Computer Corpora and the Law: a New Approach to the Translation of Legal Terms

Gill Philip, CILTA, University of Bologna

KEYWORDS

ABSTRACT

The use of computer corpora for the analysis of legal language is not common practice; still less the use of parallel corpora for the comparison of legal terminology. The Bononia Legal Corpus project (BoLC) began two years ago, and now as the first stage reaches completion, some important issues must be addressed. This paper will present findings from the first stage of the project, in which European Community directives and judgements have been analysed to identify actual translation equivalents in Italian and English. The main points to be discussed are the clarification of concept boundaries when dealing with terminology in different languages, and the help or hindrance of dictionaries in identifying translation equivalents.

INTRODUCTION

Although the legal system in operation in Italy is based on civil law, students attending the Law Faculty at the University of Bologna follow a course in English common law as a means of studying the implementation of a system which is radically different in conceptual terms to their own. Those students who follow the course are required to have a good (upper-intermediate) knowledge of English, yet they experience language difficulties all the same which make it all the harder to critically appreciate and analyse the differences between the English and Italian systems. One of the major problems is that there is a distinct lack of translated material; the coursebooks used contain the original English text with only a summary or extended commentary in Italian, rather than a full translation. As a result of this, both staff and students are faced with the task of allocating ad hoc translation equivalents to the English terms that they encounter, and in so doing lay themselves open to the traps of over-simplification and mistranslation.

For several years, the University's Centre for Theoretical and Applied Linguistics (CILTA) has provided specialised courses in English for students attending the Law Faculty. The success of this initiative led to the Faculty proposing a translation project based on the new computer technologies; in particular corpus linguistics. A corpus was to be created from which translations could be extracted manually at first, and eventually by automatic means. This aim was later proved to be far more ambitious than it had originally appeared, but did not seem in any way impossible at the outset, especially if the latest developments in retrieval software, such as aligners and other text-matching applications, were taken into consideration. It was decided that a two-year pilot project would be carried out to ascertain the viability of the proposal, and that this first stage would involve the analysis of parallel texts which have the obvious advantage having already been translated. It would therefore have been theoretically possible to find all translation equivalents for all of the terms required and thus compile a complete glossary of the equivalents.

OUTLINE OF THE PROJECT

Data

The Directives and Judgements of the European Commission, (up until 1995 for Italian, and 1996 for English) formed the data used for the pilot project. Being parallel texts, they are translations of each other, but unlike the majority of parallel corpora, neither one is an untranslated original. This is because the European Commission has two official languages (English and French) into which all texts are translated; full translations are made from these official languages into the other languages of the Member States. It is therefore common for individual texts to have been translated twice before appearing in official documents, a fact which contributes to the distinctive *officialese* characteristic of the European Commission style.

It should be stressed at this point that European Law is based on the tenets of Civil Law. The English translation is not adapted in any way to account for the different legal system in operation in England and Wales, as it is for the national law-making bodies to ensure that European rulings be fully embodied within national legislation. This uniformity in the content of the texts further aided the project in that the translations encountered were known to be functionally equivalent in every way, and ensured that translation problems related to system differences, such as those discussed in Smith (1995), were left to the second stage of the project.

The totality of the data used in the project amounts to around ten million words, a size large enough to give reliable results, yet small enough to be manageable within the timescale permitted. The data set is also finite (all available data was used), allowing us to sidestep the thorny issue of representativity of the corpus. The texts were entered into a database and, for ease of access and analysis, divided by language and text type into four sub-corpora (English Directives, English Judgements, Italian Directives, Italian Judgements). The CUE software used (Mason 1996, 1997) was designed with English in mind, and had to undergo various changes before it could cope with Italian accented characters and some features of punctuation.

Methodology

The use of corpora is in itself indicative of a particular ideological stance: data must be analysed, and results must be replicable. The core methodology used by corpus linguists is presented in Sinclair (1991), although the principle of the 'contextual theory of meaning' (Firth 1957) actually predates the use of computers in language research by several decades.

The corpus is accessed with the help of data-retrieval software, and analysed either on-screen or on a computer printout, and is presented either in sentence format or in a KWIC (key-word in context) concordance. This second option was adopted, as it facilitates the analysis of large quantities of data, as the key-word (the *node*) appears in the middle of each line, surrounded left and right by the linguistic context in which it occurs (Clear 1993: 276-277). KWIC makes it possible to read concordances both horizontally, as text extracts, and vertically, comparing the collocations or structural patterns which recur in the context of the node word. In the context of this study, the terms to be translated were identified in large part through such an analysis of repeatedly co-occurring features.

However, the software does not do the analysis; the linguist does the real work, and it is at this stage that the ideological approach should be identified. There are, broadly speaking, two schools of thought in corpus linguistics, discussed in detail by Tognini-Bonelli (1996: 54-76), and these have a huge impact on the actual handling of the language data. On the one

hand, there are the *corpus based* linguists, who use the corpus to check or give authority pre-existing theories, and who are liable to discard, dismiss or 'correct' any data which is not in accordance with their intuitions (see, for example, Fillmore 1992: 45). The other approach is *corpus driven*, in which the integrity of the data (which is viewed as no more or less than a representative sample of language) is paramount. This does not mean that the linguist's intuitions are suppressed; on the contrary, they are often brought to light through the acceptance that the language ideals of one linguist may not necessarily be those of others, or indeed of the population as a whole.

The legal corpus project (Bononia Legal Corpus, or BoLC) was to take the corpus driven approach to study, for several reasons. Firstly, the corpus was finite, and not merely a sample of available data, therefore no doubts could be raised about its representativeness. For this reason, any findings which were not in accordance with existing reference works could not be dismissed, especially given that the project had come about because the information required was not already available or had already been questioned. A second consideration is that there were two languages in use; intuitions about language are considered acceptable only so long as the language in question is the linguist's mother tongue, which was not the case in this project. Even though all of the linguists working on the project were highly competent in both languages, none were bilingual. Furthermore, although the Law Faculty provided the word list of terms for study and translation, legal experts were not involved in the linguistic analysis. The linguists had no *a priori* terminological knowledge, therefore the only way forward was to work in as objective a way possible, through the analysis of collocations and frequency counts.

The role of the dictionary

Dictionaries, be they mono- or bilingual, specialised or general, are the first port of call for any translator, and this project was no different. It was not expected that there would be substantial differences between documented translations and those arrived at by analysis of the corpus. As the analysis proceeded, however, it became clear that the usefuless of dictionaries was limited, especially as many of the suggested translations simply did not materialise even in EURODICAUTOM, a multilingual on-line dictionary provided by the European Commission itself. Had the corpus been sampled, this could have been put down to a problem with the data selection, and subsequently ignored or explained away; but in the event these discrepancies moved the research on in the direction of data-driven translation, where translation equivalents were identified by analysing the data and matching up parallel concordance lines.

USING A CORPUS FOR TRANSLATION

The usefulness of corpora in translation studies has been discussed elsewhere (see especially Baker 1993), and translators have become avid users of specialist, often <u>ad hoc</u> corpora. Yet it is still a relatively unexplored field of academic study, with no universally established methodology other than that which governs corpus linguistics in general. These basic principles were refined in the course of the two years working on the pilot study, resulting in a standard procedure which was clear and easy to follow as well as allowing the data to be analysed comprehensively.

1. Identify the collocations

The Law Faculty produced a list of around 60 terms which they wanted us to analyse and translate, covering the areas of contract law, corporate law, and tax law. The list was bilingual in that provisional Italian translations were given to the English terms. The BoLC project was intended to check and, if necessary, correct the translations assigned to the English terms.

The word list contained a few precise items, such as *fringe benefits* but tended to favour more general expressions such as *tax* or *loan*, which seem uncomplicated until analysis begins: whereas a fairly restricted and well-defined term such as *fringe benefits* has both a well-defined semantic field and an easily identifiable translation, general terms have to be made more specific and their semantic fields narrowed down before there is any possibility of finding the correct translation. The most reliable way of delimiting the concept boundaries of very general terms is to carry out a collocations analysis to identify the words that cluster around the nodeword(s) and thus form compounds which may or may not be established legal expressions. In the case of *tax*, over 70 collocations were identified, which were reduced to around 50 that were considered to be significant (occurring more than twice in identical form). This procedure created 50 terms to search for rather than just one, but paradoxically this made the task more manageable because every one of these collocations can be seen as a restricted term, as has just been discussed. Given that three possible translations had been assigned to *tax* in the word list (*imposta*, *tassa* and *tariffa*), the collocations analysis made it possible to isolate some of the different meanings that would be translated into each of these Italian words. The importance of this can be appreciated when we consider that bilingual dictionaries give no more specific information than that *tax* can be translated as *imposta* or *tassa* if it is a noun, and as *del fisco*, *fiscale*, tributario, *d'imposta* or *delle tasse* if it is an adjective. This is of very little practical help to those who intend to translate or merely understand the meaning of the term in question, because it does not explain where the semantic boundaries of each of the possible translations lie. Even though all large dictionaries provide a list of collocations of the headword, they give no indication of the general rules which govern the choice of one expression over another, so if a term is not illustrated, no suggestions are made as to how it should be translated. By identifying the significant collocations of a

headword, especially in the case of BoLC, where the data set was finite, it is

possible not only to state what these are and find their translations, but also to make general statements about the choice of term used to translate the word *tax* in its collocations.

2. Translate the collocations

The second stage of analysis is to find the translations of the first language (L1) collocations in the second language (L2) corpus. There are three possible ways of going about this. The first of these is to translate the English collocations with the help of both specialised and general bilingual dictionaries. The second is to translate the general terms as they appear on the word list, and to do the same work as described in the previous section to identify the words which collocate with the Italian terms. The L2 collocations can then be compared to the L1 collocations, and matched up using the context to help identify parallel phrases. The third way is to use an aligner, a computational tool which is used to match up (align) parallel data which works by identifying the same position in two parallel texts, allowing them to be analysed side by side. This option was discarded because of the text types under examination; lawyers are notorious for their eternally long sentences, which are often actually paragraphs or indeed entire pages of text. Most aligners do not function below the level of the sentence, and so it was felt that an aligner would be of little help in this project.

3. Look for the translated terms

The initial analysis of *tax* was carried out using a combination of methods 1 and 2. The collocations were assigned provisional translations based on data in a multilingual Legal dictionary (Epstein 1988), a bilingual Italian-English dictionary (Ragazzini 1995), and EURODICAUTOM, the European Commission's online dictionary, provided for the use of professional translators. These documented translations were then checked against the collocations of the basic terms (*imposta*, *fiscale*, and so on).

In some cases, no translation can be found. In the analysis of *tax*, where there were nearly 8000 occurrences, 'untranslatable' collocations were discarded if they occurred less than 5 times, because the time spent searching for the translation equivalent would have been disproportionate to any results that might have been obtained. Had there been fewer occurrences, the cut-off point would have been correspondingly lower. If the translation is too frequent to be discarded, it is useful to look at recurrent features in the extended context, such as proper names or article numbers, distinctive collocations (for example *banana tax*), and fixed phrases and formulae. By translating these contextual features, it is usually possible to arrive at the translation by default. Rarely no translation equivalent can be found because the terms have been paraphrased rather than translated. As a paraphrase is not the same as a translation, the term has no

equivalent and is discarded unless a standard paraphrase is recurrently used, which was not the case in this study.

4. Compare the frequencies

When a term is found, a crude number count should be made to establish whether or not it can reasonably be the translation that is needed. If there is a clear discrepancy between the frequencies of the two terms, as was the case with *tax authorities* (132) and *autorità fiscali* (48), the translation equivalent can be considered possible, but cannot be considered definite until the other occurrences have been accounted for. The search for is then resumed until all other possibilities are exhausted. When the numerical frequency of the term in L2 matches that of L1 with more than 80% accuracy, it can be considered the main translation equivalent. Even though it is clearly not the only translation, it is usually not very productive to spend time looking for alternatives if this level of accuracy has already been reached.

5. Match up the concordance lines

As a final check, L1 concordance lines should be matched up with their corresponding L2 concordance lines. This involves the translation of contextual features as mentioned in step 4, though to a lesser degree, and it certainly need not be done for every single occurrence. The aim of this is to move a stage beyond mere number-crunching and to make sure that the translations occur where they ought to.

Because of the highly formalised nature of the language of EU legislation, some headwords not only collocate with single words, but can form quite lengthy multi-word units, as in *tax evasion and tax avoidance* / *frode o evasioni fiscali*, which are in turn found within extended formulae. Because of this, the work was considerably slowed down, as it was often impossible to analyse effectively on-screen. The monitor display, limited to 110 characters, often proved insufficient when trying to analyse the extended context of these multiple collocations. The different stylistic features of English and Italian contributed further to this problem: English nouns tend to be premodified, whereas Italian favours postmodification. In practice, this means that when a node word is being studied, the context of each collocation is visible predominantly to the right of an English term, and to the left of an Italian one. When these factors rendered the identification of parallel examples virtually impossible on-screen, the only solution was to print out the concordances (with as much as 5 lines of expanded context) and to do the analysis on paper. This somewhat defeats the purpose of using a computer corpus, as the data cannot be reselected or sorted once it is on the page.

100% accuracy is an ideal that has little success in the real world. Not surprisingly then, none of the Italian translations identified in this project could be assigned with absolute precision to their English counterpart, although many were very close (with over 95% accuracy). There can be surprises, though, as will be observed in the case study below: supposedly equivalent terms are sometimes not all that they seem.

CASE STUDY part 1 - DIRECTIVES

The information gleaned from the contextual features surrounding the legal terms proved invaluable to the translation process. This was especially true when the point of departure is a dictionary translation, and even more so given the similarities between English and Italian lexicons. *False friends* - those words which look alike but have different meanings - are a translator's nightmare. To illustrate the point we will look in detail at *tax evasion* and its translations.

tax evasion / *evasione fiscale*

The term *tax evasion* was one of the 50 collocates that were studied in order to arrive at a definitive Italian translation of *tax*. There are 64 occurrences of the term in the English data. Having carried out stage 1 on the headword *tax* to arrive at *tax evasion*, we now have to go on to stage 2 to identify the possible translations. The sources used (Epstein 1988 and Ragazzini 1995) both gave the translation *evasione fiscale*. This term was searched for in the Italian data, but only 24 examples were found, leaving a further 40 to account for, and clearly showing that there was a mismatch between the two terms.

It has already been suggested that one of the most efficient ways of searching for translation equivalents in parallel texts is by searching for the same contextual features in both languages. In the English data, there is a strong link between *tax evasion* and *tax avoidance*. A similar link can be found in the Italian data, this time between the terms *evasione fiscale* and *frode (fiscale)*. Therefore *frode* 'fraud', although not one of the terms on the original list, became a word to study because it had a probable link with either *tax evasion* or *tax avoidance*, or indeed both. The concordances which contained these two terms were studied more closely, using the subcorpora to separate the examples by language and text type (see **Data**, above). The Directives were studied first as they form the smaller data set and are more uniform in style, a fact probably due to the necessity of absolute clarity in legislative documentation.

There are thirteen occurrences of *tax evasion* in the English Directives, all of which are followed by either *avoidance* or *tax avoidance* two words after the node (the n+2 position).

```
1  t certain types of tax evasion or avoidance. measures intended to
2  ereas practices of tax evasion and tax avoidance extending across
3  mbat international tax evasion and avoidance (3) É cil Whereas th
4  unter new forms of tax evasion and avoidance, which are increasin
5  as the practice of tax evasion and tax avoidance leads to budget
6  s certain forms of tax evasion or avoidance should be prevented W
7  s certain forms of tax evasion or avoidance should be prevented,
8   ks of distortion, tax evasion and tax avoidance resulting from t
9  sks of distortion, tax evasion and tax avoidance linked to the di
10 s as its objective tax evasion or avoidance or results in a compa
```

```
11 incipal objectives tax evasion or tax avoidance; the fact that on
12  the operation has tax evasion or tax avoidance as its principal
13 d certain forms of tax evasion or avoidance, it is necessary to c
```

The Italian Directives contain nine occurrences of *evasione fiscale*, which, as can be clearly seen from the concordance lines below, are all preceded by *frode* three words to the left (the n-3 position):

```
1  la pratica della frode e dell' evasione fiscale al di là dei conf
2  lla lotta contro la frode e l' evasione fiscale sul piano interna
3  alle nuove forme di frode e di evasione fiscale, che hanno sempre
4  la pratica della frode e dell' evasione fiscale conduce a perdite
5  are talune forme di frode o di evasione fiscale; considerando che
6  are talune forme di frode o di evasione fiscale, HA ADOTTATO LA P
7  a come obiettivo la frode o l' evasione fiscale o ha come consegu
8  ttivi principali la frode o l' evasione fiscale; il fatto che una
9  are talune forme di frode o di evasione fiscale, É opportuno prec
```

In fact, all nine of the Italian examples have their English equivalent above; lines 1, 8, 9 and 12 in the English data are not accounted for here. It is immediately striking that *evasione fiscale*, the expression thought to be the translation of *tax evasion*, is not at all what it seems, despite the fact that both of the dictionaries consulted were in full agreement; the evidence above shows that *evasione fiscale* means *tax avoidance*. In addition to this, the dictionary translation of *tax avoidance* is given as *elusione fiscale*, a term which does not occur at all the data. *Frode (fiscale)* is the term which is actually used to translate *tax evasion* in the European Directives, and this is backed up examination of the concordances of *tax avoidance*: all but one of the seven is preceded by (*tax*) *evasion* at the n-1 position, and interestingly enough, the odd one out reads as follows.

(…) of tax or to avoid fraud or tax avoidance. Whereas it might appear (…)

Tax avoidance is preceded at position n-2 by *fraud*, in what would seem to be a direct translation of the term *frode*, which above has been shown to be equivalent to *tax evasion*.

tax evasion / *frode*

In accordance with the procedural steps set out in the previous section, we must look again at possible translations of *tax evasion*, given that the initial translation has been shown to belong elsewhere. It is now clear that *frode* should be studied in more detail, remembering that it may not be the only translation of *tax evasion*, as indeed *tax evasion* may not be the only translation of *frode*.

In the Italian Directives there are 34 occurrences of *frode*, compared with 13 of *tax evasion*. We saw in the exampkes above that all 13 of these are translated with the word *frode*, and that at least one other instance of *frode* is taken into

account by the multi-word expression *...fraud or tax avoidance*. This leaves twenty concordances of *frode*, which may or may not be translated into English as *fraud*.

Even though the terms *fraud* and *frode* were not amongst those on the Law faculty list, the fact that they were translations for other terms was considered adequate justification for further study, especially as the primary motivation for carrying out this research was to obtain results which would be of use to students and professors, and not as a mere translation exercise. The translation of *tax evasion* is always *frode*, yet *frode* is not always translated as *tax evasion*. It would be helpful to attempt an explanation as to why this is the case.

Apart from *evasione fiscale*, there is another, stronger collocational pattern associated with *frode* in the Italian Directives, which is that with the words *evasione* 'avoidance' and *abuso* 'abuse', either as *evasione, frode o abuso* (4 times), or as *frode, evasion o/e abuse* (11 times). From the on-screen context it is not clear if this always refers to financial matters, although the following concordance line does specify this context.

...o sono comunque all'origine di una frode, di un'evasione o di un abuso fiscale...
'or are however at the heart of fiscal fraud, tax avoidance or abuse'

What does emerge, however, is that these three things (*fiscal fraud*, *tax avoidance* and *abuse*) have a semantic prosody (Louw 1993) of 'undesirability': eleven of the fifteen contain the verb *prevenire* 'prevent' before the collocation; others contain *evitare* 'avoid' and *combattere* 'combat'. A look at the extended context shows that almost all of the occurrences contain direct reference to finance, especially visible in the use of the terms *esenzioni* and *franchigie* (both 'exemptions'), and *disposizioni* 'allowances'. In fact, 29 of the 34 refer directly to taxation; the remaining five refer to the deceitful meaning of *frode* - the intention to mislead - although there is still a financial element involved, as is the case when the *frode* is associated with fraudulent insurance claims, as here.

... e di prevedere nel caso dei danni alle cose provocati da un veicolo non identificato, dati i rischi di frode, che l'indennizzo dei danni possa essere limitato o escluso...
'...and to ensure that in the case of damage caused by an un identified vehicle, given the risk of fraud, damages be limited or excluded...'

Of the 34 examples of *frode* present in the Italian Directives, the collocation *frode fiscale* occurs a total of ten times, including the *frode e evasione fiscale* examples presented above. However we have seen that *frode* is always found in financial contexts, even when it is not coupled with *fiscale*. This rather unfortunate *semantic preference* (Sinclair 1996) only seems to perpetuates the stereotype of the Italian as a reluctant tax payer. Ragazzini (1995) makes no reference to frode's preference for financial contexts, giving the translation as '*deceit; cheat, sham*", except when found as the collocate *frode fiscale*, the only collocate

example given (translated as *tax evasion*). Even de Franchis (1988) translates *frode* as "*fraud, deceit, deception*", again separating off *frode fiscale*, and translating it as *tax fraud*.

fraud / *frode*

We saw above that *fraud* was used as a substitute term for *tax evasion* in one of the concordances referring to *tax evasion and avoidance*. A closer examination of the term will show us if it can be used to translate the 20 instances of *frode* which remain after having taken *tax evasion* into account. In the English Directives, there are 30 occurrences of *fraud*, compared with 34 for *frode* in the Italian Directives. Looking at the numbers alone, 30 and 34, it seems quite plausible to view them as equivalents, but on closer examination of the concordances it can be seen that for only seven of the English examples is there a parallel concordances with *frode*. These amount to five of the six examples not covered by either *frode o l'evasione fiscale* or the collocation *evasione, frode* and *abuso*, plus two others.

```
1 quando sussistono seri dubbi di frode. Articolo 3 Condizioni relativi a
1 e there are grave suspicions of fraud. Article 3 Import conditions The

2 identificato, dati i rischi di frode, che l'indennizzo dei danni possa
2 luded in view of the danger of fraud; Whereas it is in the interest of

3 e l'impossibilità di qualsiasi frode; considerando che, in una prima ta
3  an adequate guarantee that no fraud is possible; Whereas, as a first s

4 e necessarie per combattere la frode nel settore dell'oro a decorrere d
4 l necessary measures to combat fraud in this area from 1 January 1993.

5 sistema che riduca i rischi di frode, indicando la natura esatta dei ma
5  a system reducing the risk of fraud by indicting the exact nature of t

6 li vigenti e evitare qualsiasi frode, evasione o abuso; considerando ch
6 nted properly and to avoid any fraud, evasion or abuse; Whereas it is n

7  articolo ed evitare qualsiasi frode, evasione o abuso. 10) All'articol
7 o properly, and to prevent any fraud, evasion or abuse, Member States s
```

It was mentioned above that one of the most important collocates of *frode* in the Italian Directives was *prevenire* 'prevent'. In the English data, the most frequently-occurring collocates of *fraud* are *protect against fraud* and *prevention of fraud*. This similarity is important, but not conclusive, as the English data makes very little reference to financial *fraud*, more usually indicating deceitful actions and intent to defraud. This suggests that *fraud* and *frode* are false friends almost to the same extent as *tax evasion* and *evasione fiscale*.

CASE STUDY part 2 - JUDGEMENTS

tax avoidance / *evasione fiscale*

The results obtained from the analysis of the Directives were compared to the Judgements, to see whether the same patterns were present and if these were translated in the same way. The Judgements data set from was larger, and was expected to be less uniform in content because its function is not to impose legislation, but to document how legislation has been used in practice. *Tax avoidance* and *evasione fiscale* were analysed first of all because the analysis of the Directives data had successfully identified these as equivalents. The number of in the Judgements was very similar - 15 *tax avoidance* and 17 *evasione fiscale*. It was expected that the translation would be confirmed in this second data set, but in actual fact only 6 occurrences - less than half - turned out to be matching pairs. However there was a considerable degree of repetition of some of the multi-word units identified in the Directives, especially when inflected forms are taken into account. For example, all 17 plural occurrences of *evasioni fiscali* feature the collocation with *frodi* (as Italian has no uncountable nouns, these plurals probably correspond to *tax avoidance* in English, and not to **tax avoidances* or **avoidance of taxes*). The recurrence of these extended collocations is positive in that it validates the collocational approach to analysis, highlighting the fact that structures are repeatedly used. It does not help to explain the inconsistencies in the translations, because these units ought to have standard equivalents, and yet this is simply not the case.

The English word *avoidance* (without the modifier tax) has an overwhelmingly strong semantic preference for the financial world, and seems to be used instead of the longer form *tax avoidance* when the context clearly indicates finance or taxation. This same principle applies to Italian, as the ellipsed (singular and plural) forms evasione and *evasioni* are used when the financial context has already been specified in the text, whereas *evasione fiscale* is stated in full when there is room for doubt. Around 70% of the occurrences of *tax avoidance* were found, and all were translated as *evasione fiscale*. Importantly, not one example was found of the term *elusione fiscale*, which is the standard dictionary translation.

Where translations are not forthcoming, other mechanisms come into play. The Judgements often contain explanatory material to aid in the translation process, or comment on it. Even though *tax avoidance* and *evasione fiscale* were not often found to correspond in the Judgements, there is evidence in the data itself that this is indeed the correct translation:

Esse sostengono anzitutto che l'autorizzazione non è stata chiesta per evitare "evasioni" fiscali, bensì frodi fiscali. Esse rilevano nella comunicazione inviata dal Regno Unito alla Commissione il 15 marzo 1985, ai sensi del n.2 di detto art. 27, che non si tratterebbe di "tax avoidance", cioè di "evasione fiscale", ma di "tax evasion" cioè frode fiscale. Esse aggiungono che la versione   inglese   della

Gazzetta ufficiale del 31 luglio 1985, che pubblica la decisione del Consiglio di autorizzare il provvedimento in deroga chiesto dal Regno unito, farebbe espresso riferimento alla "tax evasion", cioè alla frode fiscale.

'They maintain above all that authorisation was not requested to avoid tax 'avoidance', but tax fraud. They point out that in the communication sent by the United Kingdom to the Commission on 15 March 1985, in accordance with the aforementioned article 27, paragraph 2, that it would not be for 'tax avoidance', i.e. evasione fiscale, but for 'tax evasion', i.e. frode fiscale. They add that the English version of the Gazzetta Ufficiale of 31 July 1985, which published the Council's decision to authorise the derogation requested by the United Kingdom, made express reference to 'tax evasion', i.e. to frode fiscale.'

It should be pointed out that although this excerpt has a counterpart in the English data, it does not contain the same explanatory information, which is clearly necessary in Italian to avoid misinterpretation, especially given the fact that translators' dictionaries translate the terms differently. It would appear that the Judgements have been more loosely translated than the Directives (though they are certainly not free translations) and this may go some way in accounting for the difficulties encountered in searching for translations in this subcorpus. EURODICAUTOM was unable to offer much more help. It did suggest some alternative terms, but they were either not present in our data, or not appropriate in the context. Although is it surprising that the Commission's own translator's aid was so limited, this is probably due to the fact that it is compiled from existing specialised and non-specialised dictionaries which, as we have already seen, are not as reliable as is generally believed.

tax evasion / *frode fiscale*

It was established from the Directives data that the term *evasione fiscale* is not translated as *tax evasion* as the dictionary claims, but that the preferred term is *frode*, or much less frequently, *frode fiscale*. The collocations of *tax evasion* are the same as those found in the Directives, despite the fact that the Judgement texts favour linguistic variation rather than the formulaic structures typical of the Directives. Given that the singular *frode (fiscale)* can also be found in the plural *frodi (fiscali)*, this too must be examined. We have already seen that *frode* does not refer exclusively to *tax evasion*, and also that it is not the best translation of *fraud*, which has a semantic preference for deceit in general rather than financial fraudulence.

The occurences of *tax evasion*, *fraud*, *frode* (singular) and *frodi* (plural) in the Judgements are set out for clarity in the table below:

English	Italian
tax evasion 48	frode 126
fraud 212	frodi 181

Table 1: Frequencies of terms in English and Italian

Examination of the data makes it clear that *fraud* is not used as a synonym for *tax evasion*, as was suggested in some of the Directives data. In fact, not once do the words *tax*, *evasion* or *avoidance* occur in the surrounding context. The important collocates of *fraud* find their counterparts in *frode* and *frodi*, but are fairly evenly divided between the singular and plural forms. We find that the singular form is preferred when talking of *the risk(s) of fraud* (14 out of the 19 occurrences appear as *rischi(o) di frode*), but it is more important to prevent *frodi* (26 out of 39) just as it was in the directives data. *Irregolarità* 'irregularities' is more common with the plural form *frodi* (nine out of thirteen), and *diritti* 'rights' appears fairly evenly between the two (16 with the singular and 18 with the plural), though no definitive statement can be made about factors affecting this choice. Similarly, the choice of singular and plural when translating *tax evasion* into Italian seems to lie with the translator's personal preference, and although the collocation with *evasione* (tax avoidance) occurs seven times with the singular *frode* and only twice with the plural, there is no obvious preference in the translation of *tax evasion* when *(tax) avoidance* is not present (ten are singular and nine are plural). In the multi-word combination *frodi, evasioni e/o abusi* 'evasion, avoidance and/or abuse', seventeen occur with all terms in the singular, and twelve in the plural, again suggesting that there is no distinct preference, even though the plural form probably be more appropriate (English uncount nouns are usually translated into Italian as plurals).

Although the collocation *frode fiscale*, the dictionary translation of *tax evasion*, was present both in the Directives and the Judgements, it amounted to less than five percent of all the occurrences of *frode*. This is despite the fact that *frode* almost always occurred in contexts of finance and taxation. This again questions the logic of dictionaries whose Italian translation of *tax evasion* is always given as the full *frode fiscale*, and who never indicate that the standalone term *frode*, if found in a financial context, will refer to the same concept.

CONCLUSION

The pilot project was successful in so far as it has shown that translation equivalents can be found in parallel texts even when an aligner cannot be used. The methodology outlined above is undeniably time-consuming and laborious yet it is accurate, a claim that the reference works consulted in the course of the project are unable to make. There are however some important points to consider before consigning these to the bottom of our bookcases: the language studied in this project was all European legislation, well known for its bizarre mixture of coined expressions, translationese and stolid bureaucracy. Such language not only bears little resemblance to the spoken and written languages of

the European Community, but is also distinct from the styles adopted in the national legislation of the various Member States. It is therefore possible that the inaccuracies of the dictionaries and glossaries consulted are only apparent, and that these works simply need updating to take account of this relatively new restricted language. At the same time we must acknowledge the possibility that the findings presented here are conclusive, based as they are on the exhaustive analysis of texts and their actual translations. It is also very difficult to dismiss as chance the huge discrepancies between the data and the reference works consulted, which never offered more than two-thirds correspondence, and inclined towards fifty percent or less, if indeed they suggested an equivalent term. All too often the translations offered were erroneous, non-existent in the data (therefore not verifiable), or they were simply not included in the examples presented. This has been bourn out in the analysis and translation of around one hundred terms, and is in no way restricted to the case studies discussed.

The only one way in which the doubts raised here can be resolved is in the continuation of this project and others like it. English and Italian national legislation will be compared in the BoLC project proper (research in progress), but this is only a drop in the ocean. Corpus-based translation work can only be to the benefit of those who use language professionally. In questioning previously unchallenged statements about language and translation, it provides the impetus for a long-overdue updating of reference works. In a specialist area such as the Law, this is not merely an ideal, but a necessity both for translation and the study of foreign-language texts, because mistranslations have the potential to seriously jeopardise the course of justice.

REFERENCES

Baker, M. (1993) 'Corpus Linguistics and Translation Studies - Implications and Applications', in Baker, M., G. Francis and E. Tognini-Bonelli (eds) (1993) *Text and Technology. In Honour of John Sinclair*, Philadelphia and Amsterdam: John Benjamins, 233-259.

Clear, J. (1993) 'From Firth Principles - Computational Tools for the Study of Collocation' in Baker, M., G. Francis and E. Tognini-Bonelli (eds), 271-292.

Epstein, R. (1988) *West's Law and Commercial Dictionary in Five Languages*. Italian Edition, Bologna: Zanichelli / West.

EURODICAUTOM available on the internet at <http://www.echo.lu/edic>

Firth, J. R. (1957) 'Personality and Language in Society' in *Papers in Linguistics 1934-1951*, London: Oxford University Press, 177-189.

Fillmore, C. J. (1992) '"Corpus linguistics" or "Computer-aided armchair linguistics"' in J. Svartvik (ed), *Directions in Corpus Linguistics*. Proceedings of Nobel Symposium 82, Berlin and New York: Mouton de Gruyter, 35-60.

Franchis, F. de (1988) *Dizionario Guiridico / Law Dictionary*,

Louw, W. E. (1993) 'Irony in the Text or Insincerity in the Writer? - The Diagnostic Potential of Semantic Prosodies' in Baker, M., G. Francis and E. Tognini-Bonelli (eds), 157-176.

Mason, O. (1996) 'Corpus Access Software: The CUE System', *TEXT*

Technology, 6 (4): 257-266.

Mason, O. (1997) 'CUE - A Software System for Corpus Analysis' in *Proceedings of the Second TELRI Seminar*, Kaunas, 85-90.

Ragazzini, G. (1995) *Dizionario Inglese-Italiano, Italiano-Inglese*, Bologna: Zanichelli.

Sinclair, J (1991) *Corpus, Concordance, Collocation*, Oxford: Oxford University Press.

Smith, S. A. (1995) 'Culture Clash: Anglo-American Case Law and German Civil Law in Translation', in M. Morris (ed), *Translation and the Law*, American Translators Association Scholarly Monograph Series, vol. VIII, Amsterdam and Philadelphia: John Benjamins, 179-197.

Sinclair, J (1996) 'The Search for Units of Meaning', *Textus*, IX: 75-106.

Tognini-Bonelli (1996) *Corpus Theory and Practice*, TWC Monographs, Birmingham: TWC.

Tognini-Bonelli (forthcoming) 'Functionally Complete Units of Meaning across English and Italian: towards a Corpus-driven Approach', in S. Granger and B. Altenberg (eds), *Lexis in Contrast*. Studies in Corpus Linguistics, Amsterdam and Philadelphia: John Benjamins.