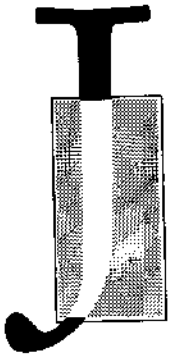
ENTREVISTA A JOHN SINCLAIR, EDITOR EN CAP DEL COLLINS COBUILD ENGLISH LANGUAGE DICTIONARY



ohn M. Sinclair és professor del Departament d'anglès modern de la Universitat de Birmingham des de 1965. Les àrees de recerca més importants en què ha treballat són el discurs oral i escrit i la lingüística computacional, amb atenció especial a l'estudi de textos. És el director del projecte COBUILD des del seu inici i l'editor principal de les publicacions del COBUILD. D'entre la seva extensa bibliografia, volem destacar per la seva importància el llibre que descriu el projecte COBUILD (Sinclair, J. (ed.) *Looking up. An account of the COBUILD Project in lexical computing*. Londres, Collins 1987), el seu darrer llibre sobre corpus (Sinclair, J. *Corpus, concordance, collocation*. Oxford, Oxford University Press 1991) i l'obra encara en premsa sobre lèxic (Sinclair, J. et al. *English word in use*. Londres, Collins).

P: Professor Sinclair, podria explicar-nos com va néixer el projecte COBUILD?

El projecte va néixer cap al final dels anys setanta amb un conveni de l'editorial Collins i la Universitat de Birmingham.

L'editorial Collins tenia una gran reputació en lexicografia bilingüe i monolingüe, sobretot de l'anglès, i estava molt interessada a elaborar altres productes i obrir

nous mercats. La Universitat de Birmingham feia aproximadament uns 15 anys que desenvolupava mètodes per crear un corpus en suport magnètic. Això passava cap a l'any 1962. Així va néixer l'acord de desenvolupar sistemes informàtics i software per fer estudis de vocabulari. I jo em vaig fer càrrec de materializar aquest acord.

En aquell moment el corpus de què disposàvem era de mots simples. Amb M. Halliday, em vaig adonar que calia disposar de textos sencers organitzats en un gran corpus per poder dur a terme amb rigor estudis sobre el llenguatge i el vocabulari; que calia crear un corpus d'almenys 5 milions de mots, o probablement encara més ampli.

Cal que aclareixi que el meu interès, de fet, no és la lexicografia pròpiament dita, sinó els contextos (*collocations*) i, per fer estudis sobre contextos, necessito una gran quantitat de text, ja que la freqüència d'aparició d'un mot no és gaire alta i fins que no es disposa d'unes 10, 15 o 20 ocurrències del mateix mot no es pot considerar que un context està suficientment establert.

Tornant al COBUILD, ens vam trobar que, d'una banda, nosaltres necessitàvem software i material en suport magnètic per a les nostres recerques, i que, d'altra banda, Collins volia un diccionari. I així, el 1980 la Universitat de Birmingham va signar un primer contracte amb Collins, de 5 o 6 anys de durada, amb el compromís de produir un diccionari a partir d'un corpus, usant els mètodes que havíem desenvolupat en el meu equip. I d'aleshores ençà la col·laboració no s'ha aturat.

El primer gran projecte que va sortir del corpus va ser el diccionari COBUILD, que ha estat amplament acceptat arreu i ha rebut molt bones crítiques. Per a l'elaboració del diccionari, partíem de les idees tradicionals de la lexicografia, però el fet de partir d'un corpus ens va fer canviar de plantejaments. El diccionari COBUILD és de fet un producte molt original, per moltes raons, gairebé totes elles lligades al fet que ha estat construït a partir d'un corpus textual. He analitzat la lexicografia dels darrers 20 anys i puc dir que el COBUILD és molt més modern que els altres diccionaris, i també molt diferent. En molts aspectes. Posem alguns casos, els exemples del COBUILD vénen dels textos i les definicions surten dels contextos. La fraseologia que conté és molta i molt important. És en gairebé tots els aspectes que el COBUILD és diferent dels diccionaris existents. I aquesta especificitat rau en el fet que prové d'un corpus, d'un corpus de textos que fa evidents els fets del llenguatge.

Quan es va acabar el període del primer contracte (1986-1987) vam constatar que tant la Universitat de Birmingham com l'editorial Collins havíem quedat satisfets del treball conjunt i vam crear una nova societat en col·laboració. Disposàvem aleshores ja d'un equip d'especialistes. Aquesta nova associació entre la Universitat de Birmingham i la casa Collins va néixer per produir altres tipus de treballs, a més dels lexicogràfics, com l'elaboració d'una gramàtica i un manual d'ús de la llengua.

De fet ara ja existeixen diverses publicacions que hem fet en cooperació. Aquestes noves obres neixen amb el propòsit fonamental de proporcionar als usuaris els beneficis

d'haver usat un corpus. Els productes que hem publicat ja són força nombrosos: tenim el *Cobuild English Course*, diverses obres breus destinades sobretot a la correcció (*sillabus*) i basades en la freqüència d'aparició de determinats fenòmens (la formació de paraules, les preposicions, els determinants, les paraules que es poden confondre, etc.), dues gramàtiques (una de més completa, l'*English Grammar*, i una altra més reduïda per als estudiants, l'*Student's Grammar*) i 6 diccionaris: el diccionari general *English Language Dictionary*, l'*Student's Diccionary*, el *Dictionary of Phrasal Verbs*, el diccionari reduït *Essential English Dictionary*, el *BBC English Dictionary* (especialment dedicat al llenguatge parlat i extret sobretot dels enregistraments de la BBC) i un diccionari escolar, que encara no s'ha publicat.

El tercer tema en què hem treballat és el de l'*English Usage Book*, que es va publicar l'any passat.

Ara treballem en la introducció d'elements gramaticals en els textos mitjançant l'etiquetatge dels mots (*tagging*) i l'ús d'analitzadors morfològics i sintàctics dels textos. Aquest sistema ens permetrà recuperar després determinades estructures a partir de la selecció de categories. Ho fem amb un software procedent de la Universitat de Helsinki.

El corpus, que al començament tenia 5 milions de mots, s'ha anat ampliant amb l'escaneritzacio de textos i amb materials de diversos estudiants. Actualment el corpus conté 120 milions de mots i és molt flexible i de fàcil accés.

Avui dia un dels propòsits del projecte COBUILD és de connectar amb una màquina que permeti seleccionar i respondre preguntes sobre 120 milions de mots de diverses procedències. Com que els textos estan classificats de diverses maneres, l'usuari pot seleccionar perfectament diversos materials a partir de diferents criteris: que procedeixein d'una font o d'una altra, escrits o orals, d'una varietat d'anglès o d'una altra, etc.

P: Parli breument dels aspectes tècnics del projecte COBUILD.

Poques coses puc dir en profunditat sobre els aspectes més tècnics relacionats amb la informàtica, perquè no en sóc especialista. Però puc dir que quan vam començar el projecte, no hi havia especialistes en corpus lingüístics i vam haver-ne de formar a partir de col·legues especialistes en computació.

En l'etapa inicial del projecte vam utilitzar l'ordinador central de la Universitat, que precisament el van comprar per al projecte. Era una màquina Honeywell. Ara ja és una eina de museu. Però després hi va haver una època crucial en aquest sentit que va ser quan la Universitat va decidir canviar l'ordinador que tenia per un de nou més apte per al projecte.

Cap el 1987 la política de petits ordinadors va esdevenir impopular en el món acadèmic i es va muntar un immens ordinador central que donés suport a centenars d'usuaris. Així els equips que necessitàvem informàtica vam començar a desenrotllar els nostres propis sistemes.

Des de 1993 treballem amb màquines UNIX. Busquem sistemes molt flexibles i fàcils de manejar, amb rutines especials i al màxim d'estàndards. Necessitem que la qualitat dels terminals sigui molt alta. Ara, doncs, tenim accés al gran ordinador central i disposem de dos grans servidors de la nostra xarxa local; un que conté el corpus i un altre gran servidor que conté el software.

P: Quin és l'estat actual del projecte?

El treball més immediat en el present és la revisió del diccionari, publicat ara fa set anys, a partir de les crítiques que hem rebut d'arreu del món, sobretot dels professors d'anglès que l'han utilitzat. El diccionari ha estat rebut en general amb gran entusiasme, però hi ha aspectes que cal millorar.

Un segon camp en el qual anem treballant és un futur diccionari de contextos (*Collocation Dictionary*), a partir de la introducció d'etiquetes (*tags*) en el corpus, etiquetatge de què abans he parlat.

També estem preparant la versió electrònica en CD-ROM dels nostres treballs. Actualment tenim un prototip que conté tres llibres: el dicionari, la gramàtica i el manual d'ús; i, a més, inclou 5 milions de paraules del corpus. D'aquesta versió experimental se n'han fet 100 exemplars, que han estat distribuïts a diverses parts del món a fi de recollir-ne les impressions i les crítiques. Serà comercialitzat per l'Editorial Collins, i espero que despertarà el mateix entusiasme que les obres impreses, si les eines per accedir-hi s'abarateixen.

Un dels grans problemes de la informació electrònica en treballs de lexicografia és com controlar la quantitat d'informació. Necessitem eines que sumaritzin la informació. Per tant considerem que cal una eina de referència que controli la informació i el vocabulari. Crec que la utilització de corpus per fer diccionaris ha fet canviar la lexicografia. Permet fer productes més sofisticats, més evidents. El lexicògraf disposa de molta informació alhora; té al seu abast sistemes d'accés a la informació més flexibles i selectius. I la introducció d'analitzadors en els textos és un pas per al tractament automàtic del llenguatge i, a la llarga, si es disposa de la mateixa anàlisi per a dues llengües, és la base de la traducció assistida o automàtica.

Un altre dels projectes importants que tenim entre mans en el marc dels acords de col·laboració amb Collins és el dels diccionaris bilingües (*bilingual bridge dictionaries*). Són diccionaris bastant originals que parteixen de la paraula en anglès i l'expliquen en la llengua de traducció. El primer *diccionari pont* que s'ha fet és l'anglès/

portuguès-brasiler. El bilingüe anglès/espanyol és en curs d'elaboració. Per fer aquests diccionaris treballem en col·laboració amb lexicògrafs natius de cada llengua. Al final del procés de treball, el nostre equip supervisa el treball i es permet discutir amb el lexicògraf natiu determinades qüestions que poden ser problemàtiques. Un dels problemes amb què ens trobem contínuament és el calc de determinades estructures de l'anglès quan es fan les explicacions en l'altra llengua. I també ens trobem amb el fet que determinades explicacions que són naturals en anglès no ho són en altres llengües. Aquests *diccionaris pont* són força originals en relació amb els bilingües tradicionals, en el sentit que intenten reproduir el procés intel·lectual que, en consultar-los, fa el parlant no natiu amb una competència reduïda en llengua anglesa.

Un exemple real seria el següent:

> **nail** /pron/, **nails, nailing, nailed**. 1. COUNT N **A nail** es una pieza metálica pequeña, con una punta en un extremo, la cual se golpea con un martillo para meterla dentro de algo. *The mirror that hung from a nail on the wall.* 2. VB WITH OBJECT AND ADJUNCT Si se dice que alguien **nails** algo en algún sitio, lo sujeta allí con un clavo. *They nail plastic sheets over their windows.* 3 COUNT N Tus **nails** son las láminas finas y duras que recubren la punta de los dedos. *He keeps biting his nails.* 4 Si dices que alguien **has hit the nail on the head**, quieres decir que con lo que ha dicho ha acertado plenamente. 5 **a nail in the coffin** de algo: busca **coffin**.

P: Per acabar, quin balanç faria de la utilització dels ordinadors en el tractament del llenguatge i de l'elaboració de diccionaris a partir de corpus textuals?

Penso que la introducció dels ordinadors en el tractament del llenguatge ha fet canviar totalment el panorama de la recerca sobre el llenguatge i el sistema de treball en lexicografia.

Els corpus textuals permeten disposar d'informació no intuïtiva i, si són suficientment amplis, donen una fiabilitat a les dades que altrament no poden tenir. Tres són al meu entendre les diferències fonamentals que la utilització de corpus confereixen al treball lexicogràfic. La primera és la major qualitat i diversitat de la informació: si es disposa d'un bon corpus es poden obtenir productes lexicogràfics més sofisticats i diversos; també es poden il·lustrar les unitats de manera més fiable, tot i que hi ha diversos aspectes que no estan encara ben resolts com, per exemple, la categorització pragmàtica de les paraules. El segon avantatge que ofereixen els corpus és la possibilitat d'establir la fraseologia: amb els corpus es poden detectar moltes estructures concurrents a cavall entre la fraseologia i la combinació freqüent de determinats mots, que no es podrien detectar sense la utilització de corpus; amb un corpus ampli de dades es pot aconseguir una variació lingüística suficient per donar compte de la realitat ex-

pressiva. Les *collocations* (concurrències freqüents) són importantíssimes per descriure les llengües. La semàntica dels mots només es pot establir a partir de la seva utilització en contextos variats. La tercera característica diferencial que ofereix el treball lexicogràfic a partir de corpus és l'evidència: els corpus permeten detectar expressions, significats i fraseologia que no s'havien detectat abans i il·lustrar els mots amb molta més fidelitat a l'ús.

La introducció de sistemes d'etiquetatge en els textos obre grans possibilitats de descripció del llenguatge, encara que cal superar determinats problemes, com el que suposa la classificació pragmàtica dels mots i les expressions.

El fet de poder disposar d'eines de tractament del llenguatge i d'aprenentatge de les llengües en suport electrònic canvia notablement les possibilitats de treball, sobretot si es pot fer ús d'aquestes eines en ordinadors personals.

Del naixement del projecte COBUILD fins ara les circumstàncies han canviat molt. És un fet que cada cop hi ha més corpus de textos i que avui dia no es pot plantejar la descripció d'una llengua sense disposar de materials que aportin l'evidència dels usos, i no es pot concebre l'elaboració de diccionaris i de gramàtiques sobre la base de la intuïció.

M. TERESA CABRÉ CASTELLVÍ
Birmingham, 11 d'agost de 1993