How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs.

October 2008.
[Pre-Press. Article to appear in D-Lib Magazine, March/April 2009, vol. 15 no 3/4. URL: http://www.dlib.org]

Rose Holley,
Manager – Australian Newspaper Digitisation Program
National Library of Australia
rholley@nla.gov.au

Abstract
This article details the work undertaken by the National Library of Australia Newspaper Digitisation Program on identifying and testing solutions to improve OCR accuracy in large scale newspaper digitisation programs. In 2007 and 2008 several different solutions were identified, applied and tested on digitised material now available in the Australian Newspapers Digitisation Program beta service http://ndpbeta.nla.gov.au/ndp/del/home. This article gives a state of the art overview of how OCR software works on newspapers, factors that effect OCR accuracy, methods of measuring accuracy, methods of improving accuracy, and testing methods and results for specific solutions that were considered viable for large scale text digitisation projects.

---

1. Introduction

Optical Character Recognition (OCR) software was first used by libraries for historic newspaper digitisation projects in the early 1990's. An experiment at the British Library with the Burney collection and a co-operative project in Australia (the ACDP) [1] with the Ferguson collection were both considered unsuccessful largely due to the difficulties with OCR technology and historic newspapers [2]. However, by 2008 both the British Library and the National Library of Australia had established large scale historic newspaper digitisation projects and felt that OCR technology had reached an acceptable level to offer good full-text searching. In addition many other national libraries, newspaper publishers, and commercial companies such as ProQuest had also embarked on newspaper digitisation in a serious way, moving from small pilot projects into mass scale digitisation. The Australian Newspaper Digitisation Program (ANDP) [3] decided that 'acceptable' OCR was still not good enough and has therefore spent some time investigating ways to improve OCR accuracy that could be applied to other large scale newspaper digitisation programs as well as its own.

2. Understanding how OCR Software works on newspapers

Before we can begin to look at measuring or improving OCR accuracy, we need to gain a good understanding of how OCR software works on newspapers in 2008. There were significant developments in OCR software technology between 1993 and 2005, notably by ABBYY Finereader [4], which resulted in OCR technology achieving better results for historic newspapers. Until recently OCR technology was developed by commercial companies and licensed to OCR contractors (many of whom are based in India). There is one notable exception to this, which is OCROpus [5] - open source OCR software that Google is helping to develop.

OCR software attempts to replicate the combined functions of the human eye and brain, which is why it is referred to as artificial intelligence software. A human can quickly and easily recognise text of varying fonts and of various print qualities on a newspaper page, and will apply their language and cognitive abilities to correctly translate this text into meaningful words. Humans can recognise, translate and interpret the text on a newspaper page very rapidly, even text on an old poor quality

newspaper page from the 1800s. We can quickly scan layout, sections and headings, and read the text of articles in the right order (which is much more difficult than reading the page of a book). OCR software can now do all these things too, but not to the same level of perfection as a human can. Also, whereas a human can read greys and colour, the OCR software is still reliant on there being a clear contrast between black and white to be able to distinguish what is text and what is background page.

Pre-processing may be carried out on newspaper pages prior to the OCR process, e.g., de-skewing and de-speckling. The first step of the OCR software is to analyse the structure of the newspaper page. It divides the page into elements such as blocks of texts (columns), tables, images, etc. The lines are divided into words and then into characters. Once the characters have been singled out, the program compares them with a set of pattern images stored in its database. It analyzes the stroke edge, the line of discontinuity between the text characters, and the background. Allowing for irregularities of printed ink on paper, each algorithm averages the light and dark along the side of a stroke, and advances numerous hypotheses about what this character is. Finally, the software makes a best guess decision on the character. This character is given a confidence rating. The encoding of this confidence is dependant on the software or schema used to represent the OCR results. Therefore, a confidence rating encoded according to the ALTO standard [6] for newspapers is an integer within the range of 0-9, 9 being very confident. A secondary level analysis may then take place at word level (since now a word is formed). The built-in English dictionaries and possibly dictionaries of other languages are checked to see if the word matches. If it does, the confidence rating of the characters may be increased. The built-in dictionaries have a complicated relationship with the algorithms and the hypotheses, and how they integrate together is usually kept confidential by the software companies.

Some OCR software has the capacity for 'training'. ABBYY has an interface that allows a human to confirm/correct ABBYY's interpretation of characters as OCR progresses, hence training it by adding shapes to its database. This is mainly useful for old fonts or for material distorted or otherwise corrupted in some typical way. Once trained, the software remembers these things and applies them in the future to other similar materials. Training may be useful, but it is incredibly time consuming and has therefore not been used for large scale text digitisation projects.

OCR software can now recognize a wide variety of fonts and layouts, but handwriting and script fonts that mimic handwriting are still problematic, as are gothic fonts.

There are two alternatives to the OCR process for large scale digitisation projects:

1. To re-key the original source (can be double or triple keyed to gain better accuracy than OCR software, and is sometimes more cost efficient for items with simple layout)
2. To use Intelligent character recognition (ICR) for recognition of handwriting.

However, neither of these are suitable for historic newspapers so are not discussed further in this article.

3. Factors affecting OCR

Now that we understand the OCR process, we can pick out the factors that will affect and influence OCR accuracy. These are outlined in Table 1 below, with suggested actions that should be taken for historic newspapers.

Table 1: Factors influencing OCR accuracy in historic newspapers.

Process Steps	Factors influencing OCR	Recommended actions for historic newspaper projects
Obtain original source	Quality of original source	• Use the original hard copies if budget allows (digitisation costs will be considerably

		higher than for using microfilm). • Hard copies used for microfilming/digitisation should be the most complete and cleanest version possible – preferably not bound. • Use microfilm created after establishment and use of microfilm imaging standards (1990's or later). • Use master negative microfilm only (first generation) or original copies, no second generation copies.
Scan file	Scanning resolution and file format	• Scanning resolution should be 300 dpi or above to capture as much image information as possible. • File format to be lossless e.g. tiff so that no image information (pixels) are lost.
Create good contrast between black and white in file (Image pre-processing)	• Bit depth of image. • Image optimisation/ binarisation process • Quality of source (density of microfilm).	• Scan the image as greyscale or bi-tonal. • Image optimisation for OCR to increase contrast and density needs to be carried out prior to OCR either in the scanning software or a customised program. • If the images are greyscale convert them to image optimised bi-tonal (binarisation). • Obtain the best source quality. • Check density of microfilm before scanning.
OCR software- Layout of page analysed and broken down.	• Skewed pages • Pages with complex layouts. • Adequate white space between lines, columns and at edge of page so that text boundaries can be identified.	• De-skew pages in the image pre-processing step so that word lines are horizontal. • Layout of pages and white space cannot be changed, so work with what you have.
OCR software - Analysing stroke edge of each character	• Image optimisation. • Quality of source.	• Optimise image for OCR so that character edges are smoothed, rounded, sharpened, contrast increased prior to OCR. • Obtain best source possible (marked, mouldy, faded source,

		characters not in sharp focus or skewed on page negatively affects identification of characters).
OCR software- Matching character edges to pattern images and making decision on what the character is.	• Pattern images in OCR software database. • Algorithms in OCR software.	• Select good OCR software.
OCR software- Matching whole words to dictionary and making decisions on confidence.	• Algorithms and in-built dictionaries in OCR software.	• Select good OCR software.
Train OCR engine.	• Depends on how much time you have available to train the OCR.	• Purchase OCR software that has this ability. • At present it is questionable if training is viable for large scale historic newspaper projects.

4. Measuring OCR accuracy

If you are going to improve your OCR accuracy, you first need to know how to measure accuracy, so that a baseline can be created against which improvements can be tested.

Measuring accuracy rates
OCR software calculates a confidence level for each character it detects. Word and page confidence levels can be calculated from the character confidences using algorithms either inside the OCR software or as an external customised process. The OCR software doesn't *know* whether any character is correct or not - it can only be confident or not confident that it is correct. It will give it a confidence level from 0-9. True accuracy, i.e., whether a character is *actually* correct, can only be determined by an independent arbiter, a human. This can be done by proofreading articles or pages, or by manually re-keying the entire article or page and comparing the output to the OCR output. These methods are very time consuming.

OCR contractors often talk about OCR confidence levels and OCR accuracy as if they were the same thing, and in practice confidence levels are often used as a substitute for accuracy because determining true accuracy is not feasible for large volumes of text. Only one contractor to whom we spoke suggested a good solution for gaining an accuracy figure (rather than a character confidence figure) for libraries with large scale text projects. We thought the contractor's idea was potentially viable, and could be resource effective and accurate, but as far as we know, no one is actually utilising it at present. The idea is as follows:

1. Determine the actual accuracy of a sample of the OCR output by a manual method.
2. Gather the OCR confidence levels for the same sample.
3. Write an algorithm that will correlate the two and provide a 'proxy' accuracy level based on both.
4. Use the algorithm to supply proxy accuracy levels at article/page level.
5. At regular intervals re-check the algorithm, since it may change with different newspapers.

We concluded that being able to identify real or proxy accuracy at article level would be the most useful thing for us rather than having a figure for the entire newspaper corpus (which was too broad), or the individual title (since most changed in layout and quality over time), or the entire page. Being able to identify accuracy levels at article level would give us the potential to be able to measure increase in accuracy now or in the future. At present the Library simply records page level character confidences in the ALTO XML file, however we intend to pursue this idea further and see if we can implement it in our own program.

Acceptable OCR accuracy rates

The Library noted that most OCR software claims 99% accuracy rates, but these are either on new good quality clean images, e.g., word documents, or when manual intervention in the OCR process takes place, so these accuracy rates are not applicable to historic newspapers. The older the newspaper, the lower the accuracy rate is likely to be, and accuracy rates are generally lower for newspapers than for books and journals. In 2007 the Dutch National Library (Koninklijke Bibliotheek) did some market research with a dozen OCR contractors and published their findings in a *D-Lib Magazine* article [7]:

"Accuracy rates, on either word or character level, should not be considered as watertight performance indicators for OCR software. Usually the quality of the OCR text says more about the condition of the original materials than it does about the performance of the OCR software. For what its worth, the rates respondents gave for newspaper digitisation projects vary from 99.8% for 700,000 newspaper pages (word accuracy, manually corrected) to 68% (character accuracy, no correction) for 350,000 pages of early 20th century newspapers."

Our own tests yielded similar results. In a sample of 45 pages to be representative of the libraries digitised newspaper collection 1803-1954, we found that raw OCR accuracy varied from 71% to 98.02% (character confidence, no correction). At the 71% accuracy rate there would be 145 incorrect characters in an average paragraph of 500 characters. Looking at this the other way round means that 29% of the paragraph would be incorrect. This makes reading and accurate text retrieval difficult, even more so if the incorrect characters are not whole words but characters within many words. The question of what is acceptable has not been answered, but in speaking to other libraries and OCR contractors, it was generally agreed for historic newspapers that when we talk about good, average and bad OCR we mean:

Good OCR accuracy	= 98-99% accurate	(1-2% of OCR incorrect)
Average OCR accuracy	= 90-98 % accurate	(2-10% of OCR incorrect)
Poor OCR accuracy	= below 90 % accurate	(more than 10% of OCR incorrect)

However, there was not consensus on whether these percentages referred to character or word confidences, and whether this was at page or article level, and many people were still confused about what accuracy/confidence meant and how it was calculated. We concluded that we could not set a baseline figure for acceptable accuracy until the method for measuring it had been firmly established and agreed.

3. Methods of improving OCR accuracy

From the very early planning stages of the Australian Newspaper Digitisation Program, the Library was preparing itself for getting poor OCR back from historic newspapers, largely due to being unable to obtain better quality original source documents and having to work with what we had. In regular IT planning meetings we brainstormed ideas to seek cost effective and realistic methods we could use to improve accuracy if it was very bad. This resulted in 13 potential methods being raised for

improving OCR accuracy. Of these methods four were already being done, and four were not viable for mass scale newspaper digitisation, but five had the potential to be further investigated and tested. They are outlined in Table 2 below.

Table 2: Potential methods of improving OCR accuracy suggested by ANDP team.

	Suggestion	Comments	Actions
1	Improve quality of original source.	Generally cannot be changed. Using hard copies instead of microfilm is not an option due to extra budget and time required. Will also not necessarily give better results. Some re-microfilming has taken place funded by ANPlan. Use master negative (first generation copy only).	Continue as is using existing master microfilm (even if poor quality).
2	Scan at 300 dpi resolution or above.	Already being done.	Continue as is.
3	Use tiff files only for OCR.	Already being done.	Continue as is.
4	Make manual adjustments to OCR image optimisation process for each page.	Not viable for cost effective mass scale digitisation.	Do not pursue.
5	Compare image optimisation software to see if we are using the best.	Worth further investigation.	Test.
6	Experiment using greyscale files instead of bi-tonal files for OCR.	Worth further investigation. Contractors claim it will increase accuracy. May add additional costs due to transport and delivery of large files.	Test.
7	De-skew every page on a vertical and horizontal grid so that image text is as horizontal as possible.	Already being done by Contractor.	Continue as is.
8	Manual intervention in the OCR process for each file to improve results.	Not viable for cost effective mass scale digitisation.	Do not pursue.
9	Use the "training" facility (artificial intelligence) in the OCR software.	Not viable for cost effective mass scale digitisation.	Do not pursue.
10	Use voting technology (i.e., use more than one OCR software solution and voting technology picks the best results).	Would add further cost, complexity and time to the process. Voting technology is more suited to dealing with 'forms' with typed and handwritten text than it is with newspapers. We think ABBY software makes the best guess decisions and we are already using it.	Do not pursue.
11	Use Australian dictionaries in OCR	Australian dictionary not developed until much later. Australian historic newspapers	Test.

	process.	use English spellings. Could input Australian gazetteer as a dictionary to test placename spellings.	
12	Clean/correct OCR text manually.	Ask Contractor to correct titles, subtitles, author and first 4 lines of article text. Not financially viable in mass scale digitisation project to correct much of the article text. Let the public view and correct the text for free as they find and read articles. May need to write moderation software for this. If we can identify poor quality articles serve these up to the public for correction. Library QA staff to correct lead story on front pages if viable.	Test.
13	Use Confusion Matrix and Language modelling post/during OCR processing	Requires software development time – could be time consuming. Applying a confusion matrix during or after OCR processing and in combination with a language model could improve results a lot. Need to find the matrix and language examples that we want to apply.	Would like to test. Don't have resource at present. Keep on list to review in future.

4. ANDP Tests and Results

Four of the five suggested solutions were investigated and tested with the assistance and advice of these scanning and OCR contractors: W & F Pascoe Ltd, DataComIT, Apex Co-Vantage, and Planman Consulting. The tests and research were done over a nine-month period in between other project tasks as the program progressed in 2007/2008. The four major solutions tested were:

1. Using Australian dictionaries in the OCR process,
2. Using greyscale instead of bi-tonal files for OCR,
3. Comparing image optimisation software to see if we are using the best, and finally
4. Correcting OCR text manually.

A fifth solution was also identified – using a confusion matrix and language modelling. This could not be tested due to a lack of resources, but the suggestion is described below.

Using Australian dictionaries in the OCR process.
Implementation of dictionaries seemed to be a potentially quick and easy way to increase OCR accuracy, which is why this solution was tested first. However, we quickly realised that using the Macquarie Australian national dictionary [8] (first published in 1981) was unlikely to make any difference to OCR results, since at the time Australian newspapers were published (1803- 1954) pure English was being spoken primarily, and there were few Australian colloquialisms that might appear in those newspapers. Jumbuck (sheep), Billabong (watering hole) and tucker (food) were the only possible exceptions we could think of. However, aboriginal place names were in wide use then, as they are today, so we asked our OCR contractor to incorporate the official Australian gazetteer of place names as a secondary dictionary into ABBYY (the primary dictionaries are already built in) and run 45 sample pages through OCR. We then compared the results with pages using the primary dictionary only and pages using no dictionaries.

Our findings were that using ABBYY primary dictionaries gave the highest accuracy, followed by using no dictionaries. Surprisingly, the worse results were obtained from using the primary and secondary dictionaries together. This was unexpected and at variance with experiences from the National Library of New Zealand, which had obtained improved results when applying a secondary dictionary populated with geographic place names. We wanted an explanation for this, but unfortunately, both our contractor and we had a very limited understanding of when and how a secondary dictionary was applied and utilised. The OCR contractor approached ABBYY about this, but ABBYY was reticent to share that proprietary software information. We suspected that the secondary dictionary was interfering with the primary dictionary or had been implemented incorrectly, but we were unable to obtain any further information about implementation of dictionaries or application of secondary dictionaries, so the results could not be explained. We repeated the tests again on different pages six months later with the same contractor but the results were the same. More time and research on the use of two dictionaries is desirable, because there were a lot of unknowns for us in the test process. In the meantime, however, we have decided to use the default primary dictionaries only in ABBYY for OCR of Australian Newspapers.

Using greyscale files instead of bi-tonal files for OCR.
OCR contractors have suggested that using greyscale files instead of bi-tonal files for OCR may result in improved accuracy due to recent developments in OCR technology. Until recently it was certain that OCR software had to have clear black and white definition (bi-tonal) to be able to identify characters properly and that OCR software could not deal with greyscale files, but some contractors disputed this. We noted there was not a consistent consensus among OCR contractors as to which files gave better results, and very little, if any, thorough or extensive testing or research had been done in this area by libraries.

Our test method was to use 45 pages from the 1803 – 1954 period of different titles representative of good, average and poor quality pages to mirror production mode. Each page was a pair (an image optimised bi-tonal and a greyscale file). The files were processed by two different contractors, in each case using the same version of ABBYY Finereader, and we also processed some pages in-house ourselves to double check results. We hoped that we would see the greatest improvements in the 'average' group of pages when using greyscale.

Results were that, although three contractors had confirmed that they could process greyscale files and that they were likely to yield better results, all the contractors were reluctant to do this in production. Their interpretation of being able to process greyscale files meant converting the greyscale files into image optimised bi-tonal files for the OCR process, rather than using the actual greyscale file for OCR. The contractors wanted to do this, because greyscale files take much longer to process, handle and return due to their large size. If greyscale files were processed, day to day the cost would be much greater than processing bi-tonal files. In spite of their reluctance, we asked two contractors to do the tests on the actual greyscale files – just so we would know if that would give better results.

The results showed there was no *significant improvement* in OCR accuracy between using greyscale or bi-tonal files. In the 'average' group of files, the total character accuracy score for bi-tonal was 93.8% and for greyscale 94%. There was a variant of +/- 3% between files in this group. This meant that sometimes the greyscale yielded a slightly better result than bi-tonal, sometimes there was no difference, and in one case the results were worse; there was no consistency in results or overall significant improvement. The small variation in overall accuracy of results simply would not warrant the extra cost involved in processing greyscale files, and doing so would not lead to uniformly improved OCR accuracy rates on every file.

Those results surprised both the Library and the contractors. The Library later learned that the results mirrored those obtained from the National Library of New Zealand [9] performed at a similar time, also on a small sample, with the same software. The ANDP therefore decided to apply OCR on image-optimised bi-tonal files only. In the future, perhaps, we will do further testing of greyscale files when OCR, storage, and transport technology have been developed further.

Comparing image optimisation software to see if we are using the best.

Newspaper pages for ANDP are being scanned from microfilm using NextStar [10] scanning software. The software produces a matching bi-tonal and greyscale file pair and can also carry out generic image optimisation processes to improve files for OCR.

It is important to note that image optimisation for OCR is very different from image optimisation for reading by the human eye. Often image optimisation for OCR results in the image being less readable by humans. The ANDP is performing image optimisation in two different ways and at two different stages of processing. The first image optimisation is done by the scanning software and converts the greyscale into a bi-tonal file (binarisation), which is then optimised for OCR. The second image optimisation is done by Library-developed in-house software on the greyscale file for delivery to the public as the reading image. There is still a lack of consensus about whether humans prefer to read a bi-tonal or greyscale file, and the Library has not done research in this area.

The Library was satisfied with the generic OCR image optimisation within NextStar software (e.g., smoothing characters, sharpening, removing dirt and noise, increasing contrast), but wanted to double check the quality of the NextStar optimisation against other image optimisation programs. Most OCR contractors offer a service to optimise image files prior to OCR using in-house proprietary or open source programs or a combination of open source and propriety software. This is because most OCR software has a very limited image optimisation program within it, and other software programs can do a better job than the OCR software can. This was another surprising thing to us. The Library had expected that recent developments in OCR technology would have entailed development of image optimisation within the OCR software itself (not the scanning software, or as a separate process).

The test method was for 45 representative greyscale pages to be converted from greyscale to bi-tonal (binarised) and then image-optimised for OCR. The resulting file would be OCR processed. There were to be four test groups:

- Control group (using current methods in NextStar)
- Contractor 1 test (using their proprietary software)
- Contractor 2 test (using their proprietary software)
- Library internal program test (using free/open source software)

The 45 pages should be image-optimised by running files through a generic program, as it would be in real production, rather than handcrafting the example for perfection. The ANDP gave itself a limit of two weeks development work to create an effective image optimisation program from free and open source software, and agreed if the in-house program could not be achieved in this time, we would continue with only the contractor's tests and the control group's test . The ANDP's internal program could not be developed within the two-week time frame, re-enforcing the fact that image optimisation is a very complex process requiring advanced algorithms and specialist technology concepts and knowledge.

The results of the contractor and control group tests showed that there was no *significant* improvement between OCR accuracy from images optimised with NextStar software (current process) and other methods. The two contractors' methods gave very slightly lower accuracy results, but this difference was negligible. We concluded that, in time and cost, it was more efficient to continue with the image optimisation process as part of the scanning component rather than implement a separate process prior to OCR. This would also keep file delivery and transport costs low.

Correcting OCR text manually.

Manual text correction has always been an effective method of increasing accuracy, but on a large scale it is not viable or cost effective (assuming you pay your OCR contractor to do it). However, as early as 2005 the Library had decided that the OCR contractor would correct the OCR in article title headings, subtitles and the first four lines of the article to 99.5% accuracy. Our decision was made in case the raw OCR was really bad and also so that in the search results it would be easier for users to quickly identify the relevance of the article. When more text correction was discussed, it was always assumed that we would have the contractors do the additional corrections. However, this idea was challenged by the lead system architect on the project, who suggested we expose the raw OCR to the public for correction by them. This was a very challenging and exciting idea that held both benefits and risks. To the best of our knowledge, no other library or newspaper service worldwide had implemented user correction of text, or even considered doing so as an option. But the system architect's suggestion was the one that excited us most. The vision was for a user to be able to view, edit, correct and save OCR text on whatever article they happened to be looking at. In addition if we knew which articles had very poor OCR, we could serve those up for correction on demand. The major benefits of allowing the public users to manually correct OCR text would be:

- OCR text could be improved and potentially whole articles made perfect (100%), therefore improving the searchability of the entire newspaper collection.
- It was likely to be cost effective and suitable for mass scale digitisation programs.
- The community could be harnessed and involved to add value to the service in a new way (similar to Wikipedia).

Questions we didn't have any answers to included:

- How hard technically would it be to implement?
- Would the public really do it?
- Should/could we moderate it?

In spite of these big unknowns, the ANDP team decided to put the idea into action. Basic OCR correction by public users was implemented and tested in the prototype search system released to State and Territory Libraries for testing in December 2007. User correction of text was positively received, though most Libraries asked if and how moderation would take place. It was then implemented in the Beta search system (without moderation), which had a soft release to the public without any publicity on 25 July 2008. In the first three months of use (July – October 2008) the public immediately began correcting OCR. We have found it quite hard to monitor what they are doing, how well they are doing it, and how it is effecting the overall quality of the data, since moderation is not yet in place and login to do it is not mandatory (it is optional) at this stage. We also have had difficulties measuring the accuracy of the OCR-corrected text. We have three methods of measuring text correction: number of lines corrected, number of correction "transactions" (i.e., pressing the "save corrections" button), and number of different articles corrected. However, it is questionable how useful any of the three methods are. We are assuming that all correction transactions are to improve text and make it right. No extra text can be added, only existing lines corrected. No text has been deliberately incorrectly changed as far as we are aware.

The results of user activity within the first 12 weeks of the soft launch (without publicity) are that 868 registered users have corrected text and approximately 390 unregistered users (total of 1,200 text correctors). 700,000 lines of text have been corrected within 50,000 articles. The top text corrector has corrected 50,000 lines of text within nearly 2,000 individual articles. Some articles have had corrections added by more than seven 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 newspaper.

User feedback returned via surveys, e-mails, phone calls and the "contact us" form has been overwhelmingly positive and interesting. Users did not expect to be able to correct OCR text. Once they discovered they could, they quickly took to the concept and method, and several reported finding correcting the text both addictive and rewarding. Users were actively correcting much more than they or we had expected to correct. In addition, our own users have the potential to achieve a 100% accuracy rate with their knowledge of English, history and context, whereas our contractors are only achieving an accuracy of 99.5% in the title headings.

The ANDP team are continuing with the experiment, and I intend to write more fully about the outcomes of public OCR correction later this year once we have more data available, have received more feedback from the public, and the Beta service has been available for six months. To date, we consider this method to be one of the most promising and radical ways for large scale newspaper digitisation projects to improve OCR accuracy.

Using a Confusion Matrix and Language Model

OCR errors, whilst not predictable, can be modelled statistically and are very different from human spelling mistakes. A confusion matrix would model these errors in order to be able to correct them and improve OCR. For example, an OCR engine may commonly mistake the letter *h* for *l i* or *m* for *in*. Thus the word 'the' would be translated as 'tlie' instead of 'the'. In the ANDP the mistranslated word 'tlie' occurs in 1 of 8 articles. But the mistranslations are more significant on words on which users may search. There are 30,000 articles with 'Sydnoy' instead of 'Sydney'. These observations of unigrams, bigrams (pairs of characters) and trigrams and what they get translated as form the basis of the confusion matrix. The matrix could be applied across the OCR as the first part of the process by identifying correction candidates.

Once a word is formed, its occurrence in a sentence in the language model can be applied. Words are more likely to occur in context; therefore, a sentence reading "the cot sat on the met" is much more likely to actually be "the cat sat on the mat". Likewise a "rich coal scam" is more likely to be a "rich coal seam", the occurrence of the word scam directly next to coal is unlikely. The confusion matrix applied with a language model has the potential to increase OCR accuracy, though not to make it perfect. The team were interested in this approach but did not have the time, resources or adequate OCR data to pursue it further. Software development would be required and thorough testing of data using the matrix and algorithms to ensure that results were improved not worsened. It was unclear how long the post OCR processing might take and how viable it may be for large scale newspaper projects. In 2008 the Impact[11] Project was established in Europe as part of the i2010 vision of the European Digital Library. It has funding of 15.5 million Euros and has 15 national library partners. Some of the aims of the project are to develop text recognition products further, develop post correction modules and investigate the most effective methods of enhancement and enrichment of OCR for mass scale digitisation projects. The confusion matrix and post language processing may fall under this research.

7. Conclusion

The experiences gained from the ANDP show that we can make OCR accuracy better than we have ever been able to till now, by using a combination of methods. It also appears that one of our best solutions to improve OCR accuracy may not be actually carried out by our OCR contractors or by us, but by our users. The best way to improve accuracy may not rely on a technical solution but on a manual method of humans correcting the mistakes of a machine. This was ruled out before as being too labour intensive, but that was before the advent of web 2.0 technologies, social networking and user involvement. If we can harness the energy and time of our users and their desire (as strong as ours) for the OCR to be improved, who knows how accurate we can get it? We are also hopeful that more research and development will be done on confusion matrix and language model post processing, since we also considered this approach to have potential but have not yet been able to investigate it further. Being able to measure the true accuracy of OCR and the degree to which it has been improved or to establish a baseline of what is acceptable is still difficult, and we have not achieved this goal yet. To date, most libraries do not record accuracy, because of the difficulties in doing so and the seemingly pointlessness of it. Or they are recording character confidence and calling it accuracy, which it is not. But perhaps we should not get too hung up on measuring percentages and instead just let the public decide what is right and wrong and then do something about it themselves, in a similar way to Wikipedia. We are using public OCR correction as a good interim strategy to improve OCR text, but maybe it will turn out to be the ultimate solution. Our investigations will continue.

8. Notes and References

[1] Australian Cooperative Digitisation Program (ACDP) website: http://www.nla.gov.au/ferg/about.

[2] The unsuccessful projects of the 1990's are reported in detail in this article:
Entlich, Richard., 2002. Where are they now? Digitising Microfilmed Newspapers. *RLG Diginews*, June 15, 2002, vol. 6, no 3.
URL: http://worldcat.org/arcviewer/1/OCC/2007/08/08/0000070519/viewer/file1572.html#faq.

And a response to this from the British Library is in this article:
Deegan, Marilyn., 2002. Digitising Historic Newspapers: Progress and Prospects. *RLG Diginews*, August 15 2002, vol. 6 no 4.
URL: http://worldcat.org/arcviewer/1/OCC/2007/08/08/0000070519/viewer/file730.html#feature2.

[3] Australian Newspapers Digitisation Program (ANDP) website: http://www.nla.gov.au/ndp.

[4] ABBYY website gives more information about ABBYY Finereader OCR software development history and how the software works: http://www.abbyy.com/company.

[5] OCRopus website: http://sites.google.com/site/ocropus/.

[6] ALTO (Analyzed Layout and Text Object) is a standardized XML format used for storing layout and content information of complex digital objects like newspapers. It is currently being used for newspaper digitisation projects at the US Library of Congress, the National Library of Australia, and the Bibliothèque nacional de France.

[7] Klijn, Edwin., 2008. The current state of art in newspaper digitisation. A market perspective. *D-Lib Magazine*, January/February 2008, Vol. 14 No. 1/2, doi:10.1045/january2008-klijn, ISSN: 1082-9873.
URL: http://www.dlib.org/dlib/january08/klijn/01klijn.html.

[8] Macquarie dictionary website: http://www.macquariedictionary.com.au.

[9] Powell, Tracy and Gordon Paynter. 2009. Going Grey? Comparing the OCR Accuracy Levels of Bitonal and Greyscale Images. D-Lib Magazine, March/April 2009, vol. 15 no 3/4. URL: http://www.dlib.org/dlib/march09/powell/03powell.html.

[10] NextScan website: http://www.nextscan.com/products/nextstar.html

[11] Impact website: http://www.impact-project.eu/home