Finding Haiku – Enhancing Findability and Accessibility of Poetry Resources in Multi-genre Collections across Different Languages

Mrugalski, Michał
michal.mrugalski@hu-berlin.de
Humboldt University of Berlin, Germany

Charvat, Vera Maria
veramaria.charvat@oeaw.ac.at
Austrian Centre for Digital Humanities and Cultural Heritage, Austrian Academy of Sciences

Börner, Ingo
ingo.boerner@uni-potsdam.de
University of Potsdam, Germany

Durco, Matej
matej.durco@oeaw.ac.at
Austrian Centre for Digital Humanities and Cultural Heritage, Austrian Academy of Sciences

Laszkovits, Sabine
sabine.laszakovits@oeaw.ac.at
Austrian Centre for Digital Humanities and Cultural Heritage, Austrian Academy of Sciences

Resch, Stefan
stefan.resch@oeaw.ac.at
Austrian Centre for Digital Humanities and Cultural Heritage, Austrian Academy of Sciences

Abstract

The refinement of data curation and selection takes a fundamental part in Computational Literary Studies. Using the exemplary case of scanning European text collections for instances of the genre “haiku”, this short presentation will illustrate the need for the methods and tools being developed in the CLS INFRA project[1] in order to enhance data findability and accessibility according to the FAIR framework (findable, accessible, interoperable, reusable; Wilkinson et al. 2016). Specifically, it will show which procedures are necessary to optimally structure literary corpora for the CLS community.

Problem Statement

To all appearances, the haiku as a well-established and strictly prescribed short form should be easy to identify in multi-genre and multilingual collections. Yet, difficulties arise from both cultural (Hokenson 2007; Johnson 2011; #niecikowska 2016) and infrastructural factors. Given the stringent rules of application for Japanese phonetics (mora[2]*kireji*[3]), syntax, and semantics (*kigo*)[4], the genre of the haiku appears to be untranslatable into European literary systems. All European haikus are merely approximations of the precise form contingent on the specific structure of the Japanese language. This impossibility of identifying the canonical version of the haiku outside Japanese literature, paired with the heterogeneity of the existing infrastructure of text collections, makes the search for sustainable data extremely time and resource consuming. The following questions arise:

Which features of different European approaches to the strict Japanese genre should be given priority in identifying a poem as a haiku?

How can the evaluation and exploration of literary text collections be facilitated in regard to these features?

Adapted Approach

To answer these questions, a number of corpora[5] from a list of literary resources (collected and curated since the beginning of the CLS INFRA project as part of a deliverable) were selected for closer examination.

Relying on metadata- and data-derived information, the corpora were scanned for both “explicit” and “implicit” (“adaptations” according to Long and So 2016) realizations.

Explicit haikus are either translations of the Japanese models or original creations labeled as “haiku”. Their findability relies heavily on metadata provided by corpus compilers/distributors. Two telling cases are Wikisource[6] and the Russian National Corpus,[7] which recognize the haiku as a distinct form, but only in annotations to separate poems and not in the inventory of genres.

As for implicit haikus, their findability depends on the addressability of certain textual features (such as: verses, syllables, and topics) ensuing from the structure of a corpus.

Example: As a result of favoring the cultural universal of verse count over the relation to the number of syllables (absent in Japanese, instead divisible in mora), an algorithm was devised that extracts three-verse poems from the DLK (German Lyrik Corpus, cf. Haider 2021). In the DLK syllables are already marked-up, which makes counting syllables possible and thus provides an additional criterion to identify possible haikus. However, the ideal structure of 17 (5+7+5) syllables could not be identified.

Two problems arise: inconsistencies in the recording of metadata and discrepancies in the structural annotation of texts in literary corpora.

Proposed Solution

Based on implementation concepts being developed within the CLS INFRA project, the following solution is proposed for the aforementioned problems: a descriptive metamodel for literary corpora will be introduced based on the “Metamodel for Corpus Metadata” (MCM; Odebrecht 2018), allowing a fine-grained genre-specific description of collections. This will be implemen-

ted in a catalogue of literary corpora, which will contain explicit statements about the various features of the individual corpora - especially information about the inner structure of the documents, e.g. the presence of explicitly annotated verses.

The model will be realized as an extension to CIDOC CRM,[8] using other well-established ontologies, e.g., FRBRoo[9] or Postdata's Ontopoetry Core Ontology[10]. This will be extended with reference resources, such as the Dewey Decimal Classification[11] to provide controlled vocabularies for categorizing the corpora and documents. Recording provenance information, the model will allow to express even conflicting statements about a resource, which can be traced back and assessed individually.

In the haiku-example, this approach will be tested by attaching information to corpora such as tokenization rules, syllable structure, verse structure, etc.

Notes

1. "Computational Literary Studies Infrastructure" (CLS INFRA, No. 101004984) is a Integrating Activities for Starting Communities (IASC)-project, launched in March 2021 and funded by Horizon Europe 2020 (Call: H2020-INFRAIA-2020-1) for the duration of 48 months. Its overall goal is to create uniform and easy access to the best European and national infrastructures for the CLS community. See also: https://clsinfra.io/(last accessed: 2022-10-26)
2. A unit of duration in Japanese used to measure the length of words and utterances; mora languages differ from syllable languages like English.
3. Cutting syllable, separating the first verse from the others.
4. Codified signals of seasons: plus 1000 lexemes contained in the *saijki*, a prescriptive list of such words.
5. The history of our searches is provisionally stored in Gitlab (please note that the access at this time is limited to CLS INFRA project members): https://gitlab.clsinfra.io:11435/cls-infra/haiku-challenge; https://gitlab.clsinfra.io:11435/cls-infra/haiku-challenge/-/issues
6. https://wikisource.org/wiki/Main_Page (last accessed: 2022-10-26).
7. https://ruscorpora.ru/new/search-poetic.html (last accessed: 2022-10-26).
8. https://cidoc-crm.org/ (last accessed: 2022-10-26).
9. https://cidoc-crm.org/frbroo/home-0 (last accessed: 2022-10-26).
10. https://postdata.linhd.uned.es/ontology/postdata-core/documentation/index-en.html (last accessed: 2022-10-26).
11. https://www.oclc.org/en/dewey.html (last accessed: 2022-10-26).

Bibliography

Haider, Thomas (2021): " Metrical Tagging in the Wild: Building and Annotating Poetry Corpora with Rhythmic Features " . Proceedings of the European Association for Computational Linguistics, arXiv:2102.08858

Hokenson, Jan Walsh (2007): "Haiku as a Western Genre. Fellow-Traveller of Modernism". Eysteinsson, Astradur/ Liska, Vivien (eds.), Modernism. Amsterdam/Philadelphia:, vol. 2, 693-714.

Johnson, Jeffrey (2011): Haiku poetics in Twentieth-Century Avant-Garde Poetry. Langham et al.: Lexington Books.

Long, Hoyt / **So, Richard J.** (2016): "Literary Pattern Recognition: Modernism between Close Reading and Machine Learning". Critical Inquiry 42 (Winter 2016): 235–67.

Odebrecht, Carolin (2018): "MKM – ein Metamodell für Korpusmetadaten. Dokumentation und Wiederverwendung historischer Korpora", Dissertation. Humboldt-Universität zu Berlin, Sprach- und literaturwissenschaftliche Fakultät, Berlin. doi: https://doi.org/10.18452/19407

Śniecikowska, Beata (2016): Haiku po polsku. Genologia w perspektywie transkulturowej. Toruń: Wydawnictwo UMK.

Wilkinson, Mark D. / **Dumontier, Michel** / **Aalbersberg, IJsbrand J.** / **et al.** (2016): "The FAIR Guiding Principles for scientific data management and stewardship". Sci Data 3, 160018. doi: https://doi.org/10.1038/sdata.2016.18