Some issues in the mark-up of handwriting in a learner corpus

Harold SOMERS
Centre for Computational Linguistics,
UMIST, PO Box 88
Manchester M60 1QD
harold.somers@umist.ac.uk

Abstract

In learner corpora, as in any corpus, mark-up is an important issue. One aspect of learner corpora so far largely ignored, however, is the specific question of handwriting and in particular how to mark-up handwriting anomalies, especially with learners whose native language uses a different writing system. In this paper we pose some open questions about what aspects of a learner's handwriting might be of interest and how these features can be marked-up in a learner corpus of handwriting. We begin by considering some general issues related to the acquisition of a second writing system, and consider whether and how handwriting anomalies should and could be marked up in a learner corpus. We consider the relevance of mark-up guidelines for manuscripts in general, and are guided by the importance of handwritten material in second language acquisition. A number of examples taken from a small corpus of English essays handwritten by Arabic-speakers are presented.

1. Introduction

Delegates at this Workshop are probably in agreement that Learner Corpora, that is, corpora of language produced by second or foreign language (L_2) learners, can be used to track linguistic features of the L_2 use, whether lexical, grammatical or stylistic. These may include over- or underuse of specific features, incidence of errors, and influence of the native or first language (L_1). Studies may be purely descriptive, may focus on data as evidence of psycholinguistic aspects of second-language acquisition (SLA),[1] or may be aimed more at influencing teaching strategies.

One interesting aspect so far ignored in the literature on learner corpora is **handwriting** as an aspect of SLA. In particular, where the learner's L_1 uses a different writing system (WS_1), acquisition and use of a second writing system (WS_2) may be an important area for research, again from various points of view: SLA, teaching strategies and contrastive analysis in general.

This paper looks in particular at the question of **mark-up** of handwriting in a learner corpus and tries to draw attention to some issues related to this question. Amongst the most important of these is the extent to which mark-up recommendations of bodies such as the Text Encoding Initiative (TEI, Sperberg-McQueen and Burnard 1994) can be applied to such corpora, and whether it is possible to annotate such a corpus independent of any analysis of it. As a case study, we will look at examples from a small collection of English essays written by Arabic-speaking learners, collected in connection with a (more conventional) study of grammatical and lexical errors.

2. Learner Corpora

The main goal of SLA research is, as Granger (1998a:4) says, "to uncover the principles that govern the process of learning a foreign/second language", and the main tools of this research are data from language use, metalinguistic judgments, and student introspection. Learner corpora directly meet this first need, by providing a body of, usually genuinely produced, examples of L_2 usage by students. There are of course drawbacks to the use of learner corpora, and, equally, issues regarding the design and collection of material, as there are with any corpus material. It is perhaps unnecessary to rehearse these here (see, again, Granger 1998a).

---

[1] A distinction is sometimes made between "second language" and "foreign language" learning, depending on whether the language in question is learned in its native habitat or not, "a crucial distinction which is too often disregarded in SLA studies" (Granger 1998a:9). However crucial, it is not a distinction which is felt to be relevant for the present study.

The corpus which forms the basis of our case study below is by all standards a tiny one. It is a collection of 20 handwritten essays, each about 150 words in length, produced by adult learners of English, all with (Cairene) Arabic as their L_1 and WS_1. The essays, on the topic of the computer as an educational tool, were written as a follow-up to a listening comprehension exercise, and therefore tend to repeat many of the phrases and ideas expressed in the original. The corpus was collected as data for a study of lexical and grammatical interference, not explicitly for the study of handwriting. It was the task of converting the corpus to machine-readable form which alerted the current author to the specific problem of marking-up handwriting anomalies. Figure 1 shows a typical example and is a source of examples below.

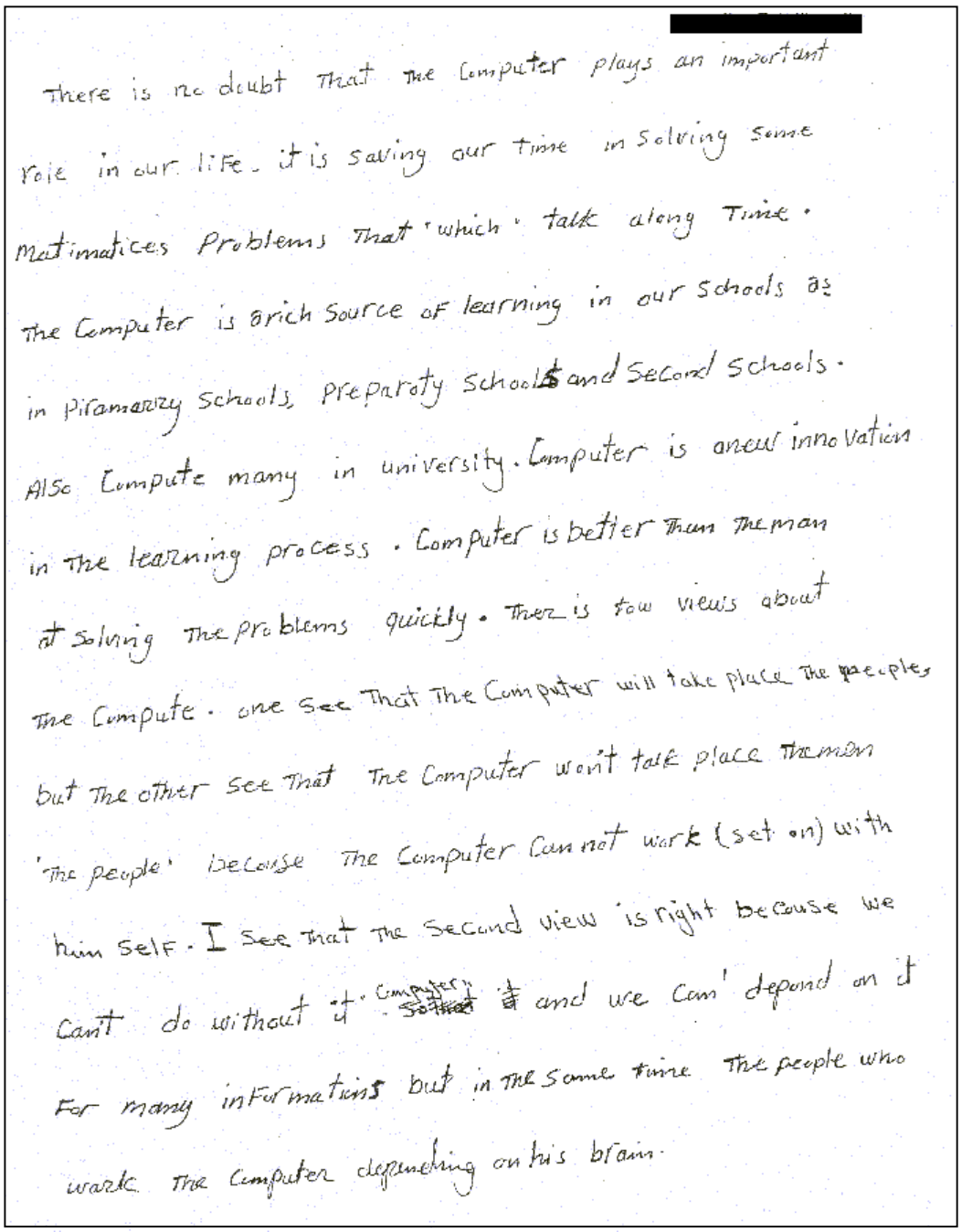
There is no doubt that the Computer plays an important role in our life. it is saving our time in solving some mathmatices Problems that 'which' talk along Time. The Computer is a rich source of learning in our schools as in Piramarry schools, preparaty schools and Secand schools. Also Compute many in university. Computer is a new innovation in the learning process. Computer is better than the man at solving the problems quickly. Ther is tow views about the Compute. one see that the Computer will take place the people but the other see that the Computer won't take place the man 'the people' because the Computer can not work (set on) with him self. I see that the second view is right because we can't do without it. Computer ~~Software~~ and we can' depand on it For many informations but in the same time the people who work the Computer depending on his brain.

Figure 1. An example of the raw corpus data.

Granger (1998a:11) notes that keyboarding is currently the most common method of data capture, and indeed the only method for handwritten texts. But this process presents special difficulties for proofreaders: "The proofreader has to make sure he [sic] edits out the errors introduced during keyboarding … but leaves the *errors* that were present in the

learner text, a tricky and time-consuming task" (*idem*, emphasis added). Granger's use of the word "errors" here raises an interesting point to which we will return.

Regarding mark-up, which can be used to record textual features of the original data, Granger worries that it is also very time-consuming, and suggests that "[i]n the case of learner corpora, which tend to contain few special textual features, this stage can be kept to a minimum, although it should not be bypassed" (*ibid*., p.12). We feel that this minimalistic approach is quite inappropriate for learner corpora, since one of the most striking features of WS_2 writing is that it contains anomalous language use that, if not annotated, might lessen the usefulness of the corpus. One obvious example is misspelling: consider that if a misspelled word is left unmarked in a corpus, its presence may interfere with subsequent attempts to analyse the corpus, whether this be by parsing or tagging, or more straightforward analyses such as concordances. Meunier's (1998) overview of tools available for corpus analysis makes no mention at all of the impact of anomalous usage on parsing, tagging and so on.

In contrast, Dagneaux et al. (1998) describe the use of learner corpora in connection with computerised error analysis in which errors in the student's text are tagged with an appropriate "error tag", which indicates the nature of the error as well as the (or rather, one of the) correct form(s).

The issue of learner's "errors" is one that should not be ignored, and is especially relevant for marking up handwriting. While some of the things that learners write are unquestionably errors, e.g. *matimatices* for 'mathematics' (Figure 1, line 3), often things are not so clear cut. One finds a whole range of grammatical infelicities, ranging from lack of number agreement, inappropriate prepositional complements, use (or omission) of articles, all of which might safely be tagged as errors, through to awkward uses of tenses, choice of near synonyms and other lexical matters, which may be more an issue of judgment. In marking up a learner corpus, it would be preferable if tags could differentiate between objective errors and anomalies, and others that more or less presuppose some analysis. For example, a tag of "wrong tense" would presumably reflect some analysis of what the student was trying to say, rather than any ungrammaticality as such. As we will see, when we come to consider handwriting, some of these issues are far from clear-cut.

3. Acquisition of a WS_2

Despite the large amount of literature in the field of SLA and in particular that of teaching English as a Second (or Foreign) Language, little has been written about "the equally important subject of how to acquire the Latin alphabet as a second writing system, or how to change from any particular writing system to another" (Sassoon 1995:5). The work from which this quote is taken seems to be an almost unique exception. For example, Swan and Smith's highly recommended collection (1987) of language-by-language essays on L_1 interference contains some examples of learners' handwriting, but generally says little about what to expect. Smith (1987:146f), discussing Arabic speakers, has a few paragraphs on "Orthography and punctuation" which includes reading, writing and spelling problems, as do Wilson and Wilson (1987:133) discussing Farsi. Thompson (1987:215) states merely that "Japanese learners do not generally have great difficulty with English spelling or handwriting", while Chang (1987:227) states that "Alphabetic handwriting … presents no serious problems for Chinese learners".

Sassoon's focus seems to be mainly on children learning English as an L_2, and her somewhat anecdotal approach is aimed at helping teachers develop strategies for overcoming problems which could be symptomatic of, or conversely the trigger for, deeper problems of linguistic and cultural assimilation. Nevertheless, her book is a good source of typical problems, with illustrations from learners with a wide variety of WS_1s.

Sassoon's book starts usefully by comparing the "rules" of writing systems, including elements such as general direction of the writing, entry point and direction for individual letters ("ductus"), heights of letters and their composite parts (ascenders and descenders), alternate letter forms (upper and lower case, word-initial, medial and final), and spacing and joining of letters and words. In each of these aspects there is the possibility of interference from the WS_1, if it differs. For some writing systems there is the additional problem of "false friends": in this case letters which look similar but have a different value (e.g. Greek 'P' with the phonetic value [r] – but cf. also Cyrillic lower-case 'т' which in handwriting resembles an 'm'). In other cases the interference is more generic, as exemplified (p. 46) by the problems Tamil writers have with joined-up writing, since their WS_1 prescribes a clockwise ductus, whereas letter forms in English often require the opposite.

Sassoon then goes on to look at a number of interesting case studies, which are informative, but her treatment of them is not systematic. Subsequent chapters focus on writing materials and posture, assessment, teaching techniques, psychological, and sociological aspects of handwriting, and, finally, typography. Although of interest, little of this is especially helpful to us in our quest for guidelines for marking up a handwritten corpus.

4. Mark-up Recommendations and Manuscripts

There is a considerable literature on the electronic mark-up of manuscripts, mostly on the World Wide Web. Much of this work is related to more or less ancient documents, though original manuscripts of modern literary works are also subject to this kind of attention. In most cases, researchers look to the TEI guidelines for some basic suggestions, and agree that SGML-type mark-up is appropriate. Often, transcribers take it upon themselves to regularise features of the original text, to make them more readable to modern scholars (e.g. Hines 1995). Most researchers find that they have to extend the TEI recommendations to meet their specific needs. The following extract regarding the work of the Electronic Text Center, University of Virginia, is typical:

> A primary goal of documentary editing is to preserve as many features of the original document as possible. To this end, we carefully transcribe each page, noting and preserving such features as line breaks, underlining, post-scripts scrawled in margins, changes in hand, and so forth. TEI includes a number of tags that enable an editor to describe these textual and non-textual features. For instance, we record information about the content and location of additions and deletions with the <add> and <del> tags, and we mark errors in the text and editorial emendations with <sic>, <corr>, and <orig reg>. […]
>
> Even as we strive to replicate the original document as accurately as we can, we also want the text to be accessible to as many users as possible, for as many uses as imaginable. Of course, simply putting the text and its accompanying images up on the web makes a rare, unique document available to millions of users. These texts are fully searchable, so that scholars can discover connections among documents that were previously unknown. (Spiro and Fay, 2001)

An interesting example of some relevance to us is the Lancaster Corpus of Children's Project Writing, which is a corpus of transcribed texts written by 8–12-year-old children, freely available on the Web.[2] The corpus includes images of the original material, which allow us to see how handwriting anomalies have been marked up.[3]

In this section, we aim to summarize the TEI recommendations on manuscript mark-up, as described in Burnard and Sperberg-McQueen (1995), and to consider how these guidelines relate to our present problem.

In a section headed "Editorial Interventions", the TEI guidelines distinguish between "correction", where the editor "believes the original to be erroneous", and "normalization", or "changes introduced for the sake of consistency or modernization of a text". In the former case, one can either mark something as corrected, with the original indicated as an attribute of the `<corr>` tag, or, conversely a self-explanatory `<sic>` tag is proposed, with the attribute `corr`. In either case, additional attributes can identify the editor responsible for the annotation, and the degree of certainty of the correction. Similarly for normalization, one can indicate the original, with the correction as an attribute (`<orig>` with attribute `reg`) or the converse. We can illustrate these tags in relation to Figure 1, the first three lines of which might be tagged as in (1) (we show here only the tags discussed so far, with the addition of `<lb/>` to indicate "line break". For the sake of illustration only, we use `<sic>` for clear-cut errors and `<corr>` for cases where we judge the text to be anomalous, though in doing so we are not necessarily advocating this distinction.

(1)
```
There is no doubt that the computer plays an important<lb/> role in
our life. <sic corr="It">it</sic> is saving our time in solving
some<lb/> <sic corr="mathematics">matimatices</sic> problems that
"which" <corr orig="talk">take</corr> <corr orig="along">a
long</corr> time.<lb/>
```

The TEI guidelines also suggest that additions, omissions, and deletions can be indicated, using the tags `<add>`, `<gap>` and `<del>`. The first of these is for text inserted, and includes attributes indicating the manner of the insertion. Figure 2 shows an obvious case where this tag could be used. In this case, the text might be marked up as in (2).

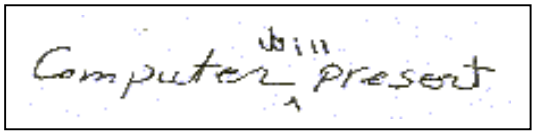

Figure 2. An example of an insertion.

(2)
```
<add addtype="superlinear">will</add>
```

The `<del>` tag is of certain interest to us. It is used to indicate "a letter, word or passage deleted, marked as deleted, or otherwise indicated as superfluous or spurious…". Attributes useful to us include the type or manner of deletion. Looking again at Figure 1, we see in line 5 that a correction has been made by overwriting (*schoold* changed to *schools*), while three lines from the bottom the words *so that it* have been crossed out and the word *"computer"* in quote marks (see below for discussion of punctuation marks) added. It is certainly a matter of consideration whether deletions should be rigorously recorded, but there is no doubt that certain

---

[2] See http://www.ling.lancs.ac.uk/lever/index.htm

[3] Ormerod and Ivanič (2000) discuss in detail the importance of non-textual "physical" characteristics of children's work, though without any reference to corpus mark-up.

kinds of self-correction can nevertheless be very revealing, as in Figure 3, which shows two examples of corrections which could indicate to a teacher or researcher a pattern of potential error: the top example shows a possible confusion regarding English morphology, while the bottom example reflects a typical Arabic-speaker's phonological error reflected as a spelling error. No matter that the learner in each case spotted and corrected the mistake. The appropriate tag in this case is presumably `<del>`, again perhaps with an attribute to indicate the manner of the deletion, as in (3) and (4), illustrating alternative possibilities.

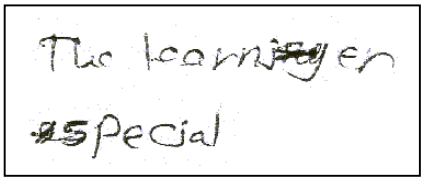

Figure 3. Two examples of revealing corrections.

(3) `learn<del type="linethrough">ing</del>er`

(4) `<del rend="/">e</del>special`

Although, as we can see, the TEI guidelines give us some initial ideas, there remain a number of interesting issues in the mark-up of learners' handwriting that remain to be addressed.

5. Marking up Learner Handwriting

Perhaps a good place to start is to consider what elements of learner handwriting might be of pedagogic interest. In other words, what kinds of things might we want to mark up? Once again we note the tension between the notion of "error" and, for want of a better word, "anomaly", which can apply at all levels.

At the most abstract level, and of comparatively little interest to us here, are aspects of the text that go beyond the question of orthography and calligraphy. Style, syntax and lexical choice, for example, are of course of interest to the researcher, but do not generally relate to the question of handwriting. As we have seen, deletion and insertion may be revealing, but are well treated in existing TEI guidelines.

Our first point of contact might be **spelling**. Although at first sight this might seem straightforward, there are some interesting interactions between orthography and handwriting. Look again at Figure 2, this time concentrating on the words *computer* and what we assume to be *present*. On close inspection, the *n* of *present* closely resembles the *r* of *computer*; so how do we know that it is not a misspelling? Our judgment is guided by the plausibility of that error for this student, and also perhaps of our knowledge of the student's WS_1. Figure 4 shows, in the space of just three lines, some of the difficulties facing us. On the first line, is that *invented* or *intented*? Is the second *m* in *humman* crossed out? Is that *depend* or *depond* or even *dopond*? And what exactly is the fourth word of the last line? These are of course the kinds of decisions that teachers have to make when assessing students, but our purposes are somewhat different. Nevertheless, we probably use the same strategies, looking elsewhere in the text to see whether the student has made similar mistakes.

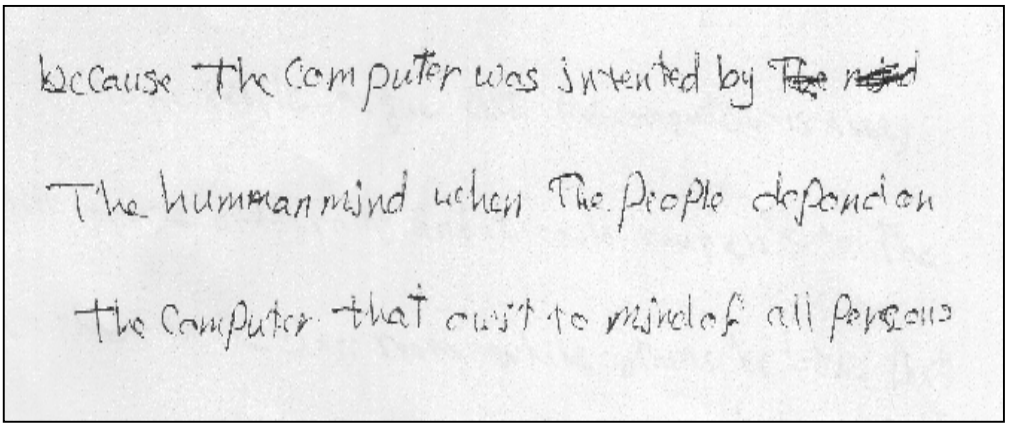

Figure 4. Some difficult decisions.

Spacing and **punctuation** is another major issue in our particular corpus. In Figure 1 we can see that the student consistently leaves a space before a full stop, which tends to be somewhat elevated. Should this be marked up or not? In the same example we see (line 4) *arich*, and (line 6) *anew* written with comparatively little space: is this significant? In this case we might be influenced by the fact that we find a similar phenomenon in other students' writing. Another feature that we find in Figure 1 in several places is an unorthodox (in terms of the target language) punctuation: quote

marks (lines 3, 11 and 13) are used as parenthetical markers here and also in other students' work. But this student also uses conventional brackets (line 11), though with unclear significance. One is reminded of the practice, taught in British primary schools, of indicating deletions with brackets and an 'X', *(thus)x*, since crossing out is for some reason discouraged.

We consider next perhaps the most basic aspect of writing, namely letter shape and choice among alternate forms, or "allographs" (Sampson 1985:25). In the Latin writing system of course, there is a significant distinction between upper and lower case, but also, especially in handwriting, a number of insignificant distinctions between alternate forms. Figure 2 for example shows two forms of the letter *r*. To complicate matters further, for just under half the letters (give or take one or two borderline cases), the difference between upper and lower case is simply a matter of scale: compare *C* and *c*, *O* and *o*, *V* and *v*, etc. In this respect our small corpus turns out to be full of difficult cases for mark-up, and not surprisingly since the WS_1 differs from the WS_2 particularly in this respect: Arabic has different letter forms for initial, medial and final, but no upper–lower case distinction. Both Wilson and Wilson (1987) and Smith (1987) mention this as a problem.

Figure 5 shows an example of a student who consistently uses a large letter form for word-initial 'C', but does use a smaller letter elsewhere (though cf. *sourCe* in line 5, and *proCess* in line 8). Some of the 'P's are rather large too. Should we mark this up, and if so, how?

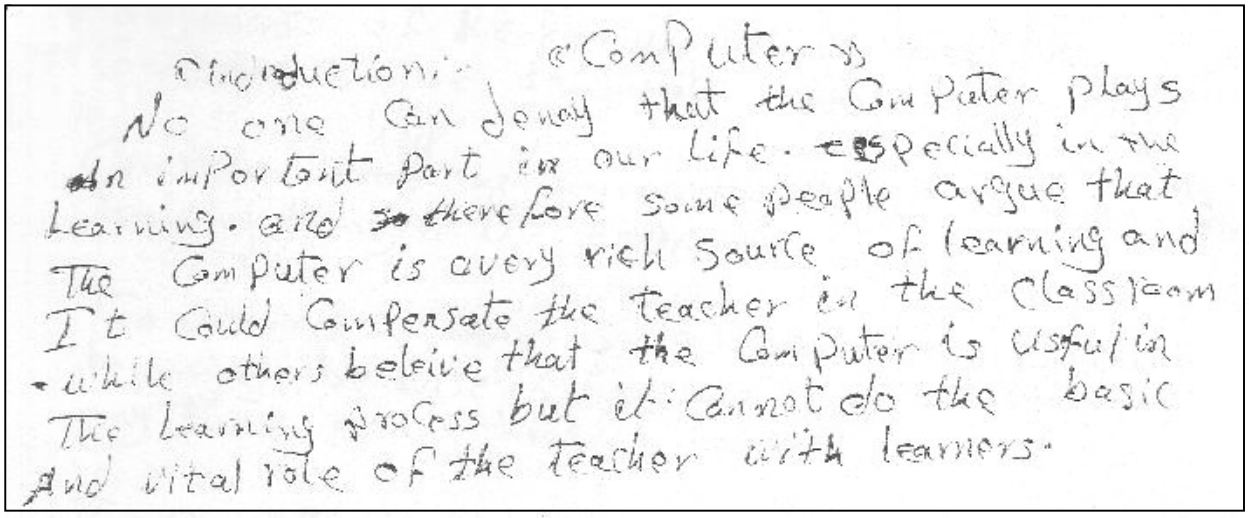

Figure 5. Failure to distinguish upper- and lower-case 'C'.

Figure 6 shows a similar case which extends to 'C' and 'W'. In this example, there is a more consistent pattern of using large letter shapes word-initially and smaller shapes elsewhere. Apart from that, the student's handwriting is comparatively neat, and the standard of English quite good. All of these factors could influence us to resist marking the letter forms as capitals.

A more intriguing case is illustrated in Figure 7. Here we can see three or four distinct 'T' shapes. The 'T' of the first word, *Teacher* is most like a conventional capital 'T'. In *cannot*, *vital*, *it's* and the second 'T' of *that*, we have a clear-cut lower-case letter. But the remaining cases are somewhat hybrid, with the crossbar on top of rather than cutting the vertical stroke, characteristic of the upper case variant, and the short exit stroke which suggests the lower case. And the 'T' in *computer* is more like a capital 'T', but slanted backwards more like the other lower-case variants. In fact this student shows a systematic variation (also in the remainder of his essay, not shown here) where what we have called the hybrid form is used word-initially but not sentence-initially. The 'T' in *computer* was probably meant to be lower case.

Notice that in all three cases we have been able rather easily to spot a more or less consistent pattern. Should this analysis of the data be reflected in our mark-up? And if so, would it not make more sense to capture the facts in the document header rather than marking up each individual case? One way to do this would be to define separate **entity references** as in (5). In this way one could also distinguish inappropriate use of apparent capitals, as in (6), illustrated in (7), showing the end of Figure 6, line 3 in transcription.

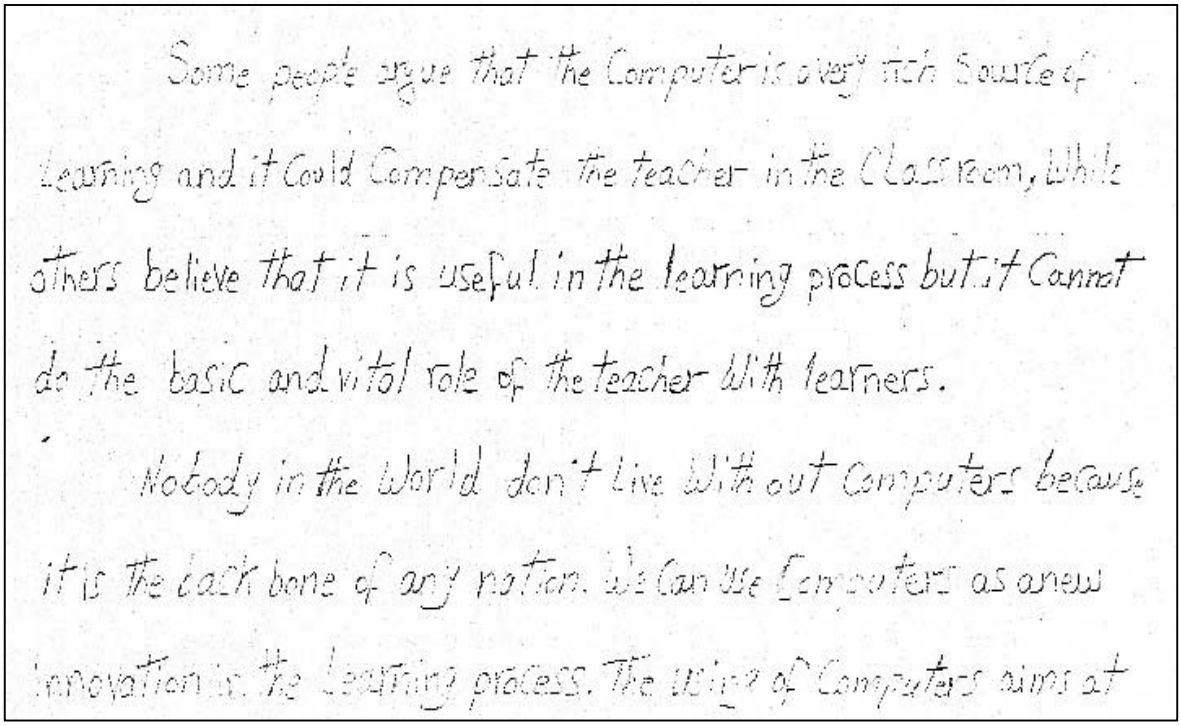

Figure 6. Consistent use of large letter shapes word-initially.

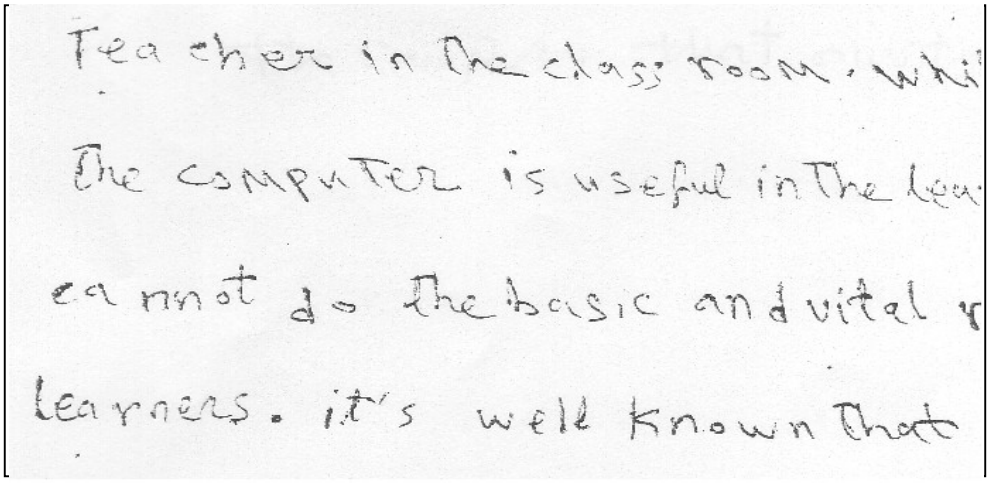

Figure 7. Three different 'T' shapes.

(5) `<!entity t1 '[medial-T]'>`

(6) `<!entity c1 '<sic corr="c">C</sic>'>`

(7) `in the learning pro&c1;ess but it &c1;annot ...`

One final example is a clear case where a global comment in the document header would be most appropriate. Figure 8 shows an example which is unremarkable except for the somewhat idiosyncratic 'R' shapes, which consistently extend over the following letter. This is clearly not ambiguous, or malformed in any sense, but may be something that one might want to reflect in the markup.

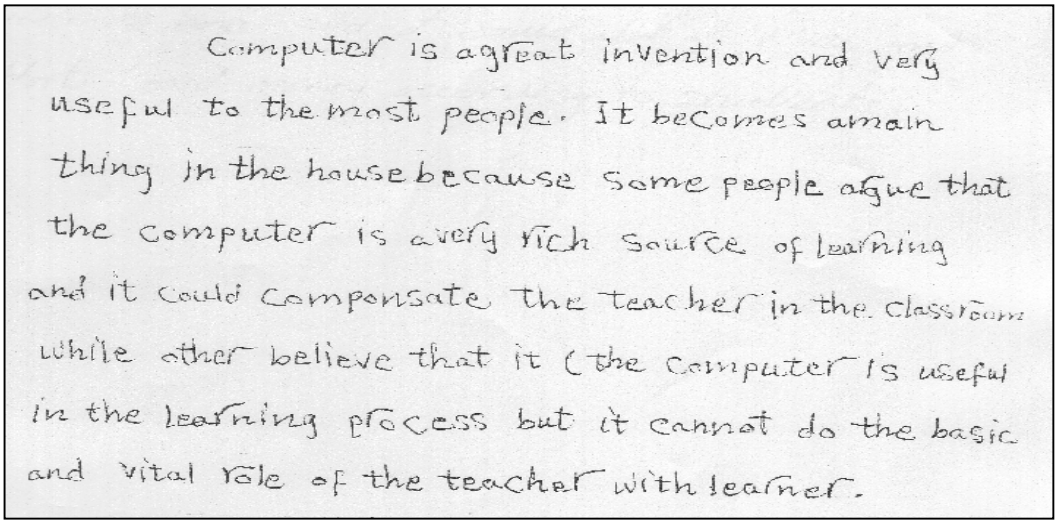

Figure 8. idiosyncratic but consistent R *shapes.*

Conclusions

Our aim in this paper has been to raise some issues, rather than provide answers. It is apparent that we can mark up whatever we like, but how should we be guided? At one extreme, we could try to mark everything imaginable, so that the text could be more or less reconstituted on the basis of the mark-up, a bit like ball-by-ball scoring in cricket or baseball. More practically, we might be guided by the use to which the mark-up is going to be put, in which case we could not separate the notion of mark-up from the analysis of the corpus. Is that a bad thing?

Whatever one decides, notice that there are other areas where these issues might arise, with different decisions, for example corpus-based research on WS1 handwriting with young children, aphasics, etc.

Acknowledgments

I am grateful to Hossam Moharam for making his data available to me, to Lou Bernard for a number of discussions on this topic, and to Sylviane Granger for extensive comments on an earlier draft. Any errors and infelicities are mine and mine alone.

References

Burnard L. and C.M. Sperberg-McQueen 1995 TEI Lite: An Introduction to Text Encoding for Interchange. Document No: TEI U 5. http://www.hcu.ox.ac.uk/ TEI/Lite/teiu5_en.htm (dated June 1995; retrieved August 2001).

Chang J 1987 Chinese speakers. In Swan and Smith (1987), pp. 224–237.

Dagneaux E., Denness S., and S. Granger 1998 Computer-aided error analysis. *System: International Journal of Educational Technology and Applied Linguistics* **26**: 163–174.

Granger S 1998a The computer learner corpus: a versatile new source of data for SLA research. In Granger (1998b), pp. 3–18.

Granger S (ed.) 1998b *Learner English on Computer*. London, Longman.

Granger S. and J. Hung (eds) 1998 *Proceedings of the First International Symposium on Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching (14–16 December 1998)*. The Chinese University of Hong Kong.

Hines P 1995 The Newdigate letters. *ICAME Journal* **19**: 158–161.

Meunier F 1998 Computer tools for the analysis of learner corpora. In Granger (1998b), pp. 19–37.

Ormerod F. and R. Ivanič 1999 Texts in practices: Interpreting the physical characteristics of texts. In Barton D, Hamilton M. and R. Ivanič (eds) *Situated literacies: reading and writing in context*, London: Routledge, pp. 91–107.

Sampson G 1985 *Writing systems*. Stanford, California, Stanford University Press.

Sassoon R 1995 *The Acquisition of a Second Writing System*. Oxford, Intellect.

Smith B 1987 Arabic speakers. In Swan and Smith (1987), pp. 142–157.

Smith N., McEnery T. and R. Ivanic 1998 Issues in transcribing a corpus of children's handwritten projects. *Literary and Linguistic Computing* **13**: 217–225.

Sperberg-McQueen C.M. and L. Burnard 1994 *Guidelines for Electronic Text Encoding and Interchange*. Chicago and Oxford: Text Encoding Initiative.

Spiro L. and C. Fay 2001 Procedures for Transcribing and Tagging Manuscripts. http://etext.lib.virginia.edu/tei/DocEdit.html (13 August 2001)

Swan M. and B.Smith (eds) 1987 *Learner English: A teacher's guide to interference and other problems*. Cambridge, Cambridge University Press.

Thompson I 1987 Japanese speakers. In Swan and Smith (1987), pp. 212–223.

Wilson L. and M. Wilson1987 Farsi speakers. In Swan and Smith (1987), pp. 129–141.