Malaysian Journal of Computing, 4 (2): 335-348, 2019
Copyright © UiTM Press
eISSN: 2600-8238 online

THE COMPARISONS OF OCR TOOLS: A CONVERSION CASE IN THE MALAYSIAN HANSARD CORPUS DEVELOPMENT

Anis Nadiah Che Abdul Rahman, Imran Ho Abdullah,
Intan Safinaz Zainuddin and Azhar Jaludin
Faculty of Social Sciences and Humanities, Universiti Kebangsaan Malaysia, Selangor
P87706@siswa.ukm.edu.my, imranho@ukm.ukm.edu.my, intansz@ukm.ukm.edu.my, azharj@ukm.ukm.edu.my

ABSTRACT

Optical Character Recognition (OCR) is a tool in computational technology that allows a recognition of printed characters by manipulating photoelectric devices and computer software. It runs by converting images or texts that are scanned beforehand into machine-readable and editable texts. There are a various numbers of OCR tools in the market for commercial and research use, which are obtainable for free or restrained with purchases. An OCR tool is able to enhance the accuracy of the results which as well relies on pre-processing and subdivision of algorithms. This study intends to investigate the performances of OCR tools in converting the Parliamentary Reports of Hansard Malaysia for developing the Malaysian Hansard Corpus (MHC). By comparing four OCR tools, the study has converted ten reports of Parliamentary Reports which contains a number of 62 pages to see the conversion accuracy and error rate of each conversion tool. In this study, all of the tools are manipulated to convert Adobe Portable Document Format (PDF) files into Plain Text File (txt). The objective of this study is to give an overview based on accuracy and error rate of how each OCR tools essentially works and how it can be utilized to provide assistance towards corpus building. The study indicates that each tool possesses a variety of accuracy and error rates to convert the whole documents from PDF into txt or plain text files. The study proposes that a step of corpus building can be made easier and manageable when a researcher understands the way an OCR tool works in order to choose the best OCR tool prior to the outset of the corpus development.

Keywords*: Optical Character Recognition, PDF to text converter, Malay text converter, Corpus development, Malaysian Hansard Corpus*

Received for review: 16-04-2019; Published: 20-12-2019

1. Introduction

Optical Character Recognition (OCR) is a tool in computational technology. It enables a recognition of printed or written characters by manipulating photoelectric devices and computer software. OCR runs by converting images or texts that are scanned heretofore into machine-readable and editable texts. There are various numbers of OCR tools in the market for commercial and research use, which are obtainable for free or restrained with purchases. An OCR tool is able to increase the accuracy of the results which relies on pre-processing and subdivision of algorithms. The existence of OCR has assisted machine translation and enabled its users to convert

This open access article is distributed under a Creative Commons Attribution (CC-BY SA) 3.0 license

files from images or PDF into their desired output formats namely plain texts, Microsoft word format and many more.

An OCR tool is competent to work in academic or non-academic sphere. It is capable in assisting human to recognise texts or characters from image or scanned texts into editable and machine-readable to further exploit or analyse the texts. According to Davenport and Kirby (2016), machines still have disadvantages compared to human being in term of the capability to elucidate unstructured information or data better regardless of being smart and advanced. This is supported by Islam *et al.,* (2017) who stated that the brain of a human has the ability to indisputably recognise characters or texts from various sources including an image which machines could not efficiently do. Due to the inadequacy of machine's capability, increasing number of studies have been put forward to convert images to machine-readable and editable format. According to Herceg *et al.,* (2005), improved exactitude or accurateness in the OCR is able correspond to the better accomplishment of machine translation in regards to corpus development. According to Afli *et al.,* (2016), previous research on OCR error connections can be encompassed into three main categories which include (i) the enhancement of visual and linguistics approach by utilising scanned images, (ii) the integration of OCR system outputs in choosing the most accurate OCR tool, and (iii) correction of OCR output (post-processing technique). In our case of corpus development, the right selection of converter is crucial as it will enhance the process of a corpus development, especially when it involves high-volume documents. According to Richter *et al.,* (2018), the conversion's rate of error would critically enhance the usability the document for further analysis.

A corpus is defined as a collection of written or spoken and machine-readable text. A corpus file needs to be editable and in plain text format to be further processed. A plain text is defined as "the intelligible form of an encrypted text or of its elements." (Merriam-webster's Dictionary, 2019). Our corpus is a raw (plain) corpus with no mark up or annotation. To create the corpus, we need to convert all PDF to txt format in high volumes. Each of the files has approximately 60-160 pages. In our case, we have retrieved 3,511 files from Malaysian Hansard portal which archived the Malaysian Parliamentary Reports from 1959 (Parliament 1) until the last download in March 2018 (Parliament 13) to develop the Malaysian Hansard Corpus (Imran *et al.,*2018). Since the amount of files to be converted are in higher volume, it requires a compatible and reliable converter to suit the needs of the research as well as to expedite the output.

The core objective of this research is to compare the performance of selected OCR tools to find the most suitable and reliable tool in converting PDF to txt data to develop a Malay diachronic and specialised corpus (the Malaysian Hansard Corpus). According to Cambridge Dictionary (2018), diachronic means something that is related to changes or evolution, especially the one that is related to language. Initially, the study started with a number of OCR software to be tested and utilised. Following the pilot conversion of the data, 4 converters were selected for further analyses.

2. Basic Criteria in selecting the OCR Tools for the Study

There are certain criteria set in the study to determine the selection of the commercial OCR tools available in the market. The selection of commercial OCR is due to its availability and user-friendly trait it possesses. The first phase includes a general search on the web to resolve for a list of OCR tools. Subsequently, the list of selected OCR tools was shortlisted based on reviews or rating on the net. Nield *et al.,* (2019) for example, reviewed and compared some excellent commercial OCR tools on the market.

In addition to that, a pilot study was done using all of the shortlisted OCR tools based on pre-set parameters to meet the needs of the study. The tested OCR tool should be able to convert multiple pages at a time, possesses the ability to do conversion for big number of files. The process was undertaken to meet the need of the study and to adapt with the conversion process of our files. The application of all tools was made based on their trial versions. Based on the pilot study, 4 OCR tools were found out to suit the criteria of the research with the ability to convert multiple documents at a higher volume at a time. Thus, this paper compares 4 different types of OCR tools those are (i) 4Videosoft PDF Converter Ultimate, (ii) PDF to Text, (iii) Readiris Corporate 17 and

(iv) ABBYY FineReader 14. Brief introduction of each of the selected converters will be explained in the following sections.

2.1 4Videosoft PDF Converter Ultimate

4Videosoft PDF Converter Ultimate (4Videosoft Studio, 2018) is a professional converter to convert PDF files to Image, Text, Word, Excel, PowerPoint, ePub, HTML and many others (TXT, Word, RTF, JPEG, PNG, GIF, BMP, PCX, TGA, TIFF). It provides four interface languages: English, Japanese, French and German. It is able to convert multiple files in high speed and great quality. This software is also compatible with most gadgets like iPhone, PSP and other portable audio and video players.

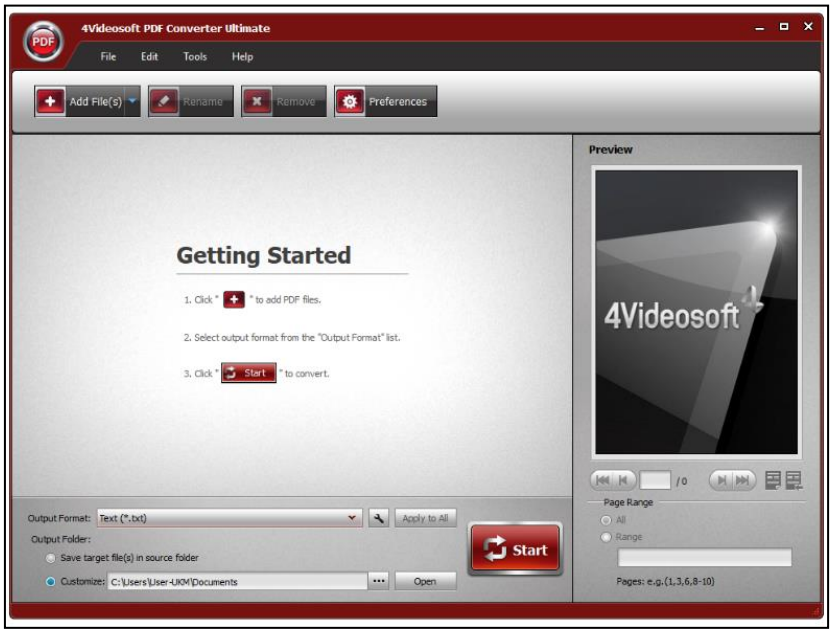


Figure. 1. The Interface of 4Videosoft PDF Converter Ultimate

2.2 PDF to Text

PDF to Text (Media4x, 2019) is an online and free software to covert PDF files into other output formats. It is able to speedily convert multiple files (up to 20 files per conversion). Results can easily be downloaded in a ZIP file. This online OCR tool supports 14 languages including Indonesian. This online OCR tool possesses other functions which include the ability to unlock, rotate, compress and merge PDF files other than multiple-format converting functions. The data submitted online will be removed after a subsequent hour of upload and conversion.

Figure. 2. The Interface of PDF to Text

2.3 Readiris Corporate 17

Readiris Corporate 17 (IRIS S.A, 2018) is a PDF and OCR publishing software. It enables conversion from PDF's, images, and texts in a various output format. Other than normal conversions, this software could also convert files into audio format like mp3 and wav as it has voice annotation and read-aloud functions. Its' other functions include creating and editing PDF files, putting annotations and comments, splitting, merging and compressing PDF files, importing and scanning from computers. It also recognises excel, numbers and calculation tables. This software supports 128 languages in its system.

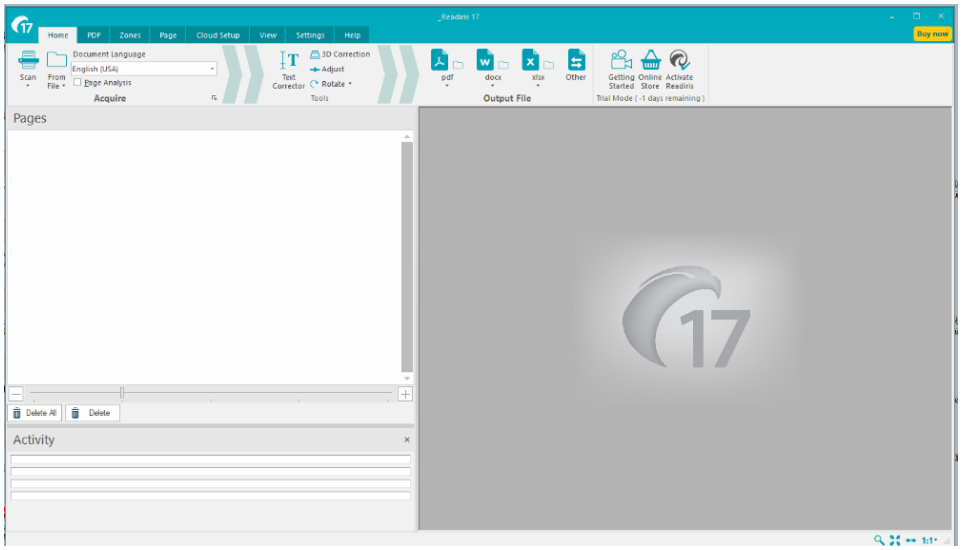


Figure. 3. The Interface of Readiris Corporate 17

2.4 ABBYY FineReader 14

ABBYY FineReader 14 (ABBYY, 2018) is claimed to be the most powerful OCR software that is available on the market for providing fast and precise text recognition (Kimari, 2018). According to Heliński, Kmieciak and Parkoła (2012), FineReader is found out to be marginally more accurate on characters level as compared to Tesseract, an optical character recognition engine in term of cleaned data. It is also capable to operate high volume data and correct labourious tasks. This OCR software supports 192 recognition languages. It also supports input formats like PDF, image formats and editable formats like DOC(X), XLS(X), PPT(X), VSD(X), HTML, RTF, TXT, ODT, ODS, ODP

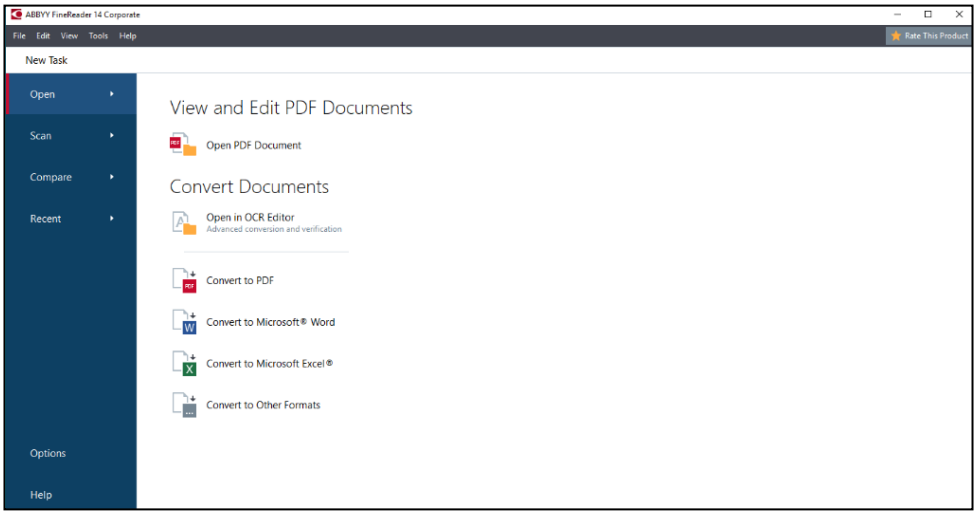


Figure. 4. The interface of ABBYY FineReader 14

2.5 Comparative Analysis

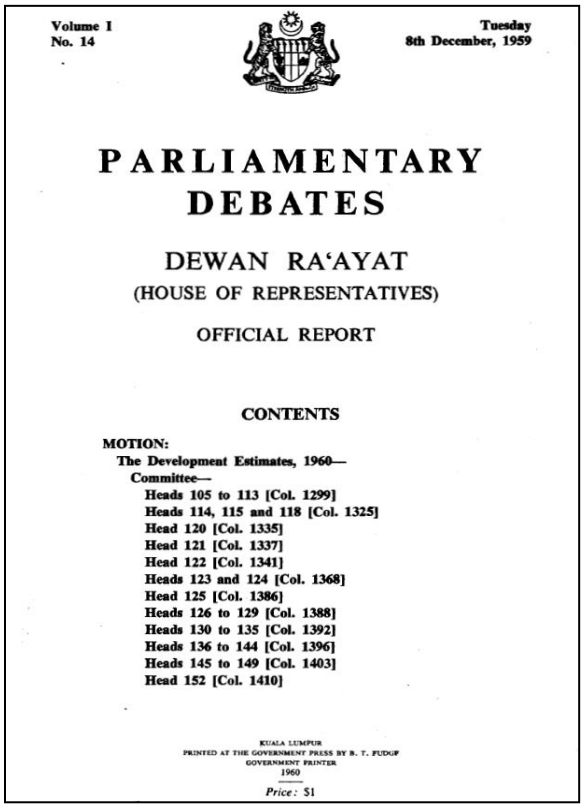
Based on the investigation on the four selected OCR tools, the comparisons are divided into several key features. The features include the availability online or downloadable version, the ability to recognise multiple languages, the ability to convert in bundles or multiple files and pages, and the fees of subscriptions or purchases. The comparison of the features can be seen in Table 1.

Table 1. General comparison of Selected OCR software

OCR Tool	**Online**	**Download**	**Multi-language recognition**	**Multiple files conversion/ bundled**	**Multiple pages conversion**	**Fee**
4Videosoft PDF Converter Ultimate		✓	4 basic languages	✓	✓	Free (limited access) Fees applicable
PDF to Text	✓		Not mentioned	✓ (up to 20)	✓	free
Readiris Corporate 17		✓	Various		✓	Free (limited access) Fees applicable
ABBYY FineReader (14)		✓	Various	✓	✓	Free (limited access) Fees applicable

3. Methodology

In general, there are 3,511 PDF files from Malaysian Parliamentary Report from the House of Representative. The total files cover Parliament 1(1959) to Parliament 13 (2018) (Imran *et al.*, 2018). Parliament 1 has become the main focus of this study. The selection of Parliament 1 was due to its general characteristics. Typically, there are several reporting formats for different Parliamentary Debates in Malaysia. This is due to the evolution of language and the shift in policy throughout the 60 years of parliamentary sessions held in Malaysia. Parliamentary Reports in Parliament 1 have consistent structure and formatting. The reports come in English and Old Malay language and the report has two columns each in one page. There are noteworthy noises from the documents that were scanned beforehand before being archived in the Malaysian Hansard Portal. The spelling of the Old Malay is different from the contemporary one which is similar to English characters. The sample of the PDF documents that should be converted can be seen in figure 5 and 6.



Volume I
No. 14

Tuesday
8th December, 1959

PARLIAMENTARY DEBATES

DEWAN RA'AYAT
(HOUSE OF REPRESENTATIVES)

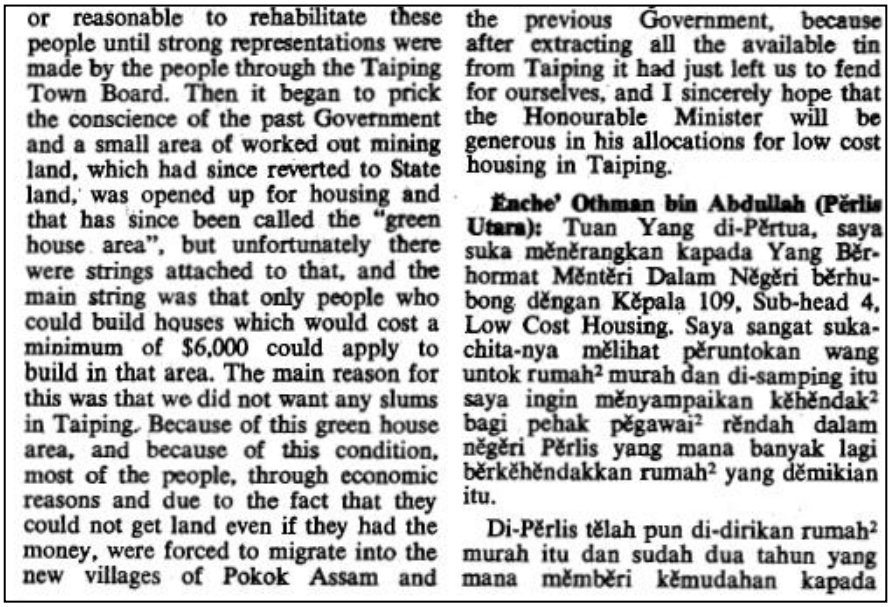
OFFICIAL REPORT

CONTENTS

KUALA LUMPUR
PRINTED AT THE GOVERNMENT PRESS BY B. T. FUDUP
GOVERNMENT PRINTER
1960
Price: $1

Figure. 5. Front Page of the Malaysian Hansard
Source: Dewan Rakyat Malaysia. (1959)

Figure 5 shows the front page of the Parliamentary Debates of House of Representative (*Dewan Rakyat*). Typically, all of the front page has coat of arms of Malaya/ Malaysia with the date, day, and volume, no of issue, the title, and contents. The very first parliamentary proceeding (Parliament 1) was convened from 1959-1964. It was officially conducted in English language thus the official reporting was done in English. Nevertheless, the speakers were free to use Malay with the permission from the Speaker.



or reasonable to rehabilitate these people until strong representations were made by the people through the Taiping Town Board. Then it began to prick the conscience of the past Government and a small area of worked out mining land, which had since reverted to State land, was opened up for housing and that has since been called the "green house area", but unfortunately there were strings attached to that, and the main string was that only people who could build houses which would cost a minimum of $6,000 could apply to build in that area. The main reason for this was that we did not want any slums in Taiping. Because of this green house area, and because of this condition, most of the people, through economic reasons and due to the fact that they could not get land even if they had the money, were forced to migrate into the new villages of Pokok Assam and the previous Government, because after extracting all the available tin from Taiping it had just left us to fend for ourselves, and I sincerely hope that the Honourable Minister will be generous in his allocations for low cost housing in Taiping.

Enche' Othman bin Abdullah (Perlis Utara): Tuan Yang di-Pertua, saya suka menerangkan kapada Yang Berhormat Menteri Dalam Negeri berhubong dengan Kepala 109, Sub-head 4, Low Cost Housing. Saya sangat sukachita-nya melihat peruntokan wang untok rumah² murah dan di-samping itu saya ingin menyampaikan kehendak² bagi pehak pegawai² rendah dalam negeri Perlis yang mana banyak lagi berkehendakkan rumah² yang demikian itu.

Di-Perlis telah pun di-dirikan rumah² murah itu dan sudah dua tahun yang mana memberi kemudahan kapada

Figure. 6. Sample of e page of Malaysian Hansard Report
Source: Dewan Rakyat Malaysia. (1959)

Figure 6 shows the middle page of the Parliamentary Debates of House of Representative. The verbatim reporting was reported purely on the exact words of the spoken reports or debates uttered by the Member of Parliament (henceforth MP). Hence, both English and Old Malay spellings can be easily seen on one page of the report. On the other hand, the middle pages of the Hansard reports also have two columns each. Each page stereotypically has the date and page number. In total, there are 182 files in Parliament 1 which cover five *penggal* or sessions. Each session consistently has two *Mesyuarat* (Meeting). In each Meeting, there are a different number of reportd which denote the day of parliamentary proceedings. The division of reports can be seen in Table 2.

Table 2. Division of Reports according to Meetings in Parliamentary Sessions

Penggal (Session)	Mesyuarat (Meeting)	Number of Reports
1	M1	10
	M2	2
2	M1	40
	M2	6
3	M1	24
	M2	22
4	M1	34
	M2	4
5	M1	32
	M2	8
TOTAL		182

For the purpose of this study, 10 samples from 182 PDF files (5.49 percent) were taken to understand the structure of each report. Each sample was taken from the first report in all Meetings. The selection of sample is based on the likelihood or random sampling by Riffe *et al.*, (2005). According to Riffe *et al.*, (2005), an individual item in a particular population has an equivalent chance to be chosen as a sample. In addition to that, a study on efficient sampling on five years' issues of consumer magazines by Riffe, Lacy and Fico (1998) proves that a selection of one document that is randomly being selected in each year, is more structured and has higher precision compared to selecting a random bigger number from overall documents. In the case of

big sample size, the study also applies the framework set by Somer (2010). Somer (2010) randomly selected 1200 files out of 42,463 newspapers articles to do content analysis which is equal to 2.83 percent of the total population. Due to that, a random sampling was chosen to be applied in this study in choosing the file samples from the parliamentary reports.

3.1 Measure for Recognition Performance through Error Recognition

The study utilises measure for recognition performance by Alexandrov (2003). The model underlines four types of errors in recognition process of an OCR which include (i) substitution, (ii) deletion, (iii) rejection and (iv) addition. Substitution occurs when one character is recognised as another. Substitution normally occurs for structurally adjacent characters. Deletion on the other hand, occurs when a character is being ignored as the OCR recognises it as noise. Rejection happens when the system could not distinguish a symbol or uncertain of the recognition.

Addition transpires when the OCR recognises one symbol as two, or when noise is being distinguished as a character or characters. According to the measures for recognition performance by Alexandrov (2003), there are several calculation that can be made. The measures include the global error rate, rejection rate, recognition rate and the level of reliability of the OCR systems.

3.1.1 The main measure: Global Rate

The main measure of this study is the main global rate or $Gerr$. $Gerr$ is determined based on number of committed errors and number of characters in the text.

$$Gerr = 100\ (ne/nc) \tag{1}$$

3.1.2 Rejection Rate

Rejection rate or $Grej$, is the measure of the registered errors which include the deletions, substitutions and additions. It can be calculated based on the following formula.

$$Grej = 100\ (nr/nc) \tag{2}$$

3.1.3. Recognition Rate

Table 3 shows the formula of recognition performance and its indicators. Recognition rate is frequently used for describing the efficacy of an OCR system. The equation is as follows:

$$Grec = 100\ (nc - nr - ne)/nc \tag{3}$$

Table 3. The formula's Indicator for Recognition Performance

No	Formula	Indicator
1	ne	Number of committed errors
2	nc	Number of all characters in the text.
3	nr	Number of rejections

3.1.4. Level of Reliability

Level of reliability shows the reliability of the overall results subsequent to the above measures. The equation is as follows:

$$Grel = 100\ (nc - nr - ne)/(nc - nr) = Grec/(Grec - Gerr) \tag{4}$$

The aim of the measurement is to maximise $Grec$ and to minimise $Gerr$. Generally, total percentage of the recognition based on $Grec + Gerr$. In order to see the types of errors in

recognition process of each OCR, the whole excerpt of required texts was copied to Microsoft Word to mark the errors based on (i) substitution, (ii) deletion, (iii) rejection and (iv) addition. The excerpt starts from the page no (1301) and ends with the beginning of the new page (1303). The errors were then marked according to colours to determine the errors. Yellow was tagged for substitution, red for deletion, blue for rejection and grey was tagged for addition. The tagging of the errors can be seen in the figure 7:

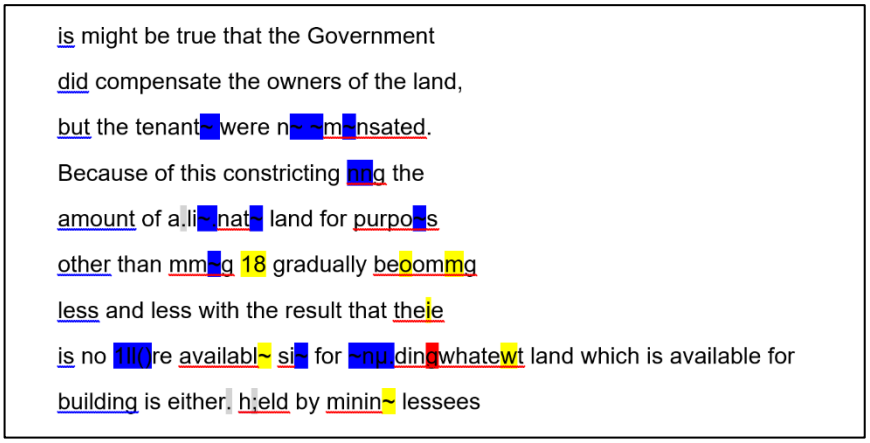


Figure 7. The sample of colour-coded tagging on OCR's recognition process

4.0. Results and Discussions

The results and discussions will be distributed according to the conversion results and the error of recognition results.

4.1. The Conversion Results

The following section will show the output of the conversion process.

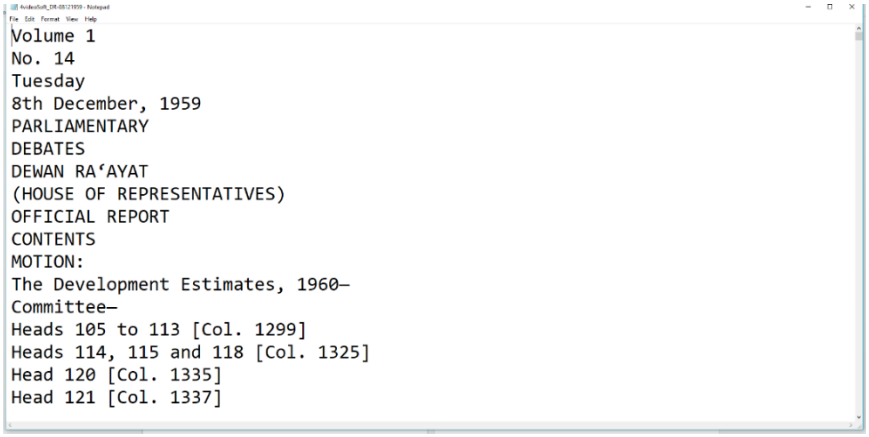

Volume 1
No. 14
Tuesday
8th December, 1959
PARLIAMENTARY
DEBATES
DEWAN RA'AYAT
(HOUSE OF REPRESENTATIVES)
OFFICIAL REPORT
CONTENTS
MOTION:
The Development Estimates, 1960–
Committee–
Heads 105 to 113 [Col. 1299]
Heads 114, 115 and 118 [Col. 1325]
Head 120 [Col. 1335]
Head 121 [Col. 1337]

Figure. 8. 4videosoft's conversion (front page)

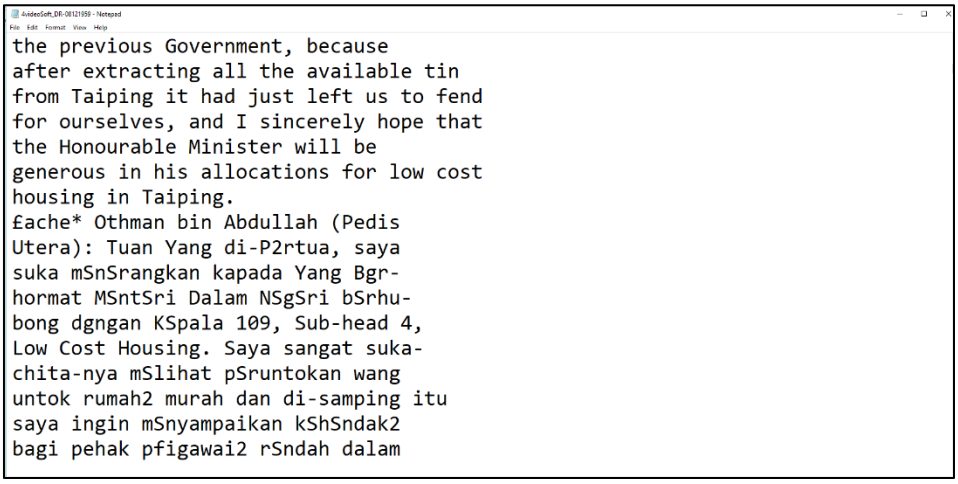


Figure. 9. 4videosoft's conversion (middle page)

Figure 8 and 9 show the sample of conversion results from 4videosoft. The converter is able to convert from PDF to plain text format including the noise. Some of the noises are converted to "■," "r" or "•." Based on Figure 10 however, it could be seen that the conversions are not precise for Malay words. The word *Mêntêri* (minister) for example, has been converted to *MSntSri* instead of *Menteri*. The word *bêrkêhêndak-kan* (equals to English word 'want'*)* has been converted to *b£rk£hgndakkan* instead of *berkehendakkan*. The other example is Ench*e'* (English word: Mr.) which is converted into £ache instead of Enche'.

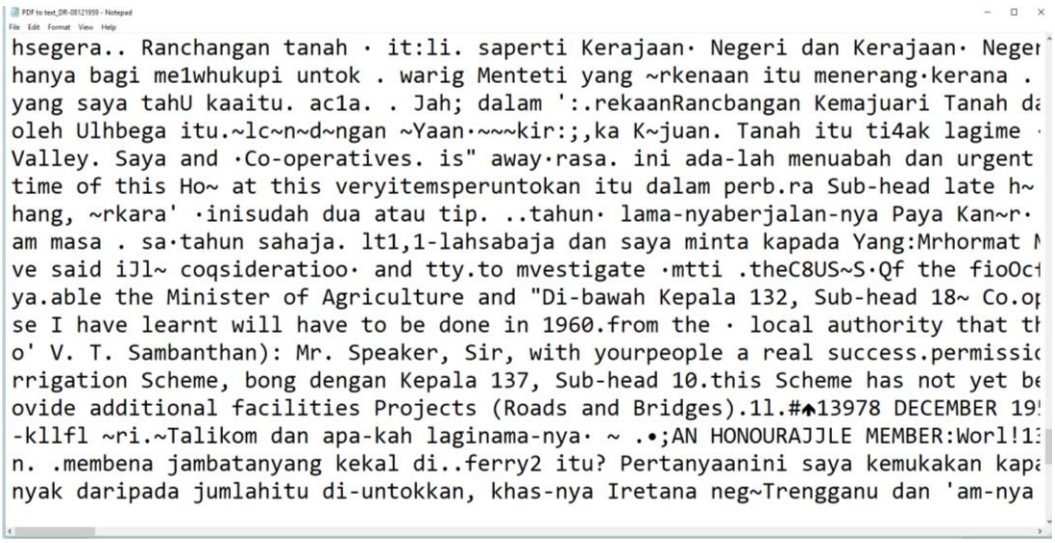


Figure. 10. PDF to Text's conversion

Figure 10 depicts the sample of conversion from PDF to Text. The converter identified the next page of the PDF and marked with the arrow symbol in target text (plain text file). There are also errors in spelling. For instance, the word *pêgawai* is converted into *P~gawa1* instead of *pegawai*. There are also characters like "#", "~*"* and "•"

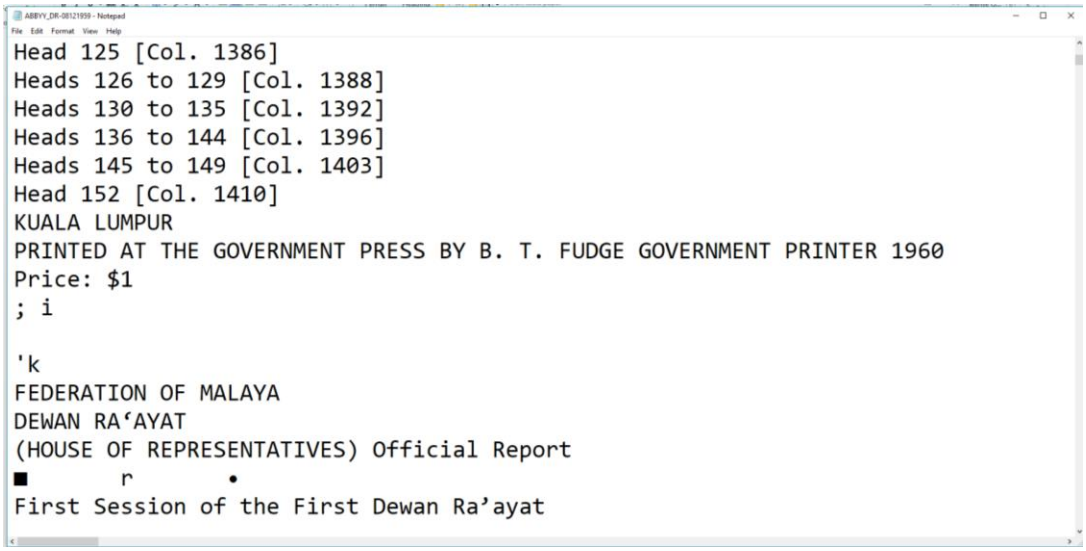


Figure. 11. ABBYY FineReader's conversion (front page)

Figure 11 shows the conversion result from ABBYY FineReader 14. Based on the conversion result, it can be seen that the converter also identified noise from the scanning of the PDF to certain characters like "k", "i", "■" and "•". The character recognition was caused by the original marks on PDF files that were previously scanned from papers earlier.

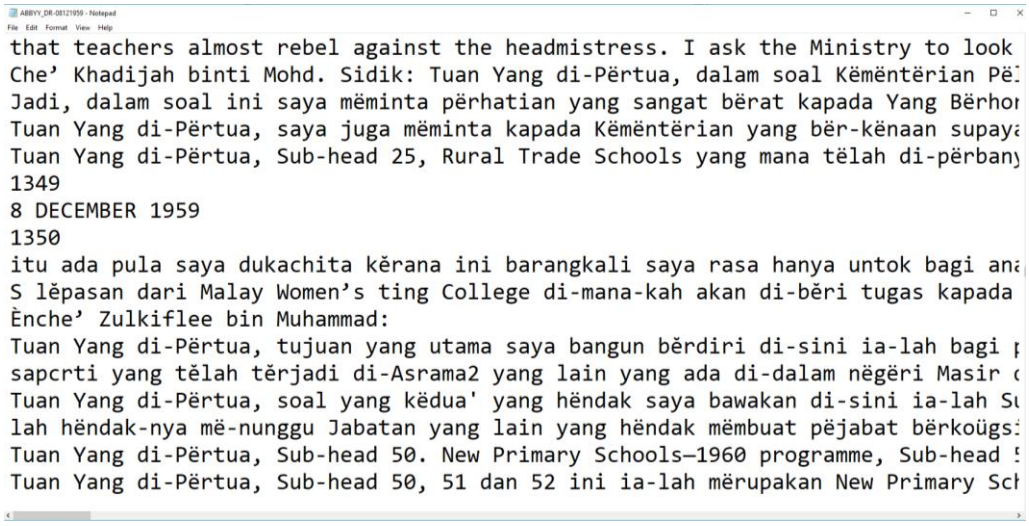


Figure. 12. ABBYY FineReader's conversion (middle page)

Figure 12 also depicts the conversion result from ABBYY FineReader 14. As the conversion language was also set to French and German, the converter was able to identify the Old Malay spellings with diacritics and hyphens. The word *mêminta* with acute *ê* for example, was converted into *mëminta* with umlaut *ë*.

4.2 Error Recognition Results

Table 4 shows the ratio of human recognition of errors based on the OCR's conversion of ABBYY FineReader 14, Readiris Corporate 17, Videosoft PDF Converter Ultimate and PDF to Text. The result is presented based on the ratio from ten samples. Based on the table, it can be seen that the converters have different character recognition on the same PDF file. PDF to Text has the highest substitution while ABBYY FineReader 14 has the lowest. Videosoft PDF Converter Ultimate has 0.7 deletion for the characters while PDF to Text has the highest deletion of 2.3. ABBYY

FineReader 14 gives no deletion to the conversion while PDF to Text has the highest rejection of 30.5. Readiris Corporate 17 gives the highest addition to the file (43.6) while ABBYY FineReader 14 gives only 2.0. Table 5 shows the performance of the 4 OCR tools based on its measures for recognition performance model by Alexandrov (2003).

Table 4. Ratio of the Human Recognition of Errors based on OCR's Total Sample' s Conversion

Tool	No of characters	Substitution	Deletion	Rejection	Addition
ABBYY FineReader (14)	4143.4	36.5	0.7	0.3	2.0
Readiris corporate 17	4187.8	37.4	0.9	12.6	43.6
Videosoft PDF Converter Ultimate	4214.4	38.8	0.9	1.9	4.8
PDF to Text	4160.7	44.0	2.3	30.5	34.1

Table 5. The Performance of OCR Tools based on Measures for Recognition Performance Model

Tool	Global error rate	Rejection rate	Recognition rate	Level of Reliability
ABBYY FineReader (14)	0.9533	0.0072	99.0394	99.0466
Readiris Corporate 17	2.2566	0.3009	97.4426	97.7366
4Videosoft PDF Converter Ultimate	1.1010	0.0451	98.8539	98.8985
PDF to Text	2.6654	0.7330	96.5847	97.3149

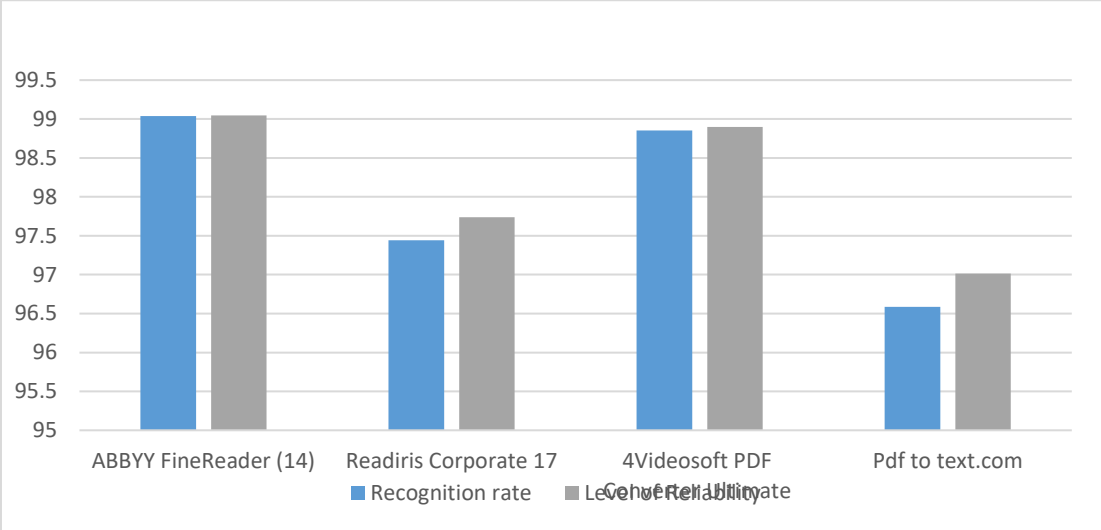


Figure 13. The Performance of OCR Tools based on Measures for Recognition Performance Model (Recognition Rate and Level of Reliability)

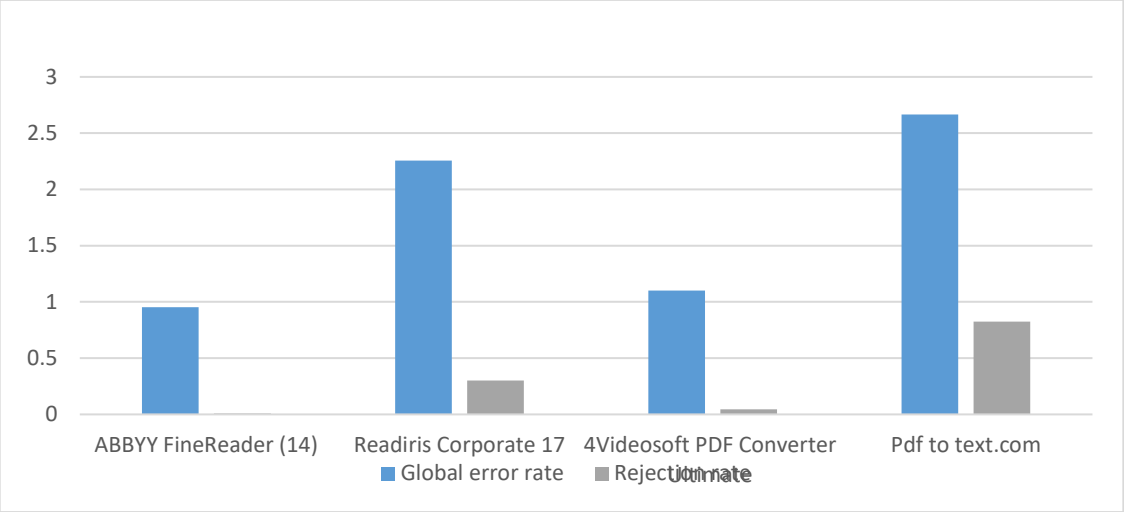


Figure 14. The Performance of OCR Tools based on Measures for Recognition Performance Model (Global Rate Error and Rejection Rate)

Based on figure 13 and 14, it can be seen that ABBYY FineReader produces the least global error rate followed by 4Videosoft PDF Converter Ultimate, Readiris Corporate 17 and PDF to Text. ABBY FineReader shows no rejection rate while PDF to Text has the highest rejection rate of all the tested OCR tools. ABBY FineReader also has the highest recognition rate while PDF to Text has the lowest. High recognition rate shows that the OCR tool is capable to recognise many different characters compared to the others. Finally, ABBY FineReader 14 is found out to have the highest level of reliability (99.1669) while PDF to Text has the lowest level of reliability of all of the converter with reliability rate of 97.0142.

5. Conclusion

This study has studied the performance of four different types of OCR tools in the case of developing the Malaysian Hansard Corpus in regard to converting a file of multiple language at a time. Based on the exploratory finding, ABBY FineReader 14 has the highest and most reliable performance based on Measures for Recognition Performance Model by Alexandrov (2003). The present results are significant as this performance of OCR test has enabled researchers to further perform the conversion of files in larger volumes to meet their individual needs. This analyses from the study may be useful to choose the most efficient converter in corpus development, especially towards utilising machine-translation approach to develop a diachronic corpus like the Malaysian Hansard Corpus.

Acknowledgement

This research is supported by Universiti Kebangsaan Malaysia under research grants AP-2014-018 and KRA-2018-005.

References

ABBYY. (2018). ABBY FineReader Version 14 for Windows. Retrieved from https://www.abbyy.com/en-apac/finereader/

Afli, H., Barrault, L., & Schwenk, H. (2016). OCR Error Correction Using Statistical Machine Translation. *International Journal of Computational Linguistics and Applications*. 7(1), pp. 175–191.

Alexandrov, V. (2003). Error Evaluation and Applicability of OCR Systems. *International Conference on Computer Systems and Technologies*. Retrieved April 13, 2018 from http://ecet.ecs.uni-ruse.bg/cst/docs/proceedings/S3/III-10.pdf

Davenport, H.T., & Kirby, J. (2016). Just How Smart are Smart Machines? *MITSloan Management Review*. 57(3). pp 20 – 26. Retrieved 27 March 2019 from https://pdfs.semanticscholar.org/a9d8/0b09f21d9d0306766d2c3ba2ce49b4b2b95b.pdf

Diachronic. (2018). In Cambridge Dictionary. Retrieved from https://dictionary.cambridge.org/dictionary/english/diachronic

Heliński, M, Kmieciak , M., & Parkoła , T. (2012). Report on the comparison of Tesseract and ABBYY FineReader OCR engines. *IMPACT*. Retrieved July 20, 2017, from http://lib.psnc.pl/dlibra/docmetadata?from=rss&id=358

Herceg,P., Huyck, B., Johnson, C., Van Guilder, L. and Kundu, A. (2005). Optimizing OCR accuracy for bi-tonal, noisy scans of degraded Arabic documents. *Visual Information Processing XIV*. 179-187. Retrieved April 4, 2019 from https://pdfs.semanticscholar.org/8ed7/76c183ba07bccd47de941077f1e3b18f962a.pdf

Imran, H.A., Anis Nadiah C.A.R., Azhar, J. (2018). Malaysian Hansard Corpus. Universiti Kebangsaan Malaysia.

Iris, S.A. (2018). ReadIris Version 17 for Windows

Islam, N., Islam, Z., & Noor, N. (2017). A Survey on Optical Character Recognition System. *Journal of Information & Communication Technology*. 10(2), pp.1-4

Kimari, K. (2018, January 03). 8 best OCR software for Windows 10. Retrieved April 12, 2018, from https://windowsreport.com/ocr-software-windows-10/

Media4x. (2019). PDF to Text. Retrieved from https://pdftotext.com

Nield, D., DeMuro, J. & Turner, B. (2019). Best OCR Software of 2019: scan and archive your documents to PDF. TechRadar. Retrieved 10 October 2019 from https://www.techradar.com/best/best-ocr-software

Somer, M.(2010). Media Values and Democratization: What Unites and What Divides ReligiousConservative and Pro-Secular Elites? *Turkish Studies.* 11(4), pp 555-577

Plaintext. (2019) In *Merriam-Webster*, Merriam-Webster. Retrieved from www.merriam-webster.com/dictionary/plaintext.

Richter C., Wickes, M., Beser, D. & Marcus, M. (2018). Low-resource Post Processing of Noisy OCR Output for Historical Corpus Digitisation. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation.* (2331-2339). Miyazaki: European Languages Resources Association (ELRA)

Riffe, D., Lacy, S., & Fico, F. G. (1998). Analyzing Media Messages: Using Quantitative Content Analysis in Research. London: Lawrence Erlbaum Associates.

Riffe, D., Lacy, S., & Fico, F. (2005). Analyzing Media Messages: Using Quantitative Analysis in Research. Mahwah, NJ: Lawrence ErlbaumAssociates.

4Videosoft Studio. (2018). PDF Converter Ultimate for Windows. Retrieved from https://www.4videosoft.com/pdf-converter-ultimate.html