Table 1

Collection	Original documents	Scanned images	Word tokens
British House of Commons Parliamentary Papers (ProQuest)	**118,526**	**6,448,739**	**4,672,737,402**
Early Canada Online (Canadiana.org)	**83,016**	**3,938,758**	**2,066,746,780**
Kew Gardens Directors' Correspondence	**24,765**		**4,550,118**
Confidential Prints Collection: Africa, Latin America, Middle East, and North America (Adam Matthew Online)	**1,315**	**140,010**	**206,389,401**
Selected Books and Documents	**306**		**5,123,458**
Total	**227,928**	**10,527,507**	**6,955,547,159**

Table 1 Caption:

The Corpus. Scanned images are the pages produced by optical character recognition and generally correspond to pages in the source texts. Word tokens are the strings that are recognized by our text processing software.