A TRANSLATOR'S WORKSTATION

EUGENIO PICCHI[1], CAROL PETERS[2], ELISABETTA MARINAI[3]

[1]Istituto di Linguistica Computazionale, CNR, Pisa, Italy
[2]Istituto di Elaborazione della Informazione, CNR, Pisa, Italy
[3]ACQUILEX Project, Istituto di Linguistica Computazionale, CNR, Pisa, Italy

ABSTRACT

A description is given of the present state of development of a workstation that has been designed to provide the translator with efficient and easy-to-use computational tools. The aim is to offer translators fast and flexible on-line access to existing dictionary databases and bilingual text archives and also to supply them with facilities for updating, adding to and personalizing the system data archives with their own material.

1. INTRODUCTION

Over the last few years, at the Institute for Computational Linguistics in Pisa, an open-ended modular set of tools, known as the PiSystem, has been designed and developed to meet the various requirements of literary and linguistic text processing and analyses. The core component of the system is the DBT, a textual database management and query system that has been implemented in different configurations to perform specific text and dictionary processing tasks. Other components can be integrated with this system kernel as required, depending on the needs of a particular application. (For a detailed description of the DBT in its various configurations see Picchi, 1991.) Within this general framework, in the present paper we describe the construction of a Translator's Workstation.

Translators need fast and flexible tools to assist them in the task of rendering an L1 text in L2, as fluently and faithfully as possible. They also need tools that are easy-to-use, relatively economic and wherever possible portable, as many translators are free-lancers and much translating work is done at home. These requirements have been borne in mind in the design of the Workstation.

The Workstation is being constructed around two main components: a bilingual lexical database system and a system that creates and manages bilingual text archives. In addition, procedures are being provided to permit the users to update the basic system archives with their own data. At present, the system languages are Italian and English; however, the procedures are designed to be generalizable: given the necessary lexical components, they could be transported to other pairs of languages. The user can also access monolingual LDBs, and invoke Italian and English morphological programs to query the dictionary and text databases or to check inflectional paradigms. The entire system is menu-driven; the translator is guided in his use of each component by a set of menus, and context sensitive Helps can be invoked to explain the functionality of each command.

2. THE BILINGUAL LEXICAL DATABASE SYSTEM

The bilingual lexical database system was first described in Picchi et al (1990); it now forms part of the MLDB, a multilingual integrated lexical database system implemented within the framework of the ACQUILEX project[1] and described in detail in Marinai et al. (1990). The lexical components of the MLDB include the Italian Machine Dictionary - mainly based on the Zingarelli Italian Dictionary -, and LDBs derived from the Garzanti 'Nuovo Dizionario Italiano', and the Collins Concise Italian-English, English-Italian Dictionary; we hope to add an English LDB shortly.

[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 natural language processing systems.

2.1 Querying the Bilingual LDB

The translator will primarily be interested in the bilingual dictionary data. Using the bilingual LDB system he can retrieve much valuable information for a given lexical item at all levels (e.g. translation equivalents, examples of usage, syntactic information, etc.) which is inaccessible using traditional dictionary lookup. The LDB query system offers dynamic search procedures that permit the user to navigate through the dictionary data and within the different fields of the entry in order to access and retrieve information in whatever part of the dictionary it is stored, specifying the language on which the query is to operate. Any lexical item or combination of items entered as a value is searched in the database with reference to its particular function in the entry and the results (i.e. number of occurrences of the item) are displayed field by field. The user can then select, view and print those results that interest him. Morphological procedures can be used in order to search the entire inflectional paradigm of a word throughout the dictionary; this is particularly useful when looking for information on the usage of a given lexical item in the example fields. A full description of the LDB query language and a complete list of all the functions implemented is given in Marinai et al. (1990).

The translator can also access and query the monolingual dictionaries maintained by the system. The different perspective on the data provided by a monolingual entry often gives a more complete view of a given lexical item and its usage than is provided by the bilingual entry alone. A procedure has thus been implemented to permit semi-automatic mapping between bilingual and monolingual LDBs. Equivalent entries from the separate dictionaries can be combined and links are created between them semi-automatically at the sense level, mainly on the basis of information that can be extracted from definitions, examples and semantic labels. In this way, we create a more complete composite entry which represents the sum of the information contained in the individual dictionaries (see Marinai et al, forthcoming). The translator can use this procedure to access, compare and scan rapidly the lexical information given for the same item in different source dictionaries.

2.2 Specializing the Bilingual LDB

In the version of the bilingual LDB that we are implementing in the Translator's Workstation, the user will also have functions available so that he can add his own information to the bilingual entry. This will be particularly useful for the translator working in a specific domain who may well accumulate information on the usage of particular terms and expressions within this discipline which is not registered in any dictionary. He can call the User Update Procedure which permits him to add to the data in the lexical entries as he wishes, as long as he respects the data representation schema.

The procedure will work in interactive mode. The user calls the lexical entry to which he wishes to add information by entering the headword on the keyboard. The structured and tagged entry is displayed on the screen. The user then invokes a Help function to display the different functions that can be used to intervene on the entry. All the information added by the user is recorded in a special User Memo Section. Within this section, he is given a choice of fields in which he can enter his data. These fields are similar to those used in the rest of the Entry schema, and consist of fields for translations, examples, translations of examples, semantic indicators, and various kinds of semantic labels: subject, usage, geographic and register codes (for a detailed description of the data representation schema we use, see Calzolari et al., 1990). With the exception of a User Note field used for free comments by the translator, purpose-written, dynamic indexing procedures will then be executed on this new data so that it becomes directly accessible for subsequent querying. In this way, the translator is able to exploit and reuse information acquired as a result of his own experience and activity.

3. PARALLEL TEXT RETRIEVAL

The considerable attention now being given to corpus-based studies means that there is also growing interest in the creation of bilingual reference corpora. Such corpora will be important sources of information in many studies of the linguistic phenomena involved in the process of transferring information, ideas, concepts from one language to another as they can provide large quantities of documented evidence on the possible realization of a concept in two languages, according to a number of contextual factors, e.g. usage, style, register, domain, etc.. The chance to access a corpus of this type would be of enormous help to the translator in his search for that elusive 'right' translation equivalent which is so often not found in the bilingual dictionary.

So far most of the systems studied to manage bilingual corpora use statistically based procedures to align the texts at the sentence level. Such programs often request the user to supply not only an SL word but also a TL candidate translation in order to construct parallel concordances. Church and Gale (1991) present a system of this type and also describe a word-based concordance tool in which the possible translations for a given word are discovered from the corpus on the basis of a pre-computed index indicating which words in one language correspond to which words in the other. Our approach to the problem is quite different. We use external evidence provided by a bilingual LDB to create links between pairs of bilingual texts on the basis of SL/TL translation equivalents. These links are then used by the bilingual text query system to construct parallel concordances for any form or cooccurrences of forms found in either of the two sets of texts. A preliminary version of this system is described in Marinai et al. (1991).

At the moment, the system runs on a small sample set of Italian/English texts chosen to be representative of different language styles and thus to provide a suitable test-bed for performance evaluation and the definition of bilingual corpus design criteria. It is now our intention to extend these archives. In the version of the system which has been implemented in the Translator's Workstation, the translator has the possibility of creating a reference corpus from his own material and adding new texts to it as they become available. An easy-to-use interface has been prepared to guide the translator step-by-step as he inputs pairs of texts to the system.

3.1 Creating a Bilingual Corpus

Given a new pair of bilingual texts, the first stage is to structure them in text database format using the DBT procedures. The texts are scanned to recognize and identify the different elements composing them. For example, word forms are distinguished from the other tokens, such as punctuation marks, numbers, line and paragraph breaks; codes are added to distinguish between full stops and abbreviation marks, between dashes and hyphens, between the different use of the apostrophe in Italian and in English, etc.. This stage is simple, rapid, and once a few preliminary instructions have been given, automatic.

Once a pair of texts is stored in DBT format, they must be input to the text "synchronization" procedure which establishes as many links as possible between translation equivalents in the two texts. This procedure is totally automatic and operates as follows. Each word form in the text selected as the Source text is input to the morphological analyzer for that language in order to identify its base lemma which is then searched in the bilingual LDB. All translations given for this lemma are read and input to the morphological generator for the TL; all the forms generated are then searched over the relevant zone in the target text. If the procedure finds more than one possible base lemma for a given form the translations for each will be read as, in the case of grammatical homography, it is quite possible that the translation equivalent does not respect the category of the source language and, in the case of lexical homography, it is presumed unlikely that the translations of the 'wrong' lemma will find a correspondence in the target text. A schema of the procedure is given in Figure 1.

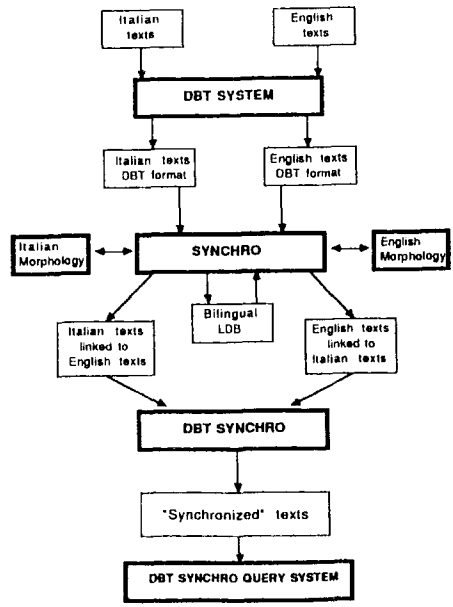

Figure 1. Parallel Text "Synchronization" Procedure

Articles, pronouns, prepositions and a small list of stop words are excluded from this search procedure as of little significance to the matching procedure and liable to create noise. When one of the translation equivalent forms is found in the searched section of the L2 text, a link - consisting of a physical address which locates the equivalent word in the L1 text - will be created. When no entry for a word in the L1 text is found in the dictionary, it may be that the form being examined is either a proper noun or a word from a highly specialised vocabulary not included in our bilingual LDB. An attempt is thus made to match such forms against any equivalent character strings in the relevant zone of the L2 text, ignoring the last characters to allow for morphological variations as, in the two languages in question, proper nouns and scientific terms frequently resemble each other. The matching procedure continues, word by word, to the end of the L1 text.

The execution of the "synchronization" procedure is rapid and totally transparent. When it is completed, the results are presented to the user in terms of the number of successful "matches" of translation equivalents between the Source and Target texts. The procedure will be considered to have "failed" if the number of matches is less than a given percentage of the total text. This procedure must be executed just once for each pair of bilingual texts, when they are added to the archives.

3.2 Querying a Bilingual Corpus

When the bilingual texts have been processed by the synchronization procedure, all the links obtained are memorized in the text archives so that they can be used by the parallel query system. The bilingual text system provides functions to query the bilingual archives and retrieve parallel contrastive contexts. The translator querying the corpus must first specify his "source" language, i.e. the language on which the search is to be performed. For each form or combination of forms he searches, the parallel source and target language contexts are constructed and displayed on the screen. The word(s) for which the contexts are being created will be highlighted and, where a direct link 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 likely position in the TL text of the translation for the SL form(s) being searched. The user can either search for individual word forms or, using the morphological generator, for all the forms of a given lemma. The indicators help him to identify the TL equivalents rapidly. Figure 2 gives examples of parallel concordances for the Italian adverbial expression **pian piano / pian pianino** which is used to attenuate or moderate the action of the verb; its translation in English is thus context-dependent.

```
DBT-Synchro (Picchi)                    Bilingual Reference Corpus      V
        {I}PIAN & ({I}PIANINO | {I}PIANO)
4 {I} estremo del campo. L' osservai con indolenza masticando uno di quei
fili d' erba coi quali le ragazze predicono il futuro. Camminava pian
pianino lungo la scarpata. Teneva una mano sul fianco e nell' altra aveva
un bastone col quale saggiava il terreno erboso. I-Dublin2.197
  {E}. I watched him lazily as I chewed one of those green stems on which
girls tell fortunes. He came along by the bank slowly. He walked with one
hand upon his hip and in the other hand he held a stick E-Dublin2.211
5 {I} "C' è tempo" rispose Corley. "Ci dovrebbe già essere, ma la faccio
sempre aspettare." Lenehan ridacchiò pian piano. "Accidenti, Corley, sai
sempre come trattarle" disse. "Li conosco tutti i loro I-Dublin6.150
  {E} enough", said Corley. "She' ll be there all right. I always let her
wait a bit." Lenehan laughed quietly. "Ecod! Corley, you know how to take
them", he said. "I' m up to all their  E-Dublin6.170
6 {I} si sarebbe aperta la strada. Sul tavolo davanti a lui giaceva un
volume delle poesie di Byron. L' aprì pian piano con la sinistra per non
svegliare il bimbo e cominciò a leggere la prima: Tacciono i venti e
immoto l'aer  I-Dublin8.493
  {E} might open the way for him. A volume of Byron's poems lay before
him on the table. He opened it cautiously with his left hand lest he
should waken the child and began to read the first poem in the book:
"Hushed  E-Dublin8.536
        Continue          Interrupt          F1 Help
```

Figure 2 Parallel Concordances for **pian piano/pianino** from the Bilingual Text Archives

"Wrong" links between falsely recognized translation equivalents that 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. We are now considering ways to filter the results so that the user has the option of viewing only that part of them which most interests him, e.g. he could choose to view only those parallel contexts in which there is no direct (dictionary established) link for the SL word being searched. During a query session, bilingual concordances can be selected for printing or saved in a separate file for future reference.

The bilingual text retrieval system is currently implemented for interactive consultation, e.g. by the lexicographer or translator. However, data derived from analyses on bilingual corpora should also provide valuable input for MT systems. For example, Nagao (forthcoming) stresses the importance of including detailed collocational information in the transfer dictionaries of such systems: there are many specific expressions which must be translated in a specific way in a given TL and knowledge of this sort improves the quality of an MT system greatly. To acquire it many collocational expressions with their translations must be accumulated and bilingual texts are important sources of such data. For this reason, we have begun to examine methods by which the results can be synthesized so that the most probable translation candidates for a given expression within the TL context can be identified (semi)automatically.

4. FINAL REMARKS

The components of the translator's workstation are in an advanced stage of implementation. We envisage a final integrated system in which the translator creates his document using one of the commercially available word processing packages from which he can access and query the bilingual (and monolingual) lexical databases or the bilingual text archives whenever the need arises. In this way, not only will he be able to consult much material otherwise inaccessible, but the speed of the system response times means that, to a large extent, it is possible to avoid that interruption to concentration so often involved when it is necessary to stop work to perform a manual look-up of a reference work.

The rapid recent technological progress in the computer hardware world means that it is increasingly possible to provide desk-top tools with large storage capacities at relatively low costs; the workstation is thus being implemented on personal computers and runs under the MS/DOS operating system.

REFERENCES

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, 90p.

Church K., Gale W. (1991), Concordances for Parallel Text, in *Using Corpora*, Proc. 7th Annual Conference of the UW Centre for the New OED and Text Research, OUP, UK, pp 40-62.

Marinai E., Peters C., Picchi E. (1990), The Pisa Multilingual Lexical Database System, Esprit BRA 3030, Twelve Month Deliverable, ILC-ACQ-2-90, Pisa, 61p.

Marinai E., Peters C., Picchi E., A prototype system for the semi-automatic sense linking and merging of mono- and bilingual LDBs, in N.Ide and S. Hockey (eds.), *Research in Humanities Computing*, OUP, forthcoming.

Marinai E., Peters C., Picchi E. (1991), Bilingual Reference Corpora: A System for Parallel Text Retrieval in *Using Corpora*, Proc. of 7th Annual Conference of the UW Centre for the New OED and Text Research, OUP, UK, pp 63-70.

Nagao M., Dictionaries for Machine Translation, *Linguistica Computazionale*, Vol VIII, forthcoming.

Picchi E. (1991), D.B.T.: A Textual Data Base System, in L. Cignoni and C. Peters (eds.), *Computational Lexicology and Lexicography. Special Issue dedicated to Bernard Quemada. II, Linguistica Computazionale*, Vol VII, pp 177-205.

Picchi E., Peters C., Calzolari N. (1990), Implementing a Bilingual Lexical Database System, in T. Magay and J.Zigány (eds.), *BUDALEX '88 Proceedings*, Budapest, 1990, pp 317-329.