

Canadian Century Research Infrastructure Project

Language Background

Root variable: Language

Synopsis of code set construction:

This code set was constructed using the expected responses from the Enumerator Guides and the Dominion Bureau of Statistics (DBS) code books, and by consulting the microdata for imputed values DBS did not expect or did not distinguish.

We consulted Statistics Canada's Mother Tongue Standard Classification, as well as language code sets from the Canadian Families Project (CFP) and the Integrated Public Use Microdata Series (IPUMS). We decided to use a modified Mother Tongue Standard Classification because:

1. It allows the possibility of multiple answers;
2. It is designed specifically for the Canadian context; and
3. The listing of languages in families allows for easy aggregation.

The question concerning languages spoken changed in 1921. In 1911, respondents were asked which languages they spoke, and enumerators recorded multiple languages in a single column. Beginning in 1921 enumerators recorded "yes" in distinct columns if respondents could speak English or French, and there was a separate column for "languages other than English or French spoken as Mother Tongue." Responses to the official languages questions ("Do you speak English? Do you speak French?") have been coded using the YES/NO root variable code set.

In many census years, DBS anticipated that respondents would give only one language as their Mother Tongue. But the imputed data demonstrates that some people gave multiple responses. The code set had to be designed to make this information readily available to researchers. We decided that multiple responses that included Official Languages would be coded separately, and other multiple languages would be amalgamated under one code, since including all the possible multiples would make the code set unwieldy.

Understanding the numeric code:

This code has seven digits.

The first digit indicates whether the respondent gave a single or a multiple response.

- 1 = Multiple response
- 2 = Single response

The second and third digits refer to the language family.

- 01 = Aboriginal languages
- 02 = Romance languages
- 03 = Germanic languages

Canadian Century Research Infrastructure Project

Language Background

- 04 = Netherlandic languages
- 05 = Scandinavian languages
- 06 = Celtic languages
- 07 = Slavic languages
- 08 = Baltic languages
- 09 = Finno-Ugric languages
- 10 = Turkic languages
- 11 = Semitic languages
- 12 = Indo-Iranian languages
- 13 = Dravidian languages
- 14 = Japanese
- 15 = Korean
- 16 = Sino-Tibetan languages
- 17 = Tai languages
- 18 = Austro-Asiatic languages
- 19 = Malayo-Polynesian languages

The fourth and fifth digits indicate groups of languages within the the Aboriginal Languages family.

The sixth and seventh digits refer to distinct spoken languages.

In the case of codes beginning with a 1, the sixth and seventh digits refer to various combinations of languages learned simultaneously.

- 1 = English and French
- 2 = English and Other
- 3 = French and Other
- 4 = Other multiples

Code set application notes:

- In cases where two languages were given, they were mapped as “Other multiples.” The only exception to this rule was for those that gave English and other, English and French and French and other, which were mapped to account for the two official languages.
- Occasionally, religious or ethnic distinctions might be mapped to a specific language set, depending on origin. For example, “Doukhobor” was mapped as “Russian.”
- In the case of Switzerland, those that gave Swiss and one of its official languages would be mapped to that language. Thus “Swiss Italian,” would become “Italian.”
- Certain geographic distinctions were mapped based on their official language. I.e. “Mexican” was mapped to “Spanish.”
- Data entry operators were instructed to enter a question mark for each illegible character in a response. These values were coded to “Illegible.”

