The German Language Worldwide: Data – Course Participation in Germany (Technical Description and Codebook)[*]

Silke Uebelmesser — University of Jena and CESifo

Matthias Huber[†] — University of Jena

Severin Weingarten — University of Jena

March 26, 2018

When using this data set, please always cite:

The official data presentation article

Silke Uebelmesser, Matthias Huber, and Severin Weingarten (2018). "The German Language Worldwide: a New Data Set on Language Learning". *CESifo Economic Studies* 64 (1), 103–121.

and the data repository

Silke Uebelmesser, Matthias Huber, and Severin Weingarten (2018). "The German Language Worldwide: Data – Course Participation in Germany". *Harvard Dataverse*, V1. doi: 10.7910/DVN/V5BWMT

[*] This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, UE 124/2-1) – 270886786.

[†] Corresponding author: matthias.huber@uni-jena.de

0 Background

We provide data from the Goethe Institut (GI), a German cultural association with institutes worldwide and in Germany. Since 1965, the GI has continuously published annual reports in which activities of each institute including data about language course and exam participation are reported. These annual reports are publicly available. We digitised this information in order to construct three datasets[1]. In the following, the technical description and the codebook for the dataset about language course participation at institutes in Germany (Uebelmesser, Huber, and Weingarten 2018c) are presented. The dataset provides detailed information about the annual number of registrations in language courses in Germany by nationality. Data are available for the period 1966 to 2014. In total, course participants with around 200 different nationalities studied the German language at the institutes in Germany over the whole time period.

1 Technical description

The dataset includes observations for each nationality of which at least for one year a positive number of students participated in a language course at an institute in Germany. The dataset is balanced for the years from 1966 to 2014 and contains in total 10,143 observations for 207 countries.

1.1 Country codes

To harmonize nationalities we use the standardised 3-letter country codes by ISO 3166 alpha-3 for currently existing countries. For dissolved countries we use the 4-letter codes provided by ISO 3166-3 for the period the countries existed. In every year we report only data for countries which exist in that year according to the ISO standardizations. For nationalities with students at any German institute at least in one year, we assign the value 0 to years where the GI did not report any students. We only assign missing values

[1] See Uebelmesser, Huber, and Weingarten (2018b,d) for the technical description of the other two datasets, and Uebelmesser, Huber, and Weingarten (2018a) for a comprehensive presentation of the data.

to countries in years in which they did not officially exist. We have to deviate from this procedure in the following cases:

1. During the start of the dissolution of the Social Federal Republic of Yugoslavia in 1991, assignment of course participants' nationality to successor states (Slovenia, Croatia, Macedonia and Bosnia and Herzegovina) and the remainder of Yugoslavia is not clearly reported. Therefore, we deviate slightly from the practice of simply assigning the country names given by the GI to their respective ISO codes officially assigned in that year. In order to get a clear demarcation in the yearly data, we assign all observations for the affected countries until the end of 1991 to the ISO 3166-3 country code for Yugoslavia (YUCS). From 1992 onwards we use the country codes for the new successor states.

   For Serbia-Montenegro, we deviate slightly from the ISO standard, which continues to use YUCS for the remainder of Yugoslavia (i.e. Serbia and Montenegro) until 2003 and replaces it with CSXX for Serbia-Montenegro only then. We assign all Yugoslavian observations to the new country code CSXX from 1992 onwards. We think that this makes more sense, because the borders of Yugoslavia changed massively between 1991 and 1992, but not at all in 2003, when the country was officially renamed. Despite the fact that Serbia was only established as a country on June 5, 2006 and the country code SRB only introduced at this time, the GI reports Serbian observations as early as 1992. We assign these to the CSXX country code, too. From 2006, Serbia-Montenegro (CSXX) splits up into Serbia (SRB) and Montenegro (MNE).

2. After the dissolution of the Soviet Union the GI did not report all successor states immediately. For 1991 through 1997, there were observations for the Commonwealth of Independent States (COMIS) and for successor states, which appeared in the annual reports gradually until 1997, even though officially independent earlier. As a consequence, it is unclear whether some residents of the former Soviet countries

were reported as COMIS "nationals" or nationals of successor states. Therefore, we summarise all observations reported as COMIS and successor states in the period from 1991 to 1997 under the code COMIS. From 1998 on we use the official ISO-3166-3 country codes for the successor states.

3. In some annual reports entries are reported as "Aussiedler", in other annual reports the entry is "Bundesrepublik Deutschland (Aussiedler)". We subsume all related entries under the country code DEU. Only in the 1981 entries for "Aussiedler" and "Bundesrepublik Deutschland" were reported separately. Therefore, we add up both numbers and also assign the country code DEU.

4. The former Belgian colony Ruanda-Urundi split up in 1961 into the countries Ruanda and Burundi but the GI reported entries for "Ruanda and Burundi" in the annual reports until 1970. Since the split happened before the introduction of the ISO country code system, no code exists for "Ruanda and Burundi". We assigned the special code RUABU until 1970. From 1971, these two countries have been reported separately.

5. Before 1999, in some years Israel and Palestine were explicitly reported together, in other years it is not clear if numbers for Israel contain numbers from Palestine. Therefore, we assign all respective observations before 1999 to the self-assigned country code ISPAL. From 1999 onwards, separate numbers are given for Israel and Palestine which we assign to the respective official country codes.

6. For Guyana, French Guyana and Suriname, several variations of names of these countries appear in the dataset which do not allow a clear distinction between these countries throughout the whole period 1966 to 2014. We assume that officials at the GI were themselves unsure about the status of these countries and where course participants were really from. We therefore summarise all respective entries under the self-assigned code GUAYA.

7. The GI reports students without/ with no specified nationality which we identify

with the self-assigned country code NONAT.

1.2 Interpolation in 1986 – 1988

For the years 1986, 1987 and 1988 some interpolation was necessary because the GI did not publish the same information as in the pre- and succeeding years. In the years 1986 and 1987 five institutes (Berlin, Bonn, Frankfurt, Düsseldorf and Iserlohn) did not report the nationalities of their students. Therefore, we calculated the average share of students of each nationality in the years 1986 and 1989 and assigned the total number of students in these five institutes (5772 students in the year 1986 and 5704 in 1987) to nationalities according to those shares.

Contrary to all other years, in 1988 the GI only published a diagram that reports numbers for the 10 nationalities with the largest numbers of participants and the total number of course participants in Germany in that year. Hence, we use the following three-step procedure to interpolate the 1988 values for the remaining nationalities:

1. We calculate the 1987, 1988 and 1989 totals without the values for the nationalities that are listed in the annual report of 1988.

2. Using these totals from 1), we calculate the shares of the remaining countries (i.e. all countries except the ten countries listed in the annual report 1988) in total course participation averaged over the years 1987 and 1989.

3. We assign for each nationality this share to the 1988 total calculated in the first step.

As a consequence, the reported 1988 values for the ten countries are not changed and the sum of all 1988 values is equal to the grand total reported in the 1988 annual report.

These interpolated years may coincide with important historical events, and therefore, provide only crude estimates for some countries. In Germany, for instance, at the end of the 1980s the number of ethnic German immigrants ("Aussiedler") increased sharply , which may reflect the increased participation in language courses in the GI: in 1986 the GI reported 614 students and in 1989 the number amounted to 5,206 students. Our interpolation smooths the increase over the years 1986 to 1989 which may not capture the exact timing of the sharp increase.

2 Codebook

year	Year
Format	Numeric
Range	1966 – 2014
Missing	0/10,143

country	Country of origin
Format	Character
Comment	ISO 3166 alpha-3 for currently existing countries
	ISO3166-3 for dissolved countries
	self-assigned codes
Missing	0/10,143

registrations	Registrations at German institutes by nationality		
Format	Numeric		
Range	0 – 5,206		
	Mean	Median	Std. Dev
	121.462	18	324.983
Missing	1,332/10,143		
Comment	Missing values are due to countries that did not exist in some years.		

References

Uebelmesser, Silke, Matthias Huber, and Severin Weingarten (2018a). “The German Language Worldwide: a New Data Set on Language Learning”. *CESifo Economic Studies* 64 (1), 103–121.

— (2018b). “The German Language Worldwide: Data – Course Participation Abroad”. *Harvard Dataverse,* V1. DOI: 10.7910/DVN/XVNUY8.

— (2018c). “The German Language Worldwide: Data – Course Participation in Germany”. *Harvard Dataverse,* V1. DOI: 10.7910/DVN/V5BWMT.

— (2018d). “The German Language Worldwide: Data – Language Learning Opportunities Abroad”. *Harvard Dataverse,* V1. DOI: 10.7910/DVN/VSQBM7.