Working Papers in

Urban Language & Literacies

Paper 189

Categorizing languages and speakers: Why linguists should mistrust census data and statistics

Brigitta Busch (*University of Vienna*)

2016

Categorizing languages and speakers:
Why linguists should mistrust census data and statistics

Brigitta Busch

University of Vienna

Abstract

This article concentrates on how census data are processed by statistical offices. It first investigates how the language categories displayed in language statistics are actually formed. After that, it considers the mechanisms by which speakers are assigned to specific language categories. With both processes, it addresses the underlying language ideologies, and the ways in which a certain view of the world is thus transported and reinforced. Finally, it discusses how these practices are linked with the paradigms of statehood and territoriality, and how they have been developed historically in the contexts of colonization and of ethno-national conflict. The paper draws on censuses and statistics in Austria to explore these questions, but the procedures used correspond to those in other countries that follow the guidelines set by the Conference of European Statisticians (United Nations 2006).

Introduction

At the beginning of his treatise on the history of the human sciences, *The Order of Things*, Michel Foucault (1994: xv) places a text by Jorge Luis Borges (1964). This text, writes Foucault, made him laugh, a laughter that shattered all the familiar landmarks of thought, "breaking up all the ordered surfaces and all the planes with which we are accustomed to tame the wild profusion of existing things", a laughter that disturbed and threatened with collapse "our age-old distinction between the Same and the Other". The text, which Borges attributes to a Chinese encyclopaedia, states that "animals are divided into: (a) belonging to the Emperor, (b) embalmed, (c) tame, (d) sucking pigs, (e) sirens, (f) fabulous, (g) stray dogs, (h) included in the present classification, (i) frenzied, (j) innumerable, (k) drawn with a very fine camelhair brush, (1) et cetera, (m) having just broken the water pitcher, (n) that from a long way off look like flies".

My concern in this article is to call into question the sometimes careless way in which statements are made about the numbers of people who speak certain languages. I refer to this well-known passage from Foucault because it raises fundamental questions, questions that confront any attempt at a categorization and classification of languages, without which it is impossible to count, numerically survey or establish quantities. Numbers and relative sizes traditionally play a significant role in language policy and language planning, especially when it comes to defining the status of languages within a state structure, for example as regional or minority languages, indigenous or endangered languages. They are also important, however, when it comes to establishing a system of proportionality based on membership of language groups, to regulate the distribution of political power and financial resources, or when it comes to analysing past developments and forming prognoses about future demographic developments, as a basis for defining education or "integration" policies. The rights of language minorities are often dependent on substantial numbers of speakers being found within particular geographical areas. In the political debate, however, language use figures can also be interpreted simplistically as evidence of ethnic affiliations, and can thus be misused, for example to criticize immigrant population groups for a lack of willingness to integrate. Even in scholarly work on language policy, there is a great temptation – though tempered with a certain malaise – to fall back on figures that have been

gathered in the context of censuses or school statistics, without accounting for how these figures were reached, and how the underlying categories were established.

Regarding data on language collected through censuses, a number of questions have been raised in the context of minority rights (for an overview see e.g. Heintze 2005). One area of questioning is the circumstances and the purpose of censuses: who commissions them (government agencies or non-governmental organizations), in what political climate does the census take place, who are the census-takers, and to what extent is anonymity guaranteed? Another concern is the kind of questions asked: are respondents asked about their 'native language' or 'mother tongue', the languages most currently spoken or the language they think in? Are answers provided as closed lists of prescribed categories, is there space for respondents to name other language themselves, is it possible to name multiple languages, and is responding to the question optional or compulsory?

This article, however, is less concerned with the process of census-taking, and instead concentrates aspects on how census data are being processed by the statistical offices. The first question that I will investigate is how the language categories displayed in language statistics are actually formed. The next question is how – by means of what mechanisms – speakers are assigned to specific language categories. In both cases, we are interested in what language ideologies underlie these processes, and how a certain view of the world is thus transported and reinforced. Finally I will discuss in which way these practices are linked with the paradigms of statehood and territoriality and how they have historically been developed in the contexts of colonization and of ethno-national conflicts. In my paper I draw on censuses and statistics in Austria to explore these questions. Austria serves merely as an example: the focus is on the procedures used, which largely correspond to the procedure in other countries that follow the guidelines set by the Conference of European Statisticians (United Nations 2006).

In these Recommendations for the 2010 Censuses of Population and Housing (United Nations 2006), the Conference of European Statisticians recommends including questions about language in the questionnaire. Language – alongside religion, specific customs and ways of life – is regarded as a characteristic of "ethnocultural affiliation" (§ 419). § 430 states that:

> Multilingual countries and countries with significant immigrant populations may wish to collect data on languages that are currently written or spoken. Depending on information needs, the following data may be collected:
> a) "Mother tongue", defined as the first language spoken in early childhood at home;
> b) Main language, defined as the language which the person commands best;
> c) Language(s) most currently spoken at home and/or work;
> d) Knowledge of language(s), defined as the ability to speak and/or write one or more designated languages.

While the questions on a) and b) are described as "relevant to understand processes of language change and to determine language regions and language groups", and only one answer is generally expected (§ 431), the questions on c) and d) are seen as "relevant to understand language practices and knowledge of languages, including official languages and languages learnt at school" (§ 432). The question about language ability is specifically mentioned in connection with migration, suggesting that the main concern here is proficiency in official languages within the state structure. It is recommended that at least one open answer box be included in questions for groups a) or b), and that classifications should be comprehensive and include, on the finest level, language groups, separate languages, and regional dialects, as well as invented languages and sign languages.

Individual states take different approaches in terms of whether and how questions are asked about languages. In Germany, no questions are asked about ethnic affiliations (and therefore none about

languages), in reaction to the fact that censuses and categorizations have been misused in German history to persecute Jews, Roma and other groups. Canada, since 2011, has chosen an elaborate combination of three language questions: knowledge of official languages, home language and mother tongue. The data are then analysed with regard to the linguistic duality English and French as well as to immigrant languages and aboriginal languages (Statistics Canada 2012). In the US, language use is mainly surveyed and mapped with a view to finding out to what extent individual language communities display "English-speaking ability" or tend towards "linguistic isolation" (United States Census Bureau 2011). In most countries, the censuses held every ten years include questions about the language or languages of the inhabitants in one form or other. The content of the questions and the focus of the analysis is mainly linked to two factors: firstly, traditional discourses about the constitution of the relevant state (e.g. whether it is defined as a unitary nation state, as multi-ethnic or multilingual), and secondly, the current issues that are seen as relevant on the language policy agenda – e.g. the recent emphasis on state languages versus languages of migration.

How language categories are formed

We take Austria as a case study for investigating the connection between language ideologies and the counting and categorizing of speakers and languages for statistical purposes, not because it is a special case, but because, on the contrary, the hidden discourses and ideologies that come to light here could certainly also be found in the practices of other states, if one were to look more closely. The first question that interests us is how language categories are formed and named. The language recorded statistically in Austria[1] is the *Umgangssprache* (the 'spoken language')[2], defined as the language "that is usually spoken in the private sphere (family, relatives, friends etc.)" (Statistik Austria 2005: 8). Statistics on population by *'spoken language'* (Statistik Austria 2007a) are led by the majority category "German only" (Table 1). This is followed by the categories covering those who have indicated other 'spoken languages' (sometimes in combination with German). The category "German only" is tacitly assumed to be the norm, and is set apart from the others as a whole. This leads to a pair of concepts defining two basic categories. As Jakobson (1990) showed in relation to the language system, dichotomous pairs of concepts are rarely symmetrical, but mostly have a hierarchical structure: priority is given to the unmarked term (in our case "German only"), while the marked term is classed as secondary and divergent. The first, unmarked category suggests that there is a normality that is above questioning, in our case that of the monolingual speaker of the majority language, German. The heterogeneous amalgam of all the others follows this and is set apart from it.

Derrida (1967) shows that, in this kind of binary logic of pairs of opposites, the dominant term is paradoxically defined by the very things that it excludes and marginalizes. In our case, the top, unmarked majority category, "German only", is constructed through the fact that any other constellation of languages is excluded from it. The unmarked category "German only" clearly reflects an idea based on language ideology: that the nation is formed in its substance by monolingual speakers of the state language. The construction of the other categories follows the same mechanism of exclusion, with all the other languages of the world being assigned to seven broad overarching categories (see Table 1): 1) languages of the recognized Austrian ethnic groups;

---

1   The most recent population data by 'spoken language' are based on figures from the 2001 census. Austria now, like other countries, relies on data from the Register of Residents (Melderegister) and other agencies; information on language use is now gathered in connection with the school system (BilDok) and in the context of surveys on levels of education (in the form of random samples).

2   The term *Umgangssprache* corresponding to the French term *langue parlée* originally used internationally in the context of censuses is commonly translated as 'spoken language' (cf. Arel 2002: 94). Therefore, in this text I will also make use of this English translation.

2) languages of the former Yugoslavia and Turkey; 3) English, French, Italian; 4) other European languages; 5) African languages; 6) Asian languages; 7) other languages, unknown.

Table 1: Statistics "Population 2001 by 'spoken language'" (Statistik Austria 2007a, translated by the author)[3]

'Spoken language'*	
Total	8,032,926
German only	7,115,780
Languages of the recognized Austrian ethnic groups [langues des groupes ethniques autrich. reconnus]	119,667
- Burgenland Croatian	19,412
- Romany	6,273
- Slovakian	10,234
- Slovenian	24,855
- Czech	17,742
- Hungarian	40,583
Languages of the former Yugoslavia and Turkey	534,207
- Bosnian	34,857
- Croatian	131,307
- Macedonian	5,145
- Serbian	177,320
- Turkish	183,445
- Kurdish	2,133
English, French, Italian	79,514
- English	58,582
- French	10,190
- Italian	10,742
Other European languages	116,892
- Albanian	28,212
- Bulgarian	5,388
- Danish	735
- Finnish	987
- Greek	3,098
- Dutch/Flemish	3,802

---

3 In this reconstruction the figures are presented as totals, not, as in the original, also divided by nationality and country of birth.

- Norwegian	569
- Polish	30,598
- Portuguese	3,197
- Rumanian	16,885
- Russian, Ukrainian, Belarusian	8,446
- Swedish	2,683
- Spanish	9,976
- other European languages	2,884
African languages	19,408
- Arabic	17,592
- other African languages	1,816
Asian languages	47,420
- Chinese	9,960
- Hebrew	1,189
- Indic [*Indisch*]	3,582
- Indonesian	451
- Japanese	1,806
- Korean	1,264
- Persian	10,665
- Filipino	5,582
- Thai	1,593
- Vietnamese	2,310
- other Asian languages	9,018
Other languages, unknown	38
*) Non-German 'spoken languages' include those mentioned alongside German	

Even at first glance it is obvious that these overarching headings do not follow any unitary ordering principle, but are formed according to criteria based on different patterns, criteria that seem somewhat arbitrary or random. For the category "Languages of the recognized Austrian ethnic groups", the decisive factor is the legal status granted to certain minorities regarded as traditional or autochthonous. The category "Languages of the former Yugoslavia or Turkey" refers to the countries of origin with which Austria made bilateral agreements for worker recruitment in the 1960s. This category may be explained by both the historical context and the relatively large number of immigrants from these areas. The category "English, French, Italian", separated from "Other European languages", brings together three languages that have only in common the quality of enjoying a high level of prestige in Austria, and are traditionally taught as foreign languages in schools. The three subsequent categories are formed on the basis of geographical and territorial aspects, with each language being assigned to the term 'European', 'African' or 'Asian'. One consequence of this allocation of particular languages to specific continents is that Spanish is classed as a European language, although it functions as an official language in 22 countries in the world, and although Mexico, the US and Colombia each have more speakers of Spanish than Spain itself.

Within each collective category, the languages included in it are listed in alphabetical order. In the category headed "Asian languages", for example, these are "Chinese, Hebrew, Indic [*Indisch*], Indonesian, Japanese, Korean, Persian, Filipino, Thai, Vietnamese, other Asian languages". The list is striking in several respects, and is as unsettling, in its way, as the taxonomy quoted at the beginning, attributed to a Chinese encyclopaedia. The first striking thing is absences. For example, Hebrew is named as an Asian language, but not Arabic. This is because Arabic – on the basis of what considerations? – has been assigned to the category "African languages". It is also noteworthy that languages are named after states. For example, it remains unclear whether "*Indisch*" (Indic) is to be equated with Hindi, or also includes other languages spoken in the national territory of India. However, what this terminological confusion clearly demonstrates is how strongly language ideologies associate or even equate 'languages' with territory, nation state and ethnicity.

Residual categories such as "other Asian languages" are present at the end of most categories, but not all. Thus the list of languages of the former Yugoslavia and Turkey gives the impression of a closed list. For Turkey, for example, the absence of a residual category means that all languages other than the two listed, Turkish and Kurdish, are ignored from the outset. The last category, "Other languages, unknown", constitutes the residual category for everything that could not be assigned to one of the preceding categories. Both the absence and the presence of a residual category mean that certain languages are treated as less relevant or irrelevant in comparison to those that are named. The very use of 'etc.', according to Butler (1990), suggests that processes of naming based on social categories cannot be finalized.

What is to be done with 'old' and 'new' languages?

The hopelessness of the attempt to assign languages to overarching categories becomes apparent not only on the continental macro-level, but also in the details, e.g. when it comes to distinguishing the languages of the six ethnic groups currently recognized by the Austrian constitution from the other non-German languages. Let us take Slovenian as an example: although this is also a language of the former Yugoslavia, Slovenian is not listed under this heading, but among the languages of the ethnic groups. The matter is further complicated by the fact that the term *"Windisch"* is also used for the Slovenian language spoken in Austria. In the 20th century, this term was used mainly to construct and ideologically justify a political differentiation within the Slovenian-speaking minority in Austria: between those who considered themselves part of the ethnic group, and those who did not wish to identify with it (Priestly 1997). The population statistics by 'spoken language' deal with this by assigning people who have stated that they speak *Windisch* to the category of "Languages of the recognized Austrian ethnic groups", but not to the category "Slovenian". They constitute a kind of mute, non-designated residual category.

What should or should not be registered and counted as 'a language', and how a language is to be classified once it has been recognized as relevant for the census, is dependent on changing political parameters and discourses. This can best be detected if the time axis is also taken into account, as is the case in another table published by Statistik Austria that lists the population with Austrian citizenship by 'spoken language' (Statistik Austria 2007b). Here we can follow how 'new' languages appear in the statistical field of vision, while 'old' ones disappear from view. Table 2 shows (without including the relevant figures) how significantly the categories of languages displayed – or their designations – have changed within just four decades.

Table 2 (Statistik Austria 2007b, translated by the author):

1971	1981	1991	2001
German only	German only	German only	German only
Other languages	Other languages	Other languages	Other languages
- Croatian	- Croatian	- Croatian	- Burgenland Croatian
			- Romany
- Slovenian	- Slovenian	- Slovenian	- Slovenian
- Czech	- Czech	- Czech	- Czech
- Hungarian	- Hungarian	- Hungarian	- Hungarian
	- Slovakian	- Slovakian	- Slovakian
			- Croatian
	- Serbo-Croatian	- Serbo-Croatian	- Serbian, Bosnian, Macedonian
	- Turkish	- Turkish	- Turkish
- Other (incl. unknown)	- Other (incl. unknown)		- Other (incl. unknown)

All four columns share the basic division into "German only" as the unmarked norm, and "Other languages", which are set apart from it. In 1971, the information in the category "Other languages" refers solely to the four languages of the groups recognized as autochthonous minorities at the time: Croatian, Slovenian, Czech, Hungarian. All the other languages are included under "Other (incl. unknown)". In 1981, Slovakian is added as the fifth language of autochthonous minorities, and at the same time the list is expanded to include Serbo-Croatian and Turkish, in order to encompass languages associated with labour migration. The term "Croatian", which refers solely to the ethnic group of the Burgenland Croats, now stands in opposition to the term "Serbo-Croatian", designating the language of the migrants.

In 2001, new breaks and shifts become visible: after the recognition of the Austrian Roma as the sixth ethnic group in 1993, Romany is introduced as a new category. What turns out to be particularly complex, however, is the naming and categorization of languages from the region of the former Yugoslavia.

With the outbreak of war, the former Yugoslavia broke up into independent nation states, which made new regulations about state languages to assert their differences from each other. The language Serbo-Croatian ceased to exist on a political level: the former common language was replaced by Croatian as the official national language for Croatia, Serbian for Serbia, Bosnian, Croatian and Serbian for Bosnia, and – somewhat later – Montenegrin for Montenegro (Busch 2010). In the Statistik Austria table, the previous name "Serbo-Croatian" is no longer used from 2001 onwards, although a number of speakers of this language continue to use this term to avoid identifying with one specific ethnicity or nationality.[4] It is not replaced, however, by the collective term "Bosnian/Croatian/Serbian", which has become internationally established; instead the languages Serbian, Bosnian and Macedonian are combined into one category, while Croatian constitutes a category of its own. Macedonian first appears in 2001, although it was already recognized in the former Yugoslavia as a separate official language of the Yugoslav Republic of

---

4   The language survey on the level of Austrian schools therefore continues to allow the designation 'Serbo-Croatian'.

Macedonia, distinct from Serbo-Croatian. The category "Croatian", newly created in connection with the languages of the former Yugoslavia, is now distinguished from the language previously referred to as "Croatian". The latter, the language of the minority group in Austria, goes under the new name of "Burgenland Croatian" in 2001. We can only speculate about why Croatian has been extracted from the collective category "Serbo-Croatian", which had existed until 1991, and is now listed and counted as a separate category. It seems likely that this is not a matter of relative sizes, but of political interests, as reflected in the support expressed by leading Austrian politicians for Croatian national independence.

What is a language, what is not a language?

The emergence and disappearance, the naming and renaming of languages in the ten-year rhythm of the Austrian census creates an impression of clumsiness, arbitrariness or randomness, but ultimately reflects a sincere effort to update population statistics, while somehow taking into account the changed political and demographic parameters. What becomes clear here is that languages are not natural, pre-existing categories, because the criteria that stipulate what is and what is not a language not cannot be 'objective' or universally valid, but are subject to constant political and ideological discussion and negotiation. Jacques Derrida (1998: 30) explains the non-countability of languages as follows: "What I meant to suggest is that it is impossible to count languages. There is no calculability, since the One of a language, which escapes all arithmetic (ac)countability, is never determined." Because every language is internally heterogeneous and externally permeable, the question of what is conceived and recognized as 'a language' inevitably remains controversial; it is the result of struggles for standardization and recognition, and can at best be answered on a temporary basis, with regard to a given historical period. Mikhail Bakhtin (1981: 270) draws the conclusion that: "A unitary language is not something given [dan], but is always in essence posited [zadan] – and at every moment of its linguistic life it is opposed to the realities of heteroglossia." The idea behind every form of categorization of languages is, ultimately, that languages are objects that can be systematically registered and catalogued according to certain criteria, and classified by means of a universally valid and ahistorical system.

In current sociolinguistic research, there is a broad consensus that the idea of languages as objects that can be distinguished from one another is an ideological construct, which has close historical links with the development of nation states, and also – conversely – with the recognition of minority languages (e.g. Gal 2001, Jaffe 2004, Blommaert 2006, Wright 2007). However, essentializing ideas about language continue to hold a firm place, as shown by Jaffe (2008), not least in discourses about language endangerment, linguistic ecology or language rights (see also Duchêne & Heller 2008). It is particularly in contexts such as these that an important role is attributed to the counting of languages and speakers. On an international level, a leading position is undoubtedly occupied by the online database Ethnologue, which states that it "contains information on 7,106 known living languages".[5] As is well known, Ethnologue is a satellite organization of the Summer Institute of Linguistics (SIL), which defines itself as a "faith-based nonprofit organization serving language communities worldwide", with its headquarters in Dallas (US). The organization's website [6] also states:

> SIL's service with ethnolinguistic minority communities is motivated by the belief that all
> people are created in the image of God, and that languages and cultures are part of the

---

5   http://www.ethnologue.com. Accessed 27 November 2014.
6   http://www.sil.org. Accessed 27 November 2014.

richness of God's creation. SIL's service is founded on the principle that communities should be able to pursue their social, cultural, political, economic and spiritual goals without sacrificing their ethnolinguistic identity.

One problem with this approach is that it is specialists called on from outside who define what is understood as a "God-given ethnolinguistic identity" (a term used in an earlier version of the website, cf. Busch 2013: 109), and what distinguishes one "ethnolinguistic identity" from others.

Pierre Bourdieu (1991: 221) used the example of regionalist movements to explore how struggles over ethnic identity and related classifications not only form new ways of looking at social reality, but ultimately produce what they purport to be describing or naming:

> Struggles over ethnic or regional identity – in other words, over the properties (stigmata and emblems) linked with the *origin* through the *place* of origin and its associated durable marks, such as accent – are a particular case of the different struggles over classifications, struggles over the monopoly of power to make people see and believe, to get them to know and to recognize, to impose the legitimate definition of the divisions of the social world, and thereby, to *make and unmake groups*.

According to Bourdieu, origin or particular marks of origin such as language or accent cannot be regarded as 'objective' criteria for group membership. Instead they provide an arsenal in the conflict over power and the distribution of power, an arsenal that only gradually creates new social realities – through the repeated invocation and interlinking of such marks:

> What is at stake here is the power of imposing a vision of the social world through principles of di-vision which, when they are imposed on a whole group, establish meaning and a consensus about meaning, and in particular about the identity and unity of the group which creates the reality of the unity and the identity of the group. (ibid.)

The manner in which the Austrian statistics discussed above deal with the languages of the former Yugoslavia suggests that battles over classifications are not only fought *in situ*, within the states or regions involved, but also have a secondary, retroactive effect on the paradigms that determine how these processes are perceived and interpreted from outside.

How speakers are assigned to the language categories

Once language categories have been established, the next question is how the persons counted are assigned to the individual categories. What interests us here is the mechanism used to 'filter' the data gathered in the census, in such a way that the allocation to categories seems unambiguous.

In the case of the Austrian population statistics by 'spoken language', every person counted only ever appears in one language category. In the census questionnaire, however, respondents were explicitly given the option of naming more than one 'spoken language', in line with international standards. So how is this multilingualism of individual speakers made to disappear in the statistics? The operation of 'disambiguation' involves two steps: firstly, people who named other 'spoken languages' in addition to German (as the footnote in Table 1 makes clear) were assigned to the corresponding "non-German 'spoken languages'". Secondly, a special procedure was used to assign those people who named more than one "non-German 'spoken language'" to a single "non-German" category.[7] The respondents themselves could not decide which category this should be, however.

---

7   The above-mentioned meta-information on the census (Statistik Austria 2005: 8) states: "In the case of two or more non-German languages, only one response was analysed (here a table of rankings was used)."

Instead this decision was made on the basis of a hierarchical "table of rankings" (*Rangordungstabelle*) (Statistik Austria 2007c: 209), in which 'spoken languages' are numbered consecutively from 1 for "German" to 54 for "world languages, other" (Table 3).[8] If more than one "non-German 'spoken language'" was named, only one is ultimately displayed in the statistics, namely the one that appears higher up in the table.

Table 3: Table of rankings

1 German	29 Portuguese
2 Burgenland Croatian	30 Rumanian
3 Romany	31 Russian, Ukrainian, Belarusian
4 Slovakian	32 Swedish
5 Slovenian	33 Spanish
6 Czech	34 European languages, other
7 Hungarian	35 Arabic
8 Windisch	36 Swahili
9 Bosnian	37 West African tribal languages
10 Croatian	[*Westafrikan. Eingeborenensprachen*]
11 Macedonian	38 African languages, other
12 Serbian	39 Chinese
13 Turkish	40 Hebrew
14 Kurdish	41 Indic [*Indisch*]
15 English	42 Indonesian
16 French	43 Japanese
17 Italian	44 Korean
18 Albanian	45 Persian
19 Bulgarian	46 Filipino
20 Danish	47 Punjabi
21 Finnish	48 Singhalese
22 Greek	49 Tamil
23 Dutch	50 Thai
24 Icelandic	51 Vietnamese
25 Ladin, Romansh	52 Asian languages, other
26 Maltese	53 languages of the (American) Indians
27 Norwegian	[*Indianersprachen*]
28 Polish	54 World languages, other

Thus in cases where respondents state that they speak more than one 'spoken language' among family and friends, this is reduced to an unambiguous classification for the sake of statistical clarity. The process used to achieve this is based on different ideas, rooted in language ideology. On the one hand, this is an ideology of monolingualism: monolingualism is assumed to be the norm, while bilingualism or multilingualism are regarded as deviations. The allocation of respondents to either one category or the other means that real people, with their multilingual, heteroglossic linguistic repertoires, are statistically 'monolingualized' to fit a nation-state doctrine.

On the other hand, this procedure results in a hierarchization of languages, influenced by linguistic ideology, a hierarchization which makes 'less important' languages disappear statistically in favour of 'more important' languages. This occurs as follows: if, for example, someone names Turkish and Kurdish as 'spoken languages' (whether or not this is in addition to German), Kurdish, ranked

---

8   In addition to this, each language is listed again in combination with German.

number 14, is deleted in favour of Turkish, ranked 13th. This is the only possible explanation for the fact that the statistics based on the 2001 census show 183,445 inhabitants with Turkish as their 'spoken language', but only 2,133 with Kurdish (Statistik Austria 2007a). This is in contrast to an estimated 80,000 to 120,000 people of Kurdish descent living in Austria [9] – though obviously not all of them speak Kurdish.

Also the example of African languages gives an idea of the kind of misinterpretations that can result from this procedure – the reduction of multiple responses about 'spoken language' so that only one of these is displayed. The statistics for the 2001 population by country of birth (Statistik Austria 2007a) show 24,480 people who were born in African countries. Somewhat more than half of these fall into the category of "North Africa", while 11,480 are counted in the category "Rest of Africa". In contrast, the population statistics by 'spoken language' for the same point in time cite only 1,816 people under "Other African languages" (meaning languages other than Arabic). So what has happened to the 9,964 people who were born in sub-Saharan Africa, the majority of whom can be assumed to use one or more African languages in their everyday life? The answer can again be found in the table of rankings: the former colonial languages, English, French and Portuguese, which continue to serve as official languages and often as languages of education in most countries of sub-Saharan Africa, are ranked as number 15 (English), 16 (French), and 29 (Portuguese), above the three categories (apart from Arabic) under which the languages of Africa are subsumed: "Swahili" (no. 36), "West African tribal languages" (no. 37), and "African languages, other" (no. 38). So, for example, a person from Mali who named Bambara, French and German as 'spoken languages' in the census would fall into the category "French" in the statistics.

The table of rankings, according to which languages are hierarchized and discarded, makes it clear how individual languages (independent of the number of speakers living in Austria) are positioned on an axis between 'near' and 'far away, 'familiar' and 'strange', 'significant' and 'insignificant' – always in relation to an imaginary 'we', which forms the centre of this concentric arrangement.

This reduction in complexity undertaken for statistical purposes corresponds to what Irvine and Gal (2000) describe as a language ideological mechanism of erasure: any form of real, everyday multilingualism that goes beyond bilingualism in conjunction with German is erased from public perception. The same process of erasure happens to speakers of those languages that are largely obscured in the statistics, regardless of their numbers, because – according to a key which is virtually inaccessible to normal users of the statistics – they have been classified as 'less important' than others. Hegemonic languages are statistically 'magnified', non-dominant ones are 'minimized'. These are the languages that are not counted, because they count for less.

Particularly striking, in this context, are the pejorative terms used in the table of ranking, such as *Eingeborenensprachen* which roughly corresponds to the English term 'tribal languages', or *Indianersprachen* (referring to indigeneous languages in the Americas), which suggest that these are 'idioms' rather than fully-fledged languages. Later on I will discuss how much the categorization of languages and speakers has, from a historical viewpoint, been shaped by colonial and racially tinged world views.

Between recognition and misrecognition: The categorized subject

Surveying and counting in the framework of the census are acts imposed by the state, which usually cannot be evaded. By counting the people living on the state's territory, the governing power takes

---

9   http://medienservicestelle.at/migration_bewegt/2013/03/19/kurdinnen-in-osterreich-durch-vielfalt-gepragt/. Accessed 27 June 2014.

notice of them as subjects – acknowledging their existence, as it were. At the same time, it subjects them to a logic of its own, dividing them into categories, making them manageable. Recognition and subjection are, according to Judith Butler (1997a, 1997b), inseparably linked: they describe the dual nature of subjectification, i.e. of that process by which we are formed as subjects and at the same time (as Butler states, with reference to Foucault) subjected to the productive power of discourses. Productive, because we only become subjects, and capable of social action, through subjection by, and self-subjection, to the power of discourses.

Not being 'recognized' basically means being socially non-existent – like the *sans papiers*, the so-called 'illegals', people with no clear status or residency. Hannah Arendt's statement (1952: 302) in relation to the stateless persons at the end of the Second World War applies to such people:

> The paradox involved in the loss of human rights is that such loss coincides with the instant when a person becomes a human being in general without a profession, without a citizenship, without an opinion, without a deed by which to identify and specify himself – and different in general, representing nothing but his own absolutely unique individuality, which, deprived of expression within and action upon a common world, loses all significance.

A person who is denied the quality of a subject is deprived, as it were, of a foundation, of a possible position from which she or he can act legitimately, can have an effect on the "common world". In this respect, censuses have an inherently ambivalent character. On the one hand, they are a form of recognition (albeit a minimal one): being noticed and counted can be taken as the basis for ascertaining needs, predicting developments, or justifying demands – for example with regard to the valorization of indigenous or minority languages. On the other hand, they constitute an exhortation, addressed to all individuals, to reveal themselves to the governing power, to declare their 'identity', to align themselves with pre-existing categories.

This particular aspect of censuses has often been the target of forms of civil disobedience.[10] It is forms of resistance such as these that Foucault (1982: 208) examines, with the aim of investigating the essence of power, or rather, different forms or techniques of power:

> To sum up, the main objective of these struggles is to attack not so much "such or such" an institution of power, or group, or elite, or class but rather a technique, a form of power. This form of power applies itself to immediate everyday life which categorizes the individual, marks him by his own individuality, attaches him to his own identity, imposes a law of truth on him which he must recognize and which others have to recognize in him. It is a form of power which makes individuals subjects. There are two meanings of the word "subject": subject to someone else by control and dependence; and tied to his own identity by a conscience or self-knowledge. Both meanings suggest a form of power which subjugates and makes subject to.

Unlike questions about such things as people's commute to work, or the kind of housing they live in, the question asked in censuses about language is of the same order as questions used to enquire about 'identity markers' such as age, gender, origin, citizenship, and in some cases religion, ethnicity or 'race'.[11] This kind of question is about who one is. Categories of linguistic affiliation gain much of their subject-constituting power from their multiple discursive interconnections, the intersectional interplay with other such 'identity categories'.

---

10 In Austria, for example, in the 1981 census, many residents of Vienna named Slovenian as their 'spoken language' out of solidarity with Carinthian Slovenes.

11 Questions about racial affiliation are still asked in the census in countries including the US and South Africa.

Each person counted in the census is exhorted to declare his or her 'identity', as defined by a number of key data. This exhortation can be understood, in the terms of Louis Althusser, as an ideological 'interpellation', a concept also used by Judith Butler in her theory of subjectification. To illustrate the mechanism of ideological interpellation, Althusser uses the well-known metaphor of a policeman interpellating a passer-by with the words 'Hey, you there!' (Althusser 1971: 17), at which the latter immediately stops, because he knows that he is the one being addressed. By subjecting himself to the authoritative exhortation and identifying himself 'of his own accord', he is recognized, in a Hegelian sense, by (state) ideology.

In the case under discussion here, the individuals subjected to the census have to state – usually under the controlling gaze of a census-taker – what language or languages they identify with. The language ('your' language, which you shall not deny) slips into the role of the interpellating ideology, with which one has to identify, and to which one must give allegiance: every language 'has' its speakers. To continue to use Althusser's terminology, language becomes that imaginary, absolute subject of ideology (which Althusser spells with a capital S), in the name of which the rituals of mutual recognition are practised. These rituals serve to assure us of our own identity: "the mutual recognition of subjects and Subject, the subjects' recognition of each other, and finally the subject's recognition of himself" (Althusser 1971: 21-22).

That every ideology, every ideological interpellation, includes not just the function of recognition but also that of "misrecognition – méconnaissance" (Althusser 1971: 16), is nowhere more clear than in the case of the mechanisms discussed above, which serve to reduce linguistic complexity for statistical purposes. This occurs, firstly, when respondents are forced to simplify their heteroglossic practices and repertoires (Busch 2012) into a small number of clear answers, which are then – by means of a kind of magic formula – ultimately reduced to a single language appearing in the statistics. Secondly, this process largely filters out those languages that are accorded less prestige or importance. In other words, individuals are classified in a different way to what they might have wished, without their knowledge and without any action on their part.

Why count languages? Looking back

Historically, the question of language use or of language affiliation in censuses is not 'innocent'. It bears the ideological burden of two questions that dominated the 19th century: nationality and the nation state on the one hand, colonialism on the other. As Dominique Arel (2002) explains, the question of whether national/ethnic identity should be surveyed was first discussed more widely at the International Statistical Congress in Vienna in 1857. A direct request for self-designations was rejected, with the reasoning that many of the respondents were not accustomed to thinking in national categories, and that subjectively tinged, unreliable responses were therefore to be expected. In contrast, questions about language were identified as the most reliable, 'objective' markers of affiliation. This led to a recommendation that a question about 'spoken language' ("   ", in the original French) be included in the census (Arel 2002: 95). Based on the assumption that every person had a single dominant language, hybrid categories were avoided, and people who named two or more languages were classified as monolingual (ibid.: 98).

Kertzer and Arel (2002) show how the need to categorize and count the population is closely linked with the emergence of the modern state, especially the colonial state. With reference to Benedict Anderson (1983), they speak of a totalizing and classificatory grid that was spread over the colonies, in order to take possession of everything that was to be found within an enclosed territory. In the territories that later became colonies, as Peter Uvin (2002) shows with the example of Burundi and Rwanda, the counting and classifying began with the arrival of the first adventurers, missionaries and ethnographers, and intensified with actual colonization. Soldiers and policemen were used to

carry out the comprehensive population surveys that began at this point, in order to raise taxes to finance the colonial system. In this system the census categories were not – as was already the custom in Europe – negotiated with key social actors, but were constituted from a European perspective, forcibly imposed, and charged with racial ideologies. As Uvin demonstrates, the colonial project of dividing up the population into essentialized groups – into Hutu, Tutsi and Twa, in the case of Rwanda and Burundi – simplified and consolidated what had previously been more complex, social embedded differentiations. The process of naming and counting led to the creation of ethnic identities and categories, but reduced or prevented social permeability and mobility. In his book, *Fear of Small Numbers,* Arjun Appadurai (2006) cautions against the danger of ethnicization. He identifies systems of counting and comparing ethnic, linguistic or religious groups, inherited from colonial history, as one of the main factors fuelling mass pogroms, civil wars and terrorism in India as well as in other regions of the world.

The question of ethnic or national affiliation took on particular importance in the 19th century in the eastern part of Europe, especially in the Russian empire and the Austro-Hungarian monarchy, where the dominant ideal was not that of a (homogeneous) nation state, as in the West, but the idea of a multi-ethnic state. In the non-Hungarian part of the Habsburg monarchy, the Constitutional Law (Staatsgrundgesetz) of 1867 ensured the fundamental right to the "preservation and care of nationality and language". Language was postulated as being constitutive of nationality, which meant that the question of language suddenly assumed a central position in the struggles over the political distribution of power. The decisive question from then on was what 'nation' could lay claim to how many speakers, and where (Wolf 2012). Even after the collapse of the Habsburg monarchy, questions of linguistic affiliation – with varying wording – continued to be asked in Austria (Brousek 1980). In the 1923 census "Zigeuner" ("gypsies") were the only group who were required to declare their ethnicity as well as language. The *Zentralevidenz* (central registry) established in the following year was later used by the National Socialist regime as the basis for the systematic persecution and murder of the Austrian Roma. In 1939, after Austria was annexed by the German empire, respondents were obliged under threat of punishment to truthfully answer questions about their ethnicity (*Volkszugehörigkeit*), and, where relevant, about the Jewish ancestry of their grandparents. We all know what dramatic consequences this had for those concerned.

Today, censuses can contribute to solidifying ethnic-national categories, and to pressure people into aligning themselves with these categories. An example showing the extent of this influence is Bosnia-Herzegovina, where the 2013 census asked about nationality, faith and language.[12] Since the common Serbo-Croatian or Croato-Serbian language has been divided up into the languages Bosnian, Croatian and Serbian, all three of which serve as official languages in Bosnia-Herzegovina (Busch 2010), and since religious belief is largely equated with ethno-national affiliation, all three questions can be considered as questions of identity. The last census took place in 1991, i.e. immediately before the outbreak of war. These census results were taken as the basis for the system of proportionality established in the Dayton Agreement of 1995, which stipulated that political functions and administrative offices were to be distributed according to national affiliation (Bosnian, Croatian, Serbian). Although the ratios are not meant to be made dependent on the results of the new census, there were calls from all sides to identify with the relevant nationality. Among Bosniak nationalists in particular, fears were expressed that many of the respondents might prefer the 'neutral' designation "Bosnian" (referring to citizenship) to the ethnic identification as "Bosniaks". Because of their sensitive character the census results relating to linguistic or ethnic affiliation have not yet been published.

As in Bosnia-Herzegovina, the 2011 census in Croatia also enquired about nationality [*narodnost*], mother tongue [*materinski jezik*] and faith [*vjera*]. For each of these questions, one possible answer

---

12 Article in the Austrian newspaper *Der Standard*, 27 September 2013.

was provided: "Croat", "Croatian", "Catholic". As an alternative, respondents could in each case tick "other" [*drugi/druga*] or "no declaration" [*ne izjašnjava se*]. An empty space was available to specify the "other".[13]

The last thing to be noted in this context is that the above-mentioned UN recommendations (United Nations 2006) continue to treat language – alongside religion, specific customs and ways of life – as a marker of ethnocultural affiliation.

Conclusion

Let us return to the classification of the animal world quoted at the beginning of this paper. The taxonomy quoted from Borges has an amusing but at the same time unsettling effect. Why is this the case? Probably because this taxonomy is not only opposed to our habitual way of thinking, but simultaneously carries it to extremes. A way of thinking that is shaped by systems of classification such as those going back to Linné, in which the animal kingdom is divided into classes, orders, genera, species and varieties. The allegedly Chinese taxonomy makes us think about the fact that every order is contingent, in other words that another order is always possible; that every order tends to lay claim to universality, i.e. to establishing itself as universally valid and ahistorical; that such order shapes the way we perceive the world, because we can only perceive what the order makes perceivable; that every order must remain inconsistent, because there are always residual categories that resist allocation; that the way categories are created and related to each other not only reflects social power relations, but also helps to establish them as self-evident facts, thus exerting a performative power; that every category is defined by what is excluded from it, and thus nurtures the fiction that the things subsumed into a category are internally homogeneous; and finally that there can never be a stable relationship between content and container, between signified and signifier, between world and discourse.

References

Althusser, L. (1971). Ideology and Ideological State Apparatuses (Notes towards an Investigation). In L. Althusser, *Lenin and Philosophy, and Other Essays*. London: New Left Books, 127–188.
Anderson, B. (1983). *Imagined communities: Reflections on the origins and spread of nationalism*. London: Verso.
Appadurai, A. (2006). *Fear of Small Numbers. An Essay on the Geography of Anger*. Durham: Duke University Press.
Arel, D. (2002). Language categories in censuses: backward- or forward-looking? In D. Kertzer & Dominique Arel (eds.), *Census and identity. The politics of race, ethnicity, and languages in national censuses*. Cambridge: Cambridge University Press, 92–121.
Arendt, H. (1952). *The origins of totalitarianism*. New York: Harcourt Brace.
Bakhtin, M. (1981). Discourse in the Novel. In M. Bakhtin, *The dialogic imagination* (ed. by Michael Holquist). Austin: University of Texas Press, 259–422.
Blommaert, J. (2006). Language Ideology. In *Handbook of Language and Linguistics* (vol. 6). Oxford: Elsevier, 510–523.
Borges, J. L. (1964). The Analytical Language of John Wilkin. In J. L. Borges, *Other Inquisitions (1937–1952)*. Translated by Ruth L. C. Simms. Austin, TX: University of Texas Press, 101–

---

13 http://www.dzs.hr/Hrv/censuses/census2011/forms/P1-WEB.PDF. Accessed 27 June 2014.

105.

Bourdieu, P. (1991). *Language and Symbolic Power*. Oxford: Polity Press.

Brousek, K. M. (1980). *Wien und seine Tschechen: Integration und Assimilation einer Minderheit im 20. Jahrhundert*. Munich: Oldenburg.

Busch, B. (2010). New National Languages in Eastern Europe. In N. Coupland (ed.), *Language and Globalization*. Malden, Oxford: Blackwell, 182–200.

Busch, B. (2012). The Linguistic Repertoire Revisited. *Applied Linguistics* 33(5), 503–523.

Butler, J. (1990). *Gender trouble: Feminism and the Subversion of Identity*. New York: Routledge.

Butler, J. (1997). *Excitable Speech. A Politics of the Performative*. New York: Routledge.

Butler, J. (1997). *The Psychic Life of Power. Theories in Subjection*. Stanford: Stanford University Press.

Derrida, J. (1967). *De la grammatologie*. Paris: Les Éditions de Minuit.

Derrida, J. (1998). *Monolingualism of the Other or The Prosthesis of Origin*. Stanford: Stanford University Press.

Duchêne, A. & Heller, M., eds. (2008). *Discourses of Endangerment. Ideology and Interest in the Defence of Languages*. London, New York: Continuum.

Foucault, M. (1982) The Subject and Power. In M. Foucault, *Beyond Structuralism and Hermeneutics*. Ed. by H. Dreyfus and P. Rabinow. Chicago: The University of Chicago Press, 208–226.

Foucault, M. (1994). *The Order of Things. An Archaeology of the Human Sciences*. New York: Vintage Books.

Gal, S. (2001). Linguistic Theories and National Images in Nineteenth-Century Hungary. In S. Gal & K. Woolard (eds.), *Languages and publics. The making of authority*. Manchester: St. Jerome Publishing, 30–46.

Heintze, H.-J.(2005). Article 3. In M. Weller (ed.), *The rights of minorities. A commentary on the European Framework Convention for the Protection of National Minorities*. Oxford: Oxford University Press, 107–137.

Irvine, J.; Gal, S. (2000). Language ideology and linguistic differentiation. In P. Kroskrity (ed.), *Regimes of language: ideologies, polities and identities*. Santa Fe: School of American Reserach Press, 35–83.

Jaffe, A. (2008). Discourses of endangerment: Contexts and consequences of essentializing discourses. In A. Duchêne & M. Heller (eds.), *Discourses of Endangerment. Ideology and Interest in the Defence of Languages*. London, New York: Continuum, 57–75.

Jakobson, R. (1960). Linguistics and Poetics. In T. A. Sebeok (ed.), *Style in Language*. New York: Wiley, 350–377.

Jakobson, R. (1990). The Concept of Mark. In R. Jakobson, *On Language* (ed. by Linda R. Waugh and Monique Montville-Burston). Cambridge: Harvard University Press, 134–140.

Kertzer, D. & Arel, D. (2002). Census, identity formation, and the struggle for political power. In: D. Kertzer & D. Arel (eds.), *Census and Identity: The politics of race, ethnicity, and language in national censuses*. Cambridge: Cambridge University Press, 1–42.

Priestly, T. (1997). On the development of the Windischentheorie. *International Journal of the Sociology of Language* 124, 95–98.

Schmidt, R. (2008). Defending English in an English-dominant world: The ideology of the 'Official English' movement in the United States. In A. Duchêne & Monica Heller (eds.), *Discourses of Endangerment. Ideology and Interest in the Defence of Languages*. London, New York: Continuum, 197–215.

Statistics Canada (2011). Linguistic Characteristics of Canadians. http://www12.statcan.ca/census-recensement/2011/as-sa/98-314-x/98-314-x2011001-eng.cfm. Accessed 10 November 2014.

Statistik Austria (2005). Standard-Dokumentation. Metainformationen (Definitionen, Erläuterungen, Methoden, Qualität) zur Volkszählung 2001. Vienna: Bundesanstalt Statistik Austria.

Statistik Austria (2007a). Bevölkerung 2001 nach Umgangssprache, Staatsangehörigkeit und Geburtsland.

http://www.statistik.at/web_de/statistiken/bevoelkerung/volkszaehlungen_registerzaehlunge
n/bevoelkerung_nach_demographischen_merkmalen/022896.html. Accessed 16 June 2014.

Statistik Austria (2007b). Bevölkerung mit österreichischer Staatsbürgerschaft nach
Umgangssprache seit 1971.
http://www.statistik.at/web_de/statistiken/bevoelkerung/volkszaehlungen_registerzaehlunge
n/bevoelkerung_nach_demographischen_merkmalen/022886.html. Accessed 18 June 2014.

Statistik Austria (2007c). Volkszählung 2001: Benutzerhandbuch.
http://www.statistik.at/web_de/statistiken/bevoelkerung/volkszaehlungen_registerzaehlunge
n/index.html. Accessed 25 June 2014.

United Nations (2006). Conference of European Statisticians' recommendations for the 2010
Census of Population and Housing prepared in cooperation with the Statistical Office of the
European Communities (EUROSTAT). New York, Geneva: United Nations.

United States Census Bureau (2011). Language Use.
https://www.census.gov/hhes/socdemo/language/ Accessed 10 November 2014.

Uvin, P. (2002). On counting, categorizing, and violence in Burundi and Rwanda. In D. Kertzer &
D. Arel (eds.), *Census and identity. The politics of race, ethnicity, and languages in national
censuses*. Cambridge: Cambridge University Press, 148–176.

Wolf, M. (2012). *Die vielsprachige Seele Kakaniens. Übersetzen und Dolmetschen in der
Habsburgermonarchie 1848–1918*. Vienna: Böhlau.

Wright, S. (2007). The right to speak one's own language: Reflexions on theory and practice.
Language policy 6, 203–224.