Carol Peters, Istituto di Elaborazione della Informazione, CNR, Pisa
Eugenio Picchi, Istituto di Linguistica Computazionale, CNR, Pisa

From Parallel to Comparable Text Corpora

Abstract

We present a bilingual corpus management system under development in Pisa. The first component of this system was a set of procedures to create and query parallel text archives; we are now studying the implementation of a second set of procedures to interrogate comparable archives. The approach followed is quite different from that used for parallel data and considerably more complex; the results are also very different. In the paper, we describe the strategy we are adopting to retrieve significant data from comparable corpora, and discuss the preliminary results.

1. Background

During the eighties, there was an enormous investment of efforts and resources in the construction of monolingual language reference corpora. At the beginning of the nineties, researchers in machine translation and bilingual lexicography began to turn their attention to bilingual text archives. The motives were similar: bilingual corpora make it possible to investigate syntactic, semantic and lexical relationships between languages rather than for a single language, and are important sources of contrastive evidence of language usage.

So far most of the studies on bilingual (and multilingual) text archives have been on parallel or translationally equivalent texts. However, a major criticism of the results of analyses based on such texts is that target texts are not true examples of natural language. As stated by Hartmann (1994) "the translated text(s) cannot by definition share the full range of linguistic features of genuine texts produced in the respective target language". For this reason, many researchers have suggested that a more reliable source for certain types of studies could be a bilingual or multilingual comparable text corpus.[1]

Comparable text archives have been described as collections of "texts which, though composed independently in the respective language communities, have the same communicative function" (Laffling 1992: 20). Unlike parallel corpora they concern a restricted sublanguage. They provide a source of data on natural language lexical equivalents within a given domain. Zanettin (1994) asserts that "being topic-related, (com-

parable) subcorpora tend to express and report similar facts: being instances of attested behaviour, they represent models against which to check hypotheses on language use".

These two kinds of bilingual corpora thus have different if complementary functions, and should be used for different scopes. For example, the parallel corpus can provide much useful data when studying the translation process or teaching translation skills, while the comparable corpus should provide reliable evidence on how lexical equivalences are rendered naturally in different languages. Both should be useful tools for the bilingual lexicographer. The problem is that the querying and extraction of significant data from a comparable corpus is far more complex than from a parallel one; an effective strategy must be found to identify semantically equivalent expressions in different languages without relying on translationally equivalent textual material. This is the issue addressed in this paper.

2. The Bilingual Corpus System

A system for bilingual corpus construction and management has been under development at the *Istituto di Linguistica Computazionale*, Pisa, for several years. The first component of this system was a set of procedures to create and query parallel text archives. The majority of systems studied to manage such corpora use statistical-distributional data. The first experiments concentrated on algorithms for text alignment mainly on a sentential basis. Other papers have suggested different ways for isolating and identifying translation equivalents within previously matched sections and without pre-alignment (see a.o. Gale and Church 1991). Our approach has been very different: we employ external evidence provided by a bilingual lexical database (LDB) and morphological procedures to create links between pairs of texts on the basis of SL/TL translation equivalents. These links are then used to construct parallel contexts for any form or cooccurrence of forms (see Marinai et al 1992).

It is now our intention to extend the scope of our bilingual corpus system by including a second component – a set of procedures for the analysis and extraction of significant data from comparable text archives. The aim is quite different: with our parallel system users can retrieve examples of specific instances of how a given word or expression has been translated in another language, depending on contextual factors; using the comparable system, they will be able to retrieve sets of natural

language examples of L2 renderings of the ideas represented by a given term in a given domain in L1, independently of any translation link.

3. Procedures for Comparable Corpus Processing

One of the basic tenets of corpus linguistics is that words acquire sense from their context, and a word used in a given sense frequently reveals not only typical syntax but typical patterns of lexical cooccurrence. We decided to exploit this fact when developing our comparable corpus procedures. Our starting point is thus the idea that a term is characterised by the words with which it cooccurs. If we can establish equivalences between several items contained in different contexts, then there is a high probability that the contexts themselves are to some extent similar. Our aim is not to retrieve precise equivalences in L2 of the L1 term under examination, but to isolate the set of contexts in the L2 corpora that has the highest probability of providing L2 correspondences to the L1 input.

The procedures that we are working on operate on sets of comparable texts in two different languages: texts from the same domain or on the same topic. We are currently working on Italian and English texts, and so far all work has been focused on nouns.[2] Given a particular term found in texts in one language (L1), the aim is to be able to identify contexts which treat the same argument in texts of the second language (L2). To do this, we isolate the vocabulary related to that term in the L1 corpus – hypothesizing that the word will be surrounded by a similar vocabulary in L2.

A term, T, is thus selected in the L1 set of texts (either set can be chosen as L1). T can be either a single lexical item or combination of lexical items. For each occurrence of T in the L1 set of texts, the system constructs a context window, containing T plus up to n lexically significant words appearing to the right and left of T, but within the same phrase, i.e. strong punctuation marks (full stops and semi-colons) act as break points in the construction of these contexts. Words contained in a stop list are not counted. The list of stop words includes functional words such as articles, pronouns, prepositions, and highly frequent insignificant words which create noise. The user can modify this list, e.g. eliminating certain frequent domain-specific terms, if necessary to improve performance.

For each cooccurrence of our keyword T in the context windows, morphological procedures identify the source lemma (or lemmas in the case of homography). The set of significant words that are found in the context windows for T make up the vocabulary, V1, that characterises T

in the particular L1 corpus being analysed. The frequencies of the cooccurrences of T are then computed and to each element of V1 is assigned its mutual information value which measures the significance of the correlation between the V1 item and T, i.e. the relative frequency of the V1 item as a collocate of T is measured against its overall frequency in the corpus in order to identify how strongly it is related to T (Church and Hanks 1989). Using the MI index as an ordering element, we list V1 in order of decreasing significance and set a threshold below which terms in V1 are not considered relevant and can be ignored. Figure 1 shows the significant collocates for the Italian lemma *bilancio* (a financial/ commercial term) found in a set of comparable English and Italian parliamentary debates. There were 552 occurrences of *bilancio* in the Italian side of the corpus. For each collocate, the first column shows the MI value, the second the frequency value, i.e. the number of times the collocate was found in the context windows for *bilancio.*

Next, using lexical tools that we have already developed, i.e. morphological analysers and generators and a bilingual lexical database (based on the Collins Giunti English–Italian general language dictionary), we construct an equivalent vocabulary (V2) in L2 of translation equivalents for the L1 set of cooccurrences (V1), i.e. for each element of V1, we create a set of translation equivalents in L2, denoted as L2 translation blocks. This is shown in Figure 2. Each L2 translation block contains the set of translations supplied by the bilingual lexical database for any member of the L1 vocabulary, together with all possible forms for each translation (generated by the morphological procedure). For example an L2 translation block for *finanziare* includes the English forms 'finance, finances, financed, financing, fund, funds, funded, funding'. To each translation block, we assign a value equal to the MI Index of the L1 term represented by this translation block. These values are used to assign weights to each block to represent the probability of occurrence in the L2 texts of any of the members of that particular translation block when searching for expressions regarding our keyword, T. A translation block referring to the L1 word being processed (T) is also created and given an arbitrarily high value. For T=*bilancio*, this contains 'balance, balances, balanced, balancing, budget, budgets'.

The procedure then searches the L2 corpus in order to identify words or expressions that can be considered as in some way lexically equivalent to our selected term in the L1 texts. This is done by searching for those contexts in L2 in which there is a significant presence of the L2 vocabulary for T. The significance is determined on the basis of a statistical procedure; this procedure uses the number of V2 items found in the context and the weights assigned to them in order to assess the

probability values for different sets of L2 cooccurrences to represent lexically equivalent contexts for T, and to establish thresholds of acceptability. Although it is clear that the process of translating the L1 vocabulary for T into L2 introduces a considerable number of irrelevant terms, this does not constitute a problem. For example, the translations found in the bilingual LDB for *fondo/fondare/fondere* included the highly relevant collocates 'fund(s), cash' but also 'land, property, country estate; bottom (of sea); seat (of trousers); road surface; dregs (of wine), grounds (of coffee)'. However, this type of noise does not normally affect the results as, for example, it is extremely unlikely that texts treating financial questions are also going to talk about 'coffee grounds', or 'trouser seats' and, in any case, for an L2 expression to be listed as a representative context for L1 it is necessary for a number of items from the L2 vocabulary for T to be present.

The contexts retrieved are written in a file and listed in descending order of (i) the number of items contained in different translation blocks appearing in the context, (ii) the sum of the MI values associated with these items. The file of results can be displayed on the screen and browsed, or printed out for further consultation.

4. First Results

Much work remains to be done in refining the search criteria and increasing the efficiency of the global algorithm in order to improve performance and, in particular, to increase precision of retrieval, eliminating as much noise as possible. However, we feel that our first results on real texts are encouraging and already demonstrate the validity of our approach. We are able to identify and retrieve contexts in an L2 set of texts which refer to a particular argument represented in L1 by a given expression (term or set of terms), without the necessity for a known translation equivalent for that expression being present.

In Figure 3, we give some examples of comparable contexts in English for the Italian lemma *bilancio*. This term had been associated in our L1 test corpus (parliamentary debates in Italian and English) with the set of Italian collocates shown in Figure 1. For reasons of space, we only print out the first 15 contexts, i.e. those calculated by the system as being most representative of the use of these terms in this particular corpus, and have eliminated contexts which are almost identical. At the beginning of each context, the number of items from the L2 vocabulary for T (highlighted in the text), and the sums of the MI and the frequency values

are given for these items. It is interesting to note that only two of these contexts contain a direct translation for T (Nos. 13 and 14).

Notes

1. We here use the terms 'parallel' and 'comparable' commonly used by computational linguists to distinguish between the two kinds of bilingual or multilingual texts. However, other researchers have proposed a distinction between bi-texts, for translationally linked texts, and parallel texts for texts that are functionally similar in situational motivation and rhetorical structure (see Hartmann, 1995).
2. In sub-language texts, the nouns bear most of the weight of topic-specificity, and the occurrence of polysemous nouns is greatly reduced when compared with general language texts.

References

Church, K.W., P. Hanks 1989. "Word Association Norms, Mutual Information and Lexicography", in: *Proceedings 27th Annual Meeting of ACL*, Vancouver, B.C., pp. 76–83.

Gale, W.A., K.W. Church 1991. "Identifying Word Correspondences in Parallel Text", in: *Fourth DARPA Workshop on Speech and Natural Languages*. Morgan Kaufmann Publishers, pp. 152–157.

Hartmann, R.R.K. 1994. "The Use of Parallel Text Corpora in the Generation of Translation Equivalents for Bilingual Lexicography" in: *Euralex-1994*, pp. 291–297.

Hartmann, R.R.K. 1995. "From Contrastive Textology to Parallel Text Corpora: Theory and Applications", paper for Jacek Fisiak Festschrift.

Laffling, J. 1992. "On Constructing a Transfer Dictionary for Man and Machine", in: *Target* 4(1), pp. 17–31.

Marinai, E., C. Peters, E. Picchi, 1992. "A Project for Bilingual Reference Corpora", in: *Acta Linguistica Hungarica*, 41 (1–2). Akadémiai Kiadó, Budapest, pp. 1–15.

Zanettin, F. 1994. "Parallel Words: Designing a Bilingual Database for Translation Activities", in: A. Wilson and T. McEnery (eds.), *Corpora in Language Education and Research: TALC 94*. Lancaster Univ. UK, pp. 99–111.

```
    552bilancio
0000000     20
500.000   552|BILANCIO|BILANCIARE (balance, budget;
                                  to balance, to budget)
  9.533     3|ATTINGERE (to draw on)
  8.164     8|ISCRITTO|ISCRIVERE (registered; to register)
  7.353     6|ASSEGNARE (to assign)
  7.318     6|STANZIARE (to allocate)
  6.654     9|FINANZIARE (to finance)
  6.574     6|SINGOLO (single)
  6.486     6|RICONVERSIONE (reconversion)
  6.172     6|STANZIAMENTO (allocation)
  6.108     5|CONSENTIRE (to allow)
  5.928     8|TOTALE (total)
  5.751    10|PREVEDERE|PREVISTO (to forecast; forecast)
  5.606     8|DESTINARE (to intend to use for)
  5.504     5|PRIMO (first)
  5.188     9|PROGETTO|PROGETTARE (plan; to plan)
  5.005     5|IMPORTO|IMPORTARE (amount; to import)
  4.911     4|SPECIFICARE (to specify)
  4.819     7|CONCEDERE (to concede)
  4.651     5|FONDO|FONDARE|FONDERE (fund; to fund; to melt)
  4.644     4|CONTRIBUTO (contribution)
  4.635     4|DECIDERE (to decide)
```

Figure 1 - Significant collocates for *bilancio*

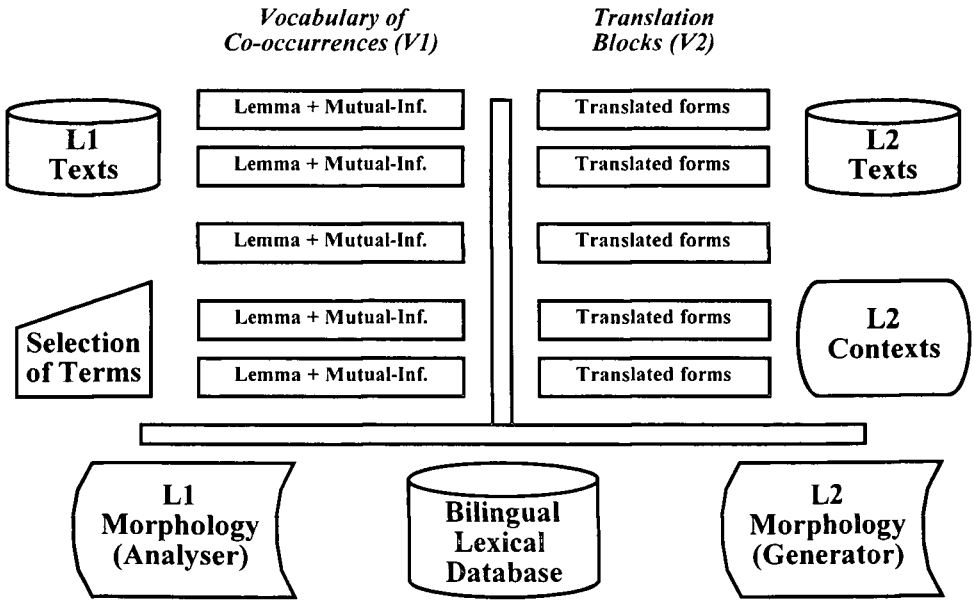

Figure 2 - Comparable Corpus Query System

5 29.728 37 1) ...to ECU 19 786 000 (or 8,6 of the **total allocation**). The **main** thrust of **projects financed** through SACC is in the... "051.0022".35
4 22.959 26 2) ...transfer were listed as the **main cross**-sectoral priorities for the **allocation** of **funds** and the selection of ... "258.0020".24
4 22.603 20 3) ...Structural Funds. Accordingly, ESF **funds amounting** to ECU 33 million were **allocated** to **programmes** there during the... "297.0040".26
4 22.417 18 4) ...information, long delays occur in **allocating** and using **funds granted** under regional development **programmes** to... "297.0006".13
4 22.083 29 5) ...of funds The Commission has **decided** to co-**finance** an **initial** series of 42 **projects** in the context of LIFE... "051.0030".13
4 22.007 27 6) ...proposal was needed for the whole **programme**. **Financing** decisions for **individual projects** were taken by the chief... "090.0021".23
4 22.007 27 7) ...on the basis of an operational **programme** rather than **individual projects**. The Social **Fund** therefore no longer... "095.0019".35
4 22.007 27 8) ...can confirm that the choice of **individual projects financed** by the ERDF under an operational **programme** and the... "320.0015".33
4 20.501 28 9) ...submitted their regionalization **plans** within the **specified** period? Are these **plans based** on appropriate and..."185.0043".14
4 20.252 24 10) ...of action as part of operational **programmes** or major **projects**. The **granting** of **funds** is subject to compliance with... "016.0024".48
4 20.084 26 11) ...committees have prepared **programme funding** proposals **based** on the many **projects** submitted, selected by... "106.0011".35
4 19.923 21 12) ...Could it provide a complete list, **specifying** the **amounts granted** for each **project**? 2. Are the appropriations... "333.0023".15
3 510.756 567 13) ...calculated on the basis of a **forecast** supply **balance** for each marketing year. If the **import** quotas fixed for... "127.0018".44
3 510.402 567 14) ...and the additional quota) will be **based** on the **forecast** supply **balance** for the year, calculated on the basis of... "127.0018".35
3 20.581 18 15) ...is therefore not informed of the **financing allocated** to each **individual** operation nor of its beneficiary. The... "185.0038".32

Figure 3 - Comparable Contexts for *bilancio*