John Sinclair (1933-2007). In memorian

Aquilino Sánchez
(Universidad de Murcia)
<asanchez@um.es>

Conocí personalmente a J. Sinclair en 1983, en la Universidad de Murcia. Hacía un largo periplo por diversas universidades de Europa, América y Asia, de la mano de la editorial Collins, difundiendo las bondades del incipiente corpus *Cobuild* y las obras lexicográficas que de él derivarían. El principal bagaje que llevaba consigo era un disquete de 3,5", que contenía unos pocos ficheros con concordancias de algunas palabras inglesas, extraídas precisamente del corpus *Cobuild*.

Reconozco que ya existía por mi parte una marcada predisposición favorable al uso de los ordenadores en la investigación lingüística. Esa predisposición se había iniciado 10 años atrás en el tiempo, cuando cursaba un máster en *Applied Linguistics* en la universidad de Georgetown, Washington D.C. El tema, por tanto, encontraba ya en mí un terreno preparado y abonado. Hecha la exposición del proyecto por parte del profesor Sinclair, quedaba aún por superar el reto más importante: encontrar un ordenador en el que mostrar a los presentes los ficheros contenidos en el preciado disquete de 3,5". El único lugar en el que era posible visualizar los ficheros era el Centro de Procesamiento de Datos de la universidad, en una pequeña pantalla en blanco y negro, valiéndonos del MS DOS.

Aunque ya contábamos con varios precedentes en la utilización de los corpus para determinadas investigaciones (Kaeding, 1897, intentaba estudiar la distribución de las frecuencias de las letras y de las secuencias de éstas; Thorndike, 1921 y M. West, 1953, se proponían seleccionar las palabras más frecuentes con fines didácticos, etc. Véase Sánchez, 1995, McEnery, 1996), el tema distaba mucho aún de ser popular entre los lingüistas. Y menos aún de ser tomado como uno de los paradigmas fundamentales en la investigación lingüística. El entusiasmo y convencimiento de J. Sinclair eran patentes y contagiosos. La exposición bastó para darme el empujoncito decisivo y unirme al reducido grupo de entusiastas de los corpus. Muy pocos años después inicié yo mismo un largo proceso para elaborar un corpus similar al *Cobuild*, pero aplicado al español.

En 1995, tras haber logrado la financiación necesaria, se finalizó el proyecto *Cumbre*, corpus representativo del español, de 20 millones de palabras. Probablemente, tal proyecto no habría nacido si no hubiera mediado este primer y decisivo encuentro con J. Sinclair.

Escocés de origen (John había nacido en Edimburgo, en cuya universidad inició su docencia), accedió muy joven –con 31 años- a la cátedra de 'Modern English Language', en la universidad de Birmingham. Sus primeras publicaciones están relacionadas con el análisis del discurso (*Towards an Analysis of Discourse*, 1975, junto con M. Coulthard) y con sus trabajos sobre las colocaciones. Sin embargo, fue la lexicografía la que le llevó más directamente a lo que luego vino en llamarse la lingüística del corpus. En la década de los setenta se unió el grupo de trabajo que la editorial Collins formó para elaborar un diccionario novedoso y diferentes obras referidas a la enseñanza del inglés. La fuente de datos para ese diccionario sería precisamente el *Birmingham Corpus*, de cuya dirección fue responsable J. Sinclair. Elaborar un corpus en aquellos años era una tarea ímproba, sobre todo porque era necesario contar con una abultada aportación económica (sólo el escáner usado para escanear los textos del corpus *Cobuild* costó unos 100.000 euros). El diccionario, *Collins Cobuild English Language Dictionary*, se publicó finalmente en 1987 y marcó un camino que el trabajo lexicográfico ya ha asumido como propio: el uso de grandes recopilaciones textuales, debidamente planificadas, ordenadas y codificadas, como el instrumento ideal del lexicógrafo para detectar voces y acepciones nuevas, o incluso para identificar usos que la lexicografía tradicional no había recogido. Naturalmente, la utilización de los corpus fue posible porque había surgido otra herramienta que facilitaba su aprovechamiento: el ordenador y su capacidad para procesar la palabra electrónica.

El profesor Sinclair ha sido fundamental y quizás decisivo en la creación y consolidación de un método de trabajo al que los estudios lingüísticos no estaban habituados: la perspectiva empírica. Si la teoría generativista de Chomsky había centrado la atención en la formulación de una teoría de base cognitiva e hipotético-deductiva, cuya finalidad es explicar racionalmente cómo la mente humana genera y procesa el lenguaje, la perspectiva empírica se centra en la observación del output lingüístico para determinar lo que es normal en la lengua. La primera se basa en la introspección y aduce unos pocos ejemplos, a menudo elaborados *ad hoc*; la segunda critica la pobreza de esos datos y fundamenta sus conclusiones en miles o decenas de miles de ejemplos extraídos del uso real, tal cual se ofrecen en los corpus. Nunca antes del advenimiento de la lingüística del corpus –con la ayuda de los ordenadores- un lingüista había sido capaz de tener ante sí tal volumen de evidencias sobre el uso de una lengua concreta. En cuestión de segundos, y a partir de un corpus de 10, 20, 100 ó 200 millones de palabras (debe recordarse que 100 millones de palabras equivalen a unos 1.000 libros de 250 páginas cada uno), el investigador puede acceder a cientos o miles de ejemplos que dan fe de cómo los hablantes de una lengua expresan un determinado

mensaje, o de cómo una determinada forma o lema se comporta en el uso lingüístico. Como consecuencia de este nuevo método de trabajo, la manera de estudiar las lenguas ha experimentado un cambio radical en los últimos 15 años. Además, pocas disciplinas dentro de la lingüística han tenido un crecimiento tan ostensible y en tan corto periodo de tiempo como la *lingüística del corpus*. Las listas de correos, los foros y las reuniones científicas o congresos centrados en los corpus y en sus aplicaciones llenan las agendas de los especialistas, no mes tras mes, sino semana tras semana, casi día tras día.

Estudiar el lenguaje implica estudiar las relaciones entre forma y significado. Puede concluirse, pues, que el estudio de las formas conducirá a una mejor comprensión de cómo se conforma, se codifica y se transmite el significado. La gran ventaja de los corpus es que hacen posible que el lingüista no reduzca su trabajo investigador a la introspección o a lo observado en la producción lingüística de unos pocos hablantes. La ampliación de las muestras de lengua a miles o millones de hablantes permite afianzar determinadas conclusiones, que de otra manera serían tildadas de subjetivas.

J. Sinclair no muestra apego a la creación de nuevas teorías lingüísticas. Prefiere los hechos observables para construir sobre ellos. Si nos atenemos a lo que confesaba en el prólogo publicado en el mismo año de su muerte (Sinclair, 2007), su mayor preocupación era el significado tal cual se podía detectar en los textos y la manera como éste se hermanaba con la formulación más tradicional de la gramática:

> 'I worked for many years under the illusion that all the valuable meanings in sentences were conjured up by the syntax, and it took some years more before I realised that steps towards formalism in language description are steps away from meaning' (Sinclair, 2007).

Y en realidad Sinclair no se adscribe a corrientes lingüísticas asentadas. Los 'culpables' de tal actitud son los corpus, en cuanto suministradores de datos que se resisten a asignaciones preconcebidas o a encuadres ideológicos prediseñados. Sus obras tampoco constituyen tratados contra una u otra determinada teoría lingüística. Solamente introducen cuñas de evidencias que parecen exigir planteamientos nuevos y revisión de algunas 'verdades'. Como él mismo afirma,

> 'the theories available to me did not alert me at all to the strongly recurrent patterns found in a corpus nor explained them.'

Eso es lo que hace que Sinclair se posicione frente a las diferentes teorías lingüísticas 'with increasing suspicion' (Sinclair, 2007). John Sinclair prefiere actuar como testigo de lo que ocurre y observa en el uso lingüístico. Y a partir de sus observaciones intenta llegar a conclusiones fundamentadas, aunque para ello tenga que salir de los caminos habituales en lingüística, o de las directrices de teorías consolidadas, o de enfoques establecidos y promovidos por reconocidos 'gurús' de las ciencias del lenguaje.

Las dos áreas en las que se percibe con mayor claridad la influencia de Sinclair son, a mi entender, la lexicografía y la lingüística del corpus, y dentro de éste ámbito, la cuestión de las colocaciones y su función en la conformación del significado y en la desambiguación semántica de cada una de sus ocurrencias, cuando los significados posibles son varios.

En cuanto a la lexicografía, las innovaciones promovidas por Sinclair han quedado reflejadas en el diccionario *Collins Cobuild English Language Dictionary (CCELD)*. La más conocida –no la única- de tales innovaciones es la formulación de las definiciones, como puede comprobarse en estos ejemplos:

Milk: **Milk** is the white liquid produced by cows, etc. *(+ example)*

Light: **Light** is the thing that lets you see things, and that comes from the sun, moon… *(+ example)*

Diver: A **diver** is a person that works or explores under water… *(+ example)*

Jest: If you do something **in jest**, you do not mean it seriously, but want to be amusing.

Formulaciones que contrastan, a su vez, con el tipo de definiciones tradicionales, tal cual se percibe en el diccionario *Merrian-Webster* (2000):

Milk: a white or yellowish fluid secreted by the mammary glands of female mammals…

Light: something that makes vision possible.

Diver: a person who stays under water (as in salvage work) for long periods by having air supplied from the surface.

Jest: an act intended to provoke laughter.

Quien consulta un diccionario, no suele reparar en innovaciones: por lo general, carece de las referencias necesarias para ello. Los usuarios tienden más bien a *creer* en lo que dicen los diccionarios. De ahí que pasen desapercibidos rasgos tan importantes como el orden en que se presentan las acepciones (atendiendo a su aparición histórica o a su frecuencia); el hecho de que las definiciones vayan seguidas de ejemplos ilustrativos o no (para avalar la veracidad de lo definido); la presencia o ausencia de todos los significados de la voz definida (un corpus representativo permite detectar los significados con mayor facilidad y fiabilidad); o la comprobación de si la manera como se formula la definición induce por sí misma a encuadrar el uso de la voz dentro de su contexto natural. Las definiciones del *CCELD* incluyen precisamente las innovaciones reseñadas: (i) se ordenan atendiendo a la frecuencia de uso (algo que sólo puede llevarse con la información suministrada por un corpus representativo); (ii) van seguidas siempre de un ejemplo de uso real (extraído del corpus); (iii) se recogen todos los significados registrados en el corpus (aunque es preciso reconocer que los 8 millones de palabras de la primera versión del corpus *Cobuild* no garantizan el deseado nivel de exhaustividad);

(iv) las definiciones se formulan no solamente de manera más sencilla y transparente (a veces imitando el modelo de definición habitual en libros infantiles), sino con oraciones completas que permiten situar la palabra definida dentro del contexto oracional que le es más propio.

Desde su obra en el 75, sobre el análisis del discurso, Sinclair llega pronto a la convicción de que la palabra, el lema, tal cual se presenta en los diccionarios (como elemento aislado del contexto en que es usado) no se corresponde con la función detectada en la realidad del uso: las palabras cobran sentido pleno solamente en el discurso. Cabe concluir, por tanto, que el significado no reside en la palabra, sino en el discurso o en el texto, lo que equivale a afirmar que el significado reside en unidades más amplias que la simple palabra (Almela, 2006), precisamente porque sólo en esas unidades actúa el contexto. De ahí que los textos nunca sean ambiguos, mientras que las palabras sí pueden serlo –y de hecho muchas lo son. Si el significado estuviera incrustado en las palabras, la ambigüedad léxica no existiría, ya que cada término conllevaría un solo significado. El problema se pone claramente de manifiesto en los términos polisémicos; en estos casos solamente el contexto permite que el oyente seleccione en cada caso el significado o acepción pretendido por el hablante.

La función desambiguadora del contexto no se hermana bien con alguno de los principios seguidos en lexicografía, como sería el de la 'sustituibilidad' (la definición de una palabra debe ser sustituible por la palabra definida). Sinclair cuestiona esta norma porque las palabras no cobran sentido hasta que no están enmarcadas en el texto en que se usan, y en consecuencia no cabe la posibilidad de pretender que una definición pueda ser reemplazada por la palabra definida, puesto que las palabras en un diccionario se presentan, típicamente, descontextualizadas. Partiendo de esta premisa, la manera de definir propuesta por Sinclair en el diccionario *Cobuild* cobra sentido: el objetivo es formular la definición de manera que sugiera o implique el contexto de uso real del término definido. Quizás es preciso admitir que no siempre se consigue este objetivo en el diccionario, pero el método apunta en la dirección adecuada. 'If you do something **in jest**, you do not mean it seriously…' incluye una estructura habitual de uso (*to do sth in jest*), de igual manera que 'A **diver** is a person that works or explores under water' (*a diver is a person*…) repite una estructura habitual en el uso de 'diver'. Por el contrario, el diccionario *Webster* presenta definiciones al estilo de la lexicografía tradicional ('**Jest**: an act intended to provoke laughter'), propiciando, eso sí, la sustitución plena de la definición por la palabra definida. Así, la frase:

*It was obviously made **in jest** and yes it had an element of humor*.

Puede ser sustituida por

It was obviously made **(in jest)** [as *an act intended to provoke laughter]* and yes it had an element of humor.

El proyecto de Sinclair tiene sus riesgos; y el más importante se refiere a los usuarios de diccionarios: ¿están habituados los usuarios a este tipo de definiciones? El fracaso parcial del proyecto *Cobuild* (la editorial clausuró el proyecto a finales de los noventa) no fue un elemento alentador.

El segundo aspecto en el que ha destacado el profesor Sinclair se enmarca ya más directamente dentro de la lingüística del corpus. La obra clave en la trayectoria investigadora de Sinclair es para muchos *Corpus, concordance, collocation*, publicada en 1991. Probablemente ha sido también su obra más influyente y la que para algunos ha constituido casi el libro de cabecera. Sinclair enuncia lo que suele llamarse 'the open choice principle' y 'the idiom principle' y aporta un sinfín de datos que avalan su tesis. El 'open choice principle' predice la selección de palabras individuales atendiendo a las restricciones de combinación impuestas por la gramática de la lengua. Es pues, un principio que funciona *de arriba abajo*, justo al contrario que el 'idiom principle', que actúa *de abajo arriba* y es de naturaleza léxica. El primero no es suficiente para producir discurso por parte del hablante. La potencialidad comunicativa se complementa con el 'idiom principle'. En estilo llano y fácilmente comprensible –cosa infrecuente en el mundo académico- argumenta que la *potencialidad creativa* en el uso de la lengua es bastante baja, a menudo prácticamente nula (afirmación poco grata para los generativistas). Por el contrario, lo más frecuente es expresarnos echando mano de recursos ya preestablecidos y almacenados en nuestra mente. La producción lingüística de los hablantes está plagada de 'trozos de lengua' (*chunks of language*), segmentos de oraciones prefabricadas, que se adquieren como conjuntos ya disponibles previamente y así se almacenan en nuestra memoria; sólo recurrimos a la gramática cuando estos elementos 'pre-establecidos y almacenados' no son suficientes para formular lo que queremos decir. En realidad, el hecho no es despreciable ni debe minusvalorarse: el recurso a los 'segmentos de lengua' pre-establecidos y almacenados en la mente favorece la fluidez de la comunicación.

Considerado desde otro ángulo, 'the idiom principle' no hace sino confirmar la intuición que Sinclair ya había expuesto años antes: algunas palabras tienden a ir acompañadas de otras palabras específicas, eje central en torno al cual gira el concepto de *colocación*. Actualmente nadie parece dudar de esta verdad, que podría formularse con la frase de que 'las palabras tienen *amigos* (otras palabras), siendo variable la intensidad de tal *amistad* (escala de colocación variable)'. Según Sinclair (1991:112),

> 'the idiom principle is as important as grammar in the explanation of how meaning arises in text.'

A principios de los noventa, Sinclair se retiró de la vida académica. Pero tal retirada fue solamente parcial. En 1995 se instaló en Italia, en la Toscana tan querida por muchos anglosajones, donde fundó el *Tuscan Word Centre.* Desde esta sede privilegiada

continuó con su labor docente, organizando cursos en torno a la lingüística del corpus y promoviendo el estudio científico del lenguaje. Quienes han asistido a tales cursos son unánimes en reconocer la generosidad intelectual de Sinclair, quien supo rodearse de jóvenes investigadores ávidos de saber y nunca escatimó sus conocimientos e intuiciones para incentivarlos e introducirlos en la investigación mediante corpus. Para ello se valía de sus cualidades de liderazgo y de persuasión; a ello contribuían su claridad de ideas, su visión de los proyectos de futuro y la sencillez y transparencia con que exponía las ideas aparentemente más intricadas. Como otros ya han afirmado en sus escritos laudatorios, Sinclair ha sido una figura de corte académico indiscutible y el verdadero padre de los corpus lingüísticos y de la lingüística del corpus. Así ha pasado a la historia y, probablemente, así será recordado en la historia de la lingüística.

REFERENCIAS

Almela, M. 2006. *From Words to Lexical Units: A Corpus-Driven Account of Collocation and Idiomatic Patterning in English and Spanish*. Peter Lang: Frankfurt am Main Berlin.

McEnery, T. and A. Wilson. 1996. *Corpus Linguistics*, Edinburgh, Edinburgh UP.

Sánchez, A. et al. 1995. *CUMBRE. Corpus lingüístico del español contemporáneo. Fundamentos, metodología y análisis*, Madrid, SGEL s.a.

Sinclair, J. 1975. (with R M Coulthard) *Towards an Analysis of Discourse: the English Used by Teachers and Pupils*, Oxford: Oxford University Press.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. 2004. *Trust the Text: Language Corpus and Discourse*. London: Routledge.

Sinclair, J. 2007. Preface, *International Journal of Corpus Linguistics*, 12, 2, p. 156.

PUBLICACIONES MÁS SOBRESALIENTES DEL PROF. JOHN SINCLAIR

Sinclair, J. 1966. 'Taking a Poem to Pieces', in *Essays on Style and Language* ed. Fowler R, London: Routledge, 68-81; reprinted in *Linguistics and Literary Style*, ed. Freeman D M, New York: Bolt, Rinehart and Winston.

Sinclair, J. 1972. *A Course in Spoken English*: *Grammar*, Oxford: Oxford University Press.

Sinclair, J. 1974. 'English Lexical Collocations', *Cahiers de Lexicologie*, Paris: Institut des Professeurs de Français a l'Étranger.

Sinclair, J. 1975. (with R M Coulthard) *Towards an Analysis of Discourse: the English Used by Teachers and Pupils*, Oxford: Oxford University Press.

Sinclair, J. 1980. 'Computational Text Analysis at the University of Birmingham', in Johansson, S (ed.) *Newsletter of the International Computer Archive of Modem English*, Bergen: The Norwegian Computing Centre for the Humanities, 13-16.

Sinclair, J. 1982. (with D. Brazil) *Teacher Talk*, Oxford: Oxford University Press, first part reprinted as J. Sinclair *The Structure of Teacher Talk*, *Discourse Analysis Monographs* No. 15, Birmingham: English Language Research, University of Birmingham, 1990.

Sinclair, J. 1982. 'Reflections on Computer Corpora in English Language Research', in *Computer Corpora in English Language Research*, Johansson S, (ed.) Bergen.

Sinclair, J. 1984. 'Lexicography as an Academic Subject', in Hartmann, R. K. K. (ed.) *LEXeter 83 Proceedings*, Lexicographica Series Maior No. 2, Tubingen: Max Niemeyer Verlag, 3-12.

Sinclair, J. 1985. 'Lexicographic Evidence', in Ilson, R. (ed.) *Dictionaries, Lexicography and Language Learning*, ELT Documents 120, Pergamon, 81-94.

Sinclair, J. (Ed.). 1987. *Looking up. An Account of the Cobuild Project in Lexical Computing*, Collins, London.

Sinclair, J. 1987. (ed.) *Looking Up*, London: HarperCollins.

Sinclair, J. 1987. *Collins COBUILD English Language Dictionary*, (1st edition).

Sinclair, J. 1987. 'Grammar in the Dictionary', in Sinclair, J M (ed.) *Looking Up*, London: HarperCollins, 104-15.

Sinclair, J. 1987. 'The Nature of the Evidence', in Sinclair, J M (ed.) *Looking Up*, London: HarperCollins, 150-59.

Sinclair, J. 1988. 'Sense and Structure in Lexis', in Benson, J., Cummings, M. and Greaves, W. (eds.) *Linguistics in a Systematic Perspective*, Amsterdam: John Benjamins, 73-97.

Sinclair, J. 1990. *Collins COBUILD English Grammar.* London: HarperCollins.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. 1992. 'The Automatic Analysis of Corpora', in Svartvik J (ed.) *Directions in Corpus Linguistics*, Proceedings of Nobel Symposium 82, Stockholm 4-8 August 1991, Berlin and New York: Mouton de Gruyter, 379-97.

Sinclair, J. 1993. (with M. Hoey and G. Fox) (eds.) *Techniques of Description: Spoken and Written Discourse, a Festschrift for Malcolm Coulthard*, London: Routledge.

Sinclair, J. 1994 (ed. with M. Hoelter and C. Peters), *The Language of Definition: The Formalization of Dictionary Definitions for Natural Language Processing*, vol. 7 in the series *Studies in Machine Translation and Natural Language Processing,* Brussels: The European Commission.

Sinclair, J. 1996. 'The Empty Lexicon', *International Journal of Corpus Linguistics*, 1,1; 99-119.

Sinclair, J. 1996. 'The Search for Units of Meaning', *TEXTUS*, 9, 1, Special Volume Merlini L. and Sinclair J. (eds.) *Lessico e Morfologia*, 75-106; reprinted in Corpas G. (ed.) *Las Lenguas de Europa: Estudios de Fraseologia, Fraseografia y Traduccion*; Granada: Editorial Comares, SL, 7-38.

Sinclair, J. 1999. 'The Lexical Item', in Weigand, E. (ed.) *Contrastive Lexical Semantics* Amsterdam/Philadelphia: John Benjamins, Vol. 17 of the series *Current Issues in Linguistic Theory*, 1-24.

Sinclair, J. 2001. Preface. In *Small Corpus Studies and ELT*, eds. Mohsen Ghadessy, Alex Henry and Robert L. Roseberry, vii-xv. Amsterdam/Philadelphia: John Benjamins.

Sinclair, J. 2003. Corpora for Lexicography. In *A Practical Guide to Lexicography*, ed. P. Van Sterkenberg. Amsterdam: John Benjamins.

Sinclair, J. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Sinclair, J. 2007. Preface, *International Journal of Corpus Linguistics*, 12, 2, p. 156.

Richard M. Hogg Prize

ISLE intends to offer an annual Richard M. Hogg Prize for a paper on any research-related topic in English language or English linguistics.

Richard Hogg

Richard Hogg was Smith Professor of English Language and Medieval Literature at the University of Manchester from 1980 until his death in 2007. He was the General Editor of *The Cambridge History of the English Language* (6 vols, 1992-2001), one of the founding editors of the journal *English Language and Linguistics*, and well known for his work on Old English, on phonology, and on English dialects. Click here to read an obituary by Nigel Vincent in *The Guardian*. A list of Professor Hogg's publications is accessible from this database at the University of Manchester. It is hoped that his *History of English dialectology* and *Grammar of Old English*, 2, *Morphology* will be completed by friends and colleagues.

Eligibility

The Prize will be awarded in open competition. The competition is open to any individual who is both:

1. an early-career scholar, defined as a registered student not yet in possession of a doctoral degree, or a post-doctoral scholar within two years of the award of the doctorate at the time of submission; and
2. a member of the Society (membership can be applied for at the time of submission).

It is expected that most candidates will be students on a doctoral degree programme (PhD) or recent graduates of one, but undergraduates and master's students are not precluded from submitting a paper. Joint or multiple authorship is acceptable so long as all authors meet the two conditions above. Authors should submit a letter from their supervisor, or from a person of similar standing, attesting to their status and that the submission is their own work.

The paper

Candidates may write on any research-related topic in English language or English linguistics. In awarding the prize the committee will take into consideration the originality of the submitted paper and the theoretical and/or empirical contribution it makes to the discipline.

<http://www.englang.ed.ac.uk/isle/richard-hogg-price-html>