Meta
Journal des traducteurs
Translators' Journal

META

Bilingual Comparable Corpora and the Training of Translators

Federico Zanettin

Volume 43, numéro 4, décembre 1998

L'approche basée sur le corpus
The Corpus-based Approach

URI : https://id.erudit.org/iderudit/004638ar
DOI : https://doi.org/10.7202/004638ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (imprimé)
1492-1421 (numérique)

Découvrir la revue

Citer cet article

Zanettin, F. (1998). Bilingual Comparable Corpora and the Training of Translators. *Meta*, *43*(4), 616–630. https://doi.org/10.7202/004638ar

Résumé de l'article

Cet article montre comment un petit corpus bilingue, qu'il soit en langue de spécialité ou en langue générale, peut être utilisé par des enseignants pour développer des activités pédagogiques qui améliorent la compréhension de la langue source des étudiants et qui développent leur habilité à produire un texte en langue cible.

BILINGUAL COMPARABLE CORPORA AND THE TRAINING OF TRANSLATORS[1]

FEDERICO ZANETTIN
Università degli Studi di Bologna, Forlì, Italy

Résumé

Cet article montre comment un petit corpus bilingue, qu'il soit en langue de spécialité ou en langue générale, peut être utilisé par des enseignants pour développer des activités pédagogiques qui améliorent la compréhension de la langue source des étudiants et qui développent leur habilité à produire un texte en langue cible.

Abstract

This article demonstrates how small bilingual corpora of either general or specialised language can be used to devise a variety of structured and self-centred classroom activities whose aim is to enhance the understanding of the source language text and the ability to produce fluent target language texts.

INTRODUCTION

One of the major developments in linguistic research has come from the possibility of studying vast amounts of text through computer enhanced tools, particularly through text retrieval and concordancing programs. The basic investigation procedure for querying text corpora consists in producing multiple concordance lines, the so-called Key Word In Context — or KWIC — concordance, for a specified string of characters — a word, a lemma or a phrase. The citations thus obtained can be sorted to reveal recurring clusters of words. The analysis of these recurring patterns highlights the behaviour of actual language in context, and complements and sometimes challenges the information provided by standard reference tools such as dictionaries and grammars.

In this article I discuss the application of such tools to the training of translators. In particular, I provide examples of features which can be investigated using small bilingual comparable corpora (English - Italian) showing how these can be used for activities aimed at developing comprehension and production skills.

1. CORPORA AND TRANSLATION

A wide array of different types of corpora have been constructed for use in the field of translation. They reflect the criteria according to which they are designed and the purpose for which they are created.

A first type is the multi-source-language monolingual "comparable" corpus, consisting of two sets of texts, one originally written in language A and one of similar texts translated into language A from a variety of different languages (Baker 1995, 1996; Laviosa 1997 and this volume). Its value is mainly theoretical, what is investigated is the linguistic nature of translated text, independently of the source language.

A second kind of corpus used in translation is the bilingual (or multilingual) corpus. Language pairs are put together either on the basis of "parallelism" or/and "comparability." Parallel bilingual corpora consist of texts in language A and their translation into language B, and/or vice versa. The relationship between texts is directional, i.e. it goes from one text (the SL text) to the other (the TL text). Comparable bilingual corpora consist of texts in the languages involved, which share similar criteria of composition, genre, and topic.

Bilingual corpora have been mainly used for terminology extraction (e.g. Dryberg & Tournay 1990; Laffling 1992) and machine-aided translation (see Somers 1993). Much recent research in MT aims not so much to create a system able to perform the job of translating a given text automatically, but to implement computerised tools to assist human translators in their work. Parallel corpora can also be treated as "translation memories," from which translators can retrieve chunks of translated language in order to speed up their work and ensure accurate and consistent translations.[2] Parallel and comparable bilingual corpora have also been used for language learning and the training of translators (Johns,[3] Barlow 1996; Gavioli forthcoming; Zanettin 1994, forthcoming).

As regards parallel corpora, it has been observed that translated texts cannot represent the full range of linguistic possibilities of the target language and that they may reflect the stylistic idiosyncrasies of the source language and of individual translators (Teubert 1996: 247; Picchi and Peters 1997). However, the comparison between large numbers of texts and their acknowledged translations can show how equivalence has been established by translators under certain circumstances and provide examples of translation strategies. If such corpora are sufficiently varied and large, looking at recurring linguistic choices made by translators allows general patterns to be perceived. Learners can thus notice "preferred ways of putting things" (Kennedy 1992), and generalize from the aggregation of sets of individual instances.

The other type of corpus which may be relevant to translator training is the bilingual comparable corpus. This can be defined as a collection of texts composed independently in the respective languages and put together on the basis of similarity of content, domain and communicative function. The two collections are "typically unrelated except by the analyst's recognition that the original circumstances that led to the creation of the two [sets of] texts have produced accidental similarities" (Hartmann 1980: 38).[4]

The practice of collecting texts in different languages on the basis of similarity of type, content or function was common in translation research and training before the word corpus came to mean almost exclusively a collection of **electronic** texts. Snell-Hornby (1988) collected a corpus of printed English and German public notices, and Shäffner (1996) discussed how bilingual English German corpora composed of, for example, tourist brochures or instruction manuals can be used for the training of translators.[5] They argue that, by looking at multiple instances of texts belonging to the same text-type, prototypical features can be identified and can provide a "profile" of the type of text which translators are going to produce in the target language.

Technological advances have now made it easy to construct small bilingual comparable corpora. Teachers, students and researchers, as well as professional translators, can make use of corpora and text analysis software as a complementary resource to printed books and other materials. Comparable corpora can be created from a variety of sources: collections of texts distributed in electronic format (e.g. newspaper archives on CD-ROM, the Internet, etc.), or even from scanned or typewritten material. Criteria for creating comparable corpora depend on the homogeneity of texts, both within

and across languages, in terms of features such as subject domain, author-reader relationship, text origin and constitution (i.e. "single" or "joint"/"composite" texts), factuality, technicality and intended outcome (i.e. communicative function) (see Biber 1993).

The size of comparable corpora can, of course, vary depending on how well they meet these criteria. In the case of texts with a high degree of technicality, written by subject specialists for their peers, it is highly likely that relevant information can be derived even from a few texts. Collections of newspaper articles downloaded from CD-ROMs can be used to make larger comparable corpora (see Zanettin 1994 for details on the design of these corpora).

Figure 1 lists types of small bilingual comparable corpora which were "handcrafted" with limited resources at the School for Translators and Interpreters at the University of Bologna.

Text type	Source	Size
Newspaper articles (by section, by topic, by date), "Olympics '92" corpus, "Business" corpus, "Foreign news" corpus, "Racism" corpus	Newspaper archives on CD-ROM (e.g. *The Independent, The Daily Telegraph, Il Sole 24 Ore, La Stampa*), on-line newspaper archives (e.g. *The Electronic Telegraph, La Repubblica On-Line*), typewritten texts	100,000 - 1,000,000 words
Medical literature (journals, textbooks), "Hepatitis" corpus, "Hydrotherapy" corpus	Scanned and typewritten texts	50,000 - 100,000 words
Tourist brochures and guides	The Internet, scanned texts, typewritten texts	under construction

Figure 1
Some small bilingual comparable corpora

These corpora have been analysed using freely available programs (e.g. Tact, Corpus Wizard) and commercially available software (e.g. MicroConcord, Wordsmith Tools, Longman MiniConcordancer, DBT).[6]

2. CORPORA IN THE TRANSLATION CLASSROOM

Translation is a text production activity in which many factors influence the source and target communicative situation. The analysis of texts produced in comparable communicative situations can help learners investigate the respective expectations, experience and knowledge of the linguistic communities involved. In the process of establishing equivalencies between comparable sets of texts, learners acquire information about the way in which discourse is laid down in the two languages. They can use the attested evidence which corpora provide and create new texts which are partly made of citations from the target language adapted to the new occasion. By looking for coincidence across languages and manipulating units which approximate the concepts and functions they want to convey, they engage in a meaning-creation activity and

develop procedural skills. Instead of dealing with isolated words and rules for combining them into a text, learners can refer to and make use of wider chunks of language, multi-word units and recurring word patterns which are attested in a comparable corpus of texts in the target language. There are a number of ways in which comparable bilingual corpora or data derived from them can be used in the translation classroom.

2.1. Using the Corpus to Translate

In Zanettin (forthcoming) I describe an experiment carried out at the School for Translators and Interpreters at the University of Bologna, in which a group of trainee translators used the Olympics corpus as a primary tool in translating one short Italian sports article into English. By looking for homographs (i.e. proper names), cognates and perceived equivalents, learners were able to evaluate the respective behaviours in the two languages of similar discourse units, and to draw from a selection of citations in the target language suitable candidates to be included in the translation, in adherence to the linguistic and genre conventions of the receiving culture.

For example, sports articles are replete with figurative language. It may be difficult to establish whether an expression such as *"salire il gradino più alto del podio*" (to climb onto the highest step of the podium") would sound "native-like" if translated literally (Figure 2):

L'ammiraglio Straulino, con a prua Niccolo' Rode, negli anni 50 vince tre campionati mondiali, dieci europei, **sale il gradino piu' alto** del podio alle Olimpiadi finlandesi del 1952 e quattro anni dopo conquista la medaglia d'argento ai giochi olimpici australiani.

Figure 2

(from: "Timonieri leggendari: Dodo Gorla e Agostino Straulino" by Antonello Cherchi, *Il Sole 24 Ore*, 11/10/1992)

Looking at a 65,000 word corpus of articles on the 1992 Olympic games taken from *Il Sole 24 Ore*, we find that this expression is recurrently used in Italian to mean "win the gold medal." A search for the word *podio* within a context of five words of *gradino* produced the concordance lines in Figure 3:

Search word: PODIO + GRADINO (5 words to left or right)
dieci europei, sale **il gradino piu' alto del podio** alle Olimpiadi finlandesi del 1952
appena conquistato **il gradino piu' alto del podio** nella 4x100 (Ashford, Jones,
Oggi favoriti per **il gradino piu' alto del podio** Scarpa e Josefa Idem; outsider
la gara era mista) **sul gradino piu' alto del podio**, la cinesina Zhang Shang. La
, costretto a cedere **il gradino piu' alto del podio** alla Germania. Appassionante il
ecutiva ai Giochi, **sul gradino piu' alto del podio** ben cinque volte nell'edizione

Figure 3

(corpus: 65,000 words from *Il Sole 24 Ore* of 1992 about Olympic Games)

A search for the word "*podium**" in a 250,000 word corpus of English articles taken from *The Independent* and *The Daily Telegraph* of 1992 dealing with the same topic produced 22 citations (see Figure 4):

Search word: **PODIUM***

Sort: 1st word to left then 2nd word to left

about noon yesterday Michael Carruth was standing on a **podium** in the boxing arena here, listening to the

'You can't finish anywhere better than an Olympic Games **podium**' By COLIN GIBSON KEVIN YOUNG last

if Redgrave had never stepped on to the Olympic medal **podium** he would have been identified as one of

he wanted to make sure he said farewell from the medal **podium**. What the former Army sergeant could

three major championships I have stood on three medal **podiums**. 'Next year I will be running around

dreaming about what I am going to do when I get on that **podium**,' he says engagingly, 'and I have never

ijan. At the presentation the three men embraced on the **podium** for the last time as team-mates before

hat Young was that good. I'm proud to have been on the **podium**, though, and it's a great way to finish a

risingly, all four were expected to take their place on the **podium** at Barcelona. Edgington conceded that

tably, received huge acclaim; Skah shuffled over on the **podium** to take, rather than shake, his hand and

y stand just one game away from a definite place on the **podium** after crushing Australia 98-65 in the first

moralised the best in the world. Only as he stood on the **podium**, the first Briton to earn an Olympic gold

be good enough to win. But I was proud to stand on the **podium** after a race like that. 'It's a great way to

ORE WHEN the Great Britain team stood proudly on the **podium** in Seoul, with the Union Flag fluttering

uch harder. You can't finish anywhere better than on the **podium** of an Olympic Games.' Ever since his

Childerley, belying his carefully mapped-out route to the **podium**. Childerley Page -2- Printed Tue Jul 05

s to win the gold medal and was twice called back to the **podium** by the enthusiastic crowd. The overall

Figure 4

The corpuses confirm that *podium* indeed refers to the same concept as *podio* and can be used figuratively, but also show that it does not occur in conjunction with *the highest step* to mean *win the gold medal*. In this case the evidence provided by the corpus may help avoid a literal translation which, even if possible in principle, may have the effect of sounding awkward and non-native-like.

Furthermore, some recurring collocational patterns can be seen in the English citations. The phrase "stand on the podium" appears in six lines out of 22. In lines 3, 4, and 5

we find the phrase "medal podium" and in line 22 "two Americans stood on the winner's podium." A translator may thus decide to use this information and write, for example, that "Admiral Straulino ... stood on the medal podium..." or adhere more closely to the source text and simply write that he won the gold medal, depending on the target communicative context or on target textual constraints.

2.2. Using the Corpus to Learn about Terminology and Content

Corpora made up of specialised texts can be a useful source of terminology and content information (Pearson 1996; Bowker this volume). In the classroom, comparable corpora can be used to confirm translation hypotheses and to suggest possible solutions to actual translation problems related to a specific text. They can also provide a means to investigate similar domains or subdomains across languages. A specialised comparable corpus can offer information about terminology and concepts, and about the attestedness of expressions within a certain context. Gavioli (forthcoming) describes how learners constructed and used a comparable corpus of medical literature — the "hepatitis corpus" (see Figure 1), to acquire knowledge about a specific subject matter and use that knowledge, together with enhanced sensitivity towards target language features, to understand the source text and produce a translation. One of the main feature of medical texts is the high percentage of specialised terminology, often of Latin or Greek origin, resulting in many "look-alike" terms in English and Italian. By using a corpus of comparable texts, terms in the target language can be easily identified and their collocates compared to those of the corresponding terms in the source language to reveal diverging patterns. Consider the following sentence, taken from an article about hepatitis C published in an Italian medical journal:

> *Due mesi dopo il trapianto fu eseguita una* ***biopsia epatica*** *che evidenziò alterazioni ascrivibili a rigetto.*

The term *biopsia epatica* seems apparently unproblematic, since *biopsy* and *hepatic* are both English words belonging to medical language, as any dictionary can reveal.

A search for *biopsy* in a corpus of 30 articles on the same subject in English, however, shows that the phrase *hepatic biopsy* does not occur, while *liver biopsy* would appear to be a more satisfactory translation equivalent, with 39 occurrences. A comparison between the words *liver* and *fegato*, and between *hepatic* and *epatic** shows that in English the adjective *hepatic* collocates only with generic words such as *disease lesion* or *failure*. In Italian, on the other hand, the adjective *epatico* is preferred to compounds with the word *fegato*, the only exception being *trapianto di fegato*.

2.3. Using the Corpus to Explore Texts

A third approach to comparable corpora involves using them to investigate a particular genre and/or topic area, as in pre- or post-translation activities. Comparable corpora can be the source of a potentially endless "serendipity process" (Johns 1988), as one word or phrase leads to another, depending on the learner's intuition and individual proficiency, interests or needs. Comparable corpora provide learners with a means of testing the relationship between items of language which they perceive as holding some kind of similarity or equivalence. Learners can be instructed to look for similarities between languages and to compare words and phrases by identifying categories which have a strong formal resemblance, such as proper names and cognates (see Partington 1995), or which are proposed as translation equivalents in dictionaries.

For example, a search for the proper name *Mitterrand* in a comparable corpus of foreign news articles about France (from *The Independent, the Daily Telegraph,* and *Il Sole 24 Ore* of 1992, of about 152,000 in English and 102,000 words in Italian) shows some differences in the way Mitterrand is talked about in the two languages. The search produced 152 and 188 concordance lines in Italian and in English respectively (0.13% in each corpus component). Figure 5 lists all two-words clusters occurring in these concordance lines with a frequency higher than ten.

1	...francois mitterrand...	41	1	...mr mitterrand...	59
2	...di mitterrand...	21	2	...francois mitterrand...	23
3	...il presidente...	21	3	...president mitterrand...	23
4	...mitterrand e...	16	4	...president francois...	18
5	...presidente francois...	14	5	...m mitterrand...	12
6	...a mitterrand...	11	6	...mitterrand has...	12
7	...mitterrand che...	11	7	...mitterrand said...	12
8	...mitterrand ha...	11	8	...mr mitterrand's...	11

Figure 5
Two-word Clusters from MITTERRAND

A first observation is that the Italian texts seem to be much more at ease in calling politicians by their full or last name, e.g. *Francois Mitterrand* or simply *Mitterrand*, whereas the English texts preferably call him *President Mitterrand or President Francois Mitterrand* and *Mr* or *M Mitterrand*. This may seem somehow counter-intuitive, as it is known that Italians are generous in using titles, whether it be *dottore* (anybody with a university degree), *professore* (any teacher in secondary education) or *geometra* (land surveyor).

This observation can be followed up by a search for all contexts where the words *French* and *President* (Figure 6) and *francese* and *presidente* (Figure 7) respectively co-occur.

In the Italian corpus we find that the phrases *Il presidente francese Mitterrand* or *Il presidente francese Francois Mitterrand* are quite common, while the English texts rarely use the adjective *French* and prefer to avoid long phrases such as *the French President Francois Mitterrand* or *the French President Mitterrand*. A further activity may be to scan the full texts around citations to investigate the structural position of these references. What emerges is that English articles use the name *Mitterrand* without a title only in headlines. The body of the article usually introduces him as *President Mitterrand* and later on refers to him as *Mr Mitterrand* or *the President*. In the Italian articles he is as often first mentioned as *Francois Mitterrand* as *il Presidente Mitterrand*, and subsequent references are to *il presidente/il presidente francese*, or simply to *Mitterrand*. There are of course exceptions, but general patterns of reference can be observed and compared. These can then be used as stylistic models to relate to when writing in the foreign language.

This corpus can also be used to highlight differences in the lexis used to introduce direct speech. Most of the time, President Mitterrand is quoted as saying something. In

Search word: PRESIDENT+ FRENCH (5 words to left or right)

Sort: 1st word to right then none

he had only a brief meeting with the **French President** and was given finger-wagging lectures on the need to support

in borrows to defend sterling in ERM, **French President** champions European union: Andrew Marshall reports from Paris on

uly. It will be the full fig. The **French**, from the **President** downwards, have obviously woken up to the potential for

d a convulsion within the **French** body politic. **President** Francois Mitterrand's seven-year term has three years to run: even

. But the German Chancellor and the **French President** issued no formal statement after their meeting, despite convening a

nsequences for **French** agriculture', he said. **President** Mitterrand gave his full support to M Beregovoy's stance before flying

re, centrepiece of **French** culture and also of **President** Mitterrand's immense programme of artistic and architectural

the Bill submitted to the **French** people by the **President** of the Republic authorising the ratification of the treaty on

nth. M Rene Monory, a Centrist, was elected **president** of the **French** Senate yesterday after a battle between opposition

IM By FRANK KANE JEAN Franois Theodore, **president** of the **French** stock market ruling body, the Societe des Bourses

abash land By Our Paris Correspondent THE **president** of the **French** Red Cross and three government advisers resigned

bilisation,' said Simone Veil, a **French** former **president** of the European parliament yesterday. The **French** government is

move against this wave of xenophobia. The **president** of the **French** rugby federation last week pointed to a pile of letters

hich caused the damage. Bernard Lapasset, **president** of the **French** Rugby Federation, said yesterday he hoped the South

Figure 6

English, quotations seem to be mostly introduced by the verb *say*. A concordance of all occurrences of pronouns, *Mr*, *Mrs*, *Ms*, and *Pr** followed or preceded by a quotation mark (within a span of seven words on each side) allows the learner to compare how direct speech is reported in the Italian and the British articles.

A search for *ha* (has) in the vicinity of a quotation mark in the Italian texts identifies a list of functionally equivalent verbs to "dire" (in the third person of *passato prossimo*). By scanning the many concordance lines thus generated (a few hundred) and grouping them through different sortings, recurring lexical choices can be isolated and further investigated. In the English texts, beside *say* (which accounts for more than a third of the occurrences), and *add* (about a fourth), the following verbs are recurrently used in conjunction with reported speech: *admit*, *announce*, *ask*, *insist*, *declare*, and *complain*. In Italian the most frequent are: *dire*, *dichiarare*, *aggiungere*, *affermare*, *chiamare*, *concludere*, *insistere*, *ribadire*. *Dire* and *aggiungere* are the most common, but they account for less than half of the total verbs used to introduce direct speech, unlike

Search word: PRESIDENTE + FRANCESE (5 words to left or right)

Sort: 1st word to right then 2nd word to right

stricht l'ha spezzata anche Francois Pe'rigot, **presidente** della Cnpf, la Confindustria **francese**, che dopo essersi astenuto dal

into solo un **francese** su tre. Jacques Delors, **presidente** della Commissione di Bruxelles ha "ringraziato per la Francia, per ersi al mercato _ afferma Micha Spierenburg, **presidente** della filiale **francese** della banca d'investimento britannica

tto di legge presentato al popolo **francese** dal **presidente** della Repubblica autorizzando la ratificazione del trattato sull'Unione

primo ministro **francese**, Pierre Be're'govoy. Il **presidente** della Repubblica, Francois Mitterrand, ha deciso di affidare la

francese, sara' a Parigi per esaminare con il **presidente francese** Francois Mitterrand i prossimi passi da compiere per la

" sull'Europa PARIGI _ La mina vagante che il **presidente francese** Francois Mitterrand ha lanciato nel campo dell'opposizione

UROPA OCCIDENTALE Didascalia FOTO-01 Il **presidente francese** Francois Mitterrand arriva al seggio del referendum per l'

ato di non avere il monopolio dell'iniziativa, e il **presidente francese** ha detto che altri Paesi, Italia in testa, hanno gia'

attacco contro l'Alleanza. All'inizio del 1991 il **presidente francese** Mitterand e il cancelliere tedesco Kohl decisero di costituire

ile la situazione politica nella zona del Golfo. Il **presidente francese** Mitterrand, secondo due leader curdi che lo hanno n duro attacco contro Francois Mitterrand. Al **Presidente francese**, che la sera prima aveva criticato Londra per il voltafaccia su

Figure 7

Corpus: foreign news articles about France (from *The Independent, The Daily Telegraph,* and *Il Sole 24 Ore* of 1992, about 152,000 in English and 102,000 words in Italian)

their English dictionary equivalents *say* and *add* which account for more than 60%. Some of the verbs recurrently used in Italian and in English have cognates or other accepted dictionary equivalents which do not always seem to have the same distribution or frequency. For example, dictionaries give as equivalents for *affermare* a cognate (*affirm*), together with *assert* and *state*. However, while a search for *affermare* produces a considerable list of occurrences, a search for forms of the verbs *assert*, *affirm* and *state* does not produce a single citation.

Translators can draw the conclusion that *affermare* could be used to translate verbs such as *say* and *declare*, while to translate it into English *affirm* or *assert* is not the safest bet in this context. Of course, no corpus can tell us that something is not possible (i.e. cannot be said) — but it can provide some evidence as to how likely a word or expression is to occur, i.e. how routine it is.

Comparison between related words can also highlight aspects of text conventions in similar genres which vary across languages.

For example, if we compare a concordance of *ha detto* (Figure 8) with one of *said* (Figure 9), we can immediately see that a pattern emerges.

Figures 8 and 9 both show concordances obtained by putting quotation marks as context words within a short span to the left of the keywords in order to select fewer

HA DETTO + "* (10 words to the left)

14 lines out of a total of 54 "ha detto"

dicembre a Strasburgo. "Stiamo studiando la possibilita' _ **ha detto** il portavoce dell'Asaja _ di organizzare forme di protesta

ttersi dalla guida della Commissione. "Perche' dimettermi _ **ha detto** il presidente della Commissione _ sono in buona salute e

unitario. "Non abbiamo paura di una guerra commerciale _ **ha detto** Luc Guyau, presidente della Fnsa, la piu' importante

quadrato e dare battaglia. "Il mio avviso di convocazione, **ha detto** ieri Emmanuelli, e' stato dato alla stampa il 9 luglio, porta la

bra si sia scelta la seconda via. "La riduzione dei tassi _ **ha detto** ieri Sapin _ e' diventato un obiettivo di tutti i Paesi europei.

) e realizza all'estero il 48% del giro d'affari. "La societa' — **ha detto** il presidente Pierre Conso — sara' lo strumento di sviluppo

tro la fine dell'anno. "Sono piu' ottimista di dieci giorni fa _ **ha detto** Delors, aggiungendo che l'incontrodei ministri degli Esteri

zione nel suo Paese. "Se la Russia fallisce le sue riforme, **ha detto** Eltsin, l'unica alternativa e' la dittatura. Chiediamo quindi

diritto di contare sull'aiuto della comunita' internazionale", **ha detto** il leader russo, evocando immagini di camicie rosse o

te e' che tale forza diventi il piu' multinazionale possibile", **ha detto** Mitterrand. La scelta franco-tedesca e' importante e

ullato. "Non e' il miglior accordo per gli agricoltori europei" **ha detto** un portavoce dell'associazione spagnola dei coltivatori

on Delors. "Prendero' parte ai negoziati con gli Stati Uniti"

Figure 8

instances. A few lines where *ha detto* or *said* appeared in connection with reported speech have been manually deleted.

When *ha detto* is used to introduce direct speech, it is often embedded in the quotation (lines 1-8), while in English *said* usually occurs in a sentence's final position (lines 1-21).

All this goes to show that Italian newspaper articles have different stylistic conventions than English ones. In many cases, learners may not be aware of these differences: they may assume that their knowledge of the language is adequate to provide an appropriate translation. For instance, words like *prices* and *prezzi* may not seem to present any problems as translation equivalents. A comparison of two concordances (Figures 10 and 11) may however reveal some interesting information. The figures show one out of five randomly selected occurrences in a corpus of about 500,000 words for each language taken from Italian and English quality newspapers.

What is noted here will depend on the learners and the particular task at hand: for instance, some may see that while *prezzi* is usually preceded by the definite article *i*, *prices* is not. Others may note instead that nominal groups such as *l'aumento dei prezzi* or *la discesa dei prezzi* are more common in Italian, while verbal groups such as *prices soared* or *prices fell* are more common in English. Others may enlarge their vocabulary by observing

SAID + "* (7 words to the left)

26 lines out of a total of 404 "said."

elf- control. 'They put themselves under pressure,' Astre **said**. 'And the addition of all these mistakes led directly to their loss of

turning point. 'Perhaps there was a refereeing error,' he **said**. 'But how did we get into that situation in the first place? We got

tuff, he was ready. 'The race fit my style today,' Indurain **said**. 'I needed a good race and I got a good time. Sure I gained an

meras next year. 'That's probably my last match here,' he **said**. 'I would have liked the chance to play one of the top guys, but

low to the head and they were told 'it's over',' Mr Mauroy **said**. 'Of course, the number of farmers is going to diminish. Don't put

elors equals socialism, protectionism and selfishness,' he **said**. 'We cannot go on like this. We don't need another Sun King in

French territory. 'It is absurd, illogical and unpleasant,' he **said**. 'We had total faith in the Agnellis in Italy, and now they come

efforts have topped the ratings. 'I am a little tired,' Indurain **said** after 3,983 kilometres (2,474 miles) of racing through seven

nmnent was probably 'seeking compensation', Mr Mauroy **said**. Commenting on whether France would use the Luxembourg

nt runners, 5min 6sec behind. 'We've limited the damage,' **said** Lino. 'We expected to lose two minutes to Panasonic, who set

he fall in property prices had reached 'excessive levels', **said** Mr Sapin. The moves are an attempt to reassure banks and

uled out a mark revaluation. 'I don't see that,' Mr Tietmeyer **said**. Mr Waigel also pledged that Germany would continue to meet

Figure 9

the range of verbs which collocate with prices: prices go up, shoot up, increase, soar, rise and skyrocket or go down, abate, fall, plunge, tumble, and plummet. Often, it is not so much that a concordance reveals aspects of contrastive analysis which went formerly unnoticed (even if this can at times be the case). Rather, a concordance gives striking visibility to certain discourse structures (Stubbs 1996) and can thus present the translator trainee with clear patterns of use.

3. TOWARDS A "TRANSLATOR TRAINEE WORKSTATION"

Contrastive analysis of comparable corpora can reveal how similar ideas and concepts are expressed in similar texts in different languages. The analysis may regard stylistic preferences related to conventions of rhetorical and propositional structure, figurative language, lexico-grammatical features and collocational patterns. Comparable corpora provide information on the way discourse is realised as text in different languages, which is to be taken as an indication of recurrent patterning rather than as a normative statement.

Corpora reveal regularities, not rules; evidence that emerges from data is subject to interpretation and always needs to be verified against larger corpora. Learners should also be aware that the evidence provided by, for example, a corpus of quality press articles may

Search word: PREZZI. Sort order: 2nd word to left, then 1st word to left

recente, sensibile abbassamento dei **prezzi** sui mercati tedeschi ha reso meno
tto Saddam", che si e' abbattuto sui **prezzi** petroliferi dall' inizio di
e di flessione si e' avuto anche sui **prezzi** del provolone, mentre il burro ha
riforme. Gli stock aumentano, alcuni **prezzi** scendono e inevitabilmente
di attendersi ulteriori aumenti dei **prezzi**". Secondo la fonte, l' incertezza
lettriche. Concedendo l' aumento dei **prezzi** chiesto dall' Enel, il Governo
erso il terzo round e un aumento dei **prezzi** petroliferi in una condizione
ella degli interessi; l' aumento dei **prezzi** si scarichera' su BoT e CcT e
usa prima del recupero autunnale dei **prezzi**. Il movimento rialzista non era
l mercato comunitario, la caduta dei **prezzi** e la necessita' per la
in malinconici ammassi; il calo dei **prezzi** internazionali del frumento fa
a scorta di manovra ogni volta che i **prezzi** scendevano sotto un livello
na allo Sme il tasso di crescita dei **prezzi** in Gran Bretagna dovra' scendere
a del mercato del Golfo e crollo dei **prezzi**, anche sull' onda della caduta
nsieme, hanno generato un crollo dei **prezzi** e reso necessarie delle misure d'
he da carne). Domanda molto debole e **prezzi** nettamente cedenti, per contro,
Medio Oriente ha mantenuto elevati i **prezzi** delle scadenze lontane e non ha

Figure 10

differ from one drawing on the popular press. The extent to which features of a specific corpus can be generalised would be a useful subject for further investigation.

Comparable corpora offer no single translationally established correspondence, but present a repertoire of recurring collocational or structural features which provide a basis for establishing equivalence between stretches of texts in two languages. Frequency can provide an indication of the expectations the prospective audience might have about the linguistic features of a target text in a particular genre.

Translation activities based on bilingual corpora can be integrated into the curriculum of trainee translators and provide a means of learning about aspects of language that are otherwise not easily detectable.

Concordances can be printed out and examined in the translation class as an aid to the analysis of source and target text features. Corpora can also be accessed directly by the learners while performing a specific translation task. A "translator trainee workstation" comprising a word processor, bilingual corpora and facilities for bilingual concordancing together with other resources may constitute a valuable aid in the training of translators. Software for bilingual concordancing of parallel corpora is already available on the market (e.g. *MultiLingual Concordancer for Windows*, *Wordsmith Tools Text Aligner*, *Paraconc*) and bilingual corpora of various kinds and sizes are being created, possibly to become available to the general public in the form of electronic support or on-line facilities through the Internet.[7] While statistical alignment techniques (see McEnery and Wilson 1996) can be used in dealing with parallel corpora, different methods have to be adopted for providing bilingual concordances of comparable texts. Since they are compa-

Search word: PRICES. Sort order: 1st right, 2nd right
tions against the steep rise in food **prices**. A court in Marrakesh sentenced
tonnes might be delivered, squeezing **prices** acutely. In that event, smaller
day, but said no relief from soaring **prices** and food queues could be expected
will step in themselves to maintain **prices**, as they did in the wake of the
eir nadir because of the hike in oil **prices**. At the same time inflation and
ven in the 1973/74 collapse, nominal **prices** did not fall. The `wealth effect"
s. Properties will stick rather than **prices** fall." Barclays Bank said that t
p 3.9 points at 1,775.6. Bombay: **Prices** fell sharply for the second day
are traded publicly. Last month junk **prices** fell sharply on news that Campeau
market drifted in sympathy. Bombay: **Prices** fell on heavy speculative selling
poll discloses a decline in selling **prices** for the second successive quarter
weekend sales in New York and Paris **prices** for post-war French artists soare
ominated a moderate session in which **prices** generally fell back from early ad
delayed to the end of 1991." House **prices** have already begun to fall across
ial said. Jaguar, which dropped its **prices** in North America last year, now
yield the best opportunities, as the **prices** of companies fall. Jon Moulton,
st's cocoa crop indirectly pushed up **prices** on the London Futures and. Option
ors. Page 22 World Markets New York: **Prices** plunged and by the

Figure 11

rable at a global level, no one-to-one correspondence would be found between pairs of sentences. What could be retrieved from a comparable corpus are similar textual environments. Picchi and Peters (1997) describe how correspondences can be established between clusters of dictionary equivalents through a bilingual lexicon and computational techniques, somehow automating part of the process for establishing relationships between bilingual textual occurrences.

While all language learners could benefit from activities based on comparable corpora, it seems to me that trainee translators would be particularly motivated to use them, as they not only need to enhance their linguistic competence but also acquire specific skills related to translation, e.g. cross-linguistic mediation, accuracy in text production, and the ability to process text in electronic format.

Notes

1. A version of parts of this paper was presented at the XVIII AIA Conference (Genova 1996). I am grateful to all those who offered comments and suggestions on that occasion.

2. Many well-established commercial translation systems, such as Trados Translator's Workbench, are based on the concept of translation memories. For more information on machine translation systems see http://www.sslmit.unibo.it/zanettin/cattools.htm

3. For a description of pedagogical applications of multilingual parallel corpora see "http://sun1.bham.ac.uk/johnstf/lingua.htm"

4. According to Hartmann, a third kind of bilingual texts, labelled "adaptations," is exemplified by the production of advertising copy for different countries or by authoritative multilingual versions of international law.

It is characterized by simultaneous formulation with reference to a common source. While equivalence is not here based on a "textual" identity (from "original" texts to "secondary" texts) but on conceptual identity, it can be argued that all translations are adaptations to the "respective conventions of the two languages for the purpose of conveying an identical message to receivers of sometimes very different cultures" (Hartmann 1980: 38).

5. The terminology used for bilingual texts and corpora is not yet fully established. For example, Snell-Hornby and Shäffner talk about "parallel texts." See also the entry for "Corpora" in the *Dictionary of Translation Studies* (Shuttleworth and Cowie 1997).

6. For further information see Corpus Linguistics Page at http://www.sslmit.unibo.it/zanettin/cl.htm

7. For some examples of on line bilingual concordances see http://www.sslmit.unibo.it/zanettin/cltrans.htm

REFERENCES

BAKER, M. (1995): "Corpora in Translation Studies: an Overview and some Suggestions for Future Research", *Target*, 7, pp. 223-243.

BAKER, M. (1996): "Corpus-based Translation Studies — the Challanges that Lie Ahead", H.L. Somers (Ed.), *Terminology, LSP and Translation*, Philadelphia/Amsterdam, John Benjamins Publishing Company.

BARLOW, M. (1996): "Parallel Texts in Language Teaching", S. Botley, J. Glass, T. McEnery and A. Wilson (Eds), *Proceedings of Teaching and Language Corpora 1996. UCREL Technical Papers*, 9, Lancaster, UCREL, pp. 45-56.

BIBER, D. (1993): "Representativness in Corpus Design", *Literary and Linguistic Computing*, Vol. 8, No. 4, Oxford, Oxford University Press, pp. 243-257.

DRYBERG, G. and J. TOURNAY (1990): "Définition des équivalents de traduction de termes économiques et juridiques sur la base de textes parallèles", *Cahiers de lexicologie*, 56-57, pp. 261-274.

GAVIOLI, L. (forthcoming): "Corpora and the Concordancer in Learning ESP. An Experiment in a Course for Interpreters and Translators", G. Azzaro and M. Ulrych (Eds), *Lingue a confronto. Atti del XVIII Convegno AIA, Genova, 30 Settembre-2 Ottobre 1996, vol. II*, Trieste, EUT.

HARTMANN, R. R. K. (1980): *Contrastive Textology. Comparative Discourse Analysis in Applied Linguistics*, Heidelberg, Julius Groos Verlag.

JOHNS, T. (1988): "Whence and Whither Classroom Concordancing?", E. van Els *et al.* (Eds), *Computer Applications in Language Learning* (Foris).

KENNEDY, G. (1992): "Preferred Ways of Putting Things with Implications for Language Teaching", J. Svartvik (Ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*, Berlin, Mouton de Gruyter, pp. 335-373.

LAFFLING, J. (1992): "On constructing a Transfer Dictionary for Man and Machine", *Target*, 4, pp. 17-31.

LAVIOSA, S. (1997): "How Comparable can 'Comparable' Corpora Be?", *Target*, 9 (2), pp. 289-319.

PARTINGTON, A. (1995): "'True Friends are Hard to Find': a Machine-assisted Investigation of False, True and just Plain Unreliable 'Friends'", *Perspective Studies in Translatology*, 95 (1), pp. 99-112.

PEARSON, J. (1996): "Teaching Terminology using Electronic Resources", S. Botley, J. Glass, T. McEnery and A. Wilson (Eds), *Proceedings of Teaching and Language Corpora 1996. UCREL Technical Papers*, 9, Lancaster, UCREL, pp. 203-216.

PICCHI, E. and C. PETERS (1997): "Reference Corpora and Lexicons for Translators and Translation Studies", A. Trosberg (Ed.), *Text Typology in Translation Studies*, Amsterdam and Philadelphia, John Benjamins Publishing Company.

SCHÄFFNER, C. (1996): "Parallel Texts in Translation", *Paper presented at Unity in Diversity?, International Translation Studies Conference, Dublin City University, 9-11 May 1996*.

SHUTTLEWORTH, M. and M. COWIE (1997): *Dictionary of Translation Studies*, Manchester, St Jerome Publishing.

SCOTT, M. (1996): *Wordsmith Tools*, Oxford, Oxford University Press.

SNELL-HORNBY, M. (1988): *Translation Studies. An Integrated Approach*, Amsterdam and Philadelphia, John Benjamins Publishing Company.

SOMERS, H.L. (1993): "Current Research in Machine Translation", *Machine Translation*, 7, pp. 231-246.

STUBBS, M. (1996): *Text and Corpus Analysis*, Basil Blackwell.

TEUBERT, W. (1996): "Comparable or Parallel Corpora?", *International Journal of Lexicography*, 9 (3), pp. 238-264.

MCENERY, A. and T. WILSON (1996): *Corpus Linguistics*, Edinburgh University Press.

WOOLLS, D. (1997): *MultiLingual Concordancer for Windows*, Collins COBUILD.

ZANETTIN, F. (1994): "Parallel Words: Designing a Bilingual Database for Translation Activities", A. Wilson and T. McEnery (Eds), *Corpora in Language Education and Research: a Selection of Papers from Talc 94. UCREL technical papers*, 4, Lancaster, UCREL, pp. 99-111.

ZANETTIN, F. (forthcoming): "Swimming in Words: Corpora, Translation, and Language Learning", G. Aston (Ed.), *Learning with Corpora*.