A socio-demographic Database fOr Varieties of English (DOVE) v1.1

Documentation

Katharina Ehret
https://orcid.org/0000-0003-1117-3495

07.11.2025

Preface

This document provides general definitions for all language-external triggers and the additional descriptive information comprised in DOVE v1.1. It also provides a list of individual definitions and details on deviating operationalisations of the language-external triggers whenever they had to be derived via some other variable. Thus, this document is complementary to the data provided in DOVE. It does not contain the sources and copyright, nor any numerical data.

Contents

1 Geography

1.1 Geographic spread

Geographic spread (SPREAD) in DOVE refers to country size in square kilometres for national varieties. For regionally restricted varieties it refers to within-country territories in square kilometres. For islands and island states only the land area is counted (water area is not included, where applicable/known).

1.1.1 Comments and search terms

- Ireland, Irish English: The sum of the Republic of Ireland and Northern Ireland
- Scotland, Hebridean English: Outer Hebrides and Isle of Skye
- Scotland, Scottish English: Scotland
- England, Orkney and Shetland English: Northern Isles
- England, England, English dialects in the North: Northern England
- England, English dialects in the Southwest: South West England
- England, English dialects in the Southeast: South East of England
- England, English dialects in the Midlands: Leicestershire, Nottinghamshire, Shropshire, Warwickshire
- Wales, Welsh English: Wales
- Canada, Newfoundland English: Newfoundland island only
- Great Britain, British English: Great Britain
- Canada, Canadian English: Canada

1.2 Coordinates: Longitude and latitude

Longitude and latitude in WGS84 format (LONGITUDE, LATITUDE as listed in *the electronic World Atlas of Varieties of English* (eWAVE, Kortmann et al. (2020)). For non-eWAVE varieties the longitude and latitude of the linguistic/administrative centre is provided. All values were rounded to two decimal places.

- Great Britain, British English: London
- Canada, Canadian English: Ottawa
- England, English dialects in the Midlands: Birmingham
- Trinidad, Trinidadian English: Port of Spain
- Scotland, Hebridean English: Stornoway

1.3 Region and macro-region

The geographic world region and macro-regions. Regions are taken from eWAVE (Kortmann et al. 2020) or were added based on the geographic location of the varieties. Macro-regions are largely identical with regions but importantly, subsume the regions *Australia* and *Pacific* under *Oceania*.

2 Contact

2.1 Number of contact languages

The number of other languages spoken in a territory (and/or adjacent territories) is often used as a proxy for language contact (Lupyan & Dale 2010; Nichols 1992). In DOVE, the number of contact languages is operationalised in two ways. First, as the number of major contact languages spoken by ten percent of the total population (CONTACT_LANGUAGES), and second, as the number of all first languages spoken in a country irrespective of speaker numbers (ALL_L1).

In the literature, it is rarely discussed whether other languages counted as contact languages are spoken by a sufficiently large number of speakers for structural borrowing or interference to occur. It is a well-known fact in contact linguistics that structural borrowing or interference may only take place if the number of speakers of the source language is sufficiently large. Specifically, structural borrowing can potentially occur in cases of long-term contact involving a large number of fluent bilingual speakers of the source language who are fluent in the target language as well (Thomason & Kaufman 2023: 47–50; 67). In this context, the source language is maintained. In another context, namely, a situation of language shift which may result in the replacement of the source language by the incoming target language, a large number of speakers needs to shift to the target language without having sufficient access to the target language. This typically results in imperfect learning and in source language interference (Thomason & Kaufman 2023: 47–50). Thus, as major contact languages, only those languages spoken in a country – apart from the respective English variety or varieties – are counted which are spoken by a substantial number of speakers. Specifically, contact languages are therefore defined here as languages spoken by at least ten percent (after rounding to the nearest whole number) of the total population. If the percentage of speakers was not given, it was calculated based on the number of speakers per language and the total population of the respective country for the same year and rounded up to two decimal points. Note: All first language (L1) and second language (L2) English varieties as well as English-based pidgins and creoles are counted as English varieties.

Largely following the typological literature (Shcherbakova et al. 2023; Lupyan & Dale 2010), 'all L1 languages" spoken in a given country/territory is also included as an alternative literature-motivated approximation for language contact. All L1 languages spoken in a country excludes sign languages, artificial languages, extinct languages and languages recorded for bookkeeping. This count includes English varieties/creoles/pidgins.

Unless otherwise indicated, the number of contact/all L1 languages for within-country territories (e.g. North of England) are the ones given for the entire country. For overseas territories (e.g. Falkland Islands) and smaller independent islands in the British Isles (i.e. Orkney, Shetlands, Channel Islands, Isle of Man), if no other data is available, the data of the sovereign state (Britain, France, Colombia) is provided.

2.1.1 Comments

- For Irish English, the sum of the contact languages/all languages in the Republic of Ireland and Northern Ireland (UK) is reported.
- For Hebridean English, data for Scotland is used.

3 Demography

Disclaimer: The concepts of mother tongue/native/first language speaker are highly problematic (Cheng et al. 2021) and hard to define, especially and also in the inherently multilingual context of English varieties. Acknowledging these conceptual issues, the labels are nevertheless adopted in this dataset for reasons of operationalisation and feasibility in collecting speaker information. No (d)evaluative judgements are made regarding proficiency, identity, or ethnicity of the speakers.

3.1 Native speakers

The number of people who speak English as one of their mother tongues, native languages, home languages, first languages, main languages, or primary languages in a given country or territory were counted as native speakers (NATIVES). Thus, the number of native speakers comprises only the number of speakers in a given country or territory, in which the variety is primarily spoken or originates from. The labels listed above are treated as equivalent and are based on the range of labels used in census data and other governmental or non-governmental institutional surveys. For English creole and pidgin varieties, the number of native speakers refers to the number of people who speak the given creole/pidgin variety as one of their mother tongues, native languages, home languages, first languages, main languages, or primary languages. For these pidgin/creole varieties, if applicable, the number of people speaking English as one of their mother/native/home/first/primary languages in the same territory is also provided. The number of speakers comprises only the number of speakers in a given country or territory, in which the variety is primarily spoken or originates from, thus excluding, for instance, diaspora speakers.

For some varieties the number of native speakers was approximated based on other variables. These exceptions and deviating individual definitions are listed below.

Unless otherwise indicated, in cases in which the total population count was not included in the data, the population totals for a country were retrieved from World Bank Open Data (`https://data.worldbank.org/`, last accessed 08.05.2023).

3.1.1 Exceptions and individual definitions

- Southern Ireland (Republic of Ireland), Southern Irish English (used for Irish English, not in eWAVE): The number of people who speak a language other than English/Irish at home was subtracted from the number of people born in Ireland.
- Ireland, Irish English: The sum of the number of native speakers reported for Northern Irish and Southern Irish English.
- Hebrides, Hebridean English (FRED variety, not in eWAVE): Approximated by counting all people who report speaking English as their main language and also speak, read and

write Gaelic/speak Gaelic/speak and read Gaelic. Gaelic as main language speakers are also included as it can be safely assumed that they speak English as well (McLeod 2006). This rather narrow definition includes only bilingual Hebridean English-Gaelic speakers in analogy to the definition applied by (Sabban 1982). Calculated on tables for Na h-Eileanan Siar and Eilean a' Chèo (Isle of Skye). For Isle of Skye information on who speaks, reads and writes Gaelic/speaks Gaelic/speaks and reads Gaelic was taken as proxy as cross-tabulated data at this level is not accessible.

- Isle of Man, Manx English: The sum of all residents by parish who are 60 years and older as approximation for native speakers based on the assumption that Manx English is typically spoken as L1 by older Manx ethnic, rural speakers (Draskau 2012). Ethnicity could not be extrapolated by parish and age.

- Canada, Canadian English (ICE variety, not in eWAVE): Speakers of English as Mother tongue, sum of the following categories "single responses for English" and "multiple responses for English and French", as well as "English and non-official language(s)", "English, French and non-official language(s)".

- New Zealand, New Zealand English: As conservative approximation, the number of native speakers is based on the percentage of the population who speak English only.

- Ghana, Ghanaian English: Based on Magnus Huber (p.c., 18.06.2023) approximately 10% of the population who is at the age of language acquisition is currently acquiring some form of English as a native language. These children's parents mostly do not have English as their native language (Afrifa et al. 2019). Thus, this percentage must drop with increasing age and the population above the age of 30 did not grow up with English as their L1 or, if so, in negligible numbers. The age of 30 seems reasonable given the fact that the peak for fertility rate by age group in Ghana is around 30 or 25-29 years of age depending on region. Based on this information, the number of native speakers was approximated by a step-wise decreasing percentage for the population 0-29 years of age (0-5 yrs: 10%, 6-10 yrs: 9%, 11-15 yrs: 8%, 16-20 yrs: 6%, 21-25 yrs: 4%, 26-29 yrs: 2%, 30 yrs and above: 0%).

- Sri Lanka, Sri Lankan English: Approximated by calculating 0.05% based on Liyanage (2022) of the total population in 2021.

- Philippines, Philippine English: Calculated based on English speaking households and average number of persons per household.

- Trinidad, Trinidadian English: Approximated via educational attainment based on information on Standard English competence and the sociolinguistic situation in Trinidad(Deuber 2014; Youssef, Valerie & Winford 1999). Specifically, the completion rate of upper secondary education (both sexes, in percent) was calculated for the total population in 2024.

- Wales, Welsh English: Calculated based on the number of Welsh native speakers which were subtracted from the number of people speaking English/Welsh as L1 (the census conflates this information in the questionnaire).

- England, English dialects in the Midlands (FRED variety, not in eWAVE): approximated as the number of speakers who report English (or Welsh in Wales) as their main language, are born in the UK, aged 45 years and above, and have received no education beyond GCSE (selected census categories: “Level 1” and “Level 2 qualifications”, “no qualification”), for the East Midlands and the West Midlands. See Appendix A for details and rationale.

- England, English dialects in the North: approximated by the number of speakers who report English (or Welsh in Wales) as their main language, are born in the UK, and have an “English only” OR “British and English only” national identity, for the North East and North West regions, plus the following data for Yorkshire and the Humber. Number of Speakers who report English (or Welsh in Wales) as their main language, are born in the UK, aged 45 years and above, and have received no education beyond GCSE (selected census categories: “Level 1” and “Level 2 qualifications”, “no qualification”), for Yorkshire and the Humber. Note: differences in census variable selection are due to data protection. See Appendix A for details and rationale.

- England, English dialects in the Southwest: approximated by the number of speakers who report English (or Welsh in Wales) as their main language, are born in the UK, aged 45 years and above, and have received no education beyond GCSE (selected census categories: “Level 1” and “Level 2 qualifications”, “no qualification”), for the South West. See Appendix A for details and rationale.

- England, English dialects in the Southeast: approximated by the number of speakers who report English (or Welsh in Wales) as their main language, are born in the UK, and have currently “working class occupations” (selected categories: “Skilled trades occupations”, “Caring, leisure and other service occupations”, “Process, plant and machine operatives”, “Elementary occupations”), for the South East. See Appendix A for details and rationale.

3.2 Non-native speakers

All speakers in a given territory or country who speak a given variety of English as (one of their) second or additional languages are counted as non-native speakers of English, disregarding their level of competence/proficiency unless otherwise stated (NON_NATIVES, labelled below as ”L2”). In the case of pidgins and creoles, the term non-native speaker refers to all speakers in a given territory who speak the pidgin/creole as (one of their) second or additional languages, disregarding their level of competence/proficiency unless otherwise stated. In cases in which the sources provide fine-grained information on the level of competence/proficiency in the target language, this information is included as NON_NATIVES_WELL for the categories “speaks a language well” and ”speaks a language very well” and NON_NATIVES_NOTWELL for the category “speaks a language not well”. The number of non-native speakers comprises only the number of speakers in a given country or territory, in which the variety is primarily spoken or originates from.

For some varieties the number of native speakers was approximated based on other variables. These exceptions and deviating individual definitions are listed below.

When percentages were provided, the number of speakers was calculated as the given percentage of the total population (in the respective or most recent year), unless otherwise indicated. If the total population count was not included in the data, the population totals for a country were retrieved from World Bank Open Data (`https://data.worldbank.org/`, last accessed 08.05.2023), unless otherwise indicated.

3.2.1 Exceptions and individual definitions

- Scotland, Scottish English L2: Approximated by the number of people who “Speaks, reads and writes English”, “Speaks but does not read or write English”, “Speaks and reads but does not write English”. The other categories were not counted as speakers (“Understands but does not speak, read or write English”, “Reads but does not speak or write English”, “Other combination of skills in English”, “No skills”, “Not applicable”).

- Hebrides, Hebridean English L2 (FRED variety, not in eWAVE): Non-native speakers are approximated by counting all people who speak, read and write English, speak and read or speak English for Na h-Eileanan Siar and Eilean a' Chèo minus the number of native speakers.

- Northern Ireland, Northern Irish English L2 (used for Irish English, not in eWAVE): Based on proficiency in English, Main language is not English with various categories: “Can speak English very well”, “Can speak English well” , “Cannot speak English well”.

- Southern Ireland (Republic of Ireland), Southern Irish English L2 (used for Irish English, not in eWAVE): Total number of residents who speak English ”very well”, ”well”, and ”not well” as proxy for non-native speakers.

- Ireland, Irish English L2: The sum of the counts from Northern and Southern Irish English.

- Canada, Canadian English L2 (ICE variety, not in eWAVE): Approximated by subtracting the number of English mother tongue speakers from the number of speakers with knowledge of official languages (”English”, and ”English and French”).

- New Zealand, New Zealand English L2: As a conservative approximation, the number of non-native speakers is calculated based on the total number of English speakers minus the English-only speakers who were counted as native speakers.

- Ghana, Ghanaian English L2: Approximated by the number of people who are literate in English minus the number of native speakers above the age of 6. The literacy rate in percent is provided only for the population above the age of 6. Thus, the percentage was calculated for the part of the population only who is 6 or above. For this, the number of people 0-5 was subtracted from the total population.

- Uganda, Ugandan English L2: The number of non-native speakers was approximated by educational attainment using the sum of all speakers who had some formal education for NON_NATIVES. Speakers with no formal education were not considered. Numbers were calculated from percentages for the population ten years and older.

- India, Indian English L2: Approximated by counting the number of speakers who speak English as a second or third additional language.

- Singapore, Colloquial Singapore English L2: Approximated based on speaker counts for Singlish, eWAVE, `https://ewave-atlas.org/languages/57` (Kortmann et al. 2020), and subtracting the number of speakers who speak only English at home.

- Trinidad, Trinidadian English L2: The number of non-native speakers was estimated as the percentage of the population who did not complete upper secondary education following the rationale outlined for native speaker estimation above. The approximate estimation is based on the completion rate of upper secondary education, both sexes, in percent, in 2011. This percentage was calculated for the total population of 1.5 million in 2024.

- Australia, Australian English L2: Proficiency in spoken English for those who declared to use another language than English as their main language at home. Tabulated by year of arrival in Australia. The categories "Speaks English very well", "well", and "not well" were counted. The numbers for "non-stated" were excluded.

- England, Scotland and Wales, British English L2 (ICE variety, not in eWAVE): Approximated as the sum of non-native numbers for England and Wales plus Scotland. The number of people speaking English not as their main language "very well or well" and "cannot speak well" for England and Wales. For Scotland, proficiency levels refer to literacy (reading and writing) and speaking (see entry for Scottish English L2). As these are not comparable to the estimates for England and Wales only a total number of non-natives was calculated.

- England, English dialects in the Midlands L2 (FRED variety, not in eWAVE): Non-native speaker were approximated through various variables: Country of birth (3 categories = Born in the UK) by national identity (9 categories, selected: "English only identity", "English and British only identity"), current occupation (10 categories, selected: "Skilled trades occupations", "Caring, leisure and other service occupations", "Process, plant and machine operatives", "Elementary occupations") by English proficiency (4 categories, selected: "Main language is not English (English or Welsh in Wales): Can speak English very well or well") by region. In this case, less proficient speakers were excluded. Regions: East and West Midlands. See Appendix A for details and rationale.

- England, English dialects in the North L2: Non-native speaker were approximated through various variables: Country of birth (3 categories = Born in the UK) by national identity (9 categories, selected: "English only identity", "English and British only identity"), current occupation (10 categories, selected: "Skilled trades occupations", "Caring, leisure and other service occupations", "Process, plant and machine operatives", "Elementary occupations") by English proficiency (4 categories, selected: "Main language is not English (English or Welsh in Wales): Can speak English very well or well") by region. In this case, less proficient speakers were excluded. Regions: North West and Yorkshire and the Humber. Settings for the North East vary due to data protection: Non-native speaker were approximated through various variables: Country of birth (3 categories = Born in the UK) by national identity (9 categories, selected: "English only identity", "English and British only identity"), by English proficiency (4 categories, selected: "Main language is not English (English or Welsh in Wales): Can speak English very well or well") by region. In this case, less proficient speakers were excluded. Regions: North East. See Appendix A for details and rationale.

- England, English dialects in the Southwest L2: Non-native speaker were approximated through various variables: Country of birth (3 categories = Born in the UK) by national identity (9 categories, selected: "English only identity", "English and British only identity"), current occupation (10 categories, selected: "Skilled trades occupations", "Caring, leisure and other service occupations", "Process, plant and machine operatives", "Elementary occupations") by English proficiency (4 categories, selected: "Main language is not English (English or Welsh in Wales): Can speak English very well or well") by region.

In this case, less proficient speakers were excluded. Regions: South West. See Appendix A for details and rationale.

- England, English dialects in the Southeast L2: Non-native speaker were approximated through various variables: Country of birth (3 categories = Born in the UK) by national identity (9 categories, selected: “English only identity”, “English and British only identity”), current occupation (10 categories, selected: “Skilled trades occupations”, “Caring, leisure and other service occupations”, “Process, plant and machine operatives”, “Elementary occupations”) by English proficiency (4 categories, selected: “Main language is not English (English or Welsh in Wales): Can speak English very well or well”) by region. In this case, less proficient speakers were excluded. Regions: South East. See Appendix A for details and rationale.

For the varieties listed below, it is assumed that there are no non-native speakers. Given the geography of the territories in which these varieties are spoken and the nature and/or status of the varieties, it is highly unlikely that there are non-native speakers above a threshold of 10% and that these varieties are acquired as second or additional language (variety).

- Isle of Man, Manx English L2 (Draskau 2012).

3.3 Net migration rate

The net migration rate per 1000 population (migration) is taken as an indicator for a country’s socioeconomic status but it can also be interpreted as an indicator of contact (for positive values). It is defined as the number of incoming migrants minus the number of outgoing migrants per year (here for 2024), “divided by the number of person-years lived by the population of the receiving country over that period” (UNdata source: World Population Prospects Glossary, `http://data.un.org/Default.aspx`, last accessed 07.11.2025).

3.4 Gross national income per capita

The Gross National Income per capita (GNI) is taken as an indicator for a country’s standard of living, and is one component of the Human Development Index (United Nations Development Programme, `https://hdr.undp.org/`, last accessed 07.11.2025). Unless otherwise stated, for within-country territories (e.g. North of England) the GNI for the entire country or the next largest administrative region is given (e.g. England). Similarly, for independent islands in the British Isles (e.g. Isle of Man) the GNI for the entire country is given, if no other data is available.

3.4.1 Comments

- Ireland, Irish English: the mean of the GNI for the Republic of Ireland and Northern Ireland (UK) is given.

3.5 Education

Education (EDUCATION) is operationalised as mean years of schooling and is an indicator of the educational level in the population of a specific country. It is also part of the Human Development Index (United Nations Development Programme, `https://hdr.undp.org/`, last accessed 07.11.2025). Unless otherwise stated, for within-country territories (e.g. North of England) the educational level for the entire country or the next largest administrative region is provided (e.g. England). Similarly, for independent islands in the British Isles (e.g. Isle of Man) the educational level for the entire country or next largest administrative region is provided (e.g. Scotland for the Hebrides), if no other data is available.

3.5.1 Comments

- Ireland, Irish English: the average value of the Republic of Ireland and Northern Ireland (UK) is provided.

3.6 Literacy rate

Literacy rate (LITERACY) refers to the percentage of adult literacy (i.e. all people aged 15 years and above, for both sexes) in the population of a country. This means that for varieties located in smaller within-country territories, the literacy rate for the entire country is used. In cases where several values were available, the most recent value was used. Note that the literacy rate does not always necessarily refer to the literacy rate in English.

3.6.1 Comments

The literacy rate for the following varieties is calculated based on the information of literacy below level 1 OECD (2019). In other words, people with a literacy rate below level 1 were not counted as literate.

- Australia, Australian English
- New Zealand, New Zealand English
- United States, Colloquial American English
- Canada, Canadian English
- England, all dialects spoken in the North/Midlands/Southwest/Southeast as well as Manx English
- Scotland, Scottish English/Hebridean English
- Wales, Welsh English

Furthermore for

- Ireland, Irish English: the literacy rate of Northern Ireland/UK and the Republic of Ireland is provided.
- Hong Kong, Hong Kong English: the literacy rate for English is provided.

4 Additional descriptive information

4.1 Language type

Language type (LANGUAGE_TYPE) is a theoretical construct describing the sociohistorical background including the contact historical situation of the varieties. It was taken from the *electronic World Atlas of Varieties of English 3.0* (eWAVE, Kortmann et al. (2020), `https://ewave-atlas.org/`, last accessed 11.08.2025) for varieties included therein. Descriptions for additional varieties were mainly derived based on relevant literature or the category descriptions in eWAVE. eWAVE lists the following language types: traditional low-contact L1 varieties (L1T), high-contact L1 varieties (L1C), high-contact indigenous L2 varieties (L2), for detailed definitions of the types see Kortmann et al. (2020).

- Hebridean English: Traditional low-contact L1 variety as Hebridean English can be considered a Scottish English subvariety (which is labelled L1t in eWAVE) (Kortmann et al. 2020).
- British English: High-contact L1 varieties as these can be considered standard varieties (Kortmann et al. 2020).
- Canadian English: High-contact L1 varieties as these can be considered standard varieties (Kortmann et al. 2020).
- English dialects in the Midlands: Traditional L1 variety as described in the *Freiburg Corpus of English Dialects* (FRED, https://freidok.uni-freiburg.de/proj/1, last accessed 11.08.2025) and in analogy to all the other dialects in Kortmann et al. (2020).
- Trinidadian English: Indigenised L2 variety based on the description of this variety type in eWAVE (Kortmann et al. 2020) and Deuber (2014).

4.2 Language IDs

ID (ID) was taken from Kortmann et al. (2020) for varieties listed therein, and is a unique identifier for each variety. For additional varieties an ID was generated and manually added.

4.3 Language names

The language names (NAME) was taken from Kortmann et al. (2020) for varieties listed therein. For additional varieties an intuitive language name was generated based.

4.4 Glottocode

Glottocodes (GLOTTOCODE) was taken from Kortmann et al. (2020) for varieties listed therein. For additional varieties, glottocodes were retrieved from *Glottolog 5.0* (Hammarström et al. 2024) if a match could be found.

4.5 Country IDs

The country IDs (COUNTRY_ID) were taken from Kortmann et al. (2020) for varieties listed therein. For additional varieties, the two-letter country code ISO 3166 was manually added based on `https://www.iso.org/iso-3166-country-codes.html` (last accessed 15.08.2025).

4.6 Abbreviation

Abbreviations (ABBR) were taken Kortmann et al. (2020) for varieties listed therein, and provides an abbreviation of the variety name. For additional varieties, an intuitive abbreviation was chosen and manually added.

4.7 Corpus information

Corpus information (CORPUS) indicates whether a given variety is part of a corpus and which corpus was used to study this variety in various related publications. Currently, these corpora are the *International Corpus of English* (ICE, `https://www.ice-corpora.uzh.ch/en.html`, last accessed 15.08.2025), the *Freiburg Corpus of English Dialects* (FRED, `https://freidok.uni-freiburg.de/proj/1`, last accessed 15.08.2025), and the *Santa Barbara Corpus of Spoken American English* (SBCSAE, `https://www.linguistics.ucsb.edu/research/santa-barbara-corpus`, last accessed 15.08.2025).

4.8 eWAVE information

eWAVE (EWAVE, yes/no) indicates whether a given variety is also described in Kortmann et al. (2020).

References

Grace Ampomaa Afrifa, Jemima Asabea Anderson, and Gladys Nyarko Ansah. The choice of English as a home language in urban Ghana. *Current Issues in Language Planning*, 20(4):418–434, 2019.

Lauretta S. P. Cheng, Danielle Burgess, Natasha Vernooij, Cecilia Solís-Barroso, Ashley McDermott, and Savithry Namboodiripad. The Problematic Concept of Native Speaker in Psycholinguistics: Replacing Vague and Harmful Terminology With Inclusive and Accurate Measures. *Frontiers in Psychology*, 12:715843, 2021. doi: doi:10.3389/fpsyg.2021.715843.

Dagmar Deuber. *English in the Caribbean: Variation, Style and Standards in Jamaica and Trinidad.* Cambridge University Press, Cambridge, 2014.

Jennifer Kewley Draskau. Manx English. In Bernd Kortmann and Kerstin Lunkenheimer, editors, *The Mouton World Atlas of Variation in English*, pages 48–57. De Gruyter Mouton, Berlin/Boston, 2012. URL `https://doi.org/10.1515/9783110280128`.

H Hammarström, R Forkel, M Haspelmath, and S Bank. Glottolog 5.0, 2024.

Arthur Hughes, Peter Trudgill, and Dominic Watt. *English accents and dialects: An introduction to social and regional varieties of English in the British Isles.* Routledge, 2013.

Bernd Kortmann, Kerstin Lunkenheimer, and Katharina Ehret, editors. *The Electronic World Atlas of Varieties of English.* 2020. URL `https://ewave-atlas.org/`.

Indika Liyanage. English Language Education Policy in Sri Lanka: Historical Developments, Current Realities and Future Challenges. In Ee Ling and Anne Pakir, editors, *English in East and South Asia: Policy, Features and Language in Use*, Routledge Studies in World Englishes. Routledge, London/New York, 2022.

Gary Lupyan and Rick Dale. Language structure is partly determined by social structure. *PLoS ONE*, 5(1):1–10, 2010.

Wilson McLeod. Gaelic in contemporary scotland: contradictions, challenges and strategies. *Anuari: Revista de Recerca Humanística i Científica 17 [Europa parla (II): Llengüües no romàniques minoritzades d'Europa]*, pages 83–98, 2006.

Johanna Nichols. *Linguistic Diversity in Space and Time.* University of Chicago Press, Chicago, IL, 1992.

OECD. *Skills Matter: Additional Results from the Survey of Adult Skills. OECD Skills Studies.* OECD Publishing, 2019.

Annette Sabban. *Gälisch-Englischer Sprachkontakt: zur Variabilität des Englischen im gälischsprachigen Gebiet Schottlands. Eine empirische Studie.* Groos, Heidelberg, 1982.

Olena Shcherbakova, Susanne Maria Michaelis, Hannah J. Haynie, Sam Passmore, Volker Gast, Russell D. Gray, Simon J. Greenhill, Damián E. Blasi, and Hedvig Skirgård. Societies of strangers do not speak less complex languages. *Science Advances*, 9(33):eadf7704, 2023.

Sarah Grey Thomason and Terrence Kaufman. *Language contact, creolization, and genetic linguistics.* Univ of California Press, 2023.

Youssef, Valerie and James Winford. Grounding via Tense–Aspect in Tobagonian Creole: Discourse Strategies Across a Creole Continuum. *Linguistics*, 37(4):597–624, 1999.

Appendices

A Traditional dialects in England

Regarding their level of non-standardness, traditional dialects in the British Isles in the *electronic World Atlas of Varieties of English* Kortmann et al. (2020) seem to have basilectal NORM (Non-Mobile Rural Male) speakers as their baseline. Furthermore, they can be classified as "those conservative dialects of English which are, for the most part, spoken in relatively isolated rural areas by certain older speakers and which differ considerably from Standard English, and indeed from one another" (Hughes et al. 2013: 34). Based on this profile of non-standard traditional dialects in the British Isles and expert judgement (p.c. David Britain, 05.06.2025), speaking a traditional English dialect is connected to a mixture of variables which include being born locally and having a basic level of education (possibly GCSE but usually not higher), being around a certain age (possibly 45 years of age or older), and probably having strong local roots which can be approximated via national identity. The acquisition of these traditional dialects by outsiders, if this happens at all, depends on several variables. Precisely, one prerequisite is that they are born in the UK and that they are exposed and in contact with a substantial number of dialect speakers (p.c. David Britain, 05.06.2025). It is further reasonable to assume that outsiders must have strong local roots, i.e. connect with the local identity, and frequently have a working class background/occupations. Against this background, speakers of traditional dialects were approximated through various combinations of census variables, depending on their availability. The most restrictive approximation was always applied.

A.1 Native speakers

- East and West Midlands

  Census variables (above age of 45, no education beyond GCSE/age 16):

  - Country of birth (3 categories): Born in the UK
  - Main Language (11 categories) = English or Welsh
  - Age (8 categories): Aged 45 or above

  - Highest level of qualification (7 categories): Level 1 and Level 2 qualifications, no qualification
  - 2,154,916 (Total L1/UK born: 8.85 m). The total population in the Midlands is 10.8 million.

- Southwest

  Census variables (above age of 45, no education beyond GCSE/age 16):

  - Country of birth (3 categories): Born in the UK
  - Main Language (11 categories) = English or Welsh
  - Age (8 categories): Aged 45 or above
  - Highest level of qualification (7 categories): Level 1 and Level 2 qualifications, no qualification
  - 1,171,642 (out of total L1/UK born: 4.95m)

- North: North East and North West

  Census variables approximating working class speakers are not available. Census variables used are for identity as proxy for strong local ties:

  - Country of birth (3 categories): Born in the UK
  - Main Language (11 categories) = English or Welsh
  - National Identity (9 categories) = English only OR British and English only
  - 3,071,107 (Total L1/UK born: 8.67m)

- North: Yorkshire and the Humber

  Data on Identity is not available. Census variables approximating working class speakers are used instead:

  - Country of birth (3 categories): Born in the UK
  - Main Language (11 categories) = English or Welsh
  - Age (8 categories): Aged 45 or above
  - Highest level of qualification (7 categories): Level 1 and Level 2 qualifications, no qualification
  - 1,126,060 (out of Total L1/UK born: 4.66m)

- Southeast

  Census variables approximating working class speakers:

  - Country of birth (3 categories): Born in the UK
  - Main Language (11 categories) = English or Welsh
  - Occupation (current) (10 categories): Skilled trades occupations, Caring, leisure and other service occupations, Process, plant and machine operatives, Elementary occupations

  - 1,142,549 (total L1/UK born: 7.49m)

- East Anglian English (here: Norfolk, Suffolk, and Cambridgeshire)

  No proper data available at this level due to data protection.

  - Country of birth (3 categories): Born in the UK
  - Highest level of qualification (7 categories): Level 1 and Level 2 qualifications, no qualification
  - 760,530 (out of Total population ca. 1177843)

A.2 Non-native speakers

Non-native speakers were approximated through various variables for all of the traditional dialects listed below, unless otherwise specified: Country of birth (3 categories = Born in the UK) by national identity (9 categories, selected: "English only identity", "English and British only identity"), current occupation (10 categories, selected: "Skilled trades occupations", "Caring, leisure and other service occupations", "Process, plant and machine operatives", "Elementary occupations") by English proficiency (4 categories, selected: "Main language is not English (English or Welsh in Wales): Can speak English very well or well") by region. In this case, less proficient speakers were excluded.

- Midlands
- Southeast
- Southwest
- North West and Yorkshire and the Humber
- North East

  Less restrictive data not available due to data protection. Non-native speakers were approximated through various variables (approximating identity)

  - Country of birth (3 categories = Born in the UK)
  - national identity (9 categories, selected: "English only identity", "English and British only identity")
  - English proficiency (4 categories, selected: "Main language is not English (English or Welsh in Wales): Can speak English very well or well"). In this case, less proficient speakers were excluded.