

MODALIDADES TECNICAS PARA LA REALIZACION DE UN DICCIONARIO DE CONCORDANCIAS Y FRECUENCIAS

El empleo de calculadores electrónicos para la elaboración de los datos lingüísticos, cuya utilidad se discutía hace quince años¹, forma parte hoy de los instrumentos habituales de muchas disciplinas lingüísticas y literarias². Este rápido desarrollo ha sido determinado por la difusión de los centros de cálculo universitarios y científicos que facilitaron el uso de los « computers » a la total comunidad de los investigadores y de los estudiosos, y por la notable cantidad de medios que los países tecnológicamente más adelantados empeñaron, entre 1950 y 1965, en la traducción automática y en la *Information Retrieval*. Estas investigaciones, aun si no alcanzaron en la mayoría de los casos la realización de su objetivo principal, produjeron un conjunto de métodos y de conocimientos que, no sólo contribuyeron mucho al desarrollo de la lingüística aplicada matemática y computacional, sino influyeron en el incremento de muchas escuelas lingüísticas³.

Pero, donde el uso del calculador ha conocido su mayor afirmación, es en el sector del escrutinio de los textos, por lo que sería hoy difícil pensar en selecciones completamente manuales⁴.

Los « corpora in machine readable form » se hacen cada vez más amplios y numerosos, por lo que, y al menos por cuanto se refiere a las lenguas primarias de cultura, resulta siempre más fácil al investigador encontrar ya elaborado electrónicamente el texto que desea examinar⁵.

El escrutinio electrónico de textos, si bien incomparablemente más fácil y rápido del manual y tradicional, implica sin embargo operaciones muy caras por trabajo y gastos; en primer lugar la perforación (o sus equivalentes) y la lematización. Por esto, hay que evitar — por medio de la coordinación nacional e internacional de los programas de investigación — la duplicación de las indagaciones, garantizando con la adopción de « standards » comunes de registración y análisis, la comparación

técnica y científica de los materiales elaborados y de los resultados conseguidos por distintos investigadores⁶.

La experiencia señaló que el único medio eficaz es el de establecer una red internacional de bancos nacionales de datos lingüísticos, cada uno de los cuales garantiza la adopción de un único « standard » de elaboración de los textos por parte de los propios usuarios y, al mismo tiempo, está en grado de ejecutar todas las complejas operaciones que se necesiten para hacer posible en concreto el intercambio de textos en « machine readable form » con los demás bancos de la red.

Las elaboraciones con que se produjeron las concordancias y los índices de las obras poéticas de César Vallejo, se efectuaron según los procedimientos y los « standards » adoptados y experimentados por la División Lingüística del C.N.U.C.E., que gestiona el banco italiano de datos lingüísticos y participa en la creación de la red internacional. Ésta, en el momento en que escribimos, empieza a actuar y a tener consistencias efectivas⁷.

Antes de describir brevemente las distintas fases operativas del método de escrutinio, nos parece útil precisar algunas locuciones empleadas a menudo en la práctica cotidiana de los escrutinios lexicales automatizados. No entendemos dar definiciones rigurosas ya que, en algunos casos, no existen, sino sólo indicar intuitivamente su contenido, así como el uso confirma, con el objeto de facilitar y simplificar la descripción de las fases sucesivas de nuestras elaboraciones.

Llamamos *palabras* (u *ocurrencias*) las sucesivas unidades gráficas de las cuales se compone un texto: para el programa de elaboración, la unidad gráfica corresponde a una o más letras (o caracteres equivalentes) entre espacio o signos de puntuación⁸. El conjunto de las indicaciones que localizan la palabra en el texto constituye su *referencia*: v. gr. los números de la página o de la línea, o bien del canto, octava, verso, etc.

Indicamos como *record* la unidad de tratamiento registrada en un soporte aceptado por el calculador (cintas magnéticas, discos, etc.), que representan un elemento basilar de la elaboración electrónica y que se compone, a su vez, de un cierto número de elementos de información (*campos*) conectados entre ellos. Por ejemplo, podemos conjeturar que un texto se registre como una serie de *record-palabra*; cada palabra del texto se registra en un *record* (indicado desde este momento como RP), que comprende tres campos distintos: el *campo-palabra*, en que cabe exclusivamente una palabra del texto; el *campo-referencia*,

que contiene la referencia de esta palabra; el *campo n.º-record*, en que se indica el número progresivo del *record* en la serie ordenada de los mismos.

Llamamos *selección* a la operación con que se ordenan algunos *records* según una secuencia establecida (ascendente o descendente), que se basa en una particular clave de ordenación. Por ejemplo, en una guía del teléfono, los usuarios aparecen por orden alfabético, y la clave de ordenación es su apellido-nombre. Pero, considerando siempre en una guía del teléfono a los usuarios en una secuencia distinta, aparece la del elenco por calles. Aquí la clave de ordenación es la calle y el número del domicilio. Como podemos ver, la clave se puede componer por medio de varios campos, unidos entre ellos por una exacta *jerarquía*: en el interior de la secuencia alfabética por calles (campo n. 1), los usuarios se encuentran sucesivamente ordenados por el número de sus domicilios (campo n. 2)⁹.

En un texto suficientemente amplio, no todas las palabras son distintas: algunas se repiten una o más veces. Llamamos *formas gráficas* las palabras « distintas » en el texto. El elenco de las formas gráficas nos indicará, pues, todas las palabras distintas en el texto examinado: al lado de cada forma gráfica podemos registrar el número de sus presencias en el texto, y esto se denominará *frecuencia absoluta* de la forma. Por ejemplo, en la frase *quien no me escucha me ofende*, las palabras son seis, pero las formas gráficas son cinco: *me* tiene frecuencia 2; *quien*, *no*, *escucha*, *ofende* tienen frecuencia 1. Como veremos, de vez en cuando a la misma forma gráfica corresponden unidades lingüísticas distintas: v. gr. la forma gráfica *daño* puede ser la primera persona del presente indicativo del verbo *dañar* y el sustantivo sinónimo de *avería*, *perjuicio*, *menoscabo*. Decimos, pues, que la forma gráfica *daño* es homógrafa, y corresponde a dos *formas lexicales* distintas.

Llamamos *lema* aquella voz basilar que representa, en un diccionario, las distintas formas de un texto. Por ejemplo, a las formas *decid* y *diremos* corresponde el lema *decir*; a las formas *id* y *fueron*, el lema *ir*; a las formas *bello*, *bella*, *bellos*, *bellas* el lema *bello*.

Hemos definido *lematización* la operación con la cual colocamos cada forma, y relativas ocurrencias suyas, bajo el lema respectivo¹⁰.

Indicamos con *record-palabra lematizado* (RPL) un RP al

cual se unió — en un campo especial — el lema de la palabra en él contenida.

La *frecuencia absoluta del lema* es el número de sus ocurrencias en el texto, y corresponde a la suma de las frecuencias de sus formas.

En la lista de los lemas (o de las formas) por *frecuencia descendente*, los lemas (o las formas) están ordenados del más frecuente al menos frecuente: lemas (o formas) que tienen frecuencia igual se ordenan alfabéticamente dentro de la misma *clase* de frecuencia.

Consideramos *contexto* de una palabra un conjunto de palabras contiguas a ella, escogidas por supuesto entre las que preceden o siguen inmediatamente".

Llamamos *record-palabra contextualizado* (RPC) un RP al que se unió, en un campo determinado, el contexto relativo a la palabra registrada en su campo-palabra. Desde ahora en adelante, este conjunto se llamará *palabra-exponente* del RPC.

Definimos *contextualización* de un texto las operaciones con las cuales se reimprimen, en una cinta nueva, todos o algunos RP de una cinta-texto, con la adjunta de su contexto, transformándolos en RPC.

La contextualización se dice *integral* si *todos* los RP han sido reimpresos y transformados en RPC. En otro caso se llamará *selectiva*. Un ejemplo muy frecuente de contextualización selectiva se da cuando se contextualizan sólo las palabras denominadas llenas o *semánticas* (sustantivos, adjetivos y adverbios calificativos, verbos), mientras se eliminan las demás, o sea las llamadas palabras *vacías* o *sinsemánticas* (artículos, pronombres, preposiciones, conjunciones, etc.).

Llamamos *concordancias* de un texto el *conjunto ordenado* de todos los RPC que nacen de la contextualización. Si la clave de la ordenación es el orden alfabético de las palabras-exponentes, hablamos de *concordancia de las formas*; al contrario, se define *concordancia de los lemas* la ordenación alfabética de los mismos. Cada operación descrita en el texto se indica con un código (Op. 1; Op. 2; etc.).

Op. 1. - Para obtener un texto legible por parte del calculador, se usa reproducirlo en fichas o en cintas de papel con una máquina perforadora, o bien directamente en cinta magnética

o en disco por medio de un terminal¹². Para las obras poéticas de César Vallejo utilizamos fichas en que reproducimos las palabras, la puntuación, la división tanto por colección como por verso, tanto por página como por líneas (ficha-texto). Grabando en ficha mediante perforación los caracteres de la imprenta, seguimos las normas convencionales propuestas por la División Lingüística del CNUCE de la cual, como ya dicho, empleamos los programas y los procedimientos, así como los elaboradores IBM/155 y IBM/360-167.

Op. 2. - Las fichas texto perforadas se introducen en una máquina llamada « verificadora », y el texto se « escribe » nuevamente por una persona distinta de la que ejecutó la primera perforación. Si hay discrepancia, la máquina se para y el operador controla si esto se debe a una falta suya o de quien perforó el texto la primera vez.¹

Op. 3. - El calculador « lee » las fichas, es decir traduce los agujeros de cada una de ellas a los elementos correspondientes de su propio lenguaje interior, descomponiendo el texto en las distintas palabras que, en nuestro caso, son las unidades elementales de la elaboración¹³.

Según va leyendo el texto y lo descompone, el « computer » lo registra palabra por palabra en una cinta magnética (« cinta-palabra »), y da a cada una de las palabras un número progresivo que la localiza de modo unívoco. Al mismo tiempo, con una máquina impresora que está conectada con él, imprime la llamada « lista-texto », que corresponde al texto original así como se perforó en las fichas, reproducido con los caracteres que la máquina impresora posee¹⁴.

Op. 4. - La lista-texto reproducida por el calculador se compulsa con el texto original¹⁵. Las faltas que se encuentran se evidencian en hojas especiales con un gráfico oportunamente dispuesto, donde se indica también el número progresivo que localiza la palabra o, más en general, la información errónea.

Op. 5. - El operador, copiando exáctamente el gráfico, perfora correctamente, sobre fichas especiales de rectificación, las palabras y las informaciones equivocadas la primera vez¹⁶.

Op. 6. - El « computer » copia la cinta-palabra, corrige las faltas y, al mismo tiempo, imprime otra edición nueva de la lista-texto.

Op. 7. - El calculador copia las palabras del texto corregido, uniendo a cada una de ellas una porción del contexto en que se halla, delimitada según las convenciones que el investigador estableció, eligiendo entre las posibilidades que un programa flexible de concordancias — como es el programa tipo del CNU-CE — le ofrece. Este programa permite establecer cada vez que es necesario el número máximo de caracteres que pueden componer el contexto. A menudo nos encontramos con dos exigencias opuestas en el acto de elegir este número. De un lado preferiríamos incluir un contexto muy amplio, para que el ejemplo se configure como más significativo¹⁷. Del otro, nos preocupamos de contener el contexto dentro de unos límites razonables, para hacer más rápida la lectura, o para que resulten más manejables las concordancias¹⁸. De acuerdo con la experiencia anterior, fruto de trabajos antecedentes, hemos creído oportuno adoptar contextos limitados a un número máximo de 120 caracteres, lo que equivale, en media, a 12-15 palabras¹⁹.

Para delimitar el contexto, el programa tiene en cuenta factores distintos (v. gr.: la puntuación, el fin o el inicio de un verso, etc.)²⁰. Por esta razón la palabra de la cual se reproduce el contexto no se encuentra siempre en el centro de la línea, sino que su posición está condicionada por la presencia, a su derecha o a su izquierda, de tales factores. También podemos especificar los elementos del texto que se quieren « contextualizar » o no, e indicar y clasificar los elementos que tienen función de « límites » del contexto²¹.

Op. 8. - Las palabras con sus propios contextos se ordenan alfabéticamente.

Op. 9. - El calculador imprime la lista de las concordancias de acuerdo con sus formas, copia a la vez en una cinta magnética, formas y contextos numerados progresivamente y produce, por cada forma, una ficha que contiene la misma.

Op. 10. - Las indicaciones de los lemas y las probables observaciones accesorias se escriben a mano en la lista de las concordancias al lado de cada forma.

Op. 11. - Se perforan estas indicaciones en las *fichas-forma* preparadas por el « computer ».

Op. 12. - La cinta de las concordancias por forma se copia agregando, a cada palabra, el lema perforado en la ficha relativa.

Op. 13. - Se ordenan las palabras por lema, forma y referencia.

Op. 14. - Se imprimen las concordancias provisionales por lema.

Op. 15. - El investigador realiza algunos controles en las concordancias lematizadas para averiguar la exactitud de la lematización. Si se encuentran equivocaciones se repite de nuevo el ciclo de las operaciones 11-15 hasta que no se localicen más faltas.

Op. 16. - De la cinta de las concordancias — ya ordenadas por lema — se obtiene otra cinta que contiene sólo los lemas por orden alfabético, con sus propias frecuencias. Al mismo tiempo, se imprime el elenco alfabético de los lemas.

Op. 17. - La cinta de los lemas se ordena otra vez según el orden descendente de las frecuencias.

Op. 18. - Se imprimen los lemas así ordenados.

Op. 19. - De la cinta de las concordancias, ya ordenadas por lema, se imprimen las concordancias definitivas, eliminando interjecciones, artículos, preposiciones y conjunciones.

Op. 20. - Por medio de la técnica de la impresión en *off-set* se compone el volumen, reproduciendo, con obvia reducción de sus dimensiones, los tabulados de las concordancias, los de los lemas en orden alfabético y los de los lemas en orden de frecuencia.

Las cintas magnéticas que contienen el texto corregido, las concordancias y las frecuencias por forma, las concordancias y las frecuencias en orden por lemas provisorios y definitivos, se custodian en los archivos de la División Lingüística del CNUCE, a disposición de los estudiosos que deseen elaboraciones ulteriores.

Notas

- 1 Véanse, a propósito de esto, la sección *Elettronica e Letteratura* a las páginas 89-151 del *Almanacco Letterario Bompiani*, Milano, 1962; «preprints» dados en el Convenio *Kolloquium über der Literarischen Analyse und der Lexicographie* efectuado en Tübingen en 1960 y los *Actes du colloque international sur la Mécanisation des recherches lexicologiques* (Besançon, 1961) publicados en los «Cahiers de Lexicologie» 3, 1962, 1.
 - 2 Véanse por ejemplo: A. Zampolli (ed.), *Linguistica Matematica e calcolatori*, Firenze (1973); B. Quemada, *Lexicology and Lexicography*, en T. A. Sebeok (ed.) *Current Trends in Linguistics*, Vol. 9, La Haye (1972), 395-476; R. I. David, A. R. Duckert (eds.), *Lexicography in English*, New York, 1973.
- Estiloeestadística y estadística lingüística: L. Doležal, R. W. Bailey (eds), *Statistics and style*, New York, 1969; U. Bortolini, G. Tagliavini, A. Zampolli, *Lessico di Frequenza della Lingua Italiana Contemporanea*, IBM, 1971; y la serie *The Romance Languages and their Structures*, dirigida por A. Juilland por el editor Mouton.
- Dialectología: C. Grassi, *Perspectives de l'emploi de l'elaborateur électronique en géographie linguistique et en dialectologie*, en A. Zampolli (ed.), *Linguistica Matematica e Calcolatori*, Firenze 1973, 233-239; W. N. Francis, G. M. Rubin, J. Svartvik, *A Method of Computer-produced graphical representation of dialectal variation in initial fricatives in Southern British English*, en «ICCL», 1969; Cr. Wood, *Dialectology by computer*, en «ICCL», 1969.
- Demología y folklore: A. M. Cirese, *Inventaires et répertoires lexicaux, formulaires et métriques des chants populaires italiens*, en A. Zampolli (ed.), *Linguistica Matematica e Calcolatori*, Firenze, 1973, 209-239.
- Bellas artes y arqueología: E. Chouraqui, *The Index and Filing System Used by the 'Inventaire Général des Monuments et des Richesses Artistiques de la France'*, en «Computers and the Humanities» 7 (1973) 5, 273-286; D. Vance, J. Heller, *Structures and Content of a Museum Data Bank*, en «Computers and the Humanities», 6 (1971), 2, 67-84; P. E. Arias, *Mise en fiche des vases grecs; problèmes et discussions*, en A. Zampolli (1973), 243-248.
- Ciencias Históricas: L. Fossier, *Ordinateur et histoire médiévale*, en A. Zampolli (1973), 269-298; J. H. Silbey, *Clio and Computers: Moving into Phase II, 1970-1972*, en «Computers and the Humanities» 7 (1972) 2, 67-80; E. Shorter, *The Historian and the Computer*, Englewood Cliffs, 1971.
- Lenguas clásicas: S. V. F. Waite, *Computers and the Classical Literature: 1971-1972*, en «Computers and the Humanities», 7 (1972), 2, 99-104; el boletín: «Calculi» y la revista «Revue du Laboratoire d'Analyse Statistique des Langues anciennes» de Lieja.
- Fonemática y fonética: A. Zampolli, *Studi di statistica fonemática dell'italiano*, Bologna (en prensa); A. Newell, J. Barnett, J. W. Forgie, C. Green, D. Klatt, J. C. R. Licklider, J. Munson, D. R. Reddy, W. A. Woods, *Speech understanding systems*, Amsterdam, 1973; Y. Ch. Morin, J. Friedman, *Phonological Grammar Tester underlying theories*, Ann Arbor (Mich.), 1971.
- Psicolingüística: R. E. Anderson, E. R. Coover, *Wrapping up the Package: Critical Thoughts on Application Software for Social Data Analysis*, en «Computers and the Humanities», 7 (1972) 2, 81-95; R. S. Wachal, O. Spreen, *A Computer-Aided Investigation of Linguistic Performance*, 1970; G. Sankoff, *A Quantitative paradigm for the study of communicative competence*, Austin, 1974; J. B. Carrol (1973).
- Musicología: J. Wenker, *A foundation for computational musicology*, en J. L. Mitchell, 267-280; D. Cantor, *A computer Program that accepts common musical relation*, en «Computers and the Humanities», 6 (1971), 2, 100-103; P. Grossi, *Computer and Music*, en «International Review of the Aesthetics and Sociology of Music», Zagreb, V (1974), 1, 279-286.
- Filología: D. J. Froger, *La critique des textes et son automatisé*, Paris, 1967; C. Segre, A. Zampolli, *Le concordanze dell'Orlando Furioso*, en *Atti del Convegno di Studi Ariosteschi* (en prensa).
- Sintaxis: A. Zampolli, *Problemi di Linguistica Applicata Computazionale*, Pisa, 1974, cap. 2; M. Kay, *Syntactic Analysis*, en A. Zampolli (ed.), *Computers, Texts and Structures*, Pisa (en prensa).

Semántica: R. Rustin (ed.), *Natural Language Processing*, NY, 1973; R. G. Schank, K. M. Colby, *Computer Models of Thought and Language*, SF, 1973; W. Woods, *Meaning and Machines*, en A. Zampolli (ed.), *Computational and Mathematical Linguistics, Proceedings of the V ICCL*, Pisa, 1975. Obras de carácter general son las siguientes: P. L. Garvin, B. Spolsky, *Computation in Linguistics*, Bloomington and London, 1966; R. Wisbey (ed.), *The Computer in Linguistic and Literary Research*, Cambridge, 1971; A. J. Aitken, R. W. Bailey, N. Hamilton-Smith (eds.), *The Computer and Literary Studies*, Edinburgh, 1973; A. Zampolli (ed.), *Linguistica Matematica e Calcolatori*, Firenze, 1973; J. L. Mitchell, *Computers in the Humanities*, Edinburgh, 1974; A. Zampolli (ed.), *Computational and Mathematical Linguistics, Proceedings of the V ICCL*, 1975; A. Zampolli (ed.), *Computers, texts and structures*, Pisa (en prensa).

- 3 Por lo que se refiere a la historia de la traducción automática y a su fracaso, véanse: J. Bar-Hillel, Y. A. Wilks: *Grammar, meaning and the machine Analysis of Language*, London, 1972, 1-13; *Language and Machines, Computers in Translation and Linguistics*, Washington, D.C., 1966; *Feasibility study on fully automatic high quality translation*, University of Texas, 1971. Teniendo en cuenta la influencia que ha ejercido la traducción mecánica sobre las escuelas lingüísticas, se aconseja ver: K. Brockhaus: *Automatische Übersetzung*, Pergamon Press, 1971; G. Mounin: *La Machine à traduire*, The Hague, 1964; A. Ljudskanov: *Mens und Maschinen Übersetzung*, Leipzig, 1973.
- 4 J. Raben, habló en este sentido de *The death of hand made concordances*, New York, 1969, y M. Spevack, en el artículo *Concordances: Old and New*, en «Computer Studies in The Humanities and Verbal Behavior», IV (1973), 1, 17-19, ha podido afirmar que «We now have literally hundreds of concordances and verbal indexes, perhaps thousands». El índice de mayor éxito obtenido por la llamada «lexicografía asistida por los calculadores» se localiza sin duda alguna en la adopción del escrutinio electrónico efectuada por las grandes empresas como las de los diccionarios históricos nacionales: el *Tesoro delle Origini* y el *Dizionario storico della Lingua Italiana dell'Accademia della Crusca* (cf. A. Duro: *Elaborations électroniques des textes effectuées par l'Accademia della Crusca, pour la préparation du Dictionnaire Historique de la langue italienne*, en A. Zampolli (ed.) *Linguistica Matematica e Calcolatori*, Firenze, 1973, 53-75; el *Trésor de la Langue Française de Nancy* (cf. R. Martin: *Travaux du Centre de recherche pour un trésor de la langue française* (T. L. F. de Nancy) en A. Zampolli, op. cit., 103-110; El *Thesaurus* del Instituto de lexicología neerlandesa (cf. F. De Tollenaere: *Travaux de l'Institut de lexicologie néerlandaise*, en A. Zampolli: op. cit., 29-40); el diccionario rumano del siglo XVI (cf. F. Dimittescu: *Projet d'un dictionnaire de la langue roumaine du XVI Siècle*, en A. Zampolli: op. cit., 41-48), los diccionarios históricos de las distintas épocas de la historia de la lengua inglesa (cf. R. V. Bailey: *Lexicography in English*, en R. I. McDavid, A. R. Duckert: op. cit.), el proyecto para un diccionario español (cf. J. Del Campo, I. Gonzalez, M. T. Molina, F. Marcos: *The automatic Syntactic Analysis as an Aid in Dictionary Making* en A. Zampolli (ed.): *Computational and Mathematical Linguistics, Proceedings of the 5th ICCL*, Firenze, 1975), etc.
- 5 Las reseñas más completas de los proyectos en curso son las siguientes: para las lenguas clásicas, la antecitada revista «*Calculi*», y, para las lenguas modernas la anteriormente mencionada «*Computers and the Humanities*», que dedica cada seis meses un número a la sección «*Directory of Scholars Active*», en la cual aparecen las respuestas a un cuestionario enviado a todas las Universidades y los Institutos de investigación del mundo, para tener al día los elencos y los estados de los proyectos de trabajos con calculador en las diversas disciplinas humanísticas (lingüística, literatura, filosofía, ciencias sociales, música, bellas artes, etc.).
- 6 El problema de la comparación de los datos cuantitativos producidos por los escrutinios hechos por investigadores distintos ha sido subrayado y discutido por muchos. Por ejemplo, ya en 1963, Ch. Müller escribía: «L'un des préceptes élémentaires de toute statistique est que l'on n'opère valablement que sur des populations, des individus et des caractères bien définis. Aussi peut-il sembler hasardeux de faire des statistiques sur les 'mots', puisque les linguistes sont d'accord pour constater qu'aucune définition satisfaisante de cet élément du langage ne peut être proposée». En particular, por las dificultades de comparar

- 13 En otras investigaciones, ejecutadas en medida lingüística de amplio rayo, tales unidades están constituidas por grafemas, fonemas, sílabas, morfemas, sintagmas, etc..
- 14 Como ya dicho para las perforadoras, aun las máquinas de imprimir, generalmente destinadas al trabajo en las haciendas y casas comerciales, disponen igualmente de un número limitado de caracteres, por lo que no ofrecen variedad de cuerpos o de series tipográficas, aunque si con varios recursos es posible representar de modo unívoco cualquier carácter del texto original.
El CNUCE, teniendo en cuenta el gran número de investigaciones lingüísticas emprendidas, hizo preparar algunas « cadenas tipográficas », con el objeto de obviar al menos en parte estas limitaciones.
- 15 En una primera revisión del texto que se perforó, se reveló muy útil el examen esmerado de los elencos de formas y de sus frecuencias relativas. Este procedimiento permite, entre otras cosas, identificar las formas « imposibles » en la lengua o en el texto investigados que, procediendo mayormente de faltas en la perforación, se presentan normalmente una sola vez. El mismo escrutinio permite también la identificación de incoherencias y discordancias en la grafía de formas particulares, que presentan la posibilidad de variantes gráficas. Este hecho es importante en la perforación de textos en edición crítica, cuando se quiera averiguar su grado de exactitud.
- 16 Normalmente para perforar un texto de unas 10.000 palabras se necesitan, más o menos, ocho horas de trabajo por parte de una mecanógrafa-perforadora. Otras tantas se emplean para controlar el texto perforado y corregir las faltas. La lectura de la lista-texto, que equivale aproximadamente a una corrección de pruebas de imprenta, corresponde a cinco o seis horas más. Frente a estos tiempos « largos » están los « breves », que son necesarios para las fases automáticas del escrutinio: las varias listas de frecuencias y las concordancias ordenadas por formas se obtienen en sólo media hora, a la que se debe añadir otro cuarto de hora para la impresión de los resultados. Aparece lógico, pues, que muchos esfuerzos se dirijan a mejorar y aligerar la fase de preparación del texto que se destina al calculador, pero importa sobre todo registrar los textos según criterios técnicos y científicos uniformes, así que puedan ser utilizados por distintos estudiosos en elaboraciones e investigaciones sucesivas.
- 17 No parece oportuno plantear aquí en términos científicos el problema de lo que se entiende por « contexto » de una palabra, problema que fácilmente nos llevaría a disquisiciones entre lingüística, lógica, psicología, etc. Con objeto de evitarlo, empleamos la palabra « significativo » en lugar de términos más técnicos y apropiados. En práctica, la « suficiencia » o « insuficiencia » de un contexto se juzgan en relación con el uso que se quiera hacer de él.
- 18 La facilidad de la consultación es de gran importancia para la mayor parte de los compiladores que usan auxilios electrónicos para componer concordancias. Ellos se proponen dar al futuro usuario no tanto un instrumento que le evite la consultación del texto, sino un instrumento para reconocer — antes de consultar el texto — las ocurrencias « interesantes » para la investigación. Con el « computer », en efecto, no se quieren producir las concordancias de la « opera omnia » de un autor o de un período histórico completo. El escrutinio resulta casi siempre integral y, en definitiva, mucho más pormenorizado que en las concordancias tradicionales. Por esta razón, resultan más numerosos y varios los sectores potenciales de interés de los futuros usuarios (gramatical, lexicográfico, documental, etc.). El contraste entre la exigencia de ampliar y la de reducir el contexto de las palabras se ejemplifica bien en el procedimiento de lematización manual. El tiempo de « lectura » de las concordancias resulta más o menos proporcional con la amplitud del contexto, sobre todo si la palabra exponente no se encuentra siempre rigurosamente en el centro de éste. De otra parte, recurrir al texto para completar un contexto insuficiente, implica mucho gasto de tiempo: los contextos potencialmente ambiguos deben, al menos, poderse interpretar claramente como tales, que obliga a recurrir al texto para evitar interpretaciones erróneas. Un tercer factor importante, y más en programas de extensión mayor, es el tiempo-trabajo del calculador indispensable para obtener los contextos, ordenarlos alfabéticamente e imprimirlos. Este tiempo, naturalmente, aumenta más o menos en proporción con la amplitud de los contextos.

- 19 Nuestras experiencias y las de las Instituciones que adoptan nuestros procedimientos, tuvieron como objeto las concordancias por formas, en las cuales el investigador debe estar en grado de individuar y dividir las formas homógrafas, asignar los lemas y, a menudo, elegir cuáles contextos eliminar de las elaboraciones sucesivas y cuáles conservar, en cambio, en la documentación lexicográfica después de la conclusión del escrutinio. Muy a menudo un contexto comprende una frase entera, a menudo un período entre dos puntuaciones fuertes. Para el italiano, el recurso del texto resulta necesario, en media, una vez cada cien ocurrencias. Además, esta amplitud permite imprimir el contexto y su referencia, al menos la topográfica, en una sola línea del «computer», con notable ahorro de papel y tiempo de imprenta.
 - 20 Los criterios de delimitación de los contextos vienen, desde tiempo, unificándose en los varios centros de lexicografía que utilizan complejos mecanográficos. Prácticamente pueden agruparse en los tipos siguientes:
 - a) En los sistemas *kwic-index*, perfeccionados sobre todo para aplicaciones documentales, todas las palabras de los títulos que componen las listas bibliográficas que hay que elaborar o, con más frecuencia, solamente las palabras «lexicales» que allí aparecen, se ponen en orden alfabético y reciben, como contexto, los títulos (o parte del los títulos) en los cuales ocurren. Los programas de *kwic-index*, hoy los más difundidos, no sirven para compilar concordancias con fines lexicográficos por distintas causas.
 - b) La palabra exponente se encuentra siempre en el centro de su contexto, es decir precedida y seguida de un número igual de espacios o palabras. El programa resulta de fácil compilación, y la composición de los contextos es muy rápida. A menudo, con este método se acompaña la elección de no lematizar por el evidente propósito de explotar la velocidad de la máquina y eliminar cualquier intervención humana.
 - c) El contexto está constituido por una unidad entera de referencia: verso, párrafo, inciso, etc.
 - d) Los límites del contexto se señalan, en fase de «pre-edición»: el texto se divide en fragmentos por medio de contraseñas que se perforan y mantienen en las elaboraciones sucesivas. Cada fragmento funciona como contexto para todas las palabras que le componen.
 - e) El contexto está constituido siempre y sólo por todas las palabras comprendidas entre dos signos de puntuación. Los tipos *c*, *d*, *e*, poseen una característica común y exacta: el texto se divide en *sintagmas* sucesivos, y todas las palabras de un sintagma tienen el sintagma completo en lugar de contexto, es decir poseen todas el mismo contexto.
 - f) Se elige el contexto según la naturaleza de la palabra: para palabras gramaticales éste es a menudo un «trinomio» en el cual la palabra gramatical se encuentra en el centro; para las preposiciones se toman las dos palabras sucesivas, etc. De esto se deduce que las palabras con las que se debe construir el contexto se encuentran ya clasificadas. Por lo tanto se usa este método para concordancias de lemas o en la última parte del escrutinio, y no como instrumento de trabajo, por ejemplo, en la fase de lematización. Aparecen interesantes, según lo dicho, las posibilidades que hay para automatizar, por lo menos parcialmente, el análisis sintáctico escogiendo para cada palabra la parte final de la estructura a que pertenece, y que se juzga interesante como contexto para la categoría gramatical a la que pertenece la palabra.
 - g) El calculador nos da con cualquier método (la mayor parte de las veces de tipo *d*) un contexto que aparece sobreabundante, y que se reduce después a las dimensiones deseadas por medio de fichas con las cuales se comunican al «computer» las palabras que el estudioso, después de un cuidadoso examen, ha decidido eliminar para aligerar el contexto.
 - h) El contexto se regula teniendo en cuenta ciertos signos, como la puntuación, el cambio de referencias, etc. El contexto se construye, como en el tipo *b*, para cada palabra, por lo que varía de una palabra a la sucesiva, pero — diferentemente del tipo *b*, y semejantemente a los tipos *c*, *d*, *e*, — está regulado por la presencia de elementos bien definidos, por lo que la palabra se puede encontrar colocada de diverso modo en el contexto: al inicio, en el centro o al final, según los diferentes casos.
- El algoritmo en nuestro programa posee potencialmente la capacidad de generar contextos de todos los tipos, aunque si el tipo *f* no ha sido experimentado

suficientemente, por la razón de que usamos las concordancias en la fase de lematización y antes de todo análisis.

- 21 En el programa nuestro los elementos del texto se deben clasificar en:
- elementos que tienen que poseer un propio contexto y estar presentes en los contextos de los otros (por ej.: las palabras «lexicales»);
 - elementos que deben formar parte de los contextos de los otros, pero no reciben contexto propio (por ej.: los signos de puntuación);
 - elementos que no deben entrar en los contextos ni recibir contexto propio; en general se trata de «códigos» introducidos en el texto para efectuar algunas operaciones del programa (por ej.: los signos de división en páginas y en líneas);
 - elementos que tienen que recibir contexto propio, pero no entrar en los contextos de los otros (por ej.: las variaciones, en el caso que se quieran tener en cuenta en la elaboración de un texto con aparato crítico).
- Los elementos del texto se deben dividir por segunda vez en los siguientes dos grupos:
- elementos cuya función es la de «delimitar» el contexto (por ej.: un signo de puntuación o el final de un capítulo);
 - elementos cuya función no es la de delimitar el contexto (por ej.: una palabra).

En nuestro algoritmo, los límites se pueden distribuir en nueve clases. Los elementos de la novena clase son límites de contexto que no se pueden pasar: quiere decir que si cuando se construye la parte derecha de un contexto se encuentra un límite de clase n. 9, el contexto no debe ir más allá de ningún modo hacia la derecha, aunque si se encuentra espacio a disposición. A menudo se definen límites de clase n. 9 los contrasignos de un cambio de capítulo: esto es así porque no se juzgan útiles, para el contexto de la palabra final de un capítulo, las palabras iniciales del capítulo siguiente y al contrario.

Los elementos de las otras clases, de la 1 a la 8, son límites de contexto que se pueden pasar. Si cuando construimos la parte derecha de un contexto encontramos uno de estos límites, el contexto podrá pasarlo, volver hacia la derecha y continuar solamente después de encontrar un límite de clase igual o mayor cuando va hacia la dirección opuesta. Los diferentes signos de puntuación se pueden atribuir a clases diversas según su fuerza: por ejemplo, el punto final, la interrogación, la admiración a la clase n. 3; los paréntesis, las comillas, los dos puntos y el punto y coma a la clase n. 2; el guión y la coma a la clase n. 1. En las concordancias de César Vallejo hemos considerado como límite el final de una línea; como sin barrera (sin otras subdivisiones de clase) las constituidas por los varios signos de puntuación.

Bibliografía

- A. J. Aitken, R. W. Bailey, N. Hamilton-Smith, (eds.): *The Computer and Literary Studies*, Edinburgh, 1973.
- R. V. Bailey: *Lexicography in English*, en R. I. McDavid, A. R. Duckert, op. cit.
- J. Bar-Hillel, Y. A. Wilks, *Language and Machines, Computers in Translation and Linguistics*, Washington D. C., 1966.
- J. Bar-Hillel, Y. A. Wilks: *Feasibility study on fully automatic high quality translation*, University of Texas, 1971.
- J. Bar-Hillel, Y. A. Wilks: *Grammar, meaning and the machine. Analysis of Language*, London, 1972, 1-13.
- K. Brockhaus, *Automatische Übersetzung*, Pergamon Press, 1971.
- A. Chandor, *Dizionario di Informatica*, Bologna, 1972. (trad. dall'inglese di G. Rappelli).
- J. Cohen, *Structure du langage poétique*, Paris, Flammarion, 1966.
- R. I. David, A. R. Duckert, (eds.): *Lexicography in English*, New York, 1973.

entre ellas listas de frecuencias lematizadas con criterios distintos, se vean: A. A. Lyne: *The Problem of no-comparability of word-frequency counts*, comunicación presentada en el III Congreso Internacional AILA (1972), y *L'élaboration des listes de fréquence*, en «Cahiers de Lexicologie», II (1973), 23, 83-108; A. Zampolli, *Problemi di linguistica applicata computazionale*, Pisa, 1974, 25-28 y G. Engwall, *Fréquence et distribution*, Stockholm, 1974. En particular, para los problemas relativos a la lematización del español, véase A. Juilland, E. Chang-Rodríguez: *Frequency Dictionary of Spanish Words*, The Hague, 1964, XXVIII-XXXV.

- 7 La «División Lingüística» del CNUCE, Instituto del CNR, tiene por finalidad promover las investigaciones en el sector de la lingüística matemática y computacional, y de auxiliar los proyectos italianos en este sector. Los Institutos que emplean los procedimientos del CNUCE son hoy más de 80, y han elaborado con el calculador unos 8.000.000 de palabras en 20 lenguas, para investigaciones lexicológicas, lexicográficas, estadísticas, filológicas, literarias, demológicas, dialectológicas, de lingüística histórica, de lingüística teórica, etc. Todos los textos han sido registrados y analizados con criterios técnicos y científicos «standard» y, por lo tanto, constituyen una gran biblioteca electrónica que trabaja con los mismos programas. Al mismo tiempo la «División Lingüística» del CNUCE colabora en la realización de un banco internacional de datos lingüísticos, formado por Centros nacionales que disponen de recursos científicos y tecnológicos para el recíproco cambio de los materiales en «machine readable form».
- 8 Los caracteres codificados que constituyen el texto registrado en las fichas se dividen, para el programa, en dos grandes categorías funcionales:
 - a) *separadores*, o sea códigos que separan e interrumpen las palabras (por ejemplo: el espacio, los signos de puntuación, etc.);
 - b) *letras*, o sea códigos que componen las palabras (por ejemplo: letras, signos diacríticos, apóstrofes, etc.).
 Una palabra se define con una serie, sin solución de continuidad, de letras entre dos separadores.
- 9 En la definición de *record* y de *selección* he seguido casi fielmente las voces *record* y *ordinamento* del *Dizionario di Informatica* di A. Chandor (Trad. del inglés por G. Rapelli), Bologna, 1972. He usado también el término *selección* para los procesos de ordenación con el calculador (mientras el Diccionario citado reserva *selección* sólo para el ordenamiento con máquinas de fichas) para respetar el uso que prevale entre los usuarios lingüísticos de los calculadores.
- 10 Para las dificultades derivadas del proceso de lematización del estado actual del concepto de *palabra* en las teorías de las diversas escuelas lingüísticas, véase; U. Bortolini, C. Tagliavini, A. Zampolli: *Lessico di frequenza della lingua italiana contemporanea*, 1971, XLII y siguientes.
- 11 Los lectores ópticos actuales se encuentran en grado de leer sólomente documentos impresos con caracteres especiales, ya que el lector óptico universal capaz de leer cualquier cuerpo de caracteres es aún una utopía.
- 12 Normalmente el número de los caracteres diversos que hay que representar pasa del centenar (se comprenden en la cuenta letras del alfabeto, cifras numéricas, signos diacríticos o de puntuación; todo esto en blanco y negro, letras redondas y cursivas, mayúsculas y minúsculas, etc.), pero las máquinas perforadoras hoy empleadas permiten distinguir, por medio de oportunas combinaciones de dos o tres agujeros, sólomente 48 o, al máximo, 64 signos distintos. Por esta razón, elaborando el texto de la Constitución italiana, tuvimos que adoptar — como haríamos para cualquier texto no numérico — una codificación compleja que permitiría establecer una relación unívoca entre cada uno de los agujeros de la ficha y los caracteres que hay que representar. Para este fin empleamos, integrándolos entre ellos, dos entre los métodos más comunes. El primero consiste en la adopción de lo equivalente funcional del que en las máquinas de escribir es la tecla de las mayúsculas. Entre 64 códigos disponibles se eligen unos con función de *clave*; cada uno de los códigos que quedan adquiere un significado distinto según la última clave que precede. Claro está que el número total de los caracteres que se pueden representar por este método corresponde al producto del número de los caracteres usados como clave multiplicado por el número de los demás.

- J. Del Campo, I. González, M. T. Molina, F. Marcos: *The automatic Syntactic Analysis as an Aid in Dictionary Making*, en A. Zampolli (ed.): *Computational and Mathematical Linguistics, Proceedings of the 5th ICCL*, Firenze, 1975.
- F. De Tollenaere: *Travaux de l'Institut de lexicologie neerlandaise*, en A. Zampolli, 29-40.
- F. Dimitrescu: *Projet d'un dictionnaire de la langue roumaine du XVI siècle*, en A. Zampolli, 41-48.
- L. Dolezel, R. W. Bailey, (eds.): *Statistics and Style*, New York, 1969.
- A. Duro: *Elaborations électroniques des textes effectuées par l'Académie de la Crusca, pour la préparation du Dictionnaire Historique de la langue italienne*, en A. Zampolli (ed.): *Linguistica Matematica e Calcolatori*, Firenze, 1973, 53-75.
- G. Engwall: *Fréquence et distribution*, Stockholm, 1974.
- A. Juilland: *Foreword, Frequency Dictionary of Spanish Words*, The Hague, 1964.
- A. Ljudskanov: *Mens und Maschinen Übersetzung*, Leipzig, 1973.
- A. A. Lyne: *The Problem of no-comparability of word-frequency counts*, comunicación presentada en el III Congreso Internacional AILA, 1972.
- R. Martin: *Travaux du Centre de recherches pour un Trésor de la langue française (T.L.F. de Nancy)*, en A. Zampolli, 108-110.
- J. L. Mitchell: *Computers in the Humanities*, Edinburgh, 1974.
- G. Mounin: *La Machine à traduire*, The Hague, 1964.
- J. Raben: *The death of hand made concordances*, New York, 1969.
- R. Rustin, (ed.): *Natural Language Processing*, New York, 1973.
- M. Spevack: *Concordances: Old and New*, en « Computers and Verbal Behaviour », IV (1973), 1, 17-19.
- C. Vallejo: *Obras poéticas completas*, Lima, 1968.
- J. Wenker: *A foundation for computational musicology*, en Mitchell J. L., 267-280.
- R. Wisbey, (ed.): *The computer in Linguistics and Literary Research*, Cambridge, 1971.
- A. Zampolli, (ed.): *Linguistica, Matematica e Calcolatori*, Firenze, 1973.
- A. Zampolli, (ed.): *Computational and Mathematical Linguistics, Proceedings of the 5th. ICCL*, Firenze, 1975.
- A. Zampolli, (ed.): *Computers, texts and Structures*, Pisa (en prensa).
- Actes du Colloque International sur la Mécanisation des recherches Lexicologiques*, Besançon, 1961, en « Cahiers de Lexicologie », 3, 1962, 1.
- Elettronica e Letteratura*, in « Almanacco Letterario Bompiani », Milano, 1962, 89-151.
- El Diccionario rumano del siglo XVI*.
- Dizionario Storico della Lingua Italiana dell'Accademia della Crusca*.
- L'élaboration des listes des fréquences*, en « Cahiers de Lexicologie », II (1973), 23, 83-108.
- El Thesaurus del Instituto de lexicología neerlandesa*.
- Preprints del Kolloquium über der Literarischen Analyse und der Lexicographie*, Tübingen, 1960.