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

Paul Holmes-Higgin, Sibte Raza Abidi and Khurshid Ahmad
University of Surrey

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 system that allow its users to use 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, author attributes for generating word lists, concordances, contextual examples etc. A lexicography/terminology management system, *System Quirk*, is described that can support such a *virtual* organisation of texts within a corpus.

Introduction

There are open questions in corpus linguistics related to how texts should be selected and, perhaps, more importantly for what purpose. Some argue that lexicographers and linguists should choose the texts themselves with some advice from teachers of English (Sinclair and colleagues in Sinclair 1987), whilst 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).

We hope that the discussion of how text is organised and, indeed, how representative text is chosen, will motivate the reader to consider various parameters that can label a text. These parameters may include the medium in which the text is delivered - books, magazines, journals, leaflets, letters; the genre of the text, fiction or non-fiction, whether it is imaginative or informative, persuasive or instructional. The register and the domain of the text are equally important parameters. Furthermore, there are some atomic features of a text including author's age and sex, publication period, language variety and so on. (One might consider the use of 'contextual correlates' described by Halliday to categorise texts in terms of their tenor and field, given that the mode of the language in the text corpora is textual).

The LOB 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.

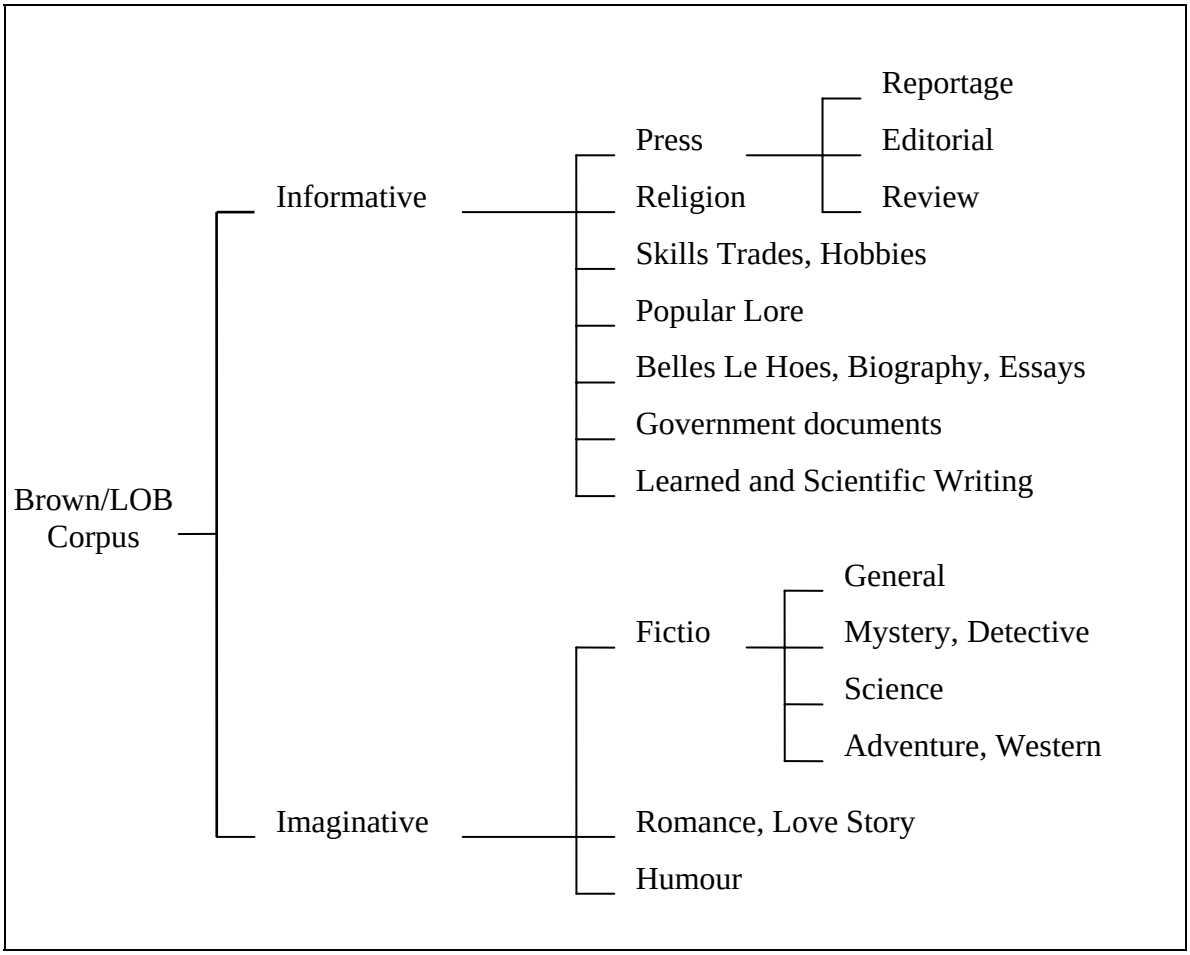

Figure 1: LOB Corpus Structure

Biber[1] has added two more categories to those used in the LOB corpus whilst discussing variation across speech and writing samples of English. First of Biber's additions is professional letters written in academic context comprising only administrative matters, the second of his categories is personal letters written to friends or relatives. The first category is defined as 'informational and interactional' and the second as from 'intimate to friendly' (Biber 1988: 67). 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

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

"sufficient and relevant textual evidence" (Renouf[2] 1987: 1) for the production of "the first wholly new dictionary for many years" (Sinclair 1987: vii. The COBUILD corpus contains 20 million words of current English in its computer store. The COBUILD corpus excluded certain categories of text included in LOB, such as poetry, and excluded drama as not an example of 'naturally occurring 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 text typology. The structure of the Birmingham Collection (Renouf, 1987:23-32) is shown in Figure 2.

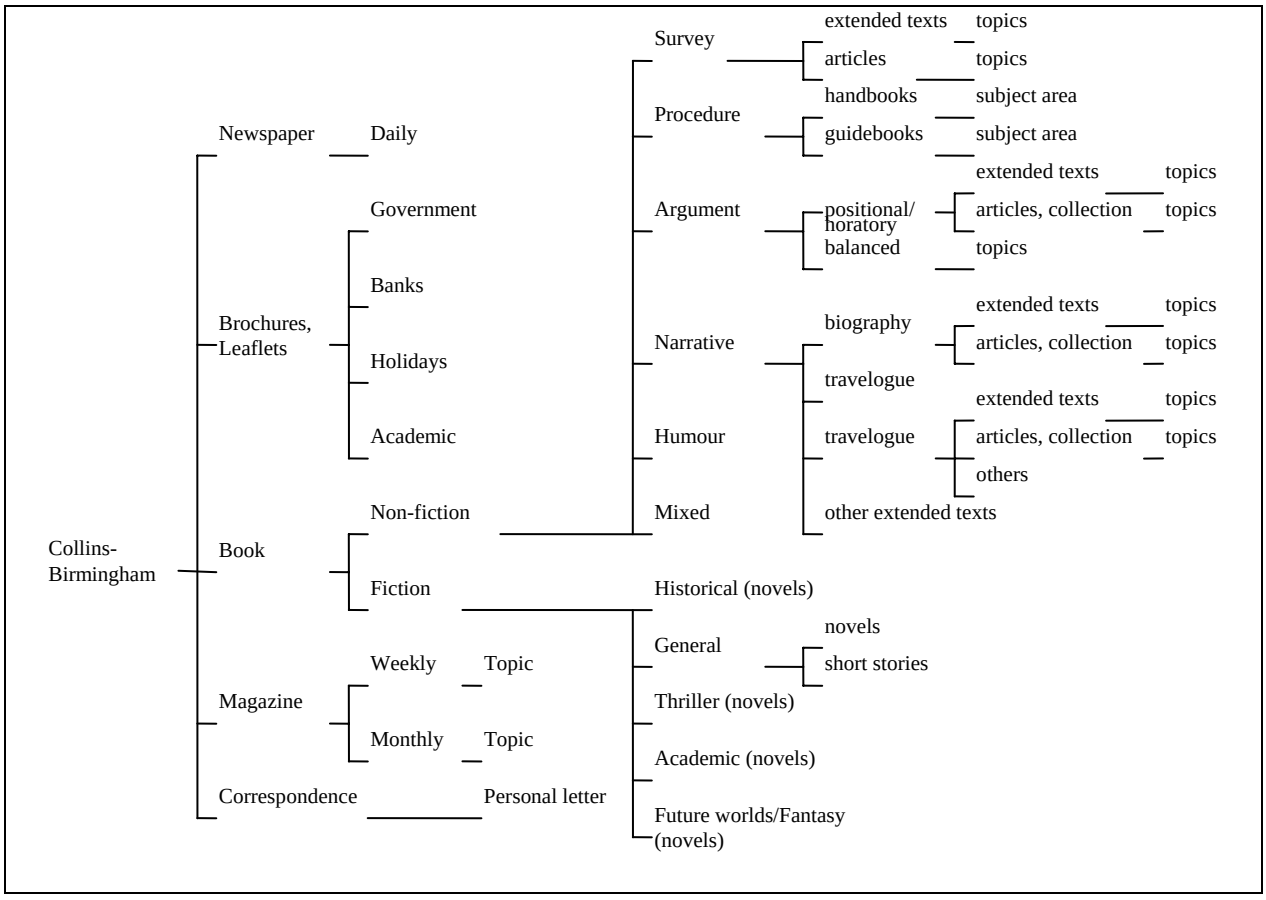

Figure 2: The Structure of the Collins-Birmingham Collection of English Texts

Note the fine-grained organisation of books: positional and horatory texts, when 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; the Birmingham collection distinguishes between handbooks and guidebooks and the distinction between a variety of narrative texts, travelogue, biography and autobiography is maintained.

[2]Renouf, Antoinette (1987). 'Corpus Development'. In (Ed.) John Sinclair (1987). pp1-40.

Summers[3] 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 judgments 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'. Recall from Figure 1 that LOB distinguishes between 'academic discourse' from 'press reportage', 'press editorial' from 'arts' and so on. The 'topic driven' texts in the in 10 superfields. Texts in Longman/Lancaster are divided into metacategories, informative and imaginative, subdivided into superfields and, like LOB corpus, informative texts comprise books, newspapers and journals, unpublished and ephemera, and imaginative texts, mainly works of fiction in book form.

There are four 'external factors' that form the basis of text categorisation in Longman/Lancaster: 'region', including language varieties; 'time', diachronic corpus containing text published between 1900-1980s; 'medium' includes the 'sources' of texts books (80%), periodicals (13.3%) and ephemera (6.7%); 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 'axiomatic' features of texts in Longman/Lancaster include author's gender and country of origin, target age and gender, number of words in total, title, and so on. Most text types in Longman/Lancaster are about 40,000 words long and no whole texts were 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, which also has some whole texts).

Longman/Lancaster Corpus design is such that half of the 20 million words are derived from carefully selected texts (c. 15 million) - the 'selective texts' and the other half is the randomly-selected individual titles collectively known as the 'microcosmic texts'. Like the LOB corpus, Longman/Lancaster have used a book catalogue - Whittaker's Book in Print - and selected texts originally published in English (in English-speaking countries) before 1900 excluding dictionaries and reference works and works for children.

Historical linguists and etymologists keen to study language change and eager to investigate the life cycle of a 'loan word' can benefit from a text corpus containing texts that were written/published at different times covering decades, and in some cases 2000 years. Indeed, the diachronic and dialectal Helsinki Corpus of English

---

3

Texts comprises 2100 text samples ranging from 2500 to 10,000 words, claimed to be written between 850 - 1720. The dialects studied include south west Midlands, Mercian, Cornish, Devon, Somerset, Wiltshire and South Avon (see, for instance, Rissanen[4] (1991) for the diachronic organisation of Helsinki Corpus and Ihalainen[5] (1991) for dialectal variation within the corpus.

Virtual machine: A computer designed to replicate copies of its entire hardware/software interface so that two operating systems can be run on a single machine (AHD, pp1996).

Biber has discussed the typology of English texts at length (Biber[6], 1988, 1989) and has attempted to shift the basis of text typologies from practical to linguistic grounds. Functional criteria of text types is based on one or two particular functional dichotomies, such as formal versus informal, involvement versus detachment, integration versus detachment, and the use of these points of distinction to describe text 'types'. Whilst Biber does not detract from the *utility* of functional distinctions, he argues that these typologies leave much to be desired 'in identifying the salient LINGUISTIC differences among texts in English'. (1989: 5).

Biber has pointed to the considerable linguistic variation in a given functional type of text whilst there is a systematic co-occurrence of linguistic features across the functional types. The linguistically grounded typology is based in 'sets of syntactic and lexical features that co-occur frequently in text', the so-called 'dimensions' of variation identified empirically by multivariable practicable methods, factor analysis to be precise: similar texts in each type are maximally similar in their linguistic characteristic, while the different types are maximally distinct from one another' (Biber, 1989: 5). Such a typology implies important functional differences as lexical and syntactic features are used to indicate common functions.

Biber uses the dimensional statistics to group LOB and London-Lund texts into 'clusters': 'Texts that are similar with respect to the dimension but very different with respect to other dimensions are likely to be grouped into different clusters' (1989: 17). the texts in LOB and London-Lund corpora appear to form eight 'distinct' clusters and their dimensional scores on each of the five dimensions indicate an interacting functional nature of each cluster. Each cluster is made up of either dominant dimension or the absence of such characteristic.

Table 1 shows the composition of the clusters according to dimensionality:

---

[4] Rissanen (1991). On the history of *that*/Zero as object clause links in English. In (Eds.) Kavin Aijmer and Bengt Altenberg (1991). pp 272-289.

[5] Ihalainen, Ossi. (1991). A point of verb syntax in South-Western British English: Analysis of a dialect Continuum. In (Eds.) Kavin Aijmer and Bengt Altenberg. pp280-302.

[6] Biber, Douglas (1989). 'A Typology of English Texts'. *Linguistics*, Vol. 27, pp3-43.

CLUSTERS **Interaction**	**Involved** **-v-** **Informational**	**Narrative** **-v-** **Non-narrative**	**Elaborate** **-v-** **Situated ref.**	**Overt** **Persuasion**	**Abstract** **-v-** **Non-abstract**
1. Intimate Interpersonal	Extremely involved	Not marked	Situated	Not marked	Non-abstract
2. Informational	Less involved	Not marked	Situated	Not marked	Non-abstract
3. Exposition/ Scientific	Extremely informational	Non-narrative	Highly elaborated	Non-persuasive	Extremely abstract
4. Learned	Extremely informational	Non-narrative	Highly elaborated	Non-persuasive	Moderately abstract
5. Narrative/ Imaginative	Moderately involved	Extremely narrative	Situated	Not marked	Non-abstract
6. General expository	Informational	Narrative	Not marked	Not marked	Not marked
7. Situated Reportage	Not marked	Non-narrative	Situated	Non-persuasive	Non-abstract
8. Involved Persuasion	Moderately involved	Non-narrative	Moderately elaborated	Extremely persuasive	Moderately non-abstract

Table 1: Clusters expressed in terms of their dimensions

Coding, Description and Representation of Texts

Computer programs are essential for the creation of a text corpus, and are equally important for retrieving one or more texts, or one or more text fragments.

The corpus creation programs are based on their programmers' understanding of the salient properties of the texts that are to be stored in a corpus, and on their understanding of how a corpus is to be structured. The programmers, in turn, learn about the properties and the structure from the organisers of the corpus: these organisers can be literary experts, linguists, lexicographers, information scientists and so on.

Once a corpus is created then a potentially well-defined group of users access the stored texts through another set of computer programs. These retrieval programs are based on a model of the user's requirements as understood by programmers working under the direction of the corpus organisers.

The creation and the subsequent usage of a text corpus appears then to depend on the shared knowledge of the corpus organisers, the programmers and the end-users. This shared knowledge has to be simple and formal enough such that it can be used in the corpus creation and retrieval programs. This shared knowledge has to be articulated with a zero-intelligence device, a computer, in mind. Furthermore, as the texts are to be stored in an electronic medium as distinct from the usual 'graphic substance' used in the print medium that is in the form of marks on a surface. The graphetics-physical properties of the symbols that constitute writing systems have to be transformed from print to the electronic medium for the purposes of storage, whilst the reverse process has to e performed for retrieval. The efficient storage and retrieval of texts depends on how texts and text corpora are dealt with at three overlapping yet clearly identifiable tasks: the *coding* of the texts; the *description* of the texts; and the *representation* of

texts. In other words, how texts are to be *coded* or, more precisely, encrypted for use in an electronic medium, how a particular class of texts and text corpora are to be *described*, and how a class of texts and text corpora is to be *represented* on a computer system. The coding, description and representation comprise some of the shared knowledge mentioned above and it is the articulation of the knowledge that will concern in the following subsections.

Description of Texts

The 'description' of individual texts or a group of texts is essentially a set of conventions used to describe some particular texts. Corpus organisers attach a variety of *descriptive* labels to individual texts. These labels may be used to express, say, functional typology: Longman's typology is at some variance with the Birmingham Corpus of English, and the two are at variance with de Beaugrande and Dressler (1981) and with Sager, Dungworth and McDonald's typology.

The description of a text in a corpus is usually at a meta-level as compared to the coding of the text in terms of much of its attributes. This description can be the description of the text's functional type, for example, whether the text is imaginative, informative, evaluative or phatic: this description may be what Sager, Dungworth and McDonald call the 'description ..... of traditional forms' (1980: 148-181); this description may involve the subject domain members responsible for producing the text. The advent of tagged corpora allows for a lexico-grammatical description of the text in that all known and current tokens in the text are tagged according to the categories and features by given grammar.

Therefore, in a modern computerised corpus of texts one may find meta-textual data, that is, additional data that is used to code and to describe each individual text in a corpus. The question here is this: can this meta-textual data be used to build a taxonomy that contains the functional typology at its apex or functional typology as the superordinate level, text-forms at a subordinate level, domain-specific data subordinated to the text form, and the rest of the attributes at the instance level. then again, one may regard the domain or subject specificity as being the superordinate level rather than the functional typology, or indeed any other attribute.

We believe that almost all of the corpus management systems are used in large measure by an enthusiastic, well-motivated community that shares the knowledge of the meta-textual data with the developers of corpora and software engineers that build the corpus management systems. The corpus and software designers, and the users, have a tacit understanding of how the meta-textual level data has been used in specifying a taxonomy. The shared knowledge mentioned above is the shared knowledge of the taxonomy.

However, the establishment and the consolidation of corpus linguistics would mean that methods may be developed by a team that has little or no contact with those identifying new techniques or building tools for analysing corpora. The growth of corpus linguistics means that the first victim of the success would be the shared

knowledge of text taxonomies. The loss of the shared knowledge also may give rise to the opinion that a corpus management system has some privileged data about a given taxonomy and yet may not have any data about any other taxonomy.

We believe that a careful and systematic approach for collecting the meta-textual data about individual texts in a corpus together with the development of programs that can manipulate this meta-textual data to create a number of different taxonomies and programs that are capable of providing access for a user to all the texts for a given taxonomy, is not only necessary but holds the key to a successful user-driven exploitation of text corpora.

The flexible approach to text typology can be computationally grounded by predicating that texts in a given corpora can be virtually-linked to each other. This means that a corpus can be configured according to the choice of its end-users by exploiting this so-called virtual link.

The access to a group of related-texts in our system - System Quirk - is provided on par (?) with the definition of virtual link, a link in telecommunications that may be realised by the use of different circuit configurations during transmissions.

4. 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-store 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-store, whereas the attribute-based system can keep the texts anywhere. The main differences in the to approaches are in text retrieval, and in this 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 atribute-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 straight-forward.

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 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, sub-corpora can be easily retrieved. The aim of virtual corpus management 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, 1991:80).

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 main screen for the Virtual Corpus Manager is shown in Figure 3.

Figure 3: Virtual Corpus Manager main screen.

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

Earlier in Figure 2 we showed the structure of the Collins-Birmingham corpus which incorporates a static organisation of texts. The hierarchy has text type (including 'Newspaper', 'Brochures', 'Book', 'Magazine' and 'Correspondence') at the metalevel 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 hierarchy.

According to the above corpus hierarchy the non-fiction texts are firstly distinguished between the paticular 'topics' of the texts, and afterwards the original distinction between narrative, survey and argument texts is maintained. By modifying the original corpus hierarchy in this way the user can now retrieve all texts non-fiction texts of 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 hierarchies of the same texts is illustrated in Figure 4. In a dynamic fashion the users can define their own text classification hierarchy or 'corpus hierarchy' from the list of pragmatic attributes, where each level of the hierarchy corresponds to one of these attributes. Additionally, the user is also allowed to include only relevant values of an pragmatic attribute in the corpus hierarchy (Figure 4b).

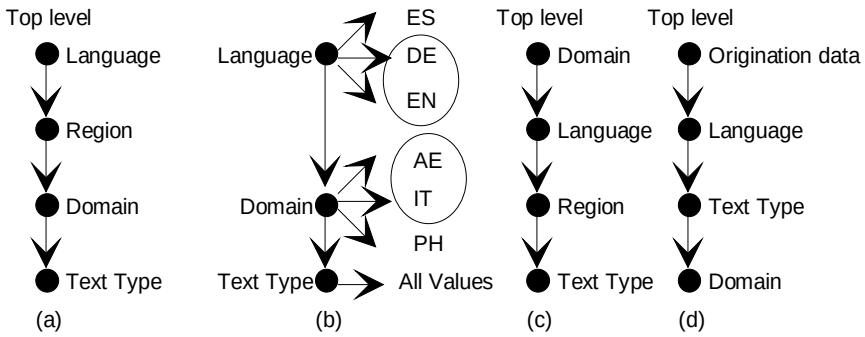

Figure 4 a-d: AE refers to Automotive Engineering; IT to Information Technology; PH to Physics

This results in a corpus hierarchy that is specific to the users' requirements, as opposed to a common defined hierarchy for all users. For instance, translators may like the top-most level to be 'language', whereas specialist text users may want a hierarchy that has 'domain' as the entry point in the corpus (Figure 4c), similarly 'origination date' with a specification of a range of dates would be the text classification basis for diachronic oriented text research (Figure 4d).

The navigation mechanism implemented in the virtual corpus manager is novel and has three main advantages:

a) The navigation is based on a user-defined hierarchy, so various profiles of the corpora can be viewed by changing the corpus hierarchy (Figure 5a).
b) At each level more than one path can be chosen to browse down concurrently (Figure 1Figure 5b). For example, just the American (US) and British (GB) variants of English could selected.
c) At any level only known values for texts are available for determining the path. This ensures that the user may not take a path that leads to a dead-end. For instance, at the language level texts are 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.

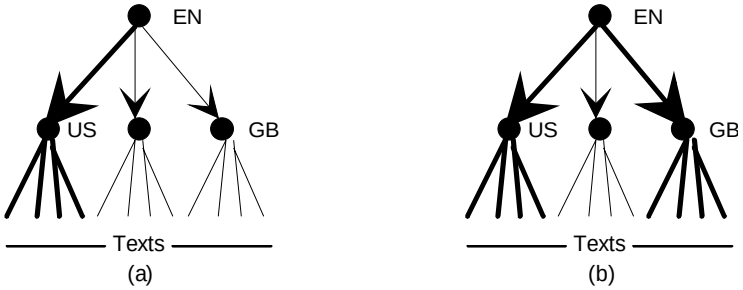

Figure 5 a-b

The browser for navigating a virtual hierarchy is shown in Figure 6.

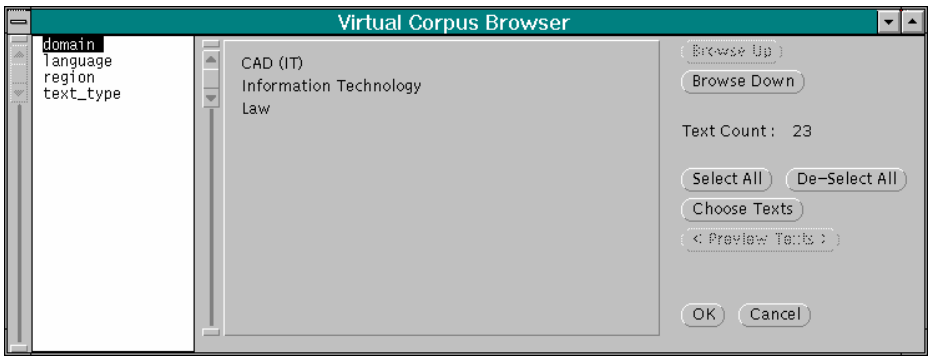

Figure 6: Virtual Corpus Browser with the virtual hierarchy defined in Figure 4.

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. The collection of texts retrieved that satisfy these constraints could be envisaged as a 'constrained' corpus. We argue that, the actual corpus containing all texts can be considered as the 'mother corpus', whereas the constrained corpus, which in fact partitions the corpus based on certain user defined constraints, can be regarded as the 'daughter corpus'. Furthermore, our approach for 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.

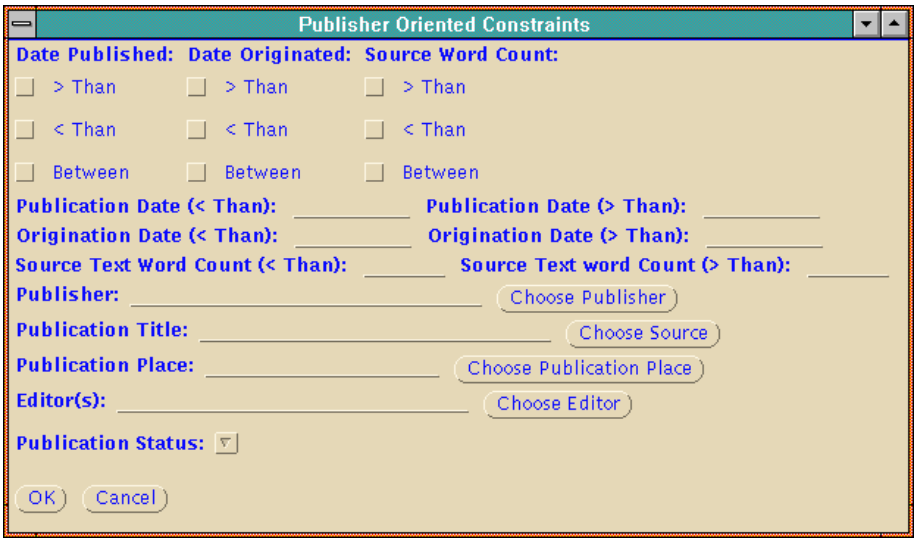

Figure 7: Text selection by attribute query.

5. Conclusion

The above discussion covered the various exemplar corpora used extensively in corpus linguistics together with our views on corpus hierarchies. We focussed on how this hierachy can be made flexible such that each individual user of the corpus can impose his or her own structure on thecorpus for the purposes of pursuing their own

invstigation. 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 virtual corpus manager will introduce some degree of objectivity in that one can evalute the efficacy of one type of typology against another.

Bibliography

Ahmad, K., Davies, A., Hughes, M., Fulford, H., Rogers, M., and Thomas, P. (1990), A methodology for building multilingual termbases and special-purpose lexica, Translator's Workbench Project, ESPRIT II, No. 2315, Report for Workpackage 1.1, Guildford: University of Surrey

Ahmad, K., Fulford, H., Griffin, S., and Holmes-Higgin, P., (1991), 'Text-based Knowledge Acquisition —A Language for Special Purposes Perspective'. In (Eds.) I.M. Graham & R.W. Milne, Research and Development in Expert Systems VIII, pp 148-162. Cambridge: Cambridge University Press.

Ahmad, K., Fulford, H., Holmes-Higgin, P., Rogers, M., and Thomas, P. (1990), 'The Translator's Workbench Project', in Translating and the Computer 11, ed. C. Picken, London: Aslib

Aijmer, K. and Altenberg, B. (Eds). (1991). *English Corpus Linguistics -- Studies in Honour of Jan Svartik*. London: Longman Group.

Garside, R., Leech, G. and Sampson, G. (Eds). (1987). *The Computational Analysis of English*. London: Longman Group.

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.

Leech, G. (1991). 'The state of the art in corpus linguistics'. In (Eds) Aijmer and Altenberg. pp 8-29.

Tompa, W.F. and Raymond, R.D. (1991). 'Database Design for a Dynamic Dictionary'. In (Eds.) Susan Hockey and Nancy Ide. *Research in Humanities Computing: Selected Paper from ALLC/ACH Conference, Toronto (June 1989)*. Oxford: Clarendon Press. pp 257-272.

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