Running head:
English accents and their determinants

Full title:
English accents and their determinants*

Authors:
Martijn Wieling[a], Jelke Bloem[b], Kaitlin Mignella[a], Mona Timmermeister[c], R. Harald Baayen[d,e] and John Nerbonne[a,f]

Affiliations:
[a]Department of Humanities Computing, University of Groningen
[b]Department of Dutch Linguistics, University of Amsterdam
[c]Department of Pedagogical and Educational Sciences, Utrecht University
[d]Department of Quantitative Linguistics, University of Tübingen
[e]Department of Linguistics, University of Edmonton
[f]Freiburg Institute of Advanced Studies, University of Freiburg

*We are very grateful to Mark Liberman for his post on Language Log inviting native U.S. American English speakers to rate the speech samples for the native-likeness.

Address for correspondence:
Martijn Wieling, University of Groningen, Department of Humanities Computing, Oude Kijk in ‘t Jatstraat 26, 9712 EK Groningen, The Netherlands,
wieling@gmail.com

Abstract

In this study we investigate determinants of the strength of foreign accents in English pronunciation. We use pronunciation data from more than 800 speakers with a variety of language backgrounds and analyze the data with an eye to assessing the presence of a critical period in second language learning. In our dataset, speakers with a non-Indo-European native language had a clear breakpoint at the age of 6, whereas speakers with an Indo-European background had only a minor breakpoint around the age of 16. However, resampling the data in an attempt to verify the results showed that both language groups showed a mirrored bimodal pattern. In sum, our study does not support the existence of a stable critical period within which a second language can be learned with a high degree of proficiency, but rather a complex interaction between various social, educational and maturational factors.

Key words

Second language learning, Critical period hypothesis, Piecewise regression, Mixed-effects regression, English pronunciation

1. Introduction

There is no doubt that the age at which one starts to learn a second language (L2) is an important factor in the final L2 proficiency of the speaker. According to Ellis and Bogart (2007) "there is a clear inverse correlation relating age and L2 ultimate attainment of r = -0.6 to -0.8 across studies," and Long (1990) proposes that it be taken as axiomatic that L2 speakers never reach native ability (acknowledging that there may be very rare exceptions). It is also clear that the ultimate success in pronouncing the L2 is not determined solely by the time of first exposure to the L2, but may also be influenced by factors such as L1 background (Suter, 1976; Purcell and Suter, 1980), the frequency of use of the first language (Flege, Frieda & Nozawa, 1997; Piske, MacKay and Flege, 2001), and length of residence in the L2 country (Bialystok, 1997; Rasinger, 2007, p. 161).

The critical period hypothesis (CPH; Penfield and Roberts, 1959; Lenneberg, 1967) states that there are maturational constraints limiting the age at which a native language has to be acquired in order to master the language. The CPH has also been extended to second language acquisition and posits that after a certain maturational point L2 learners are not able to reach native-like proficiency in the L2. In general this hypothesis has been operationalized as a non-linear influence of age of L2 onset on the ultimate attainment level of the L2 (compared to native speakers). Several possible shapes of the non-linear relationship have been considered (Birdsong, 2006), but a common recent view is that L2 acquisition is characterized by a steady decrease in ultimate attainment with increasing age of acquisition until the end of the critical period, followed by the (almost complete) absence of an effect of age of acquisition on ultimate attainment of the L2 (DeKeyser, Alfi-Shabtay & Ravid, 2010). Importantly, DeKeyser et al. (2010) accept that there is a monotonically decreasing

relationship between language proficiency and the age of acquisition. We return to the issue of how the idea of a monotonic decrease ought to interact with the CPH briefly below.

In their influential paper, Johnson and Newport (1989) argued for the existence of a critical period in L2 acquisition on the basis of age effects before the critical period (which they set at puberty), but not after the critical period. However, Birdsong and Molis (2001) were not able to corroborate these findings in a subsequent replication study. Several other studies also failed to find results consistent with the CPH. For example, Flege, Munro and MacKay (1995) found a clear linear effect of age of L2 onset on foreign English accent ratings, with no support for any non-linearity. In addition, Hakuta, Bialystok and Wiley (2003) did not find a (practically important) non-linearity in the effect of age of L2 onset on speakers' self-reported English ability. On the other hand, DeKeyser et al. (2010) provided support for the presence of a critical period at the age of eighteen in two groups of native speakers of Russian. A frequent point of criticism of the studies supporting the CPH in L2 is that the age used to denote the end of the critical period seems to be quite variable (Muñoz and Singleton, 2011; Piske et al., 2001). This naturally impedes attempts to test the CPH.

Another, more severe problem, which occurs frequently in studies both supporting and opposing the CPH (e.g., Bialystok and Miller, 1999; Johnson and Newport, 1989, DeKeyser et al., 2010), is that the analyses employed are generally not suitable for their purpose (Vanhove, 2013). For example, binning of age groups is highly subjective and results in a loss of statistical power. Furthermore, the absence of a significant difference in performance between two (arbitrary) age groups does not imply there is no difference, it may simply indicate that the sample sizes were too

small to detect significant differences. Finally, comparing correlation coefficients (assessing the relationship between the age of L2 acquisition and performance) between two age groups is likewise problematic, as the correlation coefficient *r* does not contain information about the degree of influence, i.e., slope of the line relating attainment to age, but only how much the data points scatter around the line, i.e. the reliability. To see differences in slopes, regression coefficients must be examined. Vanhove (2013) therefore recommends using the piecewise regression technique introduced by Baayen (2008, Ch. 6.4) which explicitly tests for the presence of a non-linearity in the effect of age of L2 onset on the L2 performance.

In general, studies investigating the CPH in L2 have focused on a single second language (mainly English) and only one or at most L1 backgrounds. There are some exceptions (e.g., Stevens, 1999; Chiswick and Miller, 2008), but these studies have used census data, in which the dependent variable was a self-reported measure of English ability. Focusing on a small set of L1 backgrounds severely limits the ability to investigate the contribution of this important variable (Piske et al., 2001). In this study, therefore, we investigate the performance with respect to L2 (English) pronunciation for more than 800 speakers with varying L1 background (over 170 native languages). To test for the presence of a non-linear effect of age of English onset (which would support the CPH), we use the piecewise regression technique (Baayen, 2008, Ch. 6.4).

Recent studies appear to be generous in their attempts to detect critical period (CP) effects in interaction with the monotonic deterioration effects mentioned above (i.e. Flege et al. 1995, and Ellis and Bogart, 2007). Birdsong (2006) notes that all of the curves below (Figure 1) are taken as indications of a CP even though the end of the CP should be accompanied by a *deterioration* in language learning ability. We are

aware that variable interactions can potentially be quite complex in general but fail to see why Birdsong's form A ("the hockey-stick") should be taken to confirm the CPH even if it clearly suggest an age-related discontinuity. Form A suggests that there is a critical period after which the normal deterioration in language learning ability is arrested. We shall not pursue the issue here further, however, and we continue in the tradition of considering a discontinuity in the age of onset effect as supporting the CPH.

A second important hypothesis that has been proposed concerning foreign accents is that accents result from the ENTRENCHMENT of pronunciation patterns used in the first language (Flege et al., 1995; Flege, 2002). In this view, older second language learners are increasingly likely to use the categories and automated, highly coordinated patterns of pronunciation from their first language when pronouncing their second language. Entrenchment also suggests that language structure should influence how strong an accent is. If the first language is structurally similar to English, then retaining its features should distort English pronunciation less than if the first language is structurally very different.

For testing the CPH, one could investigate lexis, morphology, syntax or any other linguistic level. We focus here on pronunciation – phonetics and phonology – because it is notoriously difficult for second-language learners to approach native-like competence in pronunciation (Munro and Mann, 2005; Abrahamsson and Hyltenstam, 2009). Because native-like pronunciation depends on very fine muscular coordination, we think that we are more likely to detect a critical period with respect to pronunciation than with respect to other linguistic abilities (Scovel, 1988). We concede, however, that Scovel's argument establishes only plausibility. Other studies

may show that language proficiency depends differently on age with respect to other linguistic levels, possibly in ways that provide evidence supporting the CPH.

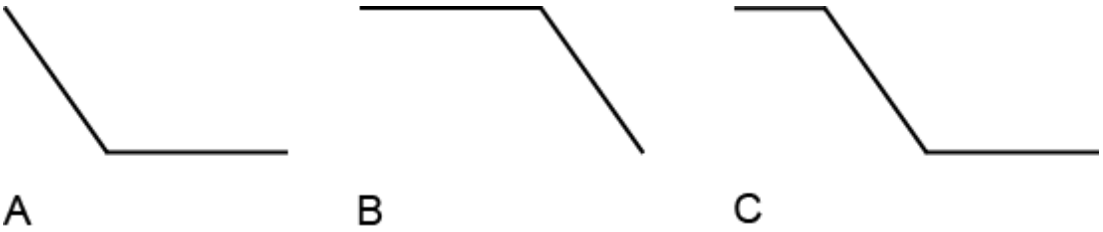

Figure 1. Three curves depicting language proficiency as a function of age of L2 onset (taken from Birdsong, 2006). We note that the curve in A indicates that the general deterioration of language learning ability (left part of curve) appears to be arrested after a critical age, and similarly for the second, lower discontinuity in C. If the CPH claims that language learning abilities deteriorate sharply after a CP, then curve A would appear to contradict the CPH.

2. Material

2.1. The Speech Accent Archive

Our dataset consists of data from the Speech Accent Archive (SAA; Weinberger and Kunath, 2011). The SAA is available at http://accent.gmu.edu and contains a large sample of speech samples in English from people with various language backgrounds. Each speaker reads the same paragraph containing 69 words in English:

Please call Stella. Ask her to bring these things with her from the store: six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.

Speaker-specific information was also collected and includes native language (people who are balanced bilinguals are excluded), other languages spoken, country of birth, age, gender, age of English onset (AEO; i.e. defined as being the age when first exposed to sustained English language input), cumulative residence length in an English-speaking country (LR), and learning style (i.e. naturalistic or academic). All speech samples are transcribed according to the International Phonetic Alphabet.

In 2010, we extracted all available 989 transcribed samples from the Speech Accent Archive including speaker information. As there were only three speakers who were younger than 18, we excluded these from the dataset. Of all 986 adult speakers, 180 were native speakers of English, of which 115 were born in the United States. In this study, we focus on the 806 non-native English speakers. There were slightly more men (439: 54.5%) than women (367: 45.5%) in this dataset. The average age of these speakers was 32.7 (SD: 12.3). The average age of English onset of these 806 non-native English speakers was 12.3 (SD: 7.4), while the mean residence length in an English-speaking country of these speakers was 7.7 years (SD: 11.7). A minority of the non-native English speakers (11.7%) learned English in a naturalistic (as opposed to an academic) setting.

2.2. Additional information

In addition to the speaker-related information available in the Speech Accent Archive, we obtained information with respect to the countries in which the speakers were born. For each country, we obtained the population size, the gross national income (per inhabitant), and the average number of years of education in 2011 (UNDP, 2011, Statistical annex). Furthermore, we grouped the language(s) of the speakers in Indo-

European (IE) languages versus non-Indo-European (non-IE) languages on the basis of the Ethnologue (Lewis, 2009). There were a total of 171 different native languages in our dataset. We undertook the division into IE versus non-IE in order to obtain a first, rough indication of the effect of the L1 on L2 learning. Thus, we used a genealogical division as a proxy for a variable that would best be construed as typological. Ideally one would identify which typological features were likely to be influential (use of tone, aspiration, diphthongization, size of vowel inventory, etc.) and test for those directly, but this would have gone beyond the scope of the present study.

3. Methods

As it is obviously not feasible to obtain the degree of foreignness (i.e. foreign accent ratings) for all 986 speech samples on the basis of elicited native speakers' judgments, we use an automatic method to calculate these ratings. We then use these ratings as the dependent variable in a mixed-effects regression model, where we explicitly test if the effect of age of English onset on these ratings is non-linear (which would be essential evidence for a critical period; Bialystok and Miller, 1999).

3.1. Automatically calculating foreignness ratings

The Levenshtein distance (LD) algorithm is able to calculate pronunciation distances between two transcribed strings by calculating the number of substitutions, insertions and deletions to transform one string into the other (Levenshtein, 1965). For example, the Levenshtein distance between two accented pronunciations of the word Wednesday, [wɛnzdeɪ] and [wɛnəsde] is 3 as can be seen in the alignment below:

w	ɛ	n		z	d	e	ɪ
w	ɛ	n	ə	s	d	e	
			1	1			1

The Levenshtein distance has been successfully used for comparing pronunciations in dialectometry (Kessler, 1995; Heeringa, 2004; Nerbonne and Heeringa, 1997; Wieling, Heeringa & Nerbonne, 2007; Wieling, 2012) and matches perceptual dialect distances well (Gooskens and Heeringa, 2004). Unfortunately, the basic Levenshtein distance algorithm is quite crude and only distinguishes same from different sounds (i.e. substituting two completely different sounds, such as [u] and [ɛ] is not distinguished from substituting two more similar sounds such as [u] and [o]). To make the pronunciation comparison procedure more linguistically sensible, Wieling, Prokić and Nerbonne (2009) proposed a method to incorporate (automatically obtained) sensitive sound distances in the Levenshtein distance algorithm and showed that this approach improved the alignment quality significantly. The procedure is based on calculating the Pointwise Mutual Information (PMI; Church and Hanks, 1990) and works by counting how often two segments correspond in alignments and comparing this to how often they would correspond by chance. Segments which correspond more frequently than would be expected get a low distance, while the distance is high for segments which correspond less frequently than expected. Wieling, Margaretha and Nerbonne (2012) then showed that the underlying sound (vowel) distances were linguistically sensible and corresponded well to acoustic vowel distances, with correlations ranging from $r = 0.63$ to $r = 0.76$ for several datasets. Applying this method to our example alignment yields the following

associated costs (and a total pronunciation distance between the two pronunciations of 0.081):

w	ɛ	n			z	d	e	ɪ
w	ɛ	n	ə		s	d	e	
			0.031		0.020			0.030

Wieling et al. (forthcoming) showed that the PMI-based Levenshtein distance is a valid measure of how native-like accented pronunciations are. Using audio samples from the Speech Accent Archive, they obtained human native-likeness ratings for 286 speech samples. In their study, 1143 participants judged 41 speech samples on average, resulting in consistent judgments (Cronbach's alpha: 0.85). For each of the 286 distinct transcribed speech samples, the PMI-based Levenshtein distance was calculated with respect to the transcriptions of 115 speech samples of native American English speakers. Subsequently, these 115 distances were averaged and represented the distance from that speaker to the "average American English speaker". Wieling et al. (forthcoming) reported a correlation between the PMI-based Levenshtein distance and the human native-likeness judgments of r = -.78 (p < .001). When log-transforming the Levenshtein distances, this correlation increased to r = -.81 (p < .001). The correlation is negative as higher native-likeness implies a lower pronunciation distance. Given that this correlation was also very close to how well individual raters agreed with the average native-likeness ratings (r = .84, p < .001; Wieling et al., forthcoming), the PMI-based Levenshtein distance can be used as a valid measure of non-native-likeness (i.e. the strength of foreign accent).

We use the Levenshtein distance as a measure of pronunciation difference in accents because it is sensitive to all the segmental variations that accents are associated with, not merely typical ones such as [t]:[θ] or [s]:[θ], and because it is sensitive to the frequency with which segments are inserted, deleted or modified. It thus provides a global measure of difference in segmental realization. In the following we will use this measure of foreign accent strength as our dependent variable.

3.2. Testing for a non-linearity of age of English onset: Breakpoint analysis

To test for a non-linear effect of age of English onset (AEO) on foreign accent strength, we followed the piecewise regression approach described by Baayen (2008, Ch. 6.4) and recommended by Vanhove (2013). In an iterative procedure, this procedure finds the best possible breakpoint (i.e. the age of English onset after which the effect of this predictor is different than before this point). Subsequently, the model including this breakpoint is compared to the simpler model without a breakpoint. If the former is an improvement over the latter, this indicates the relationship between age of English onset and the foreignness ratings is non-linear. In our procedure we considered breakpoints at an age of English onset between 1 and 30 years.

As there may be structural variability linked to the countries and/or languages of the speakers, we opted for mixed-effects regression modeling (see e.g., Baayen, Davidson & Bates, 2008; Baayen, 2008, Ch. 7), which is able to take into account the (possible) structural variability linked to country of birth and native language. We tested all possible random slopes and intercepts, and only included those which provided a significant improvement in goodness of fit. We assessed the improvement in fit by comparing the value of the Akaike Information Criterion (Akaike, 1974) of both models. A reduction of at least 2 indicates that the higher complexity of the new

model is warranted. The mixed-effects regression approach has been used previously in combination with the Levenshtein distance (for dialectal pronunciations) by Wieling, Nerbonne and Baayen (2011).

In the following section, we discuss the determinants of the automatically determined foreign accent ratings. The analyses and results described in the following section may be reproduced with the paper package (including the data, R analysis codes, graphs and numerical results) accompanying to this manuscript. The paper package can be downloaded from the Mind Research Repository (http://openscience.uni-leipzig.de) or the first author's website.

4. Results

We used the log-transformed PMI-based Levenshtein distance between the average American English speaker and each of the 806 non-native English speakers as our dependent variable. While our results showed clear support for the inclusion of country of birth as a random-effect factor (i.e. having structural variability associated with it), this was not the case for native language (an AIC *increase* of 0.4). This might seem strange at first sight, but there can be much variation between speakers of the same language in different countries. For example, Canadian French is significantly different from continental French (Walker, 1984) and their English accents may be as well. In addition, we tested, but did not find support for by-country random slopes.

Table 1 shows all significant factors and covariates of the best model for our data on the basis of 805 speakers. One speaker was excluded, as the residuals of the initial fitted model revealed a single extreme outlier during the model criticism phase (Baayen, 2008). Excluding this outlier increased the explained variance of our model from 42.5% to 43.6%. To compare the relative effect of each predictor fairly, we

added a measure of effect size by specifying the increase or decrease of the dependent variable when the predictor increased from its minimum to its maximum value (following the approach of Baayen et al., 2008). Besides these fixed-effect predictors, the random-effect structure only consisted of a random intercept per country. We assessed if there were additional interactions, but none improved the fit of the model shown in Table 1. We first discuss the predictors other than age of English onset (and its interactions), which serve as controls to rule out potential confounding explanations.

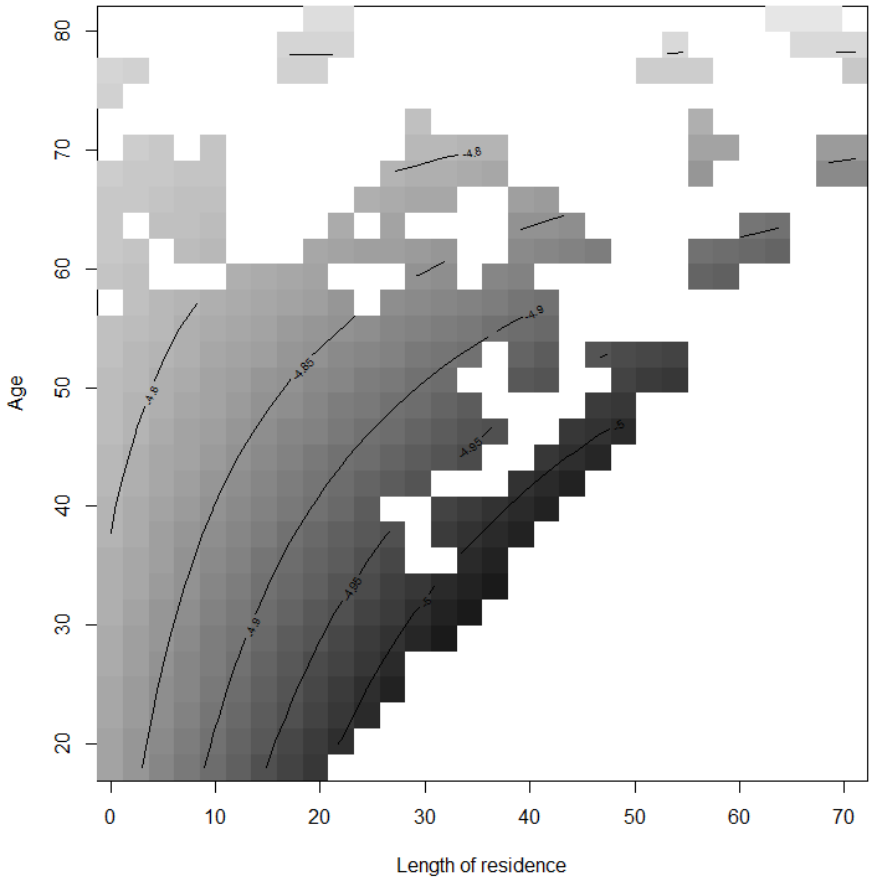

Figure 2. Interaction between length of residence and age. Darker shades of gray indicate a pronunciation closer to that of an average native American English speaker. The beneficial effect of increasing length of residence clearly diminishes with age, as the black contour lines are further apart for increasing age.

Length of residence and speaker age interacted significantly as can be seen in Figure 2. Darker colors in this graph represent pronunciations which are more similar to average native American English, whereas lighter values indicate the opposite. Clearly the beneficial effect of length of residence is dependent on age. For younger speakers a longer length of residence has the strongest effect (e.g., for a 20-year old speaker the lines are only a short distance apart), whereas the effect is smaller for older speakers (e.g., for a 50-year old speaker the lines are further apart). The significance of the individual predictors indicates that the effect of length of residence is significant for the mean value of speaker age, and vice versa.

We observe a clear effect of the number of languages spoken besides English. The more languages spoken, the more native-like the pronunciation of the speaker. In addition, the average number of years of education per country was a significant predictor, with speakers from countries with longer average education having a more native-like American English pronunciation. As this variable correlated highly, r = 0.83, with the average gross national income, this measure does not necessarily reflect education only, but also incorporates the wealth of a country. We did not observe a significant effect of natural as opposed to academic learners, gender or population size.

We are now in the position to consider whether the effect of age of English onset (AEO) changes early in the lifetime. Using the aforementioned breakpoint analysis, we initially identified a breakpoint at an AEO of 6 yielding an improved fit of the data compared to the model excluding the breakpoint. The model with the breakpoint was better than the simpler model assuming a linear effect of AEO, as the AIC reduction was equal to 11.2. Figure 3 illustrates the location of the initial

breakpoint by visualizing the non-linear relationship between age of English onset and the log-transformed PMI-based Levenshtein pronunciation distance.

	Estimate	Std. err.	*t*-value	Eff. size
Intercept	-4.62501	0.02634	-175.61	
Length of residence (centered)	-0.00617	0.00087	-7.05	
Age (centered)	0.00232	0.00065	3.58	
Length of residence:Age (both centered)	0.00013	0.00003	3.80	0.272
Nr. of other languages spoken	-0.01537	0.00552	-2.79	-0.077
Avg. nr. of years of education per country	-0.02028	0.00284	-7.15	-0.232
AEO (shifted, IE speaker, before BP of 16)	0.01427	0.00199	7.18	0.200
AEO (shifted, IE speaker, after BP of 16)	0.00950	0.00171	5.56	0.342
AEO (shifted, non-IE speaker, before BP of 6)	0.04384	0.00961	4.56	0.219
AEO (shifted, non-IE speaker, after BP of 6)	0.00577	0.00122	4.74	0.219

Table 1. Significant fixed-effect factors and covariates of the final model. A positive estimate indicates that a higher value for this predictor increases the strength of the foreign accent, while a negative estimate indicates the opposite effect. Predictors where the absolute *t*-value is greater or equal than 2 are significant ($p < 0.05$). Effect size indicates the increase or decrease of the dependent variable when the predictor value increases from its minimum to its maximum value (i.e. the complete range). The interaction between age and length of residence is visualized in Figure 2. The total effect size of this interaction is based on the combined effect the two variables have on the dependent variable (only values pairs which occur in the dataset are considered to determine the minimum and maximum effect). For the Indo-European (IE) speakers the significant breakpoint was identified at an age of English onset (AEO) of 16. The breakpoint for the non-IE speakers was located at an AEO of 6. The final four lines of table show the varying influence of AEO before and after the breakpoint for each group. The values of AEO were shifted by subtracting the breakpoint (per group) from the original AEO (conform the piecewise regression approach reported in Baayen, 2008, Ch. 6.4). See the text for further details.

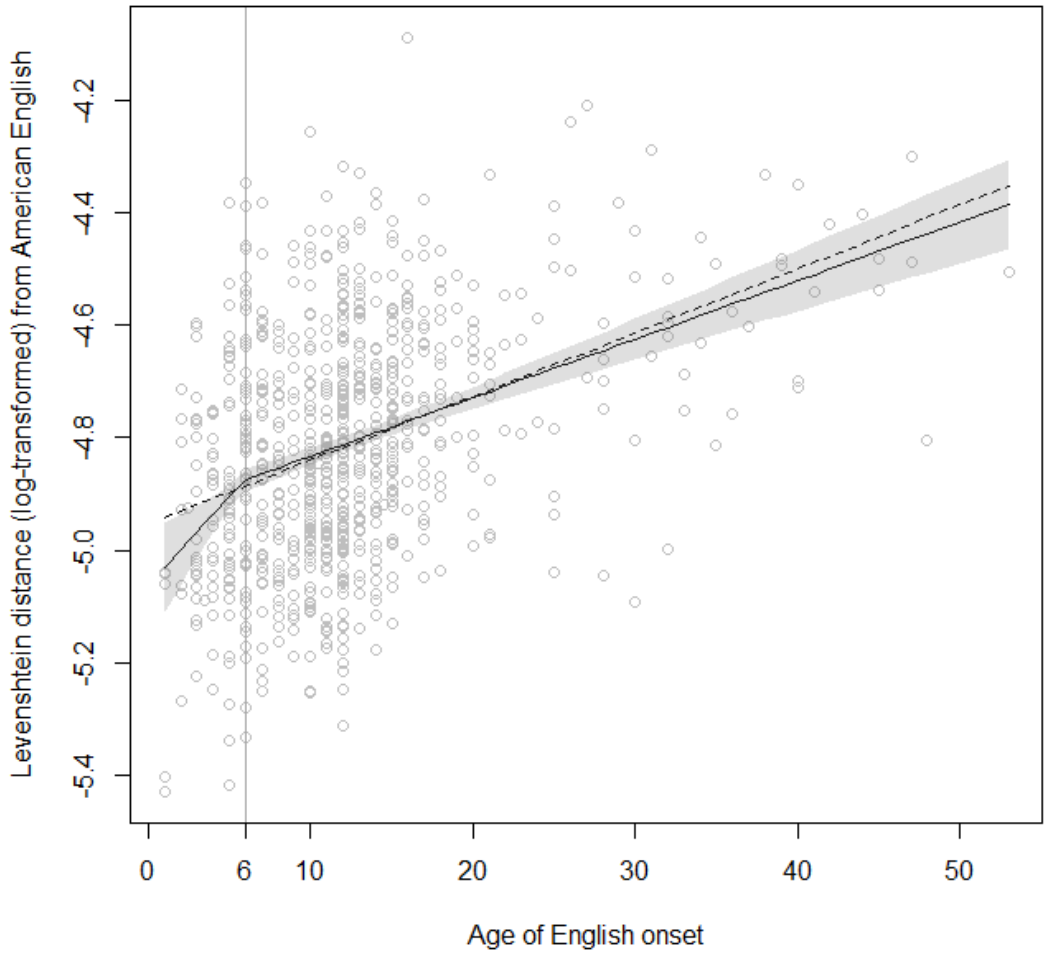

Figure 3. The breakpoint based on the full dataset is located at age of English onset is equal to 6. The shaded area indicates the 95% confidence interval of the solid partial regression lines. The dashed line shows the regression line for a model without breakpoint (providing a worse fit to the data than the model with breakpoint). A higher value of the dependent variable rating indicates pronunciations which are less similar to native American English. Further investigation revealed that the breakpoint was not very stable (see Figure 4).

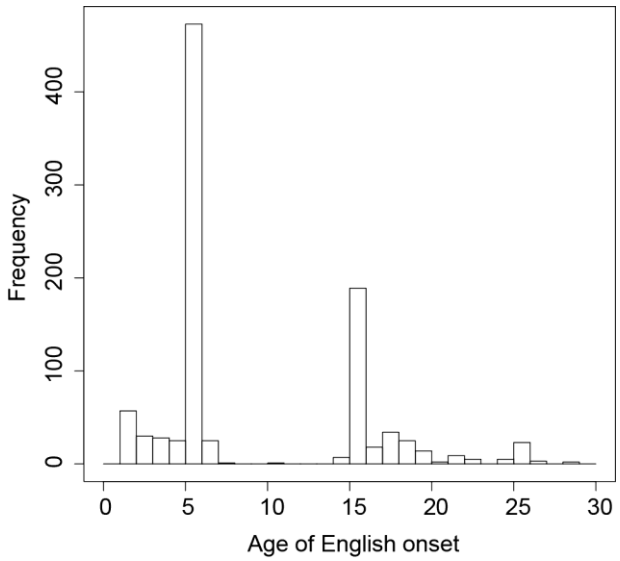

Figure 4. Histogram visualizing the distribution of significant breakpoints in the validation step. There appears to be a separation of two groups: one with a breakpoint around 6 and the other around 16.

Figure 3 also shows that there is considerable variability in pronunciation distance for any given AEO, especially for lower values of AEO. To validate that there is a signal in the noise, we carried out a validation step in which we applied the analysis on 1000 new equal-sized datasets (generated by bootstrapping: random sampling with replacement from the original dataset). In a majority of cases, the breakpoint was indeed located around an AEO of 6. Surprisingly, a sizeable minority of bootstrap runs suggested a breakpoint closer to the age of 16. No significant breakpoint was obtained in only 2.4% of the cases. Figure 4 shows a histogram of this distribution.

We investigated whether the two breakpoints (one around 6, and one around 16) might reflect differences between language groups. We therefore modified the breakpoint analysis in such a way it allowed for two separate breakpoints: one for the Indo-European speakers (438 speakers: 54.3%) and one for those having a non-Indo-European native language (368 speakers: 45.7%).

This analysis resulted in two separate breakpoints (included in our final model reported in Table 1). For Indo-European speakers the optimal breakpoint was located at an AEO of 16, while it was still equal to 6 for the non-Indo-European speakers. The new model resulted in an improved fit to the data compared to the model with only a single breakpoint (the AIC was reduced by 30.6). Figure 5 visualizes the two separate breakpoints. While the breakpoint seems almost unnecessary for the Indo-European speakers (see Figure 5, right), it is supported by a reduction in the AIC of 3.9. The horizontal dotted lines show the average (log-transformed PMI-based) Levenshtein distance. Note that the higher Levenshtein distance for the non-Indo-European speakers compared to the Indo-European speakers is implicit in our model (shown in Table 1) as the piecewise regression technique (Baayen, 2008) involves shifting the age of English onset values such that the breakpoint is positioned at an AEO of 0 (i.e. the breakpoint value is subtracted from the original AEO values). As the two groups (IE vs. non-IE) have a different breakpoint, this means that non-Indo-European speakers with an age of English onset of 6 are compared to Indo-European speakers having an age of English onset of 16. Obviously, this is not an informative comparison. However, a clear significant difference between the two groups ($p < 0.001$) is observed for a model including only a linear effect of (the unshifted) age of English onset for both groups.

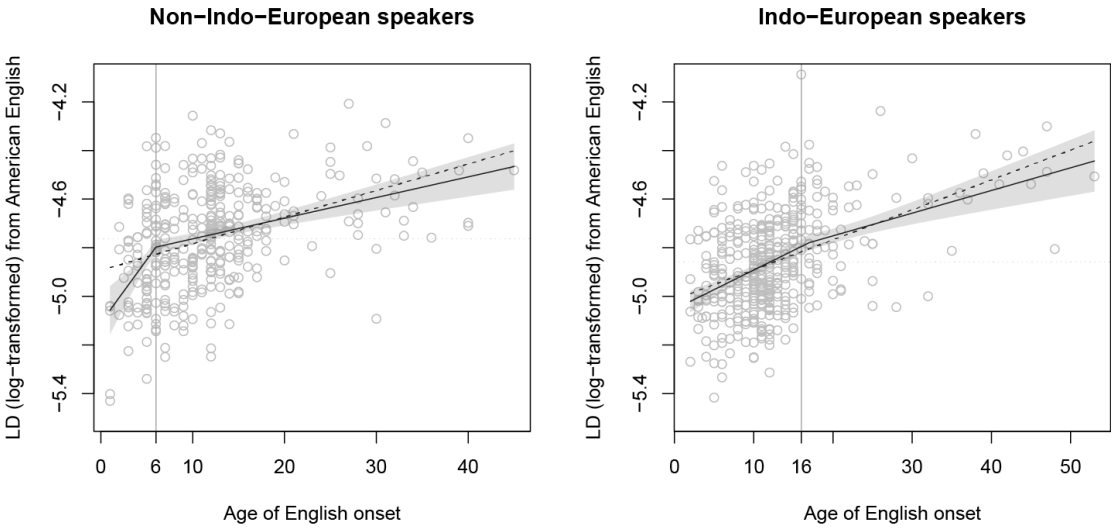

Figure 5. The breakpoints for the non-Indo-European speakers (left; age of English onset: 6) and the Indo-European speakers (right; age of English onset: 16) are marked by the vertical bar. The dotted horizontal lines indicate the average Levenshtein distance (LD) per group (significantly higher for the non-Indo-European speakers: their pronunciations are less similar to the average native American English pronunciation than those of the Indo-European speakers). The shaded area indicates the 95% confidence interval of the solid partial regression lines. The dashed line shows the regression lines for a model without breakpoints (which provides a worse fit to the data than the model with breakpoints).

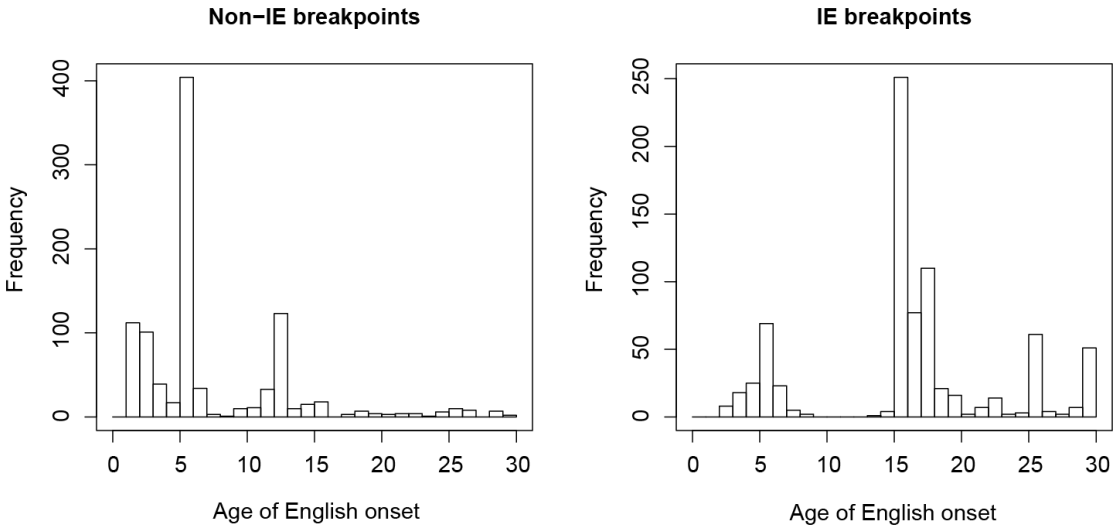

Figure 6. Histogram visualizing the distribution of significant breakpoints in the validation step for Indo-European and non-Indo-European speakers.

In similar fashion as for the single breakpoint, we validated the two breakpoints using bootstrapping. For the non-Indo-European speakers a breakpoint was significant in 99% of the cases, but for the Indo-European data the breakpoint was significant in 78.3% of the cases, which is not surprising given the small change in the slope of the regression line before and after the breakpoint. Figure 6 summarizes the distributions of the breakpoints. For non-Indo-European speakers, most of the breakpoints are at an AEO of 6, but a minority is around 12. For Indo-European speakers, the pattern reverses, with a minority at 6 and a majority around 16. In other words, both language groups show a bimodal pattern, with mirrored locations for the primary modes. For both groups, we are dealing with mixture distributions, with both early (majority for non-IE, minority for IE) and late (minority for non-IE, majority for IE) breakpoints.

An indication of what might be at issue here is provided by an ancillary analysis (model not shown here, but included in the paper package) in which the breakpoints are completely removed from the model specification. In this analysis, a contrast emerged that did not reach significance in the analysis with the two breakpoints. This contrast concerned natural as opposed to academic learners. Speakers who learned English in a natural setting (a minority) had a more native-like American English pronunciation than those who learned English in an academic setting. The removal of breakpoints primarily affects the non-Indo-European speakers (see Figure 5 where the dashed line is relatively far away from the piecewise regression lines for the non-Indo-European speakers). For this language group, the pronunciation distance is overestimated for values of AEO below 6. This overestimation is compensated for in the analysis without breakpoints by means of a global downward shift for L2 learners acquiring English in a natural setting.

Especially at an early age of English onset, our data shows that non-Indo-European speakers are more likely to learn English in a natural setting.

To clarify whether our results might be contingent on the use of our automatic pronunciation method, we also used the (log-transformed) average native-likeness scores assigned by native American English speakers as the dependent variable (using a subset of 272 samples of non-native English speakers reported by Wieling et al., forthcoming). The initial breakpoint analysis identified breakpoints comparable to those in the analysis based on the Levenshtein distance (i.e. a breakpoint at 6 for the 125 non-Indo-European speakers and a breakpoint of 17 for the 147 Indo-European speakers). The bootstrapping procedure revealed that the breakpoint for non-Indo-European speakers was significant in 98.3% of cases and relatively robust. In contrast, the breakpoint for the Indo-European speakers was much more variable and not significant in 24.4% of the cases. Figure 7 shows the breakpoint distributions. The main difference is the absence of a second minor mode for the non-Indo-European distribution. This is probably due to human ratings being based not only on segmental overlap, but also on the fine phonetic detail concerning the realization of stress, pitch, and segmental duration. The final model for the native-likeness ratings was comparable to the one reported in Table 1. The only difference (besides the slightly different distributions of breakpoints obtained via the bootstrap procedure) was that natural learners were more likely to have a more native-like pronunciation than academic learners (t = 2.2). (For brevity, the model is not shown here, but included in the paper package.) In sum, results on the basis of the native-likeness judgments generally pointed in the same direction as the results on the basis of the log-transformed PMI-based Levenshtein distance.

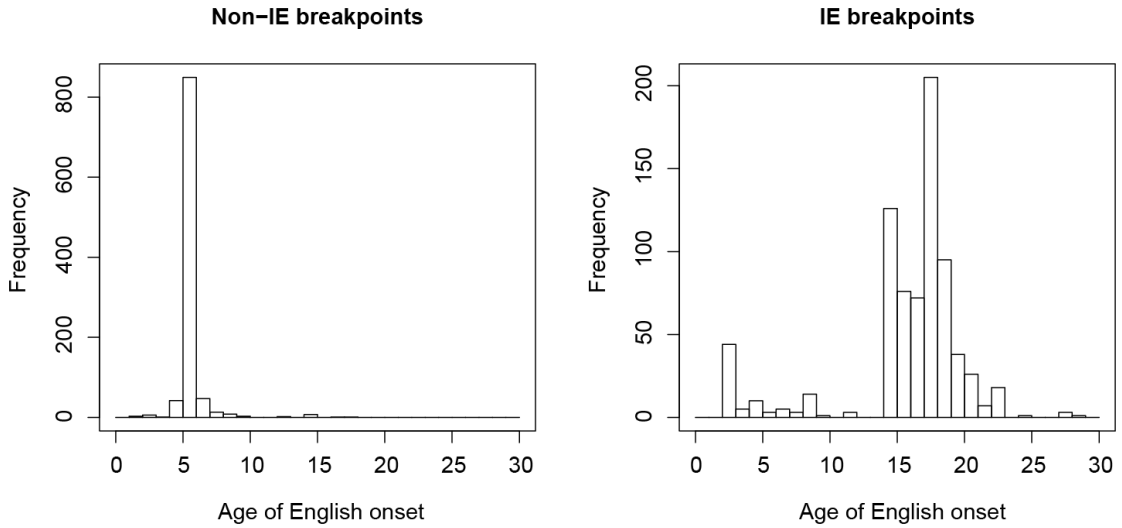

Figure 7. Histogram visualizing the distribution of significant breakpoints in the validation step for Indo-European and non-Indo-European speakers (on the basis of the native-likeness judgment data).

5. Discussion

In this study we have used piecewise regression (Baayen, 2008) applied to a large set of pronunciation data of more than 800 non-native speakers of English and a validated computational measure of pronunciation difference[1] to test whether we could identify a specific age after which the influence of age of English onset was different than before. The presence of such a breakpoint could be considered evidence for the presence of a critical period in second language acquisition.

An initial analyses suggested two different breakpoints, one for speakers with an Indo-European background around the age of 16, and one for speakers with a non-Indo-European background located around the age of 6. Subsequent validation with bootstrap samples indicated that the two language groups are likely to be mixture distributions with an early and a late mode. For the non-IE speakers, the dominant mode is early, for the IE speakers, the dominant mode is late. When non-segmental information is included in the evaluation, the two language groups have more distinct

profiles, with an early breakpoint for the non-IE group, and a late breakpoint with only a hint of an early breakpoint for the IE group.

Speakers with a late breakpoint (typically found among the IE L1 speakers) show this breakpoint around the age of 16, which in many countries where IE languages are spoken is at the end of the period of mandatory schooling. This late breakpoint therefore suggests that speakers learning English without the benefit of being taught English at school (i.e. after the age of 16) tend to approximate native English pronunciation less successfully. This might be taken to indicate that the late breakpoint is not a biological or maturational breakpoint, but rather a cultural breakpoint, bearing witness to educational institutions having some success in teaching children how to pronounce English. We return to this issue below.

Speakers with an early breakpoint are typically found among the non-IE L1 speakers. This breakpoint, around 6 years of age, suggests that children coming into contact with English only by the time they go to primary school (around the age of 6 or 7) are at a disadvantage compared to children with experience of English before that age. Irrespective of the precise circumstances under which this very early exposure to English took place, these early learners have a marked advantage compared to the later learners. This early breakpoint is compatible with the hypothesis of a critical period for L2 acquisition.

Thus, we are faced with a paradox: an early breakpoint that might fit the critical period hypothesis, and a late breakpoint that appears not to. In fact, the emergence of two breakpoints instead of one might be taken as evidence against a biological critical period for L2 learning. After all, it is unclear why the same species would have different critical periods, depending on the L1 language family.

A further argument against a critical period is that the breakpoint pattern that we observed differs from all three patterns in Figure 1. We see no evidence for a stable state. As AEO increases, pronunciation distance increases. This holds across the full range of AEO values. The breakpoints indicate that before the breakpoint, the effect of a later AEO is more deleterious than after the breakpoint. In other words, there is no evidence of a window in time during which the ability to absorb a second language is stable.

At the same time, the early breakpoints indicate that there is an advantage to coming into contact with English as early as possible. This fits well with Flege's entrenchment hypothesis (Flege et al., 1995; Flege, 2002) and with studies indicating that infants quickly lose sensitivity to subtle phonetic contrasts that do not occur in their speech environment already in their first year (Werker & Tees, 1984). However, since the data available to us post-date early infancy, this early rapid decay in sensitivity to nonnative phonetic contrasts cannot provide a full explanation, even though it may help explain why full native fluency is out of reach for a large majority of L2 learners.

Even if the decay in phonetic sensitivity (as an auxiliary psycho-physiological postulate to entrenchment) cannot be a sufficient explanation for our results, the entrenchment hypothesis is still interesting because we can add to it the plausible postulate that the effect of entrenchment might depend in detail on the speaker's L1. This could help in explaining the result that the speakers of non-IE languages showed a different and stronger maturational break than the IE speakers. If we invoke entrenchment, we might explain this differential effect of L1. This line of reasoning suggests further research. We introduced the distinction Indo-European vs. non-Indo-European as a proxy for structural differences in phonology, even while admitting that

the genealogical distinction was psycholinguistically untested. But the distinction might be serving rough proxy for structural differences, which might be partially hiding a variable with a psycholinguistically potent and differentiating role. This suggests that a study of the same data with a careful treatment of phonological structure in the L1s would be insightful. The strength of the CPH in that case would depend on an interaction between age and the structural similarity of the languages involved.

More important for understanding the present findings are the maturational changes in the frontal lobes that have been found to begin to take place in the fourth year of life. Unlike regions of the brain responsible for motor and sensory processing, as well as speech and language development, the prefrontal cortex (PFC) has a particularly protracted developmental trajectory (Gogtay et al., 2004; O'Hare & Sowell, 2008) well into early adulthood (see also Best, Miller & Jones, 2009). These maturational changes come with a qualitative shift in how learning takes place. Before these maturational changes set in, children are challenged by response selection. Learning is predominantly discriminative (Ramscar & Gitcho, 2007; Baayen et al., 2011), with little guidance from reflection about alternative options. Improvements in the signaling between the PFC and the anterior cingulate cortex (ACC, cf. Yeung, Botvinick & Cohen, 2004) make it possible for children to become aware of and think about different alternatives. Their working memory improves, as do their executive functions. Best et al. (2009) point to several cognitive discontinuities. The ability to inhibit responses improves prominently during preschool years, and changes less once school has been entered. On the other hand, working memory emerges in pre-school years, but real improvement is seen during the school years. The ability to plan ahead develops well only by the time children reach adolescence.

These physiological changes in the brain have important consequences for the learning of a second language. When exposure to the L2 takes place before the development of the PFC and ACC, learning of the L2 will proceed in a way similar to the learning of L1, without conscious reflection. Once the systems for conflict monitoring and resolution develop, children become increasingly aware of the different alternatives offered by different languages for the expression of one's thoughts. Education strengthens specifically children's "prefrontal" skills, changing them even more from exclusively implicit unsupervised learners into partially self-monitoring, supervised learners. As pointed out by Ellis (2006), many aspects of second language learning can be understood from the perspective of discrimination learning. However, the different breakpoints observed in the present study show that qualitative changes in learning may also be involved. The older a child is, the more the maturation of executive functions will enable shortcutting implicit discriminative learning, resulting in patterns of language use that diverge from those in the target L2. Interestingly, Ramscar & Gitcho (2007) argue that the late development of the ACC and PFC in humans enables the emergence of conventionality in language: young children are unable to selectively attend to the input, and selectively control what they are going to say. As a consequence, their language will come to mirror closely that of the speakers in their environment.

These maturational changes shed further light on the breakpoints that we observed in our data. The late breakpoint (around 16 years of age) highlights not only a time at which most learners have completed compulsory education. This is also the time around which the development of the PFC and ACC is nearing completion. As a consequence, late learners not only miss out on the benefits of education, they also

have cognitive skills that stand in the way of faithfully absorbing the language they seek to learn.

By contrast, the early breakpoint (around 6 years of age) indicates a watershed between a period of predominantly implicit discriminative learning with relatively little executive control, and a period in which education and developmental changes combine to strengthen thinking about different options and selecting rationally between them. Although these higher-order cognitive skills are invaluable in modern societies, they adversely affect the learning of a second language.

When the breakpoint is early, we see a strong effect of the maturational changes, indicated by an initial steep pre-breakpoint slope, followed by a relatively flat post-breakpoint slope. When the breakpoint is late, the difference in slope is less pronounced. This suggests that the consequences of the maturational changes in the PFC and ACC in early childhood have more far-reaching consequences than the late consolidation of the PFC and ACC in late adolescence, as expected.

One question for which we do not have a satisfactory answer is why the dominant breakpoint is early for the non-IE speakers, and late for the IE speakers. The distinction between these two groups in the present study is a pragmatic one, based on the intuition that IE speakers will have an L1 that is more similar to English, and that these speakers will live in a country that is culturally more similar to countries where English is the dominant language (UK, USA, Canada, Australia, and New Zealand). We take the effect of the IE/non-IE distinction to primarily reflect distances on these three dimensions (i.e. linguistic distance, geographical distance, and cultural distance). Besides entrenchment (see discussion above), one potential reason for the early breakpoint for non-IE speakers is a difference in quality of education, with a greater emphasis on rote learning and reduced access to courses in which English as

spoken by native speakers is heard. As speakers with an early AEO below 6 years of age in non-IE speaking countries are probably coming from socio-economically privileged families, the large slope for non-IE speakers before the breakpoint may in part reflect a decrease in the quality of the education in English received.

In summary, the present study does not support the existence of a stable critical period within which a second language can be learned with a high degree of proficiency. Instead, our results indicate that proficiency is best understood as the outcome of a complex interaction between socio-economic status, educational practice, and the delayed onset and prolonged maturation of the prefrontal cortex and anterior cingulate cortex.

References

Abrahamsson, N. and K. Hyltenstam (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning*, *59(2)*, 249-306.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19(6)*, 716-723.

Baayen, R. H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge University Press.

Baayen, R. H., Davidson, D.J. and D.M. Bates (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390-412.

Baayen, R. H., Milin, P., Filipovic Durdevic, D., Hendrix, P. and M. Marelli (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review, 118,* 438-482.

Best, J. R., Miller, P. H., and L. L. Jones (2009). Executive functions after age 5: Changes and correlates. *Developmental Review, 29(3),* 180-200.

Bialystok, E. (1997). The structure of age: In search of barriers to second language acquisition. *Second Language Research, 13(2),* 116-137.

Bialystok, E. and B. Miller (1999). The problem of age in second-language acquisition: Influences from language, structure, and task. *Bilingualism: Language and Cognition, 2,* 127-145.

Birdsong, D. (2006). Age and second language acquisition: An overview. In M. Gullberg and P. Indefrey (eds.) *The cognitive neuroscience of second language acquisition,* Blackwell, London, pp. 9-49.

Birdsong, D. and M. Molis (2001). On the evidence for maturational constraints in second-language acquisition. *Journal of memory and language, 44(2),* 235-249.

Chiswick, B. R. and P. W. Miller (2008). A Test of the Critical Period Hypothesis for Language Learning. *Journal of multilingual and multicultural development, 29(1),* 16-29.

Church, K. W., and P. Hanks (1990). Word association norms, mutual information, and lexicography. *Computational linguistics, 16(1),* 22-29.

DeKeyser R., Alfi-Shabtay, I. and D. Ravid (2010) Cross-linguistic evidence for the nature of age effects in second language acquisition. *Applied Psycholinguistics, 31,* 413–438.

Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied Linguistics, 27(1),* 1-24.

Ellis, N. C. and P. Bogart (2007). Speech and language technology in education: The perspective from SLA research and practice. *Proceedings ISCA ITRW SLaTE*. Farmington, PA (USA), pp. 1-8.

Flege, J. E. (2002) Interactions between the native and second-language phonetic systems. In: P. Burmeister, T. Piske, A. Rohde (Eds.), *An Integrated View of Language Development: Papers in Honor of Henning Wode*, Wissenschaftlicher Verlag Trier, Trier, pp. 217-244.

Flege, J. E., Frieda, E. M. and T. Nozawa. (1997). Amount of native-language (L1) use affects the pronunciation of an L2. *Journal of Phonetics, 25(2),* 169-186.

Flege, J. E., Munro, M. J., and I. R. MacKay (1995). Factors affecting strength of perceived foreign accent in a second language. *The Journal of the Acoustical Society of America, 97,* 3125.

Gogtay, N., Giedd, J. N., Lusk, L., Hayashi, K. M., Greenstein, D., Vaituzis, A. C., Nugent III, T. F., Herman, D. H., Clasen, L. S., Toga, A. W., Rapoport, J. L. and P. M. Thompson (2004). Dynamic mapping of human cortical development during childhood through early adulthood. *Proceedings of the National Academy of Sciences, 101(21),* 8174-8179.

Gooskens, C. and W. Heeringa (2004). Perceptive Evaluation of Levenshtein Dialect Distance Measurements using Norwegian Dialect Data. *Language Variation and Change, 16(3),* 189-207.

Hakuta, K., Bialystok, E. and E. Wiley (2003). Critical evidence a test of the critical-period hypothesis for second-language acquisition. *Psychological science, 14(1),* 31-38.

Heeringa, W. (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. PhD Thesis, University of Groningen.

Johnson, J. S. and E. L. Newport (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive psychology, 21(1)*, 60-99.

Kessler, B. (1995). Computational dialectology in Irish Gaelic. In: *Proceedings of the Seventh Conference on European Chapter of the Association for Computational Linguistics*, pp. 60-66.

Lenneberg, E. H. (1967). *Biological foundations of language*. Wiley, New York.

Levenshtein, V. (1965). Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR, 163*, 845-848. In Russian.

Lewis, M. P. (2009). *Ethnologue: Languages of the World.* SIL International, Dallas. Online version: http://www.ethnologue.com.

Long, M.H. (1990). The least a second language acquisition needs to explain. *TESOL Quarterly, 24*, 649-666.

Muñoz, C. and D. Singleton (2011). A critical review of age-related research on L2 ultimate attainment. *Language Teaching, 44(1)*, 1-35.

Munro, M. and V. Mann (2005). Age of immersion as a predictor of foreign accent. *Applied Psycholinguistics, 26(3)*, 311-341.

Nerbonne, J. and W. Heeringa (1997). Measuring Dialect Distance Phonetically In: J. Coleman (ed.) *Workshop on Computational Phonology.* Special Interest Group of the Association for Computational Linguistics. Madrid, pp. 11-18.

O'Hare E. D. and E. R. Sowell (2008). Imaging developmental changes in gray and white matter in the human brain. In: C. A. Nelson and M. Luciana M (eds.) *Handbook of Developmental Cognitive Neuroscience*. Cambridge, MA: MIT Press, pp. 23–38.

Penfield, W., and L. Roberts (1959). *Speech and brain mechanisms*. Princeton University Press, Princeton, NJ.

Piske, T., MacKay, I. R. A. and J. E. Flege (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics, 29(2),* 191-215.

Purcell, E. and R. Suter (1980). Predictors of pronunciation accuracy: A reexamination. *Language Learning, 30,* 271-287.

Ramscar, M. and N. Gitcho. (2007). Developmental change and the nature of learning in childhood. *Trends In Cognitive Science, 11(7),* 274-279.

Rasinger, S. M. (2007). *Bengali-English in East London: A study in urban multilingualism* (Vol. 11). Peter Lang, Switzerland.

Scovel, T. (1988). *A time to speak: A psycholinguistic inquiry into the critical period for human speech*. New York: Newbury House.

Suter, R. W. (1976). Predictors of pronunciation accuracy in second language learning. *Language Learning, 26,* 233-253.

Stevens, G. (1999). Age at immigration and second language proficiency among foreign-born adults. *Language in Society, 28,* 555-578.

UNDP (2011). *Human Development Report 2011. Sustainability and Equity: A Better Future for All*. Palgrave Macmillan, New York.

Vanhove, J. (2013). The critical period hypothesis in second language acquisition: A statistical critique and a reanalysis. *PLOS ONE, 9(7),* e102922.

Walker, D. C. (1984). *The pronunciation of Canadian French*. University of Ottawa Press, Ottawa, ON.

Weinberger, S.H. and S.A. Kunath (2011). The Speech Accent Archive: towards a typology of English accents. *Language and Computers, 73,* 265-281.

Werker, J., and R. Tees (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7*, 49-63.

Wieling, M. (2012). *A Quantitative Approach to Social and Geographical Dialect Variation*. PhD dissertation, University of Groningen.

Wieling, M., Bloem, J., Mignella, K., Timmermeister, M. and J. Nerbonne (forthcoming). Measuring foreign accent strength in English. Validating Levenshtein distance as a measure. *Language Dynamics and Change*.

Wieling, M., Heeringa, W. and J. Nerbonne (2007). An Aggregate Analysis of Pronunciation in the Goeman-Taeldeman-van Reenen-Project Data. *Taal en Tongval, 59(1)*, 84-116

Wieling, M., Margaretha, E., and J. Nerbonne (2012). Inducing a measure of phonetic similarity from dialect variation. *Journal of Phonetics, 40(2)*, 307-314.

Wieling, M., Nerbonne, J. and R.H. Baayen (2011). Quantitative Social Dialectology: Explaining Linguistic Variation Geographically and Socially. *PLOS ONE, 6(9)*, e23613. doi:10.1371/journal.pone.0023613.

Wieling, M., Nerbonne, J., Bloem, J., Gooskens, C., Heeringa, W. and R. H. Baayen (2014). A cognitively grounded measure of pronunciation distance. *PLOS ONE, 9(1)*, e75734.

Wieling, M., Prokić, J., and J. Nerbonne (2009). Evaluating the pairwise alignment of pronunciations. In: Borin, L. and Lendvai, P. (eds.) *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, pp. 26-34.

Yeung, N., Botvinick, M. M. and J.D. Cohen (2004). The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychological Review, 111*, 931-959.

[1] Of course, other computational pronunciation distance measures could be used, such as the one proposed by Wieling et al. (2014). Given the high correlation of this measure with the Levenshtein distance ($r = 0.89$), however, it is not likely the results would be very different.