EU Language Skills (Narrative Documentation)

Marco Civico

A model of language acquisition

In order to replicate the environment of the European Union, the agents are provided with a set of characteristics collected from actual databases. In the following subsections I explain how the database that feeds the model is constructed. The final database draws from two different databases, the Special Eurobarometer 386 and the linguistic distance database compiled by Dyen et al. (1992).

The Special Eurobarometer 386 survey

A number of properties are assigned to agents using data from the "Special Eurobarometer 386: Europeans and their Languages" (European Commission, 2012). The Eurobarometer surveys are conducted periodically on behalf of the European Commission and investigate many issues throughout its member states. They focus particularly on the citizens' perception and expectations towards the intervention of the European Union and the challenges that it faces. The topics covered by the survey are numerous, ranging from air quality, gender equality and democracy to sports, trade and climate change.[1] In particular, the Special Eurobarometer 386, carried out in 2012, is the latest survey concerning languages (similar surveys were carried out in 2001, 2005 and 2006) and provides information about language skills and the attitude of EU citizens towards multilingualism and language services. It covers the then 27 member states (the 28 member state, Croatia, joined the EU in 2013). The survey includes information collected from 26,751 EU citizens from different social and demographic groups, aged 15 or older and residing in a EU member state, with a view to making the results of the survey as representative of the whole EU population as possible.[2] In addition to information about languages, the database includes

---

[1] For a list of the topics covered by the Eurobarometer surveys, see https://ec.europa.eu/COMMFrontOffice/publicopinion/index.cfm

[2] The sampling procedure is clearly detailed in the appendix to the Eurobarometer 386. In short, the data were collected through a multistage random sampling by drawing from each country a number of sub-units with probability proportional to population size and density. These random sub-units were then once again sampled at random down to household level. To ensure the right coverage of the EU territory, sub-units were collected from each "administrative regional unit" according to the EUROSTAT NUTS II classification. Finally, interviews were conducted in person in the respondent's home and in the appropriate national

general demographic information, such as age, sex, profession and education. The survey collects data from roughly 1000 respondents per member state, except Cyprus, Luxembourg and Malta, with only about 500 respondents each. In the interviews, respondents were asked about the languages they speak, their motivation to learn foreign languages, the difficulties they encountered, the situations in which they resort to foreign languages, the impact of translation in their life, and so on.

For the purposes of the model presented here, I use a section of this database. For the practical reasons already explained, four countries had to be excluded from the analysis, namely Finland, Hungary, Estonia and Malta. In the model, each agent reproduces the linguistic, social and demographic profile of an actual respondent from the survey. In particular, agents have the following properties:

1. country of residence;
2. nationality(ies);
3. age;
4. age when they finished education;
5. mother tongue(s);
6. profession (one of eight categories);
7. living condition (countryside, small city or big city);
8. first, second and third foreign language (if any) and related level of fluency (from 1, basic, to 3, very good).

As the focus is on the impact of language acquisition on linguistic disenfranchisement given the language regime, I look exclusively at the competences of citizens in the official languages of the EU countries considered, disregarding all other languages they might now. As a consequence, agents that spoke none of these languages were also excluded, leaving a database of 21,890 observations.

The linguistic proximity matrix

In addition to the ones mentioned above, agents have an additional property, whose value depends on their mother tongue(s). Based on their native language, each agent takes on a vector of values that defines the linguistic proximity between their native language and all official languages of the environment simulated. The notion of linguistic proximity (or, equivalently, that of linguistic distance) is certainly interesting, but it poses many issues, in that it is highly dependent on the method used to compute it. For the purposes of the model developed here, I use the concept of linguistic proximity with a view to having a proxy of the subjective perception of each individual when it comes to picking a

---

language.

language that is closer to her native language and that she could learn relatively faster.

The information about linguistic distances comes from a different database, adapted from Dyen et al. (1992). The linguistic distance between languages is estimated using the lexicostatistical method, first introduced by Swadesh (1952). The method works by and large as follows:

- in the first phase, one prepares a list of basic terms that exist in all the languages that one wants to compare and collects the related terms;
- in the second phase, one looks at the terms for the same meaning across languages and establishes whether they are *cognates*, i.e. they descend from the same root;[3]
- in the third phase, one goes on to compute the percentage of cognate words within the list considered across pairs of languages.

It is often more intuitive to speak of *linguistic proximity* rather than linguistic distance. The value of the linguistic proximity index goes from 0 (no cognate words in the list considered) to 1 (all words considered are cognate). Intuitively, the higher the value of the index, the closer the languages under study. For example, Swadesh (1952) finds that English and German are connected by 57.8% of the 200 words that he considered (or, equivalently, have a proximity index of 0.578), while French and English are connected by 23.6% of the words (or have a proximity index of 0.236).

For the purposes of the model, it was necessary to create a 20 by 20 matrix that would include the pairwise linguistic proximity indexes for the Indo-European EU official languages. This choice stems from the fact that the original database by Dyen et al. (1992) only includes Indo-European languages. Therefore, the linguistic proximity index for some EU official languages, namely, Hungarian, Finnish, Estonian and Maltese, was not available. Therefore, these languages were excluded from the model. As a consequence, in order not to skew the results, Hungary, Finland, Estonia and Malta were also excluded from the database. The linguistic proximity indexes are graphically represented in figure 1.[4]

Simulating language learning

In the setup phase of the simulation, agents are created and residence is assigned proportionally to the actual distribution throughout the EU. As residence is more or less uniformly distributed in the original database (roughly 1000 observations per country), the model samples randomly out of it, in order

---

[3] It should be noted that two terms in two different languages are considered cognates if and only if they are descended from the same ancestors, and not if a language simply borrowed the term from the other.

[4] It should be noted that the specific location on the graph does not carry any particular meaning. It only indicates the relative position of languages with respect to one another.

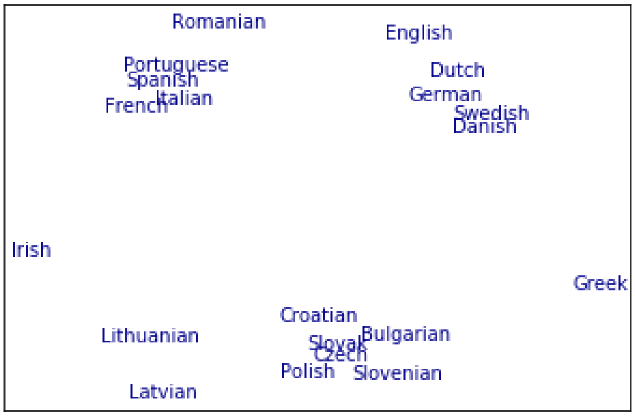

Figure 1: Language proximity map of the official languages of the European Union (Indo-European family).

to replicate the actual distribution of residents by country.[5] All other properties are assigned by randomly selecting a respondent from the database and assigning her properties to an agent with the same residence, allowing for repetitions. When the simulation is launched, agents are asked to make a decision about learning a language. In the simplest case, an agent does not speak any EU language (other than her own). The agent is then asked, with a certain probability, to start learning a new EU language (I will discuss later how the agent selects the language to learn). In case the agent already knows one or more foreign languages, she is asked to look at her level of fluency in them (which, as said, goes from 1 to 3). If she speaks a foreign language at a level of fluency lower than 3, she is asked, with a certain probability, to go on learning it until she is proficient in it (i.e. she reaches level 3).[6] If the agent's foreign languages are all at level 3, she picks a new one with a certain probability, based on the rules explained below. This process goes on as a long as an agent knows less than three foreign languages at level 3.[7]

---

[5] In the simulation performed, I sample 50000 agents (with repetitions) with the following distribution: Austria, 915; Belgium, 1177; Bulgaria, 707; Cyprus, 123; Czech Republic, 1088; Denmark, 589; France, 6633; Germany, 8513; Greece, 1059; Ireland, 502; Italy, 6144; Latvia, 191; Lithuania, 276; Luxembourg, 64; Netherlands, 1741; Poland, 3846; Portugal, 1037; Romania, 1955; Slovakia, 554; Slovenia, 211; Spain, 4751; Sweden, 1026; United Kingdom, 6898.

[6] In case the agent speaks more than one foreign language at the same level of fluency, she picks one of them based on the rules explained below.

[7] I am aware that this is a relatively extreme choice, given that not everyone is willing (or simply has the chance) to learn three foreign languages. However, I can make two considerations to justify this choice. First, this number is certainly a limit, but only a very small minority of agents get to learn three foreign languages in the time allowed for the simulation. Therefore, it should rather be seen as a possibility to learn more languages, should an agent be able to do so. Second, having a high cap to the number of languages can be interpreted as a context in which the acquisition of language skills is highly encouraged, which is the scenario

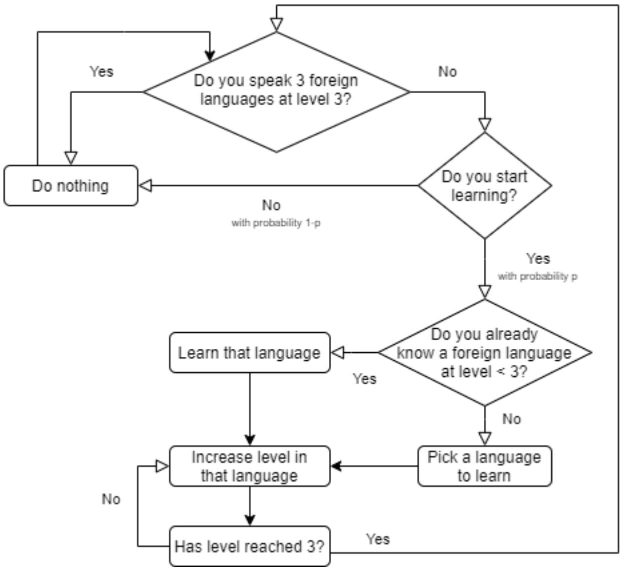

Figure 2: Flowchart of agents' behaviour.

A property that does not belong to any agent but to the environment is the language regime. I consider three types of scenarios, that is, monolingual (monarchic), trilingual and hexalingual (oligarchic with, respectively, n=3 and n=6).[8]

In the first iteration, agents are asked to start learning a language with a certain probability *p*. Once they start learning a language, they keep learning until they become proficient in it. At every time step, an agent learns the language she has picked, that is, her level of fluency in that language increases. The speed at which she acquires skills in that language (that is, the speed at which her level in that language increases from 0 to 3) depends directly on the proximity between the language that she is trying to learn and the closest language among her native language(s) and the foreign language(s) in which she is fully proficient. The shorter the distance between the two languages, the fewer the time steps required by the agent to reach proficiency in the new language. When the agent has reached proficiency, she switches her status to "not learning" and the process starts again, that is, she will not be learning a new language, unless, again with probability *p*, she is not asked to do so. Every agent is programmed to be able to learn up to three foreign languages in total, including the ones that she already knows from the start. The behaviour of agents is summarized in Figure 2.

---

that I am interested in exploring.

[8] In the simulated database, the six most spoken languages as native or foreign language (at a proficient level) are, in descending order, English, German, French, Italian, Spanish and Polish. These are the languages taken into consideration for the three language regimes.

The languages are learned in succession and not in parallel. This is due to two reasons: first, I find that it is more reasonable for an individual to focus on the acquisition of one language at a time; second, as agents can base the choice of the language(s) to learn on their skills in all the languages in which they are proficient (native or non-native), becoming proficient in a foreign language can influence the pattern of future choices.

As has been explained before, if an agent already knows one or more foreign languages at a less-than-proficient level, it is assumed that she is currently learning it and will keep doing so, starting from the one that is closer to proficiency. When asked to pick a new language to learn, agents can then use one of three strategies, which is selected before the simulation is launched and applies to all agents:

1. they can pick the language that has the highest number of native (L1) and non-native (L2) speakers;
2. they can pick the language that is closest to (one of) their native language(s) or to any other language in which they are fluent, that is, the language they will learn in the shortest amount of time;
3. a combination of strategies 1 and 2, that is, they can pick a language close to their own and having a relatively high number of L1 and L2 speakers.

During the simulation, the model updates and keeps track of a number of values. It keeps track of the total number of L1 and L2 speakers of every language. This information is crucial and affects the model in two ways. At the micro level, it affects individual agents' decision about the language they should learn, if they are using strategies 1 or 3 in the list above. At the macro level, this information is necessary for the system to establish the OWLs of the language regimes. Indeed, the three regimes consider, respectively, the one, three or six most spoken languages. After the simulation, I use the data generated to calculate the level of disenfranchisement, which is a direct consequence of the language regime. Finally, given that the model keeps the original information regarding the individual properties of the agents, it is possible to identify the socio-demographic categories most likely to be disenfranchised.

References

Dyen, I., Kruskal, J. B., and Black, P. (1992). An indoeuropean classification. a lexicostatistical experiment. *Transactions of the American Philosophical Society*, 82(5):iii–132.

European Commission (2012). Special eurobarometer 386. european and their languages. Technical report.

Swadesh, M. (1952). Lexicostatistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society*, 96:452–463.