R. Rossini Favretti, Bologna

Using Multilingual Parallel Corpora For The Analysis of Legal Language: The Bononia Legal Corpus

Abstract

This paper intends to describe our experience in building multilingual comparable law corpora here at the University of Bologna. It is the first report of an ongoing corpus-driven research project named Bononia Legal Corpus, or BOLC, which is involved in investigating legal language in English and in Italian.

In the construction of the corpus, two subcorpora have been envisaged: one English, the other Italian. According to our plans, carrying out the project will take three years: 1997-1999. As to the dimension of the corpus, we set 10 million words as the smallest target for each component.

Before the first sample, a preliminary pilot corpus (6 mw for each subcorpus) was stored.

Keywords: corpus linguistics, multilingual parallel corpora, legal language.

1. Introduction

This paper intends to describe our experience in building multilingual comparable law corpora here at the University of Bologna. It is the first report of an ongoing corpus-driven research project named Bononia Legal Corpus, or BOLC, which is involved in investigating legal language in English and in Italian.

BOLC is an interdisciplinary project. It is based on the cooperation of experts in linguistics, computer programming, and law. The presence of legal experts in the team has proved particularly important in drawing up specimen bibliographies, in the selection and classification of texts (on cultural evidence), in transmuting parallel corpora into comparable corpora, and in the selection of key concepts and key words. The presence of John Sinclair as a consultant has also been of critical importance.

---

1 As computer experts I wish to mention F. Tamburini and E. Martelli for their valuable work in formatting corpus documents and writing the software which is used in our research.

2 As a legal expert I wish to mention in particular Prof. A. Di Pietro.
In the construction of the corpus, two subcorpora have been envisaged: one English, the other Italian. According to our plans, carrying out the project will take three years: 1997-1999. As to the dimension of the corpus, we set 10 million words as the smallest target for each component, that is:

<table>
<thead>
<tr>
<th>Sample</th>
<th>10 m.w.</th>
<th>10 m.w.</th>
</tr>
</thead>
<tbody>
<tr>
<td>first</td>
<td>English</td>
<td>Italian</td>
</tr>
<tr>
<td>subsequent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>each</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The samples will be restricted to readily available material; 90% or more is probably already in electronic form.

Before the first sample, a preliminary pilot corpus (6 m.w. for each subcorpus) was stored.

2. A legal corpus

Why a legal corpus? For the relevance that collocation has in legal texts. It has been argued that when we think of law, we usually think of first of words. This is an important point since most juridical words do not have a direct connection with counterparts in the actual world. They cannot be defined in terms of ordinary factual counterparts; concordance and collocation, therefore, play a fundamental role in the definition of meaning.

English and Italian legal texts are representative of two different legal systems and of the differences existing between the common law system developed in England and the civil law system, based on the Roman law, developed in Italy.

A further point is to be made as to the representativeness of the corpus design. In the last few years, the situation has been made more complex with the development of European Community law, that body of law created by and consequent to the association of European countries. European Community law not only applies within the Community as an organization, but it also binds member states together, and creates rights and duties for legal persons therein.

In our project, European legislation is seen as a reference for the contacts taking place there among the different legal systems, and for the transnational approach which is implied in the coexistence and cooperation of different nationalities. In the building up of our corpus, we chose to refer to European legislation, in a preliminary pilot corpus, in order to create a conceptual framework to be used as a first-level reference. We chose to refer to the secondary Community legislation even if reference can also be made to primary legislation, that is to the Treaties that established the communities.

Secondary legislation consists of regulations, decisions, directives, recommendations and opinions made or given by the Commission. In particular, our focus was on directives since they can be implemented by domestic legislation and can produce direct legal effects in member states. At a further stage, directives make possible the identification and selection of legal texts which have been enacted "on specific issues," in different European countries. They make it possible to transmute parallel corpora into comparable corpora.

Data are now limited to English and Italian. An extension which will include French and Spanish is envisaged.

A short description of the documents. A directive usually begins with an extensive set of header information, indicating the date, the reference number and its coding conventions. The text follows. After the proper text there is possibly some trailer information. The text data have been divided from the header and the trailer data. Two documents have been distinguished.

The first document (data file) contains the legal text and the reference number (CELEX).

The second document (reference file) contains all other data that could be useful in performing law-targeted rather than language-targeted searches.

During the first phase, while building the pilot corpus, the corpus was not tagged in order to simplify the entire building process. In the future, the tagging process could be performed with some differences between subcorpora, in consideration of the specific requirements presented by Italian, as, for example with accents and apostrophes.

3. A case study: tax and duty

In the analysis of the pilot corpus some important methodological aspects emerged. I will comment briefly on one of the most significant collocations with regard to the problems presented in interlinguistic comparison. As a case study, I will consider the English word: tax. It refers to a situation which can be considered common both to the U.K. and to Italy, and can be assumed to apply to other European countries as well, and whose presence can be accounted for with the extension of our corpus. In all countries, taxes are levied on income and expenditure by central and local governments, but different

3 "Collocation is the occurrence of two or more words within a short space of each other in a text. The usual measure of proximity is maximum of four words intervening. Collocations may be dramatic and interesting because unexpected, or they can be important in the lexical structure of the language because of being frequently repeated." J. SINCLAIR, 1991, p. 170.
categories are employed in how taxes are defined. It was our hypothesis that the categories would emerge from the analysis of the data.

Let us consider the following concordance of the word tax, selected automatically from our corpus, where there are 1204 citations altogether:

Article 3.1. The term 'withholding tax' as used in this Directive shall:
- the calculation of a toll, tariff, tax, fee, charge, premium, tax, or any other contribution in connection with the furnishing of a service or the performance of a work or with the purchase or sale of goods or services, including the use of the goods or services, by the party who is required to pay the tax or charge, or by the party who is responsible for the payment of the tax or charge.
- the calculation of a tax, fee, charge, premium, tax, or any other contribution in connection with the furnishing of a service or the performance of a work or with the purchase or sale of goods or services, including the use of the goods or services, by the party who is required to pay the tax or charge, or by the party who is responsible for the payment of the tax or charge.
- in particular, in cases where:
  - taxes on goods and services, including the use of the goods or services, by the party who is required to pay the tax or charge, or by the party who is responsible for the payment of the tax or charge.
  - taxes on goods and services, including the use of the goods or services, by the party who is required to pay the tax or charge, or by the party who is responsible for the payment of the tax or charge.
- in particular, in cases where:
  - taxes on goods and services, including the use of the goods or services, by the party who is required to pay the tax or charge, or by the party who is responsible for the payment of the tax or charge.
- in particular, in cases where:
  - taxes on goods and services, including the use of the goods or services, by the party who is required to pay the tax or charge, or by the party who is responsible for the payment of the tax or charge.

The first position to the left of tax is occupied:
- by a noun in 102 instances, which is roughly 9%:
  - turnover tax (68)
  - income tax (10)
  - consumption tax (6)
  - export tax
  - vehicle tax
- by a noun-group in 366 cases:
  - value added tax (264)
- by an -ing form in 23 cases:
  - a withholding tax (11)

In a number of cases (94), tax is qualified by a prepositional phrase:
- tax on exports/exportation (14)
- tax on profit or loss (10)
- tax on services (9)
- tax on sales (1)

Through postnominal qualifications, the objects and services upon which taxes can be imposed are generally specified. As we can see in the examples, specification can take place both in postmodification and in prenominalisation: tax on export or export tax.

It is interesting to observe that adjectives occur as modifiers in a limited number of instances, 19 examples out of 1204, which is lower than 2%.

In the examples where the word tax is not associated with modification or qualification, the first position on the left is occupied:
- by a preposition:
  - after tax (18)
  - before tax (3)
  - subject to tax (11)
  - liable to tax
  - the amount of tax (10)
duty and tax
a toll, tariff, tax, excise duty and other charges

Tax and duty are important collocates. Their co-occurrence is strong also when there is a modifier at N-1 position and when tax is in the plural form:

- turnover tax and excise duty
- turnover tax, excise duty/duties and other consumption taxes

Another point can be observed. Adjectives rarely occur at N-1 position when the word is in the singular form; they dominate the pattern, however, when it is in the plural form. We find classifying adjectives such as internal, national, local, direct, direct, specific, various to specify the class the tax belongs to. The examples suggest that tax, in its singular form, has three different meanings:

1) a general, indefinite meaning in the first instances, when followed by a noun and used as a modifier,
2) a general collective meaning, in the second group of instances, when there is no modifier at N-1 position, either on the left or on the right,
3) a specific one, when there is a modifier at N-1 position, on the left.

There is a hyponymic relation between 3 and 2, which can be exemplified by such pairs as turnover tax and tax.

A quick scan through the concordance of tax in its plural form shows that the collocates present a limited overlapping between singular and plural forms:

- turnover taxes
- national taxes
- domestic taxes
- internal taxes
- specific taxes
- direct taxes
- export taxes
- import taxes
- various taxes
- local taxes
- all/those/these taxes

Adjectives dominate the pattern and the specific value prevails. They are used to mention specific taxes, referred to in their unspecified amount. In the singular form the modification tends to refer to goods and services on which taxes can be imposed, while in the plural form, reference is made to a more general distinction, drawn on conceptual lines.

Both in the singular and plural form, duty is a significant collocate:

- any taxes, duties or charges
- duties or/and taxes
- taxes, duties and levies
- vat, taxes, duties or charges
- national taxes and charges

Duty collocates with tax, but the lexical environment of the two words is different. It is clear from a superficial glance that the first position to the left of duty is also occupied by a noun in a high number of instances. But their most prominent collocates do not overlap, as the concordance below illustrates:
Excise or customs does not collocate with tax, nor does turnover collocate with duty. On closer inspection of the collocations duty and tax, we see that in the first group of examples, the noun groups, where tax and taxes occur, reference is made to direct taxation, while in the second group of instances, where duty and duties occur, reference is made to indirect taxation.

...
A further observation is to be made. Position N-1 to the left of *imposta* is generally occupied by definite article and the noun is generally modified on the right. N-1 and N-2 positions on the right are generally occupied by post-modification. The different uses of the article and the different structure of the noun group play an important role in a concordance which cannot be overlooked, and which will be the object of further analysis.

If we examine the corpus and focus on the instances where the enumerations occur, we find *taxes, duties and charges* mentioned as different types of contributions in English. The word *charges* tends to occur as a general, inclusive term, at the end of the enumeration; the most prominent collocates are *duty* and *tax*. It is interesting to compare *duty with tassa*, which one might expect to be its equivalent. But we see that the occurrences of *tassa* are definitely lower. There are 31 citations of *tassa*; every third one is reproduced here:

If we consider the collocates, we find that the word *tassa* is modified by adjectives, such as *automobilistica*, and noun phrases and prepositional phrases, such as *di circolazione*, *sulla circolazione*, *di attrezzaggio*, *di soggetto*. The reference to direct and indirect taxation is not made in the distinction drawn in Italy between *imposta* and *tassa*. Different conceptual categories are applied in English and in Italian. *Tassa automobilistica*, which finds its equivalents in the corpus data both in *vehicle tax* and in *vehicle duty*, is something paid for a consideration of value. A payment is due in return for services. An outstanding feature of Italian tax law is the distinction made with regard to contributions levied on a person with or without regard to personal services or advantages conferred on that person by law. The word *tassa* occurs when the payment is meant as a counterpart of personal or general services.

In the plural form, the word *tasse* presents a more general value, which includes unspecified taxes and charges of many kinds, and finds its equivalent in *charges*: Tasse di effetto equivalente charges having equivalent effect

By the words *tax, duty and tassa, imposta* different kinds of contributions are identified. Even limiting our analysis to positions N-1, N-2, N-3, we can discover relevant differences in the prominent collocates, and distinctions can be made between the different law systems. This brings to the fore another notable feature of our research. Legal language is deeply rooted in the institutional and historical traditions of a country. Meanings are bound to institutional practices. Further distinctions emerge and our analysis is to be extended to include not only tax, *duty*, and *tax*, but also other words such as *charge, rate, fee*, etc., highlighting both their collocation and the equivalence that can be identified for translation. It should also include a closer examination of the cases in which difficulties may arise from the lack of syntactic correspondence between English compounds such as *tax-payer or chargeable event*, and Italian nouns and noun groups such as *contributore o fatto generatore dell'imposta*. This is a further step, which can be made using corpora, and through practices which are not so different from the ones used in bringing forth and highlighting the equivalence between *tax and tassa* – an equivalence, it should be noted, which is not mentioned in any dictionary.
References


Feng Zhiwei, Beijing

A study of translation equivalence in the Chinese-English text of Plato's Republic

Abstract

Based on the English-Chinese parallel corpus of Plato's Republic, we study the translation equivalence in the Chinese-English text of the corpus. Our study concentrates on the first chapter "Convention under Attack" and is focusing on three levels: paragraph level, sentence level and word level. The paper describes the agreement and disagreement between English and Chinese in these three levels. Finally an experiment of word alignment based on the corpora is introduced.

1. English-Chinese Parallel Corpus of Plato's Republic

Plato's Republic is a very famous and important philosophical monography. It is a sprawling work, and it also sprawled to China. In 1957, a Chinese translation from Greek was published, the translator is Wu Xianshu. Wu's translation was rather archaic and abstruse. In 1986, another Chinese translation version (also translated directly from Greek) was published, the translators are Guo Binhe and Zhang Zhuning, both being professors of the history of philosophy. Guo/Zhang's version is easy and smooth, it is very popular in China and it has attracted a lot of Chinese readers. We decided to take Guo/Zhang's version as our study material. It was then manually put into electronic format as a corpus of Plato's Republic in Chinese translation. In 1996, Professor John Sinclair sent me a diskette containing a machine-readable corpus of the English translation of Plato's Republic, translated by Benjamin Jowett, and published in 1901 by Colonial Press. At the same time, Dr. Wolfgang Teubert sent me the book of this English version. With the help of Dr. Wolfgang Teubert and Professor John Sinclair, we set up a Chinese-English parallel corpus of Plato's Republic, where both versions were translated from the same language – Greek. The volume of the Chinese version is 443 136 bytes, and the volume of English version is 653 318 bytes.

Based on this parallel corpus, we started to study the translation equivalence in the Chinese-English text of Plato's Republic. Our study was concentrated to the first chapter "Convention under Attack". For the first chapter, the Chinese volume is 43 648 bytes, and the English volume is 67 200 bytes.