Written translation support in the 2011 Census: Evidence used to rank languages in England and Wales according to estimated relative volume of need and use.

1. Executive summary

It is a statutory requirement that all households fill in their 2011 Census questionnaire. To facilitate this, translation support was provided. All languages were supported via a national helpline. Fifty-six languages were supported by written translation. Evidence of foreign language use and translation need was required in order to guide the decision of which languages should be supported. This evidence was primarily obtained from four sources: school census data for England; national insurance number allocations to overseas nationals made in England and Wales; and from data provided by two translation agencies utilised by the UK government. Languages were ranked by inferred volume from these four sources and any language present in the top 40 from any of these sources was selected to produce an initial list of 51 languages. An additional three languages were added following an analysis of country of origin of those recently granted asylum/exceptional leave. A final list of 56 languages was produced following consultation with the National Centre for Languages and through fulfilling commitments made by the census stakeholder management and communications team.

2. Introduction

The census produces population estimates based on the direct census count adjusted for under-coverage estimated from the Census Coverage Survey (CCS). The accuracy of the population estimates has dependence upon the census response rate and variability of response rates sub-nationally. There is evidence that the response rate of non-'White British' persons in the 2001 Census was lower than for 'White British' groups[1]. Furthermore ethnic and cultural diversity in England is increasing. According to the 2001 Census, 13 per cent of the population was other than 'White British'. In 2007, estimates for the proportion of non-'White British' in England had increased to 17 per cent[2]. It was therefore a challenge for the 2011 Census to engage with non-white groups to ensure that ethnic and cultural diversity was not a barrier to participation. This included catering for language diversity.

Before the 2011 Census, there was no data source that could provide an unbiased measure of the proportion of the population whose first language was other than English. School census data suggests that the proportion of pupils in state primary schools whose first language was not English in 2009 was 15 per cent[3]. Thus, there was a potential that language may be a barrier to a significant proportion of households in completing their census questionnaire. Certain households may have had difficulty in completing their census questionnaires because of lack of proficiency in English. Other households may be proficient in English but required engagement to motivate them to complete their census questionnaire. Therefore, there were two factors to consider in the provision of language support: overcoming physical need and promoting inclusion of all communities and cultures.

Oral language support for almost all languages was provided through a national helpline. Written translations of the questionnaire and information leaflet were available on request and from the online help service for the fifty-six languages. Evidence of foreign language use and translation need was required in order to inform the decision of which languages written translations and web self-help booklets should be available in. Thus data that relates to translation need and foreign language use was considered. Data on translations for UK residents interacting with government departments was thought to be the most direct measure of need available. Data on the proportion of telephone translations by language provided by two major suppliers of translation to the UK government was obtained. The most direct measurement of the relative proportion of households speaking languages other than English in the home was considered to be from the school census, which collects data on 'first language' from pupils educated in state schools in England. In addition to these data sources, data on national insurance number allocations to foreign nationals were obtained from the Department for Work and Pensions (DWP). Though this measures country of origin rather than language directly, it is likely to provide an indication of the relative size of inward migration and ensure that the impact of the recent ascension of eastern European countries into the EU is captured in our analysis.

Fifty-one languages were selected from these four sources with the criteria that the language was in the top 40 languages ranked by proportion in at least one source. This list was validated against Home Office data on country of origin of those granted asylum or exceptional leave, by reference to community languages provided by the National Centre for Languages, the 2009 Census Rehearsal list of languages, advice from external experts and discussions with census stakeholder management and communication. Following this process, it was decided that written translation support would be provided for 56 languages. Welsh versions of the questionnaire were developed for Wales.

3. Data sources

Before the 2011 Census, there were no data sources that could provide an unbiased estimate of foreign language use or the requirement for translation. Thus the approach was to obtain evidence from several sources for the ranking of foreign languages by their use or need for translation. Validity of the ranking is inferred from the degree of consistency between the data sources and by reference to external experts. This section describes the data used and their limitations.

3.1. Translation data

We used data relating to the proportion of telephone calls received for oral translation by language from two suppliers of translation services to the UK government. The first provided actual proportions of calls for the first nine months of 2009. The second provided a ranking of the top sixty languages based on proportions of calls using 2008 data. Both sources included all telephone translations provided in the UK to government and private organisations. There is the potential for bias in these data based on the relative make up of the customers of these organisations and the nature of their business, which was not disclosed. For example, if a large proportion of all translations are provided to the Department of Health, then one might expect an age bias (under the assumption that older people require more healthcare than younger people). Despite the potential for bias, there is strong agreement between the data sources: 39 of the top 40 languages on either list are shared. An additional potential limitation is that these data sources provide a ranking of spoken languages. This limitation was overcome through consultation with experts who are able to infer the written language from the spoken.

3.2. School census data

The school census is conducted each term and covers maintained schools in England. Data on 'first language' has been collected each year since 2008. The proportion of pupils by 'first language' in January 2008 has been published by the Department for Education[4]. A ranking of the proportion of pupils by first language was produced using 2009 spring term school census data, which was filtered to exclude those pupils outside the 5-16 age group. This dataset is expected to provide a picture of languages used in the home in England, but not on the proficiency of English in those homes. Though the dataset has a large coverage (over 6.5 million pupils), there is reason to believe that school census data provides a biased picture: that the school aged population is not representative of the whole population. The proportion of pupils in the school census whose ethnicity is 'White British' is lower than that

estimated for the population of England. Probable reasons for this include distinct fertility rates between ethnic groups and the impact of private sector education (the school census covers maintained schools only). Of particular concern was the possibility that recent young economic migrants may not be fully represented in the school aged population. To ensure that these were represented in our analysis, we obtained data on national insurance number allocations to overseas nationals.

3.3. National insurance number allocations to overseas nationals

A national insurance number (NINo) is generally required by any overseas national looking to work or claim benefits in the UK, including the self employed and students working part time. It contains rich data on young migrants who may not be represented in the school census data: 80 per cent of NINo allocations in 2008 were to persons between 18 and 34 years of age[5]. NINo allocations data covering England and Wales was obtained from the Department of Work and Pensions for each calendar year from 2004 to 2009, with data for 2009 covering the first three months only. This data does not reflect the size of the resident migrant population as there is no account of emigration. A weighted average was obtained from the 5 years data by arbitrarily assuming that 30 per cent of migrants leave the UK per year. This was applied on a yearly basis, with no discount on the 2009 data, resulting in a 30 per cent discount on the 2008 data and 51 per cent discount on the 2007 data, *et cetera*. This produced a ranking of countries and not a ranking of languages. Languages were inferred from the 'country of origin'. This was initially done through internet based research and finalised by referral to linguistic experts. A ranking of languages was not directly produced, but rather a rank assigned equal to the ranking of countries to all major indigenous languages used in that country. This ranking did not exclude countries whose official language is English. This produced a list of 41 non-English languages that were spoken in the 40 countries that made up this estimate of the largest source of migrants.

3.4. Those granted asylum or exceptional leave by the Home Office

We used Home Office data[6] on those granted asylum as refugees and those granted exceptional leave, humanitarian protection or discretionary leave for the period 2002 to 2008 as a validation check of our list (see section 5). As for NINo allocation data, this source does not capture the present pool and only identifies country of origin. In this case, estimates for the size of the pool were based on the assumption that 10 per cent of those granted asylum leave per year. Again language was inferred from country. Groupings of countries were excluded from the analysis.

4. Production of an initial list of written languages for translation support

Three distinct ranks of languages were produced, one each from the volume of use for 2008 and 2009 telephone translation data (obtained from distinct sources, both providing translation services to the UK government) and from 2009 School Census data on the proportion of pupils by 'first language'. These were converted into a ranking by written language with the following considerations: data was combined for dialects that shared the same written language, for example Sylheti is a dialect that shares the same written language as Bengali; languages that have more than one script, notably Punjabi, had both written languages assigned a common rank, which displaced languages ranked lower. A fourth, indirect, ranking of languages was produced via a ranking of 'country of origin' from the NINo allocation data. This fourth rank assigned each a language rank equal to the rank of the country of origin with which the language was associated. In cases where the same language was spoken in many countries, the rank of the highest country was used. As this fourth rank is crude, it is represented as blocks, with block 1 being a language used in a country within the top 10 countries of origin by volume, block 2 being in the 11-20 country rank, *et cetera*. The top 40 languages from all four lists were combined producing an initial list of 51 languages (table 1).

Table 1: Ranking of languages from the four major data sources used (see text for details).

Language	2009 translations	2008 translations	2009 School Census	2005-9 NINo allocations
Akan			20	Block 4 (31-40)
Amharic	35	30		
Arabic	16	7	8	Block 4 (31-40)
Bengali	9	14	4	Block 1 (1-10)
Bulgarian	20	31		Block 2 (11-20)
Cantonese	27	22	30	Block 2 (11-20)
Czech	3	19	39	Block 3 (21-30)
Dutch			35	Block 3 (21-30)
Farsi/Persian	24	8	18	Block 4 (31-40)
French	18	12	11	Block 1 (1-10)
German	34	34	27	Block 2 (11-20)
Greek		37	32	Block 4 (31-40)
Gujarati	22	33	5	Block 1 (1-10)
Hindi	26	28	21	Block 1 (1-10)
Hungarian	15	29		Block 2 (11-20)
Igbo			36	Block 1 (1-10)
Italian	28	26	22	Block 1 (1-10)
Japanese	40	39		
Korean	32	27		
Kurdish (Kurmanji)	25	6	31	Block 4 (31-40)
Kurdish (Sorani)	33	24		Block 4 (31-40)
Latvian	30	35		Block 3 (21-30)
Lingala	38	40	34	
Lithuanian	8	18	25	Block 1 (1-10)
Malay				Block 4 (31-40)

Malayalam			23	Block 1 (1-10)
Mandarin	17	2	19	Block 2 (11-20)
Nepali	37		24	Block 3 (21-30)
Pashto	29	21	17	Block 4 (31-40)
Polish	1	1	6	Block 1 (1-10)
Portuguese	4	4	9	Block 2 (11-20)
Punjabi (Gurmukhi)	11	10	2	Block 1 (1-10)
Punjabi (Shahmukhi)	11	10	2	Block 1 (1-10)
Romanian	10	13	38	Block 1 (1-10)
Russian	6	5	29	
Shona			26	Block 4 (31-40)
Shqip/Albanian	31	32	16	
Sinhala				Block 3 (21-30)
Slovak	2	17	33	Block 1 (1-10)
Somali	7	9	7	
Spanish	14	15	14	Block 2 (11-20)
Swahili	39	38	28	
Swedish				Block 4 (31-40)
Tagalog/Filipino			15	Block 2 (11-20)
Tamil	19	23	10	Block 1 (1-10)
Thai	36	36	40	Block 4 (31-40)
Tigrinya	21	20		
Turkish	13	3	12	Block 3 (21-30)
Urdu	5	16	1	Block 1 (1-10)
Vietnamese	23	25	37	
Yoruba			13	Block 1 (1-10)

5. Validation and consultation to produce a final list of languages

There is high consistency between the four data sources: 25 out of the 51 languages are ranked within the top 40 in all four data sources. The four data sources fall into three categories, the contributions of each are discussed below.

Telephone translation data is expected to be the most direct measure of need. We were concerned that the nature of the activity of government departments and businesses who commissioned the translations could potentially bias the data. However, data from two distinct sources are highly consistent with other. Indeed they are also highly consistent with the other two data sources. Only four languages on this initial list are present in the top 40 languages by translation only. These are Amharic (spoken in Ethiopia and Egypt), Japanese, Korean and Tigrinya (spoken in Eritrea and Ethiopia).

School census data on first language captures language use. This does not necessarily imply a need for translation. However, it was important to include all languages that are in common use in households within England and Wales in our list as it was believed that this would encourage participation in the census. The school census data identified seven languages not represented from the translation data. These languages were also identified as being spoken in countries from which most migration occurs (NINo allocations, table 1).

The NINo allocation data does not produce a ranking of languages directly, as languages have to be inferred from 'country of origin'. It was felt that this was an important data source to capture potential language use as the most mobile population group are young adults without children. Such migrants may be poorly represented in the school census. Consistent with this, east European languages tend to be ranked higher from the NINo allocation data than from school census data (table 1). There are three languages whose appearance in the initial list is solely derived from the NINo allocation data. These are Malay (spoken in Malaysia), Sinhala (spoken in Sri Lanka) and Swedish. No other languages were identified from a single source making the NINo allocation data the least consistent of the data sources.

The high consistency between data sources serves as a validation that we have captured all languages of major use in England and Wales. A small number of languages are included based on evidence of a single type: four based on translations data only and three based on NINo allocations only. As a further validation check, Home Office data on those granted asylum as refugees and those granted exceptional leave, humanitarian protection or discretionary leave for the period 2002 to 2008 was obtained. Although the numbers of persons represented by this data is small, they may represent a migration of persons from countries not captured by the NINo allocation data and whose need for translation is greater. Consistent with this view, 11 countries ranked in the top 20 countries of origin by number of successful asylum/exceptional leave applications are not ranked in the top 40 by NINo registrations. All but two of these countries have indigenous languages on our initial list. These languages include Amharic and Tigrinya, which were captured by translation data only, as well as Lingala, Somali and Swahili. The two countries in the top 20 countries of origin for asylum/exceptional leave which had indigenous languages not in our list were Serbia and Montenegro and Uganda. The Serbian Cyrillic script and Bosnian/Croatian roman script were added to our list to cover Serbia and Montenegro. Luganda was added as the indigenous language of Uganda.

An additional validation was to ensure that all 33 languages supported in the 2009 Census rehearsal were included. The selection of languages for the census rehearsal used the same methodology as presented here with the exception that only one source of translation data was used and that all sources used earlier data. All languages supported in the census rehearsal are present in table 1.

Finally, the National Centre for Languages and census stakeholder management and communication were consulted. The National Centre for Languages produces a list of community and modern languages that are taught in school[7]. Community languages include Urdu, Chinese, Arabic, Bengali, Russian, Punjabi, Turkish, Japanese, Gujarati, Portuguese, Greek, Hebrew, Persian, Polish and Dutch. Hebrew is the only language from the community languages that is not covered. Yiddish was added on the advice that its inclusion would have more impact in engaging the Jewish community. The final language to be added was Pahari, which resulted from census stakeholder management and communications engagement with Kashmiri groups. Pahari is a language used in parts of Pakistan including Kashmir. Though Pakistan was ranked third as a country of origin for NINo allocations, Pahari was not originally selected as a language inferred from Pakistan as it appears to be a relatively minor language in that country as a whole[8]. Consistent with the stakeholders point of view, however, Pahari is ranked 44th in the 2009 School Census data.

This process resulted in the 56 foreign languages that were supported by written translation in the 2011 Census. Indigenous languages to the UK were also supported. The 56 foreign languages are listed below:

Akan	Amharic	Arabic	Bengali
Bosnian/Croatian	Bulgarian	Cantonese	Czech
Dutch	Farsi/Persian	French	German
Greek	Gujarati	Hindi	Hungarian
Igbo	Italian	Japanese	Korean
Kurdish (Kurmanji)	Kurdish (Sorani)	Latvian	Lingala
Lithuanian	Luganda	Malay	Malayalam
Mandarin	Nepali	Pahari	Pashto
Polish	Portuguese	Punjabi (Gurmukhi)	Punjabi (Shahmukhi)
Romanian	Russian	Serbian	Shona
Shqip/Albanian	Sinhala	Slovak	Somali
Spanish	Swahili	Swedish	Tagalog/Filipino
Tamil	Thai	Tigrinya	Turkish
Urdu	Vietnamese	Yoruba	Yiddish

6. References:

1. http://www.ons.gov.uk/ons/guide-method/census/census-2001/data-and-products/quality-of-the-census-data/imputation-rates/index.html
2. Population Estimates by Ethnic group 2001 – 2009, 18 May 2011, ONS Statistical Bulletin

3. http://www.education.gov.uk/rsgateway/DB/SFR/s000843/index.shtml
4. http://www.education.gov.uk/rsgateway/DB/SFR/s000786/sfr09-2008lang.xls
5. http://research.dwp.gov.uk/asd/asd1/tabtools/nino_allocations_aug09.pdf
6. http://www.homeoffice.gov.uk/rds/immigration-asylum-stats.html
7. http://www.cilt.org.uk/home/research_and_statistics/language_trends/community_languages.aspx
8. http://www.ethnologue.com/show_country.asp?name=pk

Acknowledgements

The authors would like to express their gratitude to David Elgar at *Languageline* and James Williams at *thebigword* for providing data; Antony Sanderson at Surrey County Council and Anthea Gupta at the University of Leeds for providing advice on the relationship between spoken and written languages; and Youping Han and Anne-Marie Graham at the National Centre for Languages for advice, including how to include analysis of asylum seekers in our work.

Neil A. Hopper and Enliz D'Souza
Census Methodology
Methodology Directorate
Office for National Statistics