



ARTICLE

Identifying the dialectal background of American Finnish speakers using a supervised machine-learning model

Ilmari Ivaska¹ , Mirva Johnson²  and Tommi Kurki¹

¹Department of Finnish and Finno-Ugric Languages, University of Turku, 20014, Finland and ²Department of German, Nordic, and Slavic+, University of Wisconsin-Madison, Madison, WI 53705, USA

Corresponding author: Ilmari Ivaska; Email: ilmari.ivaska@utu.fi

(Received 11 August 2022; revised 19 April 2023; accepted 5 May 2023)

Abstract

This study presents results of two experiments using supervised machine-learning models to examine individual Finnish speakers' dialectal backgrounds. Data come from interviews conducted with heritage speakers of Finnish in northern Wisconsin and are compared to data from the Finnish Dialect Syntax Archive. The models were constructed and then, following successful validation testing, used to identify the dialectal background of five individual American Finnish speakers. Results showed individual variation in dialectal backgrounds and some correlation to speakers' likely language input. Our approach offers a new methodological tool for examining speakers' dialectal backgrounds in situations of language contact.

Keywords: dialect; Finnish; heritage language; language contact; supervised machine-learning

1. Introduction

American Finnish developed as a contact variety with English after people migrated from Finland to North America in the early twentieth century, and as their descendants acquired Finnish as a heritage language. Earlier research suggests that the dialectal background of speakers' parents or grandparents is often identifiable in their Finnish, though their speech still recognizably differs from Finnish spoken in Finland (Jönsson-Korhola & Lindgren 2003:408–409). The concept of a base dialect from which an immigrant language later develops after individuals migrate is difficult to establish due to the number of variables involved, including potential for varied input across a range of registers and dialects. Studies of moribund heritage language speaking groups, or groups in the final generation of transmission, often make comparisons between the region from which a community's original settlers emigrated and the dialect common in the region at the time (Bousquette & Putnam 2020:202). Some studies have examined the extent to which specific dialect features

© The Author(s), 2023. Published by Cambridge University Press on behalf of The Nordic Association of Linguists. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

endure in the speech of later generation heritage speakers of German (Litty, Evans & Salmons 2015), Norwegian (Hjelde 2015), and Spanish (Otheguy, Zentella & Livert 2007), and suggest that in some instances dialectal variation and language change interact with the regional English, resulting in regionally distinctive minority language dialects. However, there are still difficulties in determining the base dialect from which these new varieties develop, in the same way that researchers face challenges identifying homeland language varieties to use as baselines for comparison in the study of heritage languages (Polinsky 2018:10–17). This study offers both initial analysis of data from heritage speakers of Finnish, as well as another tool for approaching the question of how to identify an individual speaker's dialectal background. This methodology can be used and built upon not only to estimate the dialectal make-up of a speaker's speech, but the dialects they likely received as input. Knowing the kind of input that a heritage speaker of a language receives has important implications for better understanding the features that an individual speaker may have due to the setting of acquisition (Pascual y Cabo & Rothman 2012:451). By extension, understanding variation in individuals' dialectal background within a community can inform understanding of the language spoken on the community level.

In this paper we use computational methods to examine the Finnish dialectal background of five individual speakers of American Finnish in Wisconsin. Building on previous quantitative approaches to Finnish dialects, we categorize a speaker's dialect by simultaneously comparing their usage of multiple linguistic features identified in earlier research on Finnish dialects. Contrary to previous studies, we define usage of different features in distributional terms: we look for each feature (e.g. the singular first person pronoun) and the degree to which individuals use variants of that feature (e.g. *mä/mie/miä/minä*). Hence there are no categorical distinctions in the dialectal traits; a speaker may instead have a larger proportion of variants characteristic to dialect X along with some variants characteristic to dialect Y. This reflects what speakers and researchers reported as indicative of Finnish in North America: the dialect of a speaker's parents or grandparents as a base but with varying degrees of influence from surrounding Finnish and English speakers.

Our research questions are: (i) How does the language of American Finnish speakers relate to the regional variation of spoken Finnish in Finland? (ii) How does their language relate to the locations from which speakers' families migrated? To address these questions we run two experiments. In the first experiment, we construct a model that predicts the probability of an individual speaker belonging to the different dialect groups. In the second experiment, we construct another model that places an individual speaker on a map of Finland based on transcripts of their speech. Both models are trained on data from the Finnish Dialect Syntax Archive (in Finnish, Lauseopin arkisto, henceforth LAX; see Ikola, Palomäki & Koitto 1989 for an overview). Notably this approach allows for the individual sociolinguistic backgrounds of speakers to be taken into account in the analysis and considered within the scope of the speakers' setting of acquisition as a heritage language.

This article is structured as follows. Section 2 grounds our approach with an overview of previous work on Finnish dialects, background on the study of heritage speakers of a language, and American Finnish as a contact language. Section 3 introduces the datasets and highlights the features examined in the models



Figure 1. Finland’s dialect areas: (1) Southwest, (2) Southwest transitional, (3) Häme, (4) South Ostrobothnia, (5) Central/North Ostrobothnia, (6) Far North, (7) Savo, (8) Southeast. The dark-grey areas are predominantly Swedish-speaking. Map by Tommi Kurki.

constructed. Sections 4 and 5 present results and discussion of the experiment, followed by concluding remarks.

2. Theory and literature review

2.1 Previous study of Finnish dialects

Finnish dialects have traditionally been divided into eastern and western varieties (Rapola 1969, Wiik 2004) and further subdivided into eight dialects (Itkonen 1964, 1989). The traditional eight dialect categories include Southwest, Southwest transitional, Häme, South Ostrobothnia, Central/North Ostrobothnia, Far North, Savo, and Southeast. See Figure 1 for their geographical distribution.

These dialect groups have been distinguished based on the distribution of a range of linguistic features. One of the more well-known surveys of Finnish dialects was conducted by Lauri Kettunen, later compiled and spatially mapped into his *Dialect Atlas of Finland (Suomen murteet: Kettunen 1940a,b)*.¹ Many of the more dramatic dialectal differences leveled off by the nineteenth century and further leveling has occurred since. Several studies have applied quantitative methods to examine Finnish dialects and their collective variations (Wiik 2004, Leino,

Hyvönen & Salmenkivi 2006, Hyvönen, Leino & Salmenkivi 2007, Leino & Hyvönen 2008). Most recently, modern population genetic approaches to dialectal variation have confirmed the computational soundness of the traditional eight-dialect categorization (e.g. Syrjänen et al. 2016).

We build on this quantitative work confirming the eight-way distinction and use it to categorize and map the speech of individual speakers. Previous approaches use aggregated survey data from Kettunen's Dialect Atlas (1940a) as input and basis for the dialectal distinctions, rather than the speech of individuals (e.g. Honkola 2016, Syrjänen 2021). The Atlas only reports the most frequently used variant of a feature per municipality, rather than capturing the full range used in a municipality and even within a particular dialect. While additional discussion of feature variation is given in Kettunen's explanation for the Dialect Atlas (Kettunen 1940b), this additional variation is for the most part not included in recent quantitative approaches to dialects. This means that studies using the Atlas as a foundation are restricted from examining social variables of speakers, as responses were not recorded for individuals (see Honkola 2016:27). We build on these previous approaches and present a model that examines the speech of individuals and defines dialect categorization in distributional terms. This is more in line with some other quantitative approaches to the study of Finnish in contact settings by Lainio (1989), whose model similarly included multiple possible realizations of dialect features per dialect for categorization. This allows for the natural complexity of the speakers' language to be captured and more overlap between dialect features to be taken into account.

By using a distribution-based model and training it using actual speech data from the LAX corpus, natural inter-speaker and intra-speaker variation is inherently included in our model, and is thus more suitable for the study of Finnish in the North American context. The model that serves as a point of reference for the regional variation of spoken Finnish is built on actual speech samples from multiple individuals, and so this approach further allows for social variables to be considered for individual speakers. This is an important factor for the study of American Finnish, a contact language that speakers mostly acquired as a heritage language.

2.2 Heritage languages and speakers

A heritage language broadly defined is a language that an individual culturally identifies with because of a family or community tie (Polinsky & Kagan 2007:369). A heritage speaker of a language is narrowly defined as an individual raised in a home where a language other than the dominant community language was spoken, later resulting in some degree of bilingualism in the heritage and societal majority language – whether balanced or (usually) unbalanced (Scontras, Fuchs & Polinsky 2015:3). Many heritage speakers do (and did) not receive formal, written instruction in the heritage language (especially in the North American context), and thus the setting of acquisition, frequency of usage, and exposure to different registers and dialects has an impact on which language features they use most frequently in their own speech. Polinsky & Kagan (2007) suggest that heritage speakers can be identified on a continuum of proficiency in order to best capture the nuances and differences between speakers' abilities and contexts of acquisition (2007:370–372).

The nature of the input that heritage speakers receive is inherently different from that of speakers acquiring the language in a majority language context (Pascual y Cabo & Rothman 2012:451). This raises the complex question of what are appropriate baselines for comparison in the study of heritage languages. Polinsky (2018) discusses nuances of the issue suggesting that depending on the community and research questions, age-matched homeland speakers from the time of emigration or balanced bilinguals within the speech community may be appropriate alternatives, though there may also not be a relevant baseline available for comparison, as is often the case for endangered languages with few remaining speakers (2018:10–17; 329–333).

2.3 American Finnish as a contact variety in northern Wisconsin

American Finnish developed as a contact language in the generations after Finnish-speaking individuals migrated to North America in the early twentieth century. It contained influence from contact with English, but also features common in the Finnish dialects spoken in the places from which individuals first migrated. The majority of migrants came from Ostrobothnia and other western municipalities, meaning many also spoke Swedish or otherwise had traces of Swedish contact in their speech. Children of Finnish migrants born in North America grew up hearing Finnish spoken by their parents and within their smaller communities, but not used in official contexts or taught in public schools – common for minority languages past and present. While there are many similarities across speakers of Finnish in North America, this is not a formalized speech variety. Despite some general tendencies, there are many local and individual variations, especially in lexicon and pronunciation (Jönsson-Korhola & Lindgren 2003:409).

Researchers who have interviewed and spoken with many American Finnish speakers comment that the Finnish spoken by Finnish-Americans tends to have the dialect of a speaker's parents or grandparents as foundation, but also clear influence from both English and the Finnish spoken by other speakers in their community (Jönsson-Korhola 1993:110; Jönsson-Korhola & Lindgren 2003:408–9; Männikkö 2004:4). For this reason, while there are similarities across many speakers of American Finnish, there are also differences that tend to be individualistic. Many variations are due to different amounts of regular exposure to Finnish, as well as the locations from which a speaker's parents or grandparents first migrated (and the related dialect that they spoke). American Finnish has been documented and studied by researchers focusing on contact effects (Larmouth 1974, Hirvonen 1992, 2005, Virtaranta et al. 1993), as well as code-switching and borrowing (Poplack 1980, Lauttamus 1991, Halmari 1997, Männikkö 2004). There have been few studies on American Finnish using quantitative approaches (aside from Poplack, Wheeler & Westwood 1989 and Hirvonen & Lauttamus 2000), mostly because the bulk of the fieldwork was conducted before such models and approaches were being widely used.

Observations about the grammatical structure of American Finnish tend to follow patterns seen in heritage languages across different types of contact settings. While some English influence is present, speakers tend to have more standard-like usage of cases (especially directional cases) than those of typical learners of Finnish

as an additional language (Martin 1993:97). Larmouth reports that directional cases (illative, elative, inessive, ablative, adessive, and allative) were preserved on nouns and pronouns in the speech of second and third generation Finnish Americans in Minnesota, while use of the partitive and accusative cases varied (Larmouth 1974:365–366). This early observation of the endurance of case marking is in line with recent work that contradicts stereotypes of heritage speakers as losing case marking. Łyskawa & Nagy (2020) observed >90% normative case usage across heritage speakers born in Toronto of Russian, Polish, and Ukrainian parents in data collected in a naturalistic setting (2020:148). Other work suggests processes of restructuring of the dative and accusative cases in multiple speaker communities of heritage German, rather than explicit loss (Yager et al. 2015; Bousquette 2020:491–492). The endurance of case marking and other features suggests that while American Finnish may differ noticeably from Finnish spoken in Finland, the language is nevertheless similar enough that case marking, consonant gradation, and other classic features of Finnish dialects can reasonably be expected in the speech of fluent or comfortable speakers, and to a greater degree than with typical non-native speakers of Finnish.

American Finnish is one of many language contact varieties spoken in Wisconsin during the nineteenth to twentieth centuries, and previous work on other immigrant languages in Wisconsin has indicated identifiable homeland dialect features, as well as variation therein. Previous research on Wisconsin German speech communities established in the early 1800s has shown that individual German dialects are not especially identifiable in individuals' speech; rather, there is a mixing of distinct regional and standard features in a way that would not be found in a European setting, and have become a regional marker for the area (Litty, Evans & Salmons 2015:184). This was in contrast to previous work on German, which assumed retention and preservation of dialects from the time of migration (e.g. Eichhoff 1985:234). This mixing is likely due to natural language change over time, whereby languages and dialects spoken in contact situations lose much of their original variation as a 'leveling', or reduction of marked variants, occurs, yet some features are maintained if they are in some sense 'simpler' (Trudgill 1986:98). It is possible that the setting of acquisition of German as a heritage language in later generations played a role in the mixing of standard and regional features; some speakers learned to read and write standard German yet spoke Low German or other variants within the home (Bousquette 2020:486–489).

Haugen (1969) examined differences between home and community dialect features in Norwegian spoken by immigrants and their descendants in the Midwest from the 1930s onward (1969:337–360; 26). Haugen notes that some Norwegian Americans had difficulty understanding differing Norwegian dialects spoken by neighbors and spouses, contributing to a switch to English for some (1969:349). While many dialect features were preserved in the speech of later generation American Norwegians, there was still a leveling of some marked variants, evidence of contact with English, and on the whole, dialect preservation could only be assessed 'for certain features . . . and for certain individuals, rarely if ever for whole communities' (1969:360). However, often there was still a community influence on language norms to some degree. Natvig (2022) outlines how local business and church services were conducted in the community's Norwegian heritage language in Ulen, Minnesota, until language shift to English had occurred on a community-wide

level. Examining linguistic features in heritage communities at different stages of language shift (with corpora from different time periods), in conjunction with analysis of changing social patterns and rates of language shift, is another potential application for this model.

The situation of American Finnish differs from Wisconsin German and Norwegian in that there were fewer generations over which the language was maintained. The settlement of Finnish communities happened later, and most Finnish-speaking communities in northern Wisconsin were smaller than the state's German-speaking communities and had a comparatively shorter period of bilingualism (Johnson 2022:25–26). For heritage speakers of the second and third generations, their own family or co-workers often provided the majority of their Finnish input. This suggests there was not a sufficiently long period of community-wide bilingualism for a leveled, regional dialect of the minority language to develop, though there may have been some community-level norms and influences. Importantly, this study is examining the dialectal background of individual American Finnish speakers rather than suggesting a cohesive American Finnish dialect amongst these speakers.

3. Data and methodology

3.1 Data 1: Heritage languages and bilingualism in Wisconsin

Primary data for this study come from five interviews conducted in 2016 and 2017 with speakers of American Finnish in northern Wisconsin. The speakers are from three rural communities established around the same time and with similar population sizes over time, economic bases in farming, and timelines of language shift. Interviews consisted of personal history questions related to language usage, picture identification tasks, the 'Frog Story' narrative task, sentence translations, and some free conversation. Interviews each lasted about 50–90 minutes. Interviews were transcribed into ELAN and variants coded for the dialect features under study, first using an automated R script and then hand-checked. These particular speakers were chosen because of their minimal Finnish language instruction (as children or adults) and the degree of comfort they expressed in conversing in Finnish.

Table 1 presents each speaker's background and their reported connection to Finland. We consider the first generation as born in Finland and relocated after the age of 10. Those of the second generation have at least one parent who migrated from Finland during adulthood. Those in the third generation have at least one grandparent who migrated from Finland as an adult. Section 3.1.1 offers more detailed discussion of each speaker's background.

3.1.1 Speaker biographies

Family background and social context of language acquisition played a role in the variety of Finnish these individuals acquired. This section presents speaker biographies (using pseudonyms) that highlight the variability of individual speakers' language input over the course of their lives while still maintaining anonymity. All speakers were over the age of 65 when interviewed, and reported using Finnish more frequently in their youth than they did at the time of the interview. Each had

Table 1. Background of American Finnish speakers

Speaker (pseudonym)	Reported generation	Reported ancestry (city)	Reported ancestry (dialect area)
Marie	2nd generation	Jalasjärvi, Eurajoki	South Ostrobothnia, Southwest
Laura	2nd generation	Kurikka, Lappajärvi	South Ostrobothnia
Gerry	2nd generation	Jalasjärvi, Ylistaro	South Ostrobothnia
Don	3rd generation	Ylistaro, Alavus, Isokyrö, Kuortane	South Ostrobothnia
Ron	3rd generation	Kuortane, Rantsila	South Ostrobothnia, Central/North Ostrobothnia

been born and raised in northern Wisconsin and most had little occasion to read Finnish, reporting difficulty reading most texts.

Marie. Marie's parents came from Jalasjärvi and Eurajoki (in the South Ostrobothnian and Southwest dialect regions). She was born in 1940 and did not speak English until starting school. She has not taken formal classes, though spent a few weeks volunteering at a Finnish language summer camp in her thirties. She has traveled to Finland twice for short visits, both times in later adulthood.

Don. Don's parents were born and raised in northern Wisconsin, but his grandparents were from western and northern Finland (Ylistaro, Alavus, Isokyrö, Kuortane within the Savo, Häme, and South Ostrobothnian dialect regions). He was born in 1948 and spoke mostly Finnish in his seasonal job as a laborer (from the age of eighteen into his early twenties). Don has never been to Finland but tries to speak Finnish whenever possible.

Gerry. Gerry was born in 1935 and his parents both migrated from near Vaasa. Gerry worked on the family farm as an adult and his mother lived nearby; she did not speak any English and Gerry's spouse (with no Finnish background) learned enough Finnish to help her with shopping. Gerry did not start working outside the home until he was thirty-five years old. He visited Finland a couple of times as an older adult and occasionally speaks Finnish with his cousins on the phone.

Ron. Ron was born in 1935 and his parents were both born in Wisconsin. They rarely spoke Finnish to Ron, but often spoke Finnish with each other. Ron learned Finnish while living with his grandmother. He often spoke Finnish with others while working (in his twenties and thirties). He also spoke Finnish with his spouse's parents, who lived on the neighboring property and had migrated from Jalasjärvi and Eurajoki. He traveled to Finland on two occasions in his forties.

Laura. Laura was born in 1931. Her father migrated from Lappajärvi and her mother from Kurikka, both in the province of Vaasa. She recalls taking a six-week Finnish language course as an adult. She briefly worked in a job where she needed to read and write in Finnish, but was in that job for less than a year. She spoke mostly Finnish with her spouse and traveled to Finland once as an older adult for a few weeks. She still occasionally calls her cousins in Finland.

3.2 Data 2: Finnish Dialect Syntax Archive

Models in this study use data from the Finnish Dialect Syntax Archive (LAX) as input. These data consist of 171 transcribed interviews annotated manually in terms of morphology and syntax, with informants from across Finland (1,037,999 tokens). Most of the data was recorded in the 1960s and the majority of informants were born between 1870 and 1890. This time period is roughly when the majority of the Wisconsin Finnish speakers' parents or grandparents were born in Finland. Many of them migrated to America as adults or in later childhood. This means that the Finnish which Finnish American informants heard and were most exposed to is from roughly the same time period as the speech recorded in the LAX corpus. This particular collection is thus a reasonable comparison, and more similar to the speech that American Finnish speakers would have had as input, compared with more recently collected interviews. The analysis makes use of the version provided by the Language Bank of Finland (Kielipankki) via the Korp user interface.²

3.3 Dialect features examined in model

While some Finnish language features are considered neutral and widespread, others are linked to particular region(s) (Mantila 1997, 2004; Nuolijärvi & Sorjonen 2010). An example of a neutral, widespread feature is final vowel deletion and deletion of *i* from unstressed diphthongs in words such as *puna(i)nen* 'red' (Mielikäinen 1986:234). Some features that were originally regional then spread to become used widely in spoken language, such as using the passive verb form with the first person plural pronoun (Mielikäinen 1986:235). These features can be considered part of a general spoken language (in Finnish: *yleispuhekieli*). Importantly, much of the movement that supported development of this general spoken language occurred from the 1950s to the early 1970s. This was a period of significant emigration and rapid rural depopulation as over 600,000 people left farming for newly created industry jobs in cities. Thus the language situation was relatively stable throughout the 1950s, but by the 1960s and 1970s there was rapid movement from rural to urban centers. This is notably after the time period in which individuals were interviewed for the LAX corpus. Most of those interviews were conducted in the 1960s and targeted older, non-mobile, rural individuals. This means that the speech documented was prior to rapid change, and captured examples of what had been relatively stable regional dialects.

To examine the dialect background of American Finnish speakers, we categorized the speech of speakers into the traditional eight dialectal categories based on the presence or absence of particular dialect features. We considered Finnish dialects in distributional terms, meaning that several of the classic features were assessed along with the degree to which they each contributed to the speaker's dialect make-up. In practice, this means that a speaker's dialect identification using the model is based on the consistency and frequency with which they use particular dialect variants over other varieties. We chose specific dialect features for the model based on their usage in previous studies (Rapola 1969, Itkonen 1989, Lainio 1989, Wiik 2004, Nuolijärvi & Sorjonen 2010), but also considered pragmatic factors which impacted the feasibility of coding – both within the data and in

Table 2. Distribution of variants of inessive case forms

	-ssA	-sA	-ss	-s	-hnA
Southwest		X		X	
Southwest transitional			X	X	
Häme	X	X	X	X	
South Ostrobothnia		X		X	X
Central/North Ostrobothnia		X	X		
Far North	X	X	X		
Savo	X		X		
Southeast	X		X	X	

operationalizing the features so that they could be extracted from the LAX corpus in a fashion that makes comparison between the datasets possible.

Dialect features under study included those that are considered classic staples in the differentiation of Finnish dialects, spanning morphological, lexical, and phonological features. We included what are considered classic distinguishing features (such as /ts/ and /d/ variables, first person pronouns, and intrusive schwa vowel) and added additional features that helped improve the accuracy of the model. The model does not include some features that rely heavily on transcriber judgment or otherwise would have been difficult to operationalize between datasets (such as other lexical items, as well as diphthong opening and reduction). For each feature, every possible realization in a speaker's transcript was identified and coded.

The rest of this section outlines the features used in the model and their realizations across different dialects. Importantly, we use classifications common for the dialect regions in the 1960s and prior to the post-urbanization movement which contributed to rapid language change and dialect leveling. Section 4 addresses the usage of the features by individual speakers, especially those which contributed most to the model's overall classifications.

3.3.1 Morphological

Inessive case form. While the standard form *-ssA* is attested across all varieties of spoken Finnish today, it was unattested in some regions in the 1960s. In Finnish, it is one of six locative cases and used to indicate place, as in example (1), or attachment, as in example (2), as well as some expressions of time, seen in example (3):

- (1) *Asun Suomessa.* 'I live in Finland.'
- (2) *Sormus on sormessa.* 'The ring is on the finger.'
- (3) *Luin kirjan kahdessa päivässä.* 'I read the book in two days.'

Table 2 indicates the distribution of usage of each variant, showing how the same variant may be found across multiple different dialects. The R scripts, anonymized

datasets, and distribution tables for all dialect features included in the study are available in an online repository.³

The variants of this case form with one *-s* as well as with final vowel apocope are considered identifying features of western dialects. The deletion of the final vowel is considered a feature of the Southwest and South Ostrobothnian dialect groups. The variant *-hnA* refers to an ending found only in the South Ostrobothnian dialect on words with a possessive suffix such as *talohmani* for *talossani*, ‘in my house’. This particular usage was not attested in our data from American Finnish speakers.

Verb plural forms. The first person plural verb suffix in Standard Finnish is *-mme*. While this is often used in spoken language, the final vowel is sometimes *a/ä* (an alternation henceforth indicated as *A*) in some western dialects. Furthermore the passive is often used in spoken Finnish more broadly, resulting in sentences like those in examples (4a) and (4b).⁴ In sum, the possible verb plural forms are *-mme*, *-mmA*, *-VVn*.

- (4) a. Me menn-ään kaupp-aan. = Mene-mme kaupp-aan.
 we go-PASS store-ILL go.pres-1PL store-ILL
 ‘We are going to the store.’
- b. Me olt-iin sie-llä. = Oli-mme sie-llä.
 we be-PASS.past there-ADE be.past-1PL there-ADE
 ‘We went there.’

3.3.2 Lexical

First person pronouns (singular and plural). The first person pronouns are classic features and their variants serve to differentiate the northern, eastern, and western dialects. In the majority of spoken Finnish, the standard *minä* is used for the first person pronoun, as well as the shortened form *mä*. In addition to this form, *mie* is often found in the Southeast and North dialects while *miä* is attested in some parts of the Savo and Southeast dialects. Thus the four forms coded for were *minä*, *mä*, *mie*, and *miä*.

The plural first person pronoun *me* is common across dialects and Standard Finnish alike. In addition, the form *myö* is attested in the Central/North Ostrobothnian, Savo, and Southeast dialects while *met* is a feature of the North dialect.

3.3.3 Phonological

Consonants: /d/ variable. The /d/ variable has played a central role in distinguishing between eastern and western dialect areas. The consonant alternates between *d*, *ð*, *r*, *l*, *j*, and deletion. We group *j* with deletion as they are part of the same change process, and both are prevalent in the North, Savo, and Southeast dialects. The *l* variant is distinctive of the Häme dialect.

Consonants: /ts/ variable. The /ts/ variable has played a similarly key role in accounts of dialect differentiation and has a wide range of realizations. The variants included in the model include *ts*, *tt:tt*, *tt:t*, *ht*, *ss*. The *tt:t* variant is found in the Southwest dialects and in some of the Southeast dialects and indicates a singular *t* instead of *ts*

in places where the weak grade of consonant gradation occurs. The *ss* form is mostly attested in the Southeast dialect. Of most importance for the present data, *ht* is found predominantly in the Central/North Ostrobothnian and Savo dialects.

Vowels: word-final -eA. There are at least ten variants attested of the word-final, unstressed vowel, four of which are sufficiently widely attested to be a valuable feature for the model. Examples of words with word-final *-eA* to which these changes apply are *korkea* ‘high/tall’ and *pimeä* ‘dark’. While *-eA* is present across standard language and dialects alike, *-iA* is realized in all dialects except those in the southwest. The form *-ee* is attested in the Savo, Southeast, Häme, and Southwest/mixed dialects; *-i(i)* is found in the Savo and Southwest dialects.

Vowels: word-final -OA. The distinction between word-final, unstressed *-OA* and *-UA* is a matter of a widening of difference between the two vowels in the cluster. Common examples include *maitua* for *maitoa* ‘milk’ and *pallua* for *palloa* ‘ball’. The fronted *-UA* variant is attested only in the North and Central/North Ostrobothnian dialects and is a distinguishing factor of the region.

Schwa and schwa following /h/. This dialect feature consists of the addition of a vowel between consonants of different places of articulation and is considered distinctive to the North, Savo, and Ostrobothnian dialects. In Finnish, this is almost exclusively following */l/* such as in words *jalaka* for ‘foot’ and *melekein* for ‘almost’. However, a notable exception is in the *nh* cluster in words like *vanaha* for ‘old’. A related feature with slightly different coverage is the insertion of a schwa following */h/* in words like *lehemä* and *ihiminen*. This is particularly prevalent in the North, Central/North Ostrobothnian and South Ostrobothnian dialects (rather than Savo). In the data, possible instances of schwa were coded as either present (+) or absent (–).

We calculated the overall frequency of each of the feature variants discussed above and divided that value by the overall occurrences of the said feature in an individual speaker’s speech sample. For example, if a speaker uses the *-ssA* variant of inessive fifteen times, the *-ss* variant three times and does not use any of the other variants, the values are $= 15/18 = 0.833$ and $1/18 = 0.167$ for these two variants, and 0 for all the other variants of the inessive case. The resulting data include each informant as one observation and each feature variant with a value between 0 to 1 as a predicting variable, resulting in 35 predicting variables.

3.4 Methodology: Supervised machine learning using Random Forests

The methodological procedure comprises two sets of experiments, which, in turn, both have a validation experiment followed by an exploratory experiment focusing on the dialectal background of the heritage language speakers of Finnish. In the validation experiments we train two models. The first model predicts the dialectal background of individual speakers of Finnish based on their use of the above-discussed dialect features in their speech. The model outputs, for each of the eight dialect regions considered, the probability of the speaker hailing from that region. The second model, in turn, applies a double regression approach, as it predicts separately the longitudinal and latitudinal coordinates of each speaker’s dialectal

home location, based on the use of the same above-discussed dialect features. Hence the output comprises the coordinates for the speaker. Both models are trained and tested using data from the Finnish Dialect Syntax Archive (LAX). Then, provided that the thus obtained models fare relatively well in their respective tasks of the validation experiment, we use these models in the exploratory experiments to study the dialectal profile of five individual American Finnish speakers.

Both models are based on Random Forests (henceforth RFs) as a supervised machine-learning algorithm (for the algorithm, see Breiman 2001). RFs are essentially large ensembles of inference trees, and they can be used for either classification, where the goal is to sort the data at hand into classes of a categorical response variable, or in regression, so as to predict the value of a numerical response variable (for various linguistically oriented examples on the use of RFs, see e.g. Tagliamonte & Baayen 2012, Ivaska 2015, Deshors & Gries 2016, Ivaska & Bernardini 2020, Ivaska & Ivaska 2022). In RFs, the values of the predicting variables (here, the distributions of the studied dialect features) are used to split the data (here, individual speakers of Finnish) according to the given response variable (here, either the dialect region or the coordinates of their geographical location). For example, if speakers hailing from the Savo region consistently use the pronoun *mie* for the first person relatively more often than speakers from other regions, the model learns to use this information and uses it when splitting the data into different branches of the inference tree. Such trees are constructed by randomly choosing a subset of the studied features to find the best predicting variables. This random sampling is repeated many times to find those predicting variables that are most consistent in the task at hand. This way, the models rely more on the consistently well-predicting features and place less emphasis on the less consistent features. In RFs, these weights can be explored in terms of the variable importance measure, a metric that helps in understanding the degree to which different included variables (here, dialect features) contribute to the outcome of the obtained models. In the present study, we use the *ranger* implementation of RFs (Wright & Ziegler 2017). For the maps, we use the *rnatrualearth* and *rnatrualearthdata* packages (South 2017a,b), and for measuring the distances between the true and predicted locations, we use Haversine great circle distances as implemented in the *geosphere* package (Hijmans 2021).

4. Results

4.1 Validating the models to be used

The results of the validation experiments suggest that the classifier fares relatively well in assigning a speaker to one of the dialect regions according to their data. The overall classification accuracy for the dialectal background of the LAX data is 84% with a baseline of 20%.⁵ As can be seen in the confusion matrix in Table 3, the model struggles with speakers representing the northern Finnish dialects (in Finnish: *peräpohjalaiset murteet*) as well as those representing the Southwest/mixed dialects (in Finnish: *lounaiset välimurteet*). As far as the northern Finnish dialects are concerned, we believe that the poor accuracy stems from the limited data, as there are only nine speakers from the north of Finland. As for the

Table 3. Confusion matrix of the classification of the LAX informants

Predicted True	Häme	Far North	Central/North Ostrobothnia	Ostrobothnia South	Savo	Southeast	Southwest	Southwest transitional
Häme	31	0	0	0	1	0	3	0
Far North	0	2	3	0	1	0	3	0
Central/North Ostrobothnia	0	0	17	0	0	0	2	0
Ostrobothnia South	1	0	0	7	0	0	2	0
Savo	0	0	0	0	34	0	0	0
Southeast	0	0	0	0	1	15	3	0
Southwest	0	0	0	0	0	0	33	0
Southwest transitional	4	0	0	0	0	0	4	4

Southwest/mixed dialects, the results are actually in line with earlier research: the region has not been included in some of the traditional descriptions of Finnish dialects (e.g. Kettunen 1940a,b), and when it has been included, it has been characterized as a combination of features from the Häme and the Southwest/mixed dialects – the very groups the classifier erroneously suggests. All in all, notwithstanding these caveats, we believe that the classifier can be used reliably to approximate the dialectal background of individual speakers of Finnish.

Figure 2 shows the fifteen dialect features that contribute most to the model's outcome. As can be seen, the three most important features include the use of morpheme *-ssA* as the inessive case marker, the use of *minä* as the singular first person pronoun, and the use of schwa. As mentioned above, the *-ssA* variant was traditionally present in the dialect regions of Häme, North, Savo, and Southeast. As for the use of *minä* as the first person singular pronoun, it has typically been connected to Ostrobothnian as well as Savo and Häme dialects, whereas the use of the schwa vowel has been considered typical especially for North, Ostrobothnian, and Savo dialects. In other words, an individual speaker's preference (or dispreference) for these features is central to the model for predicting their dialectal background. The same logic applies to all features included in the model, but the measure of variable importance indicates how much weight it carries, in relation to other considered features.

As for the second task related to the geographical location of an individual speaker, this model is also relatively successful, with a median distance between actual and predicted location of 42 km. Figure 3 represents the actual recorded location of the informants in the LAX data, together with the predictions of each individual speaker. The discussed tendencies of the classification task are also visible here: predictions are in general relatively close to the true recorded location of the informants, but informants from northern Finland, in particular, are located further south. What is more, the map clearly shows that the predictions obtained by the model are typically somewhat conservative, in that they are invariably closer to the center of the map than the true location. In sum, though, it can be said that the

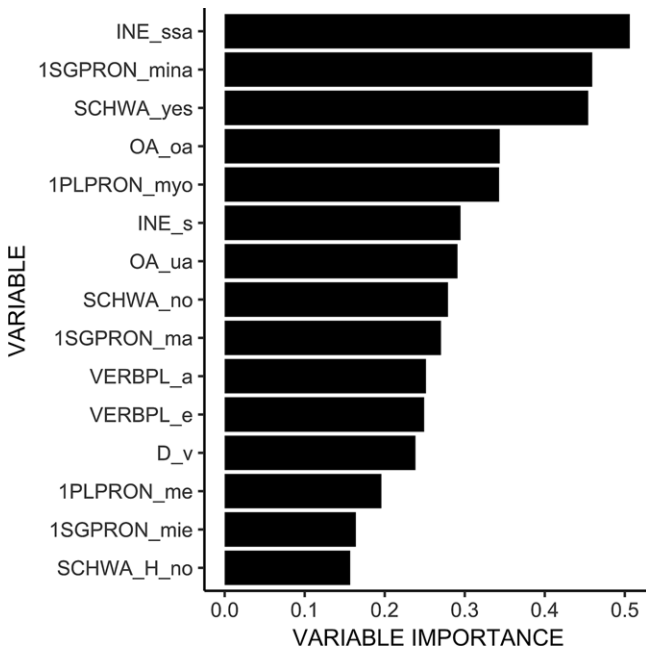


Figure 2. The fifteen most important dialect features for distinguishing speakers from different dialect regions. Abbreviations used in the model: 1PLPRON = first person plural pronoun; 1SGPRON = first person singular pronoun; D = /d/ variable; INE = inessive case; ME = standard first person plural pronoun; MINA = standard first person singular pronoun; OA = word-final -oa; SCHWA = schwa vowel; SCHWA_H = schwa vowel inserted following /h/; VERBPL = first person plural verb ending.

predictions do reflect the actual location of the informants relatively well and, hence, that this model can also provide useful information on the dialectal basis of individual American Finnish speakers.

When the fifteen most contributing features (see Figure 4) of the second experiment are inspected in greater detail, we can see that although their order and the relative importance of the features differ to some degree, fourteen of the features are the same as in the first model. It is thus safe to say that, with these data and the features included in these experiments, the most distinguishing features for a speaker’s dialectal background include the way they express the inessive case, the first person pronouns, the phonological representation of the word-final /OA/ of Standard Finnish, and the use of schwa, and to a lesser degree also the phonological representation of the Standard Finnish /d/.

4.2 Predictions of dialectal background and geographical location of American Finnish speakers

Following successful validation experiments, the model was then used to assess the dialectal background of American speakers in the Wisconsin data, as well as the geographical location in Finland that most closely corresponds to their speech. The results of the first experiment with the data from American Finnish speakers

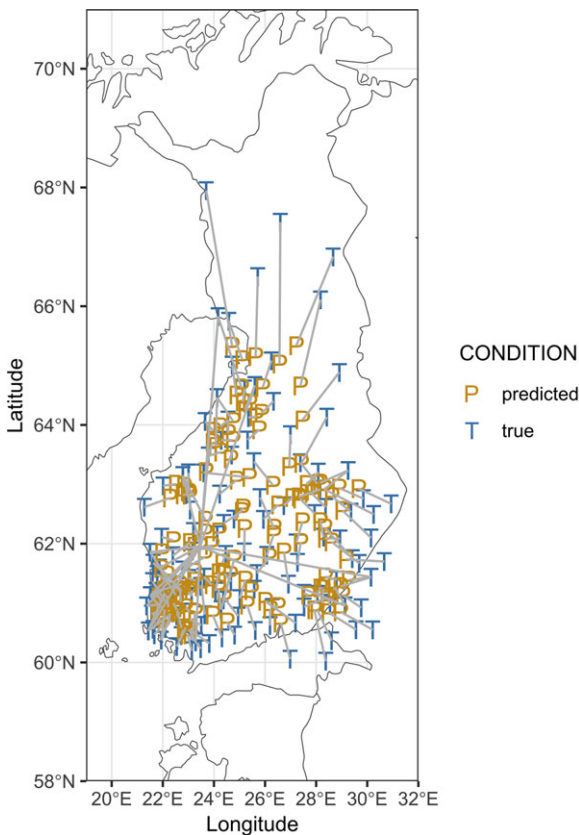


Figure 3. True and predicted locations of the LAX informants.

indicate that the speakers were identified as having a majority of features consistent with the Häme and Savo dialects, as elaborated in the qualitative analysis in Section 4.2.1. Results of the second experiment showed that American Finnish speakers could be located on a map of Finland, with the model placing them somewhat more to center at the baseline. In both models, the Savo and Häme dialects could be considered the most neutral dialects, as they tended not to be associated with marked variants for the dialect features under study. In practice, this means that unless a speaker used features considered marked by the model, their results were in the default location in the center of the map and with highest probability of the Savo and Häme dialects. The results of a hypothetical Standard Finnish speaker are given in Figure 5 as a point of reference for the model. The results for each American Finnish speaker should not necessarily be compared for their closeness to Standard Finnish, because American Finnish speakers for the most part were not actually exposed to Standard Finnish or formal education in Finland. Rather, this offers a point of comparison for better understanding of the high probability of Savo and Häme as an underlying baseline across results from the model. Note, however, that this is not to say that a high distribution of the Savo and

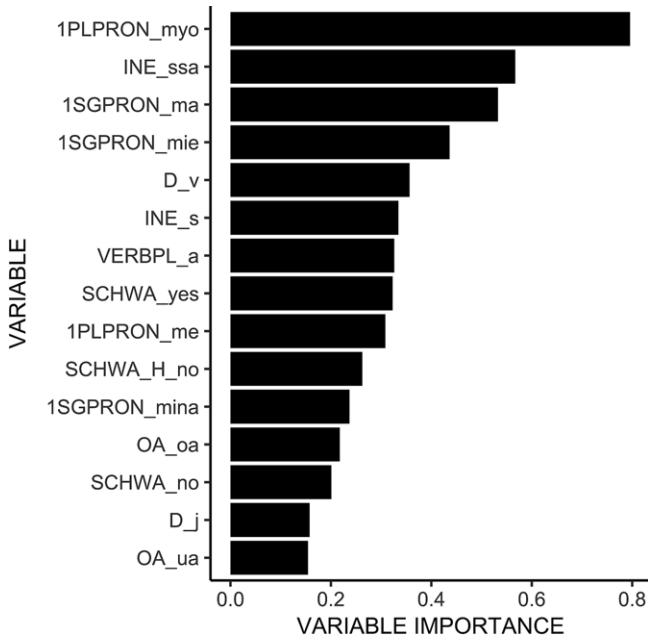


Figure 4. The fifteen most important dialect features when predicting a speaker's geographical location.

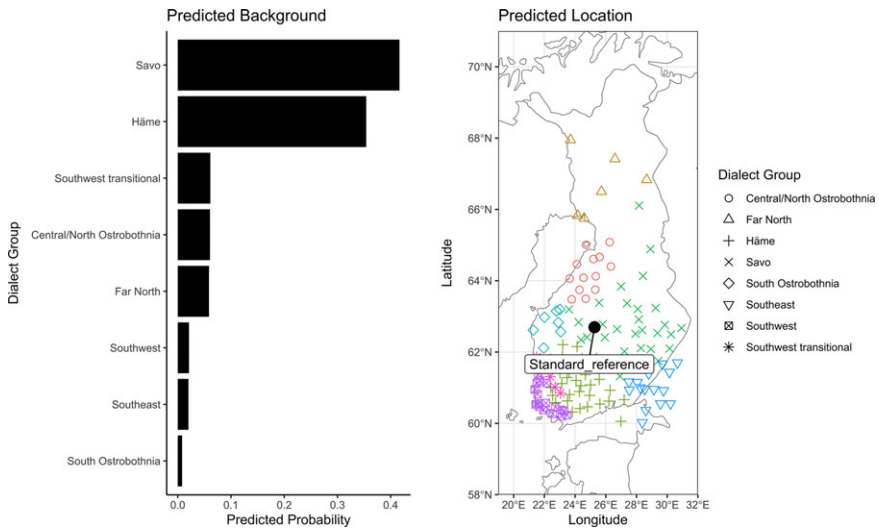


Figure 5. Standard Finnish speaker as a point of reference. The two T points in the Gulf of Finland are islands that belonged to Finland until WWII. The one further east is the island of Seiskari (coordinates 60.02302N, 28.37754E), and the one a little further west is the island of Suursaari (coordinates 60.055833N, 26.983889E). Both of these locations are included in the LAX data.

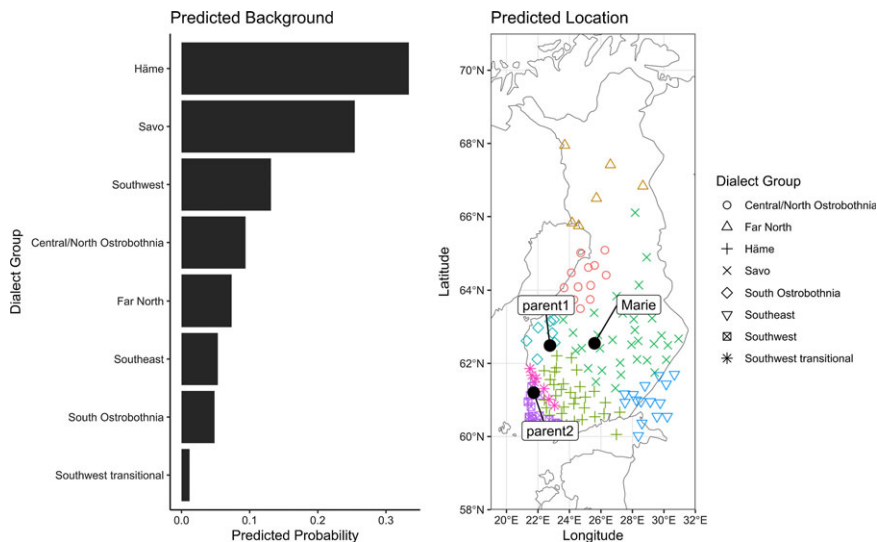


Figure 6. Marie's dialectal distribution.

Häme dialects would necessarily be due to usage or influence of Standard Finnish, but merely that they commonly make use of the same variants.

The rest of the section offers an overview of each informant's language usage, focusing on observations about their use of dialect features of highest significance in the model, as well as their ancestry. The distribution of dialectal background probabilities suggest that individual differences are evident and traceable within the data.

4.2.1 Qualitative analysis of individual speakers

Marie. Marie's dialect was categorized as mostly Häme and Savo, but also with twice as high a percentage of Southwest features compared to other speakers (Figure 6). The Southwest features are clearly seen with the frequency of *-iA* for /eA/ and *-s* for the inessive case in her speech. She furthermore had an instance of *myö* and mostly deleted instances of the /d:/ /d/ variable. Her frequent use of schwa following /h/ (present nine times out of a total of twelve) likely also contributed to her higher probability of a Southwest dialect background. This prediction is in line with her own family background. One of Marie's parents was from Eurajoki, near the traditional boundary of the Southwest dialect.

Laura. Laura's dialect was identified as mostly Savo, but with twice as high a categorization of the Central/North Ostrobothnian dialect compared to other speakers (Figure 7). This was likely in part because of her frequent use of schwa, *UA* for *OA*, and an instance of *myö*. She furthermore used *-mmA* for the first person plural verb suffix the majority of the time and frequently inserted the schwa vowel. She had a parent from closer to the Central/North Ostrobothnian dialect area, which perhaps influenced her dialectal background.

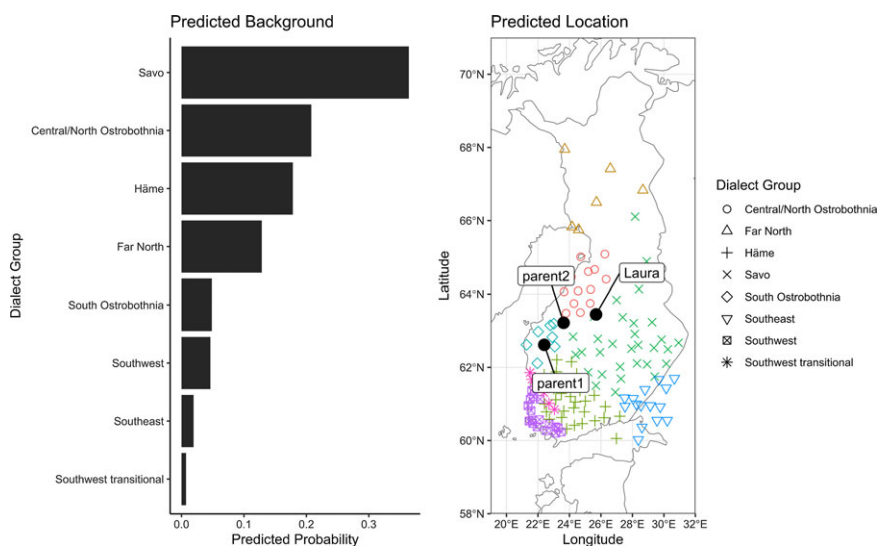


Figure 7. Laura's dialectal distribution.

Gerry. Gerry was identified by the model as having the highest probability of a Häme dialect by a fairly wide margin, with Southwest transitional as second highest (Figure 8). This may be in part because two of the features contributing most to the Häme dialect were the standard form of the first person pronoun, the standard form of the first person plural verb suffix (*-mme*), and not having an added schwa vowel. For each of these three features, Gerry used the standard forms in about 50% of all possible occurrences, a more even division than observed with the other speakers. On the whole he used few marked variants aside from *-iA* for *-eA* and *tt* for *ts*: in about half of all possible occurrences. Unlike Marie and Laura, Gerry's family background did not seem to have as much evident influence on his speech within this sample; his parents both migrated from South Ostrobothnia.

Don. Don was identified by the model as having the highest probability of a Häme or Savo dialect (Figure 9). This was likely because of his high usage of *-iA* compared to other *-eA* variants, frequent use of schwa and schwa following /h/, and exclusive use of *UA* for *OA*. Don was of the third generation and his grandparents had migrated from Savo, Häme, and South Ostrobothnia. He reported speaking a lot of Finnish both at home and while working as a young adult. Neither he nor Ron (another third generation speaker) had a single instance of a first person plural verb suffix, though the extension of the singular third person verb form across all persons is reportedly one of the more common variations between American and Standard Finnish (Martin 1993:98). This particular feature did not carry much load within the model and likely did not strongly impact results.

Ron. Ron was identified by the model as most likely having a Häme or Savo dialect (Figure 10). Much of this was likely because of his frequent use of *-iA* compared to other *-eA* variants, his frequent use of schwa, use of a *myö* form of the first person plural pronoun, and frequent use of the standard *minä* for the first person pronoun.

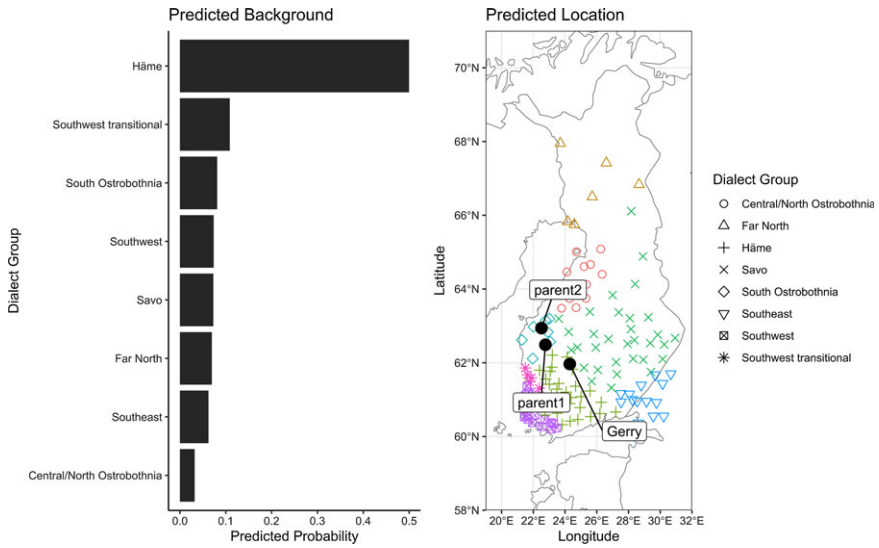


Figure 8. Gerry's dialectal distribution.

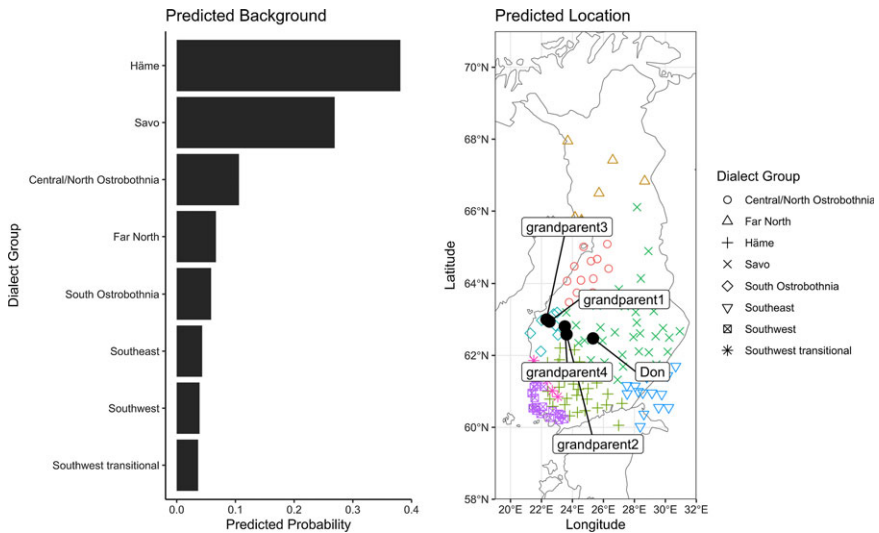


Figure 9. Don's dialectal distribution.

Ron was also of the third generation and two of his grandparents were from South Ostrobothnia and Central/North Ostrobothnia, though this was not evident in his speech. Ron commented that he often spoke Finnish both with his spouse's parents, who lived nearby, as well as at work, indicating that he likely had varied input from speakers of different backgrounds. Like Don, this sample of his speech did not have a single instance of a first person plural verb suffix.

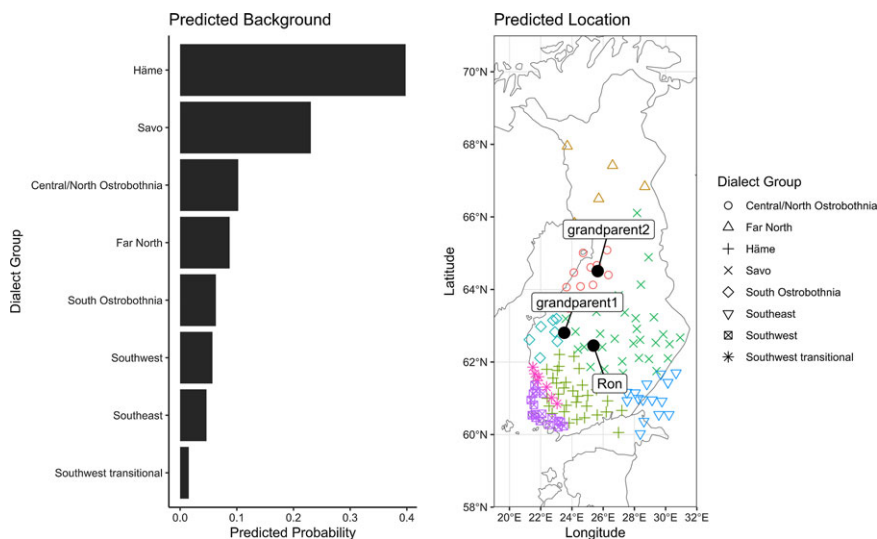


Figure 10. Ron's dialectal distribution.

5. Discussion and conclusions

Our results suggest the following answers to our research questions: (i) the dialectal background of individual American Finnish speakers is identifiable and relatable to the regional variation of Finnish in Finland, and (ii) in some instances, the speech of American Finnish speakers correlates with the regions from which their ancestors migrated. The dialect features which carried the greatest weight in the model are also those found in Standard Finnish: *-ssA* for the inessive case and *minä* for the first person pronoun. These variants overlap with particular dialect varieties, especially the Savo and Häme dialects, which is why those had the highest probabilities in the model across all American Finnish speakers. It is less that these American Finnish speakers are likely to be speakers of these dialects so much as there were fewer features in their speech indicating that they were more likely to be speakers of a different dialect. The dialect features examined were selected because of pragmatic factors like the operationalizability of the included features, and added until a sufficiently high classification accuracy was achieved and the addition of other features did not offer significant classification improvements.

In this dataset, those American Finnish speakers who had the fewest features indicating a specific dialectal background other than Häme and Savo were those of the third generation. This does not suggest that they had a shared dialect or speech variety, rather that these individuals had fewer marked dialect features – in line with what typically occurs over time in language contact settings (see Trudgill 1986). Second generation speakers in this dataset exhibited enough dialect features for their speech to be categorized as belonging to particular dialects; in two of those instances, the categorization related to where their parents had migrated from. Closer examination of which feature variants are most frequent across all Finnish dialects, and how these in turn are used by American Finnish speakers, is a potential line of

future research. Individuals in this dataset indicated limited experience in reading or writing Standard Finnish. While this is likely common across other speakers of their generations, individual sociolinguistic backgrounds are important to consider in conjunction with quantitative analyses of their speech.

While individual influences are difficult to capture fully, by embedding individual speaker variation in the model from the initial validation against data from the LAX corpus, this model offers a more realistic comparison of speakers' actual language use and dialectal background. Better understanding of speakers' dialectal backgrounds gives information on the input that speakers likely received, and in turn allows for better assessment of reasonable baseline varieties for comparison, which is especially relevant for heritage language research. This approach to dialect categorization is continually refinable with additional data and allows for social variables to be taken into account on an individual level, giving a more holistic account of speaker variation and language change. Methodologically, an interesting avenue for future research is to explore the different types of response variables in the modeling. For instance, instead of labels of dialect regions or coordinate-based location, the map could be divided into a dense grid, so that the model would predict the most likely area for each speaker (for various methodological possibilities in geolocation, see e.g. Gaman et al. 2020).

Supplementary material. The supplementary material for this article can be found at <https://doi.org/10.1017/S0332586523000057>

Acknowledgements. We would like to thank participants at the Finnish Conference of Linguistics in Turku in 2022 as well as colleagues at the University of Turku for their insightful comments, questions, and discussion of this study. We would especially like to thank Joseph Salmons and three anonymous reviewers for their valuable comments for improving and tightening this manuscript. Any remaining errors are ours alone.

Competing interests. The authors declare none.

Notes

1 Kettunen's division has only seven regions with Southwest transitional and Häme dialect groups combined. Otherwise, his divisions are similar to Rapola's.

2 The Korp user interface can be accessed here: <https://korp.csc.fi/korp/>.

3 The R scripts, anonymized datasets, and distribution charts for all dialect features included in the study are available in an online repository: <https://osf.io/y5cq4/>.

4 Abbreviations in examples follow the Leipzig Glossing Rules. In addition to these: PASS = passive, ILL = illative, ADE = adessive, pres = present tense, past = past tense,

5 As mentioned above, the model produces a probability for each region. In the validation experiment, the region with the highest probability is considered the predicted value.

References

- Bousquette, Joshua. 2020. From bidialectal to bilingual: Evidence for multi-stage language shift in Lester W. J. 'Smoky' Seifert's 1946–1949 Wisconsin German Recordings. *American Speech* 95(4). 485–523.
- Bousquette, Joshua & Michael T. Putnam. 2020. Redefining language death: Evidence from moribund grammars. *Language Learning* 70(S1). 188–225.
- Breiman, Leo. 2001. Random Forests. *Machine Learning* 45(1). 5–32.

- Deshors, Sandra C. & Stefan Th. Gries. 2016. Profiling verb complementation constructions across New Englishes: A two-step random forests analysis of *ing* vs. *to* complements. *International Journal of Corpus Linguistics* 21(2). 192–218.
- Eichhoff, Jürgen. 1985. The German language in America. In F. Trommler & J. McVeigh (eds.), *America and the Germans: An assessment of a three-hundred year history*, 223–240. Philadelphia: University of Pennsylvania Press.
- Gaman, Mihaela, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhainen, Tommi Jauhainen, Krister Lindén, Nikola Ljubešić, et al. 2020. A report on the VarDial Evaluation Campaign 2020. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, 1–14. Barcelona, Spain: International Committee on Computational Linguistics (ICCL).
- Halmari, Helena. 1997. *Government and codeswitching: Explaining American Finnish*. John Benjamins.
- Haugen, Einar. 1953/1969. *The Norwegian language in America* (2 vols.). Bloomington, IN: University of Indiana Press.
- Hijmans, Robert J. 2021. geosphere: Spherical trigonometry. <https://CRAN.R-project.org/package=geosphere>.
- Hirvonen, Pekka. 1992. Vowel and consonant length opposition in American Finnish: An example of language attrition. In Jussi Niemi (ed.), *Studia Linguistica Careliana: A Festschrift for Kalevi Wiik on the occasion of his 60th birthday*, 21–38. Joensuu: University of Joensuu.
- Hirvonen, Pekka. 2005. Finnish in the Upper Midwest. In Dennis Preston, Brian D. Joseph & Carol Preston (eds.), *Linguistic diversity in Michigan and Ohio: Towards two state linguistic profiles*, 217–243. Ann Arbor, MI: Caravan Books.
- Hirvonen, Pekka & Timo Lauttamus. 2000. Code-switching and language attrition: Evidence from American Finnish interview speech. *SKY Journal of Linguistics* 13. 47–74.
- Hjelde, Arnstein. 2015. Changes in a Norwegian dialect in America. In Johanna Bondi Johannessen & Joseph C. Salmons (eds.), *Germanic heritage languages in North America: Acquisition, attrition and change*, 283–298. John Benjamins.
- Honkola, Terhi. 2016. *Macro- and microevolution of languages: Exploring linguistic divergence with approaches from evolutionary biology*. University of Turku: PhD dissertation.
- Hyvönen, Saara, Antti Leino & Marko Salmenkivi. 2007. Multivariate analysis of Finnish dialect data: An overview of lexical variation. *Literary and Linguistic Computing* 22(3). 271–290.
- Ikola, Osmo, Ulla Palomäki & Anna-Kaisa Koitto. 1989. *Suomen murteiden lauseoppia ja tekstikieloppia* [The syntax and text grammar of Finnish dialects]. Helsinki: Finnish Literature Society (SKS).
- Itkonen, Terho. 1964. Sananrajaisten äänneilmioiden synkroniaa ja diakroniaa [The synchrony and diachrony of sound phenomena at word boundaries]. *Virtittäjä* 68(3). 225–225.
- Itkonen, Terho. 1989. *Nurmijärven murrekirja* [Nurmijärvi's dialect book] (Suomalaisen Kirjallisuuden Seuran Toimituksia 498). Helsinki: Finnish Literature Society (SKS).
- Ivaska, Ilmari. 2015. Longitudinal changes in academic learner Finnish: A key structure analysis. *International Journal of Learner Corpus Research* 1(2). 210–241.
- Ivaska, Ilmari & Silvia Bernardini. 2020. Constrained language use in Finnish: A corpus-driven approach. *Nordic Journal of Linguistics* 43(1). 33–57.
- Ivaska, Ilmari & Laura Ivaska. 2022. Source language classification of indirect translations. *Target: International Journal of Translation Studies* 34(3). 370–394.
- Johnson, Mirva. 2022. Politics and cooperatives: Verticalization in rural Finnish–American communities of the Upper Midwest. In Joshua R. Brown (ed.), *The verticalization model of language shift: The Great Change in American communities*, 25–51. Oxford: Oxford University Press.
- Jönsson-Korhola, Hannele. 1993. Lauserakenteesta toisen ja kolmannen polven kielenkäytössä [Regarding phrase structure in second and third generation speakers' language usage]. In Pertti Virtaranta, Hannele Jönsson-Korhola, Maisa Martin & Maija Kainulainen (eds.), 102–127.
- Jönsson-Korhola, Hannele & Lindgren, Anna-Riitta (eds.). 2003. *Monena suomi maailmalla: Suomalaisperäisiä kielivähemmistöjä* [Diverse Finnish around the globe: Language minorities of Finnish origin] (Tietolipas 190). Helsinki: Finnish Literature Society (SKS).
- Kettunen, Lauri. 1940a. *Suomen murteet, IIIA: Murrekartasto* [Finland's dialects, IIIA: Dialect map]. Helsinki: Finnish Literature Society (SKS).
- Kettunen, Lauri. 1940b. *Suomen murteet, IIIB: Selityksiä murrekartastoon* [Finland's dialects, IIIB: Explanations for the dialect map]. Helsinki: Finnish Literature Society (SKS).

- Lainio, Jarmo. 1989. *Spoken Finnish in urban Sweden*. Uppsala: Centre for Multiethnic Research.
- Larmouth, Donald Eilford. 1974. Differential interference in American Finnish cases. *Language* 50(2). 356–366.
- Lauttamus, Timo. 1991. Borrowing, code-switching, and shift in language contact: Evidence from Finnish–English bilingualism. In Muusa Ojanen and Marjatta Palander (eds.), *Language contacts east and west* (Studies in Languages 22), 32–53. Joensuu: University of Joensuu.
- Leino, Antti & Saara Hyvönen. 2008. Comparison of component models in analysing the distribution of dialect features. *International Journal of Humanities and Arts Computing* 2. 173–187.
- Leino, Antti, Saara Hyvönen & Marko Salmenkivi. 2006. Mitä murteita suomessa onkaan? Murreseston levikin kvantitatiivista analyysiä [What dialects are there in Finnish anyway? Quantitative analysis of the spread of dialect vocabulary]. *Virtittäjä* 110. 26–45.
- Litty, Samantha, Christine Evans & Joseph Salmons. 2015. Gray zones: The fluidity of Wisconsin German language and identification. In Peter Rosenberg, Konstanze Jungbluth & Dagna Zinkhahn Rhobodes (eds.), *Linguistic construction of ethnic borders*, 183–205. Bern: Peter Lang.
- Lyskawa, Paulina & Naomi Nagy. 2020. Case marking variation in heritage Slavic languages in Toronto: Not so different. *Language Learning* 70(S1). 122–156.
- Männikkö, Hanna. 2004. ‘–mä oon BUSY BUSY LADY, YOU KNOW–’: *Koodinvaihdon rakenteet ja funktiot jälkipolvien amerikkasuomessa*. Ms., Department of Language and Communication Studies, University of Jyväskylä.
- Mantila, Harri. 1997. Johdanto [Introduction]. In Seija Makkonen & Harri Mantila (eds.), *Pohjoissuomalaisenpuhekielen sosiolingvistinen variaatio* [Sociolinguistic variation in northern spoken Finnish], 1–23. Oulu: University of Oulu.
- Mantila, Harri. 2004. Murre ja identiteetti [Dialect and identity]. *Virtittäjä* 108(3). 322–322.
- Martin, Maisa. 1993. Muoto-opin seikkoja [Matters of morphology]. In Pertti Virtaranta, Hannele Jönsson-Korhola, Maisa Martin & Maija Kainulainen (eds.), 97–101.
- Mielikäinen, Aila. 1986. Nykysuomen murtuvat murrerajat [The fracturing dialect boundaries of modern Finnish]. *Kielikello* 2. 12–17.
- Natvig, David. 2022. The Great Change and the shift from Norwegian to English in Ulen, Minnesota. In Joshua R. Brown (ed.), *The verticalization model of language shift: The Great Change in American communities*, 85–114. Oxford: Oxford University Press.
- Nuolijärvi, Pirkko & Marja-Leena Sorjonen. 2010. *Miten kuvata muutosta? Puhutun kielen tutkimuksen lähtökohalta murteenseuruhankkeen pohjalta* [How to describe change? Departure points for spoken language research underlying the dialect project]. Helsinki: Kotimaisten Kielten Tutkimuskeskus.
- Otheguy, Ricardo, Zentella, Ana Celia, & Livert, David. 2007. *Language and dialect contact in Spanish in New York: Toward the formation of a speech community*. *Language* 83(4). 770–802.
- Pascual y Cabo, Diego & Jason Rothman. 2012. The (il)logical problem of heritage speaker bilingualism and incomplete acquisition. *Applied Linguistics* 33(4). 450–455.
- Polinsky, Maria. 2018. *Heritage languages and their speakers*. Cambridge University Press.
- Polinsky, Maria & Olga Kagan. 2007. Heritage languages: In the ‘wild’ and in the classroom. *Language and Linguistics Compass* 1(5). 368–395.
- Poplack, Shana. 1980. ‘Sometimes I’ll start a sentence in Spanish y termino en español’: Toward a typology of code-switching. *Linguistics* 18(7–8). 581–618.
- Poplack, Shana, Susan Wheeler & Anneli Westwood. 1989. Distinguishing language contact phenomena: Evidence from Finnish–English bilingualism. *World Englishes* 8(3). 389–406.
- Rapola, Martti. 1969 [1947]. *Johdatus Suomen murteisiin* [Introduction to Finland’s dialects]. Helsinki: Finnish Literature Society (SKS).
- Scontras, Gregory, Zuzanna Fuchs & Maria Polinsky. 2015. Heritage language and linguistic theory. *Frontiers in Psychology* 6. <https://doi.org/10.3389/fpsyg.2015.01545>.
- South, Andy. 2017a. *rnaturalearthdata*: World vector map data from natural earth used in ‘rnaturalearth’. <https://CRAN.R-project.org/package=rnaturalearthdata>.
- South, Andy. 2017b. *rnaturalearth*: World map data from natural earth. <https://CRAN.R-project.org/package=rnaturalearth>.
- Syrjänen, Kaj. 2021. *Quantitative language evolution case studies in Finnish dialects and Uralic languages*. University of Tampere: PhD dissertation.

- Syrjänen, Kaj, Terhi Honkola, Jyri Lehtinen, Antti Leino & Outi Vesakoski. 2016. Applying population genetic approaches within languages: Finnish dialects as linguistic populations. *Language Dynamics and Change* 6(2). 235–283.
- Tagliamonte, Sali A. & R. Harald Baayen. 2012. Models, forests and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change* 24. 135–178.
- Trudgill, Peter. 1986. *Dialects in contact*. Oxford: Basil Blackwell.
- Virtaranta, Pertti, Hannele Jönsson-Korhola, Maisa Martin & Maija Kainulainen. 1993. *Amerikansuomi* (Tietolipas 125). Helsinki: Finnish Literature Society (SKS).
- Wiik, Kalevi. 2004. *Suomen murteet: Kvantitatiivinen tutkimus* [Finland's dialects: A quantitative study]. Helsinki: Finnish Literature Society (SKS).
- Wright, Marvin N. & Andreas Ziegler. 2017. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software* 77(1). 1–17.
- Yager, Lisa, Nora Hellmold, Hyoun-A Joo, Michael T. Putnam, Eleonora Rossi, Catherine Stafford & Joseph Salmons. 2015. New structural patterns in moribund grammar: Case marking in heritage German. *Frontiers in Psychology* 6. <https://doi.org/10.3389/fpsyg.2015.01716>

Cite this article: Ivaska I, Johnson M, and Kurki T. Identifying the dialectal background of American Finnish speakers using a supervised machine-learning model. *Nordic Journal of Linguistics*. <https://doi.org/10.1017/S0332586523000057>