THE GILL CORPUS

J.M. Gill*

The corpus of text was created as part of a study on braille contractions. The braille system, in England, utilises 190 abbreviations and contractions. The use of those contractions is governed by a complex set of rules, many of which depend on pronunciation or meaning of the word. The project involved studying the effect of these contractions on:

(i) ease of learning

(ii) reading speed

(iii) writing speed

(iv) space saving

(v) ease of production.

Space saving is a significant factor since braille is very bulky. For instance one copy of the Bible takes up 72 volumes (about a quarter of a cubic metre). Ease of production is also a factor because of the increasing use of computers to translate the ink print to contracted braille. A reduction in the number of rules dependent on pronunciation or meaning would reduce the cost of producing braille by computer-assisted methods.

---

* Warwick Research Unit for the Blind, University of Warwick, Coventry, England.

One part of this project involved measuring the frequency of use of text strings. The Brown corpus was used but it had the disadvantage that it was not typical of the material read in braille by blind people in the U.K.

Therefore a new corpus was created in an attempt to be more representative. Some of the material was in the form of short documents which had been requested by blind people for transcription into braille. These could be divided into:

Agendas and minutes	15%
Instruction booklets	6%
Employment	14%
Students' handouts	11%
Leisure (e.g. record sleeves)	7%
Songs and poems	1%
Recipes	2%
General Information (e.g. government leaflets)	9%
Religious	1%
Timetables	1%
Accounts	4%
Correspondence	8%
Miscellaneous	21%

Added to this were samples of short stories and books, both fiction and non-fiction, which had been transcribed into braille by the Royal National Institute for the Blind.

The corpus contains 1030 short pieces in the English language giving a total of 2,561,308 words (a word being defined as an alphanumeric character string delimited by spaces or punctuation).

The corpus is available, for research purposes, on digital tape

from Louis Bernard, The Archive, Oxford University Computing Service, 13 Banbury Road, Oxford OX2 6NN, England.