A dynamic model for reference corpora structure definition

Fabio Tamburini

CILTA - University of Bologna
Piazza S.Giovanni in Monte, 4, I-40124, Bologna, Italy
f.tamburini@cilta.unibo.it

Abstract

A representative corpus of written Italian – CORIS – constructed at the Centre for Theoretical and Applied Linguistics of Bologna University (CILTA) is available on-line. Considering the importance of the comparability of reference corpora in interlinguistic studies, a further corpus – CODIS – was designed. Aimed at specialist needs, CODIS presents a dynamic and adaptive structure providing for the selection of the subcorpora pertinent to a specific research project and allowing the researcher to define the size of each subcorpus. CODIS is designed to be dynamically adapted by the scholar to different comparative needs by a careful combination of small corpus chunks of various types and sizes. The chunk sizes were carefully selected in order to allow for various combinations creating subcorpora of different sizes, ranging from 0 to the maximum size of each CORIS subcorpus. This fine granularity provides a wide range of corpora composition options, satisfying almost all comparative needs.

1. Introduction

A synchronic corpus of written Italian – CORIS – was constructed at the Centre for Theoretical and Applied Linguistics of Bologna University (CILTA) in the period 1998-2001 and is now available on-line. The project aimed at creating a representative general reference corpus of contemporary written Italian designed to be easily accessible and user-friendly. CORIS contains 100 million words and will be updated every two years by means of a monitor corpus. It consists of a collection of authentic texts in electronic form chosen by virtue of their representativeness of written Italian. It is aimed at a broad spectrum of potential users, from Italian language scholars to Italian and foreign students engaged in linguistic analysis based on authentic data and, in a wider perspective, all those interested in intra- and/or interlinguistic analysis.

The expansion of the EU, as well as the globalisation of markets, has given rise to the need for accurate multilingual studies. In the domain of interlinguistic analysis, corpora comparability plays a vital role. Following the taxonomy of corpora proposed by Sinclair (1996) or Teubert (1997), and considering the definition given by some scholars, such as Rayson & Garside (2000),

"Comparability is of interest too, since the corpora should have been sampled for in the same way. In other words, the corpora should have been built using the same stratified sampling method and with, if possible, randomised method of sample selection", or Teubert (1996),

"Comparable corpora are corpora in two or more languages with the same or similar composition.",

it appears to be quite clear that corpora used for multilingual studies have to be based on common choices in terms of textual varieties, composition and especially the size of the different subcorpora. Examining the structure of a number of reference corpora in different languages, a great variety of corpora may be seen. Table 1 shows some reference corpora and the composition of their written sections, in terms of textual varieties and their respective proportions. The choices underlying the construction of each corpus are quite different, leading to corpus structures which differ in terms of subcorpora composition, especially with regard to the textual varieties, and the proportions, in terms of number of words, between the different parts of the corpora. These structural differences between corpora make any comparative study quite complex, potentially introducing biasing effects in the results obtained.

Considering the crucial role to be played by the comparability of a reference corpus, it was decided to provide the option of creating an alternative corpus structure, making it adaptable to the needs of different researchers. This flexible structure, by allowing the researcher to dynamically redesign the corpus composition, should prevent the biasing effects introduced by a direct comparison of the results obtainable using the original, unmodified corpus structures. CORIS aims to become the reference corpus for Italian, providing a dynamic corpus structure, and allowing for the wide comparability that every reference corpus should permit.

To improve comparability in CORIS, a further corpus – CODIS – was designed. Aimed at specialist needs arising in interlinguistic analysis, CODIS presents a dynamic and adaptive structure that allows direct comparison with almost any other reference corpus.

Before describing the technical issues concerned with the dynamic structure of CODIS, it is necessary to consider corpus representativeness. Before designing CORIS, a preliminary study was conducted to design a corpus structure that satisfied all the representativeness criteria that a reference corpus should have, as outlined in (Biber, 1993; Váradi, 2001). This study, described in detail in Rossini Favretti (2000) and Rossini Favretti et al. (forthcoming), considered various parameters connected with textual varieties, circulation, text permanence, sampling methods, etc. An in-depth study of quantitative and qualitative parameters resulted in the design of a well-balanced corpus in its subcorpora definitions and sizes, based on a representative sample of modern Italian. The
overall CORIS structure is outlined in Table 1. By allowing the modification of the fundamental parameters defining the corpus structure, we make it possible to carry out comparative studies between languages, but we lose the original work done to define corpus representativeness. Thus, scholars using CODIS for designing an Italian corpus for their comparative needs have to be prepared to justify their choices, because our original design criteria and representativeness may no longer be valid, thus diminishing the reliability of the newly generated corpus.

| Corpus                                           | Composition             |
|--------------------------------------------------|-------------------------|
| CORIS – 100Mw – Italian                          | Press 38 Mw - 38%       |
|                                                  | Fiction 25 Mw - 25%     |
|                                                  | Academic prose 12 Mw - 12% |
|                                                  | Legal and admin. prose 10 Mw - 10% |
|                                                  | Miscellanea 10 Mw - 10% |
|                                                  | Ephemera 5 Mw - 5%      |
| BNC – 90Mw – English                             | Books 52.5 Mw - 58.6%  |
| Written section                                  | Press 27.8 Mw - 31.0%  |
|                                                  | Miscellanea 7.4 Mw - 8.3% |
| Bank of English – 385 Mw - English               | Newspapers 188.5 Mw - 48.9% |
| Written section                                  | Magazines 98.5 Mw - 25.6% |
|                                                  | Books 75.7 Mw - 19.7%   |
|                                                  | Academic Prose 14.2 Mw - 3.7% |
|                                                  | Ephemera 8.1 Mw - 2.1%  |
| LSWE – 28Mw – English                            | News 10.6 Mw - 37.7%   |
| Written section                                  | General Prose 6.9 Mw - 24.6% |
|                                                  | Academic Prose 5.3 Mw - 19.0% |
|                                                  | Fiction 5 Mw - 17.8%    |
| Corpus de Referencia del Español Actual (CREA) – 122Mw – Spanish | Press 59.8 Mw - 49%   |
| Written section                                  | Books 59.8 Mw - 49%     |
|                                                  | Ephemera 2.5 Mw - 2%    |
| INL 38 Million Words Corpus 1996 – 38 Mw – Dutch | Newspapers 12.4 Mw - 32.7% |
| Written section                                  | Legal texts 12.9 Mw - 33.9% |
|                                                  | Varied composition 12.7 Mw - 33.4% |
| The Oslo Corpus – 22.3 Mw – Norwegian            | Newspapers/Magazines 10.6 Mw - 47.5% |
|                                                  | Fiction 3.8 Mw - 17.0%  |
|                                                  | Factual prose 7.8 Mw - 35.0% |
| Corpus de Referência do Português Contemporâneo (CRPC) – 92 Mw – Portuguese | Newspapers 55 Mw - 60.8% |
| Written section                                  | Books 20.5 Mw - 22.6%  |
|                                                  | Periodicals 7 Mw - 7.7% |
|                                                  | Dec. of Sup. Court of Just. 1.8 Mw - 2.0% |
|                                                  | Miscellanea 3.9 Mw - 4.3% |
|                                                  | Leaflets 0.3 Mw - 0.3%  |
|                                                  | Correspondence 0.1 Mw - 0.1% |
| Croatian National Corpus – 30Mw – Croatian       | Newspapers 11.4 Mw - 38% |
|                                                  | Magazines 5.4 Mw - 18%  |
|                                                  | Textbooks 6.0 Mw - 20%  |
|                                                  | Prose 6.6 Mw - 22%      |
|                                                  | Imaginat.-factografic texts 0.3 Mw - 1% |
|                                                  | Essays, (speeches), etc. 0.3 Mw - 1% |
| Czech National Corpus (SYN2000) – 100Mw – Czech   | Journalism 60.0 Mw - 60% |
|                                                  | Informative Prose (arts, 25.0 Mw - 25% |
|                                                  | social science, law,     |
|                                                  | technology, adminstration,|
|                                                  | economics,…)            |
|                                                  | Imaginative Prose (poetry, 15.0 Mw - 15% |
|                                                  | drama, fiction,…)        |
| The Bank of Swedish – 50 Mw – Swedish            | Newspapers 40.0 Mw - 80% |
|                                                  | Novels 10.0 Mw - 20%     |

Table 1: The composition of some reference corpora. Not all of them claim to be reference corpora, but each of them is the most important corpus for the respective language, or the only one available.
2. CODIS dynamic structure

As seen above, multilingual research projects aimed at comparing linguistic evidence in different languages using corpora have to face problems of comparability among the corpora. To obtain consistent results, it is common practice to use corpora with roughly the same subcorpus composition, but the heterogeneous structures of the different reference corpora currently available often do not allow such comparative studies.

To overcome this problem, we designed a new corpus - CODIS - that allows for the selection of the subcorpora pertinent to a specific research project and also the size of every single sub-corpus. CODIS is designed to be dynamically adapted by the scholar to different comparative needs.

The key idea is to split each subcorpus into a set of smaller chunks and to combine them to obtain a specific subcorpus size. If the chunk sizes are carefully selected using a power2 rule, it will be possible to obtain any subcorpus size ranging from 0 to the maximum size. To avoid generating an excessive number of chunks, a minimum chunk size must be chosen, corresponding to the minimum resolution of the adaptability process. Considering that a modern reference corpus consists of at least 30 million words, 1 million words seemed a reasonable choice for the minimum granularity of the adaptation process. Having set this key parameter, it is possible to split every subcorpus using a power2 rule. Splitting the total size of the subcorpus using chunks of any size ranging from 0 to the original subcorpus size, having a resolution of 1 million words.

Unfortunately, due to the different original sizes of each CORIS subcorpus, the number of chunks resulting from the splitting of the different subcorpora will differ, giving rise to the problem of choosing the complete CORIS structure for individual needs. Thus, for reasons of simplicity, it was decided to apply a linear sampling of the documents as described in figure 1, that takes as an example the subcorpus Miscellanea.

Figure 2 shows the web interface designed for CODIS queries, similar to the one designed for CORIS. The main difference is in the "Subcorpora selection" box: by selecting the corresponding checkbox, the researcher can combine the various chunks into which the CORIS subcorpus is split to build the required corpus structure, in terms of subcorpus selection, but, more importantly, in terms of their overall sizes. In the example shown in Figure 2, the corpus used for the query consist of 30Mw of Press documents, 20Mw of Fiction, 11Mw of Academic Prose, 3Mw of Legal and Administrative Prose, 3 Mw of Miscellanea and 1 Mw of Ephemera.

3. Conclusions

This paper presented CODIS, a corpus based on CORIS, a synchronic reference corpus for written Italian, designed to be dynamically adapted to different comparative needs. Using the web interface, researchers can generate a complete Italian corpus to satisfy their needs, by selecting CORIS chunks of various types and sizes. This fine granularity creates an extremely flexible corpus structure that can be adapted to almost any comparison with other reference corpora in different languages. CODIS, as well CORIS, are available free on the web for research purposes. For further information refer to the CILTA website, http://www.cilta.unibo.it.
Figure 2: The CODIS query web interface: the checkboxes allow for the construction of a wide variety of subcorpora, starting from the original CORIS structure, as outlined in table 1 and table 2.

4. References
Atkins, S., Clear, J. & Ostler, N. (1992). Corpus design criteria. Literary and Linguistic Computing, Vol. 7, No. 1: 1-16.
Biber, D. (1993). Representativeness in corpus design. In Literary and Linguistic Computing, Vol. 8, 4: 243-257
Rayson, P. & Garside, R. (2000). Comparing corpora using frequency profiling. In Proceedings of the workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics - ACL2000 (pp. 1-6). Hong Kong.
Rossini Favretti, R. (2000). Progettazione e costruzione di un corpus di italiano scritto – CORIS/CODIS, In Rossini Favretti, R. (Ed.), Linguistica e informatica: corpora, multimedialità, percorsi di apprendimento (pp. 39-56). Roma: Bulzoni.
Rossini Favretti, R., Tamburini, F. & De Santis C. (forthcoming). CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model. In Wilson, A., Rayson, P. & McEnery, T. (Eds.), A Rainbow of Corpora: Corpus Linguistics and the Languages of the World. Munich: Lincom-Europa.
Sinclair, J. (1996). Preliminary recommendation on corpus typology. Expert Advisory Group on Language Engineering Standards (EAGLES). #EAG-TCWG-CTYP/P.
Teubert, W. (1996). Comparable or parallel corpora? In International Journal of Lexicography, Vol. 9, 3:238-64
Teubert, W. (1997). Language resources for language technology. In Tufis, D. & Andersen, P. (Eds.), Recent advances in Romanian language technology. Bucharest: Editura Academiei Române.
Váradi, T. (2001). The linguistic relevance of Corpus Linguistics. In Proceedings of the Corpus Linguistics 2001 Conference - CL2001 (pp. 587-593). Lancaster University.