eftab200

# Using Linguistic Tools and Resourcesin Cross-Language Retrieval

Carol Peters* and Eugenio Picchi**

*Istituto di Elaborazione della Informazione - Consiglio Nazionale delleRicerche
Via S. Maria, 46, 56126 PISA, Italy
carol@iei.pi.cnr.it

**Istituto di LinguisticaComputazionale - Consiglio Nazionale delle Ricerche
Via della Faggiola, 32, 56126 PISA, Italy
picchi@ilc.pi.cnr.it

## Abstract

A system to process bilingual/multilingual text corpora is described. Thesystem includes components for cross-language querying on parallel (ietranslation equivalent) and comparable (ie domain-specific) collections oftexts in more than one language. Both sets of procedures are dependent on lexical resources (bilingual lexicaldatabases) and linguistic tools (morphological procedures). The system wasoriginally designed to meet the requirements of various types ofcontrastive language studies. However, we are now studying applications to cross-language retrieval.

## Background

In the last few years, natural language processing (NLP) techniques andtools have been incorporated into information retrieval (IR) systems withvarying degrees of success (Smeaton, 1992). The recent emergence of thefield of Cross-Language Information Retrieval as an independent area of interest has clearly reinforced this trend. In order to be successful,cross-language applications frequently need access to methodologies andresources that were originally studied and constructed for NLP purposes,such as morphological analysers and generators, computational lexicons, various kinds of procedures for textanalysis, etc. The integration of such resources into typical IR processesimplies an exchange of know-how and viewpoints between the twodisciplines.

## The PiSystem

At the Institute for Computational Linguistics (ILC-CNR) in Pisa, there has been intensive work overthe last decade on the development of an integrated complex set of mono-and bilingual lexicon and text management and analysis tools, known as thePiSystem. These tools are designed to meet the needs of all kinds of literary and linguistic textprocessing tasks. The core component of the system is the DBT (TextualDatabase) search engine. The DBT system has been implemented in variousversions to process and analyse different kinds of structured and unstructured texts. There is also a client-serverversion running on Internet, known as DBTNET. A morphological engine and apart-of-speech tagger and lemmatizer for Italian are also provided. (Forfull details and demos, see our Web site: http://www.ilc.pi.cnr.it/dbt/episystem/index.htm).

Part of the PiSystem consists of a setof components designed for multilingual or cross-language applications.This includes a system for the acquisition, management and querying ofmono- and bilingual lexical databases (LDBs), morphological data and rules for Italian,English, French, Latin, and a system for bilingual/multilingual textmanagement, which has separate procedures for cross-language querying onboth Parallel and Comparable Text Corpora.Both sets of procedures use other components of the PiSystem:

- DBT
- Bilingual Lexical Database
- Morphological Analysers and Generators

We define *ParallelCorpora* as collections of translationally equivalent textsregarding general language or sub-languages of different types (e.g. for a given author, a particular domain,etc.), and *Comparable Corpora* as homogeneous sets of texts from pairs or multiples of languages with thesame communicative function; they must share certain basic features, whichcould be period, author, style, genre, register, but generally they referto the same domain; they thus regard sub-languages rather than general languages. The two types of corporaprovide different kinds of contrastive data: parallel corpora provide dataontranslation equivalents; comparable corpora give information on naturallanguage lexical equivalents within a given domain[1].

Our corpus procedures were initiallydeveloped to process Italian/English texts but are extendible to otherlanguages. We are now adding modules to handle French texts.The system was originally studied with a number of human-orientedapplications in mind: bilingual lexicography; language learning activities;translating and translation studies; cross-linguistic studies. We are now studying possible applications to CLIR activities.

## ParallelText System

Our parallel text system has been described elsewhere (Marinai et al, 1991,1992) and we will not go into details concerning its implementation here.Its distinguishing feature is that it is based on the use of linguisticresources rather than thestatistical-distributional data employed by the most well known proceduresfor this type of application (see, for example, Warwick and Russell, 1990,Gale and Church, 1993, Brown et al, 1993). Thus, it does not perform asentence-based alignment but uses a bilingual electronic dictionary andmorphological analysers and generators for Italian and English to linktexts and construct parallel contexts through the recognition oftranslation equivalents in pairs of texts.

Our procedures function in two distinct stages:
1. Links are created betweentranslation equivalent pairs in texts being processed. This stage isperformed just once for each new pair of parallel text added to thearchives;
2. The query system uses the links to construct the parallel contexts,in real time, for any form or cooccurrence of forms contained in thetexts.

Full details on how the text linking algorithm operates, how the 'search zone' (area of L2 corpus searchedfor L2 equivalents of a given L1 word) is calculated, how false links areeliminated, and under what conditions failure is signalled, are given inMarinai et al. (1992). It is sufficient to note that, from the user viewpoint, this operation is simple, rapidand, once a few preliminary instructions have been given, automatic. Thereis no need to pre-process the texts, manually indicating, for example, matching sections of the text. Neither is it important that every possibletranslation match between two texts is recognised; the translation links are just the means to construct theparallel contexts as accurately as possible. However, the fact that thesystem uses and evidences information on translations derived from ageneral purpose bilingual dictionary facilitates the recognition and retrieval of 'new' information, ie real worldtranslations for words or expressions not given in the bilingual dictionary(as shown in Figure 1 below).

When the system is queried to retrieveparallel contexts for any word or combination of words from the corpora,the L1 context is constructed with the searched word(s) in the centre;these words are highlighted andduring creation of the context, any translation links to the L2 textassociated with the other words in the context are read. If the searchedwords have an associated link, this is used to identify directly thecorresponding word(s) in

**QUERY = sicurezza5 Contexts Found**
{I} La<u>prese</u> come un <u>avvertimento</u> e, voltatosi <u>indietro</u>verso la <u>casa</u> nella quale era <u>entrata</u> la <u>giovane</u>, per avere la **sicurezza** dinon essere <u>osservato</u>, traverso' di<u>corsa</u> la <u>strada</u> senza aspettare. L'<u>ansia</u> e l'andatura veloce lo facevano <u>ansimare</u> - =FE - I-Dublin6.39
{E} light rainfell. He <u>took</u> them as a <u>warning</u> and glancing <u>back</u> towardsthe <u>house</u> which the **<u>young</u>** woman had <u>entered</u> to see that he was not **<u>observed</u>**, he <u>ran</u> eagerly across the<u>road</u>. <u>Anxiety</u> and his swift run made him <u>pant</u> =FE - E-Dublin6.105

[1] Unfortunately, there is still no general consensus between the appliedlinguistics and the computational linguistics communities on the definitionof these two terms. The applied linguists tend to use them differently from us. Hartmann

{I} Gabriel risecon un certo nervosismo e dovette dare un tocco al cravattino peracquistare **sicurezza**, mentre zia Kate si piegava in due dalle risa tanto le era piaciuto loscherzo =FE - I-Dublin7.128
{E} Gabriel **laughed** nervously and patted histie reassuringly **while** Aunt Kate nearly doubledherself, so heartily did she enjoy the joke. =FE - E-Dublin7.152
{I} Bessie che si occupa di loro." "Benone", ripetè zia Kate. "E' unabella **sicurezza**, una ragazza come quella, una sucui poter contare! C'è quella Lily, invece, =FE - I-Dublin15.232
{E} "Besides, Bessie will look after them." "To be sure", said Aunt**Kate** again. "What a comfort it is to have a**girl** like that, one you can depend on! There'sthat Lily, =FE - E-Dublin15.262
- DBT -- E.Picchi-----------------------------------------------------------------

**Figure 1: ParallelContexts for Italian term *sicurezza* in "TheDubliners"**

the L2 text, which will also behighlighted and used as the central point for the construction of the L2 context; other words that have been linked in thepaired contexts can be optionally evidenced in a different colour. The sizeof the context and the maximum distance between cooccurrences of items searched can be specified by the user. Figure 1 shows the firstthree parallel contexts found for the Italian word *sicurezza* in a parallel corpus consisting of the original text of "The Dubliners" byJames Joyce plus an Italian translation.

When there is no directly linked L2 formfor the L1 word being searched (as is the case in the figure), then all the linksfor words found in the source context are used to calculate an 'average'value which identifies the central point around which the relative L2context will be constructed. The calculation of this 'average' value allows for the possibility of unevenconcentrations of matched words in the contexts. 'Wrong' links betweenpotentially falsely recognised translation equivalents which disturbcontext calculation are identified and eliminatedby the query system. The two linked forms which are closest to the pointcalculated as the middle of the target text context are evidenced in adifferent colour, as indicators of the likely position of the translationequivalent. In the figure, the word searched and the indicators of its position in the target context are shownin bold. Thus, for the first parallel context, we see that the position ofthe L2 equivalent of *sicurezza* ("*per avere la sicurezza*" is translated by "to see that") is found between "young" and"observed"; in the second, (where the concept of *sicurezza*is translated by the adverb "reassuringly") the translation is indicated aslying between "laughed" and "while"; in the third (*sicurezza* = comfort) it falls between "Kate" and "girl". In the first pair ofcontexts only, each word for which a translation link to a word in theother language text has been created is underlined in order to show theinformation the system uses to construct the parallel context and to indicate the position of the target languageequivalent.

The system has been tested on all kindsof texts: scientific articles, extracts from text books, on-flightmagazines, short stories, novels, poetry.It is easy to evaluate the results; the system is successful if, for eachoccurrence of any form or occurrence of forms in the texts in one language,correct parallel contexts are constructed for all the translation equivalents contained in the texts in thesecond language. The only factor that may affect performance is if the userrequests a very small context size. The system default value is 25 tokenseach side of the keyword(s) (the items searched); it is advisable not to request contexts smaller than 15tokens as, in this case, the system may not have sufficient links tocalculate the parallel contexts accurately.

It is now our intention to testthis system in a translation-training application.

The approach described has beencriticised as being costly in terms of resources and only extendible toother pairs of languages if the necessary lexical/linguistic data areavailable. Our answer is that, in the first place, such resources(dictionaries and morphologies) are now widely available in thecomputational linguistics community for most European languages and, ingrowing numbers, for non-European languages. It wouldbe foolish not to use them. Secondly, it is false to believe that alignmentprocedures can be considered as totally language-independent. McEnery andOakes (1996), for instance, show how alignment methods of the typeexemplified by Gale and Church (*op.cit.*) are both language and domain dependent. These authors report recentliterature in which various suggestions are made as to how this type ofalignment can be improved by introducing some surface linguistic knowledgesuch as the notion of language specific cognate words, ie pairs of tokens in a given language which share'obvious' phonologic or orthographic and semantic properties. They go on to describe methods to employ this kind of cognate informationin order to enhance statistical sentence alignment results, although theyalso note that the incorporation of too much cognate information willresult in noise and a degradation of results.

---

(1995), forinstance, has proposed a distinction between bi-texts, for translationallylinked texts, and parallel texts for texts that are functionally similar insituational motivation and rhetorical structure.

The recent trend in parallel corpusprocessing is thus to move towards the use of other than purely statisticaldata; this appears to support our claim that if lexical and morphologicalcomponents are used in procedures for bilingual text alignment, performance in terms ofretrieval precision can be improved.

## ComparableText System

We recently decided to extend the scope of our bilingual corpus system byincluding a set of procedures for the analysis and extraction ofsignificant data from comparable text archives. The aim was different butour interests were still oriented towards human-oriented language learning activities:while with our parallel system the user can retrieve examples of specificinstances of how a given word or expression has been translated in anotherlanguage, depending on context, argument, stylistic considerations, etc., using the comparable system,he/she will also be able to look for natural language examples of L2lexical equivalents of a given word or expression in L1, independently ofany direct translation link. By definition, comparable text archives regard special domains or sublanguages; they arethus of particular interest for studies or applications regardingterminology and for technical rather than literary tests.

Our procedures operate on sets of comparable texts in two differentlanguages. We are currently working on Italian and English texts, and sofar all work has been focused on nouns; in sub-language texts, it is mainlythe nouns that bear the weight of topic-specificity, ie the technical message. The approach is based on the assumptions that (i) words acquire sense from their context, (ii) wordsused in a similar way throughout a sub-language or special domain corpuswill be semantically similar. It follows that, if it is possible toestablish equivalences between several items contained in two different contexts, there is a high probability thatthe two contexts themselves are to some extent similar.  It is important tostress that our aim is not to retrieve precise equivalences in L2 of the L1term under examination, but toisolate the set of contexts in the L2 corpora that has the highestprobability of providing L2 correspondences to the L1 input. Given aparticular term or set of terms found in the texts in one language (L1), weattempt to identify contexts which treat the same argument in the texts of the second language (L2). To do this, weisolate the vocabulary related to that term in the L1 corpus - hypothesising that the word will be surrounded by a similar vocabulary inL2.

A term, T, is thus selected in theone set of texts (denominated as L1 - either set can be chosen as L1). For eachoccurrence of T in the L1 set of texts, the system constructs a contextwindow containing T plus up to 'n' lexically significant words appearing tothe right and left of T, but within the same phrase, ie strong punctuation marks (full stops and semi-colons)act as break points in the construction of these contexts. The value for'n' is set by the user. Words contained in a stop list are not counted.This list includes functional words such as articles, pronouns, prepositions, and also highly frequent,insignificant words which would create noise. The stop list can be modified by the user so that certain frequent terms specific to theparticular domain can be eliminated if necessary to improve performance. Inthe current version of the system, we are accepting just nouns and verbs asbeing relevant for our purposes; we have made tests in which we accepted only nouns butwe appeared to lose some significant information. We have also tested with varying values for 'n' in order to try toestablish experimentally the optimum size of the context window in terms ofsignificance of results weighed against processing times (clearly, thelarger the context window, the longer the time needed to calculate the significant vocabulary for a giventerm).

For each cooccurrence of our keyword Tin the context windows, morphological procedures identify the source lemma.The set of significant words found in the context windows for T make upthe vocabulary, $V_1$, that is considered to characterise T in theparticular L1 corpus being analysed. The frequencies of the cooccurrencesof T are then computed and to each element of $V_1$ is assigned its mutual information value which measures the significanceof the correlation between the $V_1$ item and T,ie the relative frequency of the $V_1$ item as a collocate of T is measured against its overall frequency in thecorpus in order to identify how strongly it is related to T (see Church andHanks 1990). Using the MI index as an ordering element, we  list$V_1$ in order of decreasing significance and  set a threshold below which termsin $V_1$ are not considered relevant and can beignored. Figure 2 shows the significant collocates for the Italian noun*libertà*  found in a set of comparable English and Italian parliamentarydebates. There were 198 occurrences of *libertà*.

| | |
|---|---|
| 0000000 | 16 |
| 500.000 | 198\|LIBERTA' (freedom) |
| 11.259 | 4\|CONDIZIONARE (to condition) |
| 10.094 | 30\|FONDAMENTALE (fundamental) |
| 9.358 | 5\|ESPRESSIONE (expression) |
| 8.965 | 5\|SINDACARE (to inspect/control) |
| 8.722 | 3\|EFFETTIVO (effective) |
| 8.619 | 4\|PIENA\|PIENARE (full\|to fill) |
| 8.573 | 4\|INDIVIDUARE (to identify) |
| 8.550 | 4\|BENEFICIARE (to benefit by/from) |
| 8.204 | 3\|LIMITAZIONE (limitation) |
| 7.696 | 3\|GARANZIA (guarantee) |
| 6.672 | 5\|PRINCIPIO\|PRINCIPIARE (principle\|to begin) |
| 6.155 | 4\|CITTADINO (citizen) |
| 6.040 | 5\|RISPETTO\|RISPETTARE (respect\|to respect) |
| 5.975 | 6\|POLITICA (politics/policy) |
| 5.746 | 5\|TRATTARE\|TRATTATO (to treat\|treaty) |

**Figure 2:** *libertà* **- 198 occurrences - 16significant collocates**

For each collocate, the first columnshows the MI value, the second its frequency value, ie the number of timesthe collocate was found in the context windows for *libertà*; in this case, weaccepted both nouns and verbs and 'n' was set at 5. An indication of themeaning of the collocates in English has been given between brackets foreasier understanding.

Next, using our lexical tools (e.g.morphological analysers and generators and bilingual lexical database), weconstruct an equivalent vocabulary ($V_2$) in L2of translation equivalents for the L1 set of cooccurrences ($V_1$), ie foreach element of $V_1$ we create a set of L2translation equivalents, denoted as L2 translation blocks.

Each block contains the entire setof translations supplied by the bilingual lexical database for a member ofthe L1 vocabulary (no distinction is made for sense), together with all possible forms for each translation (generated by themorphological procedure). For example the L2 translation block for theItalian lemma *garanzia* includes the English forms*guarantee, guarantees*, *security, securities*, *surety*, *sureties*. To each translation block, we assign a value equal to the MI Index of theL1 term represented by this translation block. These values are used toassign weights to the translation blocks to represent the probability ofoccurrence in the L2 texts of any of the members of that particular translation block when searching forexpressions regarding our keyword, T. Direct translations of the termitself are also assigned an arbitrarily high value as being the most probable L2 representative of T. An L2 stoplist is also applied at this point,again in order to eliminate as much noise as possible from the itemscontained in the translation blocks; basically, we eliminate very common L2words.

The procedure then searches the L2corpus in order to identify words or expressions that can be considered asin some way lexicallyequivalent to our selected term in the L1 texts. This is done by searchingfor those contexts in L2 in which there is a significant presence of the L2vocabulary for T. The significance is determined on the basis of astatistical procedure; this procedureuses the number of $V_2$ items found in the context and the weights assigned to them in order toassess the probability that any given L2 cooccurrence represents alexically equivalent context for T, and to establish thresholds ofacceptability.

**ComparableContexts**

```
     6      535.181 930          1)     is also a result of the fact that international
*rules**requiring* strict safety *standards* for*passenger* *vessels* apply only to those
operating internationally. Because   *"FXAC93207ENC.0003.01.00".30
     5      530.616 1000    2)     Council Directive of 30 November on the minimum *health*
and safety *requirements* for the use by *workers* of *personal* protective equipmentat
the work-place (1)    * "FXAC93207ENC.0042.01.00".22
```

```
     5      528.9811000   14)    27 October 1992) \\ (93/C 327/22)\Q\ *Subject*: Proposal
for a *Council* *regulation* on security *measures* applicable to classified
informationproduced or      * "FXAC93327ENC.0013.02.00".9
     5      528.1311000   31)    purpose. They have given their fullbacking to
*Resolution* 787 of the United *Nations* Security*Council*, which *stepped* up sanctions
against the Federal Republic of* "FXAC93350ENC.0040.03.00".25
     5      37.766 223           45) Communities (12 March 1993) \\(93/C 280/85) \Q\
*Subject*: *Council* *Directive* laying down *health**rules* for the production and
placing on the market of raw milk, heat* "FXAC93280ENC.0043.03.00".9
     5      37.432 217           46)  to an installation using ionizingradiation. To
*increase* the *health* *protection* of such *workers* the*Council* adopted, on 4 December
1990, Directive 90/641/       *"FXAC93065ENC.0017.01.00".75
     5       34.725216           47)    States and the IAEA in theCommission Standing
*Working* Group on the *transport* of *nuclear**materials* cover the *need* for mutual
information between the parties.Since   * "FXAC93327ENC.0024.01.00".26
     5      34.132 141           48)    for a Directive on thereorganization of working
*time* setting minimum *requirements* atCommunity *level* for *protecting* the *health*
and safety of workers. Theprovisions   * "FXAC93086ENC.0036.01.00".38
     4      524.1361000   52)    the G7 summit in Munich that thedesign of RBMK *nuclear*
*reactors* does not offer the same safety *guarantees* as do nuclear reactors of a more
modern design. The        *"FXAC93327ENC.0028.02.00".28
     4      30.083 214           205)  using ionizing radiation. Toincrease the
*health* *protection* of such *workers* the *Council* adopted,on 4 December 1990,
Directive 90/641/Euratom on      *"FXAC93065ENC.0017.01.00".75
----------- DBT --E.Picchi -------------------------------------------
```

The comparablecontexts are ordered by (i) no. of significant collocates, (ii) presence ofdirect translations of the term searched, (iii) MI value, (iv) sum of frequency values. Column 4 shows theirranking order.

**Figure 3: Comparablecontexts for *sicurezza* in Italian/Englishparliamentary texts**

Although it is clear that theprocess of translating the L1 vocabulary for T into L2 introduces a numberof irrelevant terms (all dictionary provided translations are accepted) andonly some of these are eliminated by the L2 stop list, this does not normally affect the resultsas, if an L2 context is to be accepted as representative of a given L1term, it is necessary for a number of items from the L2 vocabulary for T tobe present.

The results are written in a fileand ranked in descending order according to (i) the number of items in thecontext coming from the L2 vocabulary, iecontained in different translation blocks, (ii) whether a directtranslation of the term being searched is included, (iii) the sum of the MIvalues associated with these items. The file of results can be displayed onthe screen, saved, or printed out for further consultation The user can also enlarge a selected context byclicking on it so that he/she can refer to the entire piece of text towhich it belongs in the underlying corpus.

Figure 3 shows the results of aquery on our comparable corpus for the Italian noun *sicurezza*. For reasons of space, we have printed out only 10 contexts, just to give an idea of the kind of results we obtain using thismethod. The first context contained 6 items from the L2 vocabulary for theterm being searched including a direct translation of it; the next 43contexts (2-44) contained 5 items from the L2 vocabulary, including direct translations - this wasgenerally "safety" but in contexts, 14 and 31 we find "security"; Nos. 45 -48 showexamples of contexts in which there was no direct translation of the termitself, just the presence of 5 items from the L2 vocabulary. The reader canjudge for him/herself to what extent he/she feels that the contextrepresents the concepts of safety/security in this corpus. Context 52 was the first context containing just 4items from the L2 vocabulary including "safety" and No. 245 was the firstcontext with four L2 vocabulary terms which did not include a directtranslation of *sicurezza*.

Our test corpus up until now has been a set of parliamentary debates inEnglish and Italian of approximately 1 million word forms per language.This corpus has been useful for testing but is not entirely satisfactoryfor our purposes; although stylistically homogenous, the lexicon is rather too general and thus not suitable forstudies on terminology as was our original intention; first results oftesting on this corpus were given in Picchi, Peters (1996). We have nowmade some trial runs on a set of newspaper articles from the same year: *Corrieredella Sera* (approx. 300,000 word forms) and the *Independent* (approx. 600,000 word forms). As expected, wefind that in order to obtain significant results the texts must be eitherhighly homogeneous, or of large dimensions. In the case of the newspaperarticles, the corpus was not sufficiently large to give us many interestingterms with a reasonably high frequency in order to calculate a significant vocabulary. (In fact, thiscollection of texts had been collected for a different purpose: to studythe behaviour of certain neologisms over languages). In the next two figures, we give the results obtained for the Italianterm *lavoro* (translated by the dictionary as:work, job, task, labour), one of the most frequent terms found in thesenewspaper archives.

**DBT - Comparable texts**

**from**

**Corriere della Sera 1994**

163lavoro

0000000 16

500.000 163|LAVORO (work, labour, job, task)
11.772 4|SUBORDINATO (subordinate)
10.100 5|AUTONOMO (autonomous)
8.907 3|MASTELLA (Mastella - Italian Minister of Work)
8.262 8|DIPENDENTE (employee)
8.203 3|CONTRATTO (contract)
7.794 3|RELAZIONE (report)
7.734 6|CAPITALE (capital)
7.681 3|OCCUPAZIONE (occupation)
7.563 8|MINISTRO (ministry)
7.526 3|FINANZA (finance)
7.478 3|ESIGENZA (necessity)
7.302 3|ORGANIZZAZIONE (organization)
7.145 3|PRINCIPALE (principal)
7.109 3|IMPRESA (agency)
6.940 3|SINDACATO (union)
6.938 6|TRATTA (trade)
6.879 9|POSTO (position)

**Figure 4:** *lavoro* **- 163 occurrences - 18 significant collocates**

**DBT - Comparable texts**

**from The Independent 1994**

4 521.295 176 1) in the home. Inability to switch off' from ***work***, *experienced* by 83 per cent, may be a *major* stress on MPs' family*relationships*. Looking for an * IND Health:

4 521.035 176 2) team-working ability, motivation anddrive. We *need* graduates to ***work*** in thefollowing broad skill areas: *business* analysts, engineers, *finance** IND Enterprise 94:

4 520.757 185 3) told him, effectively been defeated forthe ***Labour*** leadership, as a result ofmachinations by a group of *trade* *union* barons who *controlled* the* IND Profile:

4 27.823 19 4) designers and marketers versed in the latest*business* *school* techniques. *Firms* of' the market seek out zany*employees* with out-of-the-ordinary views* INS

4 27.435 15 5) than purely reactive. We are beginning tosee *trade* *unions* as an important voice on the *business* pages, thepersonal *finance* pages, the health page * IND

4 27.387 15 6) Last month Phase closed after five *issues*. It faced a familiar small*business* problem - not enough *capital* to *finance* its early losses,* IND

4 27.387 15 7) after five issues. It faced a familiar small*business* *problem* - not enough *capital* to *finance* its early losses,which were much higher than * IND

3 515.407 178 8) hours a year more than the industrial average. *Employees* of the*major* oil companies form only a quarter of the workforce. The rest of thehard ***labour*** * IND

3 515.272 173 9) for current **work**, the current work itself, *accounts*, expenses and any correspondence.Clicking on items in the folders automatically loads the * INDComputers:

3 515.207 173 10) throughout england and wales in a varietyof *occupations* including environmental ***work***,youth, arts, marketing and *finance*. Contact: Come along to * INDEnterprise
3 515.207 173 11) high and must be reduced, and massreduction of *employment* in the hitherto most stable sectors of tertiary*occupations* public *employment*, banking and * INS Age of extremes:
3 515.207 173 12) in the hitherto most stable sectors oftertiary *occupations* public employment, banking and *finance*,office-***work*** became common.* INS Age of extremes:

**Figure 5: Comparablecontexts for *"lavoro"* in Italian/Britishnewspapers**

Figure 4 gives the list ofsignificant collocates for "*lavoro*"in the Italian newspapers, and Figure 5 shows theresults obtained when we searched for contexts in the English newspaper archivescontaining a significant number of the L2 vocabulary items which were derivedon the basis of this list. In Figure 5, directtranslations of the term searched are shown in bold, members of the L2vocabulary are indicated between asterisks.

### Evaluating Our Results

The comparable corpus querysystem is still under development, and it is clearly more difficult toevaluate its performance than it is for the parallel system. While it is easy to check whether contexts have been constructed for eachdirect translation derived from our bilingual dictionary for the L1 termsearched, an objective evaluation of contexts that contain no directtranslation but just a relevant number of items from the L2 vocabulary is not so easy. It would be necessary togo manually through the entire L2 set of texts looking for other contextsthat reflect the same concept but which were not retrieved by the system to assess with some degree of accuracy where it hasfailed. This would be a difficult and time consuming task, and one which wehave not attempted so far. However, one test that we do make to evaluatesystem performance is to construct the L2 vocabulary excluding direct translations of the L1 term ofinterest (T). We then retrieve our comparable contexts and look to see ifany of these do contain direct translations of T, despite the fact thatthese were not searched specifically.An example is shown in Figure 6.

This figure shows the results of aquery on our comparable corpus for the Italian lemma *libertà*, using the L1 vocabulary shown in Figure 2. We show here the first 12contexts, ie those calculated by the system as being most representative ofthe use of this term in this particular corpus. We excluded thetranslations of the term given in the bilingual dictionary (*liberty, freedom*) from theconstruction of these contexts. The fact that a direct translation of*libertà* appears in a number of the results(nos. 2, 3, 4, 5, 11 and 12) is encouraging.

### Discussion and Prospects

This approach to the problem ofidentifying cross-language lexical equivalences over homogeneous sets oftexts for different languages has several merits: it allows usto disambiguate, to a considerable extent, both the L1 term being analysedand the target language terms provided by the dictionary; it permits us toretrieve lexically equivalent cross-language expressions even when the L2context does not contain a dictionary derived translation of the L1 term; it provides a ranking ofour results.

**DBT - (Comparable Texts)Search*libertà***
4 24.842 20 1) to make changes in funding *policy* in relation to *cities*, whilestill *respecting* the *principle* of =FE"FXAC93207ENC.0014.01.00".14
3 26.512 36 2) Respect for human rights and*fundamental* freedoms is *guaranteed* in the Member States by *effective*systems of =FE "FXAC93145ENC.0030.01.00".32
3 24.462 38 3) as long as that legislation*guarantees* that the relevant *fundamental* *principles* and freedomsenshrined in the
=FE"FXAC93288ENC.0024.01.00".27
3 23.830 38 4) as guardian of the Treaties. *Respect* for human rights and*fundamental* freedoms is *guaranteed* in the Member =FE"FXAC93145ENC.0030.01.00".32
3 22.806 40 5) in the Treaty, such as the *principle*of non-discrimination, *respect* for the *fundamental* freedomsenshrined =FE "FXAC93040ENC.0013.01.00".50
3 22.806 40 6) that all Community law ought to*respect* this *principle* because it is a *fundamental* right. Despite the=FE "FXAC93162ENC.0013.01.00".40

3 22.806 40 7) countries must be based on such*principles* as *respect* for international law, human rights and*fundamental* =FE "FXAC93327ENC.0036.02.00".26
3 22.741 41 8) in the drawing up of all *policies* isa *fundamental* *principle* of Community legislation. In view of this: 1.=FE "FXAC93065ENC.0034.01.00".16
3 22.741 41 9) in Portugal. This discriminatory*policy* is in clear breach of the most *basic* *principles* and provisionsof national =FE "FXAC93095ENC.0014.03.00".19
3 22.512 40 10) nationality which conflict with the*fundamental* *principles* of the EEC *Treaty*. It should be made clearthat =FE "FXAC93047ENC.0005.01.00".27
3 22.512 40 11) under Article 115 of the EEC *Treaty*derogate from the EEC Treaty's *basic* *principle* of freedom of movementof =FE "FXAC93264ENC.0031.01.00".32
3 22.512 40 12) guarantees that the relevant *fundamental* *principles* andfreedoms enshrined in the EEC *Treaty* will be observed. = FE"FXAC93288ENC.0024.01.00".27

**Figure 6: Comparablecontexts for *libertà* inparliamentary texts**

**L1 orMonolingual Sense Disambiguation** Although the problem of polysemy is greatly reduced in a domain specificcorpus, it is still present - to a varying degree depending on the type oftexts being treated. The construction of the L1 vocabulary whichcharacterises our term Twill permit us to obtain a clustering of the most relevant terms connectedto T. If the corpus contains a predominant sense for the term then the vocabulary should represent this sense - secondary senses that appearrarely will not cause a representative vocabulary of collocates to beconstructed. If, in the corpus, there is more than one relevant sense for Tthen we would expect two or more distinct clusterings of significant collocates. Take, to use a classicalexample, the unlikely event that our collection of texts has a significantnumber of occurrences of both the river and the financial sense of "bank".We would expect to be able to obtaintwo distinct clusterings of significant collocates with - in this extremecase - little or no overlap. This type of sense disambiguation for the L1term under exam did not appear very relevant when we started this work asour initial interest was in domain-specific sets of texts in different languages. However, we have now begunto extend the area of our interest to more general (although stillcomparable) collections, such as newspaper archives in more than onelanguage for the same period (see above) or searching over Web sites containing documents in different languages (seefinal section). Being able to perform some kind ofsense disambiguation on the L1 term is thus becoming far more important andwe intend to pursue this line of investigation in our next work, both formonolingual and bilingual text sense disambiguation.

**L2 or Target TermDisambiguation.** The second kind of disambiguation operates atthe target language level. As stated above, our procedure takes as inputall the translation equivalentslisted in the bilingual dictionary regardless of sense distinctions.Inappropriate translations are eliminated by the fact that we normally donot find them together with a significant number of items from the L2vocabulary for the term being searched. For example, if we examine all the occurrences of *sicurezza* in our parliamentary corpus we find that the sense is that of "safety", or"security" (one sense of "security" is a synonym of "safety"). This isconfirmed by the set of significant collocates for this term; the top ten are the Italian equivalents of toy, hygiene, reactor,health, nuclear, maritime, council, road, provisions, Euratom. Thebilingual dictionary gives us four separate senses for *sicurezza*: translated by safety, security, certainty, confidence. On the English sideof the corpus, we find 17 occurrences of "confidence" and just one of"certainty". However, the context for "certainty" does not appear in thelist of comparable contexts for *sicurezza* as it contains no other L2 vocabulary items; the contexts for"confidence" are ranked very low as they never contain more than two L2significant collocates for *sicurezza*. Thus, our approach helps us to identify the correct sense of the targetterms offered by the bilingual dictionary and to provide a ranking of thebest L2 matches for the L1 term searched. That this is not alwayssuccessful, however, is shown by context No. 3 in Figure 4, in which the political sense of Labourappears although it is the work sense represented by the Italian "lavoro" that we are looking for. The reason is the very high (but unsurprising)MI value assigned to the collocates "trade" and union"; perhaps this resultis not too discouraging, after all Labour was originally the party of theworkers!

The success of this approach depends on thedegree to which the L1 and L2 sets of significant collocates are trulyrepresentative of the term queried. We are thus now studying ways tooptimise the construction of this vocabulary. As has been stated, so far we have used the MutualInformation formula to compute our significant set of collocates for termssearched. This formula has been criticised as it tends to assign oversignificant values to infrequent words. We are currently implementing a differentmeasure based on likelihood ratios (see

Dunning 1993). But it is too soonyet to judge whether this will give us an improvement in performance.

Our next tests should be made on (i) a moretechnical comparable corpus which should provide us with a real test-bedfor multilingual terminology extraction; (ii) on a set of Italian and US newspaper items forthe same period but of far larger dimensions to give us a chance to studythe results of L1 term sense disambiguation.

## Applications toCLIR

In CLIR the aim is to find methods which successfully match queriesformulated in one language against documents stored in other languages.Various approaches have been/are being experimented. The best-known andtested involve multilingual thesauri. Otherapproaches attempt touse different kinds of translation techniques in order to extend thepotential range of the search: full-text translation of documents is notcurrently viewed as a realistic answer in consideration of the actual costsand limitations of MT systems; experiments on the automatic translation of queries employingbilingual/multilingual dictionaries have not given satisfactory results -queries are generally too short to permit an exhaustive non-ambiguoustranslation; likewise, concept-based approaches which attempt to achieve a matching between documents and queries at a more abstractlevel have not yet provided promising results on a large scale. At themoment, it looks like the most promising results will come from anintegration between multilingual lexicon and corpus-based methods.

When we began work on our two bilingual corpus processing systems, our maininterests were linguistic/lexicographic: applications such as bilinguallexicography, translation and language learning activities. However, we nowrecognise that both systems, employed together with our bilingual electroniclexicons, can also be applied to CLIR activities.

***Parallel Text System***
The very "tight" alignment achieved using dictionary derived translationlinks would greatly facilitate the statistical alignment of unmatchedterms. So far, this system has been tested in human-oriented applications,we now want to experiment it over a large parallel collection in order to automatically extract "new" translationequivalent data and thus augment the existing bilingual lexicon.

***Comparable Text System***
We intend to test our comparable-corpus strategy on twoapplications:

- multilingual digital library
- multilingual web search engine

In both cases, we will integrate a dictionary/thesaurus-based search withcorpus-based strategies. Disadvantages of lexicon-based systems are thatthesauri confine users to a controlled-vocabulary while general-purposedictionaries tend to be lacking in necessary technical vocabulary; the problem with most corpus-based CLIRsystems is that the acquisition of a suitable set of relevant documents on which to train the retrieval system is veryresource consuming. We hope to overcome these two problems: thecomparable-corpus strategy can be used to extend the limits of a simplequery termtranslation, and also to reduce the risk of ambiguity in the query term,and to provide a ranking of the results; at the same time, our corpus willconsist of the documents in the collection being queried and does not haveto be created ad hoc.

Queries will be translated by themultilingual lexicon but will also be expanded by applying the comparable-corpus based strategy in order to associate with each queryterm, not only its direct translations but also a vocabulary which definesits probable immediate context, in L1 and L2. In this way, we search forboth pre-identified translationequivalents and also cross-language lexical equivalences. When thedictionary or lexicon offers no translation equivalent, the search forcross-language equivalent contexts is still possible. Documents retrievedare ranked with respect to (i) translation equivalents of query terms, (ii) statistical value assigned to associatedsignificant collocates.

## References


Brown, P.F., Della Pietra,S.A., Della Pietra, V.J., Mercer, R.L.(1993). The Mathematics ofStatistical Machine Translation: Parameter Estimation. *Computational Linguistics*. 19(2):263-312.

Church, K.W., Hanks, P.(1990). Word Association Norms, Mutual Information and Lexicography.*Computational Linguistics.* 16(1): 22-29.

Dunning T. (1993). Accurate Methods forthe Statistics of Surprise and Coincidence,*ComputationalLinguistics*, 19(1).

Gale, W.A., Church, K.W. (1993). Aprogram for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1): 75-102.

Hartmann, R.R.K. (1997). FromContrastive Textology to Parallel Text Corpora: Theory and Applications. InRaymond Hickey and Stanislav Puppel (eds.),*Language Historyand Language Modelling*. Festschrift in Honor of Jacek Fisiak's 60th Birthday, Berlin: W. de Gruyter, in press.

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 TextResearch. Oxford: OUP, 63-70.

Marinai, E., Peters, C., Picchi, E.(1992). A Project for Bilingual Reference Corpora. *ActaLinguistica Hungarica*, 41 (1—2). Akadémiai Kiadó,Budapest, 1-15.

McEnery, T., Oakes, M. (1996). Sentenceand word alignment in the CRATER project. In J.Thomas and M. Short(eds.),*Using Corpora for Language Research*, Longman,London and New York, 211-231

Picchi, E., Peters C. (1996). CrossLanguage Information Retrieval: A System for Comparable Corpus Querying,.In G. Grefenstette (ed.) Working Notes of theWorkshop on Cross-LinguisticInformation Retrieval, ACMSIGIR'96, 24-33.

Smeaton, A.F. (1992). Progress in the Application of Natural LanguageProcessing to Information Retrieval Tasks, *The ComputerJournal*, 35 (3), 268-278.

Warwick-Armstrong, S., Russell, G.(1990). Bilingual Concordancing and Bilingual Lexicography. In*EURALEX 4th International Conference*, Malaga,Spain.