Adversarial single word processing in advanced bilinguals

Raoul Buurke (S2507501)
University of Groningen

Abstract

Bilinguals are known to portray asymmetric processing in their first (L1) and second language (L2). This effect was investigated for basic lexical access processes, i.e. single word retrieval. A picture naming task was designed to measure retrieval rate in terms of reaction times. This was accompanied with eye-tracking in order to index pupil dilation and consequently cognitive effort. The level of inhibitory control of participants was estimated by using a traditional Stroop task, which was expected to interact with the effect of language. The results indicate that there is no statistically reliable effect of language in processing for these highly proficient bilinguals. There is moreover no significant relation between the language effect and the measure of inhibitory control used. Alternative approaches and explanations are discussed.

Keywords: bilingualism; language processing; cognitive effort; pupil dilation; picture naming task

Background

Bilinguals are known to be quicker when performing language tasks in their first language (L1) than when they do so in their second language (L2). This is a reliable finding that has been found for numerous parts of the language system. In fact, whole introductory linguistics books are dedicated to exploiting the relevant differences, e.g. Lightbown & Spada (2013). The current study aims to replicate this effect specifically for the process of retrieving single words from the mental lexicon. There is evidence that switching between languages is costly at the moment itself (see e.g. Meuter & Allport, 1999). Armed with this information it becomes equally interesting to check whether such language effects exist in general lexical retrieval, i.e. whether there is a consistent difference in processing speed even outside so-called "switch trials".

In order to test this a picture naming paradigm is used, which is described in more detail in the *Methods* section. Results from earlier studies applying this methodology are somewhat surprising in this regard, as it is reported that naming in an L1 is *slower* than in an L2 (Ivanova & Costa, 2008). This is not intuitive and moreover contrasts with findings in different language tasks, such as translation, which shows that translating to an L2 is actually more difficult than doing so to an L1 (e.g. Kroll & Stewart, 1994). One could on common sense argue that any type of L2 processing should always be less efficient, so why these differences exist is not straightforwardly clear. This investigation will add to this puzzle of bilingual processing asymmetries by means of exploring the results of a simple lexical retrieval task, i.e. single word picture naming.

The traditional manner of investigating a bilingual effect of word processing is by looking at latencies. In the case of this study that amounts to comparing the reaction times of the single word picture naming task for each participant across languages. Recently, however, new methodology has been developed to support the methodological status quo. Assuming that greater latencies indicate a higher cognitive effort, it should be informative to also look at general measures of cognitive effort.

Pupil dilation is known to correlate with this concept (Kahneman & Beatty, 1966; Kahneman, 1973), even to the degree that high-informational efforts can be predicted from pupil dilation patterns (Kang et al., 2014). To be more specific: pupil dilation in animals, e.g. mice and primates, has been shown to correlate with physical and mental effort (Reimer et al., 2016; Joshi et al., 2016), although it should be noted that these do not necessarily equate to higher cognitive processes (Larsen & Waters, 2018). This is somewhat problematic, as pupils tend to dilate under a variety of circumstances related to activation of certain neurotransmitters through the locus coeruleus (notably norepinephrine in the sympathetic nervous system and acetylcholine in the parasympathetic system). Nonetheless, it becomes highly attractive to measure pupil dilation if one assumes an (in the very least) indirect link between cognitive effort and naming latencies. Pupil dilation is relatively cheap to incorporate in the methodology, because most eye-trackers already keep track of this data. The effectiveness of this joint methodology has been shown by Papesh & Goldinger (2012), who found that retrieval of low frequency words in a naming task variant resulted in greater latencies than for high fre-

quency words. Crucially, they also found that pupils dilated more when low frequency words were processed. The benefit of this methodology for the current study is that any findings are essentially more reliable, as the same concept or similar ones are measured in two manners, and they are expected to covary.

There are, however, reasons based on the literature to suspect possible effects that interact with the presumed cognitive effort required for bilingual processing. Previous studies have tried to delineate whether bilinguals actually store their words in a unified lexicon or two separate ones. Evidence has been accumulated for both points of view, but it should be clear that if one assumes a unified lexicon (as is done for the purposes of this paper) there is reason to also suspect an interaction of the speaker's level of inhibitory control. Kroll et al. (2008) extensively defend this position. The general idea is that an L1 must be actively inhibited in language tasks such as picture naming, as both languages in the unified lexicon are active at all times. Should this prove to hold under empirical evidence, there is reason to suspect that the unified lexicon approach is valid, although settling this debate is not the purpose of this paper.

In order to index their level of inhibitory control participants of this study also carried out a classical paradigm for measuring this mental skill: a Stroop task[1]. This task has been revised to different ends many times, although the original version is logically accredited to Stroop (1935). Participants are shown color names (e.g. *blue* or *red*), and each of these are also printed in a specific color. Congruent trials are when the print color matches the color name, while the there is a mismatch in incongruent trials. The participant simply says the color name aloud, but incongruent trials are known to produce greater latencies as participants are required to inhibit the interference from the print color. The offset between congruent and incongruent trials is known as the Stroop effect and serves as an indication of inhibitory control.

All in all there is a clear interplay of several concepts in this study. Bilinguals are known to portray an effect of picture naming latency between their L1 and L2, which is expected to covary with measures for cognitive effort (i.e. pupil dilation). At the same time, the aforementioned effects are presumed to interact with the level of inhibitory control of participants. The accompanying hypotheses are as follows. First and foremost, an effect of language is expected to occur in the sense that L2 processing is expected to be slower and more demanding. Greater latencies and pupil dilation is therefore expected for this experimental condition. The second hypothesis is that how capable a speaker is at handling different streams of information influences the effect of language, i.e. that participants with a smaller Stroop effect also portray a lesser effect of language.

[1] It should be stressed that the exact nature of the cognitive effect the Stroop task measures is disputed, although inhibitory control is considered to be a reasonable candidate. See MacLeod (1991) for an overview of the studies concerning this methodology.

Methods

Participants

The participants for this study were 8 students of a linguistics research master of the University of Groningen, who participated in the same psycholinguistics course at the time. The data for one participant (Participant 5, not reported in tables and figures below) was left out as the test results were unreliable. The information of the 7 remaining participants is summarized in Table 1. The mean age was about 24 (ranging from 21 to 29) and the male/female ratio 4 to 3. All participants were advanced learners of English as their L2 and used the language actively on a daily basis in academic settings.

Self reported L1/L2 use ratios differ across the sample. 3 participants use their L1 more than their L2, but note that these were mostly the Dutch participants who lived in the Netherlands, so this is not surprising. One Dutch participant reported to use both languages to an equal extent, which is somewhat plausible given the daily international academic setting. The Brazilian and German participant reported more use of their L2 in their daily lives. Interestingly, the Chinese participant used its L1 more than the L2 in daily life. Finally, the Italian participant reported equal use of the languages. These data are not taken into account as an interacting effect, because (1) the investigated pattern is already quite complex, and (2) the data are self reported and therefore not necessarily reliable. They are nonetheless retrospectively interesting if unexpected patterns arise, so they are still reported.

Procedure

The participants recorded themselves in pairs, i.e. each participant performed the experiment both as an experimenter and as a participant. The recordings took place in the EyeLab of the University of Groningen by means of an EyeLink 1000 (plus remote tracking system) built by SR Research Ltd. The stimuli were presented using the corresponding proprietary EyeLink Experiment Builder software, which minimized latencies between stimulus presentation and recording the pupil dilation, because both tasks were carried out by the same program. Sound recordings were moreover generated by the same software for each trial separately by means of a headset. There

Table 1
Summary of participant information

Participant	Male/female	Age	L1	Level of English (self reported)	Greater use of L1 or L2 in daily life
1	Male	29	Italian	C1	Equal
2	Female	23	Dutch	C2	Equal
3	Female	21	Dutch	C1	L1
4	Female	22	German	C2	L2
6	Male	24	Dutch	C2	L1
7	Male	25	Brazilian Portuguese	C2	L2
8	Male	25	(Mandarin) Chinese	C1	L1

were 100 trials in total, each with a corresponding pupil dilation recording and sound file. Each session started with a few practice trials to familiarize the participant with the task. After approximately every 25 trials the experiment was shortly paused and the participant was instructed to respond to the next 25 trials in the other language. For example: if participant the first 25 named pictures in Dutch the next 25 were responded to in English, and so on. In total there were therefore (approximately) 50 trials per language.

The participants performed a basic picture naming task as designed by Buttler (2018) for her MA thesis. Each trial started with a 500 ms fixation cross to focus the eyes on a central point. It was followed by a 2500 ms stimulus image presentation. The stimuli were chosen from the MultiPic database (Duñabeitia et al., 2018) and consisted of simple monochrome drawings that (ideally) indicated a single concept. Each stimulus was again followed by a 1000 ms blank screen before a new trial started. Participants were instructed to verbally respond within approximately 4 seconds and not to linger on the previous image if they could not come up with the word.

For the Stroop task the participants were asked to run an online version on their own laptop twice. This was done outside of the experimental set-up mentioned above. The implementation (see https://www.psytoolkit.org/experiment-library/stroop.html) was compiled on an online server that was made accessible to the participants only. For this variant of the task 4 colors were used, i.e. *red*, *blue*, *green*, and *yellow*. The task was to name the the print color as opposed to the color name on the screen and to press the corresponding button on the keyboard ("r" for *red*, "b" for *blue*, and so on). A trial started with a 300 ms fixation cross, which was followed by the stimulus presented for an indefinite amount of time. After pressing the key a feedback screen shown for 500 ms, which showed whether the response was "correct" or "incorrect".

There was no training session, but in total there were 40 trials per attempt, so any residual accommodation effect is presumably averaged out. The participants were moreover asked to run the experiment with at least 30 minutes in between in order to minimize a learning effect. This is possibly insufficient, as participants still benefited from being familiar with the task, but at the same time the task was already familiar from their studies. This again minimizes the learning effect, at least between the subjects.

Analysis

Reaction times were computed on the basis of the sound recordings, which was done manually by the participants themselves. The files were transcribed in Praat (Boersma, 2019) with metadata concerning the language, produced word, and whether it was a correct or incorrect trial. Clicking sounds (as an effect of opening the mouth) and any further disfluencies were ignored, and the onset in ms of the verbal response was taken as the reaction time for each trial. Incorrect or disfluent trials were omitted from analyses. A grand average reaction time per language was then calculated and used for analysis.

The raw pupil dilation data was processed in several manners. First off, the data were centered around the pupil dilation baseline per person (which was provided by the software). The resulting data were therefore always a negative or positive offset from that baseline, which itself was coded as 0. Secondly, missing data as an effect of eye blinks was automatically interpolated. These were relatively easy to deal with as (1) the programs automatically detects blinks, and (2) the data is simply null in the recording. After correcting for blinks, the data were downsampled to a sampling rate of 50 ms and aligned to the stimulus onset (i.e. the onset is 0). The time window for each trial was then set to 100 ms before and 3000 ms after stimulus onset. For each of these time windows the grand average per language was then calculated by averaging across all relevant trials, which was used for analysis.

For the Stroop task the data for analysis was straightforward. After each laptop session the application simply returned the mean reaction time for congruent and incongruent trials, and the Stroop ef-

fect was calculated by subtracting the value for incongruent trials from the one for congruent trials. The mean value of the two sessions was used as the Stroop effect value in the analysis.

Results

Pupil dilation

Table 2
Summary of pupil dilation (PD) data

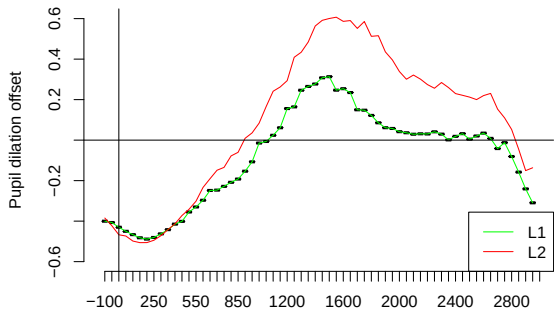
Participant	PDL1	PDL2
1	-0.065	0.105
2	-0.096	0.099
3	-0.125	0.127
4	-0.052	0.061
6	0.060	-0.054
7	-0.074	0.073
8	-0.215	0.228

The pupil dilation data is summarized in Figures 1 and 2. There is an ostensible difference between the pupil dilation curves for the L1 and L2 when averaging across participants (see Figure 1a). In order to estimate significance for this difference a paired Welch Two Sample t-test was computed between the two language conditions. Note that only the grand averages of pupil dilation offsets across participants are used (which are reported in Table 2). The mean pupil dilation difference between languages was statistically significant (t=-3.568, df= 11.975, p= 0.003).

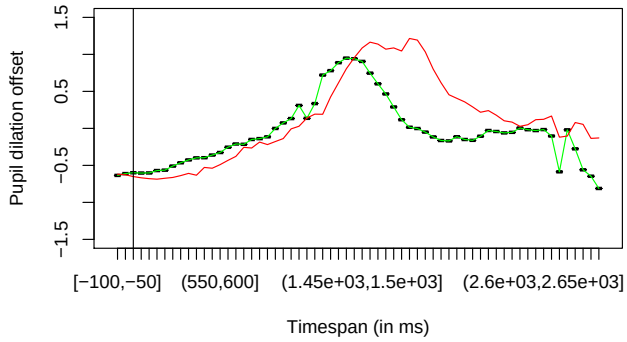
Visually inspecting the graphics for how much the red line exceeds the green line is a useful estimation of how strong the language effect is per participant. When the rest of the graphs are taken into scope, it seems there is some individual variation. Most participants show relatively uniform increases in pupil dilation after stimulus onset, but only Participant 1 shows a clear "peak". There is no clear difference between these two pupillatory patterns in persistence of the language effect, but there is one participant with clear deviation: Participant 6 (Figure 2b). For this participant the red line is in fact mostly below the green one. Another interesting case is Participant 4, who even after averaging shows almost no systematic differences between the L1 and L2 patterns.

Reaction times

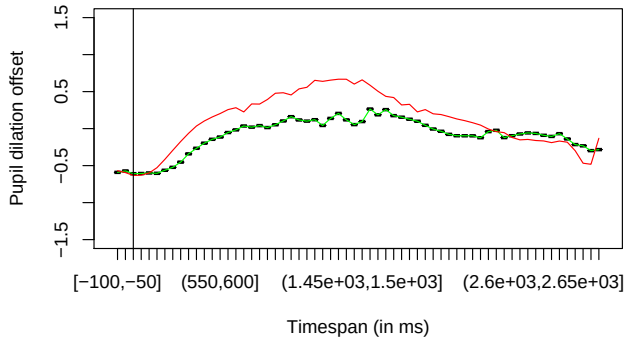
The reaction time data is summarized in Table 3. For most participant the reaction times were lower for the L1 than for the L2. The only exceptions are Participant 2 and 7. The number of excluded trials did not exceed 13% in all cases but one: Participant 8. In this case almost a quarter of the data was left out. The reliability of this case is consequently lower, and an analysis of the error trials therefore becomes relevant. This is not



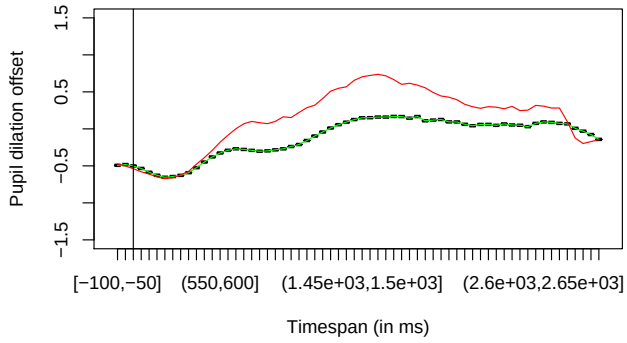
(a) *Grand average*



(b) *Participant 1*

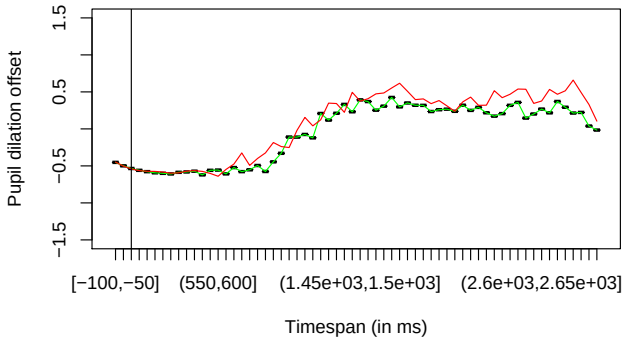


(c) *Participant 2*

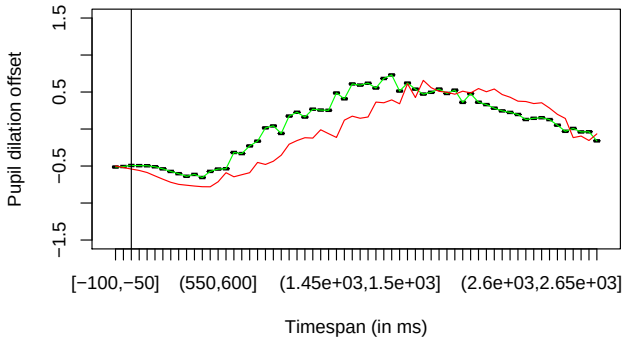


(d) *Participant 3*

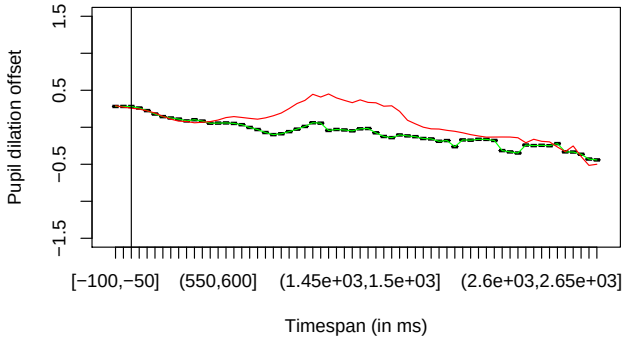
Figure 1
Stimulus-aligned pupil dilation curves per language and per subject. Green = L1. Red = L2. The vertical line indicates stimulus onset.



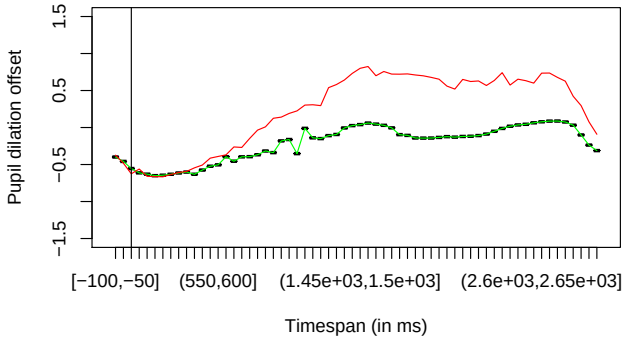
(a) *Participant 4*



(b) *Participant 6*



(c) *Participant 7*



(d) *Participant 8*

Figure 2
Continuation of Figure 1.

Table 3
Summary of reaction times (RT) and trial correctness

Part.	RT_{L1}	RT_{L2}	RT Offset (L2-L1)	# of correct trials (max = 100)
1	1609	1731	122	87
2	1562	1405	-157	88
3	1201	1329	128	88
4	1523	1617	94	89
6	1401	1538	137	89
7	1688	1650	-38	89
8	1526	1634	108	77

carried out given time and space requirements. For the purposes of this analysis the data is used as is. To test for significance a Paired T-test is computed, which yielded a non-significant difference of grand average RT between languages (t= -1.337, df= 6, p= 0.23). Note that this contrasts with the finding for the PD data, for which the analysis of grand averages yielded a significant result between languages.

Stroop task

Table 4
Pupil dilation (PD) offset and Stroop effect data

Part.	PD Offset	Stroop effect
1	0.170	239.5
2	0.195	62.5
3	0.252	19.0
4	0.113	104.5
6	-0.114	112.7
7	0.147	256.0
8	0.443	-19.0

Recall that the size of the language effect tested above is expected to be influenced by how well a speaker is able to handle two languages at the same time. In order to quantify the size of the language effect the offset between the PD offset values are computed by subtracting PD_{L2} from PD_{L1} (see Table 2) and are reported next to the respective Stroop effects in Table 4. The Pearson product-moment correlation between these values is strong, but non-significant (ρ= -0.458, df= 5, p= 0.302). Doing the same for the RT data produces similar results, but with a lower correlation (ρ= -0.119, df= 5, p= 0.799).

Discussion

This study looked into whether the used language (either L1 or L2) in bilingual speech production is a determinant of language processing speed. In order to approximate this bilinguals participated in a picture naming task while their pupil dilations were monitored. Next to this, they performed a classic Stroop task in order to obtain an indirect heuristic for their

ability to process two languages at the same time. The results are somewhat mixed and suffer in statistical terms from too few data. They are discussed below.

The used indices of a language effect in single word processing are the pupil dilation and reaction time patterns. There was a significant difference between the grand averages of the PD analysis, but not for the RT analysis. This is interesting, because it clashes with earlier findings by Papesh & Goldinger (2012). They found in a relatively similar setting that these patterns actually covary. One way to interpret the current results is to claim that these methodologies do not capture the same phenomena. This is a reasonable explanation given the current state of the field, because the relation between language production and reaction times has been confirmed much more often and more deeply. This relationship spurred whole lines of models, e.g. Levelt (1993) and successors. The relationship between pupil dilation and speech production is not well funded in the literature as of yet. Moreover, its relation to language processing has been questioned by some. Larsen & Waters (2018), for example, notes that there is often neural co-activation with pupil dilation with other processes. In a study such as this, which controls for relatively few factors, pupil dilation is therefore not a highly trustworthy measure. Next to this, the PD data is analyzed in a manner that is not informative, as only the grand averages per person were used. Information about the stimuli themselves is therefore lost, as well as the more specific patterns within each word. In the spirit of conservative conclusions, then, it should be concluded that the reaction time data is better suited to answering the hypothesis here. As a consequence, the best conclusion is that there is no significant difference in processing speed across languages for highly proficient bilinguals.

Despite having discarded the reliability of the PD data, it is interesting to look at which specific cases were unexpected across both methodologies. Participants 4 and 6 portrayed PD patterns that were unexpected given the general trend. For Participant 4 the PD data was the least informative, as the curves were noticeably jagged. It turned out after analysis that this participant drank coffee shortly before partaking in the experiment, which is problematic as caffeine is known to influence pupil dilation (Abokyi et al., 2017), especially within 60 minutes of ingesting it. This readily explains the deviant pattern.

Participant 6's pattern is still unexpected, although there is still a reliable difference indicating an effect, but simply inverted. It is possible that this is a case of language attrition, in which processing of the L2 actually becomes faster than that for the L1. Especially the lexicon is sensitive to attrition, which is extensively discussed by Schmid & Köpke (2009). There is no direct evidence of such a case in the data that is collected here, such as a negative RT offset (which would indicate faster processing in the L2). The RT data instead show that processing in the L1 is faster than in the L2. Moreover, the participant is Dutch and lives in the Netherlands, which would be a highly atypical context for language attrition. None of these considerations firmly rule out L1 attrition, but confirming it requires a set of diagnostics that is simply outside the scope of this article. Future studies can, however, certainly attempt to establish a link between PD patterns and L1 (lexical) attrition.

Recall that a masking effect of inhibitory control was expected. The Stroop task data was paired with an ad hoc measure of the language effect (i.e. offsets between PD and RT values for L1 and L2). The correlations for both were negative, although clearly stronger for the PD data than for the RT data, and they were moreover not statistically significant. Leaving aside for the moment the small and unreliable sample size, it seems that in both cases a greater Stroop effect is related with smaller offset values. This is unexpected, because the opposite is expected: a greater Stroop value indicates less skill at handling two language streams at the same time, so under our hypothesis it is expected to be matched with greater offset values (which again indicates a greater difference in language processing between the languages). It again seems that in the current setting this simple task does not suffice as a measure of the cognitive mechanism under inspection. Munakata et al. (2011) argue on the basis of neurological patterns that inhibitory control should in fact be divided into directed global inhibition and competitive inhibition, which is a distinction not made here. Yet others (Aron, 2007) dispute the concept of inhibitory control altogether. Moreover, the type of inhibitory control approximated by the standard Stroop task may not be that indicative or useful for the bilingual case. An alternative approach is, for example, to use bilingual Stroop task variant in which the color must be named variably in the L1 or L2. This more readily approximates the ability to handle two languages at once.

One final point of improvement for further investigations is the statistical analysis itself. The approach here was basic and intuitive, but did definitely not take advantage of all the data that was available. Thousands of data points were (1) downsampled and (2) iteratively averaged, which makes for straightforward analysis but extreme loss of information. Alternative methods include using Generalized Additive Modeling to fully incorporate the non-linear curves of the data, which also allows further investigation by e.g. the word types. This is an attractive method, al-

though it requires both substantial computing power and statistical knowledge. Alternatively, one can resort to bootstrapping approaches. Bootstrapping is a highly general technique that relies on resampling the data and generally only requires that the data is not autocorrelated or too few (see e.g. Chernick & LaBudde, 2014 for a succinct overview of these type of approaches). Resampling is especially useful for this case, as there are not many participants, but definitely many data per participant. By bootstrapping the pupil dilation data it is possible to do reliable statistical inference and apply a more language profile-based approach to bilingualism, which is justifiable given the many factors that are known to influence bilingual performance.

References

Abokyi, S., Owusu-Mensah, J., & Osei, K. (2017). Caffeine intake is associated with pupil dilation and enhanced accommodation. *Eye*, *31*(4), 615.

Aron, A. (2007). The neural basis of inhibition in cognitive control. *The neuroscientist*, *13*(3), 214–228.

Boersma, P. (2019). Praat: doing phonetics by computer. *http://www. praat.org/*.

Buttler, R. (2018). Cognitive effort in speech production: insights from pupillometry during fluent and stuttered speech. Master's thesis, Rijksuniversiteit Groningen, Groningen, the Netherlands.

Chernick, M. & LaBudde, R. (2014). *An introduction to bootstrap methods with applications to R*. John Wiley & Sons.

Duñabeitia, J., Crepaldi, D., Meyer, A., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2018). Multipic: A standardized set of 750 drawings with norms for six european languages. *The Quarterly Journal of Experimental Psychology*, *71*(4).

Ivanova, I. & Costa, A. (2008). Does bilingualism hamper lexical access in speech production? *Acta psychologica*, *127*(2), 277–288.

Joshi, S., Li, Y., Kalwani, R., & Gold, J. (2016). Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex. *Neuron*, *89*(1), 221–234.

Kahneman, D. (1973). *Attention and effort*, volume 1063. Citeseer.

Kahneman, D. & Beatty, J. (1966). Pupil diameter and load on memory. *Science*, *154*(3756), 1583–1585.

Kang, O., Huffer, K., & Wheatley, T. (2014). Pupil dilation dynamics track attention to high-level information. *PLoS One*, *9*(8), e102463.

Kroll, J., Bobb, S., Misra, M., & Guo, T. (2008). Language selection in bilingual speech: Evidence for inhibitory processes. *Acta psychologica*, *128*(3), 416–430.

Kroll, J. & 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.

Larsen, R. & Waters, J. (2018). Neuromodulatory correlates of pupil dilation. *Frontiers in neural circuits*, *12*, 21.

Levelt, W. (1993). *Speaking: From intention to articulation*, volume 1. MIT press.

Lightbown, P. & Spada, N. (2013). *How languages are learned*. Oxford university press.

MacLeod, C. (1991). Half a century of research on the stroop effect: an integrative review. *Psychological bulletin*, *109*(2), 163.

Meuter, R. & Allport, A. (1999). Bilingual language switching in naming: Asymmetrical costs of language selection. *Journal of memory and language*, *40*(1), 25–40.

Munakata, Y., Herd, S., Chatham, C., Depue, B., Banich, M., & O'Reilly, R. (2011). A unified framework for inhibitory control. *Trends in cognitive sciences*, *15*(10), 453–459.

Papesh, M. & Goldinger, S. (2012). Pupil-blah-metry: Cognitive effort in speech planning reflected by pupil dilation. *Attention, Perception, & Psychophysics*, *74*(4), 754–765.

Reimer, J., McGinley, M., Liu, Y., Rodenkirch, C., Wang, Q., McCormick, D., & Tolias, A. (2016). Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. *Nature communications*, *7*, 13289.

Schmid, M. & Köpke, B. (2009). L1 attrition and the mental lexicon. In A. Pavlenko (Ed.), *The bilingual mental lexicon: Interdisciplinary approaches* (pp. 209–238). Multilingual Matters Clevedon.

Stroop, J. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology*, *18*(6), 643.