

Market Consultation

**Automatic Dialect Recognition for Arabic in the
context of project CELIA**

Date: July 9, 2025

Introduction

This market consultation has been prepared by IUC DJI in anticipation of a potential (European) tender for the development and validation of a state-of-the art automatic dialect recognition (ADR) prototype to distinguish between all Arabic spoken dialects. The prototype should be suited as back-end speech analysis system for licensed software which is to be developed later if results are considered usable. Analysis results must be usable and interpretable in a forensic setting.

Contracting Authority

The contracting authority is the State of the Netherlands – specifically the **Immigration and Naturalisation Service (IND)**, part of the Ministry of Asylum and Migration

Goals

The goals of this market consultation are to:

- Gain insight into available experience and market offerings;
- Understand the visions, suggestions, ideas, and developments from market parties;
- Involve and interest market parties early in the potential tender;
- Determine the most suitable procurement strategy;
- Test internal assumptions, ideas, and proposed solutions.

Any market party that believes it can contribute is welcome to participate. During the period from July 9, 2025 to August 13, 2025 DJI would like to benefit from your knowledge and experience. The questions for market participants are listed in Chapter 3.

Document Structure

Chapter 1: Purpose and scope

Chapter 2: Information and submission instructions for the consultation procedure

Chapter 3: Consultation questions

Appendix A: Principles and conditions of the consultation

1. Context

Project CELIA

The development of the ADR system is part of the Common European Language Indication and Analysis (CELIA) project.

At present asylum authorities spend considerable time and effort investigating the identity and country of origin of asylum seekers, based on their declarations. Traditional (human) language analysis is already used to clarify the spoken language, and hence the country of origin. However, the process is cost intensive and time consuming, sometimes requiring weeks or even months. The development of automatic language assessment tools and their widespread use at European scale could substantially reduce the authorities' efforts and shorten the time needed for the examination of asylum applications. Developing a European tool accessible for all asylum authorities can significantly raise the quality of the identification process and contribute to faster procedures.

Intended application of the ADR system

If the claimed background and origin of the applicant is considered suitable for analysis by the AI system, a 'Level 1' automatic analysis will be done. If the result confirms the claimed origin with sufficient certainty, the output can be reported as is. If the result does not confirm the claimed origin or is unclear, the recording goes to a certified human analyst for a 'Level 2' indication, as a human check. When necessary, a certified human analyst can also perform a full linguistic analysis and expert report ('level 3'). This application falls under 'high risk' and will thus have to meet the criteria for high risk AI applications according to the EU AI Act.

Expected procurement process

Following this market consultation we will start an initial procurement for a research and development question, in order to gain insight in the performance of an ADR prototype under realistic forensic conditions. Depending on the results, a second procurement is expected for the development of licensed software for forensic application of an ADR system, while meeting the criteria for high-risk AI systems in migration, asylum and border control (Articles 9 to 15 of the EU AI Act) and the criteria for forensic evidence.

Expected timeframe for the research and development question:

- **4 months for preparation**, i.e. to jointly
 - o Discuss which speech data are allowed to be used (in line with GDPR and other regulations)
 - Establish frameworks for a possible end product (licensed software)
 - Establish which parts of the system are/will be owned by whom, what will be open source and what will not
 - Discuss how a reasonable amount for a license of a possible end solution can be determined, how vendor lock-in can be prevented, etc.
- **10 months for development and validation of ADR prototype**
- **4 months for further testing with IND dataset** (around 300 anonymized use case recordings) and finetuning

Expectations for the ADR prototype

1. Fully automatic, presumably using x-vector or Deep Neural Network (DNN) technique.
2. Should be able to deal with .mp3 audio samples (192 kbps, >30 seconds, usually >60 seconds) of spontaneous speech in spoken Arabic varieties (not Modern Standard Arabic).
3. Should be able to handle any spoken dialect of Arabic to the extent possible, since the true origin of the speakers to be tested is unknown.
4. Should be able to indicate the best matching language model at country level, or when this is not possible with enough reliability, the result may indicate a group of countries/regions.
5. Should communicate the level of uncertainty of the result to the user.

Extra information

Please note that within the CELIA project, an exploratory report has been prepared by the Netherlands Forensic Institute, addressing the feasibility of automatic Arabic dialect identification under the assumption that it must meet the criteria for high risk artificial intelligence systems as set out in the EU's Artificial Intelligence Act (AI Act). To this end, a summary of the relevant criteria given by the AI Act and details of available Arabic datasets and dialect recognition techniques are presented. The report also includes background information on Arabic language varieties. This report is available (for the purposes of this project only) on request; requests can be sent to CELIA@ind.nl.

2. Consultation procedure

By participating, you unconditionally agree to the principles and conditions in Appendix A.

2.1 Timeline

The IUC DJI follows this indicative schedule:

Activity	Deadline
Publication of market consultation	July 9, 2025
Deadline for submitting questions	July 30, 2025
Publication of Q&A (information note)	August 6, 2025
Deadline for submitting responses	August 13, 2025

This schedule is subject to change.

2.2 Questions

If the document raises questions or uncertainties, you may submit them via the TenderNed messaging module by the specified date. Answers will be published anonymously on www.tenderned.nl.

For more information about TenderNed please see [Dutch government's online tendering system | TenderNed](#).

2.3 Submitting responses

Participants are asked to submit answers to the questions in Chapter 3 via TenderNed. IUC DJI may invite participants for further explanation via plenary sessions or individual meetings. There is no obligation to invite all participants or provide reasons for selection.

Responses may influence the nature and scope of the future tender.

2.5 Conclusion of the consultation

IUC DJI will publish a report on www.tenderned.nl summarizing key conclusions. Participants agree that their input may be used (anonymously) in this report, which will list participating parties.

3. Questions

Market parties are asked to answer the following questions in writing, with motivation and concise responses:

Cost and feasibility questions

1. Based on the information about this assignment, what do you think are the main cost items that should be taken into account?
2. How much capacity do you think it will take to develop the ADR prototype as defined above? What level of education and/or type of experience should the developer(s) have?
3. In your opinion, is the timeframe suggested above reasonable? If not, what is in your opinion a reasonable time frame?
4. What would in your opinion be a rough estimate for the cost and investment for delivering a first prototype as described above?

Technical questions

5. What kind of ADR technology, language/data models and/or platforms would you be likely to use in the development of a prototype?
6. What risks (if any) do you foresee with respect to the use of datasets?
7. The government follows the policy of 'Open, unless': government software should be open source as much as possible, except when there are valid reasons not to make the software available. Which parts of the prototype would be based on open source and which parts would not be open source?
8. What developments do you see in the field of automatic dialect recognition methods that are relevant to this assignment?
9. To what extent would you be able to cope with the requirements for high risk AI applications as prescribed in the EU AI Act? Can you identify challenges and possible solutions?

General questions

10. What is your view on the use of a further developed dialect recognition system in migration procedures of migration authorities of the EU+ countries? What are the main points that need to be taken into account (technical, licensing, moral, legal, etc.)?
11. Do you see any opportunity to contribute to Social Return/Sustainability?
12. What risks (financial, organizational, content) do you see with regard to this assignment and what could mitigating measures entail?
13. Do you have any other comments and/or suggestions that you would like to make for this tender (including but not limited to the price-quality ratio, the sub-award criteria, etc.)?