Compiling a corpus
techniques, problems and solutions

James, G.C.; Davison, R.M.; Fang, A.C.Y.

Presented: 01/12/1991

Document Version:
Post-print, also known as Accepted Author Manuscript, Peer-reviewed or Author Final version

Publication details:
James, G. C., Davison, R. M., & Fang, A. C. Y. (1991). *Compiling a corpus: techniques, problems and solutions*. 7th International Language in Education Conference, Hong Kong.

Davison, R. M. (1991). *Compiling a corpus: techniques, problems and solutions*. 7th International Language in Education Conference, Hong Kong.

ILE CONFERENCE

Compiling a corpus: techniques, problems and solutions.

The aim of this project is to compile a corpus of one million words in the field of Computer Science, as already explained in the rationale to the project. Computer Science was chosen because it is the only subject taught in not only HKUST, but also City Polytechnic and HKU. Texts for sampling have been chosen from First Year reading lists, though these are often rather meagre in quantity, partly because the HKUST courses are still under development. Originally, it was hoped that it would be possible to include `entire books' in the corpus, but for a number of reasons, this has proven impossible. Therefore, the number of words taken from each text has been limited to approximately 6,000. These are taken in 2,000-word chunks, selected at random from the first, second and third 'thirds' of each book. Random numbers are generated using *Lotus 1-2-3*'s *@rand* command and this number is incorporated in a simple formula to convert it from a random number between 0.1 and 0.9 as created by *Lotus* to a number that is appropriate for the text involved. For example, if a book has 300 pages excluding index, preface and bibliography, then the three sections will be from pages 1 to 100; pages 101 to 200; and from page 201 to 300. Thus, one random number must be produced for each section. The formula is *(@rand*(end_of_range - start_of_range) + start_of_range; i.e. (@rand*(100-1)+1), => (@rand*99)+1*. It should also be noted that *Lotus* generates random numbers to 15 decimal places, and therefore the printed output is limited to 0 decimal places. When the random number has been calculated, a chunk of text starting with the first complete sentence from the top of that page and ending with the end of the sentence at or past 2,000 words, is photocopied, scanned and transferred to *WordPerfect* for *spell-checking* and *proof reading*. The final stage is essential, since the accuracy of the scanner is variable, producing transcriptions with anything from 50% to 95% accuracy, depending on the quality of the original print and the font used.

This, then, is the basis for text sampling, extracting and transferring, but this is by no means the whole story.

The first problem that we encountered was: what exactly is Computer Science? Looking at reading lists, we realised that some of the courses that Computer Science students have to study rove into the realms of psychology, management information systems, economics and general business. A decision was made, therefore, to attempt to restrict texts, as far as reasonably possible, to those that are more specifically "computer science" oriented in the traditional sense and Computer Science has been taken to cover such fields as: networking, databases, programming languages, artificial intelligence, expert systems, ... and other subjects that incorporate these areas to a significant extent. Inevitably, rational choices have to be made in order that protocols can be established.

The texts that have been selected are by no means just collections of formulae and diagrams, though sometimes these constitute a major part of the book. Yet, it is

certainly true that it is hard to find a passage with 2,000 words of continuous text and no interruptions. Consequently, whenever text is interrupted by formulae, programming language and diagrams, these are simply ignored and the text is allowed to jump over the obstacles. Sometimes, however, where there are many diagrams, up to twenty pages of the book may be required before the 2,000 word total is reached.

There are several other problems that have arisen, all requiring protocols, since many of them occur frequently.

Hyphenation: When a word runs over the end of the line, it will be hyphenated, but if a hypenated word runs over the end of the line at the point of the hyphen, then should there be another hyphen at the start of the next line? Actually, no text that has been scanned so far has used this helpful practice, and so it is sometimes difficult to decide if words are hyphenated or not. The protocol adopted here is that if a word is obviously hyphenated, then it is left hyphenated; if it is not obviously hyphenated, but a hyphenated example has previously occurred, then it is left as hyphenated; in other cases the hyphen is omitted. There is clearly much room for error here, but all we can do is minimise this as far as is humanly possible. Further complications arise when words exist in both hyphenated and un-hyphenated forms, e.g. *enduser*, *end-user*. Neologisms: In this context, there is also a large number of words that are not recognised by the *WordPerfect* spellchecker: some of these are legitimate, if new words, such as: *computerese*, *precompunications*, *exponentiation*, *inviolated*, *microoperations*, *misrequisitioning*,... and some are just typographical mistakes in the original, e.g. *muct* ("much"). All have to be retained, since it is considered that students will have to cope with these 'words', badly printed or otherwise in reality. An example of computerese is: The proof comes from hardware design, where we learned, three decades ago, that seat-of-the-pants logic was buggy, slow, dangerous, and hard to build, test, and maintain. Non-standard words: Another source of non-standard words is programming language. As mentioned above, running computer-language text is omitted, but where it is crucial to the sense of the sentence, it is retained and so words such as: *writeln*, *addins*, *lookaside*, *goto*,... may be found. Equally alphanumerical combinations and numbers are retained to preserve sentence structure. Where they are removed, along with formulae, a suspension mark '...' is used to indicate absence of original material.

> It is incremented by 1: [(PC) ← (PC) +1] Or it is replaced: {(PC) ← operand When in the incrementation mode, the arithmetic is modulus 1000, that is 999 + 1 → 000.
>
> This is reduced to: It is incremented by 1:... Or it is replaced:... When in the incrementation mode, the arithmetic is modulus 1000, that is, 999 + 1 → 000.

Typefaces and other formatting features are all reduced to a single common format - Roman 10cpi - since ultimately all text will have to be reduced to ASCII code before concordancing can take place. This does mean that emphasis, with italics or bold, is also lost and so reference to the original text (photocopies are preserved) will be necessary for exact verification in some cases.

Footnotes and endnotes (or foot-notes and end-notes) have been removed, along with their indicators, since it is felt that as they do contain non-textual material on occasion and as a universal protocol should be adopted for them, they cannot legitimately be

retained. Appendices are kept however, as are glossaries, since these provide explanations of text/jargon that a student would have to understand. Equally, acronyms are kept.

There now follow two extracts from randomly sampled Computer Science texts. These illustrate the kind of non-Computer Science language that is occasionally encountered:

> I want a man who knows what love is all about. You are generous, kind, thoughtful. People who are not like you admit to being useless and inferior, John. You have ruined me for other men. I yearn for you. I have no feelings whatsoever when we're apart. I can be forever happy. Will you let me be yours? Gloria.
>
> UPS's corporate culture has been described as a "cross between the Mormons and the Marines" and a "half Marine Corps and half Quaker meeting." Drivers must keep their hair short and their pants properly creased. Beards and flowing mustaches are taboo, as is drinking coffee or other beverage [sic] at your desk.