Using Machine Learning to Predict Bilingual Language Proficiency from Reaction Time Priming Data

Laura E. Matzen (lematze@sandia.gov)
Applied Cognitive Science, Sandia National Laboratories, PO Box 5800
Albuquerque, NM 87185-1327 USA

Christina L. Ting (clting@sandia.gov)
Computational Decision Science, Sandia National Laboratories, PO Box 5800
Albuquerque, NM 87185-1027 USA

Mallory C. Stites (mcstite@sandia.gov)
Applied Cognitive Science, Sandia National Laboratories, PO Box 5800
Albuquerque, NM 87185-0639 USA

Abstract

Studies of bilingual language processing typically assign participants to groups based on their language proficiency and average across participants in order to compare the two groups. This approach loses much of the nuance and individual differences that could be important for furthering theories of bilingual language comprehension. In this study, we present a novel use of machine learning (ML) to develop a predictive model of language proficiency based on behavioral data collected in a priming task. The model achieved 75% accuracy in predicting which participants were proficient in both Spanish and English. Our results indicate that ML can be a useful tool for characterizing and studying individual differences.

Keywords: repetition priming; translation priming; bilingual language processing; machine learning

Introduction

Studies of bilingual language processing have raised interesting questions about the nature of linguistic and semantic representations in semantic memory. Many open questions remain regarding the organization of multiple languages within the processing system, particularly the extent to which two languages share underlying conceptual representations and automatically activate one another during processing. Much of this research has relied on the use of priming paradigms to probe the size and nature of cross-language priming effects, as a way to understand whether the bilingual processing system shares representations across languages, or whether concepts may be represented separately. Two competing models of bilingual language comprehension, the Bilingual Interactive Activation (BIA+) model (Dijkstra & van Heuven, 2002) and the Revised Hierarchical Model (RHM; Kroll & Stewart, 1994; Kroll, van Hell, Tokowicz, & Green, 2010), have been proposed to help account for varying priming effects observed across studies.

One of the most common tasks used to probe the nature of bilingual language processing is cross-language translation priming. Masked repetition priming effects are well-established within a speaker's native language (Forster & Davis, 1984), particularly in the lexical decision task. By comparing the size of within- and across-language translation priming, researchers can begin to understand how effectively words in one language facilitate the same concept in their second language. Cross-language non-cognate translation priming effects have also been observed in cases where priming from a word in one language facilitates responses to that word's translation in another language (e.g., Grainger & Frenck-Mestre, 1998). Translation priming effects tend to be bigger under certain circumstances: for example, in more proficient bilinguals, with longer prime durations, and with priming from L1 primes to L2 targets (rather than L2 primes to L1 targets), (see Schoonbaert, Duyck, Brysbaert, & Hartsuiker, 2009 for review).

Research in this field has traditionally relied on recruiting groups of individuals with known language backgrounds and testing how priming effects manifest in these pre-established language proficiency groups. However, this existing paradigm is not without challenges. An individual's language background is hard to effectively quantify, and individuals can vary widely in their second language proficiency even within relatively well-matched groups (for review, see van Hell & Tanner, 2012). Because the size or presence of priming effects depends heavily on correctly characterizing participants' language background, it seems that efforts to establish a more individualized approach to data analysis could help the field identify more consistent findings, in turn advancing our understanding of the bilingual language processing system.

The present study seeks to establish a novel approach to bilingual language comprehension research that capitalizes on individual differences rather than averaging over them. We are interested in trying to characterize an individual's language background, without knowing it in advance, based on their behavioral responses in a cross-language priming task. Specifically, we utilize supervised machine learning techniques to identify patterns in response time data that may differentiate individuals who are proficient in the target language from those who are not. This approach represents a departure from traditional paradigms and leverages cross-

disciplinary data analysis techniques to provide a potential new avenue for the study of bilingual language processing.

In order to collect behavioral responses to words in a language an individual may not know, we employed a word-length judgment task rather than lexical decision or semantic categorization. This task has been successfully used to elicit N400 priming in a bilingual population in which L1 and L2 words were intermixed (Martin, Dering, Thomas, & Thierry, 2009), indicating that the task could still allow for contact with the word's semantics. Additionally, as in Martin et al. (2009), we intermixed trials from the two languages rather than using the more typical blocked design. This choice was made to make it less predictable at the trial level whether the upcoming word would be in a language the participant knew, thus further encouraging participants to access each word's semantics. We predicted that we would see within-language repetition priming effects for the languages in which the participant was proficient. We also predicted that proficient bilinguals would show translation priming effects, whereas participants who were not proficient in the second language would not show these effects. Furthermore, exploratory machine learning analyses will allow us to test whether other aspects of the behavioral data could reliably predict an individual's language proficiency.

Methods

This study was reviewed and approved by the Human Studies Board at Sandia National Laboratories. A total of 95 participants were recruited via Amazon Mechanical Turk (AMT). To qualify for the task, the participants had to have an approval rate >95% for prior tasks completed on AMT. A subset of 40 participants also met AMT's criteria for fluency in Spanish. Participants were paid $3-4 for their time.

Materials

The materials consisted of 30 Spanish nouns and their translations in English. The words were selected so that there were no special characters (accents, etc.) and no cognates or false cognates. We took care to select Spanish words that monolingual English speakers would be unlikely to encounter in their daily lives. Using information from the CLEARPOND database (Marian, Bartolotti, Chabal & Shook, 2012), the word lists were matched on length, frequency, and orthographic neighborhood size, as shown in Table 1. The orthographic neighborhood size across languages was minimized.

Table 1: Matched Properties of the Two Word Lists

	English	Spanish
Avg. Length	5.37	5.47
Avg. Frequency	106.93	99.04
Avg. English Orthographic Neighborhood Size	6.77	0.77
Avg. Spanish Orthographic Neighborhood Size	2.27	5.43

The words were paired in eight types of pairings: English-English repetitions, Spanish-Spanish repetitions, English-Spanish translations, Spanish-English translations, English-English unrelated, Spanish-Spanish unrelated, English-Spanish unrelated, and Spanish-English unrelated. There were a total of 240 pairs, with each word appearing in every possible pair type, four times as a prime and four times as a target. The word pairs were divided into four blocks of 60 pairs each. Each target word appeared twice in each block, once as part of a related pair and once as part of an unrelated pair. The pairs were placed in the pseudorandom order so that the two pairs that contained the same target word appeared in different halves of the block. The pseudorandom order was constrained so that there were never more than four translation/repetition or unrelated pairs in a row, and never more than two pairs of the same type (e.g., Spanish-English translation) in a row.

Procedure

After reading and acknowledging the consent form, participants completed a short language proficiency questionnaire with questions that were similar to those in the Language Experience and Proficiency Questionnaire (Marian, Blumenfeld & Kaushanskaya, 2007). They were asked to list up to four languages that they know, first in order of dominance and then in order of acquisition. They were asked what percentage of the time they are currently exposed to English and Spanish, and how much total time they have spent living or traveling in countries where Spanish or English is the dominant language. Finally, they were asked to rate their level of proficiency in English and Spanish on an 11-point scale ranging from "None" to "Perfect," the age at which they began to acquire each language (infant, child, teen, adult, or never), and which factors contributed to them learning that language. The response options included interacting with family, interacting with friends, formal language classes, reading, language tapes/learning apps/self-instruction, watching TV or movies, listening to the radio, and travel.

After completing the questionnaire, participants were shown the task instructions and an example. They were told that they would see words in English and Spanish, and that the words would sometimes be repeated or followed by the same word in the other language. They were told to press the "B" key on the keyboard if the word had 5 letters or fewer and the "N" key if the word had 6 letters or more. They were instructed to respond as quickly as possible without making too many mistakes. Finally, the participants were told that there were four blocks of words with breaks in between, and that each block would take about two minutes to complete. When they were ready to begin, they clicked on a button labeled "Start Experiment." The first six words that the participants saw were practice words and were not included in the analysis. The participants responded to every item, whether it was a prime or a target.

Behavioral Results

A total of 13 participants were excluded from the analysis, either because they did not complete the entire task, they did not provide consistent responses to the questionnaire, or because their pattern of responses indicated that they were responding randomly rather than following the task instructions. Of the remaining 82 participants, 40 were from the group that met AMT's criteria for fluency in Spanish and 42 were from the group with no specific language qualification requirements.

In the group that met AMT's criteria for fluency in Spanish, one participant rated his/her proficiency in reading Spanish at 7 ("Good"), and all of the other participants rated their proficiency at 8 ("Very Good") or higher on the 0-10 scale. Thirty-three of the participants in this group reported that Spanish was their dominant language and the first language they acquired. Three participants reported that English was their dominant language and the first language they acquired. Two participants reported that Spanish was the first language they acquired, but English was their dominant language. Two participants reported that English was the first language they acquired, but Spanish was their dominant language. All of the participants in this group reported that they had lived for at least one year in an area where Spanish is the predominant language (range 1-57 years, mean = 28.8 years). They had spent an average of nine years living in areas where English was the predominant language (range = 0-54 years). Thirty-three of the participants reported that they had spent more time living in predominantly Spanish-speaking areas than in predominantly English-speaking areas, and 17 reported that they had never lived in an area where English was the predominant language.

In the group of participants that was recruited without the use of AMT's Spanish fluency qualification, all of the participants reported that English was their dominant language, and all but one of the participants reported that English was the first language they acquired (one person reported that their first language was Mandarin). There were 21 participants who reported that they did not know any Spanish at all. Another 15 participants reported that they had learned some Spanish as a teen or adult, primarily through formal language classes or self-instruction, but they rated their Spanish proficiency at 3 ("Fair") or below. Three participants reported that they began learning Spanish as teenagers and gave themselves intermediate fluency ratings (5-7). Finally, three participants rated their Spanish proficiency as 8 or higher. One of these participants reported that they started learning Spanish in infancy, one in childhood, and one as a teen. The participants reported that they had spent an average of 37.7 years living in predominately English-speaking areas (range 25-70 years) and an average of 3 years living or traveling in predominantly Spanish-speaking areas (range 0-23 years).

For our analyses, we grouped all of the participants who rated their Spanish proficiency as 8 or higher into the "proficient" group, regardless of whether or not they had AMT's qualification for Spanish proficiency. There were a total of 42 participants in this group, 39 from the batch that required the AMT Spanish qualification and three from the batch that did not. All of the participants who rated their Spanish proficiency at 7 or lower (40 participants) were placed in the "non-proficient" group. One of these participants was from the batch that required the AMT Spanish qualification and the 39 were from the batch that did not.

We began with a traditional analysis of the priming effects for each experimental condition. The participants' average response times were calculated for each condition. Only correct trials were included in the analysis. Trials with response times (RTs) of less than 200 milliseconds were excluded, as were trials with RTs that were more than three standard deviations higher than that participant's mean response time (unless those trials had RTs that were less than 6 seconds). A total of 111 trials out of 19,680 were excluded due to having unusually short or long response times. For each participant, the priming effect for each condition (English-English, Spanish-Spanish, English-Spanish, Spanish-English) was calculated by subtracting the average RT for the targets in the repetition or translation pairs from the average RT for the targets in the unrelated pairs. Figure 1 shows the average size of the priming effects across participants.

A 2 (Spanish Proficiency) x 4 (Priming Condition) ANOVA showed that there was a significant effect of proficiency group ($F(1,240) = 16.77$, $p < .001$), a significant effect of condition ($F(3,240) = 90.04$, $p < .001$), and a significant interaction between the two ($F(3, 240) = 5.11$, $p < .01$). The participants in the proficient Spanish group had a significantly larger priming effect than the other group for both the English-English ($t(67) = 3.50$, $p < .001$) and the Spanish-Spanish condition ($t(67) = 3.59$, $p < .001$). For the two cross-language conditions, neither group showed a priming effect and the two groups did not differ significantly from one another (all ts < 1.12, all ps > .13).

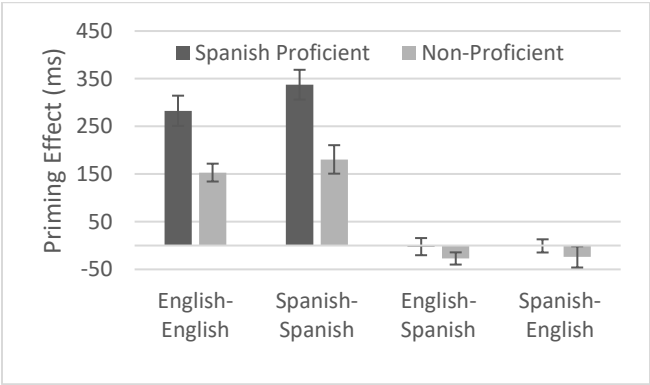

Figure 1: The average magnitude of the priming effects. Error bars show the standard error of the mean.

A Model of Bilingual Language Proficiency

The priming effects (PEs) showed that there was a significant difference between participant groups in the English-English

and Spanish-Spanish priming conditions. A potential application of this result may be to learn a function that maps the priming effects of known participants to their corresponding proficiency labels so that we can use the priming effects from new participants to predict their proficiency.

More generally, *classification* is a standard supervised machine learning (ML) task that follows a *train* and *predict* paradigm. During the training phase, labeled data is used to build a model (i.e., a learned function) that maps an input (typically numerical feature vectors) to an output (labels). During the predict phase, the model is used to infer the labels of new data (James, Witten, Hastie, & Tibshirani, 2013).

An advantage of this approach is that ML algorithms can usually handle very high dimensional data (e.g., the individual PEs or RTs) compared with standard statistical analyses of behavior, which look at averages (e.g., average PE or RT for a particular condition). A disadvantage of this approach is that ML algorithms are often considered "black boxes", providing very little interpretability as to how the model arrives at its prediction.

A linear Support Vector Machine (SVM), on the other hand, is a simple but successful ML algorithm that yields insights as to how the individual features (e.g., PEs and RTs) contribute to the predicted output (Boser, Guyon, & Vapnik, 1992; Cristianini & Shawe-Taylor, 2000). In its simplest form, the objective of an SVM is to find a hyperplane that separates the labeled data into the two distinct classes (extensions for multiclass problems exist), while also maximizing the distance between the hyperplane and the nearest point from either group (hard-margin). The coordinates of the vector orthogonal to the hyperplane form the weights (coefficients) of the model. From the weights, it is possible to do two things. First, we can determine feature importance according to the relative magnitude of the weights. Second, new data items can be labeled depending on which side of the hyperplane they fall (computed by taking the dot product with the orthogonal vector).

For our application, we use the Linear Support Vector Classification (LinearSVC) class available in Python's Scikit-learn 0.23.1 with default parameters. Scikit-learn 0.23.1 is used throughout our ML workflow for data preprocessing, feature engineering, and model validation (Pedregosa et al., 2011).

Data Preprocessing

Using the same criteria as in the prior section, participants were assigned proficiency labels based on their survey responses. Specifically, 42 participants were labeled as "proficient" in Spanish and 40 participants were labeled as "non-proficient" (English proficiency is assumed).

Each participant was associated with a list of 240 RTs for each of the 240 target words in the experiment. Across all participants, the mean RT for the target words was 825 ms and the standard deviation was 231 ms. Target words with a mean RT that was more than three standard deviations above this mean were removed from the dataset for all participants. Only one target word was excluded based on this criterion, leaving us with 239 RTs for each participant. Then, to account for different baseline RTs for different participants, each participant's RTs were normalized from 0 to 1.

We note that this approach for preprocessing the data for input into the SVM differs from the approach for cleaning the data for the behavioral analysis. In the behavioral analysis, each participant's data is cleaned by removing individual trials with incorrect responses and/or unusually short/long responses. Thus, each participant is left with a *different* set of RTs and PEs after cleaning the data. However, for input into the SVM, each participant must be represented by the *same* set of features, necessitating a different approach to removing anomalous data.

Feature Engineering and Selection

From the 239 normalized RTs, we construct feature vectors that are used as input into the SVM as follows. The first feature set simply represents the 239 normalized RTs. The second feature set represents the PEs. Each English target word appears in two PEs (English-English and Spanish-English); similarly, each Spanish target word appears in two PEs (Spanish-Spanish and English-Spanish). Therefore, for the 60 target words in this study, we have 120 PEs. Because one target word was excluded, we are left with 119 PEs.

Given a set of features, a standard next step in a machine learning workflow is to perform some type of *feature selection* technique to reduce the number of features, i.e. reduce the dimensionality. Reducing the number of features, particularly when the number of features exceeds the number of samples, can improve the accuracy of the model.

Univariate feature selection is one of the simplest techniques to reduce the number of features and works by selecting the best set of features based on univariate statistical tests such as a chi-squared test or an ANOVA. We will use an ANOVA to compute the p-value between the label and features to select the m best features according to the lowest p-values.

Model Validation

In a deployed setting, we would apply our SVM model that has been trained on the 82 participants of known proficiency to make predictions on new participants of unknown proficiency. However, without validating the model first, it is not possible to know how good the new predictions are. Therefore, a cross-validation test is usually performed first, in which part of the labeled data is withheld during training and used to test (validate) the performance of the model during prediction. Many methods exist to split the data into train/test sets. Perhaps most common is the k-fold cross validator, which splits the data into k consecutive folds. Each fold is then used once as the test (validation) set, while the remaining $k - 1$ sets form the training set. We use $k = 5$ and perform 10 runs of each of the cross-validation experiments.

Finally, the model (i.e. m best features) with the highest mean *balanced accuracy score* is selected. The balanced accuracy is defined as the average accuracy obtained on each

class (non-proficient, proficient) and is used in place of accuracy when there is a class imbalance (Brodersen, Ong, Stephan & Buhmann, 2010).

Results Using Priming Effect Size

We begin with prediction results using the PEs as the features. Figure 1 shows the mean and standard deviation (SD) of the balanced accuracy as a function of the *m* best PE features used to train the SVM. We achieve the highest accuracy of 0.68 (SD = 0.11) with *m* = 98 features. For comparison, an accuracy of 0.62 (SD = 0.10) is achieved using the average PEs as features.

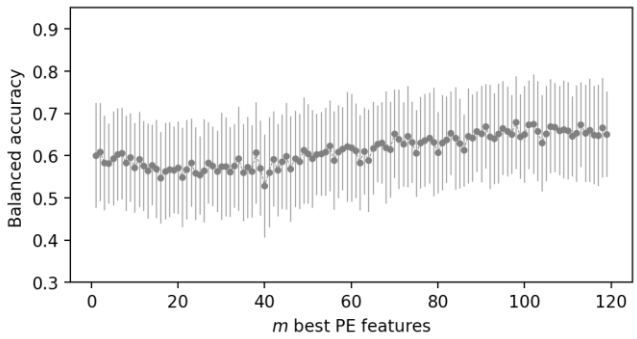

Figure 1: Mean and standard deviation of the balanced accuracy as a function of the *m* best priming effect (PE) features used to train the SVM. Best performance (mean accuracy = 0.68) is achieved at *m* = 98.

In Table 2, we also show the mean and the standard deviation (parentheses) of the confusion matrix for the best-performing model using *m* = 98 features. The confusion matrix shows the class-level prediction accuracy. From these results, we can see that the model predicts the non-proficient participants with slightly higher class accuracy than the proficient participants.

Table 2. Mean and standard deviation (parentheses) of the confusion matrix for the best performing PE model.

		Predicted Group	
		Spanish Proficient	Non-proficient
Actual Group	Spanish Proficient	0.61 (0.16)	0.39 (0.16)
	Non-proficient	0.26 (0.16)	0.74 (0.16)

We would also like to understand how the different PEs contributed to the proficiency prediction of the SVM. Figure 3 plots the mean values for the two metrics for significance for each of the 119 PE features. The SVM weights correspond to weights *after* feature selection. If a feature is not chosen it is given a weight of 0. In general, the features with the highest SVM weights also have small *p*-values. This result supports the intuition that features with lower *p*-values should also contribute more predictive power (higher weights) to the SVM model. Interestingly, three of the top four most predictive features (by either metric) correspond to the words CUELLO, LLUIVA, and PILLOW. All three of these words are six letters long and contain the digraph 'll,' which was considered to be a distinct letter in the Spanish alphabet prior to 2010 (Real Academia Española, 2010). In our word length task, participants were asked to press one button for words that were five letters or shorter and another for words that were six letters or longer. Given this task and the relatively recent removal of 'll' from the Spanish alphabet, these three words may have been tricky for the proficient Spanish speakers. It is notable that the model identified these three stimuli as the ones that were most effective for differentiating between the two groups of participants.

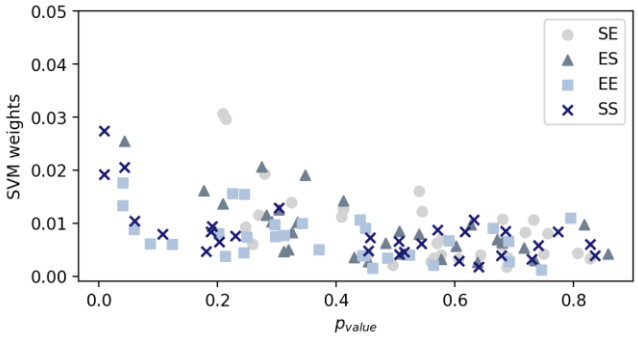

Figure 3: Priming effect (PE) feature significance. Features with low *p*-values (significant according to the univariate statistical test) and high coefficients (significant according to the model) are the most predictive.

Results Using Response Times (RTs)

Next, we repeat our analysis using response times (RTs) as features for the SVM. Figure 4 and Table 3 show the prediction performance the SVM classifier using RTs as features. Overall, we achieve better performance using RTs, compared with using PEs, as features. We achieve the highest balanced accuracy of 0.75 (SD = 0.09) with *m* = 175 features. For comparison, a balanced accuracy of 0.66 (SD = 0.11) is achieved using the average RTs as features.

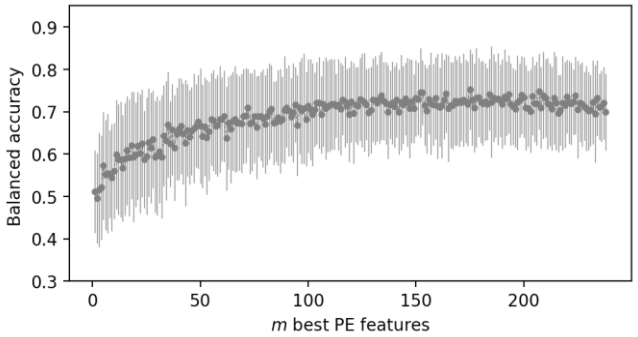

Figure 4: Mean and standard deviation of the balanced accuracy as a function of the *m* best response time (RT) features used to train the SVM. Best performance (mean accuracy = 0.75) is achieved at *m* = 175.

Table 3. Mean and standard deviation (parentheses) of the confusion matrix for the best performing model.

		Predicted Group	
		Spanish Proficient	Non-proficient
Actual Group	Spanish Proficient	0.74 (0.15)	0.26 (0.15)
	Non-proficient	0.23 (0.13)	0.77 (0.13)

As with the PEs, we would like to understand how the individual features contribute to the ability of the SVM to predict participant proficiency. Figure 5 plots the *p*-value and the mean SVM weight for each of the 239 RT features. Once again, in general, RT features with higher SVM weights have smaller *p*-values, indicating that features with lower *p*-values tend to contribute more predictive power (higher weights) to the SVM model.

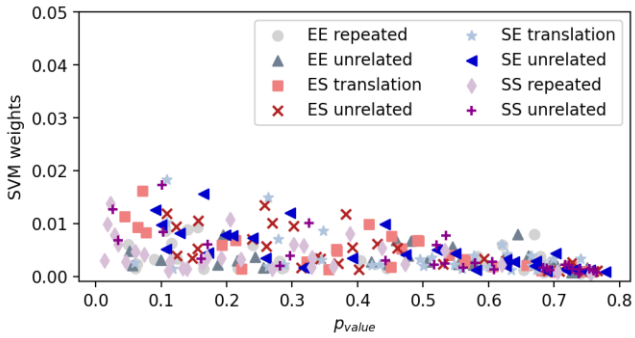

Figure 5: Response time (RT) feature significance. Features with low *p*-values (significant according to the univariate statistical test) and high coefficients (significant according to the model) are the most predictive.

We also examined which participants were misclassified in the highest-performing version of the model. Interestingly, there were six proficient Spanish speakers who reported that they started learning English before learning Spanish and that English was their dominant language. Four of those participants were consistently misclassified by the model, which placed them in the non-proficient group 90-100% of the time. Another participant in this group was misclassified 30% of the time. Only one participant in this category was always classified as being proficient in Spanish, and that was also the only participant who reported that they learned both English and Spanish beginning in infancy. The others in this subset began learning Spanish later in childhood or as teenagers. Although some of these participants may have simply overstated their Spanish proficiency, this pattern suggests that age of acquisition could be a key factor in the RT effects that are identified by the model.

Discussion

This study employed a repetition and translation priming paradigm to test the efficacy of using machine learning techniques to characterize an individual's language proficiency based on priming data. Our analyses showed within-language repetition effects for both languages, with priming effects that were larger for proficient Spanish speakers. However, we observed no priming effects for translations, suggesting that our effects were driven by the wordform and/or response priming, rather than semantic priming. On the surface, these findings may provide weak support for the RHM model (Kroll & Stewart, 1994) because we do not see facilitation between translation equivalents even for people who are proficient in both languages. However, our experimental paradigm and non-semantic task may have encouraged shallow processing. Unlike a classic priming paradigm, where participants see a prime and then respond to a subsequent target, the participants in our task responded to every word with no differentiation between primes and targets. Due to this design and the intermixing of within-language and cross-language pairs, the participants may have been less likely to make predictions about which word would come next, which could reduce the effect of semantic priming. In future research, we plan to test blocked designs where all of the targets in each block are in the same language and a more traditional priming paradigm in which participants passively read the primes and respond only to the targets. We predict that those changes to the experiment structure will produce larger semantic priming effects for proficient bilingual participants reading translated pairs.

Our machine learning analyses showed that a model trained on reaction time data and priming data can predict whether an individual participant is proficient in Spanish with high accuracy. Interestingly, for this dataset, predictions based on priming effects were slightly less successful than predictions based on the RTs alone (68% versus 75% prediction accuracy). Even though the experimental task may have encouraged shallow processing, the participants who acquired Spanish beginning in infancy displayed patterns of response times that differentiated them from the other participants. The model also revealed specific words that were more predictive of proficiency than others, indicating that this approach could also be fruitful for item analyses.

This study has several limitations. Most importantly, we based the proficiency labels on the self-reports on anonymous online participants. The majority of the participants (39 of 42) who reported high proficiency in Spanish also had a Spanish fluency qualification from Amazon Mechanical Turk, which provides some external verification of their proficiency. However, it is not clear what criteria are used to assign that qualification. In future research, it would be useful to assess the model's performance against measures of language proficiency that are more objective than self-reporting.

The word length judgment task that we used also has limitations. We were constrained to using a task that all participants could complete whether they understood Spanish or not. In future work, we aim to develop new tasks that can be completed without knowledge of the target language but that encourage semantic processing.

Overall, this study demonstrates that machine learning techniques can support a more individualized approach to

data analysis in studies of bilingualism or other individual differences. Rather than simply averaging data from all of the participants within each group and comparing the two groups, the ML approach allows us to develop a predictive model to classify participants based on their language proficiency, as instantiated in the data they produced. This can be used to identify groups of participants with different proficiency levels, rather than assigning participants to groups in advance, or to explore differences among participants with similar levels of proficiency. Finally, machine learning can be used to identify the specific stimuli that are most predictive of participant proficiency. All of these factors enable new approaches to the study of bilingualism.

Acknowledgements

This work was supported by the Laboratory Directed Research and Development (LDRD) Program at Sandia National Laboratories. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

The authors would like to thank Breannan Howell for her assistance with data collection and Michael Trumbo and Mikaela Armenta for their helpful feedback on the task.

References

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (pp. 144-152).

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010, August). The balanced accuracy and its posterior distribution. In 2010 *20th International Conference on Pattern Recognition* (pp. 3121-3124). IEEE.

Chalnick, A., & Billman, D. (1988). Unsupervised learning of correlational structure. *Proceedings of the Tenth Annual Conference of the Cognitive Science Society* (pp. 510-516). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press.

Dijkstra, A. F. J., & Van Heuven, W. J. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition, 5*(3), 175-197.

Feigenbaum, E. A. (1963). The simulation of verbal learning behavior. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought.* New York: McGraw-Hill.

Forster, K. I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 680.

Grainger, J., & Frenck-Mestre, C. (1998). Masked priming by translation equivalents in proficient bilinguals. *Language and Cognitive Processes, 13*(6), 601-623.

Hill, J. A. C. (1983). A computational model of language acquisition in the two-year old. *Cognition and Brain Theory*, *6*, 287-317.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning.* New York: Springer.

Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, *33*(2), 149-174.

Kroll, J. F., Van Hell, J. G., Tokowicz, N., & Green, D. W. (2010). The Revised Hierarchical Model: A critical review and assessment. *Bilingualism*, *13*(3), 373.

Marian, V., Bartolotti, J., Chabal, S., Shook, A. (2012). CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities. *PLoS ONE 7(8)*: e43230.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research, 50(4)*, 940-967.

Martin, C. D., Dering, B., Thomas, E. M., & Thierry, G. (2009). Brain potentials reveal semantic priming in both the 'active' and the 'non-attended' language of early bilinguals. *NeuroImage*, *47*(1), 326-333.

Matlock, T. (2001). *How real is fictive motion?* Doctoral dissertation, Psychology Department, University of California, Santa Cruz.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825-2830.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Ohlsson, S., & Langley, P. (1985). *Identifying solution paths in cognitive diagnosis* (Tech. Rep. CMU-RI-TR-85-2). Pittsburgh, PA: Carnegie Mellon University, The Robotics Institute.

Real Academia Española (2010). *Ortografía de la lengua Española*. Espasa.

Schoonbaert, S., Duyck, W., Brysbaert, M., & Hartsuiker, R. J. (2009). Semantic and translation priming from a first language to a second and back: Making sense of the findings. *Memory & Cognition*, *37*(5), 569-586.

Shrager, J., & Langley, P. (Eds.) (1990). *Computational models of scientific discovery and theory formation*. San Mateo, CA: Morgan Kaufmann.

Van Hell, J. G., & Tanner, D. (2012). Second language proficiency and cross-language lexical activation. *Language Learning*, *62*, 148-171.