There are texts for which hierarchical structures are inadequate. Shakespeare and legal texts are good examples. Their essential characteristic is that units may need to be linked to multiple units at higher levels of the tree structure rather than the single unit allowed in hierarchical structures. Such structures suggest the need to examine models described later where words are considered as atoms of data to be built dynamically into a variety of complex molecular objects.

Linked particularly with the navigation requirements described earlier is the need to generalise when describing text structures. For example, in a hierarchical text structure, any one part of the tree may usually cite any other part. The textbase can be viewed at two levels: generalisation for an abstract overview in which any type of text object cites any other type; and specialization for a more detailed representation in which a specific type of text object cites another specific type.

3 Semantic Models and Text Structures

Database models can be categorized into two main types: basic and semantic. A range of semantic models has been proposed in order to incorporate more features, constraints and abstractions than are found in the basic ones in an attempt to represent more closely the real world. These include the Entity-Relationship (E-R) Model [Chen 1976] and Taxis [Mylopoulos et al 1980]. Text because of its complex nature usually requires full semantic models to capture completely its structure and examples of Chen, Taxis and others have been developed at Newcastle.

3.1 Models for Expressing Static Aspects

The viewpoint of Chen is that database design is concerned primarily with the occurrence of entities and the relationship between them. An E-R diagram of a UK statute could be represented in the form of figure 2(a) using rectangles to denote entity-types and diamond-shapes to denote relationships. A relationship flagged '*' is mandatory, otherwise it is optional. All relationships are one-to-many (1:N) bar one. The idea of generalisation is employed with the scope of a generic entity-type being delineated by the enclosure of its associated specializations within a thickly-lined rectangle. The generic structures defined are *node* to represent all possible text units from an act through parts and schedules to subsections and subparagraphs and *text* to represent the units holding the main part of the text - section,