Active paper for active learning

Heather Brown,* Robert Harding,** Steve Lay,** Peter Robinson,***

Dan Sheppard*** and Richard Watts***

*Computing Laboratory, University of Kent at Canterbury

**Department of Applied Mathematics and Theoretical Physics,

University of Cambridge

***Computing Laboratory, University of Cambridge

Recent research into distance learning and the virtual campus has focused on the use of electronic documents and computer-based demonstrations to replace or reinforce traditional learning material. We show how a computer-augmented desk, the DigitalDesk, can provide the benefits of both paper and electronic documents using a natural interface based on real paper documents. Many electronic documents, particularly those created using the guidelines produced by the Text Encoding Initiative (TEI), include detailed semantic and linguistic information that can be used to good effect in learning material. We discuss potential uses of TEI texts, and describe one simple application that allows a student's book to become an active part of a grammar lesson when placed on the DigitalDesk. The book is integrated into an interactive point-and-click interface, and feedback is related to the currently visible pages of the book.

Introduction

Paper documents have great advantages in readability, portability and familiarity, but are necessarily static and slow to update. Much recent research has concentrated on the dynamic demonstrations, immediate feedback, and easy updating that can be provided by electronic teaching material. Although an increasing number of teaching packages make use of both paper and electronic documents, the two are typically accessed by completely separate interfaces. We have been taking a different approach and investigating the use of a DigitalDesk (Wellner, 1991; Wellner, 1993) as a means of integrating normal paper teaching material with electronic versions of the same material. Many printed books also exist in electronic form, and our goal is to allow these books to be used as natural interfaces to any additional information that may be present in the electronic version.

We have looked primarily at electronic texts prepared according to the guidelines published for the Text Encoding Initiative (TEI) (Sperberg-McQueen and Burnard, 1994) as these provide a potentially rich field of additional information to exploit. Many large projects have produced texts conforming to the TEI guidelines. The majority provide information on old books and manuscripts, but the British National Corpus (BNC) (Burnage and Dunlop, 1993) provides information on the part-of-speech of over 90 million words of modern English texts. As a concrete example of learning material based on TEI information, we have taken a publicly available text from the BNC – a simplified modern version of Lewis Carroll's *Alice's Adventures in Wonderland* (Carroll, 1994) – and have used it as the basis of a student's grammar lesson. The DigitalDesk recognizes the text on the open pages of the physical book – the one the student carries around and reads in the normal way – and matches it up with the corresponding part of the BNC text in a way that allows the book to become part of an active grammar lesson based on the currently open pages.

The following sections introduce a few of the most relevant features of the TEI, explain the example application in more detail, and outline a number of further possible uses of TEI texts in learning material.

The Text Encoding Initiative

The Text Encoding Initiative provides detailed guidelines for creating SGML (Goldfarb, 1990; Alschuler, 1995) documents that encode many different types of information about the original documents in addition to their textual content. The guidelines are designed particularly for capturing scholarly information about old manuscripts, but have been used for a wide variety of texts and projects

In (over-)simplified terms, SGML provides a syntax for adding *markup* or *tags* to a text to identify the start and end of every *element* of a document. Elements may contain nested elements, so the SGML document is a hierarchy of nested elements, Tags contain the name of the element, and start tags may also contain further information expressed by attribute-value pairs. An SGML Document Type Definition (DTD) controls the syntax by defining the names of the elements allowed and the ways in which they may be nested and combined. An associated application defines the meaning of the elements. The TEI is an application of SGML, so the guidelines consist of a DTD together with recommendations on the meaning and use of the 400 elements it defines.

A text may use any subset of the full TEI facilities, but if a particular type of information is encoded it must conform to the recommended markup. As a simple example of this, a paragraph of text may simply be coded as a single 'p' element by enclosing it in start and end tags(<p>.....</p>). Alternatively, it may be further broken down into sentences (<s>), words (<w>) and punctuation (<c>). This more detailed form is used throughout the BNC texts. Each sentence is also given a sequence number, and each word is labelled with a code indicating its part of speech. The following sample from the start of the BNC *Alice* text illustrates the level of information provided.

```
<p>
<s n=002>
<w NP0>Alice <w VBD>was <w VVG>beginning <w T00>to
<w VVI>get <w AV0>very <w AJ0-VVN>bored<c PUN>.
```

This detailed encoding of each word in a BNC text forms the basis of the application described in the next section.

Using the DigitalDesk for a grammar lesson

The DigitalDesk is built around an ordinary physical desk. It enables the user to interact with paper and electronic documents (projected onto the desk) using an LED-tipped pen. The movement of the pen is monitored by a computer connected to a video camera mounted above the desk. Harding *et al* (1997) used the DigitalDesk to develop prototype educational applications by printing documents on special paper marked with computer-readable *glyph* codes (Robinson *et al*, 1997). We extend this approach by using the information in the BNC texts to recognize pages instead, enabling previously published material to become activated too.

The normal published version of the *Alice* text is a book of about 40 pages. The BNC text covers all the running text of the book, but not the pictures, captions or page headers. When the book is placed on the DigitalDesk, the page images captured by the camera are used to match the visible running text to its BNC counterpart. This is done by recognizing the positions of the lines of text on the pages and using the pattern of ascenders and descenders in the lines to provide a match with the BNC text. Once this has been done, it is relatively simple to find the position of each word on the pages and to associate it with its part of speech as specified in the electronic version.

The user interface for the example grammar lesson has been kept deliberately simple. Once the pages of the book have been recognized, two windows of information are projected onto the desk, one either side of the book. One of these leads into a simple lesson providing information about different types of words. The other is for a grammar quiz based on the words on the page. The lesson provides information on the main types of words. Users point to menu items in the projected information to request more detailed information or examples. Figure 1 (left) shows the initial information that appears if a user asks about nouns, and Figure 1 (right) shows the further information that appears if the user points to the Proper Nouns menu item.

As Figure 1 (left) shows, the user can also ask to see all the words of a given type on the pages. Selecting this option causes all the appropriate words on the open pages to have coloured rectangles projected onto them, providing a vivid illustration of the use of the words, and integrating the paper text into the lesson. Different colours are used throughout the design to identify the different types of words. Thus the nouns shown in italics in Figure 1 are projected using the colour for nouns, and if nouns are highlighted on the pages the same colour is used. At any time users can also point to a word on one of the pages and see information about that word. Figure 2 shows the result of pointing to the word 'her' in the book.

The quiz is also used to bring the pages to life. It tests users' knowledge of the words on the pages by asking them to solve simple problems such as 'Find two adjectives on the pages', 'Find three words from the verb "to be" on the pages', or 'Is the highlighted word an adjective or an adverb?'. To solve the 'Find . . .' problems, users point to words on the pages and receive feedback on their choices. The emphasis throughout is on using the book as the source of the information, and on projecting information onto the book to relate the

Nouns
Nouns are words for things (*cat, house, star, nose, policeman, ship*). Most nouns have a singular and a plural form. The common way to turn a singular noun into its plural form is to add 's' or 'es' (*cat/cats, dish, dishes*) but some have unusual plurals (*pony/ponies, foot/feet*) and some do not change (*sheep, aircraft, food*).
More about nouns **Proper nouns (names)** **More examples** **Show me the nouns on the pages**

Proper nouns
Names of people and places are proper nouns and are written with an initial capital letter (*James, Amanda, London, Calcutta, Japan*). This includes the names given to mountains (*Everest*), rivers (*Thames*) and animals (*Rover*). Unlike ordinary nouns, we do not use an article in front of many proper nouns (use '*London*' not 'the *London*'). Most proper nouns do not have a plural form (use '*London*' not 'the *Londons*').

Figure 1: Information about nouns (left) and 'proper nouns' (right).

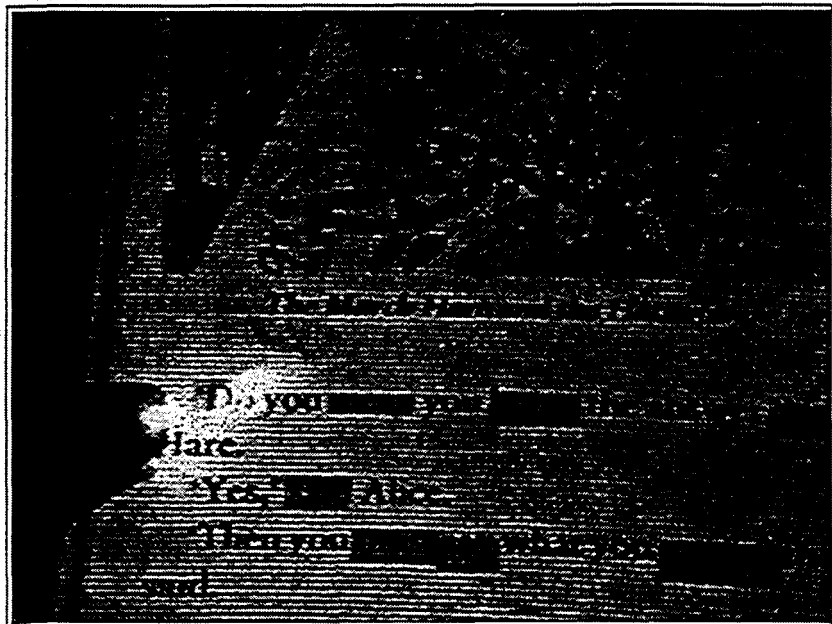| Pronoun – personal |
| --- |
| A pronoun is used instead of a noun (*she, we, his, everything, whom*)

'*her*' is a personal pronoun – one that refers to particular people or things (*I, them, yours*) |

Figure 2: Feedback on 'her'.

Figure 3: Words on the page have rectangles projected onto them.

grammar information to the 'real' text that the student is familiar with. In this way, the paper itself becomes an active part of the learning process.

Different versions of the lesson and quiz would be appropriate for different levels of knowledge. The version implemented is at an intermediate level; it contains approximately 50 different sets of information in the lesson. Adjectives, adverbs, determiners, nouns, pronouns and verbs are presented at the highest level. Information on prepositions, conjunctions, numbers, interjections, negatives, infinitives and possessives is available at a secondary level. Several of these word categories are subdivided when the user asks for more details (normal, comparative and superlative adjectives, for example, and separate categories for the verbs 'be', 'do' and 'have'). The BNC text uses 57 different word categories (Burnard, 1995), but not all of these are identified as separate categories in this lesson (adverb particles, for example, are never distinguished from general adverbs). However, the system has been designed so that the level of detail presented can easily be changed by altering the hierarchy of windows and menus, and by changing an internal grouping of the BNC categories into 'super-categories'.

It should be clear that, just as it is simple to alter the level of the grammar lesson provided, it would also be relatively straightforward to adapt the main features of this system to foreign-language teaching. The essential requirements are a book that has a corresponding electronic text containing part-of-speech information, and a DigitalDesk with the software to match the paper and electronic texts. Lessons and quizzes for different languages would need careful design – and different word categories might be needed for different languages – but the general method of involving the paper text in the lesson would be appropriate for all languages based on Latin scripts.

Using the TEI for learning material

This section looks briefly at a few of the more specialized features of the TEI guidelines that could provide valuable active paper applications for university students or even for professional linguists or historians.

In the applications suggested below, the information available would be much more complex than the part-of-speech information used in the grammar lesson. A word in a book might be associated with several different pieces of information and, conversely, a piece of information might refer to a paragraph or span of text rather than to a single word. Nevertheless, the users' interaction with the book would probably be based on the two basic mechanisms used in the grammar lesson: 'Show me . . .' menus to highlight all the areas on the open pages associated with a particular type of information, and pointing at a position on a page to see all the information associated with that point.

The printed versions used would normally be modern transcripts of old books and manuscripts.

Verse

The TEI provides many special facilities for verse. These include encoding mechanisms for line groups of different types (<lg type=sestet>), or to specify the pattern of rhyme in the group (<lg rhyme="ababcdcd">). Individual lines within a group are marked (<l>) and may have characteristics such as their metrical pattern specified (<l n=357

met=``-+ | -+ | -+``>). Segments (<seg>) within lines allow the metrical structure of the verse to be encoded down to the level of each foot or syllable (<seg type=foot met=``+-``>).

TEI texts containing this level of information represent a significant scholarly resource. Allowing direct access to them from a printed version would be invaluable to scholars or students of classical verse.

Variant versions

When multiple versions of important manuscripts exist, scholarly editions are often produced with a critical apparatus that records known variations attested by particular 'witnesses'. Typically, these identify one base or lemma version and a number of different readings provided by the different witnesses. A TEI text may record such information using an <app> (apparatus) element containing one <lem> element for the base version and one or more <rdg> elements for the other variants. Attribute values in the start tags identify the witness concerned, and give further information about the type of variation. Variations may apply to minor differences in the spelling of single words or to significant changes over a span of text. The following example from the TEI guidelines shows how variant spellings of a single word might be encoded.

```
<app>
<lem wit="El Ra2">though</>
<rdg wit="Hg" type="orthographic">thogh</>
<rdg wit="La" type="orthographic">thouh</>
</app>
```

A modern transcription of such a manuscript could be used on the DigitalDesk to identify all parts of the text with variant versions and to allow users to see either the base version or the variation attributed to one particular witness.

Old manuscripts

Many other features of old manuscripts may be recorded in TEI texts: the condition of the manuscript (damaged areas), a record of erasures and insertions in the original hand-written version, the level of certainty about the transcription of the text from hard-to-read areas, the colour of the ink used, changes in the 'hand' of the scribe, and so on. This may be supplemented with information provided by various scholars who have studied the text and have added corrections of apparent errors, expansions of abbreviations used, and suggested text to use in place of damaged areas.

Relating such information directly to the pages of a modern version could provide a valuable teaching and research tool.

Conclusion

Books, printed worksheets and other paper-based documents are likely to remain a key part of every student's learning environment. The challenge facing teachers and students is to integrate computer-enhanced activities into this environment.

We have shown that it is feasible to use the DigitalDesk as a means of accessing the information in TEI texts from ordinary printed texts. Although the existing system lacks the speed required for use in real learning situations, it is constructed from readily available components which are improving in performance all the time. One of the most time-consuming parts of the system, page recognition, can also be achieved in other ways (foe example, the barcodes regularly used for this purpose in industry), and a hybrid system could be developed in the future.

Acknowledgements

References

Alschuler, L. (1995), *ABCD . . . SGML: A User's Guide to Structured Information*, London: International Thomson Computer Press.

Burnage, G. and Dunlop, D. (1993), 'Encoding the British National Corpus', in Aarts, J., de Haan, P. and Oostdijk, N. (eds), *English Language Corpora: Design, Analysis and Exploitation*, Amsterdam: Rodopi, 79–95.

Burnard, L. (1995), *Users' Reference Guide for the British National Corpus*, Version 1.0, Oxford: Oxford University Computing Services.

Carroll, L. (retold by Bassett, J., 1994), *Alice's Adventures in Wonderland*, Oxford: Oxford Bookworms (OUP).

Goldfarb, C. F. (1990), *The SGML Handbook*, Oxford: OUP.

Harding, R. D., Lay, S., Robinson, P., Sheppard, D. and Watts, R. (1997), 'New technology for interactive CAL', *ALT-J, 5* (1).

Robinson, P., Sheppard, D., Watts, R., Harding, R. D. and Lay, S. (1997), 'A framework for interacting with paper', *Proceedings of Eurographics '97*, 16 (3).

Sperberg-McQueen, C. M. and Burnard, L. (1994), *Guidelines for Electronic Text Encoding and Interchange (TEI P3)*, Chicago and Oxford: ACH/ACL/ALLC (Association for Computers and the Humanities, Association for Computational Linguistics, Association for Literary and Linguistic Computing).

Wellner, P. (1991), 'The DigitalDesk calculator – tangible manipulation on a desk-top display', *Proceedings of the ACM Symposium on User Interface Software and Technology, Hilton Head*, November 1991.

Wellner, P. (1993), 'Interacting with paper on the DigitalDesk', *Communications of the ACM* 36 (7), July.