

Measuring the OCR Accuracy across The British Library's 2 Million Page Newspaper Archive

Simon Tanner

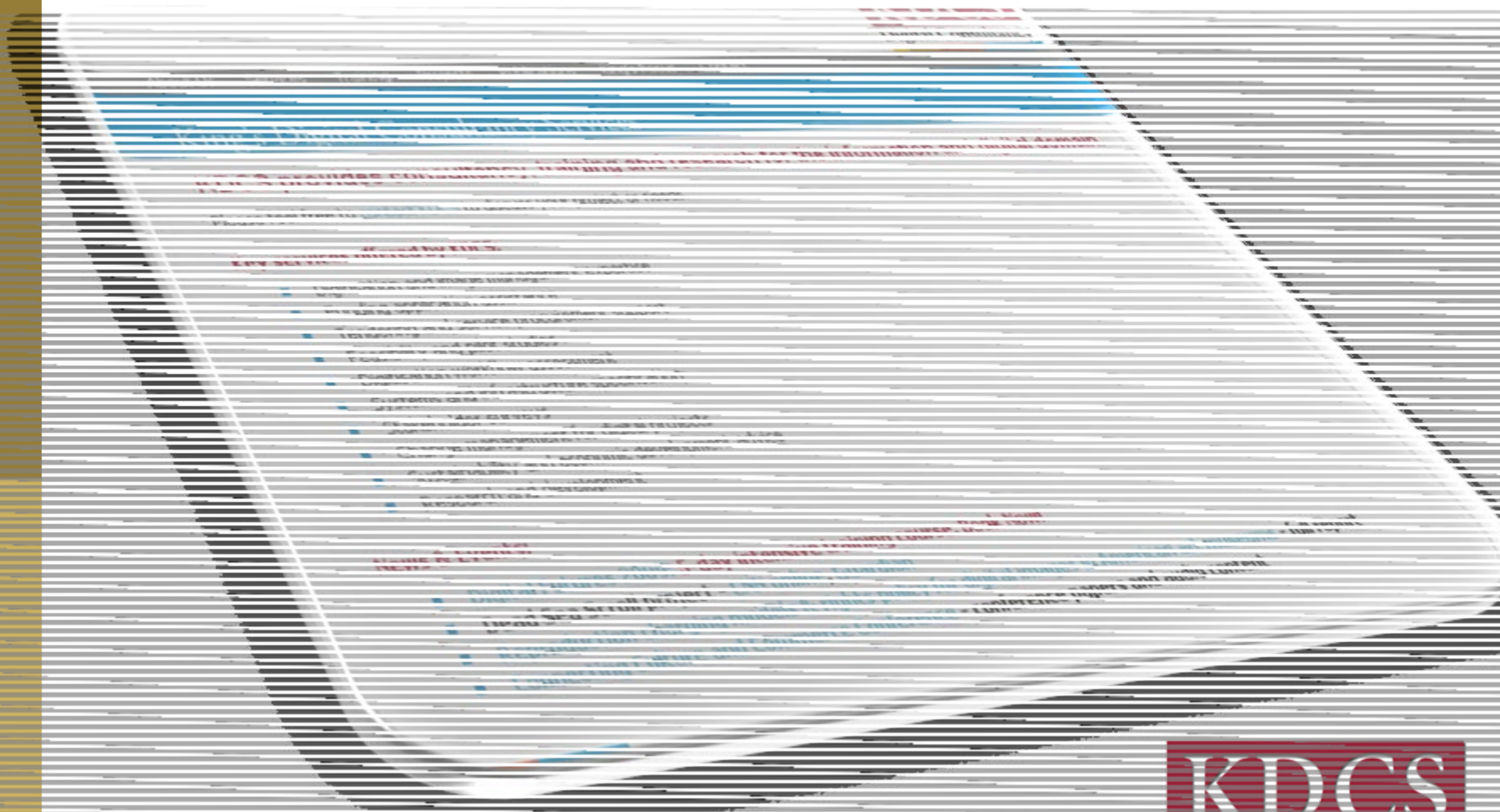
Director

King's Digital Consultancy Services

Email: simon.tanner@kcl.ac.uk



King's Digital Consultancy Services



www.digitalconsultancy.net



KDCS: Activity

- **National Library of Scotland: digital access strategy**
- **The National Archives: digitisation advice**
- **Tate Archive: digitisation business planning**
- **Text recognition study for Oxford University Digital Library**
- **OCR accuracy assessment for British Library 19th Century newspaper digitisation**
- **Mellon Foundation projects: rights and reproductions in US Art Museums (100 surveyed, 20 interviewed)**
- **National Library of Ireland: digital library strategy**
- **Digitising the Dead Sea Scrolls**
- **Digital Futures Academy**



Digital Divide Data

DDD is an internationally acclaimed non-profit that uses a sustainable, private-sector business model to break the cycle of poverty in the developing world.

We are an innovative social enterprise that pursues a double bottom-line approach to development.



A Sustainable Social Enterprise

DDD manages a sustainable business selling **outsourced digitization services** to some of the world's leading publishers, universities and archives.

DDD raises grant funding from international development agencies, governments, and private donors to fund our **social mission**.

Evaluation of OCR accuracy

- **Discovery of OCR accuracy**
- **Based upon the actual XML output in relation to the original images rather than the OCR process**
- **Looking for significant results that deliver meaningful information on OCR accuracy**

- **Thus:**
 - **Not directly assessing the capability of the OCR engine or algorithm used**
 - **Not assessing whether the supplier that was used is good or bad – that is for the BL to assess**

OCR Accuracy

- Evaluating OCR accuracy is about more than just character to character accuracy rates
 - **Character accuracy rates are misleading (more later...)**
- It is also about assessing the functionality enabled through the OCR's output
 - Search accuracy
 - Volume of hits returned
 - Ability to structure searches and results
 - Accuracy of result ranking
 - Amount of correction required to achieve the required performance

OCR Accuracy

- Character accuracy rates may be misleading:

Consider this scenario:

- 1,000 words with 5,000 characters (an average of 5 per word) excluding spaces
- 90% character accuracy means:
 - 4,500 characters correct
 - Possibly a maximum 900 words correct (90%)
 - Possibly a minimum 500 words correct (50%)
 - Reality is somewhere in between
 - Depending on the number of “significant words” the search results could still be almost 100% or near zero

OCR Accuracy Evaluation: Principles

- Seeking to find the accuracy of the OCR in the BL's 19th Century Newspaper Project
- Looking for the following:
 - Character accuracy
 - Word accuracy
 - Significant word accuracy
 - Significant words with capital letter start accuracy
 - Number group accuracy
- Sampling a significant proportion ~ 1% of pages assessed
- Aiming to find the **highest** OCR accuracy rate, not the lowest, on each page

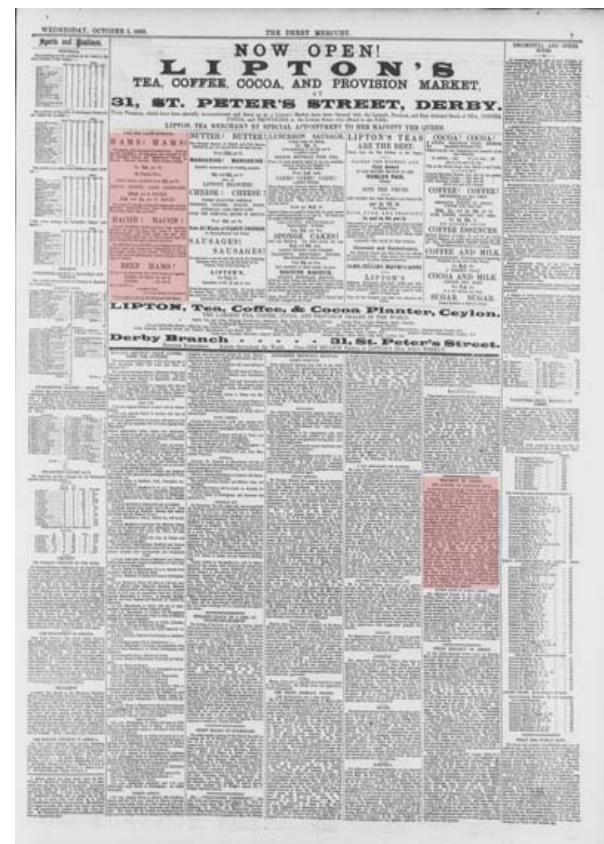
OCR Accuracy Evaluation: Method overview

Digital Divide Data

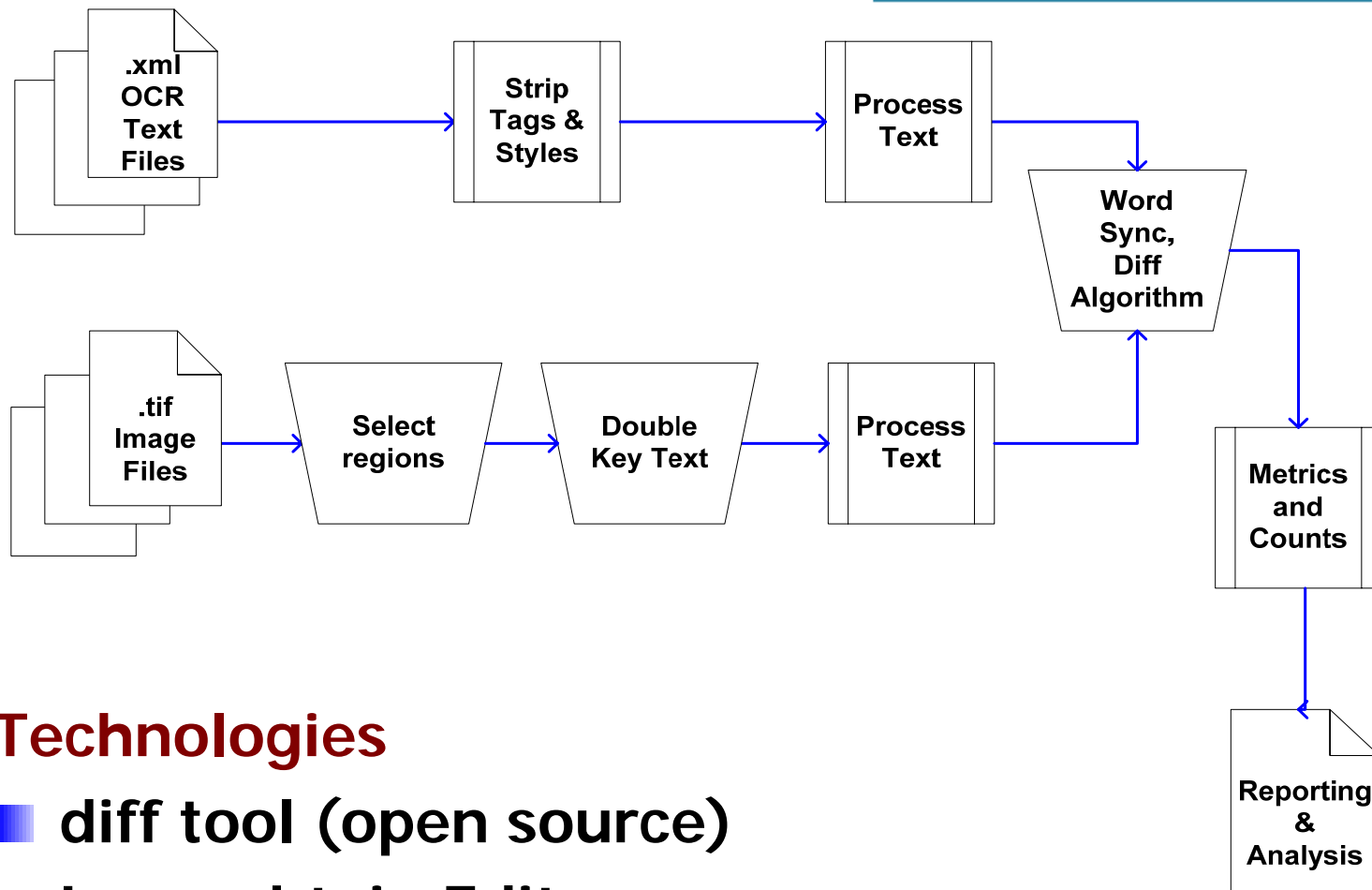
- Double rekey 200 words drawn from 2 zones per page
- Prepare the same text from the BL XML files
- Compare the double rekeyed text with the XML text

KDCS

- Create metrics
- Analyse results



OCR Accuracy Evaluation: Workflow



Technologies

- diff tool (open source)
- Levenshtein Edit Distance algorithm

Example of text for comparison

	The said Lease contains a covenant for the purchases of ale and beer from or through the Lessor which is however not of an onerous nature and can be inspected at the Office		
rekeyed		The said Lease contains a covenant for the the purebase of ale and beer from or through the Lessor which is 5 not of an onerous nature and can be in at the Office	OCR

Note:

Punctuation removed

Hyphens resolved

Spaces ignored

Tags replaced:
e.g. “'” for ‘

Significant words selected
by exclusion of stop list

OCR Accuracy Evaluation: Output

- **Excel spreadsheet with:**
 - > 40,000 lines of comparison
 - > 4,000,000 words
 - > 25,500,000 characters

- **Newspaper Title / Code**
- **Year & Month**
- **Original filename & Clip**
- **Numbers and Statistics for:**
 - characters
 - words
 - punctuation marks
 - words with capital letter start
 - number groups
 - significant words

Results

Overall averages for the 19th Century Newspaper Project:

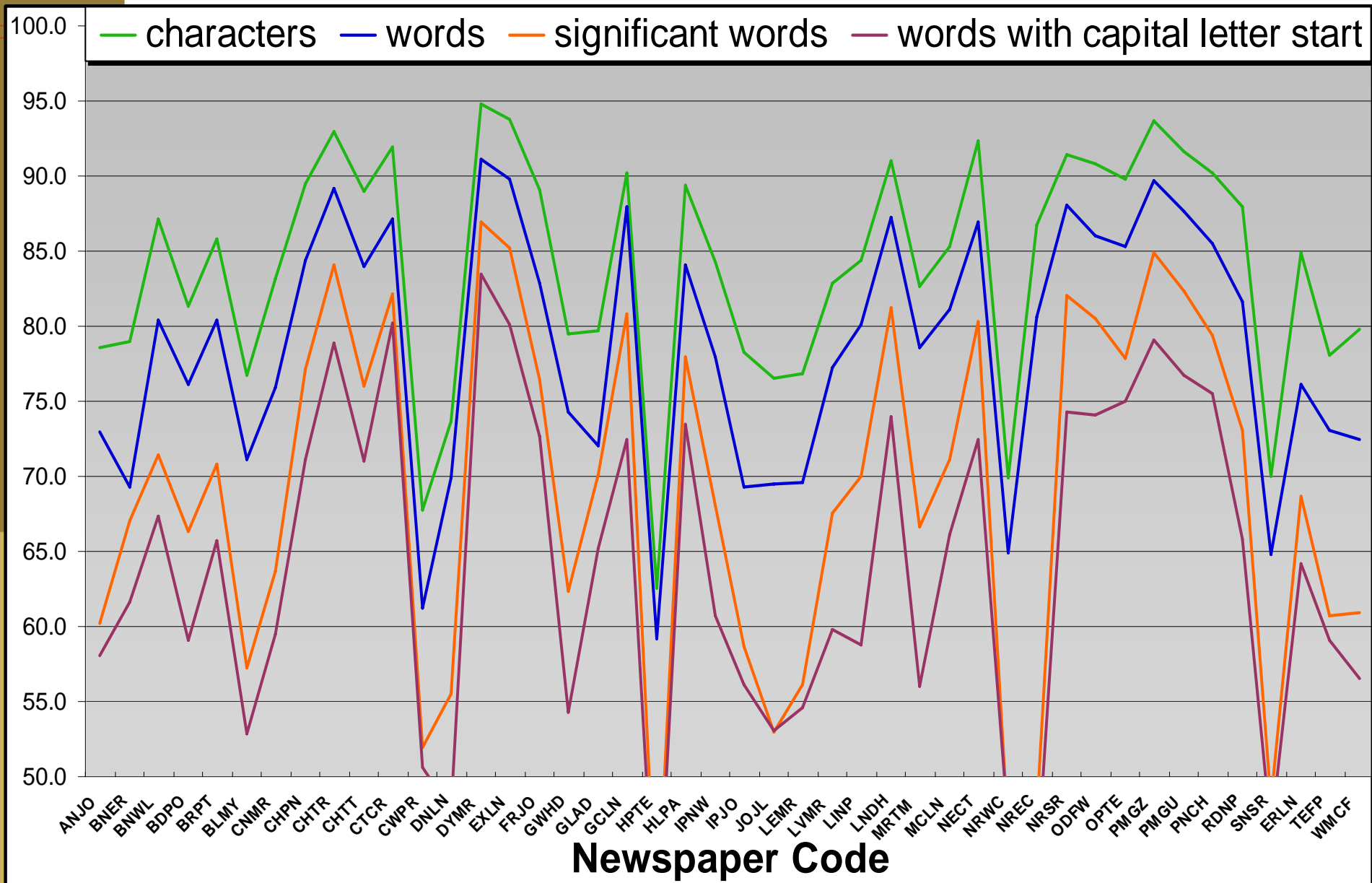
- Character accuracy = 83.6%
- Word accuracy = 78%
- Significant word accuracy = 68.4%
- Words with capital letter start accuracy = 63.4%
- Number group accuracy = 64.1%

Overall averages for the Burney Collection:

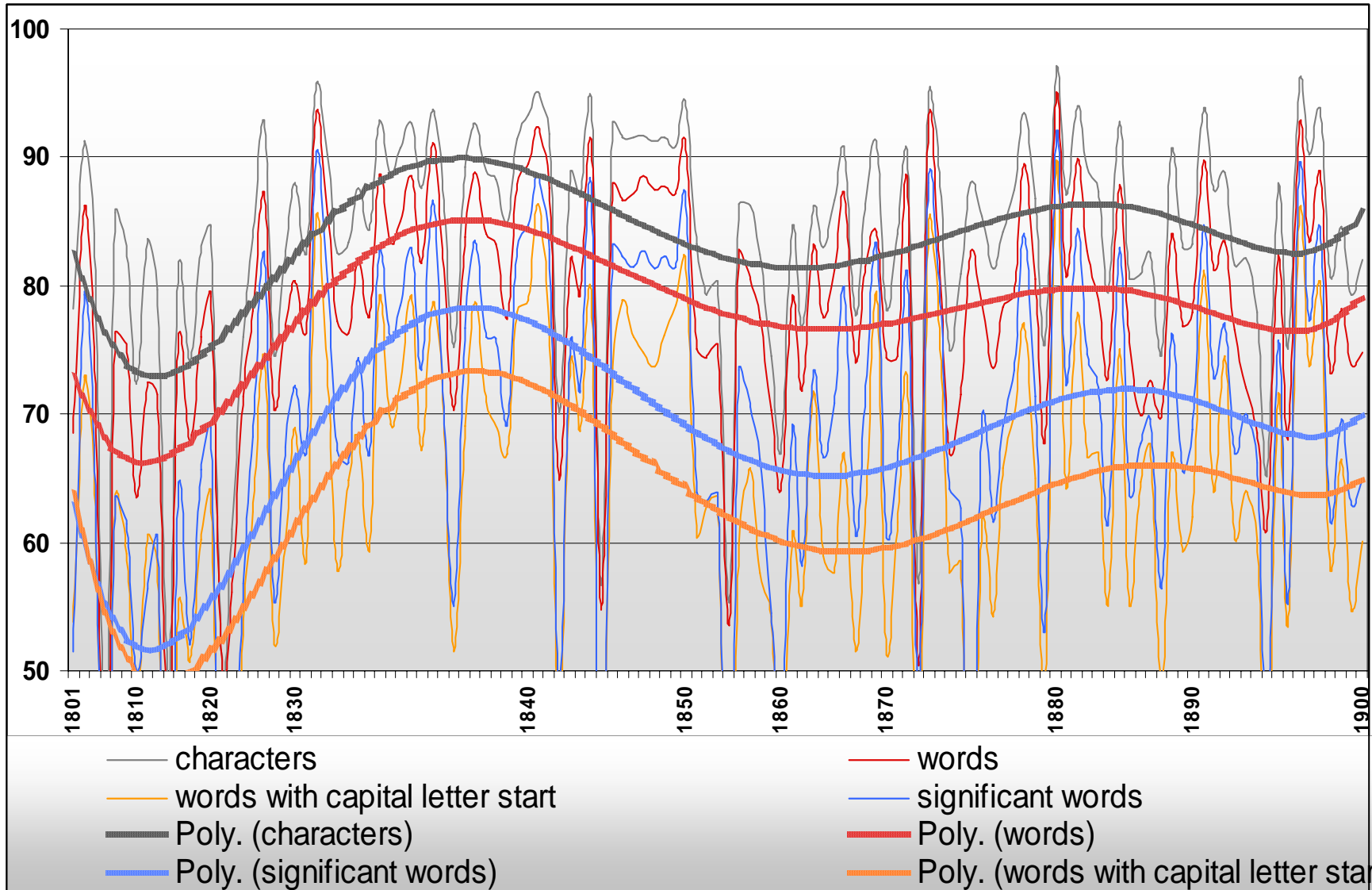
- Character accuracy = 75.6%
- Word accuracy = 65%
- Significant word accuracy = 48.4%
- Words with capital letter start accuracy = 47.4%
- Number group accuracy = 59.3%

(Note: from this point on the statistics will refer purely to the 19th Century Newspaper Project content)

Results: arranged by title

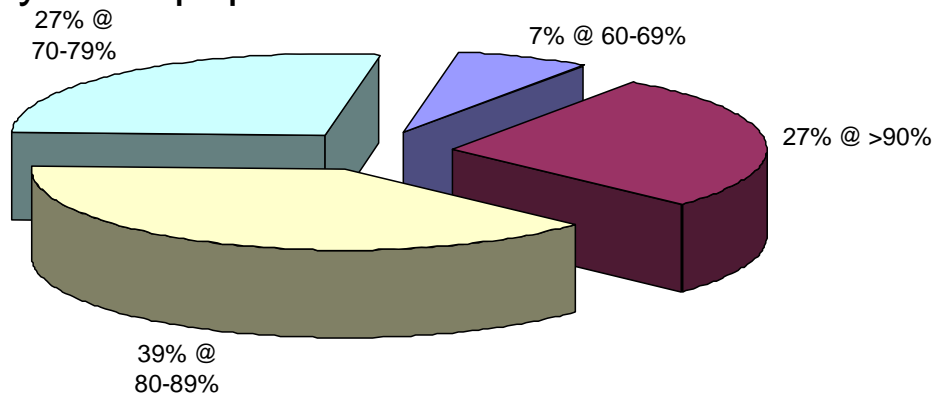


Results: arranged by date

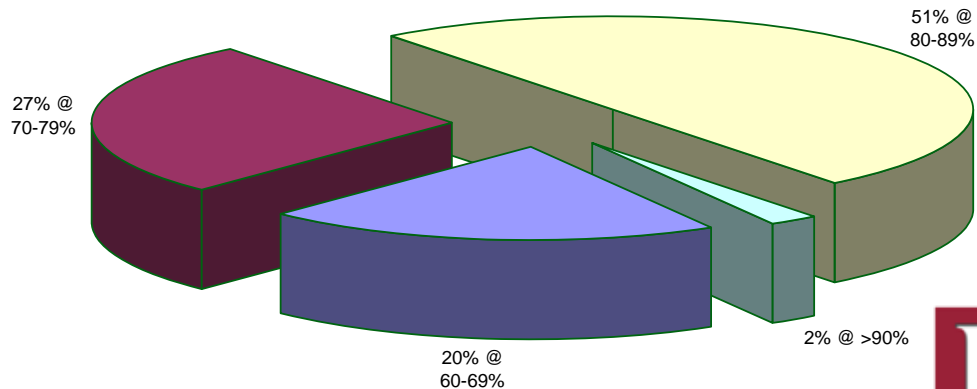


Results: Proportions by newspaper title

Proportions of OCR **Character** Accuracy by Newspaper Title



Proportions of OCR **Word** Accuracy by Newspaper Title



Conclusions

- 2/3 of the newspaper titles have an average character accuracy above 80% accuracy.
- However, 1/2 have the same for word accuracy and only a mere 1/4 have greater than 80% significant word accuracy.
- The lower the significant word accuracy the lower the likely search result accuracy or volume of search results returns.
- Should the word accuracy be greater than 80% then most fuzzy search engines will be able to sufficiently fill in the gaps or find related words such that a high search accuracy (>95-98%) would still be possible.

Conclusions

- The user experience of searching the resource is unlikely to be as satisfactory as might be desired.
- This is due to a combination of two factors:
 - the OCR process itself and
 - the content of the newspapers themselves.

We can also say:

- The methodology and the statistical model developed here has many other applications
- The data set we have is very rich and lends itself to further analysis

Something more...

- This method could be used to assess and optimise the OCR performance in relation to the original text.
- It can be used as a means of making better OCR decisions.
- Note: These results imply rather than give a definitive statement of potential search accuracy and search result ranking. **A definitive method does exist but was outside the scope of the BL evaluation delivered.**
- By comparing the number of repeat significant words and then measuring the accuracy against the OCR it would be practicable to accurately assess the search accuracy.

A new service offering

Offering

King's Digital Consultancy Services in partnership with Digital Divide Data is thrilled to offer a service to assess the true statistical accuracy of OCR for digitization projects in any European language.

Our consultative service will use these accuracy results to give you actionable information that you can use to select content, optimize your OCR process, improve your search performance, design your delivery systems and reduce the costs for your project.

Our process involves a unique method for algorithmic comparison of OCR text output and original text transcripts to determine OCR accuracy to a high degree of confidence.

Our deliverables include:

An OCR accuracy statistical report, including:

- A complete item-by-item report of the OCR accuracy achieved in relation to the original text.
- OCR accuracy expressed as the percentage of characters, words, significant words, names, place names and numeric content rendered correctly.
- OCR accuracy expressed graphically for statistical overviews of the text resources broken down by factors such as date range, publication title, language or publication type.

Actionable conclusions from this report, including:

- Analysis of the reasons for the OCR accuracy results. For example, the age, condition, imaging technology or type and state of the text in the original can all affect OCR accuracy – this statistical method will help to identify the most likely causes and improve content selection.
- Recommendations for the means by which certain features of OCR performance can be optimized. For instance, for a publication with a large proportion of person names, improving the dictionary of names would improve OCR performance. This report will identify publications that require special attention.
- Results that could enable the comparison of different OCR engines to help the selection of the right technology.
- Analysis that will aid the assessment of a vendors' promised search engine performance against the likely performance suggested by the OCR accuracy results.

Linda Thomas

115 West 30th Street, Suite 400

New York, NY 10001 USA

T +1.212.461.3700 x55

M +1.703.297.1473

<http://www.digitaldividedata.org>

Simon Tanner

Director

King's Digital Consultancy Services

Web: www.digitalconsultancy.net

Email: simon.tanner@kcl.ac.uk

