

Real-word error corpus – brief documentation

This is a corpus of dyslexic real-word spelling errors compiled as part of my PhD Research. It contains approximately 12,000 words, just under 700 sentences and 833 marked-up errors.

The corpus comprises nine files in text format (Windows). Each sentence starts on a new line with the errors marked as in the example below:

... the <ERR TARG=material> martial </ERR> on each slide ...

Full details of the compilation and content of the corpus can be found in my PhD thesis available for download from <http://www.dcs.bbk.ac.uk/~jenny/publications.html>.

Brief details of the files are given in the table below.

File	Sentences	Tokens*	Words	Errors	Description
A	27	541	533	45	Homework by a dyslexic child, aged approx. 12 years.
B	39	612	594	50	Compositions written by school leavers in the 1960's.
C	22	347	335	26	Office documents written by a dyslexic office worker.
D	29	397	387	39	Online typing experiment.
E	37	614	600	51	Dyslexic student studying IT NVQ.
F	124	2786	2759	134	Undergraduate essays written by a dyslexic student.
G	10	132	128	14	Stories written by dyslexic primary school child.
H	202	3167	3117	247	Dyslexia mailing list.
J	184	3428	3386	227	Dyslexia bulletin board.
All	675	12024	11839	833	Complete corpus

* The token count represents the number of strings in each file. Not all strings correspond to orthographic words, for example the text includes numbers and dates.

Please let me know if you find this corpus useful or if you find any errors in its compilation.

Jenny Pedler
jenny@dcs.bbk.ac.uk
Birkbeck, London University
September 2009