Melissa Baralt
Ph.D. candidate
Department of Spanish and Portuguese
Georgetown University
Research Focus: Languages and Linguistics

SUMMARY

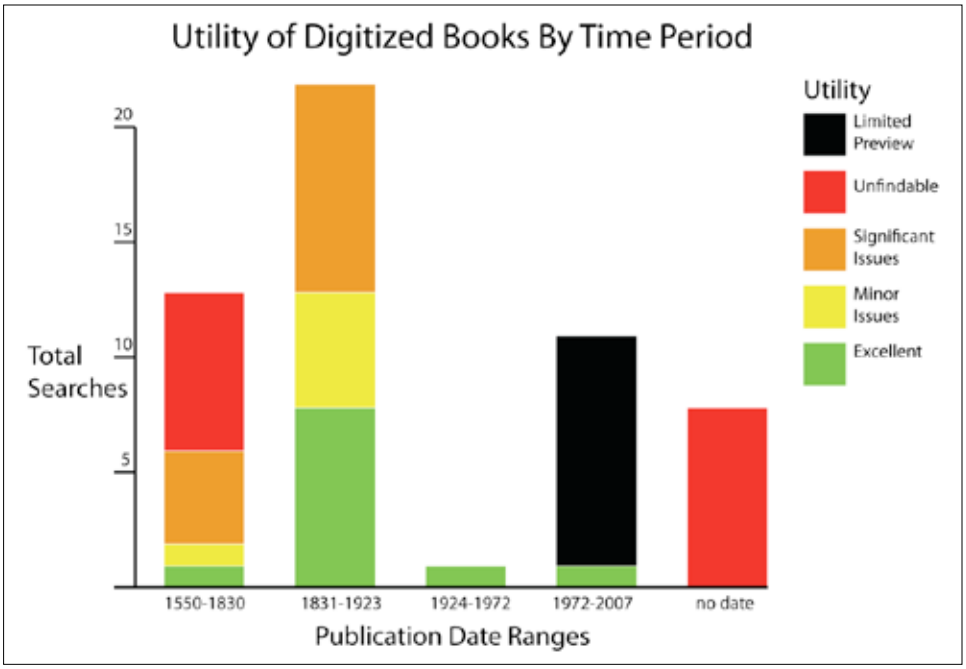
Scope

This document distills and summarizes the results of Melissa Baralt's report on the quality and availability of digitized books in linguistics and historical texts about languages. Her research reports almost exclusively on the availability of books in Google Book Search (GBS).

Overview

Over the summer of 2008, Baralt searched for digitized copies of 61 books related to linguistics and the history of languages. Her searches included a wide range of books published between 1553 and 2007. Of her 61 searches, 72 percent succeeded in finding digitized copies of the works. An overview of the results of her searches for works from different time periods can be found in the figure below.

For the purpose of this analysis, digitizations considered to have *excellent utility* were available in full, had at most one unintelligible scanned page, and were fully indexed. Books with *minor issues* were available in full, had one or two unintelligible pages, and included a range of minor issues for search dealing with character recognition. Digitizations with *significant issues* either had two or more unintelligible pages, or were hindered by widespread issues with search. 41



percent of the books included between one and eight illegible pages. 48 percent of the books had issues with search, many because of problems with non-English characters. Searches marked as *unfindable* are works that Baralt could not find in any format online. Digitizations labeled *limited preview,* while of high quality, are only available in 20 percent preview chunks. The works listed without a date are works Baralt searched for but did not find or provide publication information about.

Key Findings

72% of works were available in some fashion, primarily through GBS. Most of the works Baralt searched for were available. She rated all of these books easy to find. Other than one book in Proquest and one in Georgetown's library site, all of the books in her analysis were accessed through GBS.

Finding particular editions or translations was difficult. Most of Baralt's failed searches are the result of attempts to find copies of specific editions or translations of a given work. The numbers of unfindable works would have been much higher if she had consistently searched for a specific edition or translation of a work. Similarly, if she had always accepted different editions or translations of a work, she would have found versions of nearly all of the books. This is particularly pronounced in attempts to find original copies of works in non-English languages. Several multivolume works were mislabeled.

Primary interest and highest success was with works published between 1830 and 1924. Baralt was most interested in, and had the most success in finding, fully usable copies of works published between 1830 and 1924. She found nearly all of the books she sought during this period. She had much less success with works published before 1830, where she found less than half of the works she searched for.

Significant issues arose with character encoding and searchability in other languages. Works written entirely in English included minimal OCR problems and minimal errors in searches. Works written entirely in Portuguese and Spanish presented some problems with search. Works that included Greek and Chinese characters, as well as linguistic symbols for phonetic transcription, were particularly poor for search. In the case of a book in a Tibetan language, this problem was exacerbated by Baralt's inability to enable her computer to generate Tibetan characters to attempt to conduct a search.