Eugenio Picchi
(Istituto di Linguistica Computazionale, CNR, Pisa, Italy),
Carol Peters
(Istituto di Elaborazione della Informazione, CNR, Pisa, Italy)
Elisabetta Marinai
(ACQUILEX Project, Istituto di Linguistica Computazionale, CNR, Pisa, Italy)

The Pisa Lexicographic Workstation: The Bilingual Components

ABSTRACT: The main components of the Pisa Lexicographic Workstation are a full text retrieval system and a lexical database system; each system incorporates procedures that have been implemented to meet the specific needs of the lexicographer. The paper describes the recent tailoring of existing modules and the development of new ones with bilingual lexicography in mind. The aim is to provide a flexible, user friendly system that can be employed in all stages of dictionary compilation, from the acquisition of citation material to the formatting of the entry for printing.

1. Introduction

For some time now, a lexicographic workstation has been under development at the "Istituto di Linguistica Computazionale" in Pisa. The workstation provides a series of tools designed specifically for linguistic and lexicographic text processing tasks which can be used by the lexicographer to assist him/her in the various activities involved in the creation and revision of dictionaries. The main components of the workstation are the DBT (Data Base Testuale), a full text retrieval system that has been developed to query and analyse all kinds of texts and textual corpora (Picchi 1991), and a lexical database system that has been implemented to handle dictionary acquisition and processing activities; a morphological procedure is associated with the text and dictionary query systems. The lexicographer can use these two systems, the DBT and the LDB, to interrogate on-line text archives and electronic dictionaries and retrieve and extract reference and citation material. The core module of the system is a procedure for on-line dictionary editing which includes functions for windowing into and copying data from the dictionary and text archives, and is integrated with a structured indexing procedure that can be used to query the dictionary in compilation in order to check the regularity and consistency of the input. The present paper describes the recent specialization of existing modules and the development of new ones to meet the specific needs of the bilingual lexicographer.

2. The Bilingual Components

The bilingual modules consist of an on-line bilingual entry editor, a bilingual lexical database query system, and a system for the automatic creation and retrieval of parallel concordances from bilingual text archives (DBT-Synchro). As the languages which are currently being considered for the bilingual components are Italian and English, we have also added a morphological procedure for English to complement the program already implemented for Italian. Figure 1 shows the global structure of the integrated mono-/bi-lingual lexicographic workstation.

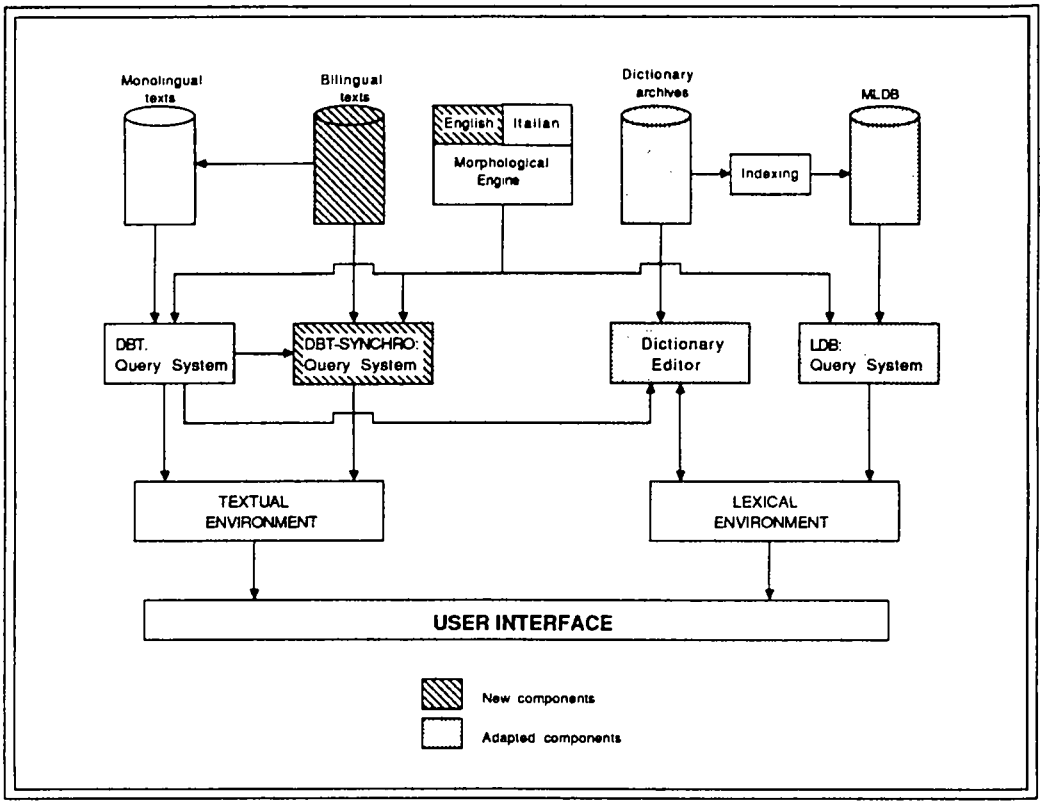

Figure 1 Integrated Monolingual and Bilingual Lexicographic Workstation

2.1. Bilingual Dictionary Editor

The bilingual editor is a specialized version of the on-line editor developed for monolingual dictionary compilation, providing functions to assist the bilingual lexicographer in creating or revising lexical entries. When the editor environment is entered, the user is presented with a basic entry template. This schema represents explicitly, in tagged fields, all the lexical information (phonetic, syntactic and semantic) which will be contained in the printed dictionary. The default entry template has been designed on the basis of experience acquired during the study of a standardized representation structure for dic-

tionary entries in the ACQUILEX project[1] (see Calzolari et al. 1990). The use of a uniform representation language ensures compatibility between entries, permits the exchange of data between different projects in a common format, and facilitates analyses over different dictionaries[2]. However, the lexicographer is free to modify the structure of the template proposed by the system to meet the particular characteristics of the dictionary or lexicon on which he is working.

The source language lexicographer, whose task it is to analyse the headword and make a first proposal of sense divisions and example material, compiles his entries filling in the appropriate data fields; they can then be saved on file or printed out and passed on to the target language compiler who is responsible for transfer into L2 but who may also propose changes to the original analysis on the basis of L2 dependent factors. The final synthesis of the entry will be the result of the joint efforts of both lexicographers. The results of each stage can be saved separately so that a trace of the development of the entry, through the several stages of analysis, transfer and synthesis, is maintained. At any moment, the lexicographers can access both monolingual and bilingual text corpora and LDBs from within the editor environment, which includes functions to window into, query and copy information from these reference archives. The structure of the entry and the main functions available for bilingual dictionary editing can be seen in Figure 2 which shows a screen dump from a work session in which the Help function has been invoked.

```
E. Picchi - Bilingual Dictionary Editor - Italian/English          CARRO
 Hdwd   1 carro                                    ═══ Help ═══
 Pron   2 ['karro]                                 '↑'    Previous field
  PoS   4 sm                                       '↓'    Next field
 NPoS   5 N                                        '"'    Duplicate field
SenNo   6 (a)                                      'C'    Change field code
Trans   7 cart, wagon;                             'D'    Delete field
SICon   8 per carnevale                            'E'    Edit field
Trans   9 float;                                   'I'    Insert new field
   Ex  10 mettere il =**= avanti ai buoi           'J'    Join two fields
SemCd  11 fig                                      'R'    Restore a field
 ExTr  12 to put the cart before the horse.        'S'    Split text field
SenNo  13 (b)                                      ALT+F8  Restore entry
SubCd  14 Astron                                   ALT+F9  Copy entry
   Ex  15 il Gran/Piccolo C=**=                    ALT+F10 Delete entry
 ExTr  16 the Great/Little Bear.                   F3      Quit entry
  Mwd  17 =**= armato                              F4      Save entry
SubCd  18 Mil                                      F5      Previous entry
MwdTr  19 tank;                                    F6      Next entry
  Mwd  20 =**= attrezzi                            F7      Call MLDB
SubCd  21 Aut                                      F8      Call DBT
MwdTr  22 breakdown van;                           F9      Call DBTSynchro
                                                   ═══ ↓ for more ═══
Select Function:
```

Figure 2: Bilingual Dictionary Editor

During a work session, parts of the dictionary being compiled can be extracted and printed or saved in separate files. In this way, particular subsets of the dictionary can be treated independently: a) a given subset of the lexicon; b) a given subset of the data fields; c) only entries satisfying certain constraints. These facilities can be adopted when it is

desired to have certain data fields compiled separately (e.g. the phonetic transcriptions) or to make certain consistency checks throughout the dictionary (e.g. for semantic label attribution).

The compiled lexical entries can be input to the parsing and indexing procedures of the bilingual lexical database system, and can then be checked interactively by the lexicographer using the LDB query system (see 2.3 below). Another procedure that can be used on the entries in LDB form to check the consistency of the data is the Reverse procedure, which operates on the data contained in the translation field of each entry in order to recognize certain incongruities between the two bilingual data sets. The procedure identifies (i) all cases where translations on one side of the dictionary are not listed as headwords on the other, and (ii) all cases where words, appearing as translations of a lemma on one side of the dictionary, when listed as headwords on the other side do not give the lemma as one of their possible translations. There are a number of reasons why symmetry is not always desirable but the bilingual lexicographer should have access to such information during the construction of the dictionary. He can then decide whether an omission is deliberate or due to an oversight, and whether to make amends or not. The Reverse procedure can also be executed in a similar way on words appearing in the Example Translation field. In each case, the procedure is run automatically and the results are printed out for verification by the lexicographer.

The completed lexical entries can be printed out in various formats, or used as input for photocomposition systems which will produce the final version of the dictionary.

2.2. Morphological Procedure for English

Our morphological system consists of a language independent set of procedures which operate on a suitably encoded description of a language in order to recognise and produce words in that language. We adopt a two-level approach rather than a generative one[3]. The language description is formulated in two files: a lexicon file containing a list of base lemmas with associated morphosyntactic information and an inflection code; a rule file containing the rules which specify the correspondences between underlying lexical items and surface forms. The program is reversible: the same lexicon and set of rules is used for recognition and for generation. Our guiding principle has been efficiency and convenience of implementation.

This module has already been implemented for Italian (see Picchi and Calzolari 1990). For the English version we have derived our list of lemmas from the headwords of two computerized dictionaries. Each lexical entry has associated morphosyntactic information. A code has been assigned (semi)automatically to each lemma and is used to invoke the rule which determines its inflection. Rules have been written to cover all regular inflections and those irregular inflections which can be grouped together into classes. Highly irregular inflections are encoded singly. Information on irregular morphology and other phenomena such as gemination has been extracted from the computerized dictionaries; we use orthographic rather than phonological information as this information is more conveniently derived from our dictionaries.

The morphological tool includes an on-line display and editor which can be used to view the generation of the word forms for any lemma in the lexicon and to add to or correct either the inflectional or morphosyntactic codes if necessary. The lexicographer

can use the generator when querying text or dictionary archives to expand any given lemma by producing the set of all its forms; the whole paradigm for the lemma can then be searched by entering a single command. The complete morphological procedure (English/Italian analyzers and generators) is also an essential component of the bilingual text retrieval system.

2.3. The Bilingual Lexical Database Query System

A major component of the set of tools implemented with the bilingual lexicographer in mind is the bilingual lexical database system. The various stages in the design and development of this system have already been described elsewhere and the bilingual LDB now forms part of a Multilingual Lexical Database System (MLDB) that is being implemented in the context of the ACQUILEX project (see Marinai et al. 1990).

The bilingual LDB query system provides dynamic search procedures which permit the user to navigate through the dictionary data and within the different fields of the entry in order to access and retrieve information of interest in whatever part of the dictionary it is stored, specifying the language on which the query is to operate. In this way, much information which is normally "hidden" in the printed dictionary can be accessed and exploited. The query system supplies a series of functions which can be used to look up lexical items or combinations of items. The user can search given items or character strings, define search functions in which items or character strings are combined by AND, OR and NOT operators, retrieve all entries satisfying the search conditions, display, print or store on file all or a selected part of the results, and define restriction rules on the results of a previous search. Searches are made on an attribute-value basis and the results are given for each field in which the item is found. The lexicographer can use the query system on existing dictionaries maintained by the system or on the dictionary under construction, which can be input to the MLDB parsing and indexing procedures and structured as an on-line database.

```
   LDB (E. Picchi)          Collins Bilingual Dictionary (Normalized) V
   Item : {I}<Trans  > CARRO              Frequency :  9

1)  CHARIOT {Hdwd}  chariot {PoS}  n {NPoS}  N {Trans}  cocchio, carro.
2)  DUSTCART {Hdwd}  dustcart {PoS}  n {NPoS}  N {Trans}  carro della
nettezza urbana or delle immondizie.
3)  FLOAT gen {Trans}  galleggiante {m}; {SI}  cork {Trans}  sughero;
{SI}  in procession {Trans}  carro; {SI}  sum of money {Trans}  somma.
4)  HEARSE {Hdwd}  hearse {PoS}  n {NPoS}  N {Trans}  carro funebre.
5)  HORSEBOX {Hdwd}  horsebox {PoS}  n {NPoS}  N {Trans}  carro or
furgone {m} per il trasporto dei cavalli.
6)  TANK {Ex}  fuel =**= {ExTr}  serbatoio del carburante. {SenNo}  (b)
{SubCd}  Mil {Trans}  carro armato.
7)  TRUCK {PoS}  n {NPoS}  N {SenNo}  (a) {SubCd}  Rail {SI}  wagon
{Trans}  carro {m} merci {SI}  inv. {SenNo}  (b) {SI}  esp Am: lorry
8)  WAGGON, wagon {HomNo}  1 {PoS}  n {NPoS}  N {SI}  horse-drawn {Trans}
carro; {SI}  truck {Trans}  camion {m inv}; {SubCd}  Rail {Trans}  vagone
9)  WRECKER {SI}  breaker, salvager {Trans}  demolitore {m}; {Usage}  Am
{SI}  breakdown van {Trans}  carro {m} attrezzi {SI}  inv.

                Continue                Select
```

Figure 3: Querying the Bilingual Lexical Database

Figure 3 gives just one example of how the LDB can be used to acquire information on a given item that is scattered throughout the dictionary and otherwise inaccessible. The Italian lemma CARRO, shown in Figure 2, was searched throughout the LDB which has been constructed for the Collins Concise English/Italian dictionary; the figure displays all those entries in which CARRO was found in the Translation field.

A procedure has also been implemented to permit semi-automatic mapping between monolingual and bilingual LDBs in the workstation. This procedure provides the bilingual lexicographer with a useful tool that permits him to examine and compare the lexical information given for the same item in different source dictionaries (for a full description see Marinai et al., forthcoming).

2.4. Bilingual Text Retrieval

The most recent component to be added to the Lexicographic Workstation is a system for the automatic construction and retrieval of parallel contexts from bilingual text archives. The importance of large language reference corpora in monolingual lexicography is widely acknowledged and it has also been asserted that such corpora are useful for the bilingual lexicographer (see, for example, Atkins 1990). However, interest is now growing in the potential of bilingual corpora as valuable sources of documented evidence on the relationships between two languages. We feel that ideally the bilingual lexicographer should have access to both sources. The monolingual corpus will be used during the analysis stage helping to achieve an accurate first breakdown of a given headword into senses and to retrieve valid examples of usage and collocations; the bilingual corpus will be employed in transfer helping to find appropriate real world translations for each usage of the L1 word suggested by the source language compiler, reflecting parallelisms or differences in sense divisions. For example, a particular sense of a word in L1, established on the basis of corpus evidence, may well have no single equivalent or set of equivalents in L2 to cover the full scope of L1. The bilingual corpus can be used to study carefully how each use is rendered in order to group the TL equivalents. It may well provide evidence which suggests an adjustment of an L1 sense division to meet the demands of L2.

Of course, this implies the availability of a high quality, sufficiently representative bilingual corpus. However, the construction of a text corpus implies a considerable investment of time and resources. Before any decisions are taken, the criteria to be adopted when assembling corpus material must be carefully evaluated. In particular, each type of text sample must be labelled, as source and target texts reveal different uses of language; it is claimed that a translation is never a true representation of the language in which it is written but rather reflects the relationships between the language of the target and that of the source. At Pisa, we have assembled a sample set of bilingual texts, selected to cover a number of different language varieties, ranging from scientific papers to poetry, from university text books to magazine articles. This set of texts was collected in the first place to provide a test-bed for our bilingual retrieval system but should also supply useful data to assist us in the definition of design criteria, which can then be used in a subsequent extension of these archives.

So far, most of the bilingual concordancing tools implemented for the lexicographer use statistically based programs to align the texts at the sentence level. But, as stated by

Church and Gale (1991), such sentence based concordance programs are not very good at showing what is not already known as the user is requested to supply the program with both a SL word and a TL candidate translation. These authors also describe a word-based concordance tool in which the possible translations for a given word are discovered from the corpus by the program, using a pre-computed index indicating which words in one language correspond to which words in the other. We have adopted a different approach which depends on the use of external evidence (extracted from a computerized bilingual dictionary) to create direct links between parallel texts on the basis of translation equivalents. In this way, we exploit already known information (dictionary translations) to access "unknown" information (real world TL renderings).

A preliminary version of our system, known as DBT-Synchro, is described in Marinai et al. (1991). It operates in two distinct steps. In the first, sets of bilingual texts are "synchronized" using morphological procedures and a bilingual electronic dictionary (based on the Collins LDB). Each word form in the text taken as the source is input to the morphological analyzer for L1 in order to identify its base lemma, which is then searched in the bilingual LDB. All translations read for this lemma are input to the morphological generator for L2 which produces all possible forms and these are then searched over the relevant search zone in the target text. When one of the translation equivalent forms are found in the L2 text, a link is created between this form and its equivalent in the L1 text. These links are then stored with the texts in the bilingual archives to be used by the query system for the on-line construction of parallel contexts. "Wrong" links between falsely recognized translation equivalents which disturb context calculation are identified and eliminated by the query system, which then recalculates the parallel contexts on the basis of those links recognised as valid.

```
D.B.T. (Picchi)          Testini di prova della sincronizzazione    V
{E}CHANGE & {E}COLOUR
  1 {E} September days continue, followed by those splendid October days
marked by golden sunset and skies which change colour from green to gold
as in the Venetian paintings of Cima da Conegliano and Titian.
E-impressi.63
  {I} le giornate di sole della fine di settembre, e le splendide
ottobrate dai tramonti dorati, dai cieli che trascolorano dal verde all'
oro, come quelli che si trovano nella pittura veneta da Cima di
I-impressi.65

  2 {E} the melancholy signs that the season is hurrying to an end.
The corn begins to change colour in the fields. The leaves are partly
green and partly the dry dirty gray colour E-impressi.134
  {I} suoi segnali melanconici di stagione che corre verso la fine. Le
piante del granoturco cominciano a cambiare colore nel campo. Le foglie
sono mezzo verdi e mezzo secche, di un grigio sporco in I-impressi.131

  3 {E} that winter is now just around the corner. The leaves on all the
trees have started to change colour. Everything begins to turn yellow or
red. Though it seems like a time of celebration E-impressi.167
  {I} e che l' inverno è ormai dietro la porta. Le foglie di tutti gli
alberi hanno cominciato a mutare colore. Tutto inizia a ingiallire o a
rosseggiare,Pare un momento di festa e di trionfo I-impressi.161

Scelta N.Contesto>                                  F1 for help
```

Figure 4: Querying the Bilingual Text Corpus

The archives are considered to be symmetric; either of the two languages can be selected as L1. The lexicographer can either search for single word forms or, using the morphological generator, for all the forms of a given lemma. For each L1 word or combination of words queried by the user, the parallel L2 contexts are constructed and displayed on the screen. The word(s) for which the contexts are being created are highlighted and where a direct link for the L1 form(s) being searched exists, the L2 matched word(s) will be highlighted in the same colour. Otherwise, the two directly linked forms which are closest to the point calculated as the middle of the L2 context will be evidenced in a different colour, as indicators of the position in the L2 context of the translation equivalent being searched. Other words which have been linked in the paired contexts can be optionally evidenced. Bilingual concordances of interest can be printed out or saved in a separate file for future reference. Figure 4 gives an extract of the results obtained from our test set of texts for a query requesting parallel contexts for the cooccurrence of CHANGE and COLOUR.

3. Final Remarks

We have given a brief description of the main computational tools which have been implemented in the Pisa Lexicographic Workstation in order to facilitate the task of the bilingual lexicographer. The entire system is implemented on personal computers running the MS/DOS operating system and is intended to run on a Local Area Network so that a team of lexicographers can work in unison, using the same tools and accessing the same reference data. At the same time, the procedures are easily transportable onto smaller desk-top systems for the lexicographer working at home. The system is menu-driven and context sensitive Helps are accessible at all times during a query session. Our main consideration has been to provide tools which are not only efficient but also user-friendly.

Endnotes

1 ACQUILEX is an ESPRIT Basic Research Action which is developing techniques and methodologies for utilising both monolingual and bilingual machine-readable dictionary sources to construct lexical components for NLP systems.

2 The study made for ACQUILEX has also been used by the Text Encoding Initiative group studying dictionary representation as part of the general TEI programme to provide guidelines and standards for the representation and exchange of texts in machine-readable form using the SGML mark-up language (see Ide et al. 1991). We will consider incorporating the final recommendations of this group into our actual data model.

3 For a description of the two-level model versus generative phonology, see Antworth (1990, Introduction).

Bibliography

ANTWORTH, E.L. (1990): PC-KIMMO: A Two-level Processor for Morphological Analysis. Occasional Publications in Academic Computing. No.16, Summer Institute of Linguistics, Dallas.

ATKINS, B. (1990): "Corpus Lexicography: The Bilingual Dimension". In: Computational Lexicology and Lexicography. Special Issue dedicated to Bernard Quemada. I. Ed. by L. Cignoni and C. Peters. Linguistica Computazionale, Vol VI.

CALZOLARI, N., PETERS, C., ROVENTINI, A. (1990): Computational Model of the Dictionary Entry: Preliminary Report. ACQUILEX, Esprit BRA 3030. Six Month Deliverable. ILC-ACQ-1-90 ,Pisa.

CHURCH, K., GALE, W. (1991): "Concordances for Parallel Text". In: Using Corpora. Proceedings of the 7th Annual Conference of the Centre for the New Oxford English Dictionary and Text Research, Oxford, UK.

IDE, N., VERONIS, J., WARWICK-ARMSTRONG, S., CALZOLARI, N. (1991): Principles for Encoding Machine Readable Dictionaries, TEI WP, A15W6.

MARINAI, E., PETERS, C., PICCHI, E. (1990): The Pisa Multilingual Lexical Database System. Esprit BRA 3030. Twelve Month Deliverable. ILC-ACQ-2-90, Pisa.

MARINAI, E., PETERS, C., PICCHI, E.: "A prototype system for the semi-automatic sense linking and merging of mono- and bilingual LDBs". In: Research in Humanities Computing. Ed. by N.Ide and S. Hockey. OUP, Oxford. (forthcoming).

MARINAI, E., PETERS, C., PICCHI, E (1991): "Bilingual Reference Corpora: A System for Parallel Text Retrieval". In: Using Corpora. Proceedings of the 7th Annual Conference of the Centre for the New Oxford English Dictionary and Text Research, Oxford, UK.

PICCHI, E. (1991): "D.B.T.: A Textual Data Base System". In: Computational Lexicology and Lexicography. Special Issue dedicated to Bernard Quemada. II. Ed. by L. Cignoni and C. Peters. Linguistica Computazionale, Vol VII.

PICCHI, E., CALZOLARI, N. (1990): "Pisa Linguistic Database". In: Literary and Linguistic Computing 1988, Proceedings of ALLC Fifteenth International Conference. Ed. by Y.Choueka. Champion-Slatkine, Paris-Geneve.

Dictionaries

GARZANTI (1984). Il Nuovo Dizionario Italiano Garzanti, Milano.

COLLINS (1985). Collins Concise English-Italian, Italian-English Dictionary, London and Glasgow.

PROCTER, P. et al. (Eds.) (1978). Longman Dictionary of Contemporary English, (LDOCE), Longman, Harlow and London.

ZINGARELLI, N. (1970), Vocabolario della Lingua Italiana, Zanichelli, Bologna.

KEYWORDS: computational and computer-aided lexicography, bilingual lexicography, bilingual lexical databases, parallel text retrieval