

MASTER'S THESIS

Corpus Wrangling: A Comparative Analysis of Available Methodologies to Overcome Data Limitation in Corpus-based Interpreting Studies

GABARRON BARRIOS, Fernando

Date of Award:
2022

[Link to publication](#)

General rights

Copyright and intellectual property rights for the publications made accessible in HKBU Scholars are retained by the authors and/or other copyright owners. In addition to the restrictions prescribed by the Copyright Ordinance of Hong Kong, all users and readers must also observe the following terms of use:

- Users may download and print one copy of any publication from HKBU Scholars for the purpose of private study or research
- Users cannot further distribute the material or use it for any profit-making activity or commercial gain
- To share publications in HKBU Scholars with others, users are welcome to freely distribute the permanent URL assigned to the publication

Abstract

This dissertation proposes a methodological approach to constructing, reusing, and sharing interpreting corpora. Interpreting corpora are valuable linguistic resources used to test hypotheses and validate interpreting theories. Data limitation in Corpus-based Interpreting Studies (CIS) has been a problem during the past decades, mainly due to the time-consuming and resource-intensive nature of collecting, transcribing, and annotating interpreting data. This problem hampers empirical research in Interpreting Studies. The focus of the present study is threefold: to describe the different approaches taken during the past decades to corpus construction, usage, and sharing in CIS; to propose a methodology (specifically “corpus wrangling”), and a documentary analysis of this methodology to overcome data limitation in CIS; and to illustrate with examples the proposed automatic methods to restructure and consistently merge interpreting corpora, in order to obtain larger, more informative, reusable, and open access datasets. Results show that: 1) data structure transformation can be applied automatically to unstructured interpreting data to make interpreting corpora compatible, reusable, and open access; and 2) corpora with similar data structures, representativeness, balance, and theoretical convergence, can be merged, thus creating larger and more informative datasets. The proposed methods could be applied to avoid the time-consuming task of encoding and managing data manually. Compatibility, reusability, and sharing of interpreting corpora could contribute to moving forward in CIS. Future studies could leverage the proposed methodology to wrangle entire interpreting corpora. This study does not seek to provide standards to overcome data limitation in CIS, it just raises a new methodological perspective on a lingering problem.