Lenguas Modernas 19 (1992), 167-172
Universidad de Chile

RESEÑAS

JOHN SINCLAIR:
Corpus, Concordance, Collocation
Oxford: Oxford University Press. 1991, 179 páginas

En los nueve capítulos de su libro, Sinclair reevalúa el lugar de la lingüística computacional y de la lingüística del corpus, prácticamente sinónimos en la perspectiva y contexto del autor, documentando adecuadamente su discusión en los distintos Apéndices del trabajo.

Del Glosario incluido al final del libro, extraemos, para mayor claridad, las definiciones ofrecidas por el autor de los tres conceptos que constituyen el eje de su presentación. Se entiende por 'corpus' "la colección de texto lingüístico de ocurrencia natural seleccionada para caracterizar un estado o variedad de una lengua"; 'colocación', a su vez, consiste en "la ocurrencia de dos o más palabras a corta distancia unas de otras en el texto"; el tercero de estos conceptos, 'concordancia', que no guarda relación con la concordancia gramatical descrita en términos tradicionales ('*concord*', en inglés), designa el "índice de las palabras en un texto" y, si bien similar a una selección de citas de una determinada frase o palabra, se diferencia de ésta por cuanto aquéllas son seleccionadas, en tanto que una concordancia, al ser generada por el computador, está constituida por todas las ocurrencias de la palabra o frase, en el corpus correspondiente, permitiendo así el acceso objetivo a cualquier patrón lingüístico que afecte al ítem investigado.

En las primeras páginas de su Introducción, Sinclair revisa someramente el desarrollo de la lingüística computacional, con particular referencia al ya bien conocido proyecto *Cobuild*, proyecto conjunto de la Editorial Collins y la Universidad de Birmingham, Inglaterra, para, a continuación, presentar su "afirmación de posición" (p. 3).

Esta afirmación, considerada provisoria, es la síntesis de las variables que, a juicio de Sinclair, deberían caracterizar los estudios descriptivos del lenguaje; básicamente, la aceptación y manipulación cuidadosa de la evidencia, donde dicha evidencia está constituida por muestras auténticas de lenguaje, las que a su vez son investigadas en unidades de dimensiones y características tales que recojan la correspondencia y la interdependencia entre forma y significado.

Serios cuestionamientos se derivan de esta postura: la evidencia aportada por el texto prevalecería sobre la intuición del investigador; el lenguaje auténtico, sobre los ejemplos inventados o adaptados; las prácticas lexicográficas tradicionales precisarían también de una revisión. Podemos así anticipar importantes consecuencias teórico-prácticas generales y, en nuestro caso particular, también pedagógicas. En última instancia, nos parece entender, el autor propicia una aproximación más bien inductiva a los estudios de lenguaje en la medida en que el computador, con las actuales y futuras bases de datos de millones de palabras de extensión y su vertiginoso procesamiento, permite un manejo de la información hasta ahora vedado a los seres humanos provistos de recursos y técnicas convencionales. Estos supuestos son reforzados a lo largo del desarrollo de este libro.

La creación y procesamiento de un corpus constituyen el tema de la siguiente sección, Capítulos 1 y 2 respectivamente. La recolección de un corpus supone, naturalmente, los correspondientes objetivos, criterios, técnicas y procedimientos. En términos generales, se sugiere que un corpus "debería ser tan grande cuanto sea posible y debería mantenerse en crecimiento" (p. 18) mediante distintos métodos aplicados a través de dos vías principales: alimentar el computador por medio de la conversión de lenguaje con técnicas convencionales de digitación o bien permitir que el computador capture muestras de lenguaje ya codificado electrónicamente (mensajes transmitidos por correo electrónico, registros electrónicos de lenguaje en los procesos de impresión, etc.) y por el más sofisticado proceso de conversión a través de *scanning* óptico.

Establecidos los criterios pertinentes (muestras de lenguaje de determinado tamaño o bien documentos completos; texto procesado por el empleo de otros códigos o texto libre de cualquier manipulación, la llamada 'política de texto simple'; determinación del tenor y modo del texto —formal o informal, escrito u oral transcrito—, entre otras especificaciones) podrán configurarse dos tipos de corpus: uno de descripción establecida, que incluye muestras de dimensión precisa clasificables en determinados géneros y relacionadas entre sí de manera conocida, en contraste con un corpus dinámico y flexible, que considere la "dimensión

histórica" del lenguaje, susceptible de ser incrementado y actualizado como también movilizado según los requerimientos del investigador. Sinclair designa estos dos tipos de corpus como 'de muestra' y 'monitor', respectivamente.

Además, este capítulo incluye una variedad de sugerencias prácticas para la creación de un corpus, información sobre el desarrollo de este campo en las últimas décadas e interesantes evaluaciones ("...existe una tentación de recolectar guiones cinematográficos y textos dramáticos [...], [ninguno de los cuales] refleja la conversación natural...", p. 18).

En lo sustancial, Sinclair desliga al lingüista del diseño del corpus, el cual, en cuanto pretende ser una cuenta del estado del uso de una lengua, debería estar sujeto más bien a la decisión de los cientistas sociales.

La siguiente cuestión es la forma de procesamiento de los textos incluidos en el corpus ya creado, donde, naturalmente, un problema básico es el de la recreación del texto dentro del computador. Se propone, a partir del registro de caracteres, una unidad de trabajo, denominada 'forma de la palabra' ("una sucesión ininterrumpida de letras", p. 28), que el usuario/lector tiende a vincular con conceptos en otro nivel de abstracción ('palabra', 'lexema', por ejemplo, en nuestro caso). Aquí, una vez más, queda de manifiesto la relativa contraposición de los datos registrados y procesados por el computador con los prejuicios del usuario/investigador, a partir de su propia intuición o de la conceptualización tomada de distintos modelos teóricos. "La relación entre datos y abstracciones [...] se transformará en un área mayor de investigación", dice Sinclair (p. 27).

Distintas técnicas de procesamiento son descritas en este capítulo, adecuadamente ejemplificadas en los correspondientes Apéndices. Se destacan las listas y perfiles de frecuencia y, principalmente, la concordancia, definida ahora como "una colección de las ocurrencias de una forma de la palabra, cada una en su propio entorno textual" (p. 32). El potencial descriptivo de las concordancias como asimismo otras cuestiones relacionadas son comentados por el autor en el capítulo siguiente, "La evidencia del uso".

Se sostiene que las principales fuentes de evidencia para la descripción tanto lexicográfica como gramatical están constituidas por otras descripciones, la introspección del investigador y la observación del lenguaje en uso. Sobre estas bases, con especial referencia a la lexicografía, Sinclair contrasta los resultados desplegados en diccionarios de reciente publicación con la información correspondiente obtenida por el procesamiento computacional de un corpus determinado.

Tal procesamiento implica al menos dos etapas previas: la determinación de una palabra que agrupe las distintas formas relacionadas y permita su registro en el diccionario ('lema', en la terminología de Sinclair) a fin de, a continuación, delimitar las concordancias respectivas.

El autor da cuenta de las dificultades tanto teóricas como prácticas que surgen en esta parte del procesamiento computacional ("...un lema no es obvio para el computador..." [p. 41]; "...cualquier método [de ordenación de citas] que se elija destaca algunos patrones [...] pero obscurece otros..." (p. 43), etc.). Sorteadas todas las dificultades, es posible contar con las concordancias requeridas que, exhaustivas o selectivas, contendrán la evidencia de los usos de un ítem determinado más allá de lo que las técnicas lexicográficas convencionales podrían haber discernido.

La comparación anunciada anteriormente se realiza a propósito del lema "decline". Se describe el riguroso análisis a que se ha sometido el lema en cuestión y el contraste en las diversas etapas de este análisis con los diccionarios mencionados. Previsiblemente, quedan de manifiesto importantes diferencias entre una y otra descripción. Por otra parte, el examen de la concordancia de "decline" revela también la correspondencia existente entre los dos significados atribuidos a dicho lema y las alternativas de transitividad verbal. Nuestro autor comenta entonces que "las distinciones gramaticales y léxicas pueden ser menores de lo que se concede normalmente" (p. 51).

Esta última afirmación constituye un planteamiento crucial en el discurso de Sinclair y a él se dedica el próximo capítulo: "Sentido y estructura en el léxico".

Con el objeto de fundamentar su hipótesis de la correlación entre sentido (aquí referido a "la contribución que una palabra puede hacer a un ítem léxico constituido por varias palabras", p. 53) y estructura, el autor comenta acabadamente la concordancia del lema "yield", registrado en el corpus central del proyecto Cobuild (ciento veinticinco ocurrencias en los más de siete millones de palabras de dicho corpus).

La evidencia estadística obtenida resulta abrumadora: el primer significado mayor del lema "yield" ("give way") es realizado estructuralmente por un verbo transitivo en treinta y tres casos; el segundo ("produce") será un sustantivo en treinta casos; se dan, por último, quince casos en que el tercer significado ("lead to") es un verbo transitivo. Otros significados menores corresponden a dieciséis ejemplos más. Todas estas cifras configuran un 70% de correlación entre sentido y estructura. El exhaustivo análisis de Sinclair

establece también que el 30% restante de ocurrencias de "yield", procesadas en su oportunidad como contra-ejemplos, es susceptible de ser explicado por otras razones: uso descuidado y desviación artística, por ejemplo.

Adicionalmente, la discusión de este estudio implica, al menos en sus aspectos metodológicos, una síntesis del aporte deductivo del investigador y de los datos procesados por el computador. Por último, suponiendo que los hechos aquí comentados representen una constante en el lenguaje, puede anticiparse una productiva fertilización cruzada entre los campos de la lexicografía y de los estudios de sintaxis. Humorísticamente, Sinclair pronostica que "el tradicional dominio de la sintaxis será invadido por hordas léxicas" (p. 65).

Se describe, en lo que sigue, un tercer estudio de concordancias. Se trata, en esta oportunidad, de ítemes calificados de "escasamente léxicos", vale decir, de palabras difíciles de aislar semánticamente. En torno al lema "set", se investigan distintas combinaciones de "set" + 'partícula'. Un hallazgo muy interesante reside en el hecho de que, si bien estadísticamente las posibilidades de ocurrencia de un verbo como "set off" son de aproximadamente una en siete millones, se encontraron setenta ejemplos de dicho verbo. Esto podría significar que, al no darse la distribución de palabras al azar que se atribuye en términos estadísticos a una muestra de estas dimensiones, tal distribución estaría determinada por otros principios. En estas circunstancias, entonces, se postula la existencia de una suerte de atracción recíproca entre ciertos ítemes léxicos, la que, a su vez, implicaría la existencia de patrones de coocurrencia susceptibles de ser identificados.

Evidencia adicional se encuentra en el estudio de la combinación "set + in", donde, una vez aislados y procesados los veintinueve ejemplos del verbo "set in", queda expuesta una serie de patrones regulares tanto estructurales como semánticos. Específicamente, "set in" tiende a aparecer en cláusulas cortas, y en posición final en la cláusula; por ora parte, los sujetos de estas cláusulas se refieren mayoritariamente a estados desagradables ("decadence", "prejudice", "anarchy", etc.). Regularidad similar es observable en los patrones de coocurrencia de "set" con otras partículas.

De esta manera se refuerza la hipótesis de la correlación entre sentido y estructura y, por otra parte, es introducido y avalado el concepto de patrones de selección léxica coordinada. En realidad, el título del Capítulo 5, recién comentado, "Palabras y frases", no transparenta la importancia y el sentido riguroso de sus contenidos.

El encuentro de léxico y gramática postulado en los capítulos precedentes puede equivaler prácticamente a una síntesis en el caso de ciertos ítemes, léxicos o gramaticales según se los mire, de los cuales ni el diccionario ni la gramática han conseguido dar una cuenta acabada, sostiene Sinclair.

A este respecto, el autor revisa las entradas de la palabra "of" en diccionarios, gramáticas pedagógicas y descriptivas. Tal información es evaluada como insatisfactoria en uno y otro ámbito. Una de las observaciones más significativas en este sentido apunta al hecho de que la ocurrencia de "of" está gobernada más bien por lo que precede que por lo que sigue, por el núcleo de un grupo nominal en los casos de postmodificación o el verbo inmediatamente anterior, por ejemplo, en tanto que las preposiciones propiamente tal tienden a integrar adverbiales dentro de cláusulas. En estas circunstancias, la hasta ahora preposición "of" podría constituir, por sí sola, una de varias 'clases de palabras' integradas por un solo miembro, configurándose de esta manera una perfecta correspondencia entre léxico y gramática.

Así no fuera más que por razones de frecuencia, cualquier observación que afecte a un ítem como "of" merece un cuidadoso examen; de hecho, "of" ocupa el segundo lugar entre las ciento trece palabras más frecuentes en un corpus de dieciocho millones. Tal volumen de datos ha determinado un método de procesamiento de concordancias parciales sucesivas que, hasta el momento de la publicación de este libro, había entregado elementos suficientes para una caracterización de "of" más elegante. Además de ocurrencias fuera de grupos nominales, las concordancias pertinentes registran abundante evidencia de la función de "of" en grupos nominales donde formalmente integra la frase preposicional en postmodificación; sin embargo, es la estructura de modificación postnominal la que porta la información imprescindible para el sentido de la frase, desplazando el rol de núcleo de la frase al sustantivo de la supuesta estructura de postmodificación. El acopio de datos y su acabada clasificación y descripción apoyan la importante argumentación de Sinclair en este capítulo.

Por otra parte, la descripción de la frase nominal aquí replanteada y la clasificación de los usos de "of" extraída de las concordancias permitirían la adecuada inclusión de este ítem tanto en la gramática como en el diccionario.

Las consideraciones en "La evaluación de instancias" (léase, la evaluación de concordancias), el capítulo siguiente, nos parecen centrales a la reflexión de Sinclair. Hasta este punto, se han establecido algunos

supuestos básicos para la postura del autor o, al menos, se han delimitado las áreas críticas en los estudios de lenguaje. Primero, las generalizaciones en dichos estudios deben estar relacionadas en forma sistemática con los datos de un corpus de dimensión adecuada; segundo, existe una vinculación estrecha entre gramática y vocabulario en el sentido de que la gramática de una lengua no está constituida únicamente por los patrones de los ítemes de vocabulario funcional sino que por los patrones de todo su vocabulario; tercero, las palabras tienden a tener más de un significado, sentido o uso, que no están distribuidos en forma uniforme en el texto.

En estos términos, el "dilema del gramático" se reduce a estudiar instancias reales pero atípicas o bien a inventar sus propias instancias. El aporte de la lingüística computacional residiría, aparentemente, en la detección de los escasos ejemplos típicos o en la determinación de tal tipicidad por cuanto "las generalizaciones gramaticales no descansan sobre un fundamento rígido sino en la acumulación de los patrones de cientos de palabras y frases individuales" (p. 100).

No se requeriría, por tanto, de la distinción categórica entre estructura abstracta y estructura real del lenguaje, implicada en dicotomías tan establecidas como 'lengua-habla' o 'competencia-actuación'. Sinclair propone más bien su propio concepto de 'estructura', equivalente a "ampliar el dominio de la sintaxis para incluir también la estructura léxica" y definido como "cualquier privilegio de ocurrencia de morfemas" (p. 104). A partir de este concepto, Sinclair hipotetiza que "la unidad subyacente a la composición está integrada por un complejo de sentido y estructura", lo cual es frecuentemente obscurecido "por las exigencias del texto" (p. 105).

Será tarea del lingüista la identificación de las "asociaciones regulares y típicas" que permitan precisar las citas que ilustren los distintos sentidos de una palabra.

Se comenta, a continuación, parte de la investigación realizada en torno a este problema, incluyendo una relativamente detallada descripción de su metodología computacional. Los hallazgos realizados sustentan un principio básico en la reflexión y praxis de Sinclair y sus colaboradores, la vinculación entre léxico y sintaxis y de ambos con la semántica.

Es éste un capítulo tan importante como complejo. Da la impresión de que algunas de las ideas expuestas habrían requerido de una discusión más detallada; en forma similar, algunos de los términos empleados admitirían cierto grado de negociación. De ahí el tono tentativo y la proliferación de citas.

El autor discute, a continuación, la relación entre léxico y semántica. Se percibe un traslapo entre estos dos ámbitos en cuanto, como comenta Sinclair, incluso en el Glosario, "los estudios léxicos apuntan a asociar los patrones formales con las distinciones de significado" (p. 154). La 'colocación', tema central del capítulo homónimo, constituye el concepto en torno al cual se busca establecer tal vinculación.

Se sostiene que el significado se realiza en el texto según dos principios mutuamente excluyentes, pero que, en su conjunto, dan cuenta de cómo surge dicho significado. Según el primero de ellos, se concibe el texto como una serie de selecciones gramaticalmente restringidas, posibles a continuación de cada palabra, frase o cláusula. El principio alterno implica que tales selecciones son limitadas, de manera que el hablante tiene a su disposición una cantidad de frases "semi-preconstruidas". El primero de estos principios, llamado de 'selección abierta', subyace a prácticamente todas las gramáticas conocidas. Al segundo, que tradicionalmente ha explicado fenómenos tales como las expresiones idiomáticas, los proverbios, los clichés, la terminología técnica, etc., se lo denomina, precisamente, 'principio de la expresión idiomática'. Por su supuesta atipicidad, las expresiones de esta naturaleza han recibido comparativamente escasa atención en los estudios del lenguaje.

Sinclair, apoyado en el estudio de textos extensos propio de la lingüística computacional, adopta una posición diferente. Destaca, con este objeto, algunos rasgos significativos de las expresiones idiomáticas (tales como extensión indeterminada, variación léxica y sintáctica interna, variación en el orden de las palabras, casos significativos de atracción léxica, correlación con determinados patrones gramaticales y ocurrencia en entornos semánticos determinados) que apoyan su valorización del lugar del principio de expresión idiomática en los estudios de texto.

El autor propone algunas generalizaciones tentativas a propósito del principio en cuestión. Por ejemplo: existiría una relación proporcionalmente inversa entre las frecuencias de las palabras y los significados independientes, vale decir, el significado de las palabras frecuentes es difícil de identificar. Podría, entonces, postularse una tendencia a la 'deslexicalización' y, en la medida en que los textos normales están constituidos por palabras frecuentes, dichos textos estarían más bien deslexicalizados y constituirían, principalmente, una aplicación del principio de la expresión idiomática.

Sinclair superpone al concepto de colocación, ahora entendido como una instancia del principio de la

expresión idiomática, una caracterización estadístico-computacional adecuada (colocación 'ascendente' y 'descendente') que permite la identificación de patrones colocacionales y describe un estudio realizado sobre la palabra "back", concluyendo que "toda la evidencia apunta a una rigidez de la fraseología, a pesar de una rica variación superficial" (p. 121).

Sinclair intitula el último capítulo "Palabras sobre las palabras", donde, aun cuando se propone aportar "a una mejor comprensión del lenguaje sobre el lenguaje", no se referirá al metalenguaje propiamente tal sino a la posibilidad de "articular una teoría de la reflexividad del lenguaje" (p. 123). Esto significa un cambio, según su propia declaración, con respecto a su posición original en relación con la lexicografía.

Tal cambio está asociado —parcialmente, suponemos— al trabajo de compilación del diccionario *Cobuild* por parte de Sinclair y sus colaboradores. Este diccionario, radicalmente diferente de otros diccionarios, incluso de publicación reciente, se caracteriza por explicar los significados de las palabras utilizando lo que sus compiladores describen como "una breve extensión del uso corriente de la lengua inglesa" (p. 136).

La significación de este recurso va más allá del mero intento de simplificar la tarea del usuario. Se postula que, entre otras consideraciones, explicaciones de esta naturaleza son susceptibles de ser asimiladas al repertorio lingüístico general del usuario y, por consiguiente, todas las posibilidades de lenguaje natural en cuanto a la formulación de inferencias e implicaciones le estarán también disponibles. Pero, por sobre todo, se atribuye a este tipo de aseveración léxica el valor de la paráfrasis, "uno de los más poderosos pero menos comprendidos rasgos del lenguaje natural" (p. 136); la posibilidad de replicar esta capacidad en una máquina establecería el vínculo con los investigadores de la inteligencia artificial.

La descripción de los recursos utilizados en la compilación del diccionario *Cobuild* y la explicación de su fundamentación son, como ha ocurrido en los capítulos anteriores, claramente ejemplificadas por el autor.

La significación de la postura teórica de Sinclair y la relevancia de su aplicación más significativa, el proyecto *Cobuild*, se transparenta en *Corpus, concordance, collocation*.

Su crítica a importantes supuestos y aplicaciones en el campo de los estudios del lenguaje, sólidamente fundamentada en los datos provistos por la lingüística computacional, constituye un aporte importante a la evaluación de la reflexión y praxis en dicho campo. Por otra parte, en vista del lugar que ocupan las ciencias de la computación con sus espectaculares logros en la particular versión de los estudios de lenguaje por parte de Sinclair, cabe anticipar un desarrollo igualmente importante de la lingüística del corpus así entendida.

La nueva descripción de la lengua inglesa que está en proceso de construcción, según se nos informa en este libro, tendrá, necesariamente, que encontrar su expresión en la investigación y aplicación en nuestro quehacer; la lingüística aplicada y la enseñanza de lenguas extranjeras, con sus respectivas metodologías, por ejemplo, podrían, desde ya, considerar las variadas sugerencias de investigación y revisión de prácticas establecidas que, explícita o implícitamente, abundan en *Corpus, concordance, collocation*.

En otra dirección, aun considerando que la lingüística del corpus aparece en este libro como de base más bien sociolingüística, el componente psicolingüístico inherente a los estudios discursivos sugiere perspectivas interesantes a los investigadores de la cognición. En forma similar, da la impresión de que el decisivo apoyo computacional en este campo permite la relación, en más de un sentido, con los estudios de inteligencia artificial.

Todo lo cual es anticipado o sugerido por Sinclair, pero, por sobre todo, descrito en términos de accesibilidad y aplicabilidad. Se trata, pues, de una obra y de un autor importantes en un ámbito en que las posibilidades de desarrollo y el volumen y calidad del trabajo por realizar son ilimitados.

PATRICIO NOVOA
UNIVERSIDAD DE CHILE