Short title: Moving to continuous classifications of bilingualism through machine learning

Long title: Moving to continuous classifications of bilingualism through machine learning trained on language production

Authors: Coco, M. I.[1,2*], Smith, G.[3*], Spelorzi, R.[4] & Garraffa, M.[3]
[1]Department of Psychology, "Sapienza" University of Rome, Rome, Italy
[2] I.R.C.S.S Fondazione Santa Lucia, Rome, Italy
[3]School of Health Sciences, University of East Anglia, Norwich, UK
[4]Department of Linguistics and English Language, University of Edinburgh, Edinburgh, UK

[*] Denotes equal contribution

Competing interests: the authors declare none.

Addresses for correspondence:

Dr Moreno I. Coco
Dipartimento di Psicologia
Sapienza, Universita' di Roma
Via dei Marsi, 78
00185, Roma,
Italy
email: moreno.coco@uniroma1.it

Dr Giuditta Smith
School of Health Sciences
University of East Anglia
Norwich Research Park
NR4 7TJ
Norwich
UK
Email: giuditta.smith@uea.ac.uk

Abstract

Recent conceptualisations of bilingualism are moving away from strict categorisations, towards continuous approaches. This study supports this trend by combining empirical psycholinguistics data with machine learning classification modelling. We trained support vector classifiers on two datasets of linguistic productions coded for type of production of Italian speakers to predict their class (i.e., "monolingual", "attriters", and "heritage"). All classes were predicted above chance (> 33%), even if the classifier's performance substantially varies, with monolinguals identified better (f-score > 70%) and attriters being the most confusable (f-score < 50%). The confusion matrices qualify that attriters are identified as heritage speakers nearly as often as they could be correctly classified, suggesting this class to sit in the middle. Clitic clusters were found to be the most identifying features for discrimination. Overall, this study supports a conceptualisation of bilingualism as a continuum of linguistic behaviours rather than sets of a-priori established classes.

Keywords: bilingualism, heritage speakers, attrition, support vector machine, classification

1. Introduction

In a globalized and highly integrated world, the boundaries of languages have become fluid and seemingly continuous. Speakers are more likely to move across countries, transfer their homeland language to their offspring and acquire other languages, with bilingual proficiency reaching native-like language abilities well after childhood (Steinhauer, 2014; Roncaglia-Denissen & Kotz, 2016; Hartshorne, Tenenbaum, & Pinker, 2018; Köpke, 2021; Gallo et al. 2021). However, bilingualism is known to substantially vary among individuals, as it is shaped by intra and extralinguistic factors such as amount of exposure, social status, and education (Gullifer et al. 2018; Haranto & Yang, 2016; Rodina et al. 2020; Bialystok, 2016; Gullifer & Titone, 2020). Consequently, research in bilingualism has progressively abandoned strict categorical approaches in favour of more nuanced ones. The increased complexity of a "winner-take-all" definition of bilingualism has created a plethora of labels to classify speakers (see Surrain & Luk, 2017 for a systematic review), sometimes leading to the same speakers being labelled differently according to whether the classification is based on language dominance, learning history, age, etc., which renders it impractical to perform consistent comparisons across different studies. Moreover, strict classifications disregard that individuals can also change their "label" during their lifetime. The most notable examples are expatriates to foreign countries who exhibit quick changes in their native language after immersion in the dominant language of the host country (so-called attriters, with attrition phenomena starting just a few years of immersion, Eckle & Hall, 2013), or early bilinguals who experience expatriation to the family homeland (so-called returnees, Flores et al. 2022). All the above taken, the cogent question explored in the present study is how separate these categories truly are, especially given that they could overlap. The effects of cross-linguistic influence associated with the attrition on the L1 by the L2, for example, are hard to

disentangle, with long-lasting effects attested bidirectionally, which implies that every bilingual may also be an attriter (Schmid & Köpke 2017a,b). Even the category of monolinguals, which may represent a gold standard, is now considered the exception rather than the norm, given the growing number of people immersed in multilingual and multidialectal societies (Davies, 2013; Rothman et al. 2022).

Critically, this debate about terminology and categorical labels in bilingualism has key implications for research practices. Most of the research on bilingualism has adopted a grouping model whereby individuals are assigned to a priori selected language groups with arbitrary cut-offs (Wagner et al., 2022). These categories have typically been used to compare bilinguals, also articulated in different categorical subtypes, to a control group of monolinguals. However, even more nuanced categorical distinctions pose several challenges. First, any a priori classification is based on some enumerable inclusion criteria (e.g., age of acquisition) but may exclude others (e.g., quantity of exposure; Marian & Hayakawa, 2020; Kremin & Byers-Heinlein, 2021; Wagner, Bialystok & Grundy, 2022). Second, empirical evidence deriving from possible class comparisons, which feed hypothetical models aiming to explain them, circularly depend on the criteria adopted to define the groups at the outset. While this issue is inherent to most research comparing groups, it bears important consequences when such groups are highly variable at their core. This is the case, for example, in comparative research about Autism Spectrum Disorder (ASD) which often employs selection criteria at the outset that are based on measures such as verbal and non-verbal intelligence. This is problematic because individuals with ASD widely vary on these and other cognitive abilities, so even if matched on standardised common measures, they may still be very different in their individual profiles (see Jarrold and Brock 2004, and references therein). As already hinted, this issue is of paramount importance for research into

bilingualism too, as the variability in the criteria used to define bilingual classes could inevitably lead to results that are difficult to replicate.

Along with other researchers, therefore, we suggest that a more fruitful conceptualisation of bilingualism would be to place each individual on one point of a continuum according to certain linguistic characteristics (Baum & Titone, 2014; de Bruin, 2019; Marian & Hayakawa, 2020; Rothman et al. 2023). A recent proposal in this direction comes from Kremin and Byers-Heinlein (2021), who suggest factor-mixture and grade-of-membership models, which evaluate individuals' bilingualism according to their intrinsic variability in language experience along a continuum of fuzzy classes. Conceptually, these models assign to each individual a composite continuous score, based on specific measures (e.g., a bilingual questionnaire), which reflects how much they belong to a monolingual or bilingual (also of multiple types) class, therefore accounting for within-group heterogeneity. In essence, this approach proposes bilingual classes, but their boundaries are fuzzy so that individuals deviating from the strict inclusion criteria could also be accommodated. The main advantage of this approach is to still investigate a diversity of factors contributing to the bilingual experience but without introducing biases that may arise when evaluations are strictly based on predetermined categories (DeLuca et al. 2019, 2020; Kałamała et al. 2022; Li & Xu, 2022). Another useful, more continuous, approach to examining bilingualism is through machine learning which has already shown promising results such as differentiating the degree of second language proficiency (Yang, Ty & Lim, 2016), qualifying its relationship to executive control (Gullifer & Titone, 2021) or uncovering its longitudinal lifelong impact (Jones, Davies-Thompson, & Tree 2021).

Yet, the concept of bilingual continuum still struggles to take off, despite its theoretical and methodological benefits, and it is still countered by attempts to establish better

boundaries of bilingual categories through richer assessments or questionnaires (see Kašćelan et al. 2022 for a review). The core objective of this study is to precisely provide empirical and computational proofs that the adoption of a categorical approach can be fallacious in bilingual research; and that instead continuous approaches better describe the true nature of bilingualism and must be adopted whenever possible.

2. Current study

The key proposition of the current study is that bilingualism distributes along a continuum, which is difficult to frame within strictly defined classes. We provide empirical proof to this proposition with machine learning classifiers trained on psycholinguistics language production data from individuals that vary in their degree of bilingualism but are conventionally identified as belonging to three specific classes (i.e., monolinguals, attriters, heritage). We demonstrate that even if we can successfully identify the class an individual was a-priori assigned to, the classification performance widely varies as susceptible to inevitable overlaps in the language production profiles of the speakers. So, individuals belonging to a class situated in the middle of two possible extremes (i.e., attriters) are identified much less accurately as their language profile is shared by other classes. We precisely take the uncertainty in the classification of bilingualism as the proof of concept about its continuous nature. If individuals of all classes were equally identifiable, then the existence of such classes would have been correctly assumed, but this is not what we find. Instead, the variability found in the language production profiles of these individuals, and consequently the inability to fully discriminate among them, is coherent with a continuous rather than categorical definition of their bilingual nature.

3. Methods

We train support vector machines (SVM), which are suited to classification problems and often used in the cognitive sciences (see Cervantes et al., 2020 for a review) on the syntactic typology of utterances produced in a question-directed image description task to predict three classes of speakers on the monolingual-bilingual spectrum (i.e., homeland monolingual residents, long-term residents or attriters, and heritage speakers). Two different datasets, both including these three different classes of speakers, are used to train and test the SVMs. The first dataset, which we will refer to as "the original dataset", has been recently published by Smith et al. (2023) and a second dataset, which we will refer to as "the novel dataset", purposely collected for the current study to ensure that our results are reliable and robust: if the SVM trained on the original dataset can predict well above chance the classes of speakers on a novel dataset, collected using the same stimuli, procedure and task, then results are highly replicable.

3.*1 The datasets: Participants*

The original dataset comprises productions from a total of 86 adult speakers of Italian (26 homeland monolingual speakers, 30 attriters, and 30 heritage speakers), while the novel dataset comprises a total of 15 adult speakers of Italian, 5 participants for each of the same 3 classes of speakers considered in our study (see Table 1 for a description of the two datasets). At the time of testing, homeland speakers were living in Italy, attriters were living in Scotland, where they had been living for a minimum of 5 years, and heritage speakers were

living in Scotland, where they had lived all or most of their lives but were highly proficient in Italian[1].

<Insert table 1 about here>

3.*2 The datasets: Productions*

All participants took part in a series of elicitation tasks, which are fully described in Smith et al. (2023). In the tasks, participants are prompted to answer a question about an image depicting two characters interacting with each other, and an object. The question is related to either one of the arguments (direct or indirect object, example in 1) or both (example in 2) and is designed to elicit an affirmative one-verb sentence with a bi- or tri-argumental verb.

(1) Preamble: In questa scena, ci sono una signora, un commesso, e un maglione.

In this scene, there is a lady, a clerk, and a pullover.

Question: Cosa fa il commesso al maglione/alla signora?

What is the clerk doing with the pullover/to the lady?

(2) Preamble: In questa scena, Marco vuole prendere oppure ridare il pupazzo a Sara.

In this scene, Marco wants to take or give back the teddy bear to Sara.

[1] Proficiency was tested through a standardised test of Italian proficiency for adults, Comprendo (Cecchetto et al. 2012), on which both bilingual populations were at ceiling.

Question: Qui Marco prende il pupazzo a Sara. Qui cosa fa?

Here Marco takes the teddy bear from Sara. What is he doing here?

Although several answers are possible, the prompt question is designed to maximise the accessibility of the target object(s), consequently creating the pragmatic environment for the use of a weak form, which in Italian is realised through the clitic pronoun (*gli* in example 3). This design is widely used and is very effective in healthy native speakers of Italian in eliciting clitic pronouns (Tedeschi, 2008; Arosio et al. 2014; Guasti et al. 2016; Vender et al. 2016, and more).

(3) (Cosa fa la bambina al bambino?)

Gli ruba la merenda

(What is the girl doing to the boy?)

*She is stealing the snack **from him***

The production rates of this structure show significant differences between monolinguals and bilinguals as well as among bilinguals, particularly when the two languages spoken are a clitic and a non-clitic language (Belletti et al., 2007; Smith et al. 2022; Romano 2021, 2022). Smith et al. (2023), which provides the original dataset used also in the current study, found a differential pattern of clitic production across three groups (monolinguals, attriters, heritage

speakers) where all types of clitics (one argument, as in 3 above, or clitic clusters) were significantly fewer in attriters compared to monolinguals, and in the heritage speakers compared to attriters. This phenomenon was interpreted as a by-product of "inter-generational attrition", where only monolinguals retain a strong preference for clitics over any other structure, attriters make use of single clitics but not of clusters, and heritage speakers, whose input is provided by attriters, mostly prefer the use of lexical expressions[2]. Building upon these insights, in the current study, we identified and coded for five types of answers according to how object(s) of the main verb was realised: "single clitic", "lexical element", "cluster 1st/2nd", "cluster 3rd", "other". "Other" comprises all types of answers that did not fall under any of the remaining categories (e.g., irrelevant answers or answers containing either an omission or a strong pronoun). Examples of the coding are provided in Table 2.

In the original dataset (i.e., from Smith, et al., 2023) we had a total of 2,688 sentences, which are divided into 760 (single clitic), 757 (lexical element), 619 (cluster 1st/2nd), 444 (cluster 3rd) and 108 (other). In the novel dataset (i.e., collected specifically for the current study) we have a total of 480 sentences, divided into 126 (single clitic), 106 (lexical element), 128 (cluster 1st/2nd), 115 (cluster 3rd) and 5 (other).

<insert table 2 about here>

[2] Lexical expressions, such as *the boy* ('il bambino'), in 'la bambina calcia il bambino', *the girl is kicking the boy*, are preferred over any form of pronoun, including strong pronouns (e.g., 'lui', *him*), which would be direct translations of the English (but refer to Smith et al. 2023 for a discussion).

3.3 Analyses

We perform three types of analyses all based on support vector machines classifiers[3] (SVM) trained to predict the class of the speaker, i.e., a three-level categorical vector of class labels (monolingual, attriter, heritage) based on the type of *answer* produced (a categorical vector of 5 levels indicating the typology of the utterance). All data processing and analyses are conducted on the R Statistical Software (v. 4.3.2, R Core Team, 2023) through the RStudio environment (v. 2023.09, RStudio Team, 2023) and using the package `e1071` (v. 1.7-14, Chan & Lin, 2011) to run the SVMs.

The first analysis only uses the original dataset, and we train the SVM classifier on a randomly selected 90% of such data and then test it on the remaining, unseen, 10%. This process is repeated 1,000 times to make sure that the classifier does not overfit the data while making full use of a relatively small data set[4]. To measure the prediction performance of the algorithm, we compute the F-score, which is the geometric mean of precision and recall defined as $F = 2 * (P * R)/(P + R)$. Precision (P) is the number of correctly classified instances over the total number of instances labelled as belonging to the class, defined as $tp/(tp + fp)$. Here, tp is the number of true positives (i.e., instances of the class correctly predicted), and fp is the number of false positives (i.e., instances wrongly labelled as members of the class). Recall (R) is the number of correctly classified instances over the total number of instances in that class, defined as $tp/(tp + fn)$, where fn is the number of false

[3] We tried also different classifiers, such as linear discriminant analysis, or multinomial log-linear neural network models, but were only able to classify two out of three classes (i.e., monolingual and heritage) because attriters are a particularly confusable class as our error analysis shows.

[4] We also followed a canonical cross-fold validation procedure whereby we partitioned the entire original dataset into 10 randomly generated folds, each containing 90% of the data for training and 10% for testing. This makes sure that the algorithm is equally trained and tested on each data point. We repeat this process 100 times to guarantee that the data is well-mixed across the folds. We obtained identical classification results (F-scores; monolingual = 0.72; heritage = 0.58; attriters = 0.42)

negatives, i.e., the instances labelled as non-members of the class even though they were. As precision, recall and F-score are relative to the class being predicted, we report separate values for each of them. To explicitly quantify the differences in classification performance for the three different classes, we run a simple linear regression predicting F-scores as a function of the to-be-predicted class (monolingual, attriter, heritage, with heritage as the reference level). The purpose of this first analysis is to demonstrate that we can successfully classify the class an individual belongs to, based on a published dataset of which we already know the characteristics (i.e., the original dataset by Smith, et al., 2023), but not with the same accuracy, indicating a continuum of linguistic behaviours across classes.

In the second analysis, we train the SVM on the original dataset but test it only on the second novel and unseen dataset, which is collected at a different time (after the original dataset) on a different set of speakers but using the same task, and report the same measures of F-score, Precision and Recall. As already said, the purpose of this analysis is to conceive a blind test that makes sure our classification results are fully replicable also on unseen data (i.e., a novel dataset). If we repeat the elicited production task with the same classes of speakers, and we observe the same level of categorization accuracy in our predictions, it means that our empirical results are highly replicable and consequently our theoretical claims are very solid (i.e., we can repeat the experiment and run the models on yet another unseen sample of the same populations and observe the same pattern).

The third analysis instead examines the impact of each type of production on the classification performance to provide a rough idea about the importance of each elicited structure in distinguishing the class each speaker may belong to. First, we aggregate both the original and the novel datasets and recode all different productions into binary vectors (0,1), indicating whether a certain structure (e.g., cluster $1^{st}/2^{nd}$) was used in that particular

sentence. This re-coding generates 5 different binary feature vectors, one for each production observed (refer to Method for a description of the coding). Then, we use a stepwise forward model-building procedure, where at each step we evaluate whether the model with the added feature is significantly better, i.e., it has a higher F-score, than the one without it. If there is no significant improvement in the F-score, we retain the model without that feature. We repeat this procedure over 1,000 iterations (randomly sampling 90% of the data for training and the remaining 10% for testing) and calculate the frequency of observing a certain feature in the final feature set according to the position it is selected to. For example, if the first feature selected, because it produced a higher F-score compared to the rest, is cluster 3rd, then it ranks first. Then, if the F-score of this model significantly improved by adding cluster 1st/2nd, this feature will be kept in the model and ranked as second (refer to Coco & Keller, 2014 for a similar approach but based on eye-movement features).

All these analyses are run on an SVM whose parameters are tuned to achieve optimal performance. There are two parameters in SVM models: *Gamma*, which shapes the decision boundaries by assembling similar data points into the same cluster, and *Cost*, which attributes a penalty to misclassification. These parameters are used to adapt the prediction plane to potentially non-linear data patterns. We extract optimal values for the gamma and cost parameters across the original dataset using the `tune.svm()` function also available in the `e1071` package. We examine a range of gamma values going from .005 to .1 in steps of 0.005. The optimal parameters obtained to model our dataset are 0.01 for gamma and 0.5 for the cost.

Finally, we visually inspect and evaluate more in-depth the performance of the SVM classifiers through confusion matrices, which tell how much the model may erroneously predict one class for another. A confusion matrix is, in fact, a contingency table, where

expected and predicted values are cross-tabulated, i.e., the number of correct and incorrect predictions is counted for each of the expected classes. In practice, confusion matrices provide insights about the type of errors that are made, e.g., whether a monolingual is more often confused with an attriter or with a heritage. In the context of our study, it is interesting to examine whether classes are univocally represented, and in case of errors, what are the most prominent switches. So, if for example, monolinguals are more confused with attriters, we can infer that these two classes share a closer production strategy than say between monolinguals and heritage.

The data and R script to illustrate the analysis supporting the findings of this study can be found in the Open Science Framework (https://osf.io/w24p3/?view_only=48f70ddee34e44a1b4ba2dd766ff9a34).

4. Results

In Table 3, we report the descriptives for F-score, Precision and Recall, regarding the classification performance of the SVM models trained and tested only on the original dataset (first analysis); and trained on the original dataset but tested on the novel dataset (second analysis, refer to section Analyses for details about their purposes). Across the board, we can predict the class of the speaker based on their typology of linguistics production with an accuracy above chance, which is 33% in our data (i.e., the SVM is trying to predict one class out of three possible classes). In particular, when training and testing were conducted using the OD, we find that the Monolingual class is most accurately classified (~72%) followed by Heritage (~58%) and finally Attriters (~42%). These results are fully confirmed, if not improved when training is done on the original dataset but testing is performed on the novel dataset (Monolingual = 79%; Heritage = 64% and Attriters = 47%).

<Insert table 3 about here>

This finding is corroborated by the linear regression on the original dataset, which confirms that Heritage and especially Attriters are predicted with a significantly smaller accuracy than Monolinguals (refer to Table 4 for the model coefficients[5]).

<Insert table 4 about here>

Our examination of the confusion matrix (third analysis) confirms that the class predicted most accurately often is Monolinguals, followed by Heritage and Attriters. The same result holds when using only the original dataset (Figure 1a), and when instead testing is performed on the novel dataset (Figure 1b). In these figures, the diagonals of the confusion matrices display all percentages of expected cases (Target, organised as columns) that were correctly predicted (Prediction, organised as a row) by the classifier. Most interesting perhaps are the misclassification errors, namely the percentages of mismatches between targets and predictions which can be read in the off-diagonal cells of the matrix. Here, we find that Attriters are misclassified as Monolinguals more often than as Heritage, whereas Heritages are misclassified as Attriters more often than as Monolinguals. This is again true for both analyses.

Finally, the feature selection analysis showed that the best classifiers needed an average of 1.88 (±0.31) types of productions to achieve the maximum F-score. Moreover, if we inspect

[5] Note, we could not repeat the linear regression for the second analysis as we are not iterating, i.e., a one-shot training-testing.

the relative importance of each feature for the classification, we find that cluster 3rd was the feature most frequently selected as first, followed by cluster 1st/2nd. The lexical element is instead the third most selected feature, and when it happened, it was usually the only one selected, i.e., adding any other would not significantly improve the F-Score (refer to Figure 1C for a visualisation).

<Insert figure 1 about here>

5. Discussion

In the present study, we tested the hypothesis that linguistic performance can be used as a proxy for bilingual categories and that their boundaries are fuzzy. By applying machine learning to a dataset of utterances, speakers were assigned to their class, out of three possible (monolingual, heritage, and attriter), with an accuracy well above chance (47-79%, where chance is 33%). This shows that specific linguistic patterns are to some extent coherent with bilingual classes created a priori, also lending empirical support to our modelling approach. However, the classification accuracy varied greatly between classes, showing that the boundaries of these classes have a degree of fuzziness, with some linguistic profiles characterising one class more strongly compared to the others. These results confirm that even if speakers can be identified to some extent as belonging to a possible category in the monolingual-bilingual spectrum based on their language production profiles, the classes consistently overlap. Critically, this is especially the case for those Italian speakers (i.e., the attriters) in the middle between a linguistic environment which was fully Italian-dominant (i.e., where they grew up) and the other which is fully English-dominant (i.e., where they now live).

Specifically, the confusion matrix and associated analysis of errors shows that the Monolinguals are closer to Attriters, which are in turn closer to Heritage. We take this uncertainty in the classification of bilingualism as a proof of concept about its continuous nature. Classification accuracy was higher for monolinguals and heritage, while lower for attriters. This is in line with predictions made by a continuous approach to bilingualism, where, considering different definitions of classes in the spectrum, we have monolinguals and heritage speakers at opposite ends (e.g., monolinguals are at the "least bilingual" end). Since the language investigated is Italian, it is theoretically expected that monolinguals will be very productive of a specific syntactic element (i.e., the clitic pronoun) that is frequently adopted in the homeland. At the other end of the spectrum are heritage speakers, who are the most dominant speakers of the second language, in this case, English and, while highly proficient, the least exposed to Italian. It seems to be the case that their language, sometimes referred to as the heritage language, is quite identifiable. This is consistent with accounts of heritage languages as being stand-alone varieties of the homeland language (Nagy, 2016; Kupish & Polinsky, 2022).

The attriter class, which displays the lowest classification accuracy, is particularly relevant for the debate of a bilingual continuum. These speakers are confused as heritage almost as frequently as they are correctly categorised, confirming there is an important degree of overlap between classes that manifests in speakers' use of language. The linguistic production of attriters is closer to heritage who were born outside of the homeland and have lower exposure to Italian than monolinguals, who like them were born in Italy.

As was stated in the methods section, the way the dataset was coded (i.e., the chosen answer strategy for the production of the direct and/or indirect object) would maximise the emergence of potential differences given that the task was designed to promote the use of a

pronominal element, which is a known area of differences between monolinguals, bilinguals, and different classes of bilinguals. Despite this, overlap between classes is still present, as is demonstrated by the high confusability rates.

Results from the present study are consistent with accounts of bilingualism as a continuous rather than categorical variable (Luk & Bialystok, 2013; Bonfieni, 2018), as the individual profiles of speakers are not univocally describable through strict boundaries, but rather behave as a continuum of discrete linguistic behaviours. The continuity of bilingual profiles also fits in well with the fact that some differences between speakers may always remain the same (e.g., whether they received, or not, inputs in a specific language as children), while others may change over time influenced by speakers' linguistic experience. Changes in linguistic boundaries across generations of speakers are to be expected and predictable because the language spoken by a speaker is constantly influenced by concurrent factors such as exposure, language dominance, environment during acquisition, and so on (Luk & Bialystok 2013; Anderson et al., 2020).

Classes in bilingual research are often determined based on a close set of apriori-defined linguistic and extralinguistic factors such as the age of first exposure, country of residence, etc., or based on self-assessment questionnaires. The latter are often reported to be subjected to enhancement bias, particularly in the case of heritage speakers (MacIntyre et al., 1997; Gollan et al., 2012; Marchman et al. 2017; Macbeth et al. 2022); the former do not fully mirror linguistic performance (de Bruin, 2021). Our study precisely confirms that there is a degree of overlap between the patterns of linguistic productions of speakers that would be assigned instead to different classes in the monolingual-bilingual spectrum. The major theoretical contribution of our novel findings is therefore the confirmation of a need to shift, whenever possible, from a priori grouping towards methodologies that either eliminate

discrete groups or can exploit explicitly such intergroup variability to better model language experience (Kremin and Byers-Heinlein, 2021) in bilingual research.

6. Conclusions

In this study, a machine learning model (SVM) trained on the typology of linguistic productions was used to predict the bilingual class a speaker may have belonged to. We did this aiming to demonstrate that class boundaries are not as clear cut and overlaps exist. Results show that classes are predicted above chance, but with a varying degree of accuracy, which depends on the apriori bilingual class a speaker was assigned to. The typology of utterances speakers produced makes it clear that (mono- and) bilingualism does not have sharp categorical boundaries, but rather it distributes on a continuum of linguistic behaviours that are shared by different classes of speakers. Heritage speakers and monolinguals seem to speak rather different varieties of Italian, while attriters seem to sit somewhere in the middle. Future research may explore how the classification behaves with larger chunks of production, for example examining the outcomes of narrative tasks. Further studies that examine the reliability of classification are also needed in other areas of linguistic research, for example in the classification of linguistic competence in neurodevelopmental disorders.
In sum, our findings strongly suggest fostering more innovative research that exploits the true linguistic environment each speaker carries to derive a continuum rather than a class-based approach to bilingual research.

Ethics: The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

Data availability: The data and script to illustrate the analysis supporting the findings of this study are available in Open Science Framework at https://osf.io/w24p3/?view_only=48f70ddee34e44a1b4ba2dd766ff9a34

References

Anderson, J.A.E., Hawrylewicz, K., and Bialystok, E. (2020) Who is bilingual? Snapshots across the lifespan. *Bilingualism: Language and Cognition, 23*, 929–937. https://doi.org/10.1017/S1366728918000950

Arosio, F., Branchini, C., Barbieri, L., & Guasti, M. T. (2014). Failure to produce direct object clitic pronouns as a clinical marker of SLI in school-aged Italian speaking children. *Clinical linguistics & phonetics*, *28*(9), 639-663. https://doi.org/10.3109/02699206.2013.877081

Baum, S., & Titone, D. (2014). Moving toward a neuroplasticity view of bilingualism, executive control, and aging. *Applied Psycholinguistics, 35*(5), 857–894. http://dx.doi.org.lib-ezproxy.concordia.ca/10.1017/S0142716414000174

Belletti, A., Bennati, E., & Sorace, A. (2007). Theoretical and developmental issues in the syntax of subjects: Evidence from near-native Italian. *Natural Language & Linguistic Theory*, *25*, 657-689. https://doi.org/10.1007/s11049-007-9026-9

Bialystok, E. (2016). The signal and the noise: Finding the pattern in human behavior. *Linguistic Approaches to Bilingualism*, *6*(5), 517-534. https://doi.org/10.1075/lab.15040.bia

Bonfieni, M. (2018) *Bilingual continuum: Mutual effects of language and cognition.* PhD thesis, University of Edinburgh.

Cecchetto, C., Di Domenico, A., Garraffa, M., & Papagno, C. (2012). *COMPRENDO. Batteria per la comprensione di frasi negli adulti* (pp. 1-85). Raffaello Cortina Editore.

Cervantes. J., Garcia-Lamont, F., Rodríguez-Mazahua, L. and Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing, 408*, 189-215. https://doi.org/10.1016/j.neucom.2019.10.118

Chang, C.-C. and Lin, C.J. (2011) LIBSVM : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1, 27:27. https://doi.org/10.1145/1961189.1961199

Coco, M. I., & Keller, F. (2014). Classification of visual and linguistic tasks using eye-movement features. Journal of vision, 14(3), 11-11.

Davies, A. (2013) *Native Speakers and Native Users: Loss and Gain*. Cambridge: Cambridge University Press.

de Bruin, A. (2019). Not all bilinguals are the same: A call for more detailed assessments and descriptions of bilingual experiences. *Behavioral Sciences*, 9(3), 33. https://doi.org/10.3390/bs9030033

DeLuca, V., Rothman, J., Bialystok, E., and Pliastikas, C. (2019). Redefining bilingualism as a spectrum of experiences that differentially affects brain structure and function. *Proceedings of the National Academy of Sciences*, *116*(15), 7565– 7574. https://doi.org/10.1073/pnas.1811513116.

DeLuca, V., Rothman, J., Bialystok, E., & Pliatsikas, C. (2020). Duration and extent of bilingual experience modulate neurocognitive outcomes. *NeuroImage*, *204*, Article 116222.

Ecke, P. and Hall, C.J. (2013). Tracking tip-of-the-tongue states in a multilingual speaker: Evidence of attrition or instability in lexical systems?. *International Journal of Bilingualism 17*(6), 734-751. https://doi.org/10.1177/1367006912454623

Flores, C., Zhou, C., and Eira, C. (2022). "I no longer count in German". On dominance shift in returnee heritage speakers. *Applied Psycholinguistics* 43(5), 1019-1043. https://doi.org/10.1017/S0142716422000261

Gallo, F., Ramanujan, K., Shtyrov, Y. and Myachykov, A. (2021). Attriters and Bilinguals: What's in a Name? *Frontiers in Psychology 12*, Article 558228. https://doi.org/10.3389/fpsyg.2021.558228

Gollan, T. H., Weissberger, G. H., Runnqvist, E., Montoya, R. I., and Cera, C. M. (2012). Self-ratings of spoken language dominance: a multilingual naming test (MINT) and preliminary norms for young and aging Spanish–English bilinguals. *Bilingualism: Language and Cognition 15*, 594–615. https://doi.org/10.1017/S1366728911000332

Guasti, M. T., Palma, S., Genovese, E., Stagi, P., Saladini, G., & Arosio, F. (2016). The production of direct object clitics in pre-school–and primary school–aged children with specific language impairments. *Clinical Linguistics & Phonetics*, *30*(9), 663-678. https://doi.org/10.3109/02699206.2016.1173100

Gullifer, J. W., Chai, X. J., Whitford, V., Pivneva, I., Baum, S., Klein, D., & Titone, D. (2018). Bilingual experience and resting-state brain connectivity: Impacts of L2 age of acquisition and social diversity of language use on control networks. *Neuropsychologia*, *117*, 123-134. https://doi.org/10.1016/j.neuropsychologia.2018.04.037

503 Gullifer, J. W., & Titone, D. (2020). Characterizing the social diversity of bilingualism using
504 language entropy. *Bilingualism: Language and Cognition*, *23*(2), 283-294.
505 https://doi.org/10.1017/S1366728919000026

506 Gullifer, J. W., & Titone, D. (2021) Engaging proactive control: Influences of diverse
507 language experiences using insights from machine learning. *Journal of Experimental*
508 *Psychology: General 150*(3), 414. https://doi.org/10.1037/xge0000933

509 Hartanto, A., & Yang, H. (2016). Disparate bilingual experiences modulate task-switching
510 advantages: A diffusion-model analysis of the effects of interactional context on switch costs.
511 *Cognition*, *150*, 10-19. https://doi.org/10.1016/j.cognition.2016.01.016

512 Hartshorne, J. K., Tenenbaum, J. B., & Pinker, S. (2018). A critical period for second
513 language acquisition: Evidence from 2/3 million English speakers. *Cognition*, *177*, 263-277.
514 https://doi.org/10.1016/j.cognition.2018.04.007

515 Jarrold, C., & Brock, J. (2004). To match or not to match? Methodological issues in autism-
516 related research. *Journal of autism and developmental disorders*, *34*, 81-86.

517 Jones, S.K., Davies-Thompson, J., and Tree, J. (2021). Can machines find the bilingual
518 advantage? Machine learning algorithms find no evidence to differentiate between lifelong
519 bilingual and monolingual cognitive profiles. *Frontiers in Human Neuroscience 15*, Article
520 621772. https://doi.org/10.3389/fnhum.2021.621772

521 Kałamała, P., Senderecka, M., & Wodniecka, Z. (2022). On the multidimensionality of
522 bilingualism and the unique role of language use. *Bilingualism: Language and Cognition*,
523 *25*(3), 471-483. https://doi.org/10.1017/S1366728921001073

Kašćelan, D., Prévost, P., Serratrice, L., Tuller, L., Unsworth, S., & De Cat, C. (2022). A review of questionnaires quantifying bilingual experience in children: Do they document the same constructs?. *Bilingualism: Language and Cognition*, *25*(1), 29-41. https://doi.org/10.1017/S1366728921000390

Köpke, B. (2021). Language attrition: A matter of brain plasticity?: Some preliminary thoughts. *Language, Interaction and Acquisition 12*(1), 110-132. https://doi.org/10.1075/lia.20015.kop

Kremin, L. V., & Byers-Heinlein, K. (2021). Why not both? Rethinking categorical and continuous approaches to bilingualism. *International Journal of Bilingualism* 25(6): 1560-1575. https://doi.org/10.1177/13670069211031986

Kupisch, T., Polinsky, M. (2022). Language history on fast forward: Innovations in heritage languages and diachronic change. Bilingualism: Language and Cognition, 25, 1–12.

Li, P., & Xu, Q. (2022). Computational modelling of bilingual language learning: Current models and future directions. *Language Learning.* https://onlinelibrary.wiley.com/doi/full/10.1111/lang.12529

Luk, G., & Bialystok, E. (2013). Bilingualism is not a categorical variable: Interaction between language proficiency and usage. *Journal of Cognitive Psychology* 25(5), 605-621. https://doi.org/10.1080/20445911.2013.795574

Macbeth, A., Atagi, N., Montag, J. L., Bruni, M. R., & Chiarello, C. (2022). Assessing language background and experiences among heritage bilinguals. *Frontiers in Psychology*, *13,* Article 993669. https://doi.org/10.3389/fpsyg.2022.993669

MackeyWF (1962) The description of bilingualism. *Canadian Journal of Linguistics, 7,* 51–85. https://doi.org/10.1017/S0008413100019393

MacIntyre, P. D., Noels, K. A., & Clément, R. (1997). Biases in self-ratings of second language proficiency: the role of language anxiety. Lang. Learn. 47, 265–287. c10.1111/0023-8333.81997008

Marchman, V. A., Martinez, L. Z., Hurtado, N., Gruter, T., & Fernald, A. (2017). Caregiver talk to young Spanish-English bilinguals: comparing direct observation and parent-report measures of dual-language exposure. *Developmental Science, 20*, Article e12425. https://doi.org/10.1111/desc.12425

Marian, V., & Hayakawa, S. (2021). Measuring bilingualism: The quest for a "bilingualism quotient". *Applied Psycholinguistics, 42*(2), 527-548. https://doi.org/10.1017/S0142716420000533

Montrul, S. (2016) *The acquisition of heritage languages.* Cambridge University Press.

Polinsky, M. and Scontras, G. (2020) Understanding heritage languages. *Bilingualism: Language and Cognition 23*(1), 4-20. doi: 10.1017/S1366728919000245

Nagy, N. (2016). Heritage languages as new dialects. In M. Jones, E. Smith, & R. Brown (Eds.), The future of dialects: Selected papers from Methods in Dialectology XV (pp. 15–34). Language Science Press.

Rodina, Y., Kupisch, T., Meir, N., Mitrofanova, N., Urek, O., & Westergaard, M. (2020, March). Internal and external factors in heritage language acquisition: Evidence from heritage Russian in Israel, Germany, Norway, Latvia and the United Kingdom. *Frontiers in Education*, *5*, 20. https://doi.org/10.3389/feduc.2020.00020

Romano, F. B. (2020). Ultimate attainment in heritage language speakers: Syntactic and morphological knowledge of Italian accusative clitics. *Applied Psycholinguistics*, *41*(2), 347-380. https://doi.org/10.1017/S0142716419000559

Romano, F. B. (2021). L1 versus Dominant Language Transfer Effects in L2 and Heritage Speakers of Italian: A Structural Priming Study. *Applied Linguistics*, *42*(5), 945-969. https://doi.org/10.1093/applin/amaa056

Roncaglia-Denissen, M. P., & Kotz, S. A. (2016). What does neuroimaging tell us about morphosyntactic processing in the brain of second language learners?. *Bilingualism: Language and Cognition*, *19*(4), 665-673. https://doi.org/10.1017/S1366728915000413

Rothman, J., Bayram, F., DeLuca, V., Alonso, J. G., Kubota, M., & Puig-Mayenco, E. (2023). Defining bilingualism as a continuum. In G. Luk, J. G. Grundy, J. A.E. Anderson (Eds.) *Understanding Language and Cognition through Bilingualism: In honor of Ellen Bialystok*, *64*, 38.

Rothman, J., Bayram, F., DeLuca, V., Di Pisa, G., Dunabeitia, J.A., Gharibi, K., ... & Wulff, S. (2022). Monolingual comparative normativity in bilingualism research is out of "control": Arguments and alternatives. *Applied Psycholinguistics* 1-14. https://doi.org/10.1017/S0142716422000315

Schmid, M.S., & Köpke, B. (2017a). The relevance of first language attrition to theories of bilingual development. *Linguistic Approaches to Bilingualism* 7(6), 637–667. https://doi.org/10.1075/lab.17058.sch

Schmid, M.S., & Köpke, B. (2017b). When is a bilingual an attriter? Response to the commentaries. *Linguistic Approaches to Bilingualism* 7(6), 763–770. https://doi.org/10.1075/lab.17059.sch

Smith G, Spelorzi R, Sorace A and Garraffa M (2022). Language Competence in Italian Heritage Speakers: The Contribution of Clitic Pronouns and Nonword Repetition. *Languages* 7(3):180. https://doi.org/10.3390/languages7030180

Smith, G., Spelorzi, R., Sorace, A. & Garraffa, M. (2023). Inter-generational attrition: first language attriters and heritage speakers on production of Italian complex clitic pronouns. *Linguistic Approaches to Bilingualism.* https://doi.org/10.1075/lab.23002.smi

Steinhauer, K. (2014) Event-related potentials (ERPs) in second language research: A brief introduction to the technique, a selected review, and an invitation to reconsider critical periods in L2. *Applied Linguistics* 35, 393–417. https://doi.org/10.1093/applin/amu028

Surrain, S., & Luk, G. (2017) Describing bilinguals: A systematic review of labels and descriptions used in the literature between 2005–2015. *Bilingualism: Language and Cognition* 1–15. https://doi.org/10.1017/S1366728917000682

Vender, M., Garraffa, M., Sorace, A., & Guasti, M. T. (2016). How early L2 children perform on Italian clinical markers of SLI: A study of clitic production and nonword repetition. *Clinical Linguistics & Phonetics*, *30*(2), 150-169. https://doi.org/10.3109/02699206.2015.1120346

Wagner, D., Bialystok, E., & Grundy, J.G. (2022). What Is a Language? Who Is Bilingual? Perceptions Underlying Self-Assessment in Studies of Bilingualism. *Frontiers of Psychology* 13:863991. https://doi.org/10.3389/fpsyg.2022.863991

609 Yang, Y., Yu, W., & Lim, H. (2016). Predicting second language proficiency level using
610 linguistic cognitive task and machine learning techniques. *Wireless Personal*
611 *Communications*, *86*(1), 271-285. https://doi.org/10.1007/s11277-015-3062-2

612

Tables and figures

Table 1. *Characteristics of study participants. for the Original Dataset (OD) and the test dataset (TD)*

OD	**Monolinguals**	**Attriters**	**Heritage Speakers**
Number	26 (female: 19)	29 (female: 18)	30 (female: 19)
Age	M = 35.57 SD =8.16	M = 39.31 SD = 11.76	M = 35.7 SD = 12.29
Years in the UK	0	M = 15.25 SD = 8.9	M = 35.4 SD = 11.98
Level of Education	Secondary: 7, University: 19	University: 29	Secondary: 10, University: 20
Schooling in Italian (years)	26	29	0
Schooling in English (years)	0	6 (HE)	30
Geographic areas of Italy	North: 10 Centre: 11 South + islands: 5	North: 8 Centre: 12 South + islands: 9	North: 7 Centre: 14 South + islands: 9
TD	**Monolinguals**	**Attriters**	**Heritage Speakers**
Number	5 (female: 3)	5 (female: 4)	5 (female: 3)
Age	M = 34.8 SD = 4.32	M = 38.2 SD = 9.52	M = 33.8 SD = 6.46
Length of residence in the UK	0	M = 8.4 SD = 2.7	M = 33.1 SD = 6.2

Level of Education	University: 5	University: 5	University: 5
Schooling in Italian (years)	5	5	0
Schooling in English (years)	0	0	5
Geographic areas of Italy	North: 1 Centre: 4 South + islands: 0	North: 2 Centre: 1 South + islands: 2	North: 2 Centre: 0 South + islands: 3

615

616

617 Table 2. *The coding strategy, with examples from the data.*

Coding	Answer type	Example
1	single clitic	Gli legge il libro To-him reads the book 's/he's reading him the book'
2	lexical element	Legge il libro al bambino reads the book to-the child 's/he's reading the child the book'
3	cluster 1st/2nd	Te lo leggo To-you it read 'I'm reading it to you'
4	cluster 3rd	Glielo legge To-him/her-it reads 's/he's reading it to him/her'
5	other	

618

619

Table 3. *Descriptive statistics of the SVM classification performances. We report the mean of F-score, Precision and Recall on 1,000 iterations of training SVMs on 90% of the data, and testing on the remaining unseen 10%. The hyphen separates the first classifier (trained and tested on the original dataset) from the second classifier (trained on the original dataset but tested on the novel dataset).*

Group	**F-score**	**Precision**	**Recall**
Monolingual	0.72-0.79	0.66-0.66	0.80-1
Attriters	0.42-0.47	0.46-0.53	0.38-0.42
Heritage	0.58-0.64	0.60-78	0.57-0.54

Table 4. *Output of a linear model predicting F-score as a function of the three classes of speakers in our study (Attriters, Heritage and Monolinguals, as the reference level).*

	Estimate	**Standard Error**	**z value**	**Pr(>|z|)**
(Intercept)	0.673	0.002	369.207	< .001
Attriters	-0.181	0.003	-70.427	< .001
Heritage	-0.113	0.003	-44.056	< .001

Figure 1. (**A-B**) Visualization of the confusion matrices about the classification performance of our models (**A**: trained and tested using the OD; **B**: trained on THE ORIGINAL DATASET tested on TD). Predictions of the model are organised over the rows while the target, i.e., expected outcome, is organised over the columns. The percentages indicate how many cases, per class, matched or not, between predictions and targets. The colours of the tiles go from white (few cases) to orange (most cases). **C:** Percentages of times a certain type of production was selected as a key feature, i.e., it significantly improved performance, by the classifier. The type of productions is depicted as colours and organized as stacked bars. Cluster glie-lo in the image refers to cluster 3rd, and cluster me-lo to cluster 1st/2nd. The x-axis indicates instead whether the feature was selected as the first or second feature. Note, all models contained a maximum of two types of production, hence, there are no further ranks.