OCR 処理生成によるテキストデータと現物書籍との照合分析（2）

【対象書籍】

『近代日本と仏教』『生きる力を体で学ぶ』『編集とはどのような仕事なのか』（以上、全てトランスビュー刊）

【調査方法】

上記 3 冊の書籍の本文 20 ページ分の PDF と関連付けられたテキストデータに対して、校正者により現物書籍を元原稿として視認照合による校正を行い、差異を赤字で記入する。

発生した赤字を、「記号」「日本語」「英語」「数字」の別にカウントし、全体の文字数で除して処理の全体精度を把握する。

照合事例から共通に読み取れるもの、特殊な事例だと思われるものを分析者が読み込み、レポートする。

【照合結果カウント結果・照合精度】

書名	校正文字数	赤字					
		記号	英字	数字	日本語	赤字合計	赤字合計／校正文字数
近代日本と仏教	13,560	43	0	0	4	47	0.3%
生きる力をからだで学ぶ	13,607	5	0	0	2	7	0.1%
編集とはどのような仕事なのか	12,670	19	1	0	4	24	0.2%
計	**39,837**	**67**	**1**	**0**	**10**	**78**	**0.2%**

1 ページあたりおよそ 600〜700 字の単行本 3 冊の本文冒頭 20 ページのテキストデータの文字数合計は 39,837 字。

赤字発生件数は 78 字。
約 99.8%の精度でテキストデータは本文を再現している。

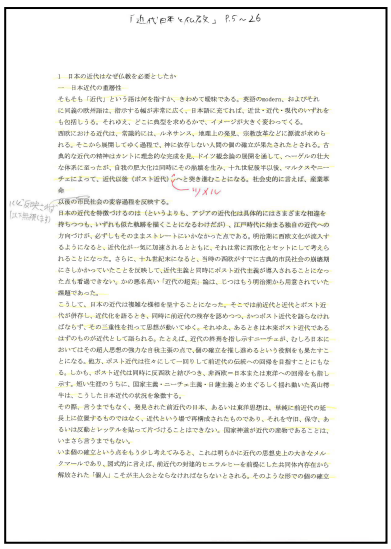
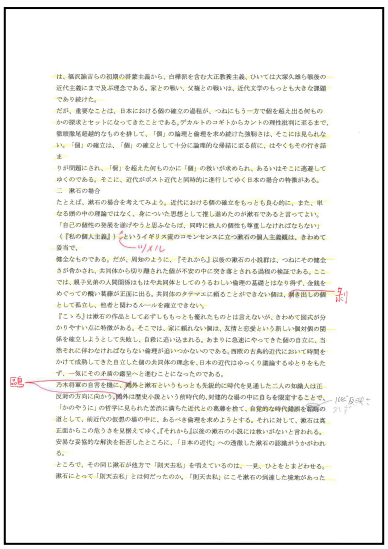
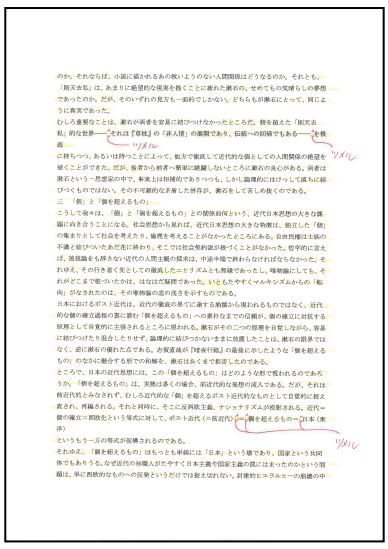
赤字合計は 78 字。
日本語（ひらがな、カタカナ、漢字）で発生した赤字は 10 字（赤字中約 12.9%）。
記号（パーレン・ナカグロ・句読点等）で発生した赤字は 67 字（赤字中約 85.9%）。

英字で発生した赤字は 1 字（赤字中約 1.2%）。
数字で発生した赤字は、なし。

記号部分の赤字の多くは本文になかったアキ（＝空白スペース）であることを考えると、ほぼ完全に本テキストデータは本文を再現している。

単純な OCR 処理で生成されたテキストだとすれば驚異的であるが、校正→修正の工程を経ているのではないかと予想される。

【テキストデータにおける一般的傾向】
ルビは読まない。
記号の赤字のほとんどは、本文になかったアキである。
ほぼ完全に本文データを再現していると言える。

『近代日本と仏教』（トランスビュー刊）先頭の 3 ページ分　赤字画像

（校正・レポート／株式会社ゼロメガ 2011/10/20）