07 October 2020

Session I:
Language recognition and name transcription

Significance of a joint European approach and cooperation

High influx of asylum seekers at EU borders and within member states

Lack of identity documents amongst a large number of asylum seekers

Concerns about **fake and counterfeit passports**

Significance of a joint European approach and cooperation

Different registration of applicants‘ names in Europe results in divergent entries in databases and difficulties in finding people in databases

Different approaches to **analysing speech recordings** in different countries goes along with respective **advantages and disadvantages**

- lack of standards for language analyses/ different depths of analyses
- lack of suitable language experts / difficulties in finding suitable professionals

Cost intensity and **long duration** of language analysis

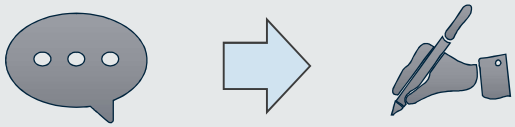
Need for cooperation for the purpose of **registration**, **identification** and **origin determination** of asylum seekers

eu2020.de

Capturing identities in Europe: One tool for a uniform transcription of names

Transcription vs. transliteration

Transcription

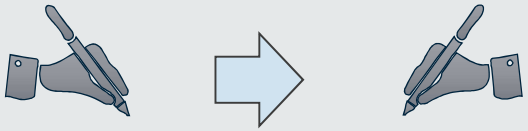


كتب [kataba]

Phonetic reproduction:
Transfer of the phonetics from one writing system to another

- Rules differ according to target language
- Phonetically correct representation of words
- The transcription process cannot be reversed 1:1

Transliteration



كتب ktb

Source script characters are represented in Latin script in full.

- Rules applied consistently
- Transliterated words are difficult to pronounce correctly
- Transliterations that represent names are difficult to work with

Common Arabic transcription standard

Although **numerous regional transliteration and transcription standards** have been developed, two of them are currently significant globally:

- DMG (Deutsche Morgenländische Gesellschaft) and related norms:
  - International Standards Organisation (ISO 233)
  - German Institute for Standardization (DIN 31635)
  - American Library Association - Library of Congress (ALA-LC)
  - The Encyclopedia of Islam, 3rd edition (EI-Three)
- IJMES (International Journal of Middle East Studies)
- As well as numerous other regionally applied standards.

If no guidelines apply, residents/officers in a specific country tend to **transcribe according to the pronunciation and spelling rules of their official language**, e.g. Mahmood (EN), Mahmud (DE), Mahmoud (FR).

We offer to use a tool-based transcription applied at German authorities, based on international Arabic transcription standards:

- mainly following the transcription rules of **EI-Three** with several simplified variations
- compatible with common software, omitting special characters, e.g. diacritics or Ayn ("ḍ", "ʿ")
- system uses **vocalization** of Modern Standard Arabic language: **a, i, u**

Transcription service (TKS)

Transcription service (TKS)

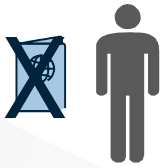
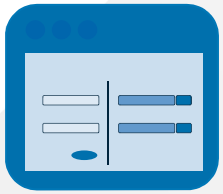
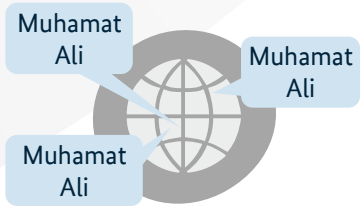
At the time of the first registration, the Latin script version of the asylum seeker's name is determined by the transcription service (TKS).

In this way a uniform/standardised transcribed version of the name is generated, which can be used **throughout Europe** and **across systems**.

This only applies to asylum seekers without identification documents (e.g. a passport), when no version of the name in Latin writing exists.

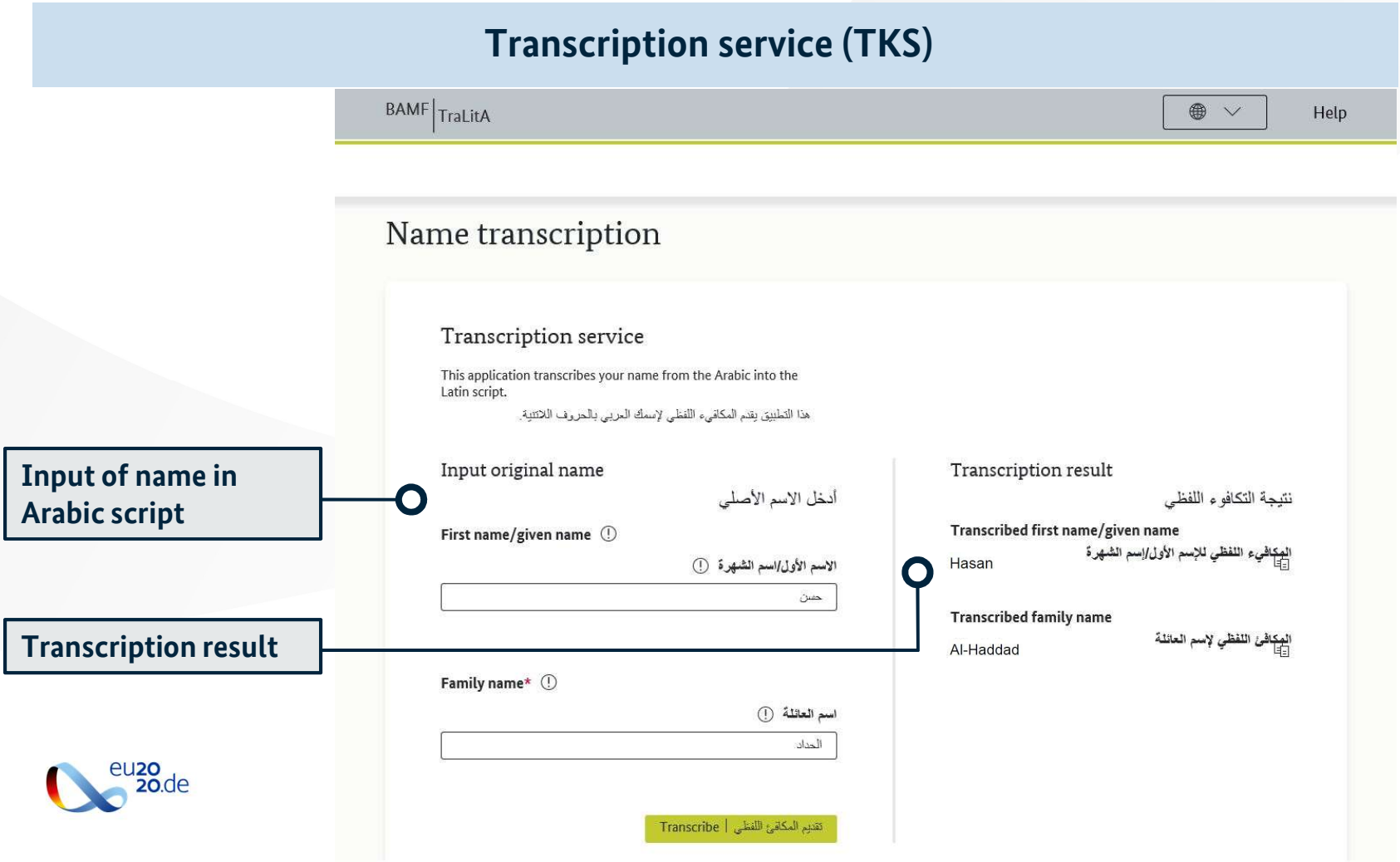
At present, the transcription tool only provides transcriptions of Arabic names. Further written languages, e.g. Persian, Russian, Georgian, can be added for transcription if necessary.

TKS process flow
Data acquisition, transmission and added value

1	2	3	4	5
				
Asylum seeker without identification documents.	Agency staff and possibly interpreter to assist with data input.	Data input and copying	Data transmission	Added value
Filing application/initial registration of core data (first and last name)	**TKS UI is opened on the intranet of agency X** **If required, asylum seeker is supported by interpreter to enter the first and last name**	**Asylum seekers without identification documents enter their first and last names via the Arabic (digital) keyboard.** **Agency staff copies the transcribed name with the "copy button"**	**Transcribed first and last names can be copied and pasted**	**Cross-country standardization of transcribed names and 'clean' data records**

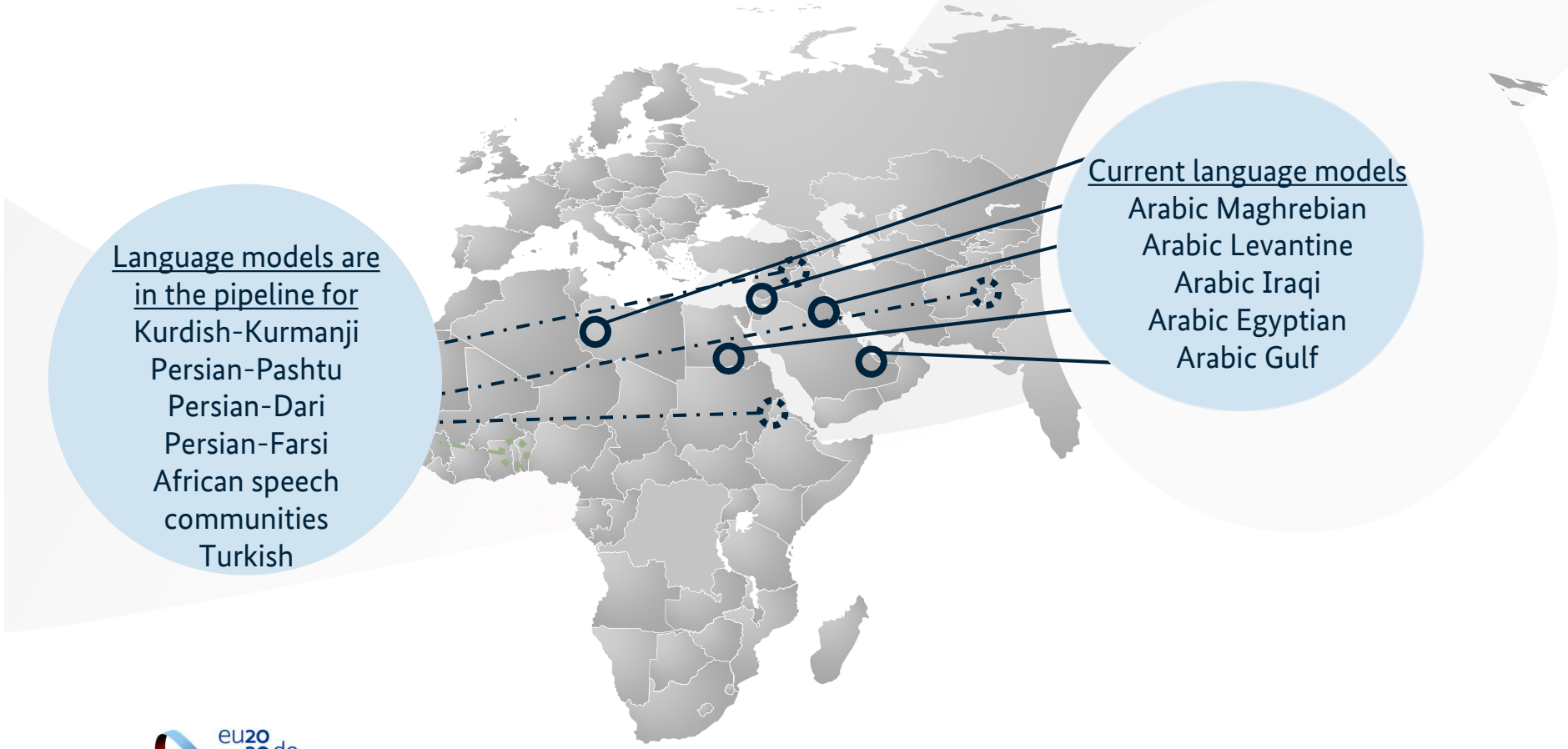
eu2020.de

TKS user interface



Language and dialect identification assistance system DIAS

DIAS language models: operative and in the pipeline



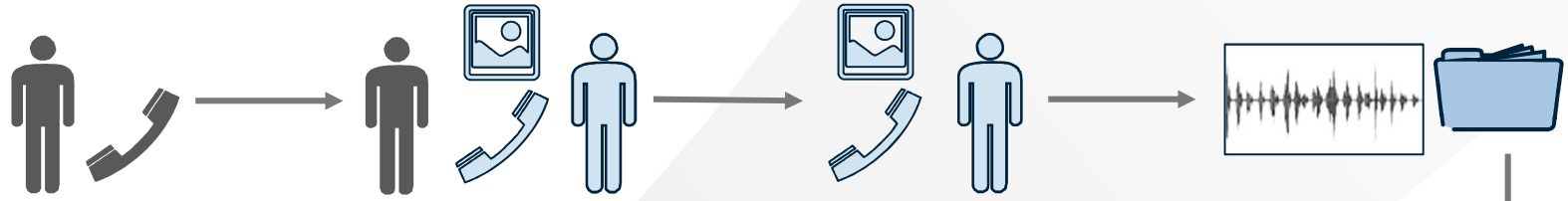
Language models are in the pipeline for
Kurdish-Kurmanji
Persian-Pashtu
Persian-Dari
Persian-Farsi
African speech communities
Turkish

Current language models
Arabic Maghrebian
Arabic Levantine
Arabic Iraqi
Arabic Egyptian
Arabic Gulf

eu2020.de

The Arabic dialect is indicated in a fast and reliable manner

Recording & storage



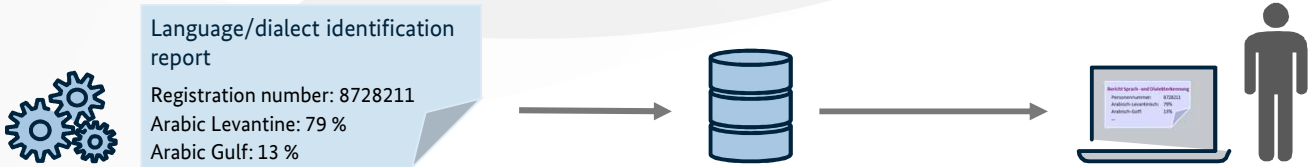
BAMF staff calls an in-house phone number and enters asylum seeker's ID data during a dialogue.

BAMF staff invites the asylum seeker to verbally describe a specific picture **over the phone** as fluently and detailed as possible.

The description is recorded electronically and then serves as the asylum seeker's country-specific language/speech sample.

Recording and ID data are saved in a central file repository.

Analysis & result



Centrally hosted language biometrics software **analyses the speech sample**: probability calculation for certain language or dialect families. The result is compiled in a report (pdf format).

The report is included in the asylum seeker's **electronic case file**.

The case worker accesses the report in the electronic file as one source of information for the overall assessment of the case records to decide the case

Flexible integration of the recording into the process

BAMF staff office

Asylum seeker

BAMF telephone network

Picture to be described

In the initial phase of the asylum procedure the asylum seeker describes the picture over the phone.

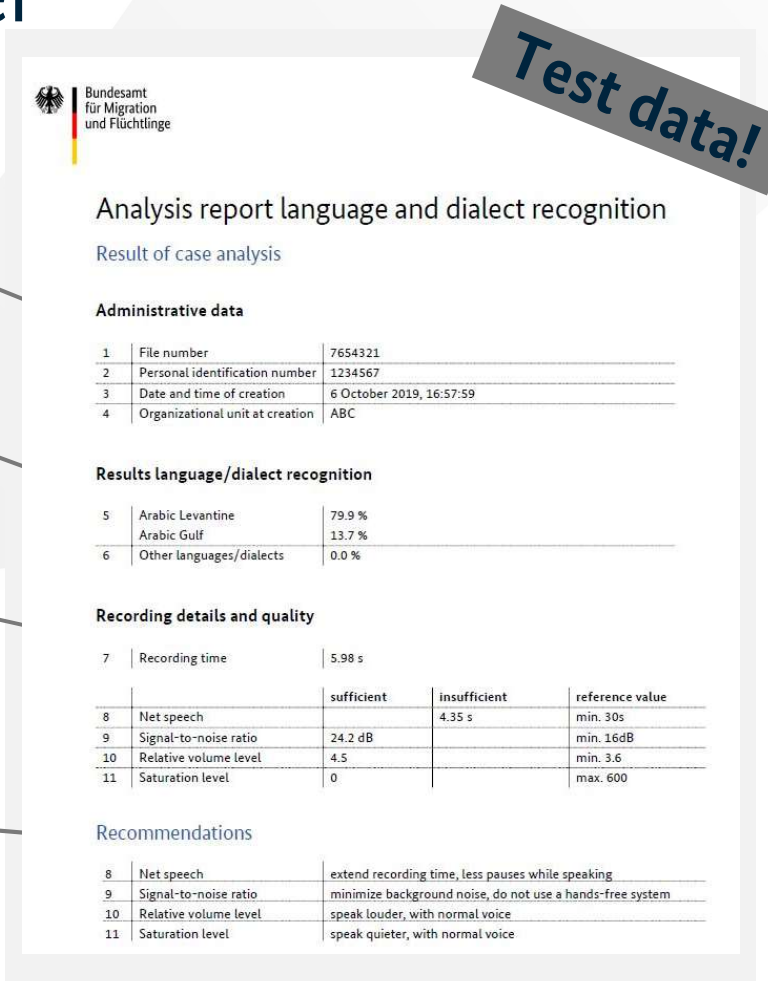
The generated report assesses the dialect as input for the case worker

Identification:
- File number, personal identification number, time, organizational unit

Result:
- Probability in %

Details & quality of recording:
- Recording time
- Net speech
- Signal-to-noise ratio etc.

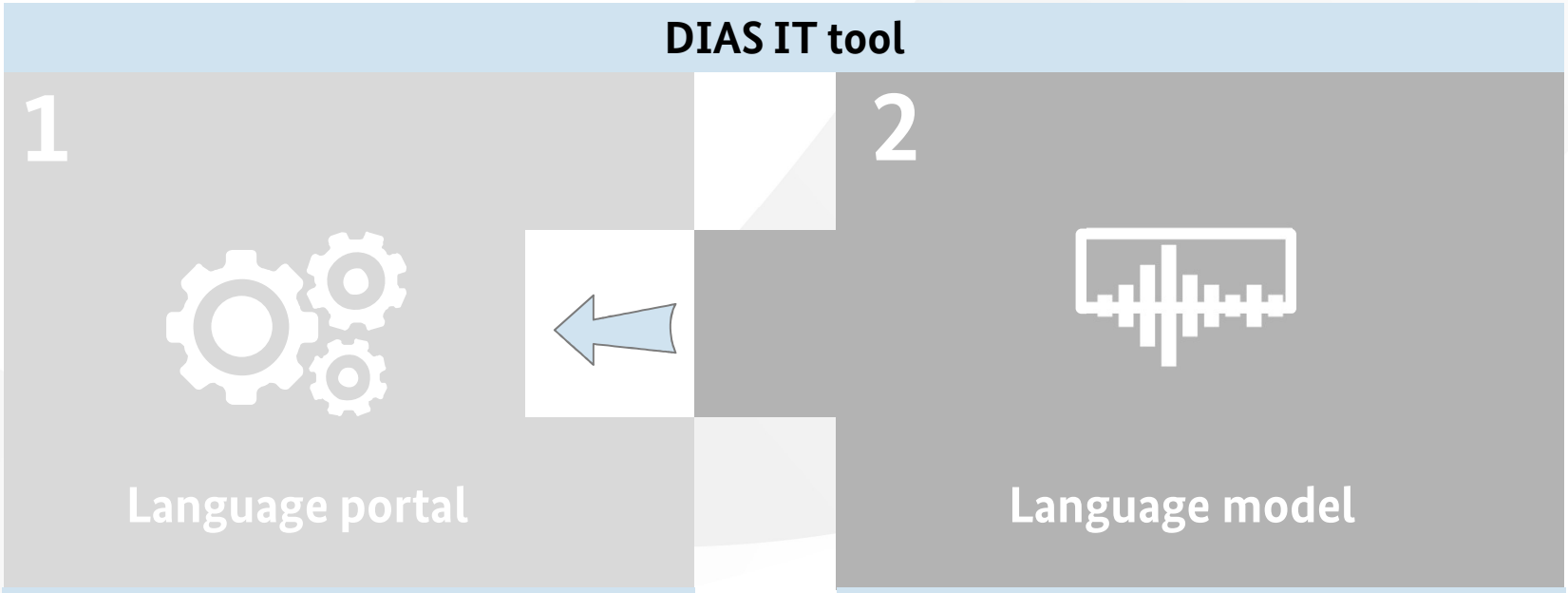
Recommendations:
- Information if recording length is ok or should be extended etc.



eu2020.de

DIAS: The technical aspects

The system consists of two key components



The language portal is based on licensed third-party software. It compares a voice recording with several reference language models.

The system's core are statistical models of the characteristics of the different languages/ dialects (language models). These language models are derived from existing background models (as reference) and from Arabic dialect models specifically trained by BAMF.

The statistical language model forms the technical basis for the DIAS tool

The language model is a collection of statistics for phonemes, sounds and acoustic characteristics for several dialects. It helps to predict an asylum seeker's spoken dialect.

Definition:

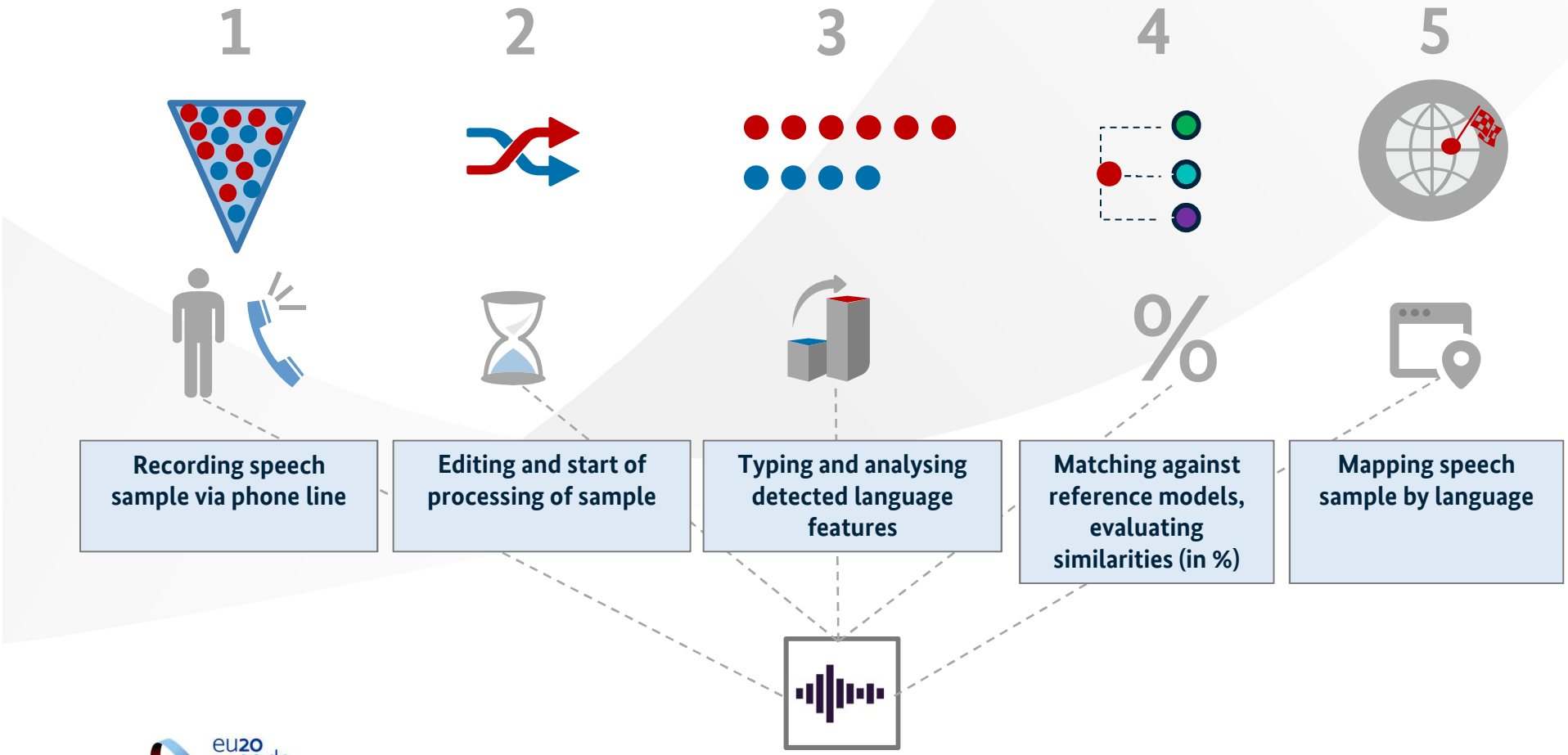
- Corpus of **language and dialect characteristics**
- Aggregate of **statistics** and **parameters**
- Abstract concept of language focussing on sequences of phonemes and acoustic characteristics ("**language as code**")

What is a language model?

Technically:

- **DIAS** examines similarities such as the frequencies of certain phonemes and their combinations.
- **Sample elements** of a language model:
  - Acoustics, Phonetics
  - Statistics about sequences of phonemes
- Use of **machine learning** techniques to classify languages/dialects

Process steps within the language model



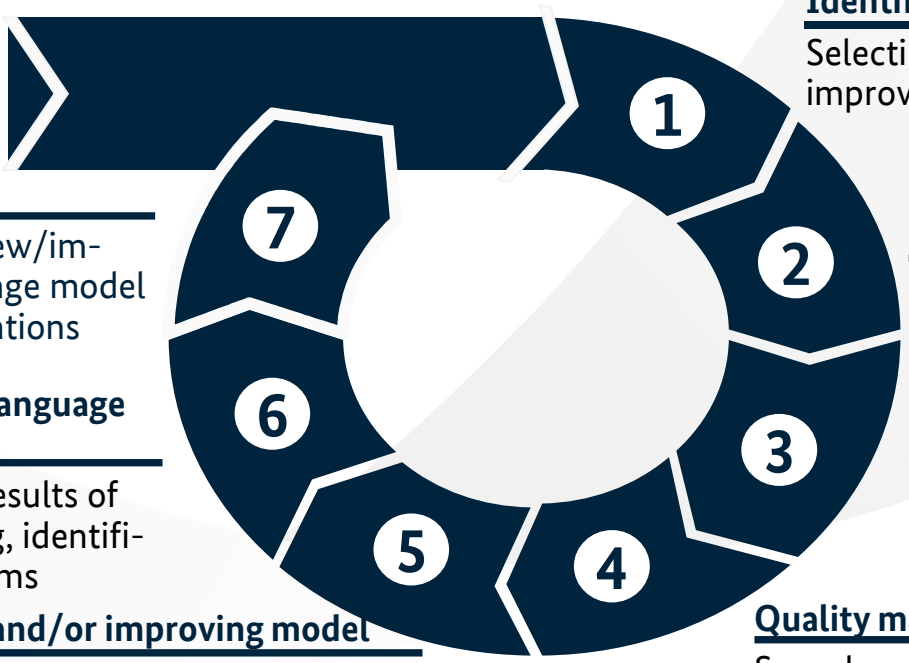
The system is based on elements from AI and computer linguistics

The editing process of the audio signal emulates the human cochlea's frequency analysis.

The next processing steps apply scientifically verified procedures: **i-vector** analysis (sound analysis, pronunciation of phonemes and sound production), **phoneme distribution** statistics as well as **syntactical analyses**.

PLDA („Probabilistic Linear Discriminant Analysis“) and **SVM** („Support Vector Machine“) are used for classification.

Improvement and expansion of DIAS



Identification of languages and dialects

Selecting languages and dialects to be improved or added

Identification of language data sources

Identifying relevant sources to procure speech/language samples

Procurement of speech samples

Procurement of suitable speech samples compliant with BAMF quality standards

Quality management

Speech samples are validated by experts to warrant (audio) quality

Developing and/or improving model

Integrating new speech samples into the language model, training and testing the expanded model

Evaluation of the language model

Evaluation of the results of training and testing, identification of action items

Roll-out

Integrating new/improved language model into live operations

Additional speech samples are key for the improvement and expansion of the language models. Searching for new sources for speech samples is an ongoing effort.

eu2020.de

DIAS: Evaluation

The DIAS tool - what does it do and what not?

The technology of BAMF's DIAS tool is complimentary, not a replacement

- The resulting report is **another resource** to assist the case worker
- The probabilities for the dialect spoken detailed in the report are a **first indication**
- Against this background the plausibility of the asylum seeker's narrative can be tested by **targeted questioning**

- The tool does **not** intend to automate existing processes
- The report provides **no** basis for the final decision.
- The tool does **not** curtail the essential personal and human interaction throughout the decision-making process.
- The tool cannot identify individuals.

Accomplishments and challenges

- Based on a short sample of fluent speech the tool provides a **fast, generally reliable, first indication** of the dialect or language spoken.
- **Implementation** of the tool is **fast and easy**, it smoothly **integrates into existing workflows**
- The system **cannot assess** the technical quality of the speech sample during the recording (e.g. too much background noise) and thus returns distorted results. A second recording might be required. This is indicated in the report.

07 October 2020

Session II: Cooperation in the field of language analysis on a European level

The challenge of identity and country of origin

Almost all processes in the field of migration and asylum start with two simple questions

Who is this person?

Where is the applicant coming from?

Clear identity and origin are no longer the rule

Experience from Germany

A significant number of asylum seekers apply for asylum without a passport or ID card

Illegally obtained and counterfeit ID cards further decrease the number of clear cases

Major challenges

The asylum procedure is more resource intensive and often takes additional time

Country of origin does not accept the asylum seeker without reliable evidence

Necessity for new methods of quickly clarifying the country of origin

eu2020.de

Language test for clarifying the country of origin

Language analysis and indications can confirm applicants' statements regarding their origin.

In refuting cases it can point towards the actual spoken language and country of origin.

Occurring challenges with full linguistic origin identification analysis

- Cost intensive
- Time consuming
- Capacity limits

Newly developed methods can yield fast results on a large scale.

eu2020.de

Fast track indications & combining strengths

human-based ‘fast-track’ language indication pre-tests

Germany has implemented software-based automatic language indication tests called DIAS (Dialect Identification Assistant)

full linguistic origin identification analysis

The combination has the potential to address the described challenge of country of origin determination

Differences between indication and analysis

Language indication

conducted mostly by **native speaker** analysts

DIAS examines similarities such as the frequencies of certain **phonemes** and their combinations

quick and preliminary analysis of the applicant‘s language, based on **short speech recordings**, no extensive expert report

used **as early as possible** in the asylum procedure,
but **not usable in court**

→ Language indications signify if a full analysis should be done, i.e. when the claimed origin is not confirmed by the language indication

Language analysis

conducted by **either linguists** with in-depth research knowledge of the language in question, or by **linguists** in **combination** with **native speakers**

profound, extensive analysis, identification of **grammatical**, **morphological**, **syntactic**, **phonetic** and **lexical features**,
based on **longer speech recordings**

qualitative assessment, e.g. “based on linguistic evidence, it is possible, likely, highly likely, unlikely [...]“

→ Full linguistic report, which delivers an overall picture of consistencies or inconsistencies with the language/dialect and claimed origin, and is valid for use in court

eu2020.de

Combining strengths in a common process

Vision

Common process for language indication and analysis accessible to member states and authorities of the European Union as well as partner countries

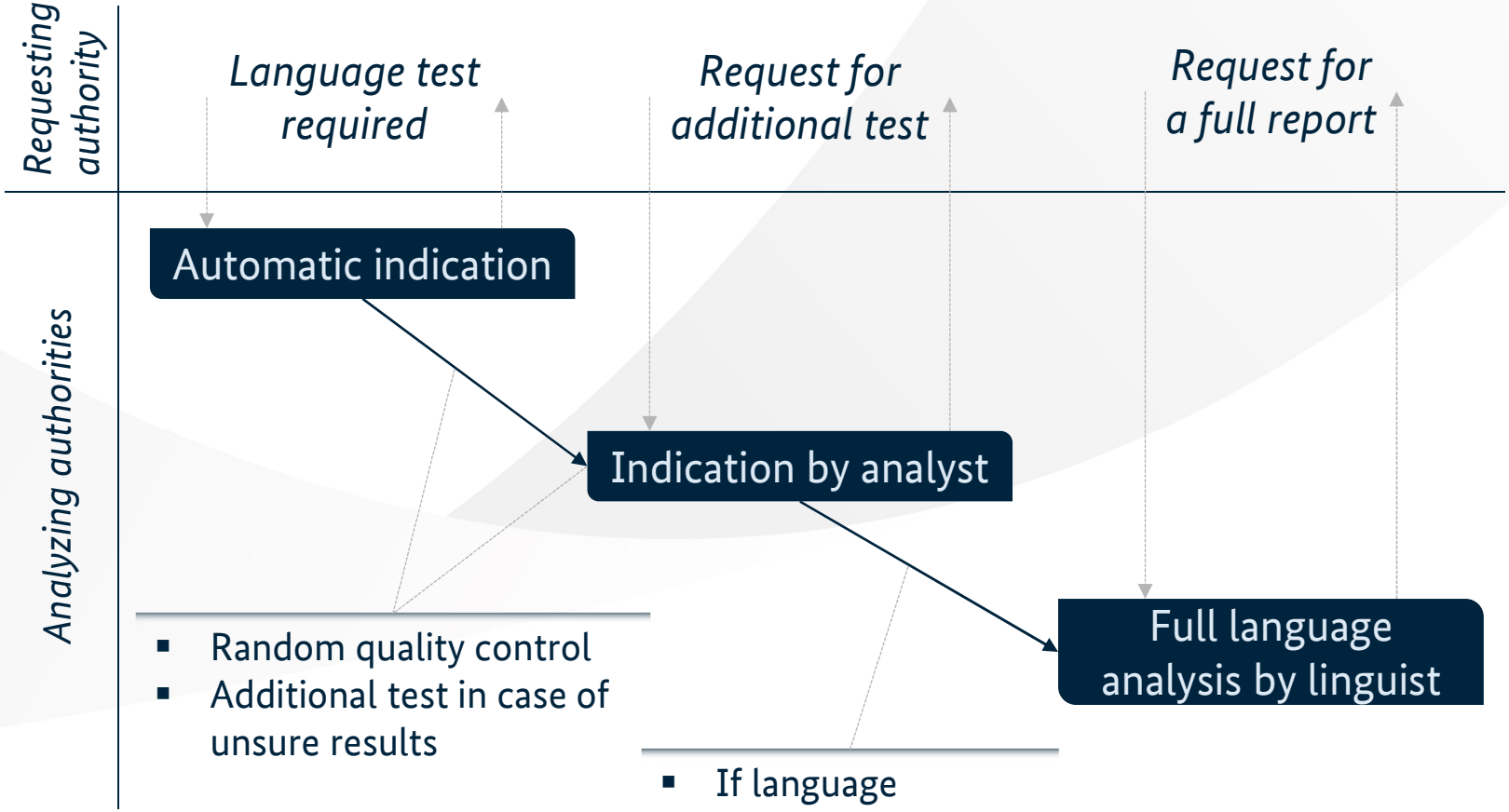
Combine

- the strong suits of human-based and automatic language indication (fast, large scale application)
- with the strong suits of human language analysis (valid in depth analysis, usable in court)

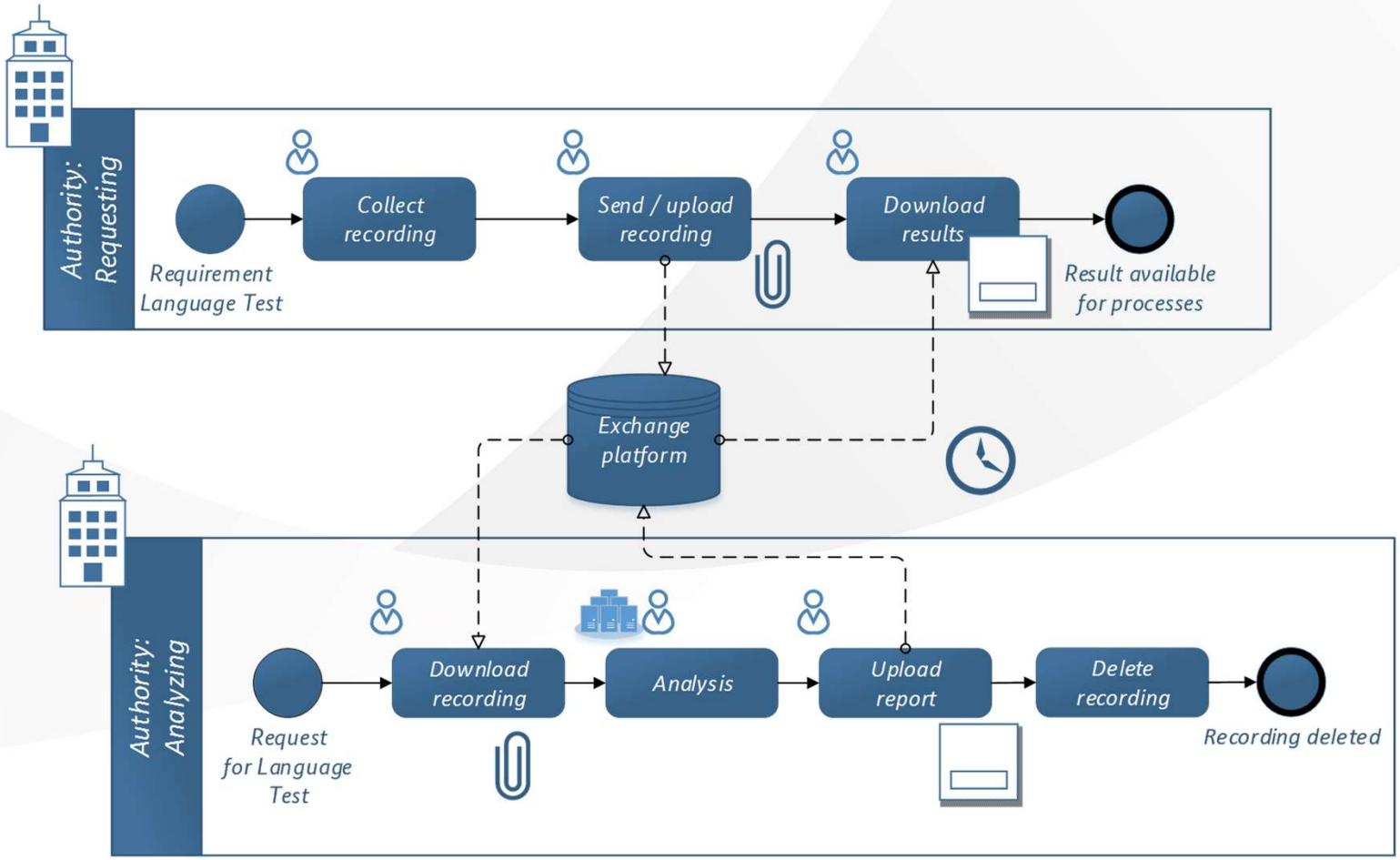
A first idea is under development together with several countries in Europe

eu2020.de

Idea for a common process



Procedure for requests



Advantages complement each other

Automatic indication

Fast available results

Large scale application

Strong indications, especially combined

Indication by analyst

Indications are helpful for interviews, internal processes and as a piece to the overall origin determination – but are no proof

Full language analysis by linguist

Can be time consuming (weeks to months)

Cost-intensive

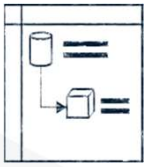
Prone to influx

valid in-depth analysis

usable in court

Pilot study to assess the feasability

Planning



Pilot preparation

Pilot

Consolidation of results

Feasability

Legal

Technical

Organisational

Contact information

We would like to take the exchange and cooperation on the project of European language analysis to the next level. If you would like to work with us on this project, please express your interest by sending us an e-mail to the following address:

IDM-S-International@bamf.bund.de

Many thanks
for your attention!