

Abbyy accuracy, overview

Hank Bromley
3 December 2007

As a follow-up to last week's comparison of ocr accuracy in normal mode vs '-fast' mode of Abbyy version 8, today I compared each of those to Abbyy version 6. Using the same sample pages and the same reporting format, I created two more documents like last week's; the full set of three is now:

- **abbyy_8_accuracy** (last week's)
compares Abbyy 8 normal with Abbyy 8 fast
this document has been revised to reflect a few additional ocr errors I hadn't noticed last week
- **abbyy_6vs8_accuracy**
compares Abbyy 6 normal with Abbyy 8 normal
- **abbyy_6vs8fast_accuracy**
compares Abbyy 6 normal with Abbyy 8 fast

The table below provides an overview of all the results. **On the average, ver. 8 normal produced the fewest errors, with ver. 6 normal producing 1.24 times as many, and ver. 8 fast producing 1.95 times as many.** But as indicated by the color coding in the table, the variability was high, with each version being the best performer on at least one page, and also the worst on at least one page.

leaf	OVERVIEW OF RESULTS			Notes
	errors in 6 normal	errors in 8 normal	errors in 8 fast	
departmentalbull00nortuoft				
0004	2	4	6	ver. 8 was mostly repeats of a single error ver. 8 recognizes mixed-case small caps
0005	3	2	5	
0006	2	2	1	
subtotal for book	7	8	12	
newvarietyofcret3716rich				
0009	2	1	1	all were mostly repeats of a single error
0012	7	7	8	
subtotal for book	9	8	9	
Jewishproblemhow00branrich				
0006	3	4	13	
0009	7	1	7	
subtotal for book	10	5	20	
totals	26	21	41	
errors per page	3.7	3.0	5.9	
errors per 1000 chars	1.7	1.4	2.7	15,000 characters total, or ~2000 per page