

Collaborative Text Correction by the Public: Summary of activity in the Australian Newspapers beta service in the first 3 months after public release (4 August – 3 November 2008)

Author: Rose Holley (Manager - ANDP)

Version: 1.0

Date: 30 January 2009

Users have actively been correcting OCR text since day 1 of beta release. There has been no time of the day or night when text correction has stopped since launch of the service. The basic ability to correct text has been given in the beta version, but no advanced power user mode to correct text or to gather statistics is available at present.

Text correction is being measured by number of lines, and by number of articles corrected. It is not possible to gather automated statistics on % correctness of articles before and after public text correction. The statistics therefore show number of text edits rather than 'corrections'. We assume that edits are making the text more correct.

In the first 3 months 868 registered users have corrected text and approximately 390 unregistered users (total of 1200 text correctors). This means that 58% of registered users are correcting text. 720,795 lines of text have been corrected within 50,887 articles. The top text corrector has corrected 59,000 lines of text within 1890 articles. Some articles have had corrections added by more than 7 users (e.g. articles in the first Australian newspaper the 1803 Sydney Gazette). This particular issue in its entirety has had several different users working on corrections (because it is difficult to read and is an important paper).

Fig 1: Summary of public OCR correction figures 4 Aug – 4 Nov 2008

Total number of lines corrected by public	720,795 lines
Total number of articles that have had text corrected by public	50,887 articles
Number of articles corrected by anonymous users	15,732
Number of articles corrected by registered users	35,155
Number of registered users doing text correction	868
Number of unregistered users doing text correction (approx)	390
Total approx number of users doing text correction	1200
% of registered users doing text correction	58%
% of correction being done by registered users	73%
% of overall users doing text correction	Unknown
Highest number of articles corrected by a single registered user	1895
Highest number of lines corrected by a single registered user	59,316
Most edited article	410 edits
Number of articles edited by multiple users with logins	1264

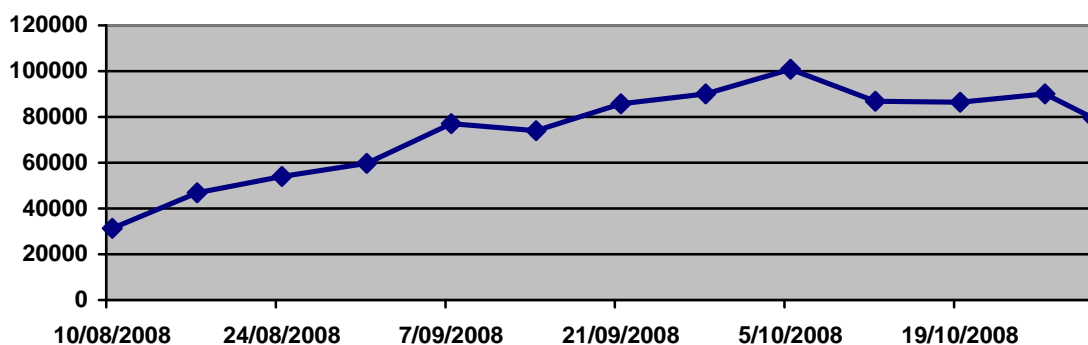
Fig 2: Greatest amount of corrections by month by line (top 5 correctors)

August 2008	Lines corrected
jhempenstall	23,623
Mrbh	9,673
Camerong	6,277
Scmorris	5,623
Maurielyn	5,372
September 2008	Lines corrected
Maurielyn	28,111
Jhempenstall	23,623
Mrbh	23,139
Fwalker13	13,141
Cmdevine	8,748
October 2008	Lines corrected
Maurielyn	24,538
Jhempenstall	23,265
John F Hall	21,886
Fwalker13	19,714
Cmdevine	16,896

Fig 3: Top 3 correctors over 3 month period 4 August – 4 November 2008, by line and by article.

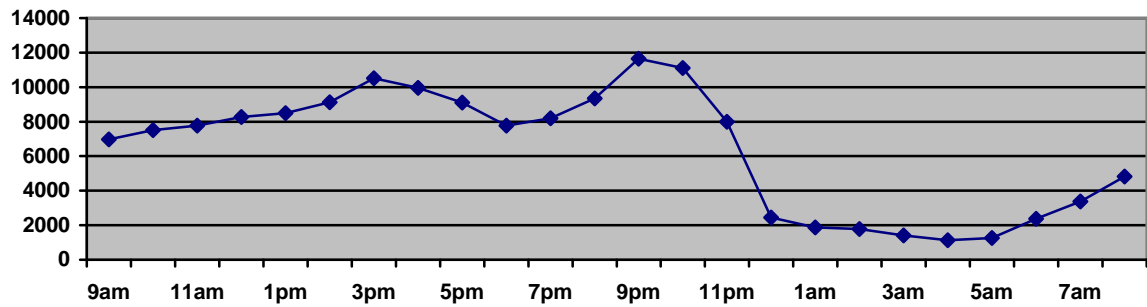
Identity	Total Number of lines corrected	Total Number of articles corrected
Jhempenstall	59, 316 lines	1,895 articles
Maurielyn	58, 021 lines	1,081 articles
Mrbh	46, 488 lines	971 articles

Fig 4: Number of lines corrected by week August – November 2008



An average of 74,045 lines are corrected per week. Highest is 100,772 in a week, lowest is 31,390 in a week.

Fig 5: OCR corrections by time of day 4 August 2008 – 3 November 2008 (number of times 'save OCR corrections' is clicked).



OCR correction rises steadily throughout the day peaking at 3pm and 9pm (with a small dip around from 6-7pm as users get their evening meal), and surprisingly continues throughout the night (though this may also be overseas users). OCR correction is occurring 24 hours a day.

Fig 6: Sample of OCR correction activity on an individual article 2 November 2008

Time	User	Activity
12:23	user 1	Corrects text in the article but doesn't know how to insert the pound symbol
12:24	user 1	Creates a comment on the same article saying that he doesn't know how to enter the pound symbol.
19:13	user 2	Enters the pound symbol in the article.
23:09	user 3	Spots an error in the correction by user 1, and fixes it.