for them. In effect, it is its own markup, saying 'Message for you' or 'Don't forget. . . '. But it is also available in white, preprinted with 'Phone Message', and with fields for the caller's name, number, date, time, and topic. The first is markup-free and has universal applicability. The second is for a special purpose, and the markup has been designed to prompt the writer not to forget key information.

Given the seemingly unattainable nature of fully re-usable markup, considerable attention has been paid to the use of logic, heuristics, and statistics to automate the process (Kelly & Abrahamson, 1991; Taghva, Condit, & Borsack, 1995; Abolhassani, Fuhr, & Gövert, 2003). However, while systems have been developed for 'vertical' applications such as news article markup (Haake, Huser, & Reichenberger, 1994), to date there is no general-purpose system available to implement the techniques for an arbitrary range of documents.

It is nevertheless true that there is usually some degree of structure evident within even simple documents, such as a blank line or indentation to indicate a new paragraph, or a large font to indicate a heading, and it is possible to develop ad-hoc systems using simple toolsets such as the Unix text tools to impose rudimentary but sufficient markup to enable documents to be opened in an XML or LaTeX editor, and leave the finer detail to human editing. Interpreting in any greater detail the arbitrary and inconsistent nature of manually-applied formatting and layout, given its high level of context-dependency, remains a subject for further work in the field of Information Retrieval. As some of the authors above remarked:

> Further, all these heuristics become useless and the difficulties we have mentioned multiply, if the device fails to zone a page properly. For example, the title-finding module will not find the title if, in a two column document, its zoning order follows the first column of text. This may occur if a document's title is right justified; also, if a floating object is zoned improperly, any object recognized by *Autotag* may be identified incorrectly. Proper zoning is a prerequisite for correct *Autotag* output.
>
> (Taghva et al., 1995, p 325)

(We identify in section 4.3.4.1 on page 220 a related problem with the use of Named Styles when dealing with wordprocessor document.)

While physical structural simplicity may be evident in some classes of documents (novels, for example), at the opposite extreme there are a number