Request For Proposal

[LONG]

Mapping and collection of scientific bilingual corpora

OPERAS AISBL

PROPOSALS DUE BY: 7 October 2022
Company Background

OPERAS is the Research Infrastructure supporting open scholarly communication in the social sciences and humanities (SSH) in the European Research Area. Its mission is to coordinate and federate resources in Europe to efficiently address the scholarly communication needs of European researchers in the field of SSH.
Project Overview

In 2020, the French Ministry of Higher Education and Research (MESR) launched the Translations and Open Science project with the aim to explore the opportunities offered by translation technologies to foster multilingualism in scholarly communication and thus help to remove language barriers according to Open Science principles.

During the initial phase of the project (2020), a first working group, made up of experts in natural language processing and translation, published a report suggesting recommendations and avenues for experimentation with a view to establishing a scientific translation service combining relevant technologies, resources and human skills.

Once developed, the scientific translation service is intended to:

- address the needs of different users, including researchers (authors and readers), readers outside the academic community, publishers of scientific texts, dissemination platforms or open archives;
- combine specialised language technologies and human skills, in particular adapted machine translation engines and in-domain language resources to support the translation process;
- be founded on the principles of open science, hence based on open-source software as well as shareable resources, and used to produce open access translations.
Project Goals

In order to follow up on recommendations and lay the foundation of the translation service, the OPERAS Research Infrastructure was commissioned by the MESR to coordinate a series of preparatory studies in the following areas:

1. **Mapping and collection of corpora**: identifying and defining the conditions for collecting and preparing corpora of bilingual scientific texts which will serve as training dataset for specialised translation engines, source data for terminology extraction, and translation memory creation.

2. **Use cases**: drafting an overview of the current translation practices in scholarly communication and defining the use cases of a technology-based scientific translation service (associated features, expected quality, editorial and technical workflows, and involved human experts).

3. **Translation output evaluation**: evaluating the output of a set of translation engines with specialised texts.

4. **Roadmap and budget projections**: making budget projections to anticipate the costs to develop and run the service.

The four preparatory studies are planned during a one-year period as of September 2022.

The present call for tenders only covers the (1) **Mapping and collection of corpora**.

Two additional calls will be released in the coming months for the following studies: (3) **Translation output evaluation** and (4) **Roadmap and budget projections**.

The (2) **Use cases** call is open from 1 September to 7 October 2022 (details available [here](#)).
Scope of Work

In order to train and evaluate specialised machine translation engines, extract relevant terminology and create in-domain translation memories, it is necessary to build up bilingual corpora of disciplinary texts extracted from scholarly publications such as journal papers, monographs, scientific blog posts and associated metadata. Some potential sources of bilingual scientific publications have already been identified: publishers, dissemination platforms, scholarly and academic networks, etc. However, due to significant differences in formats and content licensing, the collection and preparation of ready-to-use bilingual corpora require challenging operations to make them interoperable, open and eventually available to a wide audience (translators, researchers, students, etc).

Therefore, OPERAS welcomes proposals from public and private entities to identify, collect and prepare corpora of bilingual scientific texts (minimum 100,000 parallel segments for each discipline, subject to corpus availability).

Service providers are expected to propose three relevant disciplines for corpus collection, one for each of the following disciplinary domains*:

1. Life Sciences
2. Physical Sciences and Engineering
3. Social Sciences and Humanities

* Based on the ERC panel structure available here.

Disciplines for corpus collection should be suggested taking into account the following criteria: volume of disciplinary publications in France and internationally; proportion of open-access publications within the discipline; available corpora and language resources - terminology in particular; recognised translation needs within the disciplinary communities; suitability of the disciplinary language and writing style for human, semi-automatic and automatic translation; interest and accessibility of the disciplinary content for scholarly and general readers. Preference will be given to proposals covering disciplines that are relevant to contemporary social challenges.
Data collection will include full texts of scholarly publications as well as their metadata - in particular titles, abstracts and keywords. Corpora will be selected based on estimated quality, technical and licensing requirements (assistance will be provided by the project legal advisor). A list of potential resources already identified will be shared with the selected service provider.

Preference will be given to proposals including bilingual terminology extraction (minimum 200 terms for each discipline).
Target Deliverables and Schedule

1. **Comprehensive report of the scientific bilingual corpora identified for the selected disciplines and languages including information on publication types, formats, licensing, translation process and quality estimation, collection and processing requirements. The report should mention all the relevant corpora identified, including those that will not be collected for technical or legal reasons.**

2. **In-domain aligned corpora (as per the service description made by the provider in the tender response).** Segment metadata should at least include the following information: discipline, source document title and publication type, translation quality estimation, and French language variety (if available).

3. **In-domain termbases extracted from corpora (as per the service description made by the provider in the tender response).** Term metadata should at least include the following information: discipline, source document title and publication type, translation quality estimation, and French language variety (if available).

**Final Project Due:** 28 February 2023

- Bid period: 1 September to 7 October 2022
- Result notification: 18 October 2022 EOD
- Service starting date: 1 November 2022
- Expected turnaround time: 4 months
Specifications

- **Language pair**: English-French

- **Disciplines**: three disciplines as per the service description made by the provider in the tender response, one for each of the above mentioned disciplinary domains

- **Expected formats for corpora**: tmx, plain txt, and tsv

- **Expected formats for terminology extraction (if applicable)**: tbx and tsv
Existing Roadblocks Or Technical Issues

- Undetermined availability of disciplinary bilingual resources
- Varying levels of translation quality
- Copyright and licensing requirements
- Different formats across publishers, dissemination platforms and archives
- Strict time frame calculated to comply with the planning of the four preparatory studies
Budget Constraints

Budget range: €90,000-€110,000
Evaluation Metrics

OPERAS will evaluate bidders and proposals based on the following criteria:

- Experience in corpus collection, processing and alignment
- Experience in scholarly publishing corpus collection, processing and alignment
- Experience in terminology extraction
- Achievability of deliverables
- Adequacy of requested resources and expected results
Questions Bidders Must Answer To Be Considered

Service providers are asked to submit a service proposal describing the tasks that they will be able to perform in relation to the present call for tenders during a four-month period starting from 1 November 2022.

In particular, service providers are asked to include in their response the following information:

- Disciplines selected for corpus collection
- Rationale for discipline selection based on the above mentioned criteria
- Detailed description of deliverable formats, in particular their structure and the metadata included
- Provisional planning of the service tasks (Gantt chart required)
- Detailed list of work packages or tasks
- Detailed budget with distribution of effort measured in PMs across tasks
Submission Requirements

Bidders must adhere to the following guidelines to be considered:

- Only bidders who meet all 5 metrics in the evaluation section should submit a proposal.
- Proposals must be sent in by 7 October 2022. Bidders who are interested in submitting a proposal should inform Susanna Fiorini (susanna.fiorini@operas-eu.org) no later than 30 September 2022.
- Include samples and references with your proposal.
- Proposals should not be more than 4 pages. Failure to comply with this guideline will result in an automatic rejection.
- A proposed schedule must also be included and clearly expressed.
What We’re Looking For in Potential Vendors

The call is open to public and private vendors, regardless of their country of establishment.

We are particularly interested in receiving proposals from open source- and open data-friendly organisations.

We attach great value to sustainable and ethical business models.

Vendors should be able to ensure smooth communication with the steering committee throughout the entire duration of the project.
Contact Information

For questions or concerns connected to this RFP, we can be reached at:

Susanna Fiorini
susanna.fiorini@operas-eu.org