Paul Holmes–Higgin

Khurshid Ahmad

Syed Sibte Raza Abidi
University of Surrey, Guildford

A Description of Texts in a Corpus: 'Virtual' and 'Real' Corpora

Abstract

The extensive use of computer–based corpora for a range of language studies has led to the proliferation of the ways in which texts within an individual corpus are organised. Basically, the organisation reflects the immediate needs of a group of well motivated users, like lexicographers or terminologists. This means that the subsequent generation of corpus users are forced to use a classification of texts according to categories they may not be familiar with or may not be comfortable with or both. There is an urgent need to have a facility in corpus management systems that allows the users to have their own classification system to categorise texts in a corpus. That is, the users should be able to choose, for example, their own style, register, field, time–span and author attributes for generating word lists, concordances, contextual examples and so on. A component of a lexicography and terminology management system, System Quirk, is described that can support such a virtual organisation of texts within a corpus.

1. Introduction

The use of text corpora, particularly the use of computerised text corpora, has had a particularly beneficial use in the study of languages and, perhaps to a lesser extent, on the teaching and learning of languages. Some argue that lexicographers and linguists should choose the texts themselves with some advice from teachers of English (Sinclair and colleagues in Sinclair 1987), while the corpus linguistics pioneers used a random–selection approach (cf. Lancaster–Oslo/Bergen Corpus and the Brown Corpus). Still others have argued that there should be an equal mixture of deliberately selected text and randomly selected text (see, for instance, Summers 1991).

The development of a computer–based corpus of texts requires conversion of published texts onto a computer file system or a data base. This conversion can involve the *coding* of the texts, the *description* of the texts, and, where possible, the *representation* of texts. The coding, or the electronic encryption of texts, is essentially the marking–up of graphetic conventions, including layout information, character codes and so on, such that it is possible to disentangle the layout information from the content of the text.

Once a text is coded in a mark–up language, then, in principle, it is possible to reuse the text on other computer systems. The representation of texts, on the other hand, is a fairly complex matter and involves the specification of syntactic and semantic conventions by which the contexts of the texts can be represented on a computer system. Once a text is *represented* on a computer system then it would be possible for a computer program to infer new information from the text: the computer would, through the use of the conventions, *understand* the texts.

In this paper, however, we will focus on how a particular class of texts can be *described* such that these texts can be stored and retrieved without burdening the corpus user with the details of the description. The three levels of text conversion, coding, description and representation, can be construed as points along a cognitive continuum: from the simplest level, that is coding, to the most complex, that is representation. Description is of intermediate complexity in that, as we show, whilst it involves a level of detail that is much deeper than mere marking–up of texts, there is no attempt made at capturing the meaning of the texts.

The developers of text corpora describe the texts within a corpus specifically for communicating the contents of the corpus to other humans. Usually, the classification of texts is based on the imaginative versus informative dimension, something which is reminiscent of the early attempts at classifying poetry into epics and lyrics. Others would avoid this functional–literary classification and focus on the topics covered in a text. There are instances where the classification of texts concentrates on their linguistic characteristics, based on the frequency of lexico–grammatical categories, and there are classifications, particularly in the terminology literature, that focus on the informative, evaluative, phatic and directive intentions of the writer. Text linguists like to classify texts into narrative, descriptive and argumentative texts. There are numerous ways of describing the genre of a text: indeed, there are many ways of describing the term *genre* itself.

Equally, important descriptors of a text include the medium in which the text is delivered – books, magazines, journals, leaflets, letters. The register and the domain of the text are just as important parameters. There are pragmatic features of any text, like the language variant used by the writer, whether the writer used slang words or restricted himself or herself to the more acceptable sociolect of the language. The time period in which a text is prepared and published can also be used to label text: a mandatory label in a diachronic corpus. Furthermore, there are some atomic features of a text including author's age and sex, and the length of the text.

It appears, therefore, that texts in a corpus can be described through the use of a variety of labels. Indeed, one can create a hierarchy of these labels. There can be pressing lexicographical reasons for considering the medium as the apex of the hierarchy, followed by national language variant, for example, British English and American English, then by date and so on, or for gender studies students the apex would be the gender of the author, followed by date of publication, and language variants. The labels can be arranged, or in some cases have to be arranged to suit the needs of the investigator and his or her own particular niches.

The descriptive taxonomy provides the nodes and links of a network – a *tree–* that describes how texts are related to each other. The nodes are named after the labels and the links provide conduits of properties that can be inherited from the superordinate by the subordinate nodes. For example, the language variant node can be construed as a node that can navigate a user through all the texts that were written in the particular variant; the topic–node can be used to collate texts according to topics. The order of these nodes or the taxonomy then depends upon the individual investigator's niche. The taxonomy chosen by a lexicographer may not suit the needs of a grammarian, and the taxonomy chosen for stylistic studies would be wasted on a historical linguist for example. Indeed, we describe below that even in a niche area, like lexicography, there is no agreement on the descriptive taxonomy.

But no matter whatever taxonomy is chosen, in the context of a computer–based corpora this hierarchy must form the basis of the organisation of texts within a computer's file system. Any change in the taxonomy then suggests the reorganisation of the corpus at the file system level. A complex task at the best of times and, we believe, a task that should be performed by computer systems. In order to explicate the notion of a *configurable* taxonomy we have introduced the term *virtual corpus*. The adjective *virtual* has been borrowed from computing science, specifically operating systems, and is used to describe how entire resources of a computer system are replicated by a program and made available to individual users. The users of this replication are the users of a virtual machine: each believing in and having access to the whole system, whilst in *reality* such an access to and usage of machine is limited for very short intervals of time.

The notion of *virtual corpus* is similar: there is in reality only one corpus, but the users can arrange the nodes and links as they wish and create for themselves a corpus, or more accurately, a corpus organisation, based on an actually physically extant set of texts, for the duration of their use. Thus every corpus user will believe to have access to all or parts of a corpus that they have themselves configured. And, continuing the operating systems analogy, such a configurable taxonomy will have to be made available through the agency of a program, within a suite of corpus management programs, that is capable of producing this virtual corpus. The specification and operation of such a program that can create virtual corpora is the focus of this paper.

2. Structure of extant corpora: Lancaster–Oslo/Bergen, Birmingham Collection, and Longman/Lancaster Corpora

The Lancaster–Oslo/Bergen Corpus was aimed at a general representation of texts for research on a broad range of text types. The texts in the LOB corpus were selected randomly for three 'media': books, newspapers and periodicals, and government documents. Titles were randomly selected from published catalogues and the corpus was categorised into informative texts and imaginative texts. The latter category contains mainly works of fiction, ranging from detective fiction to science fiction and from adventure and 'Western' fiction to general fiction, romantic texts and humour. Figure 1 shows the structure of the LOB corpus.

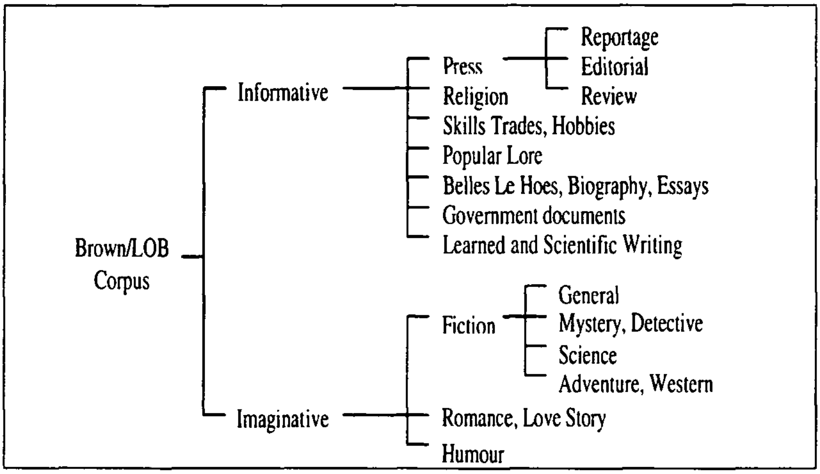


Figure 1: LOB Corpus Structure

Biber (1988 & 1991) has added two more categories to the LOB corpus whilst discussing variation across speech and writing samples of English. First of Biber's additions is professional letters written in an academic context comprising only administrative matters, the second of his categories is personal letters written to friends or relatives. The first category is classed as 'informational and interactional' and the second ranges from 'intimate to friendly' (Biber 1988 & 1991). Presumably both can be added to the informative category introduced by the designers of LOB corpora.

The Birmingham collection of English Text was compiled under the guidance of John Sinclair, in close collaboration with Collins Publishers, and served as a source of "sufficient and relevant textual evidence" (Renouf 1987:1) for the production of "the first wholly new dictionary for many years" (Sinclair 1987:vii): a dictionary not based solely on the introspection of lexicographers and their advisers but based rather on how authors of a wide variety of texts (and speakers partaking in conversation and delivering speeches with and to others) use words and phrases. The COBUILD corpus contains 20 million words of current English in its computer store. The focus of the COBUILD team was on texts published between 1960 and 1985; the team preferred general language text rather than 'technical language'. The COBUILD corpus designers, with advice from teachers of English in the UK and abroad, selected texts themselves, supplemented by a 'relevance' check through the perusal of published sales data of the texts where possible.

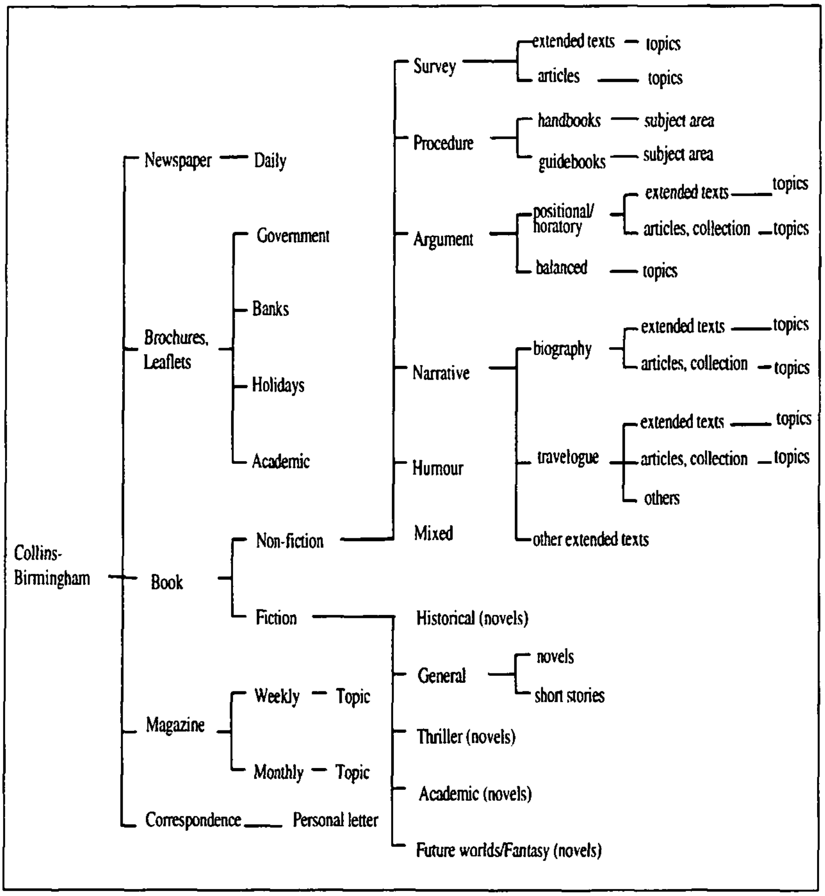


Figure 2: The Structure of the Collins–Birmingham Collection of English Texts

The text in the COBUILD corpus is not split along LOB's informative/imaginative axis, rather the textual 'medium' is taken as a base classifier: books, newspapers, magazines, brochures and leaflets, and personal correspondence are used to define the typology of the texts. The structure of the Birmingham Collection is shown in Figure 2. Note the fine–grained organisation of books: positional and horatory texts, where the 'positional' author puts forward his or her case in relation to a particular topic and the 'horatory' exhorts the reader to do or become something. Summers has argued that the motivation for creating the

Longman/Lancaster Corpus was to provide lexicographers and linguists with "an entirely new, conceived from scratch, corpus of English that could serve a number of purposes and be organised according to objective criteria" (1991: 1). The primary purpose of this 30 million word corpus was "to provide an objective source of language data from which reliable linguistic judgements about the meaning and typical behaviour of words and phrases can be made as a basis for dictionaries, grammars and language books of all kinds" (Summers 1991:3).

What distinguishes the Longman/Lancaster Corpus from the LOB or the Brown Corpus is that the former is 'topic driven' whilst the latter are 'genre driven'. Topic driven texts in the Longman/Lancaster Corpus are categorised in 10 super–fields: science (natural and pure, applied and social); world affairs; commerce and financial; arts; beliefs and thoughts; and fiction.

The lexicographic argument for choosing the topic–based approach, pioneered by Michael Rundell of Longman Dictionaries, was that "it was more likely to produce text categories that were lexically homogenous" (Summers 1991:7). There are four 'external factors' that form the basis of text categorisation: 'Region', including language varieties; 'Time', diachronic corpus containing text published between 1900–1980s; 'Medium' which includes texts books, periodicals and ephemera; and finally, the 'level' of text. For informative texts there are three levels: 'technical', 'lay' and 'popular'. Similarly, the imaginative texts were divided into 'literary', 'middle' and 'popular'. The other features of texts in Longman/Lancaster include the author's gender and country of origin, target age, number of words in total, title, and so on. Most texts in Longman/Lancaster are about 40,000 words long, with no whole texts included because the "emphasis was on many sources rather than the completeness of texts" (the length of texts appears smaller than that of Birmingham's —c. 70,000 where possible— and also the Birmingham Collection has some whole texts). The Longman/ Lancaster Corpus design is such that half of the 30 million words are derived from carefully selected texts (c. 10 million) - the 'selective texts' and the other half is the randomly selected individual titles, collectively known as the 'microcosmic texts'.

3. A virtual corpus management system

The design of corpora, and more so their management, which may include storage and retrieval of texts, navigation mechanisms, and strict integrity and security checks, determines to a large extent the efficacy of the corpora for various end users, which may be lexicographers, translators, or linguists. Most existing corpus management systems have been developed in conjunction with a particular corpus and have consequently taken a fairly literal approach to the implementation of a corpus on a computer. This has resulted in software that directly maps the structure of a corpus as described by the corpus designers to computer–based file or database management system structure. In the following section we are interested in the coding of corpora that allows different corpus designers to structure texts as they feel appropriate. We feel that any user of a corpus can be viewed as a corpus designer.

There have been two main approaches to the storage, retrieval and navigation of texts in a corpus: an explicit text taxonomy, such as LOB and Brown, in terms of file system structure; or implicit text taxonomy, such as Longman, in terms of attributes used in the text 'headers'. There are benefits and limitations with both approaches. With an explicit taxonomy, storage of texts requires a corpus management system to decide where a text should be placed in its file system, whereas the attribute–based system can keep the texts anywhere. The main differences in the two approaches are in text retrieval, and as such, it is useful to think of navigation around a corpus as highly interactive text retrieval.

An explicit taxonomy allows texts to be retrieved quickly by following the appropriate branches through the taxonomy, without needing to consider or refer to the corpus as a whole. The criteria for selecting a text from an explicit taxonomy can be viewed as a 'path' traversing the taxonomic structure. Also, an explicit taxonomy provides a means of navigation through a corpus that computer users find reasonably intuitive. In contrast, an attribute–based system may need to search for the required criteria in the attributes of all texts in the corpus, and is likely to be query–based. For user navigation, query–based retrieval usually means the user has to learn a query language, which some users do not find straightforward.

An important issue for corpus management systems is the type of retrieval requests that a user is likely to make. A frequent use of corpora is for the statistical analysis and comparison of sub–corpora, so it is important for a corpus management system to provide the facility to extract sub–corpora in an intuitive manner by a user. The retrieval benefit of using an explicit taxonomy, however, completely disappears if a number of texts (or sub–corpus) are required that occur in different parts of the taxonomy, which may be considered as the case when incomplete paths are being specified as the retrieval criteria. With an attribute–based approach, this class of sub–corpora can be reasonably easily retrieved.

The aim of virtual corpus management based on a virtual taxonomy of texts, is to provide the flexibility of the attributed–based approach, but with the intuitive functionality of the explicit taxonomy approach. This is achieved by allowing users to define a 'virtual taxonomy' for a corpus of texts, with any number of different virtual taxonomies being concurrently available over the same corpus. The term 'virtual taxonomy' has been defined by Woods in the context of descriptions of concepts in knowledge representation systems, such that whenever a system "constructs an explicit collection of concept nodes ... the result is a subgraph of the virtual taxonomy" (Woods, 1991:80). Woods' motivation for viewing a collection of 'descriptions' this way is that "although its structure is important, one never wants to make it explicit in the memory of a computer" (Woods, ibid).

System Quirk is an exemplar prototype lexical management system (Holmes–Higgin et al, 1993; previously MATE; Holmes–Higgin and Ahmad, 1992) comprising several tools (Table 1). System Quirk promotes corpus–based terminology, manages a corpus of texts that can be organised by a lexicographer or terminologist, manages a lexical data base that is based on one of the versatile data models available, can process data encoded in a language–informed format, can represent lexical and terminological data using knowledge representation schema, and can receive inputs marked up in conformance of various standards and can produce output marked–up in standards that can be processed by a range of desk–top publishing systems. The representation schemata used in System Quirk include relational tables, predicate logic, and sophisticated semantic networks, like conceptual graphs.

Organisational Tools	
Virtual Corpus Manager	organising texts in a corpus
Lexicon Exchanger	import and export lexical data in a variety of formats
Lexicon Distiller	extract smaller, specialised data banks from 'parent'
Lexicon Publisher	prints the lexical data in various formats
Customiser	user profiling tool
Analysis Tools	
KonText	analysing texts
Word Linker	analysing lexical data relationships
Elaboration Tools	
Word Browser	allows the lexica to be browsed interactively
Word Refiner	database management tool to edit word entries
Conceptual Graph Builder	allows the user to browse and build graphs from conceptual relations

Table 1: Components of the System Quirk toolset

The Virtual Corpus Manager within System Quirk has been implemented such that lexicographers and terminologists can view corpora on the basis of the 'pragmatic attributes' of the texts within a corpus. Viewing these pragmatic attributes at an abstract level, we have divided them into six categories: text, authorship, publication, language, domain, copyright status.

The user interface for the Virtual Corpus Manager is shown in Figure 3.

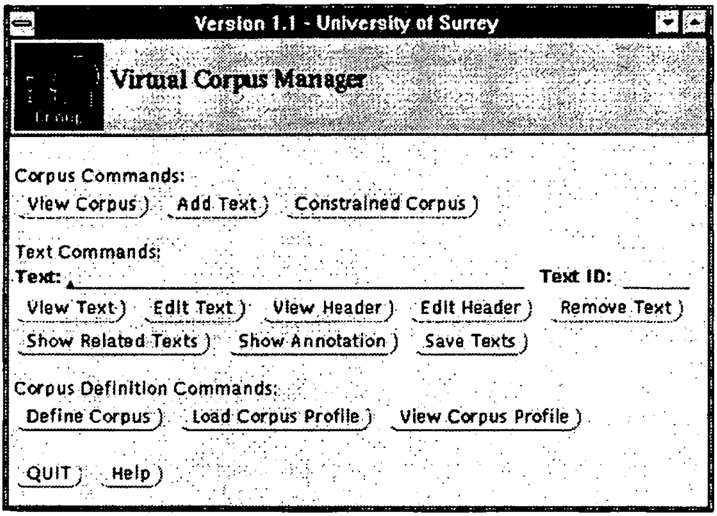


Figure 3: Virtual Corpus Manager user interface.

A configurable taxonomy introduces a shift from the usual pre–defined and explicit corpus taxonomy approach, in that it allows the definition of virtual taxonomies. The Virtual Corpus Manager supports corpora that are coded as explicit taxonomies and corpora whose descriptions are attribute–based. This is achieved by allowing texts to be stored anywhere in a file system and by maintaining attributes describing the texts. Retrieval of the texts can then be made using the attributes directly, or by imposing a virtual taxonomy over the attributes.

Earlier in Figure 2 we showed the structure of the Collins–Birmingham corpus which incorporates a static organisation of texts. The taxonomy has text type (including 'Newspaper', 'Brochures', 'Book', 'Magazine' and 'Correspondence') at the meta–level and the terminal node of the hierarchical tree usually refers to 'topics'. We argue that more than one profile of the same corpus of texts can be generated by implementing a virtual corpus taxonomy, for instance the corpus hierarchy shown in Figure 4, which is a variation of the corpus hierarchy shown in Figure 2.

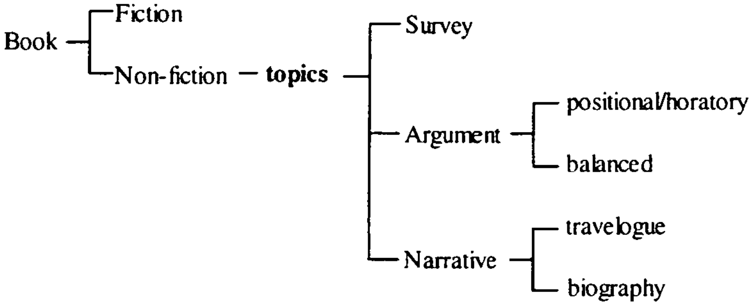


Figure 4: Section of Collins–Birmingham Corpus. The 'topic' sub–corpus from the Collins–Birmingham Corpus.

According to the taxonomy in Figure 4, the non–fiction texts are initially distinguished by the particular 'topics' of the texts, and then the original distinction between narrative, survey and argument texts is maintained. By modifying the original corpus taxonomy in this way the user can now retrieve all non–fiction texts for a particular topic. The selection of texts can be further constrained by choosing texts between 'survey', 'argument' and 'narrative', and so on. Some examples of different virtual taxonomies of the same texts are illustrated in Figure 5. In a dynamic fashion, users can define their own text classification taxonomy or 'corpus taxonomy' from the set of pragmatic attributes, with each level of the taxonomy corresponding to one of these attributes. Additionally, the user is also allowed to restrict the selection of specific values for a pragmatic attribute in the corpus taxonomy (Figure 5b).

This results in a corpus taxonomy that is specific to the users' requirements, as opposed to a common defined taxonomy for all users. For instance, translators may like the top–most level to be 'language', whereas specialist text users may want a taxonomy that has 'domain' as the entry point in the corpus (Figure 5c); similarly 'origination date' with a specification of a range of dates would be the text classification basis for diachronic oriented text research (Figure 5d).

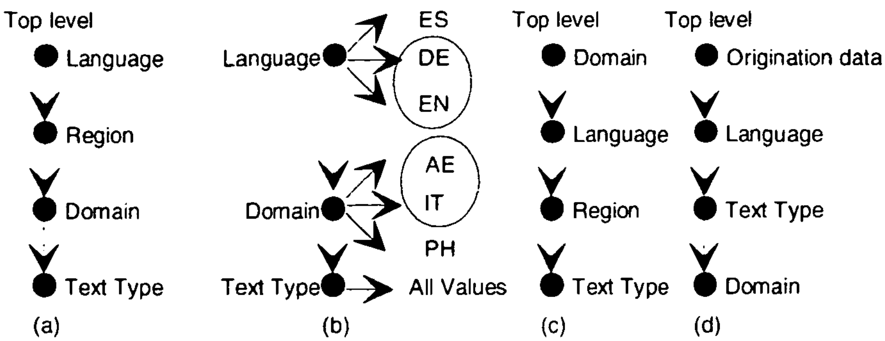


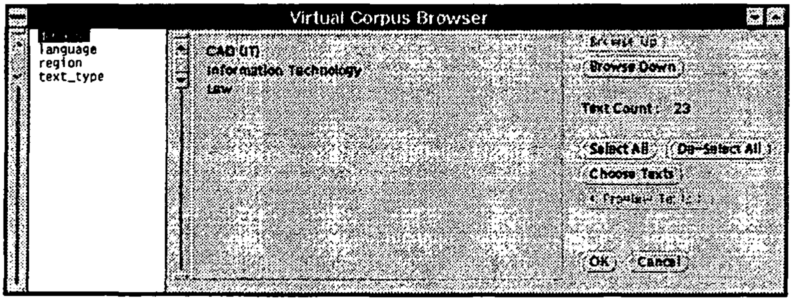
Figure 5 a–d: Example virtual corpus taxonomies. AE refers to Automotive Engineering; IT to Information Technology; PH to Physics

The navigation mechanism implemented in the Virtual Corpus Manager (Figure 6) is novel and has three main advantages:

a) The navigation is based on a user–defined taxonomy, so various views of the corpora can be supported by changing the corpus taxonomy

b) At each level more than one path can be selected concurrently, allowing sub–corpora to be browsed in parallel.

c) At any level only values for known texts are used for determining valid retrieval paths. This ensures that the user may not take a path that leads to a dead–end. For instance, at the language level texts may classified into four languages 'English', 'German', 'Italian' and 'Spanish', however when browsing down if there are no Italian texts in the corpus, this path would not be available.



Text Count : 23

Figure 6: Virtual Corpus Browser with the virtual taxonomy defined in Figure 5a.

The Virtual Corpus Manager provides a mechanism that allows the user to specify various constraints in a simple interactive manner, without recourse to a query language (Figure 7), and then retrieves all texts satisfying the user's constraints. We argue that the actual corpus containing all texts can be considered as the 'mother corpus', whereas the derived sub–corpus, which in fact partitions the corpus based on certain user defined constraints, can be regarded as the 'daughter corpus'. Furthermore, our approach to corpus management incorporates the notion that texts in a corpus can be related with other texts, for example as 'shadows' (translations), annotations and so on.

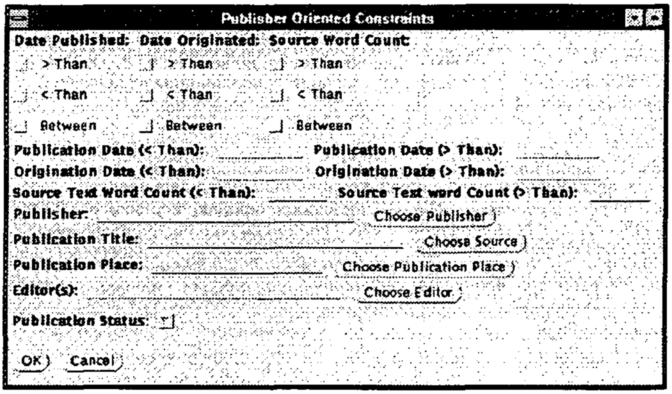

Publisher Oriented Constraints
Date Published: Date Originated: Source Word Count:
> Than > Than > Than
< Than < Than < Than
Between Between Between
Publication Date (< Than): Publication Date (> Than):
Origination Date (< Than): Origination Date (> Than):
Source Text Word Count (< Than): Source Text word Count (> Than):
Publisher: Choose Publisher
Publication Title: Choose Source
Publication Place: Choose Publication Place
Editor(s): Choose Editor
Publication Status:
OK Cancel

Figure 7: Text selection by attribute query.

4. Conclusion

The discussion above covered the various exemplar corpora used extensively in corpus linguistics together with our views on corpus taxonomies. We focused on how a corpus taxonomy can be made flexible such that each individual user of the corpus can impose his or her own structure on the corpus for the purposes of pursuing their own investigation. We believe that much of the debate on text typologies is descriptive and it is not possible to put a value on any of the text typology: the notion of virtual taxonomies and associated implementations (like the Virtual Corpus Manager) will introduce some degree of objectivity in that one can evaluate the efficacy of one type of typology against another.

References

Biber, Douglas (1988). *Variation across speech and writing.* Cambridge: Cambridge University Press (there is a 1991 paperback edition of this book from which the citations are taken).

Holmes–Higgin, P. and Ahmad, K. (1992). *The Machine Assisted Terminology Elicitation Environment: Text and Data Processing and Management in Prolog.* Technical Report CS–92–11. Dept. of Mathematical and Computing Sciences, University of Surrey, Guildford.

Holmes–Higgin, P., Griffin, S., Hook, S. and Abidi S.R. (1993). *System Quirk Reference Guide.* Final Report for Workpackage 5.5, Multilex Project, ESPRIT II, No. 5304.

Renouf, Antoinette (1987). Corpus Development. In John Sinclair (1987): 1–40.

Sinclair, J. (1987). *Looking Up.* London. Collins.

Summers, D. (1991). Longman/Lancaster English Language Corpus: Criteria and Design. Unpublished manuscript.

Woods, W.A. (1991). Understanding Subsumption and Taxonomy: A Framework for Progress. In John F. Sowa (ed.). *Principles of Semantic Networks:* 45–94. Morgan Kaufmann Publishers, California.