

Categorizing Languages and Speakers: Processes of Erasure in Data Treatment and Presentation

Catégoriser les langues et les locuteurs :
les procédés d'effacement dans le traitement
et la présentation des données

BRIGITTA BUSCH

*Universität Wien (University of Vienna) and Stellenbosch University,
brigitta.busch@univie.ac.at*

ABSTRACT ■ This article examines how census data are processed by statistical offices. It first investigates how the language categories displayed in language statistics are actually elaborated. It then considers the mechanisms by which speakers are assigned to specific language categories. It addresses the language ideologies underlying both these processes and the ways in which a certain view of the world is thus reproduced 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 (2015).

RÉSUMÉ ■ Cet article propose une réflexion sur les processus adoptés par des agences censitaires dans la génération de données. Dans un premier temps, nous examinons la construction des catégories utilisées dans des données censitaires. Par la suite, nous discutons des mécanismes adoptés pour attribuer des interlocuteurs à des catégories linguistiques spécifiques. L'étude critique de ces deux processus met en lumière la présence d'idéologies sous-jacentes et comment une certaine vision du monde est reproduite et renforcée par les données censitaires. Finalement, une discussion est entreprise sur la manière dont ces processus ont été développés historiquement dans des contextes de colonisation et de conflits ethno-nationaux. Dans ce texte, nous utilisons des données censitaires et statistiques autrichiennes pour approfondir et illustrer notre réflexion, mais nous élargissons notre réflexion pour examiner des procédures d'autres pays qui suivent, comme l'Autriche, les critères établis par la Conference of European Statisticians (2015).

KEYWORDS ■ Census data, statistics, language ideologies, Austria, Conference of European Statisticians.

MOTS CLÉS ■ Données censitaires, statistiques, idéologies langagières, Autriche, Conference of European Statisticians.

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 encyclopedia, states that “animals are divided into: (a) belonging to the Emperor, (b) embalmed, (c) tame, (d) suckling 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, (l) et cetera, (m) having just broken the water pitcher, (n) that from a long way off look like flies.”

I refer to this well-known passage from Foucault because it raises fundamental questions that confront anyone attempting to categorize and classify languages and speakers. Without such categories it is impossible to count, numerically survey or establish population size; however, they are not “innocent,” since classifications can gain a performative dimension by constructing what they claim to simply describe. Population censuses can be seen as moments in which, on a national level, such categorizations are reiterated, reproduced and as needed, (slightly) transformed. This article is less concerned with the process of data collection, which is amply discussed in academic literature (e.g. in a special issue of the *International Journal for the Sociology of Language*, 2018), and instead concentrates on how census data are processed by statistical offices. Taking Austria as an example where a data aggregation process is in place similar to that in use in many other European countries, I will first discuss problems that arise in establishing language categories for statistical purposes. A brief look back on developments over the last decades allows us to reflect on the dilemma that results, on the one hand, from trying to assure statistical continuity and, on the other, from taking into account changes in the status and naming of particular languages. I will then show the mechanisms by which speakers are assigned to specific language

categories, a process in which heteroglossic repertoires are reduced to a single language category. In particular, I am interested in the language ideologies that underlie these processes and how the ideological mechanism of erasure (Irvine & Gal 2000) introduces a bias towards languages considered more prestigious than others. Finally, I will sketch out more recent developments concerning language statistics in Austria where the traditional full census enumeration has been replaced by an approach combining register-based data with micro-census data.

Heteroglossic Practices: Unambiguous Categories

In 2021, Austria, as in other regions in Europe and the world, is preparing for a new round of census surveys of the population. Since the 2011 census, the EU has called upon the member states to harmonize their legislations. In spite of efforts towards standardisation, methodological heterogeneity still prevails and there are great discrepancies in data. Of the 47 Council of Europe member states, slightly more than half collect data on language. Many countries of central and eastern Europe continue to collect ethno-cultural data, whereas western European countries tend to abstain from such practices (Simon 2007). Furthermore, an increasing number of countries in Europe now rely on data derived from administrative registers to produce some or all of their population statistics, abandoning the traditional method of full enumeration. In the case of entirely register-based population census, language is usually dropped when the relevant information is not available from any register – as it is the case in most countries – and can then only be collected through micro-census surveys. In Austria, for example, until 2001 language was an area dealt with in the census; in 2006, a law was passed that regulates the new register-based data collection system. It stipulates that language and religious affiliation shall not be part of the data collected on individuals, unless “absolutely indispensable for the accomplishment of federal government’s tasks” (Registerzählungsgesetz 16. 3. 2006 § 1[3])¹. Nevertheless, the 2001 data on “spoken language” (Umgangssprache) are still part of the official presentation of the population structure on the Statistik Austria website² and are still in use in political and academic discourse. In post WW2 Germany, no questions were asked about ethnic affiliations (and therefore none about languages), in reaction to the fact that censuses and categorizations had been misused by the Nazi regime to persecute Jews, Roma and other groups. However, in 2017, the German micro-census includes a question on language use ‘at home.’ As Adler (2019) shows, the introduction of the language question can be directly linked to the so-called refugee crisis in 2015 and to political discourse that uses language

as a proxy indicator for social integration. Adler (*ibid.*: 246) cites the draft of the German Micro Census law passed in 2006 which says: “The recording of the language predominantly spoken in the household [...] is significant for the assessment of different dimensions of integration. It allows detailed analyses of the state of integration. In particular, cultural integration is closely related to the language spoken in the household” (Adler’s translation).

Adler criticizes the fact that multilingualism cannot be represented by the answers given by respondents as only one language can be chosen from a closed list. This is all the more problematic as in households where minority languages are present and where translanguaging practices are the rule rather than the exception.

As access to minority rights depends on numerical thresholds in many countries, the monitoring body for the Council of Europe’s Framework Convention for the Protection of National Minorities has dealt extensively with the question of how the right to free self-determination in language questions can be represented in statistical data. Documents produced by the Council (2012 and 2016) provide guidelines for language questionnaires for population censuses and other contexts to be determined. More specifically, the questionnaires should allow respondents to indicate more than one language and provide open lists of alternative answers, with no obligation to affiliate with a set category (Council of Europe 2012: 20). Guidelines propose that individuals be able also to identify themselves in different ways for different purposes, depending on the relevance of identification for them in a given situation (2012: 18). Multiple affiliations must not only be recorded but also adequately processed, analysed and displayed (2016: 16).

A similar stance is taken in the Recommendations for the 2020 Censuses published by the Conference of European Statisticians (2015) that treat questions related to language (mother tongue, knowledge, practice) under the heading “ethno-cultural characteristics.” The recommendations advocate that, since “ethnicity is multi-dimensional and is considered to be more of a process than a static concept,” ethnic (and therefore also linguistic) classification should be treated “as dynamic with movable boundaries” (*ibid.*: 702). The detailed explanations which follow, however, seem to contradict this to some extent. Depending on information needs, four questions are suggested (*ibid.*: 723): “(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) Usual language(s), defined as the ones 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 specific languages.” While the questions (a), (b)

and (c) 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 (*ibid.*: 724), question (d) is seen as “relevant to understand language practices and knowledge of languages, including official languages and languages learned at school. Questions will often refer to several languages and should thus allow for multiple responses” (*ibid.*: 727). Two challenges are also addressed in the Council of Europe guidelines: first, the problem of classification, whereby it is recommended that categories “be comprehensive and include wherever possible [...] separate languages to the finest level possible, regional dialects, as well as the reporting of invented and sign languages” (*ibid.*: 732); and second, the challenge of multiple responses to language questions when it comes to the display in tabulations and census output (*ibid.*: 725). These two issues are my focus in this article.

Bi- or multilingual speakers have always been a challenge for official statistics. Duchêne *et al.* (2018) discuss how different countries deal with this question: some allow more than one answer, but this is sometimes contested by communities that do not want to be represented in mixed categories. Whereas the authors’ overview mainly refers to studies regarding how methods of language related data collection allow for multiple answers, here I will examine the challenges that arise in processing and tabulating collected data.

Historically, the question of language use or of language affiliation in censuses is not innocent (for a historical overview, see Duchêne & Humbert 2018). It bears the ideological burden of two questions that dominated the 19th century: nationality and the nation state on the one hand and colonialism on the other. As Dominique Arel (2002) reminds us, 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 terms of 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” (“*langue parlée*,” in the original French) be included in the census (*ibid.*: 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. As Uvin (2002) demonstrates, giving the example of Burundi and Rwanda, the colonial project of dividing the population into essentialized groups simplified and consolidated what had previously been more complex, socially 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. Ethnic polarization that results in ethnic conflicts often begins with an obsession with and a misuse of ethno-linguistic maps, figures and numbers, as was the case before the outbreak of the war on the territory of former Yugoslavia (Busch 2010).

To avoid encouraging tendencies towards ethnic polarisation, the Council of Europe cautions against using only imposed, rigid language categories and suggests, albeit sometimes not wholeheartedly, allowing for multiple answers and providing the possibility for open ended lists that leave room for self-classification. Moreover, the Council calls on authorities to ensure that the heterogeneous composition of the population remains visible in the processing and presentation of collected data and that dichotomies between majorities and minorities are avoided. As Leeman (2018) shows in her analysis of the U.S. census, categories to which people can affiliate are not only the result of decades long struggles and negotiations but, in turn, they also impact on how people perceive themselves and others. With Foucault (1972), one could say that when it comes to identity related questions, census surveys contribute significantly to the construction of, and provision for, permitted subject positions rather than reflecting on a reality existing outside of dominant nation-building discourse.

Between Recognition and Misrecognition: The Categorized Subject

The surveying and counting of populations in the framework of census taking are acts imposed by the state and are usually compulsory. By counting the people living on the state's territory, the governing power recognizes those censused 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 subjectivation-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) are subjected to the productive power of discourse. Productive, because we only become subjects, and capable of social action, through subjection by and self-subjection, to, the power of discourse.

Not being “recognized” basically means being socially non-existent – like *sans papiers*, so-called ‘illegals,’ people with no legal status or residency. Regarding stateless persons at the end of the Second World War, Hannah Arendt (1952: 302) calls attention to the consequences of such lack of status:

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 and 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: an example of this is the valorization of indigenous or minority languages, which is often justified, and negotiated, by census data. 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.³ It is forms of resistance such as these that Foucault 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 (1982: 208).

Unlike questions about such things as people’s commute to work, or the kind of housing they live in, questions asked in censuses about language are 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.’⁴ 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 ‘identity categories.’

Each person counted in the census is asked to declare his or her “identity,” as defined by various kinds of key data. This question 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, 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 individuals are not likely to 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” (*ibid.*: 21-22).

That every ideology, every ideological interpellation, includes not just an act of recognition but also of “misrecognition” (*méconnaissance*) (*ibid.*: 16), becomes visible in mechanisms which reduce linguistic complexity for statistical purposes. This occurs, firstly, when respondents are forced to simplify their heteroglossic practices and repertoires into a small number of unambiguous answers, which are then, as I will show in the following, 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 a given context. 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.

Pierre Bourdieu used the example of regionalist movements to explore how struggles over ethnic identity and related classifications form new ways of looking at social reality and how such classifications ultimately produce what they are supposedly 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* (1991: 221).

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 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.*).

Census and demographic statistics have their share in the making and unmaking of groups by granting or denying a certain legitimacy through naming them, misnaming them or fading them out from the presentation of statistical results. I will discuss this issue below, briefly looking at how languages spoken in the territory of former Yugoslavia were represented at different moments in Austrian statistics.

Austrian Statistics

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 discourses and ideologies that appear there can certainly also be found in the practices of other states. The first ques-

tion that interests us is how, in the tabulation of census results, language categories are formed and named. The language asked for in the last Austrian censuses was the *Umgangssprache* (the ‘spoken language’)⁵, 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. As Jakobson (1990) showed for 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 unquestioned; 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 binary logics 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” reflects an idea based the 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 languages other than German being assigned to seven broad overarching categories (see Table 1): 1) languages of recognized Austrian ethnic groups (*Volksgruppen*); 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’

‘Spoken language’*	
Total	8,032,926
German only	7,115,780
Languages of recognized Austrian ethnic groups	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
- 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 [<i>Indisch</i>]	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

Source: Statistik Austria 2007a (my translation)⁶

* 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 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 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,” separate from “Other European languages,” brings together three languages whose only common quality is that they enjoy 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 superordinate category ‘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, with Mexico, the US and Colombia each having more speakers of Spanish than Spain itself.

Within each superordinate category, languages are listed in alphabetical order. In the category “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 from the Chinese encyclopaedia quoted at the beginning of this paper. The first thing that is striking is the absences. For example, Hebrew is listed under “Asian languages,” 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 to Do 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 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 recognized ethnic groups. The matter is further complicated by the fact that the term “*Windisch*” is also used by a certain number of respondents for Slovenian dialects 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 responded as *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 constellations and discourses. This becomes obvious when 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

Changes in language categories over four decades

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)

Source: Statistik Austria 2007b (my translation)

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 spoken by naturalized labour migrants. The term “Croatian,” which refers solely to autochthonous minority of the Burgenland Croats, now stands in opposition to the term “Serbo-Croatian,” designating the language of the naturalized migrants. In 2001, new shifts become visible: after the recognition of the Austrian Roma as the sixth autochthonous minority in 1993, Romany is introduced as a new category.

What turns out to be particularly complex in 2001, however, is the renaming and recategorization of languages from the region of the former Yugoslavia. With the outbreak of war, the former Yugoslavia broke up into independent nation states, provoking new regulations about state languages in order 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-Herzegovina, and, somewhat later, Montenegrin for Montenegro (Busch 2010). In the Statistik Austria table, the previous name “Serbo-Croatian” is no longer used from 2001 on, although a number of speakers of this language continue to use this term to avoid identifying with a specific ethnicity or nationality.⁷ 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 Austrian statistics in 2001, although it was already recognized in 1945 as the official language of the Yugoslav Republic of 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 in the 1990s by leading Austrian politicians for Croatian national independence.

The emergence and disappearance, the naming and renaming of languages in the ten-year rhythm of the Austrian census creates an impression of arbitrariness or randomness, but ultimately reflects a simple effort to update population statistics, while somehow taking into account 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 cannot be ‘objective’ or universally valid; rather, they are subject to constant political and ideological discussion and negotiation. Jacques Derrida (1998: 30) reminds us 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. The manner in which the Austrian statistics deal with the languages of former Yugoslavia (as discussed above) suggests that battles over classifications are not only fought *in situ*, within the states or regions directly concerned, but that they also have a secondary, retroactive effect on the paradigms that determine how these processes

are perceived and interpreted in other contexts outside the countries in question.

The idea behind every form of categorization of languages is, ultimately, that languages are “natural” and clearly defined 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, one that has close historical links with the development of nation states, and also, conversely, with the recognition of minority languages (e.g. Blommaert 2006; Gal 2001; Jaffe 2008; Wright 2007). Essentializing ideas about language continue, however, to hold a firm place, as shown by Jaffe (2008), in discourses about language endangerment or language rights. (See also Duchêne & Heller 2008).

How Speakers Are Assigned to Language Categories

Getting back to the example of the Austrian census, the next question is, how are persons who are counted assigned to established language categories? What interests us here is the mechanisms used to ‘filter’ the data gathered in the census in such a way that allocation to categories seems unambiguous.

In the case of the Austrian population statistics by ‘spoken language,’ every person counted only appears in one language category. In the census questionnaire (2001), 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. First, 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.’” Second, a special procedure was used to assign those people who named more than one “non-German ‘spoken language’” to a single “non-German” category.⁸ The respondents themselves could not decide which category this should be. Instead, this decision was made on the basis of a hierarchical “table of rankings” (*Rangordnungstabelle*) (Statistik Austria 2007c: 209), in which ‘spoken languages’ are numbered consecutively from 1 for “German” to 54 for “world languages, other” (Table 3).⁹ If more than one “non-German ‘spoken language’” was named, only one is ultimately displayed in the statistics, namely the one that appears highest 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	<i>[Westafrikan. Eingeborenensprachen]</i>
11 Macedonian	38 African languages, other
12 Serbian	39 Chinese
13 Turkish	40 Hebrew
14 Kurdish	41 Indic <i>[Indisch]</i>
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) Indian
27 Norwegian	<i>[Indianersprachen]</i>
28 Polish	54 World languages, other

Source: Statistik Austria 2007c (my translation)

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. On the one hand, there is an ideology of monolingualism in which 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 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 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 Austria¹⁰ – though obviously not all of them speak Kurdish.

The example of African languages also provides insight into the kind of misinterpretations that can result from this procedure. The statistics for the 2001 population by country of birth (*ibid.*) 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). 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 the language ideological mechanism of erasure: any form of real, everyday multilingualism that goes beyond bilingualism in conjunction with German is erased from public perception and representation. 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 element which is virtually hidden to normal users of the statistics – they have been classified as ‘less important’ than others. Hegemonic languages are statistically ‘magnified’ while non-dominant ones are ‘minimized.’ These are the languages that are not counted, because they count for less. Particularly disturbing, in this context, are the politically incorrect terms, shaped by colonial and racially tinged world views used in the table of

ranking, such as *Eingeborenensprachen* which roughly corresponds to the English term “tribal languages,” or *Indianersprachen* (referring to indigenous languages in the Americas).

Current Developments in Austria

The language data of the 2001 census discussed in this paper are still provided on the Statistik Austria website and are still cited occasionally because, since the adoption of a register-based system, no newer data are available regarding the linguistic composition of the Austrian population. New language data are collected only in the framework of educational documentation and are published annually by Statistik Austria. Personal language data collected by school authorities when children are enrolled allow for up to three languages (mother tongue 1, 2, 3) to be mentioned. The unusual terminology ‘mother tongue’ was implemented for legal purposes because only indicating mother tongues other than German entitles the student to supplementary hours of study in the mother tongue (*Muttersprachlicher Unterricht*) as well as in German as a second language. The figures are published on the Statistik Austria website¹¹ under the title “Students with non-German spoken language” (*Schülerinnen und Schüler mit nicht-deutscher Umgangssprache*). This table does not refer to particular languages but creates a collective category, negatively defined by what students do not have instead of using the less discriminatory term that was previously in use at the Ministry of Education: “students with other languages than German.” This category is opposed to the category tacitly assumed as the norm, although students with other first language than German or in addition to German are not negligible in number; in Vienna for example in the school year 2018/19 they outnumbered those classified as “German only.”

Conclusion

The challenge of how to represent individual and societal multilingualism in population statistics is a topic in academic literature as well as in recommendations by bodies such as the Council of Europe or the Conference of European Statisticians where the main focus is on how to collect data and how to frame language questions (Duchêne *et al.* 2018). This paper shows that the way in which language data are processed and presented is equally important when it comes to mirroring heteroglossic practices. As can be gleaned from the works of Althusser, Foucault and Bourdieu cited earlier, understanding censuses as an interpellation by which individuals are summoned to assign themselves to permitted social

(and recognized ethnolinguistic) positions, suggests that population statistics do not merely represent or describe a linguistic situation but also have a performative effect in the making and unmaking of social groups. As can be seen from the Austrian example, language categories appear in statistical tabulations in a hierarchical order informed by language ideologies and reflect the dilemma between assuring statistical continuity and considering the impact of changing political and geopolitical constellations on how languages are named and regrouped into superordinate categories. Hierarchization also plays a role when assigning individual speakers, as in the Austrian example, to a single language category, thus ‘monolingualizing’ persons with complex linguistic repertoires.

The taxonomy quoted from Foucault and Borges at the beginning of this paper has an amusing but at the same time, unsettling, effect. It destabilizes our habitual way of thinking shaped by systems of classification such as the one developed by the famous Swedish botanist, zoologist and physician Carl von Linné in the eighteenth century who divided the animal kingdom into classes, orders, genera, species and varieties. The alleged Chinese taxonomy reminds us that every order is contingent; in other words, that another order is always possible; that every order tends 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 one 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.

Notes

1. <https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=20004583> [accessed 31 March 2020].
2. <https://www.statistik.at> (such string “Umgangssprache”) [accessed 31 March 2020].
3. In Austria, for example, in the 1981 census, many residents of Vienna named Slovenian as their ‘spoken language’ out of solidarity with Carinthian Slovenes.
4. Questions about racial affiliation are still asked in the census in countries including the US and South Africa.
5. The term *Umgangssprache* corresponding to the French term *langue parlée* originally used internationally in the context of censuses is commonly translated as ‘spoken language’ (Arel 2002: 94). Therefore, in this text I will also make use of this English translation.

6. In this reconstruction the figures are presented as totals, not, as in the original, also divided by nationality and country of birth.
7. The language survey on the level of Austrian schools therefore continues to allow the designation 'Serbo-Croatian.'
8. 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)."
9. In addition to this, each language is listed again in combination with German.
10. http://medienservicestelle.at/migration_bewegt/2013/03/19/kurdinnen-in-osterreich-durch-vielfalt-gepragt/ [accessed 14 April 2017].
11. https://www.statistik.at/such_string_Umgangssprache%3A%20schuljahr.pdf [accessed 22 March 2020].

References

- Adler, A., 2019. "Language Discrimination in Germany: When Evaluation Influences Objective Counting," *Journal of Language and Discrimination*, vol. 3, n° 2, p. 232-253.
- 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, p. 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 & D. Arel (dir.), *Census and Identity. The Politics of Race, Ethnicity, and Languages in National Censuses*. Cambridge, Cambridge University Press, p. 92-121.
- Arendt, H., 1952. *The Origins of Totalitarianism*. New York, Harcourt Brace.
- Blommaert, J., 2006. "Language Ideology," in K. Brown (dir.), *Encyclopedia of Language & Linguistics. Volume 6*. Oxford, Elsevier, p. 510-523.
- Borges, J. L., 1964. "The Analytical Language of John Wilkin," in J. L. Borges (dir.), *Other Inquisitions (1937-1952)* (translated by Ruth L. C. Simms). Austin, University of Texas Press, p. 101-105.
- Bourdieu, P., 1991. *Language and Symbolic Power*. Oxford, Polity Press.
- Busch, B., 2010. "New National Languages in Eastern Europe," in N. Coupland (dir.), *Language and Globalization*. Malden, Oxford, Blackwell, p. 182-200.
- Butler, J., 1997a. *Excitable Speech. A Politics of the Performative*. New York, Routledge.
- Butler, J., 1997b. *The Psychic Life of Power. Theories in Subjection*. Stanford, Stanford University Press.
- Butler, J., 1990. *Gender trouble: Feminism and the Subversion of Identity*. New York, Routledge.
- Conference of European Statisticians, 2015. *Recommendations for the 2020 Censuses of Population and Housing*. United Nations Economic Commission for Europe (dir.). New York, Geneva, United Nations.

- Council of Europe, 2016. *The Framework Convention: A Key Tool to Managing Diversity through Minority Rights. Thematic Commentary No. 4*. The scope of application of the Framework Convention. Advisory Committee on the Framework Convention for the Protection of National Minorities. Strasbourg, Council of Europe.
- Council of Europe, 2012. *The Language Rights of Persons Belonging to National Minorities under the Framework Convention. Thematic Commentary No. 3*. Advisory Committee on the Framework Convention for the Protection of National Minorities. Strasbourg, Council of Europe.
- Derrida, J. 1998. *Monolingualism of the Other or The Prosthesis of Origin*. Stanford, Stanford University Press.
- Derrida, J., 1967. *De la grammatologie*. Paris, Les Éditions de Minuit.
- Duchêne, A. & M. Heller (dir.), 2008. *Discourses of Endangerment. Ideology and Interest in the Defence of Languages*. London, New York, Continuum.
- Duchêne, A. & P. N. Humbert (dir.), 2018. *Surveying Speakers and the Politics of Census. International Journal for the Sociology of Language*, Special Issue, vol. 252.
- Duchêne, A. & P. N. Humbert, 2018. "Surveying Languages: The Art of Governing Speakers with Numbers," *International Journal for the Sociology of Language*, vol. 252, p. 1-20.
- Duchêne, A., P. N. Humbert & R. Coray, 2018. "How to Ask Questions on Language? Ideological Struggles in the Making of a State Survey," *International Journal for the Sociology of Language*, vol. 252, p. 45-72.
- Foucault, M., 1994. *The Order of Things. An Archaeology of the Human Sciences*. New York, Vintage Books.
- Foucault, M., 1982. "The Subject and Power," in H. Dreyfus & P. Rabinow (dir.), *Beyond Structuralism and Hermeneutics*. Chicago, The University of Chicago Press, p. 208-226.
- Foucault, M., 1972. *The Archaeology of Knowledge and The Discourse on Language*. New York, Pantheon Books.
- Gal, S., 2001. "Linguistic Theories and National Images in Nineteenth-Century Hungary," in S. Gal & K. Woolard (dir.), *Languages and Publics. The Making of Authority*. Manchester, St. Jerome Publishing, p. 30-46.
- Irvine, J. & S. Gal, 2000. "Language Ideology and Linguistic Differentiation," in P. Kroskrity (dir.), *Regimes of Language: Ideologies, Politics and Identities*. Santa Fe, School of American Research Press, p. 35-83.
- Jaffe, A., 2008. "Discourses of Endangerment: Contexts and Consequences of Essentializing Discourses," in A. Duchêne & M. Heller (dir.), *Discourses of Endangerment. Ideology and Interest in the Defence of Languages*. London, New York, Continuum, p. 57-75.
- Jakobson, R., 1990. "The Concept of Mark," in L. R. Waugh & M. Montville-Burston (dir.), *On Language*. Cambridge, Harvard University Press, p. 134-140.
- Kertzer, D. & D. Arel, 2002. "Census, Identity Formation, and the Struggle for Political Power," in D. Kertzer & D. Arel (dir.), *Census and Identity: The politics of Race, Ethnicity, and Language in National Censuses*. Cambridge, Cambridge University Press, p. 1-42.
- Leeman, J., 2018. "It's All about English: The Interplay of Monolingual Ideologies, Language Policies and the U.S. Census Bureau's Statistics on Multilingualism," *International Journal for the Sociology of Language*, vol. 252, p. 21-43.

- Priestly, T., 1997. "On the Development of the Windischentheorie," *International Journal of the Sociology of Language*, vol. 124, p. 95-98.
- Simon, P., 2007. "Ethnic" Statistics and Data Protection in the Council of Europe Countries. Study Report. European Commission against Racism and Intolerance (ECRI). Strasbourg, Council of Europe.
- Statistik Austria, 2007a. *Bevölkerung 2001 nach Umgangssprache, Staatsangehörigkeit und Geburtsland*. http://www.statistik.at/web_de/statistiken/bevoelkerung/volkszaehlungen_registerzaehlungen/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_registerzaehlungen/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_registerzaehlungen/index.html [accessed 25 June 2014].
- Statistik Austria, 2005. *Standard-Dokumentation. Metainformationen (Definitionen, Erläuterungen, Methoden, Qualität) zur Volkszählung 2001*. Vienna, Bundesanstalt Statistik Austria.
- Uvin, P., 2002. "On Counting, Categorizing, and Violence in Burundi and Rwanda," in D. Kertzer & D. Arel (dir.), *Census and Identity: The politics of race, ethnicity, and language in national censuses*. Cambridge, Cambridge University Press, p. 148-176.
- Wright, S., 2007. "The Right to Speak One's Own Language: Reflections on Theory and Practice," *Language Policy*, vol. 6, p. 203-224.