Turkic Languages, Dialects, Accents:
A computational geolinguistics project

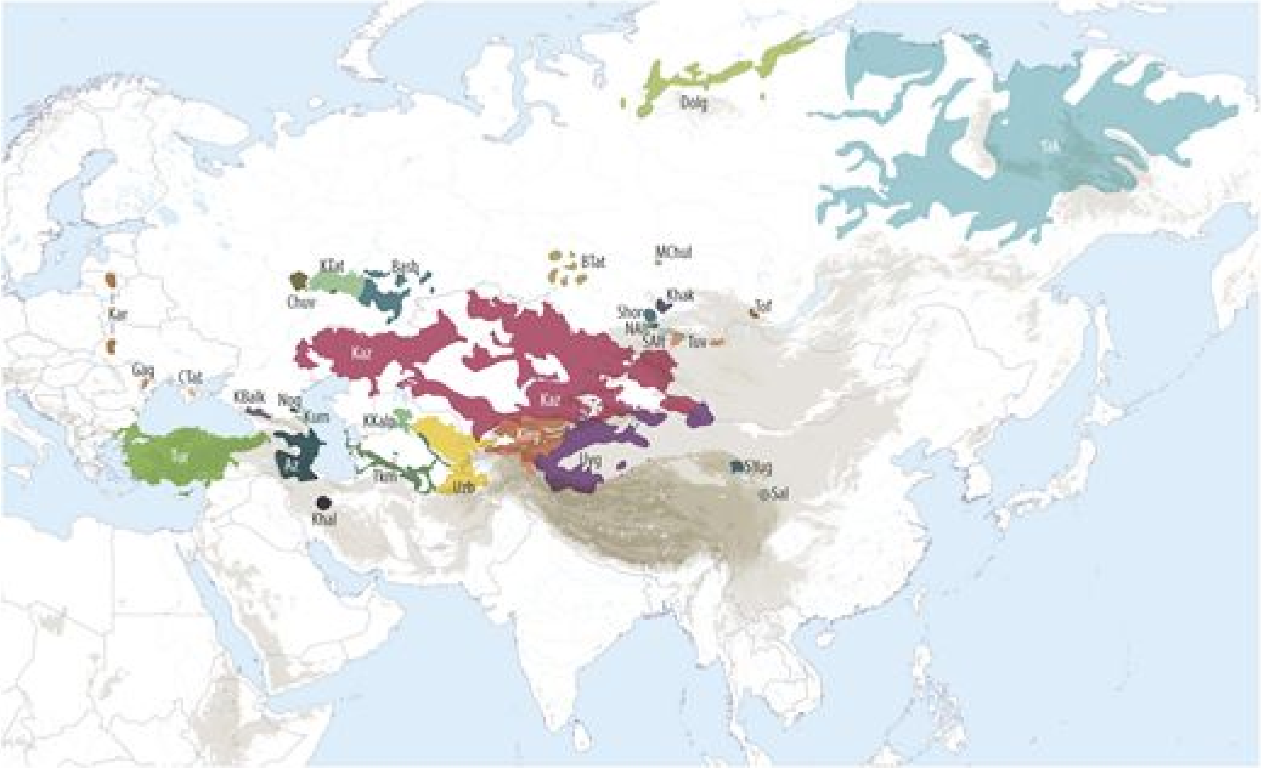

Figure from here: https://academic.oup.com/jole/article/5/1/39/5736268, see large
“Future studies in this area may benefit from better documentation of some poorly described varieties of Turkic.”

Step 1) Research questions:

-- Linguistic distance vs. Geographic distance, factors that are to do with terrain, historical events, trade relationships. Probably doable to some extent.
-- Modeling the ‘linguistic distance’ based on audio frequency analyses? Support also with corpus analyses?
-- Beyond Turkic? Mongolian, Japanese, Korean, Finnish, Hungarian? Testing the Ural-Altaic hypothesis? This is very big and requires expertise we don’t currently have :)

Step 2) Interface for data collection: Show images, collect text and speech input.

-- Do we collect anything else?
-- Open field for comments
-- Ask them to leave their contact information if we can get in touch for more later (so if we want to refine data collection or get more data we can reach them)
-- How to decide the images?

-- How many images are enough?
-- App for Android, iPhone, Browser
-- Gamification to get them contribute more (but a minimum should be set)
-- Rank for perceived linguistic distance? (once there's some data they can hear others and rank how much they understand, before an explanation/ after an explanation?)
-- Similar apps/studies?

-- https://twitter.com/sdatsbern (funded swiss dialects project has a twitter account)

-- 'Swiss Voice App': A smartphone application for crowdsourcing Swiss German dialect data: **https://www.zora.uzh.ch/id/eprint/105411/**

-- Analyzing geospatial variation in articulation rate using crowdsourced speech data : **https://core.ac.uk/display/83921046?source=2**

-- The English Dialects App: The creation of a crowdsourced dialect corpus: **https://www.sciencedirect.com/science/article/pii/S2215039017300589**

-- **https://babadada.com/** has picture dictionaries for many languages, it may be useful for selecting more modern words

-- Another commercial site, https://forvo.com/, collects word pronunciations from users

….
...

Step 3) Analyze, report, publish initial findings
Step 4) Grant application? Possible funding sources: EU, some cultural foundations?
Specification of the (pilot) application

The application should start with a small set of demographic questions and answers. and should cover information like:

- The age/gender of the speaker [form, auto-fill recommender]
- Native language(s?) of the speaker [form, auto-fill recommender]
- The language that (s)he will contribute data for (not necessarily the native language) [form, auto-fill recommender]
- If not native, the level of proficiency [Likert?]
- If native, other languages the speaker uses daily [form, auto-fill recommender]
- We may also want to gather some information about the device (it can be an indication of quality of recordings we can get) [automatic retrieval]
- … information from UK dialect app may be useful here …
- We can collect data from ALL languages of the world (no restrictions in data collection)

After this brief q&a all we want is to ask the participant to record the name of each item shown on screen. We can start with a subset of the Swadesh list. We may later want to change this to align with the wordlists collected in other projects, but should not be a big concern for the development of the application.

> Question: See the table here https://en.wikipedia.org/wiki/Turkic_languages
> Do we arrive at something else/larger/different than this table? Text reading? Sound analyses?

At the end of the session (maybe during) the app should pass the recorded data collected to a server. Server side is rather simple. All we want to receive the data and store it in a reasonable way. At the beginning we can make this only a web application,

A few additional notes:

- It is a good idea to keep in mind that in the long run we want a “internationalized” interface. At the beginning we can simply assume that the interface language is English.
- We need “unambiguous” and nice images for people to say their names. We may need to do some research and get copyright free drawings/pictures, or we can try to make them ourselves in a consistent way. But, again this can be done later/ or in parallel.
- Potentially we may also give some text to read.
- We should be fine to collect data “as much as” the participant gives. So, not to bias the data set with the order, ideally we should consider collecting the sound data in a random Order.
- We should allow the same person to record as many times as they want, and pass all versions to the server side.
- There will be other issues, like the quality of the sound (because of the equipment, or environmental noise), we may think about doing a quick analysis and giving some feedback online (e.g., asking the user to speak close to the microphone, or move to a quiet place). But, again, this is too much forward thinking. Let’s see the problems first.