*LanguageMap.eu Research Brief #2*:

# Measures on how to critically and scientifically engage with language databases and maps

## Complexities to keep in mind

While it may be tempting to take data presented by prestigious organisations, national institutions or statistics offices at face value, it is important to **contextualise** and **challenge** such data. Every method for determining language data such as speaker numbers, minority numbers, language status, or language endangerment has **limitations** and **blind-spots**. Below, you can find a selection of current European language data complexities.

### On Census Data

Compared to smaller-scale studies that can be limited in their scope, misrepresentative, or influenced by the ideologies and affiliations of participants and researchers, census data is sometimes assumed to be a 'more neutral' or 'more reliable' source of information. However:

- **Census =/= Census:** There are different types of censuses. Traditionally, door-to-door censuses were and often still are conducted in annual rotations (e.g. every five or ten years), but over the last decades, continuous statistical analyses based on already available data (e.g. registers) have risen in popularity in some countries. These methods are also often combined, as e.g. in the Slovak 2021 Census. Additionally, census data can be based on sample censuses. These are conducted with smaller 'representative' populations rather the actual population of a given country or region (as e.g. the Mikrozensus in Germany).

  **As a starting point for learning more about types, histories and censuses:**

  MacDonald, Alphonse. (2020). Of science and statistics: The scientific basis of the census1. Statistical Journal of the IAOS. 36. 1-18. 10.3233/SJI-190596.

  Matras, Yaron & Chau Príncipe, Santiago (2023). Changes to Census figures on reported 'main languages' in Manchester, 2011–2021. Manchester City of Languages

  ➔ **Recommendation: Always investigate which type of census you are looking at.**

- **Data =/= Data:** Data from two different censuses (e.g. from two different countries) need not be comparable. Without context, it could be said that there are 199 000 German speakers in Poland (2021 census), while there are 1 207 058 German speakers in Hungary (2022 census). However, who is 'a speaker'? In the Hungarian census, the question asked how many people considered German to be a 'spoken language' for them, while the Polish census specifically asked for whom German was 'a household language'. It stands to reason that there are likely even more German speakers in Poland who do not speak German at home.

  ➔ **Recommendation: Always determine which '(e.g. speaker) category' was used as the basis for calculating the overall number.**

- **Local =/= Local:** Even within a single country, census data need not align and will include varying approaches and formulations. Consider, for example, that in their estimates for UK sign language users, the British Deaf Association (BDA) specifically refers to data from the Scottish 2011 Census as the question concerning British Sign Language was "badly phrased in the Census for England, Wales and Northern Ireland" (Source).

➔ **Recommendation: Also when working locally, analyse the questions asked.**

- **Structure influences outcomes:** In the same vein, the structure of any given census will strongly impact its results. Accessibility is an important factor – is the census accessible, understandable, easy to fill out, and offers a variety of responses? The latest census results from Poland in 2021, for example, seem to indicate that the number of Silesians and Kashubians has notably decreased – however, "some activists and experts claim that this is because the new online census made it more difficult for people to select such identities" (Source). Elderly people may struggle with technological hurdles, and restrictions in potential answers can also streamline more complex realities. In Finland, Sámi people have the right to declare Sámi as their mother tongue for the Population Register – however, only one language can be selected; thus, a significant number of Sámi speakers choose Finnish and go 'unrecorded'. Lastly, it is important to note where the lines between different speaker groups and language varieties have been drawn: no accurate statistics are available for the number of Moravian Croats in the Czech Republic, for example. In recent censuses, the 'Moravian' nationality option was broad enough to also have been selected by Croats who arrived in the Czech Republic much later, e.g. after the Yugoslav conflict in the 1990s. Altogether, they came up to 16 523 in the 2021 census, while experts estimate the true number of only the Moravian Croatians to be in the hundreds.

➔ **Recommendation: Ascertain the limitations of the census in question – was it on- or offline? Were multiple answers allowed? Which ones were they?**

- **Underlying philosophies and ideologies differ:** Europe runs the gamut when it comes to positionalities towards official language recognition and data collection. The stance of Sweden is that it does not collect official language statistics because "no methods are available for determining ethnic origin that are both ethically acceptable and scientifically reliable" (Source). Belgium legally abolished language censuses in 1964 following fierce territorial disputes between different language groups. Censuses are not apolitical and are strongly connected to local political realities – in the case of Belgium, percentage thresholds of census results were used to geographically shift a language border: as soon as 50% of the local population declared that they spoke another language, the municipal administration could change the official language. To avoid future conflict, Belgium thus established four fixed language areas. Percentage thresholds have also been an issue, e.g. in Albania, as certain minority language rights are restricted to specific areas, making it difficult for more scattered populations to gain access to their language or educational rights. Such access is one of the primary

arguments of proponents of gathering language data – without official recognition, awareness, and as-up-to-date-as-possible statistics, it is very difficult to make the case for relevant public policy, which has been an ongoing challenge, e.g. in Slovenia. A symbolic court case was that of Jan Coucke and Pieter Goethals in Belgium in 1860 who were allegedly hanged without having understood the language of court proceedings and without having been able to properly defend themselves. Legal rights and representation, and "better access to public services – including education, healthcare, media and banking" are particularly relevant for the deaf community, as well as for those who are hard-of-hearing or deafblind. Next to no reliable data on the number of sign language users is available throughout Europe, and there is particularly scant data for the deafblind. Many historically underrepresented language groups find themselves in a similar situation.

➔ **Recommendation: Adopt a socio-historical and political perspective for a better understanding of EU member states' individual approaches.**

**Language data is not gathered or interpreted 'neutrally':** Political and historical factors do not only influence whether censuses are held and groups are included, but also how respondents present themselves when asked. Belonging to a certain group is often associated with real-world consequences outside of scientific inquiry, and this also factors into self-representation. While the country was still under communist rule, Albanians in Romania, for example, often declared themselves Romanians as they feared state repression and the borders to Albania were almost completely closed. Serbian speakers in Croatia are also a good example for how ethnicity numbers and speaker numbers do not always necessarily align due to various reasons (widespread diglossia, political alignments, a higher prestige of the local official language, etc.): in the 2021 census, 123 892 people living in Croatia declared themselves to be Serbian but only 49 282 declared Serbian as their native language. A last example for socio-political pressures comes from Albania: according to the latest census, the number of Albania's Macedonian minority was nearly halved. They argued that it was because "members of their community were pressured by Bulgaria to identify as Bulgarian, reportedly incentivized by promises of EU passports, particularly in eastern regions bordering Kosovo and North Macedonia" (Source).

➔ **Recommendation: Go beyond a historical perspective to current lived realities to understand which pressures are currently influencing self-presentation.**

**On Maps**

Most considerations for working with language maps are outlined in the 'About' Section of LanguageMap.eu. The most crucial recommendation is to keep in mind that languages are **not** confined to the concept of modern nation-states – most language varieties extend across borders, and only very few extend across an entire nation. Many exist in so-called 'language islands' and others in loosely connected or isolated regions. It is our hope that future maps will be able to reflect this reality more accurately. *Author: Marie-Therese Sauer (MINDS & SPARKS)*