta tendencia con trabajos sobre la correlación entre semántica y comportamiento sintáctico. En contraste, los estudios ubicados dentro de la terminología se han ocupado extensivamente de la descripción de sustantivos en detrimento de los verbos. Los motivos de esta falta de atención debemos buscarlos en la orientación aplicada de la materia y en la mejor acomodación de las unidades nominales al establecimiento de organizaciones conceptuales y definiciones (Lorente, 2002). Sin embargo, en la orientación comunicativa de la terminología, en la que nos situamos, se hace necesaria la descripción y la representación de las unidades verbales al mismo nivel que el resto de unidades léxicas susceptibles de transmitir conocimiento especializado.

*Este trabajo se inscribe en los proyectos TEXTERM (BFF2000-0841) y RICOTERM (TIC2000-1191), ambos financiados por el Ministerio de Ciencia y Tecnología.

En esta comunicación nos proponemos iniciar la descripción de los verbos contenidos en el discurso especializado sobre genoma humano en lengua española, centrándonos sobre todo en el análisis de frecuencias y en aspectos relacionados con la clasificación semántica de las unidades verbales.

2. MARCO TEÓRICO

Nos ubicamos dentro de la *Teoría Comunicativa de la Terminología* (Cabré 1999), a la que nos referiremos a partir de ahora como TCT, planteada como una aproximación lingüística a un objeto transdisciplinario, para describirlo como un fenómeno comunicativo complejo que se construye a partir de un conglomerado de elementos cognitivos, formales y pragmáticos. La TCT considera que las unidades de la terminología, fundamentalmente unidades léxicas que presentan un significado específico en el discurso de especialidad, son unidades lingüísticas, y como tales las analiza. Esto implica a) que pueden ser unidades con valor especializado, además de las unidades léxicas (las más prototípicas), tanto las unidades morfológicas o morfemas como las combinaciones léxicas o fraseología especializada; b) que el valor especializado se adquiere en contextos discursivos especializados, es decir en textos reales emitidos por especialistas que controlan el conocimiento de la especialidad.; y c) que la variación (conceptual y denominativa) es algo intrínseco a la terminología en tanto que expresión lingüística. En este marco el estudio de la terminología se aborda mediante modelos parciales diversos, no excluyentes, que nos puedan ofrecer claves para la descripción y la explicación del léxico y del discurso de especialidad dentro del lenguaje.

Desde el punto de vista de las aplicaciones lingüísticas, la TCT se interesa en la construcción de recursos lingüísticos de contenido especializado y de herramientas para la representación y la extracción del conocimiento. En esta línea, se han venido desarrollando aplicaciones como un corpus especializado textual multilingüe (Bach et. al., 1997), un extractor automático de terminología (Vivaldi, 2001), además de herramientas para el procesamiento del lenguaje natural y diccionarios computacionales. Actualmente, se está desarrollando un banco de conocimiento sobre genoma humano, constituido modularmente por un corpus textual escrito en inglés, español y catalán, una base de datos terminológica, una ontología y una base de datos documental y factográfica. Este recurso multiformato nos ofrece datos válidos para la descripción lingüística y sirve de banco de pruebas para la parte aplicada de nuestros proyectos, consistente en la creación de sistemas para la representación del conocimiento y la recuperación de información.

La línea de investigación que llevamos a cabo dentro de este marco se orienta hacia la descripción de las unidades predicativas y de la combinatoria léxica en el discurso de especialidad, para el establecimiento de generalizaciones y estrategias orientadas al diseño de aplicaciones. Los datos derivados de la descripción de textos

y de sus unidades en contexto nos permiten avanzar en la caracterización de la especificidad de la comunicación especializada y aportar datos complementarios a la descripción del lenguaje en general. La representación semántica y pragmática de las unidades que transmiten conocimiento especializado es básica para la extracción de información relevante, que supere las expectativas de precisión y recuperación que presentan los sistemas actuales.

Iniciamos la descripción de los verbos contenidos en el corpus textual de genoma humano, con las siguientes restricciones para esta primera fase del trabajo: 1) análisis cuantitativo de lemas y ocurrencias; 2) revisión y propuesta de clasificación semántica de lemas; 3) propuesta de correlación entre tipos semánticos y tipos terminológicos; y 4) análisis cuantitativo de los tipos de lemas resultantes.

Las propuestas de clasificación semántica que tenemos en cuenta aquí son la elaborada por Levin (1993) para el inglés y la usada en la codificación del proyecto CLIPS (*Corpora e Lessico Italiano Parlato e Scritto*) del Istituto di Linguistica Computazionale (CNR, Pisa)¹ para el italiano. Contrastaremos los resultados de la clasificación semántica con la propuesta de Lorente (2002), sobre clasificación de verbos del discurso de especialidad atendiendo a la transmisión del conocimiento especializado. Esto nos permitirá discernir sobre qué clases semánticas se corresponden fundamentalmente con los verbos que forman parte de unidades de conocimiento especializado del genoma.

Además de los resultados lingüísticos, que nos pueden orientar sobre las tendencias semánticas que siguen los predicados del discurso del genoma humano, el principal objetivo aplicado de este trabajo es una propuesta de codificación semántica de los verbos que se integre en el enriquecimiento del diccionario computacional del español que utilizamos para el procesamiento lingüístico del corpus.

3. EL CORPUS

El corpus que utilizamos en este trabajo está formado por la totalidad de documentos en lengua española seleccionados para el Corpus Textual de Genoma Humano, seleccionados por expertos del ámbito, que forma parte de la Base de Datos del Corpus Textual Especializado² del Instituto Universitario de Lingüística Aplicada de la Universidad Pompeu Fabra de Barcelona, que incluye también textos

¹ Este proyecto, dirigido por Nicoletta Calzolari y coordinado por Nilda Ruimy, tiene por objetivo la construcción de un diccionario computacional del italiano hablado y escrito, con codificación sintáctica y semántica. Las bases de la codificación se propusieron en el proyecto europeo SIMPLE, y en el ámbito de la semántica tiene en cuenta la propuesta de Levin (1993) e incluye la descripción de Extended Qualia, siguiendo el modelo de Pustejovsky (1995).

² La consulta del corpus se puede realizar por Internet, mediante Bwananet, herramienta de explotación del corpus construida sobre la base de Corpus Workbench (IMS-Stuttgart), y actualmente en pruebas en <http://brangaenc.upf.es/bwananet0/bwananet0a>.

de medicina, informática, derecho, economía, medioambiente, en español, catalán, inglés, francés y alemán, marcados estructuralmente y procesados lingüísticamente.

En concreto, se trata de un corpus escrito formado por 126 documentos, con un total de 975.441 palabras lematizadas, etiquetadas y desambiguadas. La interrogación sobre la categoría verbal nos da un resultado de 124.482 ocurrencias, que se corresponden a 2.516 lemas verbales distintos.

4. ANÁLISIS GENERAL

Sobre la lista de los lemas verbales se ha realizado una revisión para detectar verbos de uso exclusivo en el discurso de la genómica o de disciplinas cercanas: *alcoholar, aminoacilar, biopsiar, carboxilar, clonar, descarboxilar, fibrilar, fosforilar, hidroxilar, metastatizar, metilar, oxalatar, polimerizar, subclonar, transmembrantar, trifosfatar*. Antes de incorporarlos en el diccionario computacional, conviene asegurar que no se trate de hápax. Con este objetivo, uno de los primeros análisis cuantitativos se refiere a la frecuencia de cada uno de los lemas verbales. De los 2.516 lemas del corpus, 1.598 tan sólo presentan frecuencias que van de 1 a 9 y 680 disponen de 10 a 99 ocurrencias. Esto supone que 238 lemas (menos del 10%) tienen igual o más de 100 ocurrencias en el corpus, con un total 96.888 respecto del total de formas. Y, de los verbos considerados formal y semánticamente como “propios del ámbito”, solamente detectamos el lema *clonar*, con más de 100 ocurrencias.

Los restantes 237 lemas están presentes en diccionarios generales de la lengua. En su mayor parte polisémicos, podemos prever que, entre sus acepciones habrá sentidos especializados que, en gran parte, no estarán recogidos en estos diccionarios de referencia. Además, de acuerdo con los presupuestos de la TCT, consideramos que el conocimiento especializado se construye dinámicamente en contexto. Esto implica que una de las vías de nuestra investigación será establecer, mediante futuros análisis de concordancias, las evidencias formales que nos permitan identificar las ocurrencias de significado especializado frente a otras ocurrencias de sentido general o no marcado. Ahora, para este trabajo, nos detenemos en el análisis de los 238 lemas más frecuentes, para observar las clases semánticas a las que pertenecen.

5. CLASES SEMÁNTICAS

“If the distinctive behavior of verb classes with respect to diathesis alternations arises from their meaning, any class of verbs whose members pattern together with respect to diathesis alternations should be a semantically coherent class: its members should share at least some aspect of meaning” (Levin 1993: 14). De

acuerdo con esta idea, esta autora propone para los verbos del inglés una clasificación basada en 49 clases semánticas relacionadas con 8 alternancias de diátesis. Estos ocho patrones léxico-sintácticos son la alternancia de transitividad, la de argumentos con preposición, la de sujeto oblicuo, la de reflexivos, la de pasiva, la de sujetos posverbiales, más dos patrones más, uno con vinculaciones morfológicas y otro con complementos obligatorios. Ejemplos de las clases semánticas vinculadas son los verbos de comunicación, los de percepción, los de cambio de posesión, entre otras. Las subclases dentro de cada alternancia o dentro de clase se generan por el cruce de criterios sintácticos y semánticos.

Las ventajas de esta propuesta residen en la necesidad de localizar evidencias formales que validen las intuiciones en semántica, al tiempo que observamos algunos problemas para implementarla como sistema de codificación. Por un lado, la vinculación tan estrecha con la sintaxis del inglés, reduce las posibilidades de generalización, deseables sobre todo para el desarrollo de aplicaciones de acceso a la información multilingüe. Otra dificultad, relacionada con las necesidades de detección en contexto de sentidos especializados, está en el hecho de que al tratarse de una propuesta teórica y de alcance general, no está basada estrictamente en datos de corpus. Y, por último, la intersección entre las diversas subclases representa un problema teórico (Baker y Ruppenhofer, 2002) y un problema práctico para la codificación de los lemas si no se atiende a la sintaxis.

La clasificación de los lemas del genoma humano en español en las 49 clases semánticas de Levin (1993) ha dado como resultado que las clases más frecuentes son los verbos de creación (incluidos los de copia y los de creación de imagen), los de existencia y los de cambio de posesión. Sin embargo, tanto de la distribución de lemas como de la de ocurrencias para cada clase demuestran que hay una gran dispersión, motivada por la dificultad de codificación y por la falta de estratificación de 49 clases semánticas que presentan mucha intersección.

La codificación semántica del proyecto CLIPS (Ruimy et al., 2000) consiste básicamente en la identificación de plantillas (*templates*), que se corresponden con clases y subclases semánticas estructuradas jerárquicamente, para cada acepción de un lema. Además se incluye otra información semántica, como las *qualia* (Pustejovsky, 1995), la estructura argumental del predicado (número de argumentos, papeles semánticos y selección léxico-conceptual), la correlación con las variantes sintácticas y la relación semántica existente entre las diversas acepciones.

Se trata de un proyecto aplicado, de manera que los problemas de codificación se manifiestan prácticamente, con lo que la propuesta puede actualizarse. La complejidad informativa, aunque pueda ser parcialmente redundante, simplifica las decisiones del codificador. Otra de sus ventajas, en contraste con Levin (1993), es la autonomía entre la codificación sintáctica y la semántica, aunque la información de ambos módulos está vinculada. Pero, tal vez, uno de los elementos más positivos es la estructuración jerárquica de las clases semánticas, que facilita la codificación y la recuperación de información, con la posibilidad de establecer generalizaciones

sobre la semántica de los lemas codificados. El principal problema que presenta para los verbos de especialidad es que sus fuentes son lexicográficas, mientras que para la detección de acepciones especializadas conviene trabajar con ocurrencias de corpus textuales actualizados.

Los resultados de la clasificación de los lemas verbales del genoma bajo el modelo de CLIPS nos indican que la mayoría de lemas pertenecen al tipo eventivo de transición, mientras que el mayor número de ocurrencias corresponde a los verbos de estado. La organización entre clases y subclases nos ha permitido observar con mayor detalle qué subtipos semánticos son más frecuentes en cada grupo de clases verbales. Así por ejemplo, entre los verbos de estado, los que presentan mayor frecuencia de lemas y ocurrencias son los verbos existenciales, dentro de las acciones, los *purpose_act* y entre los actos de habla, los *reporting_events*.

Nuestra propuesta de clasificación de verbos del discurso especializado (Lorente 2002) tiene en cuenta su correlación con las UCE (Cabré 1999), es decir su capacidad para transmitir conocimiento especializado. Distinguimos entre verbos-término, que están morfológicamente vinculados al ámbito de especialidad (*clonar, descarboxilar, aminoacilar*); verbos fraseológicos, que conjuntamente con un término de categoría nominal, como mínimo, transmiten conocimiento especializado (*resecuenciar la muestra de ADN, transcribir la secuencia de ADN*); verbos conectores, cuyo significado no difiere del uso general pero que, combinados con términos, pueden formar parte de UCE (*una hebra de ADN es un polímero lineal*); y verbos discursivos, que aunque estructuran el texto no forman parte de secuencias con significado específico del dominio (*la figura 20.2 resume el ensamblaje de lambda*).

Como resultados provisionales, la correlación entre clases terminológicas y clases semánticas de los lemas del corpus de genoma se muestra en la siguiente tabla:

Lorente 2002	CLIPS
Verbos discursivos	<i>speech act, relational act, purpose act; modal events, perception, cognitive events; aspectual verbs</i>
Verbos conectores	<i>causes; relational state, constitutive state, stative possession</i>
Verbos fraseológicos	<i>existence, stative location; change of state, change of possession, change of motion; cause of change, creation</i>
Verbos-término	<i>copy creation (cause change)</i>

6. CONCLUSIONES

Presentamos en forma de síntesis las principales conclusiones a que se ha llegado, en el estado incipiente de esta investigación, sobre los verbos del discurso en español de genoma humano y sobre el establecimiento de clases semánticas.

- Los verbos-término, formalmente vinculados a la especialidad, tienen una presencia insignificante: 15 lemas sobre 2.516, y únicamente el lema *clonar* con más de 100 ocurrencias.
- Los verbos fraseológicos, que configuran unidades de conocimiento especializado, acogen el mayor número de lemas (46,21%) con un alto porcentaje de ocurrencias (33,76%). Su clasificación semántica da como resultado una alta concentración de estos verbos entre los verbos de cambio, causativos de cambio y acción, interpretable por el carácter ingeniero de la genómica frente a otros discursos de la biología.
- Las codificaciones semánticas complejas se muestran preferibles para la caracterización del discurso de especialidad, ya que la simple adscripción de los lemas a clases semánticas sólo nos ofrece una visión general de las tendencias de los predicados.
- Para la detección en contexto de sentidos especializados de la gran mayoría de verbos, que son de uso general, se precisa establecer estrategias de reconocimiento formal (patrones sintácticos o combinación léxica), mediante el análisis de concordancias, previas a la codificación de lemas.

REFERENCIAS BIBLIOGRÁFICAS

- Bach, C. et al. (1997), *El Corpus de l'IULA: descripció*. Papers de l'IULA, Informes 17.
- Baker C.F. y Ruppenhofer, J. (2002), *FrameNet's Frames vs. Levin's Verb Classes*, en Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society.
- Cabré, M.T. (1999), *La terminología: representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.
- Ruimy, N. et al. (2000), *CLIPS. Specifiche linguistiche e manuale di codifica. Livelo semantico*. Pisa: Istituto di Linguistica Computazionale, CNR.
- Levin, B. (1993), *English Verbs. Classes and Alternations*. Chicago: The University of Chicago Press.
- Lorente, M. (2002), *Verbos y discurso especializado*. Estudios de Lingüística Española (ELIES), 16 [http://elies.rediris.es]
- Pustejovsky, J. (1995), *The Generative Lexicon*. Cambridge: The MIT Press.
- Vivaldi, J. (2001), *Extracción de candidatos a términos mediante combinación de estrategias heterogéneas*. Tesis doctoral. Universitat Politècnica de Catalunya.