



Dienst Justitiële Inrichtingen
Ministerie van Justitie en Veiligheid

Answers to questions

Market consultation

Automatic Dialect Recognition for Arabic in the context of project CELIA

Kenmerk : TN 536558
Date : 6 august 2025
Versienummer : 1.0

Answers to questions

NR.	SUBJECT	QUESTION	ANSWER
1.	System	The consultation document speaks of the desired system in terms of both dialect identification (P3 "indicate the best matching language model at country level") and dialect verification (P3 "If the result confirms the claimed origin with sufficient certainty") – is the system intended to be used in both of these scenarios?	We can indeed imagine that both of these scenarios will apply. In principle, the initial question for immigration services is whether the claimed origin is confirmed by the language sample or not (dialect verification). However, this does not mean that the ADR system needs to tackle the analysis in that way; if the system gives as result the best matching language model, this can be compared to the claimed origin by the receiving party. In cases where the claimed origin is not confirmed, a secondary question relevant for immigration services is what is the most likely true origin. So this is essentially dialect identification. We are open to suggestions. See also the answer to Q3.
2.	System	Should the prototype tool include consideration of human factors (user-interface, explainability, etc) or will this be considered in the next phase of development?	This will be considered in the next phase (not during the initial procurement for a research and development question). This is what we meant to indicate when we stated in the introduction that the prototype should be suited as “back-end” for a licensed piece of software – with the front-end to be developed later.
3.	System	Related to the human-factors, are there recommendations for how results should be presented: a) A language/country/region + probability? b) A likelihood ratio based on a claimed origin?	Please note that generally, users within the asylum process are unlikely to be familiar with the Likelihood Ratio or Bayesian framework. However, for forensic purposes the result should give some indication of the strength of the evidence, so not just an answer to the question how well the findings fit the best matching dialect model. If the ADR system ends up being deployed, and assuming for now that the ADR will indicate which dialect model matches best with the questioned sample, we could imagine something like the following: (a) a best match plus an indication of the ‘score’ (level of match) generated by the ADR system, followed by (b) how likely such a result is when this is the true origin of the applicant (% correct hits), as well as (c) how likely it is that this result is false, leaving room to comment on known frequent confusions (e.g. Best match X ‘Note that this result has also been encountered for a portion of speakers originating from Y’.

			Alternatively, we could imagine that comparing the scores of the best two-three matches could be informative when formulating a verbal indication of the strength of the evidence. Again, we are open to suggestions. There may be various options to deal with the issues raised in Q1 and Q3, which may need empirical investigation. At this point, if we were to require a likelihood ratio based on a claimed origin, we anticipate practical issues with the alternative hypothesis. We note that within the set of 'alternative origins', some will inherently be more similar to the claimed variety than others. We expect that pooling data too early in the process means a loss of relevant information.
4.	System	If there are a variety of algorithms that we can use within the system, is a unified output preferred, or can we provide individual outputs for the sub-systems (with an associated risk that they may not always be in agreement with each other)?	In the R&D phase it will likely be helpful for you and us to be able to examine individual outputs for any sub-systems. However, for the final product we believe a unified output will be required, as we can't expect the receiving party to make sense of a variety of sub-results coming from systems on which they have no expertise. At this point (subject to further discussion in the 'four months for preparation' period), we envision that the prototype will be evaluated in the R&D phase on the basis of standard metrics (such as EER, Cllr) and visualisations (ROC curves, confusion matrices, scatterplots to show the correlation between score and reliability). We suggest that a unified output is required to allow comparison across prototypes.
5.	System	How hard is the input criterion (30s to 60s duration)? We would recommend recordings of at least several minutes in order to allow for a variety of algorithms to be applied.	In most cases several minutes will be available (e.g. 5-8). If a certain amount (e.g. 2-3 minutes) of net speech is minimally required, then so be it. The requirements for the recordings to be submitted for analysis in the future are yet to be determined. Thus, this input criterion is not hard.
6.	Training Data	Will data compliant with the EU AI act be made available as part of the project? (this is relevant for both the R&D and subsequent phases) e.g. any or all of the datasets listed in the report Appendix 1?	In principle, yes. Whether data is compliant with the EU AI Act, e.g. 'sufficiently representative', is hard to say (see NFI report section 6 for conclusions on this). In the '4 months for preparation' phase, we hope to decide together with suppliers which data can/will be used. We will cover the license costs for these if applicable. We are currently collecting legal advice on the question whether broadcast data and YouTube data, such as the ADI17 dataset and the GALE dataset, may be used for the final solution (cf. Advice 3 in the NFI report). Note that

			this legal question would take the GDPR and copyright regulation into account, not the quality and representativeness of the data.
7.	Training Data	Will the collection of new data form part of the project, either as part of the R&D or future tool development phases?	Only in the sense that IND recordings (which have been submitted and analyzed in the past) are being selected and anonymized for testing purposes and perhaps also training/finetuning purposes. More information on this will be supplied at a later stage. Note that these do not cover all possible varieties, but are heavily biased towards commonly claimed origins.
8.	Licensing and ownership	In response to the Open-source first recommendation, if the provided system, and associated data is completely open-source, would the Sponsor be willing to take ownership?	In principle, yes, although we are not quite sure what the implications are when saying this. We do anticipate that further development, expansion and maintenance of the software by a supplier will be required, and so we do envisage a license construction with a supplier. On the other hand, presumably the end product should not end up as a commercial product that can be sold by the supplier, while the development and data costs have been covered by government funds. Again, we hope to discuss these provisions in the '4 months for preparation' phase. We understand that this answer has a major impact on the price. When giving an estimate, please feel free to present various options or to flag certain assumptions.
9.	Licensing and ownership	Will the IND cover licensing costs of data or software components required during the R&D phase?	Yes, please include them in your answer to questions 1 and 4 if you can.
10	Compliance	Related to the Open-source question: who is responsible for compliance with EU AI act (or other relevant regulation) with regard to the use of data and subsequent application of trained models?	The EU AI Act is quite explicit on this. Given that the supplier would be the body that develops an AI system, they would be a 'provider'. As noted in the NFI report (page 10), Articles 9-12 and 15 need to be taken into account already in the development phase (R&D phase). This is regardless of the ownership. If the ADR system ends up being deployed, in our understanding there will be a certain amount of shared responsibility. Articles 11, 12 and 15 would in our view be a main responsibility of the developer/supplier.
11	Compliance	Will we have access to advice and support from the IND legal team with regard to EU AI act compliance throughout the project?	In principle, yes. However, compliance with the EU AI Act is of course new for everyone. At this point, we expect that acknowledgment of the requirements and stipulations of the EU AI Act and doing our utmost to comply with the

			requirements for a high-risk AI applications, and documenting this as well as possible, should be sufficient.
12	Other	Will the project support the costs of consultancy/hiring of a native Arabic speaker to assist us with data triage and results analysis?	Yes, if you think this would be helpful. Please include it in your answer to questions 1 and 4.
13	Other	What are the envisaged project management requirements in terms of in-person meetings, progress updates, and project dissemination (conference/meeting attendance)?	Envisaged for now are monthly online meetings with the senior expert of CELIA to discuss ongoing developments. Progress reports may be combined with the type of documentation that is required by the EU AI Act. Interim milestones and accompanying payments are yet to be defined. We further anticipate at least one in-person meeting in The Netherlands and one at the premises of the supplier. Also we hope the results will lend themselves for peer-reviewed scientific articles, to be prepared in close cooperation with the senior expert of CELIA. Note that travel and accommodation costs as well as project dissemination (writing and publishing papers, conference attendance) fall under a different part of our budget and therefore do not need to be included in your answer to the current questions.
14	Other	Will there be any training requirements (e.g. of IND staff) during the R&D phase?	No, this is not part of the R&D phase. The R&D phase is entirely focused on obtaining insight in the performance of a state-of-the-art ADR system under realistic conditions, prior to deciding on its application.