Proposal Application

Name of the applicant	Seyed Hamid Hassani, PhD
Institution	Elahé Omidyar Mir-Djalali Institute of Iranian Studies at the University of Toronto
Country	Canada

Tentative Title	“A Graded Persian-English Dictionary; Including a 5,000-word/morpheme List”
Topic	Development of a pedagogical bilingual (Persian-English) fully-tagged 5,000-word/morpheme list of today’s Persian; based on a 90-million-token corpus
Estimated duration	24 months

Abstract of proposal:

⬘ Project Summary

This postdoctoral project proposes the development of a bilingual (Persian-English) graded dictionary of today’s Persian, aimed at supporting language teaching, learning, research in Persian corpus linguistics, and broader humanistic fields. The project builds upon the applicant’s PhD dissertation titled “A Lexical-Grammatical Analysis of 5,000 Frequent Words and Morphemes of Today’s Persian” (2024), which was based on a 90-million-token, personally compiled corpus representing about 200 diverse academic and non-academic domains.

Unlike some other Persian similar resources—many of which suffer from subjective bias and some methodological inconsistencies—this project draws on a nearly balanced, clean, and genre-diverse dataset, free of thematic skew and ideal for producing generalizable findings. It will offer a reliable pedagogical tool for learners of Persian as a second/foreign language, and a research-grade resource for interdisciplinary scholarship.

⬘ Objectives & Methodology

The project aims to produce a modern, scientifically valid, and pedagogically useful bilingual, fully-tagged wordlist, featuring the following core components:

- Frequency levels (e.g., CEFR: A1, A2, B1, etc.; or: 1, 2, 3, etc.);
- Morphological classification (simple, derivative, compound, and compound-derivative) (already completed in the PhD dissertation);
- POS and LUs (lexical units) tagging (done);
- Phonological transcription of each entry (in IPA);
- Professionally recorded pronunciation of each entry and its example(s);
- Frequent English equivalent(s);
- Usage registers (e.g., fiction, non-fiction, periodicals {newspapers or magazines}, academic, spoken);
- Authentic example sentences drawn from natural context;
- Relevant morphological and lexical graphs (partially done).

While the dictionary will be compiled manually through expert linguistic judgment, the effort builds directly on the applicant’s PhD dissertation, itself based on a 90-million-token corpus assembled over nearly three decades using tools such as “MS-DOS Zarnegar” (a Persian-script word processor), MS Word, and some online open-source platforms including Wikimedia Foundation (esp. Wikipedia and Wiktionary). Although the entire corpus files are not collected in one place and in a single tool or retrievable, due to the relatively long timespan and envolving storage formats, the extracted frequent 5,000 lexical items (words/morphemes)—

organized and analyzed with expert judgment—remains intact and will serve as the reliable empirical foundation of this project. The dataset is not automatically frequency-tagged or annotated; rather, the current project will involve manual tagging and systematic expansion to create a reliable and pedagogically appropriate graded bilingual dictionary.

While the project draws upon the lexical data generated during the applicant's PhD research, it proposes a significantly novel, bilingual, and technology-enhanced approach aimed at pedagogical utility.

❖ Impact & Relevance

The final product—a fully-tagged, graded bilingual Persian-English dictionary—will serve as a high-impact resource for Persian language educators, learners (migrants, second-generation heritage speakers, and students in language education programs), lexicographers, corpus linguists, and developers of Persian NLP tools. Its pedagogical and technological applications span second-language instruction, proficiency testing, curriculum design, and automated assessment.

The dictionary can be disseminated through multiple channels: it may be published as a printed academic reference book and/or released as an open-access digital dataset (e.g., MS Word or PDF formats) accessible online. This range of dissemination formats will maximize accessibility and long-term sustainability. Although the data will be primarily designed for pedagogical and scholarly use, its structured format will make it adoptable for future NLP applications by interested developers or researchers. The fellowship will also support the internationalization of a decades-long research effort, expanding its impact across linguistic, educational, and computational domains. The project has strong potential to initiate international collaborations in Persian lexicography, corpus development, and digital humanities.

Collaboration with the Elahe Omidyar Mir-Djalali Institute of Iranian Studies —a leading institution in promoting Persian language pedagogy and Iranian linguistics— will be essential in ensuring both the scientific robustness and pedagogical relevance of this project. The institute's strong academic environment and its commitment to Persian language education provide an ideal setting for refining the project's methodology, facilitating access to relevant academic and linguistic resources, and supporting broader dissemination within academic and educational communities.

Recommendation letters can be arranged.