Using a computer database to develop and operate an on-line dictionary of neologisms

Diarmuid Bradley and Allen J. McTernan

The Spanish language is increasingly exposed to the same influences as English – mass media, scientific discoveries, technological innovation, urban culture, international markets, etc. – and, like English, is also changing. The rate of change, in terms of lexical expansion, is, one suspects, rather similar in both languages. For a number of reasons, however, Spanish lexicography is not as active as its English counterpart. One consequence of this is that bilingual Spanish/English dictionaries – which tend, in any case, to be cautious in their response to recent language change – run the danger of falling even farther behind in their coverage of up-to-date usage. That can readily create problems for students in applied languages departments, such as the one in Heriot-Watt University, in which the texts studied and the materials used in language classes are invariably contemporary and are drawn from areas such as economics, politics and social institutions, where lexical change is noticeably active in present-day Spanish.

In response to the problem, it was decided to compile a collection of such neologisms in the form of a bilingual Spanish/English glossary. Initially, the objective established was a very modest one in terms of the number of fields covered; it included, from the start, however, the intention to record a range of citations with which to document usage, morphological features and grammatical information. It very soon became obvious that a card-index system was unwieldy to use and time-consuming in operation. Accordingly, it was decided that the use of computer technology should be investigated, and the outcome, in lexicographical terms, will be described in the course of this paper.

1. Database creation and use

We should first mention, in passing, that the system has been developed within the constraints of strictly limited resources. Secondly, it should be made clear that the system is designed to assist the lexicographer in his work rather than replace him in any way; his intervention and decision-making remain essential at every stage and the following description takes that fact as understood.

The database (d/b) that has been developed is structured around a constantly expanding set of Spanish headword entries and associated illustrative citations. At the beginning of the project we experimented briefly with direct on-line input of materials into the d/b. That approach proved quite unsatisfactory, mainly for three reasons. Firstly, the d/b processing of each line or even page of material generally entailed a significant, that is to say, a costly delay, as the computer used for the project, the University mainframe system, is generally a busy machine. Secondly, quality control suffered as effectively only the keyboarder could check the material before its entry. Thirdly, it was found that proof-checking material on the screen created much more strain than that resulting from carrying out the check on a hard copy with subsequent on-line editing of the text. It was decided, therefore, in the interests of accuracy, efficiency and comfort, to prepare all materials in workfiles, employing batch mode for all processing transactions.

We propose at this point to describe the various stages, and the procedures followed in each, in the course of collecting, verifying and storing neologisms and their supporting citations in the d/b.

1.1. Entry-creation cycle. Firstly, possible neologisms are identified and marked in selected source texts (newspapers, journals, books).

1.1.1. Uncoded headword listings. Next, a simple datafile is created on the computer, listing the possible neologisms in the sequence in which they occur in the source text. At this stage the list contains no codings. The list, marked (1) in the Appendix, illustrates the usual form of such listings. When complete, the datafile is examined by a program (JOB/SPCOMP) which compares each item in the list against the existing headword fields in the d/b, searching for matching entries and their associated information. In carrying out the comparisons, this program ignores embedded spaces, diacritics, punctuation points and upper and lower-case variations, thereby ensuring that all relevant variants are also retrieved. The run generates two separate hardcopy outputs, firstly a comparison listing, containing the data just mentioned and, secondly, a provisionally coded version of the initial headword listing.

1.1.2. Comparison listing. The comparison listing, item (2) in the Appendix, echoes each of the headword entries in the initial datafile, identifying those already in the d/b and the status that has been assigned to them. Through a simple set of letter codings, shown on the left-hand margin on the printout, it identifies terms already found to be lexicalized in the major Spanish dictionaries. Thus, MM denotes items included in Maria Moliner's DICCIONARIO DE USO DEL ESPAÑOL, DR those in the most recent edition (1984) of the Royal Academy dictionary and "SU" those in the supplement appended to the 1971 edition of that dictionary. "N" indicates that the item has previously been encountered and classified as a neologism, in the sense that it is *not* contained in the dictionaries mentioned. In the case of headword items new to the d/b there are two responses. In the first category the coding PR identifies neologisms with a productive prefix. Productive prefixes are those which have been so designated in the d/b, which currently contains a list of more than 220. All lexicalized items incorporating these prefixes were entered in the d/b as part of the initial corpus. That arrangement now ensures that new terms containing such prefixes *must* be neologisms; hence items marked PR do not require manual checking.

The second category comprises all other items new to the d/b. In the printout these are preceded by a series of points, a coding which is designed to be readily visible when one is checking for new items and which serves at the same time as a space in which to record the results of the manual checks made against the two standard dictionaries used.

This program also retrieves and lists all terms nested under the same keyword – see the word *sesenta* in Appendix listing (2) –, and prints out all the citations attached to each headword form. The information thus provided greatly facilitates the evaluation of the items in the source document, for a quick comparison of resident citations and the new one being considered can establish whether the new linguistic environment of the term offers additional semantic, grammatical, morphological or syntactic information which would indicate the suitability of its inclusion in the d/b. In the case of *(-)río*, for example, the d/b already contains sufficient evidence to confirm its use as an adjunct, in a sizable number of citations, in a spread of sources and over several years. The new citation, therefore, will have to offer a distinctively new nuance of meaning or provide a particularly telling collocation, if it is to merit inclusion.

As already mentioned, the JOB/SPCOMP program also generates automatically a provisionally coded version of the original list of headwords – (3) in the Appendix. In this listing some symbols – rather than letters – have been chosen to code the words, as they can readily be seen, a desirable feature as such listings must be checked by the human eye rather than by machine. The coding is as follows: "/" = MM; "." = DR, and "," = SU. In contrast, the absence of a symbol (evident in the great majority of terms) indicates that the headword is a neologism that is already resident or a new one to be added.

At this stage the coding which had been automatically inserted may be modified in the light of the additional linguistic information provided by the new source text citation. Thus, an entry hitherto coded as recorded in MM may be reclassified as "N" as a result of evidence that the term has undergone an extension of meaning (e.g. *modelo* has acquired the meaning "fashion model') or if the existing entry is grammatically incomplete (e.g. *veterinario* should be marked *m.* and *f.* rather than just *m.*) At this stage, too, items new to the d/b will be coded in accordance with the results of the dictionary consultation. These changes are effected through on-line editing of the coded datafile.

1.1.3. Headword entry and confirmation. The fully coded file is now run against the d/b in update mode. All headwords in the file not resident in the d/b are added to it, together with a record of the code appropriate to the status of the headword. All new items are automatically assigned a reference number and a printout confirms that the entry of the new data has been satisfactorily completed.

In terms of efficiency, it should be noted that compared with the manual methods initially employed, the gain is at least tenfold in the processing of potential neologisms. There is also a clear reduction in the possibility of error. The advantages clearly increase in step with the growth of the corpus.

1.2. Citation entry. Once a headword form has been confirmed as a neologism, one or preferably more textual citations to document the usage and linguistic features of the term are entered in the CITATIONS data-set in the d/b. Here, too, the procedure employed has been designed to maximize the economy of effort and material; the system is also designed to permit the use of non-specialist clerical assistance in the data-creation aspect of the system.

The fully-coded headword listing is now used as a turn-round document, serving as the basis for a guide to be followed by keyboarding staff when creating the corresponding citation file. Such a listing may be seen at (4), which corresponds to the citation listing file (5). Entries in the list which are not neologisms or for which the potential citation in the source text is not required are scored out; the source coding and page indications are added, together with any additional instructions. At the same time the citations are delimited in the source text. With that information the typist can work through the headword list, the citations being readily located as they follow the pagination of the source text. To facilitate the task, the formatting of the citation file has been kept very simple (see (5)). The first record always contains the source code in full; it is preceded by an asterisk which signals the start of each citation entry cycle. Thereafter, only the changes in page numbers require to be entered. The first citation is then typed in. Since it has been found that citations frequently contain more than one headword, the input file is formatted to exploit that fact, as may be seen in the second citation, beginning on line 700. As each headword is encountered in the citation text it is delimited by control characters, which have a threefold function: in addition to identifying the headword form in the citation, they serve to highlight it on the VDU screen in on-line interrogation and they can also be used to control an underlining or italicising function in the final hardcopy printout.

The citation text is followed by one or more records each containing one of the headwords cited. These records list the headwords in the order in which they occur in the citation; for processing purposes they are identified by a comma which precedes each one and they coincide with the standard dictionary form in which the headwords are stored in the neologisms data set.

Thereafter, an asterisk marks the start of each new citation/headword cycle, the end of the file being signalled by a double asterisk record. When completed, the citation file is manually proof-read and corrections are entered on line. The corrected citation datafile is then run against a special program which locates and identifies any errors in format or any mismatches between the headwords in the citation file and those in the d/b. The program produces a printout of the results to facilitate any necessary corrections. Following a clean run of this program, an update run is carried out, adding the citations and headword data sets to the d/b. At this point, a printout confirms that the citation file has been added *in totoor,* if for any reason a problem is encountered, it indicates the line in the file at which the entry program was aborted and identifies the reason.

1.3. English equivalents. The addition to the d/b of the English translations of the Spanish headwords is implemented through a separate set of programs. Grammatical, morphological, usage and register details corresponding to the headword are entered at the same time. Each headword is provided with up to nine discrete descriptive fields for recording these data.

The data are prepared using another turn-round document which results from the printing of a selected listing of neologisms from the d/b, accompanied by the generation of a disk file containing these headwords, e.g. all headwords sharing an initial letter, all headwords entered after a certain date, or all headwords lacking an English equivalent in the d/b. The related citations are run off on a parallel list. The English equivalents and the other details are then prepared and entered in the disk file, which is subsequently run against the d/b in update mode, creating the appropriate entries in the d/b. A sample of this material is shown in the datafile marked (6). The English equivalents that are considered suitable for inclusion in the reverse English/Spanish dictionary are so coded in the turn-round document: the preceding oblique ("/") marks items that are to be included and the colon (":) indicates those that are excluded.

1.4. Other database functions. By using as selection markers the information in the descriptive fields attached to the headwords, specialist or restricted glossaries may be directly generated form the d/b. Another retrieval program can be used to list headword items sharing a common prefix or initial combining element; it also retrieves items sharing the same suffix or end combining element. Finally, it can list items sharing a common embedded element; thus one could print out, for example, all the phrase compounds in the d/b containing the preposition *por.*

2. Formal database structure

The database is written as a DMSII system and implemented on the Burroughs B6930 computer in the Computer Centre of Heriot-Watt University, Edinburgh. There are three major data sets: NEOLOG2 which holds the Spanish headword, CITATION which holds citations illustrating headword usage and CATEGORY which carries the grammatical and allied information. There are three other data sets concerned with cross-referencing between headwords and citations, the identification of productive prefix forms, and a "housekeeping" check of the headword files that have been entered. Finally, there is a data set ENGLISH, which contains the English equivalent of the Spanish headword or phrase.

2.1. The dataset NEOLOG2. Each record in the NEOLOG2 data set contains the headword in its standard dictionary form. The headword may in fact be in phrase form with a keyword identified for alphabetical sorting and for nesting of selected terms in print-outs. The data set also contains the dictionary-order form of the headword, with all characters in uppercase form, diacritics and punctuation removed. In the case of a phrase form, only the keyword is retained: e.g. for *los últimos sesenta* (where the keyword *sesenta* is picked out) the dictionary order form is SESENTA. This technique enables "nesting" and differentiation of different semantic usages of a headword, by presenting, in each case, the headword followed by two spaces and either a word string in which the headword appears in its particular usage or a series of numbers, corresponding to superscript numbers in conventional presentation.

Spanish has a number of dictionary-order variants in comparison with English, e.g. the ordering of *ñ* between *n* and *o*. In order to take account of this the uppercase dictionary-order form has *ñ* or *Ñ* replaced by *NZZ*. Similarly, the correct ordering of *ch* and *ll*, whether as initials or embedded in a headword, is correctly handled in the dictionary ordering.

The coding of headwords into a dictionary-order form is important in bringing together some of the typographically variant forms that are likely to occur in new usages within a language. For example the many possible variants of the loan acronym USA that occur in current Spanish, including *U.S.A, U. S. A., Usa* and *usa* are all linked by the dictionary-order form *USA*. Interrogation of the database using any one of the set will retrieve data on all such related forms.

2.2. The dataset CITATION. The CITATION data set contains entries that document the occurrence of one or more headwords that have been identified as neologisms. A text-length of up to 600 characters is allowed for. The record also contains a standardized source code and date representation for the citation and an index-number in separate fields. The use of ISO date-coding enables searches to be made for citations from a given year. Because the relationship between headword and citation is many-to-many, a cross referencing data set HW-CIT has been incorporated in the database.

2.3. The dataset CATEGORY. Provision is made for a maximum of nine grammatical, morphological, usage and register descriptive fields for each headword. These are held in the CATEGORY data set and are indexed by headword index-number.

2.4. The dataset ENGLISH. Each record in the ENGLISH data set contains the English equivalent of a headword. This is held, as in the case of a Spanish entry, in its full form (with any keyword distinguished), and in its dictionary-order form. If the English equivalent is a single word or has a keyword, the entry may be marked for inclusion in a computer-generated reverse English/Spanish dictionary.

2.5. The dataset PREFIX. The PREFIX data set currently contains a list of about 220 different productive prefix forms. These are used in the initial comparison of potential new headwords with the database. All forms embodying these prefixes which are recorded in standard dictionaries were entered initially in the database. Any new terms encountered containing any of these prefixes can therefore automatically be identified as a neologism. This check is carried out first. Further active prefixes are added as and when they are identified.

3. Applications programs

Several kinds of applications programs are used in the creation, update and analysis of the database. The primary method of entering headwords and citations in the database, testing for the presence of a particular headword and performing selective analyses of the database is to use background batch programs. On-line transaction-based programs are used for the editing of individual headword and citation entries.

In the future on-line interrogation of the database will represent a greater aspect of the usage of the content of the system, since one of the major purposes of the system design has always been to offer an on-line enquiry facility to undergraduate linguists who have no computer expertise, but who need access to information on current language usage.

A number of utility programs have been prepared for full or partial alphabetic listings or for selective interrogation of the database e.g. data on the occurrence of particular prefix or suffix forms. The use of the supplied on-line inquiry languages DMSII INQUIRY and ERGO (Extended Retrieval with Graphical Output) enable ad-hoc enquiries to be made, either for screen presentation or on hard-copy.

4. Use of computer resources

The database currently contains 86,500 headwords of which more than 58,000 are neologisms, together with 82,000 citations. The storage used by the database is approximately 50 megabytes. Applications programs and datafiles held on line amount to a further 5.5 megabytes.

Initially designed as an aid for students, the project has also proved useful in other ways. Materials have been provided on contract for a new Spanish-English dictionary that is being prepared by a London-based publisher. Listings have also been made, on the same basis, of financial and commercial terms for a specialist business dictionary. It is our intention to produce, at a later date, a "portable" version of the d/b, thereby making the materials more widely available.

Cited dictionaries

DICCIONARIO DE USO DEL ESPAÑOL (MM)
Maria Moliner, Madrid: Gredos (2 volumes 1984).
DICCIONARIO DE LA REAL ACADEMIA ESPAÑOLA (DR)
Real Academia Española, Madrid (2 volumes 1984).

Appendix

1. Initial list of potential neologisms

tremendismo
condicionante
problemático, -a
recocido
seudoneologismo
término totalmente nuevo
sesenta
tridimensional
(–) río

3. Partially coded list generated by JOB/SPCOMP program

/tremendismo
.condicionante
,problemático, -a
/recocido
seudoneologismo
término totalmente nuevo
sesenta
tridimensional
(–) río

2. Comparison listing generated by JOB/SPCOMP program

```
tremendismo                                05815  tremendismo
condicionante                              20793  condicionante
problemático, -a                           10838  problemático, -a
recocido                                   42489  recocido
                                        MM 82230  recocido, -a
                                                  seudo
seudoneologismo
.término totalmente nuevo
sesenta                                    25497  sesenta los sesenta
   Tamames 83/143         Pero la España de *los sesenta* comenzó a ser *mayoritariamente*
                          urbana y en las ciudades se impuso la *escolarización* de los menores.
                          Ya no hubo pareja trabajadora ni mujer *concientizada* que no exigiese
                          a la sociedad, como subsidio laboral, una plaza escolar lo más cercana
                          posible a la vivienda.
   D 82.12.12/15          Se presentata, en los círculos influyentes de la Barcelona a punto de
                          dejar los "felices *sesenta",* como un conspicuo demócrata liberal,
                          ferviente monárquico, convencido *juanista.*
   Huertas 78/59          El *LSD,* que tuvo una máxima explosión de popularidad por parte de
                          los medios de comunicación en la segunda mitad de *los sesenta,* es hoy
                          la única droga procedente enteramente de laboratorio.
   PU 84.03.22/2          Ya quizá por *los sesenta,* el ciudadano de a pie se había percatado de
                          que alguien, . . . se lo estaba llevando "crudo".
   C 84.06.25/113         Lo que se juega, finaliza Hayden, es una filiación histórica: Feagan
                          combate con los temas de *los veinte* y los cincuenta los logros (y los
                          fracasos) de *los treinta* y *los sesenta.*
                                                N 75477 sesenta los últimos sesenta
   P 83.02.24/26          En *los últimos sesenta,* un señor de chaqueta a cuadros blancos y
                          negros, *menuditos,* fue a verme al Gran Café de Gijón, y me llevó al
                          Hotel de Suecia para allí comerme el *tarro* con inminencias judiciales,
                          sólo porque yo había citado en un artículo a Fumasa.
                                                N 84928 sesenta los primeros sesenta
   PS 87.02.22/R–168      Profesional desde *los primeros sesenta,* algunas de sus canciones han
                          dado la vuelta al mundo.
(–)río                                          02961 (–)río
   Prado 81/65            Existen en radio la *"entrevista río"*) otras que no trataremos aquí
                          por alejarse mucho del campo informativo.
   D 84.04.06/47             Tras estrenos de muy diverso género – un espacio infantil, un docu-
                             mental y un *relato-río* canadiense – que se convierten en adelantado
                             de los escasos estrenos que trae bajo el brazo el segundo trimestre.
   P 81.07.19/L–6            Plaza y Janés, verdaderamente especializada en estas *narraciones,
                          *  generalmente río* y más fuertes que la vida . . ., ha editado reciente-
                             mente
```

4. Coded headword listing prepared for creation of citation file

índice de precios al consumo
economía del lado de la oferta
ley de Say
PC 2
intermediación
mibor
/subsanable
intertítulo
desintermediación
trading

5. Citation file (based on 4)

*Tamames 86/329
Es subordinar todo el funcionamiento económico a la evolución de un solo indicador como el IPC, el *índice de precios del consumo,* basando así la cura en un solo síntoma: la fiebre inflacionista.
,índice de precios del consumo
*
Todo ello, basado en los postulados de la llamada *economía del lado de la oferta,* favoreciendo la capacidad de producción empresarial, en una especie de resurrección de la *ley de Say* ("toda oferta crea su propia demanda").
,economía del lado de la oferta
,ley de Say
*Mediante la Red Local USERNET, la más avanzada del mercado, se consigue una perfecta integración de la información entre los diferentes usuarios de los SPERRY PC/IT, PC/HT y cualquier otro *PC* compatible.
,PC 2
*412
No es mayor la competencia extranjera cuando atienda a los grandes clientes, multinacionales o no, reduciendo al máximo el margen de *intermediación* del dinero, porque emplea los recursos y el tipo de interés del mercado interbancario de Madrid *(mibor)*.
,intermediación
,mibor
*413
Los factores que han influido en la cuenta de resultados del Banco en 1985 han sido la baja de los tipos de interés, la
desintermediación en Pagarés del Tesoro, los beneficios obtenidos en el *"trading"* de activos y la mejora del riesgo crediticio. (Pub.)
,desintermediación
,trading
**

6. Extract from entry file containing English equivalents

derrochón, -a *1adj / wasteful / extravagant
(–)estrella *ladj *2inv / star / leading / outstanding
sandwich *1nm *2pl -es / sandwich
salvaje *1adj / unauthorised / illegal
LOAPA *2acron: Devolution Harmonization Act
imperio de la ley: the rule of law
tabla de quesos / cheese-board