TITLE

« Text mining a noisy corpus of Ancien Régime French periodicals: a war story »

AUTHOR

François Dominic Laramée
Doctoral candidate in History, Université de Montréal (Canada)

ABSTRACT

Compared to the British, the Spanish and the Portuguese, the French sent notoriously small numbers of colonists to their Ancien Régime Atlantic empire. The French government's fear of depopulating the kingdom undoubtedly contributed to this phenomenon, but does it fully explain the lack of enthusiasm on the part of the French subjects themselves?

In 2006, historian Yves Landry postulated that a negative image of the colonial world in French print media might also have had a dampening influence, but no studies to date have been conducted to confirm or invalide his hypothesis. To answer the question, I have performed a dual-scale analysis, part digital and part traditional, of a corpus made up of the three main Ancien Régime French periodicals: the *Gazette*, the *Mercure de France* and the *Journal des Savants;* other sources that reached audiences as large or larger, like almanachs and the inexpensive « blue books » sold in the countryside, have proven largely silent on the matter. My study focuses on the period between 1740 — when the French print industry started growing at a rapid pace — and the end of the Seven Years' War.

However, text mining 18th-century French periodicals poses significant methodological challenges, due in no small part to the unreliable quality of the OCR'ed source material, which compounds the problems linked to the irregular spellings of the time. The digital historian working with such data must proceed with extreme caution. My method involved using Levenshtein's algorithm to identify keyword tokens that might have been damaged by OCR or « hidden » by unusual spellings; careful extraction of meaningful press articles and reconstruction of metadata by hand; and letting the data dictate the specific angles of research heuristically. It turns out that relatively simple and well-established digital methods, including token counts, metadata analysis and cooccurrence networks, were sufficient to produce glaring results unlikely to be compromised (or caused) by poor data quality. These results then guided close reading of key parts of the corpus and yielded promising interpretations.

For example, in the *Gazette,* the overwhelming majority of articles discussing the Americas presented them from a foreign, usually British, point of view; continental New France was almost invisible in peacetime; and the Caribbean colonies were shown as fraught with the constant danger of privateering and piracy. A French reader would have found precious little incentive to pack up and move across the Atlantic in this periodical — and many reasons not to. Work is ongoing, but it looks like the corpus under study largely supports Yves Landry's original hypothesis.