

# Project Charter

## PROJECT OVERVIEW

PROJECT NAME	Multilingual Semantic Embeddings (MuSE)		
PRIMARY INVESTIGATOR	Anna Yu Wang Jürgen Hackl	GRANT	CDH Research Partnership
EXPECTED START DATE	Jan 19, 2026	EXPECTED END DATE	Jun 17, 2026

## PROJECT DETAILS

PROJECT ABSTRACT	<p>MuSE aims to evaluate current methods of relating music-theoretical terms across languages to answer the following research questions: In what ways does machine translation using multilingual LLMs preserve or flatten the nuance of different music theory traditions? With that in mind, how can we reliably automate the discovery of related terms in different languages?</p> <p>To answer these questions, we will first test automated translation on a subset (7) of scholar-translated articles from <i>Music Theory Online</i> volume 30 number 4 (written in <b>Chinese</b>, <b>Japanese</b>, <b>Portuguese</b>, and <b>Spanish</b>, with <b>English</b> translations) and assess those translations both qualitatively by domain experts and quantitatively using metrics such as BLEU scores. Second, we will use the non-English articles, the full-text articles from the <i>Music Theory Online</i> archive, and a set of human-curated terms/definitions to compare methods for clustering terms and concepts across languages. This comparative analysis will identify effective methods for surfacing discursive connections within this specialized content.</p>		
PURPOSE/AUDIENCE	<p>MuSE will provide the <b><i>Music Theory in the Plural</i> team</b> with an empirically-grounded methodology for linking music-theoretical concepts across languages, as well as for decisions about whether and when to use large language models when translating domain-specific discourse. This foundation will inform the project teams' strategic direction and strengthen future grant applications. This study also provides the broader <b>digital humanities community</b> critical and notably absent comparative research on how large language models perform for translation and clustering with specialized scholarly content in comparison to established machine learning methods. The resulting papers will attempt to address questions of broader implications for the use of multilingual LLMs within scholarly domains, as well as the potential applicability of the methods discussed to other domains and research questions.</p>		
DELIVERABLES (RSE TEAM)	<ul style="list-style-type: none"><li>Draft of a short paper (ACL format or similar) comparing the quality of the automated and scholarly translations for relevant conference and/or</li></ul>	Writing (other) ▾	Must-Have ▾

	<p>publication</p> <ul style="list-style-type: none"> <li>• Draft of a short paper (ACL format or similar) on methodological comparison for multilingual clustering to relevant conference and/or journal</li> <li>• GitHub pages website (mkdocs) with a Princeton url with proof-of-concept notebooks to serve as demo for future funding.</li> <li>• A GitHub repository containing code and minimal technical documentation (install dependences, run/view experiments)</li> </ul>	<p>Writing (other) ▾</p> <p>Visualization ▾</p> <p>Code ▾</p>	<p>Must-Have ▾</p> <p>Must-Have ▾</p> <p>Must-Have ▾</p>
OUT OF SCOPE	<ul style="list-style-type: none"> <li>• Training a new large language models</li> <li>• Fine tuning an existing multilingual language model</li> <li>• Unit testing and robust documentation (e.g. API documentation, user-facing documentation, tutorials)</li> <li>• Creation of a website or hosted tool for a general user</li> <li>• Expansion beyond initial language set</li> <li>• Maintenance of published code or bug fixes after delivery</li> </ul>		
SUCCESS CRITERIA	<ul style="list-style-type: none"> <li>• We produce empirical evidence evaluating automated translation quality against scholar-translated benchmarks, and provide a recommendation on the effectiveness of embedding models versus topic modeling for discovering cross-linguistic term relationships in music theory texts.</li> </ul>		
SUSTAINABILITY PLAN	<ul style="list-style-type: none"> <li>• The code will be made publicly available under an open license so that <i>Music Theory in the Plural</i> team members and/or their future collaborators can build on it. CDH is not responsible for building, maintaining, or fixing code after the partnership concludes.</li> </ul>		
PLANNED POST-PARTNERSHIP RESEARCH ACTIVITIES (PROJECT TEAM)	<ul style="list-style-type: none"> <li>• Music Theory in the Plural Web Platform</li> <li>• Article describing the value this project brings to music theory, including music theorists as co-authors</li> </ul>	<p>Website/Inter... ▾</p> <p>Journal Article ▾</p>	<p>Must-Have ▾</p> <p>Must-Have ▾</p>

	<ul style="list-style-type: none"> <li>Grant proposal for following phase of the project based on proof of concept developed during this current collaboration</li> </ul>	Writing (other) ▾	Must-Have ▾
	<ul style="list-style-type: none"> <li>Develop an interactive platform that visualizes the multilingual semantic landscape.</li> </ul>	Visualization ▾	Should-Have ▾
	<ul style="list-style-type: none"> <li>Integrate embeddings and graph analytics using pathpyG or PyTorch Geometric.</li> </ul>	Code ▾	Could-Have ▾
	<ul style="list-style-type: none"> <li>Develop temporal and cultural overlays (e.g., "show evolution of the concept of rhythm across languages").</li> </ul>	Code ▾	Could-Have ▾
	<ul style="list-style-type: none"> <li>Compute graph metrics (centrality, clustering, modularity) to identify conceptual clusters and cross-cultural bridges.</li> </ul>	Code ▾	Could-Have ▾

## PROJECT PHASES WITH MILESTONE OUTCOMES

Phase	Key Milestones	# of Iterations
Project Initiation	<ul style="list-style-type: none"> <li>Document project milestones in GitHub</li> <li>Write GitHub issues for first phase</li> </ul>	1
Machine Translation Assessment	<p><b>RQ: Is existing automated multilingual translation good enough for this domain-specific content, or is too much nuance lost?</b></p> <ul style="list-style-type: none"> <li>One round on Notion data</li> <li>One round on Music Theory Online full-text articles</li> </ul> <p>Each round will include data preparation, choice of model, and quantitative/qualitative assessment.</p> <p>Draft relevant sections of the AI translation paper.</p>	3
Evaluating Methods for	<p><b>RQ: Which methods for clustering terms across languages</b></p>	4

Multilingual Clustering	<p><b>allow us to discover meaningful related concepts?</b></p> <ul style="list-style-type: none"> <li>Includes data preparation, choice of model, generation of networks/visualization for assessment/review of output, and refinement/application of measurement schema.</li> <li>Evaluate against a monolingual topic model.</li> <li>Write-up relevant methods section of methodological comparison paper.</li> </ul> <p>Draft relevant sections of the methods paper.</p>	
Project Wrap up	<ul style="list-style-type: none"> <li>Draft remaining sections of methodological comparison and AI translation papers</li> <li>Re-run analysis where needed</li> <li>Bundle selected visualizations into a GitHub pages site.</li> </ul>	2
Administrative Closeout	<ul style="list-style-type: none"> <li>Handoff and offboarding</li> </ul>	1

## DEPENDENCIES & RISKS

DEPENDENCIES	<ul style="list-style-type: none"> <li>A Google Doc provided by the Research Team listing 5-10 known negative examples (concepts that do not translate or are discussed differently in different traditions) and 5-10 known positive examples (concepts that we know align) across all target languages in variable combinations.</li> <li>Modifications to existing <i>Music Theory Online</i> web scraping code to include tables.</li> <li>Side-by-side scholar-translated English translations of the articles that are not published that way.</li> </ul>
POTENTIAL RISKS	<ul style="list-style-type: none"> <li>Terms/texts may cluster by language instead of semantically across language; to mitigate, identify and normalize/suppress the language-specific dimensions of the vector representation</li> <li>Due to the small number of target language articles in the existing dataset, the methods risk surfacing idiosyncrasies about the articles themselves rather than the larger music theoretical language traditions. Should this become a concern, to mitigate the Music Theory in the Plural project team will gather more examples of music theory journals/articles written in Chinese, Japanese, Portuguese or Spanish, preferably with corresponding English translations.</li> </ul>
RIGHTS + PERMISSIONS	<ul style="list-style-type: none"> <li>Permission has been obtained from the editor of <i>Music Theory Online</i> for web scraping. The Music Theory in the Plural team will secure permissions for any additional resources prior to web scraping.</li> </ul>

## PROJECT TEAM

Name	Role on Project	Dept	Role Description
Anna Yu Wang	Co-PI	MUS	<ul style="list-style-type: none"> <li>• <b>Responsible</b> for overall project direction and for providing regular feedback/testing to the RSE team during meetings and asynchronously.</li> </ul>
Jürgen Hackl	Co-PI	CEE ▾	<ul style="list-style-type: none"> <li>• <b>Responsible</b> for overall project direction and for providing regular feedback/testing to the RSE team during meetings and asynchronously.</li> </ul>
Chris Stover	External Collaborator	MUS ▾	<ul style="list-style-type: none"> <li>• <b>Responsible</b> for providing input for the overall project direction and for providing regular feedback/testing to the RSE team asynchronously.</li> </ul>
Edwin Li	External Collaborator	MUS ▾	<ul style="list-style-type: none"> <li>• <b>Responsible</b> for providing input for the overall project direction and for providing regular feedback/testing to the RSE team asynchronously.</li> </ul>
Annie Liu	Project Manager	MUS ▾	<ul style="list-style-type: none"> <li>• <b>Responsible</b> for: <ul style="list-style-type: none"> <li>○ Creating and circulating meeting agendas at least 24 hours before meetings</li> <li>○ Taking notes at meetings</li> <li>○ Maintaining project documentation and organizing Google Drive folder</li> <li>○ Updating relevant GitHub issues with decisions after meetings</li> <li>○ Managing task-tracking in Asana</li> <li>○ Assisting with data work</li> <li>○ Assisting with testing</li> </ul> </li> </ul>
Ian Peiris	Undergraduate Research Assistant	MUS ▾	<ul style="list-style-type: none"> <li>• <b>Responsible</b> for: <ul style="list-style-type: none"> <li>○ Assisting with testing</li> <li>○ Assisting with data work</li> </ul> </li> </ul>
Laure Thomps...	Co-PI, Technical Lead	CDH ▾	<ul style="list-style-type: none"> <li>• <b>Accountable</b> and <b>responsible</b> for the technical direction, design, and execution of the project.</li> </ul>

			<ul style="list-style-type: none"> <li>• <b>Responsible</b> for: <ul style="list-style-type: none"> <li>○ Identifying milestones</li> <li>○ Writing issues and acceptance criteria</li> <li>○ Developing software</li> <li>○ Moving and maintaining assigned issues on the Zenhub board</li> </ul> </li> </ul>
Hao Tan	Research Software Engineer	CDH ▾	<ul style="list-style-type: none"> <li>• <b>Responsible</b> for the design and execution of the project.</li> <li>• <b>Responsible</b> for: <ul style="list-style-type: none"> <li>○ Writing issues and acceptance criteria</li> <li>○ Developing software</li> <li>○ Moving and maintaining assigned issues on the Zenhub board</li> </ul> </li> </ul>
Rebecca Koeser	Lead Research Software Engineer, Advisor	CDH ▾	<ul style="list-style-type: none"> <li>• <b>Responsible</b> for advising on the technical design and implementation of the project.</li> </ul>
Jeri Wieringa	CDH Project Manager; DH Researcher	CDH ▾	<ul style="list-style-type: none"> <li>• <b>Responsible</b> for: <ul style="list-style-type: none"> <li>○ Facilitating project team and internal RSE team Scrum meetings</li> <li>○ Helping to define, scope, and prioritize the development work according to Agile practices</li> </ul> </li> <li>• <b>Accountable</b> and <b>responsible</b> for ensuring alignment of experiments and research goals, and leading the writing of the paper outputs.</li> </ul>

## BUDGET

There are no expenses directly associated with this Research Partnership.

## TERMS OF COLLABORATION

### DOCUMENTATION + COMMUNICATION

**Slack** will be used for most day-to-day communication among team members. Meeting notes, data, and documentation will be stored in a shared folder within the CDH Research Partnership shared **Google Drive**. **Asana** will be used for non-development-related task management; all team members are expected to check off tasks promptly to asynchronously communicate on progress. A public **GitHub** repository under the Princeton-CDH organization will be used for managing

development-related tasks (issues), code, and code documentation. **ZenHub** will be used for issue organization and reporting. For the duration of the research partnership, the project team is expected to meet at least once every other week to discuss progress and make decisions. Meeting frequency may need to be adjusted once active development ramps up to meet team needs. In the case of a breakdown in communication, the CDH Assistant Director may intervene to help find a mutually agreeable solution or, in very rare cases, negotiate the end of the partnership.

#### CREDIT + ATTRIBUTION

The project team recognizes the collaborative nature of the digital humanities and, in contrast to the single-authorship model of the traditional humanities, embraces an inclusive practice of authorship. Following the recommendation of Princeton Research Computing's [RSE Partnership Guide](#), if a publication would not have been possible without RSE project team members, then the RSEs should be included as co-authors. In the case that a single-authored work that builds on this collaboration is necessary for reasons of promotion and/or tenure, the CDH should receive clear and prominent acknowledgement as appropriate within the publication.

All student labor related to the partnership, unless part of a student's coursework, must be fairly compensated, and we encourage at minimum citation, or co-authorship if appropriate, of students and data workers in all related publications, including datasets, articles, and monographs.

The project team commits to informing the CDH of publications resulting from this collaboration within 3 months of publication. In addition, project members may be asked to be included as secondary authors on additional publications based on the collaboration produced by the CDH team.

#### PUBLICITY

The CDH Communications Manager may publicize the project and related information (such as images, links, and text) on the CDH website and newsletter, in annual reports, and on select social media (e.g. Bluesky, LinkedIn). In some cases, the Communications Manager may also work with Princeton's Office of Communications to publicize the project on Princeton media sources such as its website, social media, newsletters, magazines. To produce effective and quality publicity of the project, the Communications Manager may request additional media and information from the PI at any point before, during, or after the partnership.

#### LICENSES

Code will be made publicly available under an open license (Apache 2.0).

#### AI TOOLS

The CDH's policies on the use of AI tools for project management and software development are available here: [CDH AI Policies](#) (version 1.0)

## EXTENSIONS

The PI may request an extension beyond the allotted partnership timeframe to complete the agreed-upon deliverables by submitting a written rationale to the CDH Assistant Director, cc'ing the CDH Project Manager. Extensions may also be proposed and automatically granted by the CDH Assistant Director due to delays in the CDH development schedule.

## AMENDMENTS

Amendments to the charter should be requested in writing to all project team members and the CDH Assistant Director *and must be approved by all parties*. Changes will be recorded in the "Version History" section below. Please also update the date in the charter header.

## SIGNATORIES

PREPARED BY	Mary Naydan	Nov 21, 2025
REVIEWED BY	Anna Yu Wang Jürgen Hackl Annie Liu Jeri Wieringa Rebecca Koeser Hao Tan	Dec 3, 2025
APPROVED BY	Meredith Martin Paul Vierthaler	Dec 19, 2025

## VERSION HISTORY

UPDATED BY	Jeri Wieringa	Feb 20, 2026	<ul style="list-style-type: none"><li>Added external collaborators and undergraduate research assistant.</li><li>Added Portuguese as fourth language in the study</li><li>Adjusted CDH roles to reflect shift in project lead from Rebecca to Laure and PM role from Mary to Jeri</li><li>Adjusted language around "AI" to be more specific to Large Language Models</li></ul>
------------	---------------	--------------	--