The German Language Worldwide: Data – Language Learning Opportunities Abroad (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 – Language Learning Opportunities". *Harvard Dataverse*, V1. doi: 10.7910/DVN/VSQBM7

*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, technical description and the codebook for the dataset about the presence and type of activities of the GI (Uebelmesser, Huber, and Weingarten 2018d) are presented. The dataset contains information about the presence of the GI on the city level for the period 1965 to 2014. For each city-year combination the dataset provides data about the types of institutes and their offer of language services. Furthermore, for each city, the opening and closing years of the institute are given. Over the analysed 50 years, the GI has been present in 272 cities in 109 countries.

1 Technical description

The data includes observations for each city in which an institute was open in at least one year and is balanced for the years 1965 to 2014. In the following, we describe the variables.

1.1 Geographical Data: Country codes, city names and regions

We assign the standardised 3-letter country codes by ISO 3166 alpha-3 to currently existing countries or the 4-letter codes provided by ISO 3166-3 to dissolved countries. In the dataset, in each year each city is assigned to the country to which it belongs. Therefore, some cities are assigned to different countries over the course of time, like cities in the successor states of Yugoslavia (e.g. Zagreb), the Soviet Union (e.g. Moscow) and Czechoslovakia (e.g. Prague). Furthermore, the city Dhaka, nowadays in Bangladesh, was in Pakistan until 1970. The institutes in Ramallah and Hong Kong are assigned to the country codes

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

for Palestine (PSE) and Hong Kong (HKG) in the entire period, respectively, as there is no clear transition from one country to another country (Israel and China/United Kingdom). Finally, we assign current country codes for cities, that are in colonies, which got independent after 1965, e.g. Luanda in Angola which got independent from Portugal in 1975 is assigned to AGO for the entire period 965 to 2014.

Additionally, we assign each country to one of 12 regions according to the regional organisation by the GI which has been in place since 2008 (see Table 1).

1.2 Time

The dataset is truncated from the left side, as for the years before the GI published annual reports the data is not available. Therefore, we constructed the variable "open_before_1965" that takes the value 1 if the institute was likely to be open before 1965, and 0 otherwise. This comprises the following two cases: First, from 1986 onwards, the GI reported for each institute its year of opening, but rarely consistently over the following annual reports. Therefore, the exact year of opening cannot be determined with certainty. The variable takes the value of 1 if there was information about an opening before 1965. Second, some institutes were reported from 1965 onwards, but closed before the reporting about their opening (from 1986 onwards). These institutes may have been open before 1965. The variable takes the value 1 in this case as well.

For openings after 1965, the variable "open1" indicates the first year with public activities reported in the annual reports[2] and not the (very often inconsistent) opening year reported from 1986 onwards. The variable "close1" captures the first year the institutes was not reported any more.[3] Some institutes closed and reopened again, even several times. Equivalent to the construction of "open1" and "close1", we capture this in the variables "open2", "open3", "close2" and "close3". The variable "open1" indicates the year of opening

[2]In some cases, liaison offices are reported without any public activities. We do not assign them to be closed, but the type is set to liaison office (see Section 1.3 below).

[3]In some cases the annual reports indicate closing dates which did not match reported activities in the annual reports.

Table 1: Regions according to the regional organisation by the GI since 2008

Central Eastern Europe	Northwest Europe	Southwest Europe
Czech Republic	Denmark	Belgium
Estonia	Finland	France
Hungary	Great Britain	Italy
Latvia	Iceland	Luxembourg
Lithuania	Ireland	Portugal
Poland	Netherlands	Spain
Slovakia	Norway	
Slovenia	Sweden	
Southeast Europe	**Eastern Europe and Central Asia**	**South America**
Bosnia and Herzegovina	Belarus	Argentina
Bulgaria	Georgia	Bolivia
Croatia	Kazakhstan	Brazil
Cyprus	Russian Federation	Chile
Greece	Ukraine	Colombia
Macedonia	Uzbekistan	Peru
Romania		Uruguay
Serbia		Venezuela
Turkey		
North America	**Sub-Saharan Africa**	**North Africa and Middle East**
Canada	Angola	Algeria
Costa Rica	Burkina Faso	Egypt
Cuba	Cameroon	Iraq
Mexico	Congo	Israel
United States of America	Cóte d'Ivoire	Jordan
	Ethiopia	Lebanon
	Ghana	Libya
	Kenya	Morocco
	Madagascar	Oman
	Malawi	Palestinian Territories
	Nigeria	Saudi Arabia
	Rwanda	Sudan
	Senegal	Syrian Arab Republic
	South Africa	Tunisia
	Tanzania	United Arab Emirates
	Togo	
	Uganda	
	Zimbabwe	
Southeast Asia, Australia and New Zealand	**South Asia**	**East Asia**
Australia	Afghanistan	China
Indonesia	Bangladesh	Hong Kong
Malaysia	India	Japan
Myanmar	Iran	Mongolia
New Zealand	Nepal	Republic of Korea
Philippines	Pakistan	Taiwan
Singapore	Sri Lanka	
Thailand		
Viet Nam		

for every institute in the dataset, the subsequent variables “close1”, “open2”, etc, only take non-missing values if applicable.

Note that in some cases institutes close, i.e. institutes are not reported, and reopen after a very short period. Though, in these cases the data reflects the information of the annual reports, it is unlikely that institutes close for one or two years and then reopen again. This might be due to misreporting in the annual reports. Nevertheless, we stick to the annual reports and report them as closed for this short period.

1.3 Types: Main institutes, subsidiaries and liaison offices

For each city-year combination we indicate the status of the GI in the variable “status_institute”. The status is either closed, or if open, it takes one of three different types. The main type is an institute. Furthermore, it can be a subsidiary (“Nebenstelle” or “Außenstelle”), which is linked to a main institute or a liaison office (“Verbindungsbüro”), which is not linked to a main institute.

From 1971 onwards, the GI reported the names of the subsidiaries for each main institute with subsidiaries.[4] This was either given in parentheses after the main institute’s name or was explicitly mentioned with the subsidiary as "subsidiary of" depending on the annual report. In parentheses after the main institute’s name or explicitly indicated as “subsidiary of” when reported separately from the main institute depending on the annual report. For each main institute, we report the city names of its subsidiaries and vice versa in the variables “subsidiaries_in” and “main_institute_in”. In 1981, almost all subsidiaries were transformed to institutes, while in the following years new subsidiaries appeared. In 2007, liaison offices were introduced and all subsidiaries were transformed to institutes or in few cases to liaison offices. The type of subsidiaries did not exist any more after 2006.[5] Before the official introduction of liaison offices in 2007, we assigned the type to Havana from 2004 to 2006 and to Tehran in 2006. In both cases, the GI reported that opening is

[4] Before 1971, only the number of subsidiaries were reported for each institute. Hence, we cannot include information about subsidiaries before 1971.

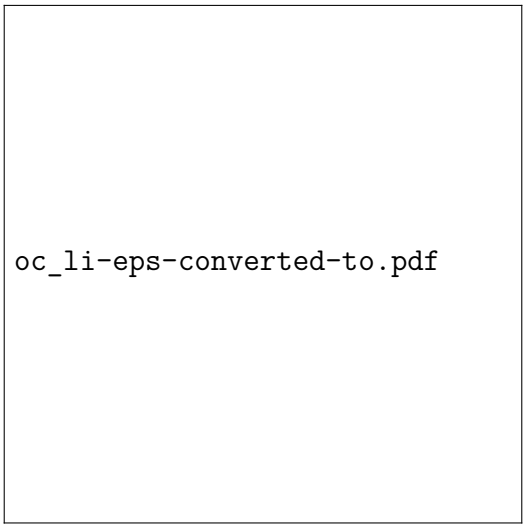
[5] Aleppo is the only exception, which was reported as a subsidiary from 2010 to 2012.

in preparation, but public activities like cultural events were already mentioned. From 2007 onwards, the reporting for both cities did not change, but the type liaison office was assigned in the annual report.

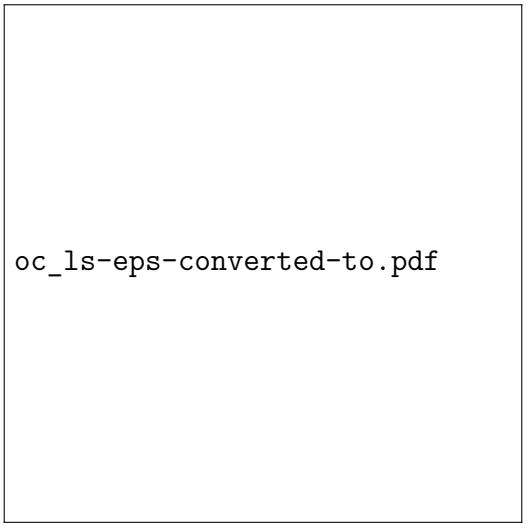
1.4 Language Services

One of the main tasks of the GI is to offer language services. Language services may be language courses, standardised language exams or both. The variable "language" takes the value 2, when an institute reports language services separately from other institutes. This is the general case. However, in some cases, the annual reports provide information on language services jointly for two or more cities, often main institutes with subsidiaries. We assume, that the reporting always refers to the first-named institutes, i.e. that first-named institutes offer language services (language=2). However, it is not clear, if not-first-named institutes offer language services. In these cases, the variable "language" takes the value 1. If there is no reporting of language services, the variable "language" takes the value 0. E.g., the annual report indicates "London (mit Nebenstelle Birmingham)" and jointly reports the numbers on language learning. We assign to London as the first-named institute the value 2 for the variable "language" and to Birmingham the value 1.

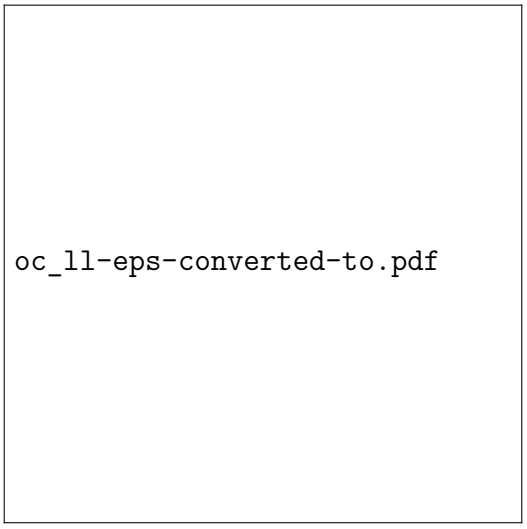
Figure 1 presents the language service offered for each type of the GI. Figure 1(a) shows that almost all main institutes offer language services. However, it also becomes evident that in some cases reporting about language services is done jointly for two main institutes or jointly for a main institute together with its subsidiaries. While this problem concerns only 1.2% of all observations of main institutes, subsidiaries joint reporting is the norm with 83.4 % of all observations (see Figure 1(b)). In these latter cases, which cover most years except for the period 1977 to 1980, it is not clear if the subsidiary offers language services as well. The four years with detailed reporting for the subsidiaries show that there is indeed no clear pattern. Some of the subsidiaries report language services, others do not. Figure 1(c) shows that there is no joint reporting for liaison offices of which less than 50% offer language services in most years.



(a) Main institutes



(b) Subsidiaries



(c) Liaison offices

Figure 1: Language services.

2 Codebook

region	Goethe-Institut region			
Format	Numeric			
Range	Value	Label	Frequency	Percent
	1	Central and Eastern Europe	450	3.31
	2	East Asia	700	5.15
	3	Eastern Europe and Central Asia	400	2.94
	4	North America	1,000	7.35
	5	North Africa and Middle East	1,200	8.82
	6	Northwest Europe	1,300	9.56
	7	South America	2,200	16.18
	8	Southeast Asia, Australia and New Zealand	800	5.88
	9	South Asia	1,100	8.09
	10	Southeast Europe	1,050	7.72
	11	Sub-Saharan Africa	1,100	8.09
	12	Southwest Europe	2,300	16.91
Missing	0/13,600			

country	Country of origin			
Format	Character			
Comment	ISO 3166 alpha-3 for currently existing countries			
	ISO 3166-3 for dissolved countries			
Missing	0/13,600			

city	Name of city with institute
Format	Character
Comment	City names
Missing	0/13,600

year	Year
Format	Numeric
Range	1965 – 2014
Missing	0/13,600

status_institute	Goethe-Institut status			
Format	Numeric			
Range	Value	Label	Frequency	Percent
	0	Closed	6,660	48.97
	1	Liaison Office	93	0.68
	2	Subsidiary	507	3.73
	3	Main institute	6,340	46.62
Missing	0/13,600			

language	Goethe-Institut language services			
Format	Numeric			
Range	Value	Label	Frequency	Percent
	0	No language services	7,299	53.67
	1	Jointly reported language services	501	3.68
	2	Language services	5,800	42.65
Missing	0/13,600			

subsidiaries_in	Subsidiaries of main institute
Format	Character
Missing	13,273/13,600
Comment	Missing values are for observations which are main institutes without subsidiaries, subsidiaries and liaison offices.

main_institute_in	Main institute of the subsidiary
Format	Character
Missing	13,102/13,600
Comment	Missing values are for observations which are no subsidiaries.

open1	Year of first opening
Format	Numeric
Range	1965 – 2014
Missing	0/13,600
Comment	Generated from “status_institute” for values 1, 2 and 3

close1	Year of first closing
Format	Numeric
Range	1966 – 2014
Missing	6,700/13,600
Comment	Generated from “status_institute” for value 0

open2	Year of second opening
Format	Numeric
Range	1971 – 2014
Missing	11,750/13,600
Comment	see above “open1”

close2	Year of second closing
Format	Numeric
Range	1973 – 2013
Missing	12,400/13,600
Comment	see above “close1”

open3	Year of third opening
Format	Numeric
Range	1974 – 2008
Missing	13,350/13,600
Comment	see above “open1”

close3	Year of third closing
Format	Numeric
Range	1982 – 2002
Missing	13,450/13,600
Comment	see above “close1”

open_before_1965	Open before 1965			
Format	Numeric			
Range	Value	Label	Frequency	Percent
	0	No	8,700	63.97
	1	Yes	4,900	36.03
Missing	0/13,600			

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.