Processing of garden-path sentences by high-proficiency bilinguals: can access to supporting dual linguistic systems help prevent lingering misinterpretations?

Ligtenberg, Kars

Processing of garden-path sentences by high-proficiency bilinguals: can access to supporting dual linguistic systems help prevent lingering misinterpretations?

Kars Ligtenberg

Supervisors: Dr Leticia Pablos-Robles & Dr Susana Valdez

Abstract

Eye-tracking reading on bilinguals has found divergent results: some have found disadvantages for bilinguals of varying proficiency as compared to monolinguals, where language proficiency scores and individual differences in cognitive control ability accounted for these differences. Others reported a bilingual advantage in cognitive control which also affected syntactic parsing beneficially as bilinguals scored higher on comprehension whilst processing garden-path sentences. However, bilingualism itself is often poorly defined, which can lead to unfair comparisons between, potentially, extremely different types of bilinguals. Therefore, we employ a strict definition of bilingualism, as well as clearly defining what language-pairing our bilinguals have and what the potential language interaction effects of the pairing could be. In this fashion, the current study assesses whether high-proficiency Dutch-English bilinguals show an advantage on sentence comprehension of garden-path sentences and whether lingering misinterpretations related to garden-path effects in Good-Enough parsing theories remain and follow the expected patterns. We employed eye-tracking with N = 20 Dutch-English bilinguals and N = 12 native English speakers, and compared their reading times and comprehension accuracy. Our results confirm the patterns suggested in recent adaptations made to Good-Enough parsing models, in which information structure and prediction are incorporated and help guide the parsing process. Additionally, we find evidence of a specific Dutch-English language interaction which surfaces as an advantage for the bilinguals in specific eye-tracking measures and sentence parts, but no further (dis)advantage between our bilingual and native English speaker group, neither in sentence comprehension nor cognitive control, was found.

Keywords: Bilingualism, Garden-Path, Eye-tracking, Good-Enough parsing, Psycholinguistics

Contents

1. Introduction

Processing of garden-path sentences is a contentious topic within psycholinguistics: there are different theories to account for how these sentences are processed and why, in many cases, misinterpretations of garden-paths due to incorrect parsing linger. The current most prominent ones are aptly summarised in Traxler (2014), including the good-enough parsing theory (Ferreira et al. (2002), Ferreira and Patson (2007) and Ferreira and Lowder (2016)), which this study employs to explain syntactic parsing. Furthermore, there have been both purported advantages and disadvantages in the processing of these sentences related to bilingualism. Usually, advantages are ascribed to a bilingual advantage in domain-general cognitive control skills, and specifically, increased inhibitory functioning (Teubner-Rhodes et al. (2016)). On the other hand, when disadvantages are observed, these are often linked to differences in proficiency in the L2 (Brothers et al. (2021)). However, the term 'bilingualism' is often poorly defined, which can lead to unfair comparison between two potentially widely varying groups of 'bilinguals'. Additionally, the existence of a bilingual advantage in itself is not widely accepted in the scientific community (see, for example, Paap et al. (2015)), nor is there uniformity in the theoretical explanations for the purported advantage. In an attempt to bridge this gap, the current study examines bilingual processing of garden-path sentences with a strict definition of bilingualism, as well as a clearly-defined theoretical framework for the expected effects thereof.

1.1. Theories on Syntactic Parsing

Within psycholinguistics, syntactic parsing is considered to be a mental process that establishes the grammatical structure and dependencies of words when interpreting a sentence in real time. Words alone are not always enough to come to a correct interpretation of a sentence: if you have merely two nouns and a verb without any syntactic information, you would not know which thematic roles to assign to each of the nouns. Syntactic parsing solves this by providing a potential analysis mechanism of the structure of a sentence by assigning grammatical structure to linguistic input. Development of theories on syntactic parsing as a mechanism largely started in the 1980's with a seminal review by Frazier (1987). This work outlines the garden-path model of sentence processing, which assumes a gradual construction of sentence structure through incorporating words into a constituent structure according to their syntactic category, as they are encountered. The garden-path model operated on two main principles: minimal attachment and late closure.The most relevant to the current study being that of late closure, which postulates that new items in a clause will preferably be attached to the existing constituent structure, instead of creating a new structure (example provided in Figure 1). In Figure 1, the principle of late closure would predict that the reader of C would have to retrace their steps in the analysis upon encountering the VP 'was correct'. Initially, 'the answer' would have been ascribed the thematic role of the (optional) object complement of the VP 'knew', as that is part of the existing structure (main clause). However, upon encountering the novel VP 'was correct', which requires a subject complement, the sentence would be ungrammatical if the NP 'the answer' is maintained as the optional object of the VP 'knew'. This entails that a new structure (embedded clause) must be built. This resembles a situation in which a pedestrian has taken a wrong turn on a garden-path and encountered a dead end, forcing them to return to an earlier node to reach the desired end. Hence the origin of the name of both the model and the effect observed in behavioural data.
Since then, syntactic parsing theories have advanced significantly, resulting in several competing models being available today. A thorough overview of the currently circulating theories is provided

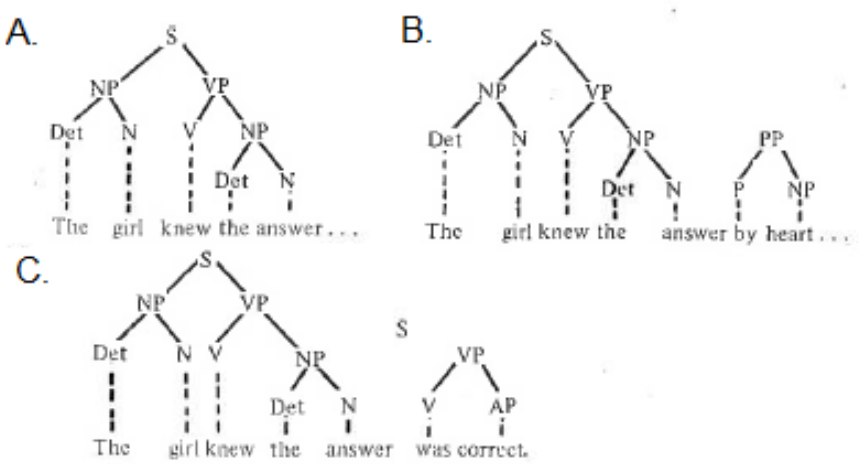

Figure 1. An example sentence, in which the transition between A and B is expected to be easy, whereas transition between A and C is expected to be problematic, adapted from the review by Frazier (1987), p. 562.

by Traxler (2014). It covers prediction-based models, which assume that readers anticipate future words based on previous context. This would explain how readers are capable of integrating new words so rapidly in their semantic and syntactic framework. There is strong evidence towards this anticipation from eye-tracking visual word paradigms, in which several objects are presented on-screen. Participants are then provided an auditory stimulus that will induce anticipatory glances (an example of such a visual word paradigm, from Hsu et al. (2021), is provided in a later section). However, how these predictions come to be has not yet fully been elucidated, and therefore theories employing anticipation are not fully capable of explaining syntactic parsing processes yet.

The model that this study employs, good-enough parsing, is largely compatible with prediction models. Good-enough parsing, as first proposed in Ferreira et al. (2002), assumes that readers do re-analyse sentences when encountering a temporary ambiguity, but do not always update their interpretation to match the licensed syntactic structure, thus leaving lingering misinterpretations of the input. This theory came to be due to a plethora of evidence from eye-tracking studies that examined the phenomenon of lingering misinterpretations: when encountered with a temporarily ambiguous sentence, or a semantically unlikely sentence, interpreters would systematically come to a syntactically wrong (but (more) likely) interpretation. Evidence can be found in, for example, Slattery et al. (2013), Christianson et al. (2017) with monolingual English speakers and in Brothers et al. (2021) with bilingual Chinese-English speakers, showing that this phenomenon is quite robust. This behaviour can be exemplified with the sentence 'While Jane dressed (,) the baby played in the crib', which is temporarily ambiguous without the optional comma in the middle. When asked 'Did Jane dress the baby?', readers would often answer 'Yes', whilst this is not the licensed interpretation of the sentence. Further evidence comes from misinterpretation of semantically unlikely passive sentences such as 'The mouse was eaten by the cheese' which would often be conflated into 'The mouse ate the cheese', which is a far likelier interpretation (Traxler (2014), pp. 608-609).

Furthermore, Ferreira and Lowder (2016) have recently enhanced their original good-enough model in an attempt to integrate prediction as well as information structure into it. Previous studies

examining lingering misinterpretations in garden-path sentences found that the effects were often localised: answers regarding the temporarily ambiguous first part of the sentence were often incorrect, whereas those concerning the disambiguating second part of the sentence were almost universally correct. To illustrate, Ferreira and Lowder (2016) provides the example of '*While Anne bathed the baby played in the crib*', in which subjects would then often answer incorrectly to the question 'Did Anne bathe the baby?' with 'Yes', whereas those asked about the baby playing in the crib would almost always answer correctly. This is consequently integrated with information structure, which poses that given information is usually located at the start of a sentence, by proposing that the good-enough processing only takes place for given information (i.e., the embedded clause). Prediction then comes into the picture in the allocation of processing resources towards the new information (i.e., the main clause) in order to integrate it. Thus the framework manages to explain why misinterpretations are observed in garden-path sentences, but only with regards to the initial clause. In sum, as many previous eye-tracking studies on garden-path sentences have found evidence for the good-enough parsing model, and it appears to be capable of explaining the behavioural effects observed, this study employs this theoretical framework as a basis for the syntactic parsing process.

1.2. Cognitive control and individual differences

Sentence processing, especially in the case of (temporarily) ambiguous sentences, is also often linked to cognitive control/executive functioning. For example, Brothers et al. (2021), also controlled for individual differences in cognitive control with a test battery included in their eye-tracking experiment with garden-path sentences. Cognitive control is generally regarded as being the ability to intentionally select thoughts, behaviours and commit neural processing power to best perform the task at hand (see for example Braver (2012) which describes a framework that attempts to explain cognitive control's ability to regulate thoughts and actions based on behavioural goals). This includes not only positive selection, but also negative selection (inhibition) of non-relevant cognitive processes. It therefore seems logical that this could have an impact on garden-path processing, as these involve a difficult parsing procedure as well as an incorrect parse that should be suppressed.

Evidence that cognitive control is likely to play a part in sentence comprehension is provided in Ye and Zhou (2008) with Mandarin Chinese speakers. Their event-related potential (ERP) study examined reanalysis of incompatible sentence representations with a clash between the syntactic analysis and the plausibility heuristic (i.e., semantic analysis). Participants were presented with a sentence word-by-word, after which a probe sentence appeared. Consequently they were asked to judge whether the experimental sentence (presented word-by-word) was semantically consistent with the probe sentence. The sentences were divided across a 2x2 condition with passive and active sentences, to examine effects of syntactic complexity, and plausible and implausible sentences, to test the main hypothesis regarding the plausibility heuristic. Cognitive control ability was assessed by means of a colour-word Stroop task (see MacLeod (1991) for a comprehensive review of the Stroop task in use as a measure for cognitive control), and depending on the performance during the Stroop task, participants were placed in either the 'high' or 'low' control group. Consequently, Ye and Zhou (2008) predicted that if there is an effect of cognitive control on sentence processing, then a P600 effect would be visible in the implausible versus plausible condition, whereas if sentence comprehension is solely guided by syntactic analysis, then no P600 would be visible (as there are no syntactic violations, but only plausibility violations). Their results revealed that cognitive control determined whether or not the expected P600 was actually observed: only the high-control group

showed an effect of syntactic complexity on the ERP responses during the experiment, but only for the implausible condition. The authors therefore concluded that inhibitory processes, which are part of cognitive control, must have been the cause for this ERP component. Consequently this would indicate that cognitive control plays an active role in sentence processing.
Additionally, there have also been studies looking at how cognitive control influences garden-path sentences specifically. Novick et al. (2014) examined training effects of non-syntactic cognitive control tasks, which target memory or attention without employing linguistic input, on the processing of temporarily ambiguous sentences with native English speakers. Their training paradigm included a conflict resolution task (i.e., inhibition of initial misrepresentation of input) in the form of a letter n-back task with lures. Additionally, they employed several working memory tasks targeting attention (running span task), verbal stimuli manipulation (letter-number sequencing task) and visual-spatial working memory (block span task). Altogether these recruited a broad span of executive functioning skills without priming any syntactic processing as they all employed sequences of letters and digits without any sentence-like structure to them. The participants were randomly assigned to either the training group or the control group, which were both assessed twice (3 to 6 weeks apart), but only the training group received 20 one-hour training sessions in between. The training group was further divided into a responders and a non-responders category based on whether they showed improved cognitive control performance over the 20 training sessions in between the assessment sessions. Their results showed significant gains in comprehension accuracy for the responder group, whereas there was no significant effect for either the non-responder or control group. Additionally, the authors collected eye-tracking data to assess the real-time performance during sentence processing. The eye-tracking measures confirmed that the responder group outperformed their pre-test selves as well as the other groups in that they showed significantly shorter regression paths for the critical disambiguating region. The authors therefore conclude that resolution of syntactic ambiguity is closely correlated to general cognitive control abilities, and that training the latter helps improve the first.
Further evidence that domain-general cognitive control can influence sentence processing and comprehension is provided by Hsu et al. (2021) through a visual-world paradigm. In this case, the paradigm employed spoken instructions, but the visual-world paradigm can also employ simple factual statements (e.g., in Altmann and Kamide (1999)). The spoken instructions or sentences are presented in combination with a visual display that has a target picture on it as well as distractors. Consequently, the eye-gaze data obtained can be used to monitor processes involving prediction and conflict resolution dependent on where and when the participants are looking whilst listening to the audio stimulus. Based on previous research (including Novick et al. (2014) discussed above), they argue that cognitive control plays a role in resolving sentential ambiguities, even if those specific cognitive processes are usually not associated with syntactic material. In fact, they even argue for a direct causal relationship between cognitive control and the revision of sentence misinterpretations. In order to test this hypothesis, Hsu et al. (2021) employed a cross-task functional-adaptation paradigm, in which the properties of cognitive engagement are varied (between conflict-based cognitive control and cross-trial attention) in a previous task to examine immediate effects on syntactic revision. This entailed that single flanker task trials, which activate the cognitive control facilities, were presented interleaved with the actual target stimuli. In these target stimuli, participants were presented with a pictures, and given a simple verbal order, which was either ambiguous: "Put the horse on the binder onto the scarf" or unambiguous: "Put the horse that's on the binder onto the scarf" (see Figure 2). Via this method the authors aimed to extricate the exact cognitive control processes that underlie syntactic parsing.

Their results showed the expected effects of incongruent Flanker trials (longer reaction times)

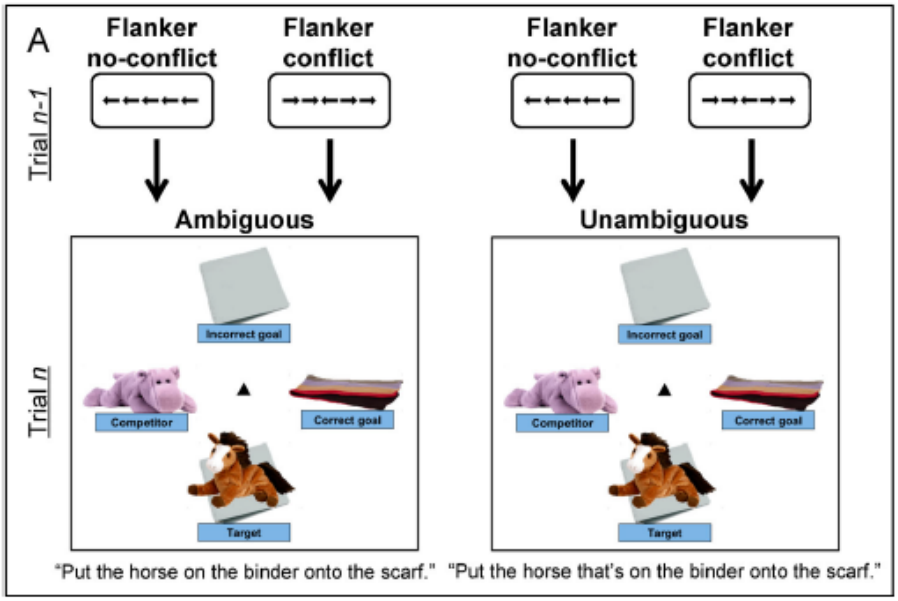

Figure 2. Experimental task for language comprehension, adopted from Hsu et al. (2021), p. 217

and more and longer gazes to the incorrect goal, as well as a lower accuracy during the processing of ambiguous sentences. Relevantly for our purposes, their analysis showed that participants were more correct on ambiguous sentences preceded by an incongruent Flanker trial as compared to the when it was preceded by a congruent Flanker trial, whilst this effect was not present in the unambiguous condition. This indicates an effect of conflict adaptation, which was confirmed by an interaction effect between previous and current trial in the ambiguous condition. This mirrors the findings of Hsu and Novick (2016), which was conducted in a similar fashion but with a Stroop task in between sentence trials. The authors therefore conclude that non-verbal cognitive control does, in fact, directly influence syntactic revision. Moreover, Hsu et al. (2021) ensured that the observation was due to cognitive control and not simply increased attention through follow-up experiments that did not display similar results to those interleaved with Flanker trials when an attention-based task was presented. In conclusion, the authors asserted that activating cognitive control prior to confrontation with a temporarily ambiguous sentence directly affected participants' processing as measured through comprehension of the instructions (as measured through recording of the click-and-drag task administered after listening to the instructions) as well as reading times positively (i.e., the former increased whilst the latter decreased).

This section has shown that cognitive control can play a major role in the processing of sentences, even if the cognitive control mechanism itself is not necessarily syntactic in nature. Priming (conflict-resolution related) cognitive control mechanisms has been shown to have a direct effect on the processing of ambiguous sentences (Novick et al. (2014), Hsu and Novick (2016) and Hsu et al. (2021)). Additionally, general cognitive control ability seemed to be able to predict performance in sentence comprehension tasks (Ye and Zhou (2008)). Finally, all the above evidence was obtained

from monolingual speakers, which shows that even in this population, the processing of complex sentences takes cognitive effort. It is therefore imperative to take cognitive control into account when working within sentence processing paradigms, as it could explain individual differences. For this reason, some cognitive control tasks that specifically target conflict resolution and inhibition are also included in the pre-test for this study, as outlined in section 2.4.

1.3. Bilinguals and Garden-Path processing

The models and studies on garden-path processing discussed above were mainly based on observations made with monolingual participants. When two language models are involved in processing, there are additional factors to be considered and potential issues that arise. Not only can different language pairs affect one another differently, for example with typologically closely- (e.g., Dutch-English) or distantly-/unrelated languages (e.g., Chinese-English). Additionally, bilingualism can develop from birth, when a child has two parents that speak different languages, or later in life due to migration or other factors. Bilinguals can be balanced, with both languages being similarly used and developed, or unbalanced, where one of the languages is far more well-developed and/or dominant than the other. Naturally, all these factors can greatly influence sentence processing in one of the languages involved, especially in complex sentences like garden-paths. Furthermore, we might expect to see specific language-interaction effects when, for example, the grammatical structures of the languages spoken differ, leading to conflicting interpretations based on which language the parsing has to be based on. Finally, there has been the long-standing debate of a domain-general bilingual advantage in cognitive control. If this is the case, then we might also expect bilinguals to be better at inhibiting the incorrect representations, as the previous section has shown an effect of cognitive control on syntactic processing. Indeed, recent behavioural and eye-tracking studies have revealed conflicting results with regards to bilingual processing of garden-path sentences: Teubner-Rhodes et al. (2016) found a bilingual advantage in the comprehension of these sentences but not in reading times of these sentences with Spanish-Catalan bilinguals. The authors attributed this bilingual advantage to an advantage in conflict-monitoring, which is part of the cognitive control facilities. Furthermore, bilinguals were also found to have an advantage in a cognitive control N-back task, in which participants have to press a button when the presented item matched the item a certain number (N) of trials back, in both accuracy and reaction times. However, this was only true for the high-conflict condition of the N-back, in the no-conflict N-back bilinguals and monolinguals performed similarly.

Different results were obtained by Jacob and Felser (2016), who paid special attention to reanalysis of garden-path sentences. Their participants were not termed 'bilinguals' by the authors, but rather L2 speakers of English, with their L1 being German; nonetheless, they were all said to be of 'upper intermediate' proficiency in English and considered it to be their primary non-native language (Jacob and Felser (2016), p. 913). The authors conducted a reading eye-tracking experiment with garden-path sentences and their unambiguous counterparts, where participants responded to comprehension questions that appeared after the target stimuli. This was done in order to compare garden-path effects across the language group condition in terms of relative slower reading times in ambiguous conditions (in the expectation that the ambiguity induces processing difficulties). Additionally, the garden-path effect was examined through assessing whether there were lingering misinterpretations in the ambiguous condition by means of comparing the accuracy of answers to the accompanying comprehension questions. Jacob and Felser (2016) employed five eye-tracking measures: first-pass reading time, regression-path, second-pass reading time, total reading time and finally, proportion of outbound regressions (first-pass reading time, regression path and total

reading time are explained further in Section 1.7, as the current study employs these as well). Their results showed that the L2 learners did have longer fixations and slower reading times in all conditions, but the garden-path effect was less compared to that in the L1 speakers of English. Comprehension accuracy for both groups were similar on filler items (89% for the L1 group and 88% for the L2 group), and both groups performed significantly worse on the experimental items, with the L1 group (72%) outperforming the L2 group (54%). The authors propose that this indicates that reanalysis was less likely to be triggered in L2 speakers, as these also showed a lesser re-reading time (i.e., shorter regression path and total reading time) of the disambiguating region as well as significantly fewer regressions as compared to the L1 group. This would also explain why lingering misinterpretations were more prevalent with the L2 group. One important aspect to note is that the authors also suggest future research to employ near-native speakers of the L2 in order to examine if these effects persist; i.e., they postulate that L2 proficiency might play a major part in the behaviour observed.

In a similar vein, Brothers et al. (2021) found universal disadvantages for their Chinese-English bilingual participants in both comprehension and eye-tracking reading measures. However, they did have several groups of bilingual participants with different proficiency levels, and the between-group differences among bilinguals were fully explained through these differences in language proficiency. This further supports the claims by Jacob and Felser (2016), who also mentioned proficiency as potentially being at the heart of the observations made. The bilinguals partaking in Brothers et al. (2021) were late L2 learners of an intermediate proficiency, from a Chinese L1 background. Importantly, the authors also conducted a battery of executive function and working memory tasks, both of which showed no correlation with garden-path effects for either the bilinguals or the monolinguals. Additionally, Brothers et al. (2021) found no significant differences in executive function between the bilinguals and monolinguals, failing to replicate previously found domain-general cognitive control advantages in bilinguals.

1.3.1. The Bilingual Advantage

The idea of an overall bilingual advantage in various fields of cognition was largely spearheaded in research led by Prof Dr Bialystok. For example, Bialystok et al. (2001) compiled a large number of studies conducted with (primarily) young children on the impact bilingualism has on cognitive effects, and found mainly advantages. Later research, by Bialystok et al. (2007) also found protective effects against dementia in older bilinguals, widening the coverage of the purported bilingual advantage. These seminal studies sparked a large debate on the existence of such a bilingual advantage, with many studies finding such a bilingual advantage, mainly in executive functioning; however, there have also been many studies that failed to either find or replicate these results. For example, Duñabeitia et al. (2014) did not find any cognitive advantages in inhibitory control in a large sample of monolingual and bilingual children. Furthermore, the meta-analysis by Paap et al. (2015) concluded that the evidence is not strong enough to support the hypothesis of an overall bilingual advantage in executive functioning. Since then, even Bialystok has somewhat revised their position as co-author in DeLuca et al. (2019), where the purported bilingual advantage is instead argued to be a spectrum of experiences caused by bilingualism which can affect brain structure and function. This would still lead to advantages in certain situations as the brain would adapt itself to be optimally efficient during processing and controlling more than one language. Nonetheless, modern studies such as Teubner-Rhodes et al. (2016) still operate under the assumption that a bilingual advantage exists, and still find evidence in favour of this hypothesis.

1.3.2. Bilingual differences

One problem that has arisen in studies pertaining to the bilingual advantage, which was also addressed in Paap et al. (2015), is the lack of uniformity and coherence among the theoretical explanations for this advantage. This is further elaborated on by de Bruin et al. (2021), who criticise the theoretical frameworks employed by studies investigating the bilingual advantage. Primarily, they focus on the need for a more thorough understanding of bilingual language control before trying to theorise on differences found in behavioural experiments. Furthermore, they stress the need for examining individual differences within bilinguals with pre-registration of types of participants, measures and analyses. These practices are especially important in the aftermath of the replication crisis, which is near-to-heart for behavioural psychology and psycholinguistics.

In an attempt to adhere to the suggestions made by de Bruin et al. (2021) to form a clear hypothesis of the mechanism *behind* the observations made in bilingual behavioural studies, this study employs the viewpoint of bilingual differences rather than advantages in processing. Furthermore, a clear theoretical framework on bilingual language control is established by employing computational linguistic models of bilingual word recognition: namely, the BIA+ model (Dijkstra and Van Heuven (2002)) and the more recent Multilink model (Dijkstra et al. (2019)), which can be seen as a continuation and further developed version of the BIA+ model. These computational models propose parallel activation and processing in both languages in the bilingual brain, which could surface as a seeming advantage in language comprehension through a multilingual analysis if the languages involved support one another in some way. The BIA+ model has already proven to be reflective of human processing through neuro-imaging data by van Heuven and Dijkstra (2010). Furthermore, Baroni (2020) argues for a general wider acceptance of artificial language models as comparative material for human language: modern natural language-processing models are sufficiently sophisticated and productive to allow for new perspectives on natural language itself. Interdisciplinary work between computational linguistics and psycholinguistic should not only be possible, but even be encouraged. Therefore, we operate under an assumption of bilingual differences based on parallel language activation as proposed through these computational model, which can surface through language specific interactions as will be discussed in more detail in Section 1.6.

1.4. Defining bilingualism

So far the bilingual advantage has been thoroughly discussed, as well as how bilinguals may differ in their overall language control and processing strategies. However, the term *bilingual* itself can be considered to be contentious within the field of linguistics. Is anyone that speaks two languages a bilingual? Is there a threshold of a certain competency before one is a bilingual? Is a 'true' bilingual only someone that has been raised in a language environment with two L1 inputs rather than an L1 and an L2? These questions are often answered differently depending on who you ask, which highlights another issue in research with bilinguals: without a clear definition of *who* is considered a bilingual, different studies with bilinguals might be comparing apples with oranges. Recently, these issues have also become apparent within the multilingualism research community, as several authors have published their view on the definition of bilingualism and how to improve upon it.

Kremin and Byers-Heinlein (2021) provide an overview of several currently existing different views of bilingualism: the two main ones being categorical and continuous approaches. These two models are rather self-explanatory; the categorical approach is based on categorisation of participants as either monolingual or bilingual, whereas the continuous approach employs a scale of bilingualism from monolingual to bilingual with varying steps of proficiency in between. Furthermore, they also

exemplify newer models that try to integrate features of both the earlier approaches. The factor mixture model, for example, allows for participants to be placed in a continuum within a category (whilst maintaining separate categories), therefore making both measures available for data analysis. The grade-of-membership model is quite similar to the previous model, but considers the entirety of monolingualism to multilingualism to be a continuum, with a 'grey area' in the centre where categories overlap. Kremin and Byers-Heinlein (2021) consequently argue for a higher level of standardisation and cooperation within research (sub)field(s), in order to facilitate comparison and the establishment of a uniform theoretical framework.
In a similar vein, Marian and Hayakawa (2021) proposed a set of more uniform methods of objectively quantifying bilingualism. They identified similar issues with the variability that exists on the spectrum of bilingualism, and the difficulties that arise in classification due to these. They mention that there are already existing methods of quantifying bilingualism by means of self-reports (i.e., Marian et al. (2007) and Kaushanskaya et al. (2020)) or the use of standardised tests. These would often be used in conjunction, as while self-reports like the LEAP-Q have been proven to effectively reflect language proficiency (see Marian et al. (2007)), the validity of these can display a lot of variation due to, for example, interpretation of scores and scales. This leads Marian and Hayakawa (2021) to conclude that ideally researchers within the field of multilingualism work together to provide an effective Bilingual Quotient (or BQ), similarly to the widely used Intelligence Quotient (IQ). They go on to discuss the benefits of such a BQ, but also the difficulties that currently still exist with developing such a heuristic.
For the reasons detailed above, this study shall clearly define what theoretical framework of bilingualism will be employed. The current study employs the factor mixture model of bilingualism (Kremin and Byers-Heinlein (2021)), allowing for within category variation whilst maintaining separate monolingual and bilingual categories. This is the most suitable model as the participants are specifically high-proficiency bilinguals, which therefore requires a scale or continuum within the bilingual category. Furthermore, we employ an adapted version of the LEAP-Q (see 2.4.1 for the adaptations made and the reasoning behind them) in order to provide more granular insights into the bilingual proficiency of participants via self-report. This allows for controlling individual differences in proficiency, as well as facilitating the identification of other potential confounds like a third language interfering with processing.

1.5. Late onset bilinguals

The ability to attain native-like language abilities in especially late onset bilinguals is often questioned, especially in critical period theories of language acquisition (as first proposed in Snow and Hoefnagel-Höhle (1978)). A recent study that conformed to the hypothesis that age of onset is a critical factor in attaining native-like abilities in the L2 is Bylund et al. (2021). This study examined L2 native-like attainment by comparing bilingualism and age of acquisition (AoA) as their research variables, to determine which effect has a stronger impact on phonetic, grammatical and lexical measures. They employed a large body of participants (N = 80) across four different groups (N = 20 for each): L1 monolinguals, L1 bilinguals (i.e., simultaneous bilinguals), L2 monolinguals (foreign language adoptees) and L2 bilinguals (foreign language immigrants that immigrated at a young age). Their results showed no standalone effects of bilingualism across the different tasks, whereas AoA was significant on six out of seven tasks. In the cases where there were still main effects of bilingualism, these always co-occurred with AoA effects. With interaction effects between bilingualism and AoA, follow-up tests were performed which also confirmed that AoA had a larger impact than bilingualism. Bylund et al. (2021) therefore concluded that age of acquisition is, in

fact, the primary determinant of L2 ultimate attainment.
However, the above study examines AoA effects primarily in light of L2 acquisition, which has some potential flaws. For example, Mayberry and Kluender (2018) make an argument that the AoA effects is often conflated with the effect of subsequent L2 learning. Instead it should concern L1 acquisition in early life, as L2 learning, regardless of when, can be influenced and potentially hampered by L1 grammatical structure and neural circuitry established during L1 acquisition. Therefore, in order to truly test the effect of AoA on L1 acquisition after early childhood, they examine the case of American Sign Language late L1 learners. Deaf children are often only exposed to language input in sign language at a relatively late age, as opposed to hearing children who will be exposed to linguistic input from birth. Whereas deaf children are capable of spontaneously learning sign language when exposed to it from birth, only a small number of deaf children are born into deaf families, and therefore most deaf children will not be born in a linguistic environment suitable to spontaneous acquisition. These late onset L1 learners thus provide crucial insights into the critical period hypothesis. Mayberry and Kluender (2018) then proceed to provide an extensive review of late L1 acquisition studies and compare these with late L2 acquisition studies. This revealed substantial differences between late L1 and late L2 acquisition, with much larger AoA effects for late L1 learning than for late L2 learning. Their findings therefore still support the critical period hypothesis, but provide a new perspective on its severity in late L2 acquisition.
For these reasons it is, once again, important to stress that our target participants need not be native-like but only highly proficient L2 speakers. Furthermore, the LEAP-Q is employed in order to account for individual differences with regards to experiential factors. This questionnaire is generally comprehensive and covers variables such as time spent in specific linguistic contexts (i.e., living in/with a country or family where L2 is spoken as a native language) as well as current usage of respective languages. Additionally, it would offer insights into when participants started learning languages to account for potential AoA effects.

1.6. Language interaction effects

Seeing as most participants are expected to be (relatively) late onset bilingual speakers of Dutch and English, factors such as language interaction effects between the L1 (which is also likely to be the dominant language) and the L2 may also be present. This is exemplified by Dussias and Scaltz (2008), who examined the effects of subcategorisation bias in L2 processing with Spanish-English L2 speakers. Subcategorisation bias assumes that certain verbs prefer certain complements, and this influences the processing of temporarily ambiguous sentences. For example, when the verbs used usually have a direct object complement, but can also have an empty object position; with the following NP actually being the subject of an embedded clause. Dussias and Scaltz (2008) employ a self-paced moving window reading paradigm, as opposed to eye-tracking, which does not allow for measuring regressive fixations to earlier portions of the sentence, but nonetheless allows for observing a garden-path effect with an increase in reading times. Therefore, their findings are crucial for evidencing L1 influence on L2 processing, which is of particular relevance to the current study.
The experiment consisted of two phases: first a monolingual Spanish control group performed a norming study to determine subcategorisation bias across 130 Spanish verbs, which contained the translation equivalents of the English verbs later used during the self-paced reading test. The Spanish-English bilingual group performed an English norming study after the self-paced reading test, to see if they had acquired a native-like subcategorisation bias for the English verbs used during the test (compared to the norming in English by native speakers in a previous study). These

norming studies showed that Spanish and English categorisation bias varied for the verbs and their translation equivalents in the respective languages, as well as showing that the bilinguals had successfully acquired the English subcategorisation bias. The second phase pertained to the previously mentioned Spanish-English bilinguals as well as a monolingual English control group, and concerned a self-paced reading paradigm with comprehension questions following each stimulus.
Their results showed that the monolingual English speakers and Spanish-English bilinguals did indeed differ in their employment of subcategorisation bias. The authors deduced that this was likely to be due to L1 influence, as the Spanish-English bilinguals did show a knowledge of the English subcategorisation bias through the norming study. Consequently, a further analysis was performed which ruled out L1 information, and this showed a main effect of bias as well as an interaction between verb bias and sentence continuation (i.e., verbs with a sentential complement). Dussias and Scaltz (2008) therefore conclude that L2 readers do employ the same structural cues to guide their reading, but also resort to L1 information to compensate for absent knowledge. So, in bilingual processing both language models (English and Spanish) are likely employed during the reading of (L2) sentences, even in high-proficiency bilinguals.
A more recent study that is also of particular relevance to our bilingual participant L1 background is Roberts and Liszka (2021), which examined self-paced reading of garden-path sentences with high-proficiency L2 learners of English from German, French and Dutch language backgrounds. Their study examined cross-linguistic effects of aspect from these differing language backgrounds in English garden-path processing. As the grammatical encoding of tense and aspect differs across English, French, German and Dutch, different interactions were expected between the L1's and L2 English. Specifically with regards to the (non)-existence of certain opposition pairs in the language. These include past/non-past, perfective/imperfective and progressive/non-progressive. The authors predicted that L1 grammatical aspect will influence L2 (English) processing of temporarily ambiguous sentences only when the progressive aspect is used, as all code the simple past similarly. Their findings confirmed this hypothesis: the L2 speakers of English, regardless of L1 background, had native-like performance on temporarily ambiguous sentences with simple past tense verbs. When the past progressive was used, however, there were strong L1 influence effects: German L1's treated both aspects similarly, as the progressive aspect is not grammatically coded in German. French L1's performed very similarly to the monolingual English control group, as French does encode the progressive through use of the perfective aspect. And finally, the Dutch L1's made up a middle ground between the French and German L1 speakers; progressive aspect is not coded in Dutch, but the 'zijn + aan het + inf' (literally '*be + to + infinitive*') construction resembles progressive aspect enough to provide an advantage over the German L1's in offline (acceptability) judgements.
In conclusion, language interaction effects do not only appear to have significant influence on L2 processing (even in high-proficiency speakers), the nature of these influences also differ depending on language background. This also complicates comparisons between garden-path processing studies with bilinguals when a categorical theory of bilingualism is adhered to: for example, Brothers et al. (2021) had bilingual participants performing worse than native English speakers across the board, but employed Chinese-English bilinguals. Due to the typological differences between English and Chinese, which are from completely different language families and environments, it is hard to compare this study with Teubner-Rhodes et al. (2016), who employed Spanish-Catalan (both Romance languages) bilinguals, and found advantages in comprehension. Therefore, it is of vital importance to be aware of specific language interaction effects when working with bilingual speakers.

1.7. Eye-tracking to measure processing

As observed in the above sections, reading studies within the field of eye-tracking generally operate under the assumption that increased fixation times and/or increased likelihood of regressions are reflective of an increased processing cost. We shall also operate under this assumption, as well as clearly defining which eye-tracking measures shall be obtained, what these measure exactly and how this is reflective of processing. Whilst a definition of eye-tracking measures may seem redundant, existing eye-tracker literature sometimes suffers from non-uniformity in the terminology employed to explain the specific measures obtained, as well as variation in which measures are thought to be relevant. Naturally, this can lead to confusion or even wrongful comparisons between eye-tracking studies that employ different measures or similar measures with conflicting names. This problem can most easily be averted by clearly discussing the theoretical background of eye-tracking measures to examine processing.

However, before proceeding to a discussion of how eye-tracking measures can be insightful into processing, a limitation must also be mentioned. Indeed, as opposed to more granular processing measures such as ERP, with eye-tracking it is hard to distinguish between different types or sources of linguistic violations (i.e., syntactic versus semantic), as discussed in Clifton Jr. and Staub (2011). Nonetheless, eye-tracking measures are very temporally detailed and are therefore still meaningful for capturing syntactic processing during reading of sentences. An extensive overview of how eye-tracking measures reflect on processing is provided in Clifton et al. (2007), which also contains a literature list of a 100 articles published about eye-tracking reflecting higher-order processing. Therefore, this also forms the basis of our definitions of the eye-tracking measures to be employed. According to Clifton et al. (2007), the standard measures employed for identifying effects of syntactic factors on eye movements during reading include: first pass reading time, go-past or regression path duration (i.e., which are two names used to describe the same phenomenon), regressions-out, second pass reading time and finally, total reading time. For this particular eye-tracking study we are interested in only three of these measures, namely: First pass reading time, go-past or regression path (henceforth regression path) and total reading time. First pass reading time concerns the fixation duration from first entry until final exit of a particular region. Regression path encompasses the sum of all fixations in a region from the point of entry until exiting it rightwards specifically (i.e., moving on to next word/region). Total reading time, as the name suggests, encapsulates the sum of all fixations on a region including all forward movement and regressions to it.

In employing these specific eye-tracking measures we are following the example set in previous seminal studies with garden-path processing (e.g., Slattery et al. (2013), Paape et al. (2018), Christianson et al. (2017) and Brothers et al. (2021)). In that they have all previously been associated with syntactic processing difficulty (see Clifton et al. (2007) and Clifton Jr. and Staub (2011)), both in studies with monolingual and bilingual participants. First fixation as a measure is also employed in some of these studies (e.g., Brothers et al. (2021)), but in this thesis this measure is consciously left out, as the critical area for disambiguation is contained in the middle of the stimuli sentences (see section 2.3 for examples) and "when regions are long and the disambiguating material is not likely to be included in the initial fixation, the first fixation measure is inappropriate." (Clifton et al. (2007), p. 349).

1.8. The current study

The current study aims to investigate bilingual processing of garden-path sentences with high-proficiency Dutch-English bilinguals as compared to English native speakers. We employ a clear definition of what is considered bilingualism as well as what theoretical framework shall be used to

explain which differences in processing are expected, and how these will be explained. We define bilingualism from the standpoint of the factor mixture model (Kremin and Byers-Heinlein (2021)), in which monolingual and bilingual are separate categories, but there are continuums contained within these two categories. This allows for distinctions to be made within the bilingual category on the basis of proficiency, which shall be further supported by evidence from the LEAP-Q, in order to account for potential individual differences within categories. The theoretical frameworks underlying our hypotheses below are the good-enough parsing model as proposed in Ferreira et al. (2002) and improved in Ferreira and Lowder (2016) to account for the mechanism governing syntactic parsing. Finally, we expect bilingual processing *differences* rather than a strict bilingual *advantage* (in executive functioning or otherwise). This expectation is based on parallel activation and interaction between both language models as proposed in Dijkstra and Van Heuven (2002) and Dijkstra et al. (2019) and evidenced in van Heuven and Dijkstra (2010) (lexical access), Dussias and Scaltz (2008) (structural cues) and Roberts and Liszka (2021) (grammatical aspect effects). Eye-tracking shall be used to gather the reading data, and will hopefully provide new insights into bilingual processing of garden-path sentences and the mechanisms that underlie this. Specifically, we employ the measures of first pass, regression path and total reading time to garner information on reading patterns of our target stimuli. First pass should evidence an increased reading time when encountered with the disambiguating region (as shown in Section 2.3.2). Regression path will enlighten how much re-reading is done, as well as in which regions of the sentence, when reanalysis occurs. Finally, total reading time shall be able to provide an overall glance of the processing costs of each region.

1.8.1. Hypotheses

The hypotheses this study operated under are as follows:

1. Dutch-English bilingual speakers will have a statistically significant advantage in comprehension question accuracy of the target stimuli as compared to the native English speakers; i.e., a lesser garden-path effect leading to lingering misinterpretations.

2. (a) Dutch-English bilingual speakers will not display universal statistically significant disadvantages or advantages in duration of eye-tracking measures in the target stimuli as compared to the native English speakers due to high L2 proficiency.

   (b) Instead, different reading patterns are expected for Dutch-English bilinguals as compared to the native English speakers due to parallel language activation, particularly surfacing through shorter reading times in the regression path measure due to Dutch L1 knowledge supporting English L2 knowledge.

3. Dutch-English bilinguals will display similar performance, as assessed by reaction time (in ms) and accuracy, to the native English speaker control group in the cognitive control tasks.

The above hypotheses have arisen from the existing literature and our perspective on the ongoing debate about bilingual garden-path processing and the bilingual advantage in general. Hypothesis 1 reflects the results of Teubner-Rhodes et al. (2016) and, to a degree, those of Brothers et al. (2021), as in the latter case proficiency fully predicted performance on comprehension and eye-tracking measures, and the current study employs strictly high-proficiency bilinguals. Furthermore, based on the computational models discussed in section 1.3.2 and the parallel activation proposed in these as well as Dussias and Scaltz (2008), which showed activation of both the L1 and L2 verb bias, we

can expect bilinguals (especially in the case of typologically close languages) to process the input in both languages; therefore having a higher chance of arriving at the correct interpretation. If this hypothesis is confirmed we can assert that high-proficiency bilinguals process garden-path sentences differently from native English speakers. Hypothesis 2a is similarly based on the previous findings by Teubner-Rhodes et al. (2016) and Brothers et al. (2021), and for the same reasons as outlined for hypothesis 1. Previous studies that found disadvantages for bilinguals in reading times and eye-tracking measures (such as Brothers et al. (2021) and Jacob and Felser (2016)) also made mention of proficiency as being an influential factor, which should be controlled for in our study through using solely high-proficiency bilingual speakers. Hypothesis 2b stems from the computational models of van Heuven and Dijkstra (2010) and Dijkstra et al. (2019) that propose parallel activation of language systems in bilinguals, as well as proven language interaction as in Dussias and Scaltz (2008) and Roberts and Liszka (2021). These lead us to expect limited advantages due to the L1 supporting the L2, especially in a measure most strongly associated with garden-path processing: the regression path. Finally, hypothesis 3 is based on the findings of Brothers et al. (2021) with regards to individual differences, as well as meta-analysis such as Paap et al. (2015) which did not find systematic bilingual advantages in domain-general cognitive control. Critically, if this hypothesis is confirmed alongside hypothesis 1, we can assume that the cause for the advantage observed in accuracy and potentially some eye-tracking measures would not be enhanced cognitive control, but rather something else (i.e., parallel activation).

2. Methodology

All procedures described in the methodology below conformed to the Leiden University Ethics code of the Faculty of Humanities.

2.1. Pilot experiment

The entire experiment was conducted in English, both the online pre-test and the eye-tracking session. However, neither the experimenter nor the supervisors partaking in this study were native English speakers, but high-proficiency L2 English speakers. For that reason, as well as to test the experimental paradigm in general, a pilot experiment was performed. This pilot experiment included two native speakers of English, as well as two high-proficiency Dutch-English bilingual speakers, who underwent the entire experimental procedure. In contrast to the actual participants, they were asked to pay special attention to the content and format of the experiment to see if everything made logical sense and (especially for the two native speakers) if the experimental stimuli were fitting. The latter criterion was of specific importance for the target (garden-path) stimuli and their corresponding comprehension questions, as these should not allow for creative interpretations to avoid incorrectly interpreting wrong answers as comprehension/revision difficulties. The experimental paradigm (both instructions and stimuli, where appropriate) was adapted based on the feedback received from these pilot participants to ensure that the actual experiment would yield high quality data.

2.2. Participants

For this study we employed 32 participants, divided into two groups: high-proficiency Dutch-English bilinguals (N = 20) and English native speakers (N = 12). These were recruited from the university population at Leiden University to assure a relatively homogeneous participant body. The bilingual group in particular was recruited among second and third year students of the English

Language and Culture Bachelor as well as the Linguistics Bachelor, to ensure a high-proficiency in English as the L2. Mean age of the bilingual group was 23.65 (*SD 8.03*) and mean age of the native English group was 25.5 (*SD 3.87*). Data on gender was not collected, as previous eye-tracking research on garden-path processing never suggested gender playing a significant role. Informed consent was obtained separately and individually before both the online pre-test (ticking a consent box) and the on-site lab session (in writing). Participants received monetary compensation conforming to the Leiden University Centre of Linguistics (LUCL) guidelines for the time they spent in the eye-tracking lab and the time spent conducting the online questionnaire, upon completion of the entire procedure.

2.3. Stimuli

The target stimuli consisted of 30 couplets of temporarily ambiguous sentences, in which the subject of the main clause is ambiguous in that it could initially also be the object of the embedded matrix clause due to the presence of an ergative verb. The unambiguous version consisted of the exact same sentence, but with an extra noun phrase (NP) inserted to resolve the ambiguity, providing a separate object for the ergative verb from the subject of the main clause. All target stimuli and their unambiguous counterparts were constructed by the experimenter, although some sentences were inspired by those of previous studies examining garden-path effects (e.g., Ferreira and Henderson (1991), Slattery et al. (2013), Novick et al. (2014), etc.). Furthermore, all target stimuli also employed solely past simple tense, as Roberts and Liszka (2021) found conflicting grammatical encoding within English and Dutch with regards to the perfective and progressive tense, and we wanted to avoid L1 interference on that account in the Dutch-English bilinguals, instead hoping for L1 support rather than conflict. An example couplet:

1. After Charles shaved the sheep stopped grazing.

2. After Charles shaved the sheep the cow stopped grazing.

The critical stimuli were intermixed with 60 filler stimuli of varying complexity. These were divided into 30 simple fillers and 30 complex fillers, in order to distract participants from the garden-path sentences. If the target stimuli would have been the only complex sentences present, an attention bias towards these sentences would likely have developed in the participants. The simple filler stimuli were adapted from stimuli previously used in an eye-tracking study by Ferreira and Henderson (1990), with some adaptations based on pilot participant feedback to make them more natural and/or more formal when necessary. The complex filler stimuli consisted of twenty sentences made up by the experimenter and checked by both the supervisors and pilot participants, and ten sentences which were also adapted from a previous study examining complex sentence structure (Johnson et al. (2011)). The stimuli were pseudo-randomised through the in-built randomiser of the Experiment Builder software, to assure that the two sentences that together make up a target couplet would never appear directly after one another. Randomisation occurred based on the anonymous participant ID assigned at the start of the eye-tracking session, so that no two participants saw the exact same order of stimuli.

2.3.1. Comprehension questions

The temporarily ambiguous sentences, and their non-ambiguous control counterparts, were accompanied by comprehension questions which were presented after each stimulus, to assess the

presence of lingering misinterpretations. The comprehension questions could be constructed in several different ways. The main distinction depended on which part of the temporarily ambiguous clause was targeted:

3. Did Charles shave the sheep?

4. Did the sheep stop grazing?

This was done to examine the effects of information structure principles, in which it is stipulated that participants pay more attention to the main clause in sentences like 1 due to associating the latter part of a sentence with novel information, as described in Ferreira and Lowder (2016). Therefore, to truly test overall comprehension, the questions should not only target the main clause but also the embedded clause of the sentences. Additionally, in the case of the example sentence provided in Example 1, and in other sentences with optionally reflexive verbs, questions directly targeting the subject of the embedded clause were also employed:

5. Did Charles shave himself?

We would therefore expect that those questions pertaining to the embedded matrix clause, which is the part of the sentence for which Ferreira and Lowder (2016) poses good-enough parsing is used, to be subject most to lingering misinterpretations. This would mean that Example 3 would be more likely to have an incorrect answer due to good-enough parsing, whereas questions such as Example 4 should be affected to a lesser degree (if at all). Furthermore, Example 5 is something of a special case, as it employs a reflexive verbs. These would be treated similarly to Example 3 in the case of the native English speakers, but the Dutch-English bilinguals are likely to display a different pattern: in Dutch, almost all reflexive verbs have an obligatory reflexive pronoun complement. Therefore, if the bilinguals do indeed activate both of their language systems in parallel, as expected on the basis of Dijkstra et al. (2019), then the activation of Dutch (and the corresponding insertion of the obligatory reflexive pronoun in the Dutch syntactic structure) could lead to more accurate representations of the embedded clause when a reflexive verb is present. Finally, the comprehension questions for the temporarily ambiguous sentences were counterbalanced to ensure an equal distribution of 'Yes' and 'No' answers. In every case, the control sentence would have the opposite correct answer from the temporarily ambiguous sentence.

2.3.2. Critical regions

For the sake of examining garden-path effects, the target stimuli have been divided into critical regions. This allows for comparison of reading measures (and patterns) between similar regions in the ambiguous and control stimuli. Usually, a garden-path effect manifests itself as an increase in fixation times in the ambiguous regions, which is often paired with regressions to previous regions to re-examine earlier syntactic content. Seeing as we only expect this effect to occur in the target stimuli, the filler stimuli were not divided into similar regions, nor was their sentence length controlled for similarity to the target stimuli. The filler stimuli shall therefore only be used for accuracy comparisons, as no meaningful eye-tracking data has been generated for these. The distribution of the regions of interest for the target stimuli follow example in Table 1.
The disambiguating region in the ambiguous condition is region 5, as marked by the italics, which corresponds to the main clause verb phrase (VP) in region 6 for the control condition. Naturally, comparing the reading measures of the disambiguating region in the ambiguous condition to the same VP in the control condition will allow us to assess processing difficulties induced by the

1	2	3	4	5	6	7
After	Charles	shaved	the sheep	*stopped*	grazing.	
After	Charles	shaved	the sheep	the cow	*stopped*	grazing.

Table 1. An example of the regions of interest used for the target stimuli

ambiguity. However, a comparison of region 5 of the ambiguous condition to region 6 of the control condition is not the only interesting data that can be retrieved. In fact, most eye-tracking studies on the reading of garden-path sentences (e.g., Slattery et al. (2013) and Brothers et al. (2021)) take more, if not all, regions into account. This is done as the ambiguity is not only expected to lead to longer fixation times on the critical regions itself, but also to induce regressions to earlier regions in order to reassess the previous syntactic information. Additionally, effects might be observed in spillover regions, for example after the ambiguous region, due to an increased wariness of participants after having encountered a previous need to reanalyse. Therefore, the current study shall examine potential effects in the eye-tracking reading measures of regions 2-7, as these are likely to include garden-path effects. Further explanations and predictions with regards to the effects in these regions is covered in section 2.6.
The full list of stimuli used, divided into critical stimuli pairs and filler sentences, is made available in Appendix A. Target stimuli in the appendix are divided into the regions as discussed above, which is indicated through a separation of regions with a pipe: '|'. Additionally, the accompanying comprehension questions and target answers are also made available for each stimulus.

2.4. Pre-test: self-reported profiency and background variables

The entirety of the pre-test was done in the online environment of Gorilla.sc, which offers free tools for scientists to build questionnaires and behavioural tasks and integrate these together into experiment trees. The experimental tree for the pre-test consisted of an informed consent form, after which some meta-data (including part of an adapted LEAP-Q) is asked for; specifically, age, native language, which languages someone speaks and language dominance. These are not only important for assessing the language background, but also allowed us to filter out participants that did not fit the inclusion criteria (i.e., being 18 years or older to conform to ethics regulations and being a native Dutch or native English speaker, depending on group), as recruitment was also done online and via posters. When participants met the inclusion criteria, they were presented with the remaining part of the adapted LEAP-Q, which covers language exposure and history. After completion of the background questionnaire, the participants were automatically re-directed to a short cognitive control battery consisting of online adaptations of a Flanker task (48 trials), Simon task (36 trials) and a Spatial N-back (go/no-go with 12 sets, each consisting of between 3 to 5 go/no-go trials) task. The trial numbers across the tasks were balanced so that they were of somewhat similar length, and the tasks themselves were presented in a counterbalanced order to participants to avoid order effects within the output data. For the cognitive control tasks reaction time (in ms) and accuracy (between 0 and 1) was measured.

2.4.1. Adaptations to the LEAP-Q

For the sake of functioning properly in Gorilla.sc's online environment, some adaptations had to be made to the original LEAP-Q format. Whilst there is a digital version of the LEAP-Q made

available by Marian et al. (2007), the Gorilla.sc platform makes use of particular pre-fashioned widgets to allow experimenters without coding experience to also build questionnaires and tasks from the ground up. This meant that some of the questions concerning 'List languages in order of X' present in the LEAP-Q had to be transformed to a drag and drop ranking from top to bottom, with a prefab of several common European languages and room for three 'others' which could be specified in a separate text-box. These had to remain 'Other 1', 'Other 2' and 'Other 3' throughout the questionnaire, rather than actually becoming the filled-in language in the text-box, as Gorilla does not allow referencing back to previous answers without creating your own lines of code. In addition, some questions regarding cultural background and environmental factors influencing how languages were learnt were left out of our adaptation of the LEAP-Q. This choice was made as we were mainly concerned with current proficiency and exposure to the language rather than the entire socio-linguistic background of all participants. The exact questions and the experimental template used in Gorilla.sc for this experiment will be made available upon request.

2.5. Eye-tracking equipment and procedure

2.5.1. Apparatus

Eye movements of the dominant eye were monitored with a desk-mounted Eyelink 1000 Plus with a sample rate of 1000 Hz. Participants were seated in front of a desk which contained a chin rest, keyboard and mouse, the eye-tracker and at the far end a display monitor, with their head rested comfortably in the chin rest. The distance between the eye-tracker and the chin rest was approximately 70 cm, and the eye-tracker itself was situated about 10 cm away from the experiment display monitor. The monitor was a BenQ Senseye 3 LED monitor with a resolution of 1024x768 and a refresh rate of 60 Hz. Camera calibration was performed at the start of each session using a 9-point grid with a validation sequence. The stimuli were presented in Times New Roman 20 pt font on the experimental display monitor present in the eye-tracking laboratory booth. Re-calibration was performed whenever necessary, as determined by observed drift during the experiment, or whenever participants decided to take a break in between blocks (as this generally involved getting out of the chin rest).

2.5.2. Procedure

Upon arrival in the eye-tracking lab, participants first received an information sheet and consent form pertaining to the experimental session specifically, after the signing of which they received a further briefing from the experimenter. Participants received instructions to move their head as little as possible during the experiment, as well as on what to do during the experiment. They were told to read the sentences as quickly but attentively as possible, and that they could press any key as soon as they finished reading the sentence to proceed to the comprehension question. Participants were told to answer the comprehension questions with either 'F' for 'NO' or 'J' for 'YES'. The comprehension questions were also presented together with a reiteration of the answer instructions on the bottom of the screen. After a suitable calibration was achieved, participants underwent a short practice block, which also allowed the experimenter to assess if the calibration was truly successful by monitoring the location of the pupil recording on the displayed sentence. The practice block contained five stimuli, during which the experimenter ascertained that the participants understood the instructions. Finally, after successful completion of the practice block, the participants were asked again if everything was clear and had a chance to ask clarification questions if necessary, after which the experiment proper began.
The experimental stimuli were presented in fifteen blocks of eight, for a total of 120 stimuli. Each

stimulus was preceded by a drift check which also focused the participants' gaze back to the centre-left of the screen, so that they would start reading stimuli sentences from the left boundary. Between each block there was a break sequence with a clearly indicated break screen that instructed the participants to either take a break if they felt like it or ignore the screen and continue to the next block if they did not. This break screen also functioned as a camera setup screen for the experimenter, therefore allowing easy re-calibration after breaks as well as an opportunity for the experimenter to announce that another calibration was required in case of drift during the preceding block. In total, the experiment usually lasted approximately 45-60 minutes.

2.6. Data analysis and predictions

The analysis performed on the data of the pre-test was mainly focused on validating the language background of the participants of both groups (i.e., presence of any second languages for the native English speakers and ensuring Dutch and English were the most dominant languages in the Dutch-English bilingual group). Additionally, it served to allow us to establish baseline cognitive control abilities through analysing the response times and accuracy of the short cognitive control task battery contained in the pre-test. The latter consequently also serves as a potential fixed effect in the proposed analysis model discussed below. Seeing as two outcome measures were obtained through our experiment, a logistic accuracy value (as determined by percentage of correct answers to comprehension questions) and continuous eye-tracking measure values (first pass, regression path and total reading time; all measured in ms), separate statistical analyses are required for each. Nonetheless, similar principles can be applied to the approach of the statistical analysis.
Recent years have seen a shift in statistical procedures within the field of psychology and psycholinguistics, with a strong focus on modeling participant and item effects without separate analysis (as would be required in the previously popular ANOVAs). This has caused hierarchical regression to be one of the most popular methods of analysis within psycholinguistics, but these are not always adhere to the same standards. Therefore, Barr et al. (2013) argued for the use of maximal models for confirmatory hypothesis testing, as the inclusion of all possible random effects in a mixed model would drastically increase generalisability of the model. However, later studies (notably Bates et al. (2015) and Matuschek et al. (2017)) indicated that this maximal modelling also has its own shortcomings: models that are potentially more complex than the data supports may fail to convergence, and a maximal model may also overfit the data and lead to a loss of statistical power. Therefore, this paper also adheres to the parsimonious mixed effects model as described in Bates et al. (2015) and Matuschek et al. (2017). All statistical data analysis and construction of the models is done in the R environment (R Core Team (2021)). The following subsections cover which fixed and random effects are included into our proposed models, and why.

2.6.1. Accuracy

Accuracy is measured through the binary (yes/no) outcome variable of the comprehension questions that accompany the stimuli. The filler stimuli are mainly present as distractors and have varying degrees of complexity, so are not a good fit for a regression model due to too much variability. Instead, these are examined from a descriptive statistical viewpoint as a means of comparison between overall accuracy in filler stimuli and overall accuracy in target stimuli. For the separate target stimuli analysis logistic regression is required, as the outcome variable is binary. Nonetheless, the effects structure of such a model still conforms to that of linear hierarchical regression models. The following are the proposed main effects (in order of importance) and the random effects that

are in accordance with out data and expectations.

Fixed effects

1. Language (Dutch-English bilingual/native English)
2. Type (ambiguous/control stimuli)
3. Question type (main clause/embedded clause)
4. Stimulus order (early/late stimuli)
5. Cognitive control (in three levels: low. mid and high, if found to be significant in pre-test results)

Random effects

1. Random intercept for participants
2. Random intercept for item
3. Random slope for item

Due to the design choices made in our experiment, discussed above and justified in the Introduction section, the fixed effects should be relatively self-explanatory. The random effects chosen to be included might warrant some explanation. We have chosen to incorporate random intercepts for both participants and items. The random intercept for participants allows the model to account for small between-participant differences with regards to general reading comprehension and language skills. The random intercept for items accounts for potential differences in participants' responses to differing items: some might be more acquainted with the context or content of a particular stimulus than others. Finally, a random slope for item is included as we expect there to be potential variability in sensitivity to several fixed effects that have to do with particular items. For one, the Language fixed effect might vary depending on item, as also shown in Dussias and Scaltz (2008), where bilingual speakers sometimes employed the subcategorisation bias of their L1 when processing their L2 depending on how well acquainted they were with the particular word or its context. Furthermore, the fixed effect of Question type could vary in its sensitivity to particular items, particularly for main clause questions accompanied by a reflexive verb in case of the Dutch-English bilinguals (see section 2.3). The final interaction that supports our choice for a random slope for item is that between item and the fixed effect of Type, which once again has to do with subcategorisation bias: some of the target verbs might be more or less likely to be followed a NP or VP complement. We chose to refrain from including a random slope for participants, as incorporating this into the model could lead to an increased Type II error, as our main independent variable of Language already depends on different sensitivities to the stimuli between participants, and we expect our two groups to be largely homogeneous within the groups themselves.

2.6.2. Eye-tracking measures

As opposed to the accuracy measure, the eye-tracking measures are continuous, as they are fixation duration measures in milliseconds. Therefore, the regression model for this analysis will be a linear mixed effects model, and it will also differ in some aspects from that as shown above. Some main effects that were relevant to accuracy, will not be relevant for the eye-tracking measures, and the reverse is also true. Notably, the fixed effects that pertain to the comprehension questions

present in the model for accuracy are not included in the eye-tracking measures model. This is due to the fact that the comprehension questions were presented after the sentence stimuli, and the eye-tracking measures are only relevant (in our study) for the actual sentence reading. As before, the fixed effects are again presented in order of importance, after which the random effects are presented and discussed.

Fixed effects

1. Language (Dutch-English bilingual/native English)
2. Type (ambiguous/control stimuli)
3. Stimulus order (early/late)
4. Subcategorisation (in three levels, similar, less probable and improbable)
5. Cognitive control (in three levels: low, mid and high, if found to be significant in pre-test results)

Random effects

1. Random intercept for participants
2. Random intercept for item
3. Random slope for participants
4. Random slope for item

A fixed effect for Subcategorisation is included on the assumption that certain target verbs used in the stimuli can be more or less likely to be followed by specific parts of speech. This is separate from potential random effects for item, which already allows variability based on the specific content of the stimuli, as it is objectively measurable. In the current study, we employed SketchEngine (Kilgarriff et al. (2014)) in order to assess which parts of speech are likely to follow the relevant verbs used in our stimuli. This assessment was done based on the enTenTen corpus (Suchomel (2020)) which contains over 36 billion English words which have already been tagged with regards to part of speech through the TreeTagger tool (based on Santorini (1990)). SketchEngine was used to select random samples of 500 items for every verb used in our stimuli, after which a concordance analysis was run to examine which parts of speech most often followed the verbs. The ambiguous target sentences have the main clause verb followed by a noun and then the disambiguating verb, whereas the unambiguous control sentences have the main clause verb followed by two nouns (see Table 1). Therefore, the concordance analysis was used to examine the parts of speech of the two words following the main clause verb within the random sample from the corpus. A ratio of relative likelihood of N-N to N-V following the verb was calculated, and consequently the subcategorisation bias fixed effect was based upon that ratio. If the ratio, and therefore the likelihood of the verb to be followed by two nouns in comparison to it being followed by a noun and a verb, was 2 or less, the verbs were classified as 'similar'. If the ratio was higher than 2 and lower than 5 the verb was classified as 'less probable'. And finally, if the ratio was higher than 5 the verb was classified as 'improbable'. The probability of the terms being based on how likely our ambiguous stimuli are to occur as compared to the control stimuli. In total, this led to 10 verbs being classified as 'similar', 8 verbs being classified as 'less probable' and 12 verbs to be classified as being 'improbable'.

Similarly to the accuracy model, a random intercept for participants is included, this time to account mainly for variability in reading patterns (regardless of group-level differences). The random intercept for items once again relates to how different items might be intrinsically harder to read or process. As opposed to the model for accuracy, we have incorporated a random slope for participants in the eye-tracking measures model as well. This was done as we do expect variability in participants' sensitivity to time during the experiment, which mainly relates to the fixed effect of Stimulus order. However, that is not the only potential interaction which justifies the choice for a random slope for participants. The fixed effect of Language could also play a part, as we allow for variability within the two categories of our language groups (following the factor mixture model from Kremin and Byers-Heinlein (2021)). Furthermore, the fixed effects of Type and Cognitive control might also interact with a random slope for participants, as different participants might engage their cognitive control facilities differently for the garden-path sentences and their control counterparts. Finally, a random slope for item is also included, as there is likely to be variability in the sensitivity of specific target stimuli to the fixed effect of Stimulus order (different target words may be more/less fatiguing or learnable). Additionally the random slope for item is justified by a possible interaction with the fixed effect of Subcategorisation, as specific words may be more or less expected in the context as well as having a differing subcategorisation bias.

2.6.3. Different eye-tracking measures and regions

The above regression model to analyse the eye-tracking measures is a general model, the eye-tracking measures , however, are not simply a single outcome measure. As discussed in section 2.3.2, there are multiple regions that need to be considered, as well as multiple eye-tracking measures (first pass, regression path and total reading time). Some regions can be more prone to displaying certain effects than others, which is why many of them need to be separately compared. Most regions can be compared on a 1:1 basis between the ambiguous and control condition. However, as shown in Table 1 by the italics, the critical region in the ambiguous condition is region 5, whereas in the control condition it is region 6 (in either case, the VP of the embedded clause). This leads to a straightforward comparison between regions 2-4 between the conditions, after which a slight shift occurs, and ambiguous region 5 and 6 are compared to control region 6 and 7 (n + 1; as an extra NP is inserted in the control condition to resolve the ambiguity). All these comparisons are done separately for these five sets of regions, as well as for the three eye-tracking measures employed for this experiment, which therefore resulted in fifteen separate regression models.

2.6.4. Predictions

Not all stimuli regions are expected to be affected in a similar fashion by garden-path effects, which therefore also leads to varying predictions as to what will be observed with regards to each eye-tracking measure in the relevant regions. For the first pass measure, the fixation duration between first entry and first exit of a region, we do not expect any main effects in regions 2-4, as no ambiguity has arisen yet, so there would not be any obvious reason for increased processing cost as evidenced through longer fixation times. When comparing region 5 of the ambiguous condition to region 6 of the control condition, on the other hand, we do expect to see a garden-path effect in the first pass measure, leading to an increased fixation time in the ambiguous condition upon encountering the critical region. A similar effect, but to a lesser degree, could be observed in the spillover region of the ambiguous condition for the first pass, as the participants might be wary due to the previous ambiguity and therefore take longer to read subsequent regions as well.
Opposed to the effects expected for the first pass measure, regression path is expected to have a ma-

jor impact on regions 2-4 in the ambiguous condition. Upon having encountered the disambiguating critical region 5, and (hopefully) realising that the previously constructed syntax is erroneous, it is very likely that the participants will have leftward regressions. Particularly, we expect most, and the longest regressions, to return to region 3, which is the initial, and likely erroneously parsed, VP. Additionally, regressions may occur to regions 2 and 4, to reassert the syntactic structure and complements of the main clause and VP. Some effects of regression path on region 5 in the ambiguous condition may also be observed, as this includes all fixations in the region until a rightward exit occurs, hence when returning to the critical VP after having regressed adds to the regression path measure.
Finally, the total reading time is largely dependent upon the previous measures, and therefore we mainly expect to observe garden-path effects in the ambiguous condition in regions 2-5, and only possibly in the spillover region 6. The strongest effect in total reading time would likely be observable in region 3 and 5 of the ambiguous condition (as compared to region 3 and 6 of the control condition), due to the VPs being the main causes of first ambiguity, and then disambiguation. This most likely makes these the hardest to process, and therefore also the ones with the longest total fixation duration.

3. Results

3.1. Pre-test

All participants completed the pre-test before they came to the eye-tracking lab. Before commencing the pre-test, participants were asked for informed consent as well as subjected to an age check to assure that all our participants were of legal age to give their informed consent (18 in the Netherlands). The descriptive measures we obtained from the pre-test are outlined below.

3.1.1. LEAP-Q

The LEAP-Q Marian et al. (2007) was used to ensure high proficiency of English in the Dutch-English bilingual speakers, as well as to observe whether the native English speakers were monolingual or if there might be interfering language competencies. The survey confirmed that for our Dutch-English bilinguals Dutch and English were the two most dominant languages, and participants have a mean exposure of 42.1% (*SD 16.23*) to English in their current daily lives. Additionally, participants that reported having spent time in an English language educational environment (N = 18) had a mean of 3.68 (*SD 2.64*) years in such an environment. The two Dutch-English participants that reported 0 years spent in an English language educational environment, instead spent an average of 21.38 (*SD 0.88*) years in an English speaking family. We believe that these results are sufficient to support our claim that the Dutch-English bilinguals were of high proficiency in English. Of the native English speakers, three participants also had a second native language, which were all divergent: one Turkish, one Filipino and one Hindi. An additional four also had language competencies in other languages with two Mandarin speakers, one German speaker and one Swahili speaker. These seven native English participants all still reported English to be their most dominant language, which was also backed up by their relative exposure to English being on average 71.57% (*SD 19.98*). Of the second languages spoken by these seven participants, only German comes from a similar language family, and the relevant participant reported a 60% exposure to English compared to a 5% exposure to German. Additionally, they rated (on a scale of 1 to 10) their own proficiency in German relatively low at a 6 for speaking proficiency, a 5 for spoken understanding proficiency and a 7 for reading proficiency. We have therefore treated this participant

as a regular native English speaker for our analyses. The participants of other L2 backgrounds were all asked about the temporary ambiguity of our target sentences in their non-English language after completion of the eye-tracking session. All participants responded that in their non-English language the temporary ambiguity still existed similarly to the English target sentences used in the study. We therefore assume that their behaviour will be similar to monolingual English speakers, as even with activation of both language systems, there should not be any inherent advantage to processing similar to the Dutch-English interaction.

3.1.2. Cognitive control tests

The short cognitive control battery consisted of three tasks that were presented in a random order to each participant by Gorilla.sc's in-built randomiser. Reaction times (in ms) and accuracy (0 or 1) were recorded for all trials within the different tasks, but for the N-back task reaction time was only relevant for go-trials (as with no-go trials participants should not respond). For every task we first examined the distribution of reaction times in order to determine the relevant statistical tests to be used for analysis. In all three tasks the reaction times were not normally distributed, hence non-parametric tests were used for comparison; specifically, Kruskal-Wallis tests for reaction times and two-sided proportion tests with continuity correction for the binary accuracy data.

In the Flanker task, we observed the expected Flanker effect (i.e., slower reaction times in incongruent trials as opposed to congruent trials) which had a significant effect at Chi-square = 5.46, df = 1, $p = 0.0195$. However, the effect was not large as the mean reaction time for the congruent condition was 543ms (*SD 227*) and the mean reaction time for the incongruent condition was 554ms (*SD 194*). The Flanker effect was also not observed in terms of accuracy, as the congruent (mean score of 0.992) trials did not differ significantly from the incongruent (mean score of 0.987) trials (Chi-square = 0.57, df = 1, $p = 0.4509$). There were no overall group level differences between the Dutch-English bilinguals and the native English speakers (Chi-square = 0.14, df = 1, $p = 0.7043$), nor for separate congruent (Chi-square = 0.21, df = 1, $p = 0.6477$) or incongruent (Chi-square = 1.09, df = 1, $p = 0.2972$) analyses by language. However, there were significant individual differences in reaction times between all participants at Chi-square 662.67, df = 31, $p < 0.001$. A consequent proportion analysis revealed that the same was not true for accuracy by participant (Chi-square = 24.25, df = 31, $p = 0.8001$), nor for accuracy by language (Chi-square = 2e-28, df = 1, p =1).

For the N-back task only reaction times of go-trials could be examined, along with overall accuracy. Due to a rare coding glitch on Gorilla.sc three participants (two Dutch-English bilinguals and one native English speaker) never got a go-trial, and were therefore left out of this analysis. In the N-back task there were significant group level differences in reaction times at Chi-square = 56.88, df = 1, $p < 0.001$ as well as individual differences at Chi-square = 200.41, df = 28, $p < 0.001$. Descriptive statistics revealed that the Dutch-English bilingual group had a mean reaction time of 599ms (*SD 155*) whereas the native English group had a mean reaction time of 467ms (*SD 160*), which is significantly faster. However, the native English group was also significantly (Chi-square = 36.70, df = 1, $p < 0.001$) less accurate with a mean accuracy score of 0.785 as opposed to the Dutch-English bilingual 0.995 accuracy score. There were also significant individual differences in accuracy between participants at Chi-square = 407.3, df = 28, $p < 0.001$.

The Simon task showed largely similar results to the Flanker task, except that the Simon effect was more readily observable. Additionally, there was one anomalous participant within the Simon task data. Namely, one of the Dutch-English bilingual participants had an accuracy score of 0 across incongruent trials, which most likely indicates either misunderstanding of, or technical difficulties

with, the task. That particular participant was therefore left out of consideration for this task as a whole. After removal of this participant, a significant effect of the Simon task was observed at Chi-square = 33.11, df = 1, $p < 0.001$ between congruent (M = 518ms, *SD 161*) and incongruent (M = 554ms, *SD 166*) trials. Additionally, accuracy significantly differed between congruent (mean score of 0.987) and incongruent (mean score of 0.903) trials at Chi-square = 38.93, df = 1, $p < 0.001$. There were no group-level differences in reaction times (Chi-square = 3.09, df = 1, $p = 0.079$) nor in accuracy (Chi-square = 2.12, df = 1, $p = 0.145$). However, similarly to the Flanker task, there were individual differences in reaction times at Chi-square = 318.88, df = 30, $p < 0.001$, without the one anomalous participant. As opposed to the results of the Flanker task, we also observed individual differences in accuracy at Chi-square = 56.40, df = 30, $p = 0.00243$.
Based on the observations above, it was clear that there were differences in cognitive control among the participants, which would have to be accounted for when analysing the experimental eye-tracking data and comprehension. Hence we did decide to include the predictor for cognitive control in the regression models for the analysis of both eye-tracking measures and accuracy, as outlined in Section 2.6. In order to determine to which cognitive control category participants would be assigned, we examined five different measures: separate RT on congruent and incongruent trials of the Flanker task, separate RT on congruent and incongruent trials of the Simon task, and a combined score for accuracy and RT on the N-Back task. The latter measure is combined, as we observed a significant group-level effect of accuracy, with a rather large difference in mean accuracy. For every measure we examined whether the mean participant score was higher or lower than the overall mean score, it if is higher they get a 'high' ranking on that measure, and vice versa. Consequently we concatenated all these 'high' and 'low' scores for each participant across all five measures, and then determined the cognitive control performance grouping based on the ratio of high to low. Participants that scored high on 4 or 5 measures were rated as 'high', participants that scored high on 2 or 3 measures were rated as 'mid' and those that scored high on 0 or 1 measure were rated as 'low'. This led to 13 participants in the 'high' group, 11 participants in the 'mid' group and the remaining 8 participants were in the 'low' group.

3.2. Eye-tracking measures overview

Before any analysis was done, we first removed any fixations less than 50 ms in duration (based on Rayner (2009) which stipulates that meaningful fixations during reading can be as short as 50ms), this was about 9.15 % of all data, including that from filler and practice items. Additionally, we determined extreme outlier cutoffs per language group as we expected potential differences in reading times and patterns. We calculated group means and standard deviations per measure and per region, both for target and control sentences separately. We then determined the extreme outlier cutoff to be the mean + 2 * SD. Table 2 shows the mean and SD of the native English speaker and Dutch-English bilingual groups, across the critical regions and three different measures employed. As mentioned in Section 2.3.2, due to the insertion of an extra noun in the control condition to ensure that these stimuli are not ambiguous anymore, region 5 of the target sentences is necessarily compared to region 6 of the control sentences. This is indicated in the table by use of the header 'Region 5-6', and with the Target reading time presented first, as target region 5 is compared to control region 6 (as these contain the same word). The same is naturally true for region 6 of the ambiguous sentence and region 7 of the unambiguous control, as these regions are also post-insertion of the additional NP.

	English native speakers (N = 12)				
First pass	Region 2	Region 3	Region 4	*Region 5-6*	Region 6-7
Target	318 (196)	259 (126)	358 (203)	308 (178)	322 (235)
Control	310 (185)	255 (142)	365 (214)	297 (180)	320 (222)
Regression path	Region 2	Region 3	Region 4	*Region 5-6*	Region 6-7
Target	431 (272)	333 (241)	472 (321)	756 (731)	2072 (1711)
Control	420 (268)	344 (264)	481 (355)	614 (773)	2049 (2327)
Total reading time	Region 2	Region 3	Region 4	*Region 5-6*	Region 6-7
Target	793 (569)	689 (435)	921 (590)	745 (519)	617 (562)
Control	756 (506)	656 (499)	921 (709)	574 (439)	514 (434)
	Dutch-English bilinguals (N = 20)				
First pass	Region 2	Region 3	Region 4	*Region 5-6*	Region 6-7
Target	297 (168)	246 (117)	332 (177)	314 (173)	324 (241)
Control	315 (200)	250 (126)	330 (186)	279 (139)	355 (259)
Regression path	Region 2	Region 3	Region 4	*Region 5-6*	Region 6-7
Target	368 (233)	361 (302)	434 (320)	675 (689)	1899 (1908)
Control	386 (276)	361 (288)	445 (359)	674 (766)	1872 (1697)
Total reading time	Region 2	Region 3	Region 4	*Region 5-6*	Region 6-7
Target	676 (468)	613 (485)	860 (597)	675 (454)	565 (452)
Control	629 (457)	551 (364)	824 (556)	563 (387)	546 (401)

Table 2. The mean fixation durations (in ms) of the three eye-tracking measures employed in our experiment, with the standard deviation in parentheses.

3.2.1. Data distribution and statistical considerations

Our eye-tracking data is almost universally right-skewed, as is evident from the distribution plots in Figure 3. However, this is not necessarily a bad thing, nor is it uncommon for eye-tracking data: this was already shown in an evaluation of reading time data for use in examining English sentence processing by Frank et al. (2013). Additionally, this finding and pattern of data distribution are supported by large eye-tracking corpora such as PROVO (Luke and Christianson (2018)), and even specifically in bilingual eye-tracking data as in the GECO corpus that even also employs Dutch-English bilinguals (Cop et al. (2017)). This right-skewed data pattern generally does also lead to a violation of the normality of residuals assumption of linear regression, which was also the case for our data. Therefore, we employed Box-Cox power transformations in order to obtain an as high as possible degree of normality so that regression analysis could be performed. Transformations were done through the EnvStats package (Millard (2013)) in R. We first ran a Box-Cox power analysis in

order to find the optimal lambda value, after which we transformed the relevant dependent variable accordingly in our regression models.

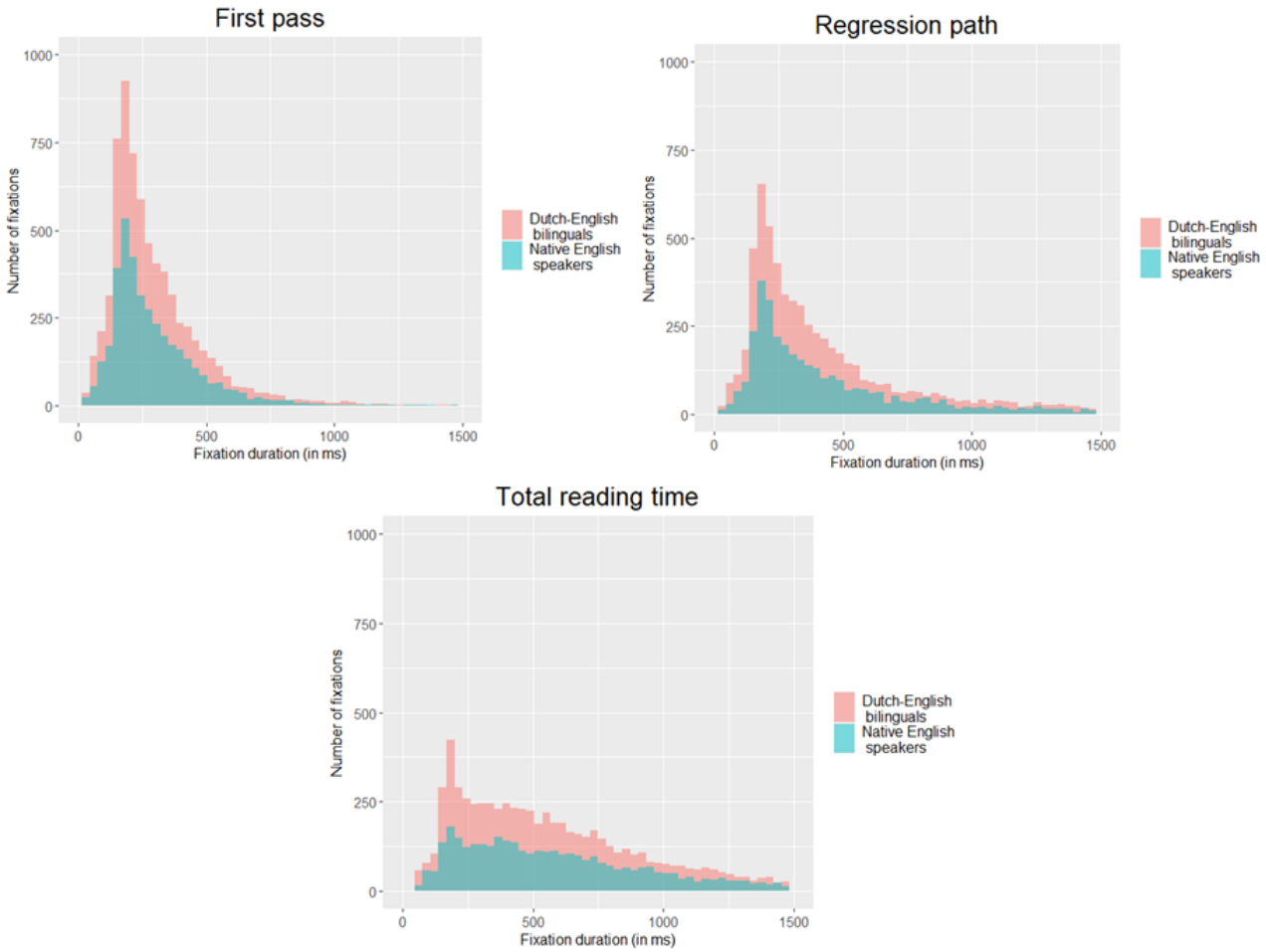

Figure 3. Histograms showing the distribution of our eye-tracking data, across the three measures we employ. Note that the total reading time does not apply to whole sentences, but to the total duration of fixations in a particular region

3.3. Eye-tracking measures regression models

For the eye-tracking mixed effects linear regression models we employed the theory of parsimonious models (Bates et al. (2015)) for constructing our initial models, as outlined in Section 2.6. After testing our null models with random slopes and intercepts for both item and participants, the model output showed that none of the random slopes were significantly contributing to model fit whilst making the models more complicated. Therefore, we excluded random slopes from our final model, but did keep random intercepts for item and participant. We also included several interaction terms between different predictors. Firstly, we included an interaction term for Language x

Type to assess if there were any group-based differences due to our experimental setup. Additionally, we included an interaction term for Language x Subcategorisation to assess if we can observe similar findings as Dussias and Scaltz (2008) did in their study, where bilinguals employed their L1 subcategorisation bias to supplement their L2 subcategorisation bias. Furthermore, an interaction between Cognitive control x Type was included to assess if the performance during the pre-test cognitive control battery was reflected in a higher ability to inhibit the temporary ambiguity in the ambiguous stimuli. Finally, we also included an interaction term for Cognitive control x Stimulus order, as we expected there could be an effect of attention, which is considered part of cognitive control, playing a part in potential fatigue or learnability effects.
For each of the eye-tracking measures below we built regression models per region, for regions 2 up to and including 7. The baseline maximal parsimonious models were used to assess normality of residuals, which were all either normal or near-normal in their distribution. Consequently, we used the buildmer package in R (Voeten and Voeten (2021)) to establish minimal models in which all terms were significant for analysis. We fed the entire maximal parsimonious model into the buildmer function, which then outputs a buildmer class object that contains the desired minimal model. This was done separately for each region within each eye-tracking measure, as we expected to see distinct effects across different regions and different measures. For example, the first verb in the sentence (Region 3) is not yet subject to a temporary ambiguity by the time of the First pass, but may be re-read after encountering the second verb in Region 5 in the ambiguous condition as the participant had to reassess whether the noun in Region 4 is the object of verb 1 or the subject of verb 2. Hence, we do not expect Type to have much of an effect in Region 2 in the First pass measure, but it may turn out to be significant in the Regression path measure. Post-hoc analysis was performed in R, in which we employed the 'emmeans' package (Lenth (2022)) for analysis of significant main effects and the 'phia' package (De Rosario-Martinez (2015)) for analysis of significant interaction terms.

3.3.1. First pass

The significant terms of the final models of the First pass measure are shown in Table 3. Post-hoc analyses of the significant terms of Region 2 revealed that while the Language x Type interaction term is significant in the model, but the groups within the interaction are not. The Cognitive control x Stimulus order interaction term does have a significant post-hoc result in which the Low cognitive control group has higher reading times for Late stimuli at Chi-square = 6.6590, df = 1, $p = 0.0296$. In the model for Region 3, there was a significant effect of Stimulus order where estimated marginal mean of the Early stimuli (M = 14.9, SE = 0.207) is lower than that of the Late stimuli (M = 15.3, SE = 0.207) with a Beta estimate of -0.359 (SE = 0.14), t-value = -2.560 at $p = 0.0105$. Region 4 showed an effect of Subcategorisation, which in the post-hoc analysis showed an effect between the Less probable category (M = 18.4, SE = 0.408) being significantly higher than the Improbable category (M = 17.2, SE = 0.385) with a Beta estimate of 1.19 (SE = 0.304), t-value = 3.901 at $p < 0.001$ and the Similar category (M = 16.6, SE = 0.394) with a Beta estimate of 1.77, t-value = 5.585 at $p < 0.001$. Post-hoc analysis of the main effect of Type in Region 5-6 showed that the Control sentences had a significantly lower mean (M = 16.9, SE = 0.239) than the Ambiguous sentences (M = 17.5, SE = 0.238) with a Beta estimate of -0.6, t-value = -2.724 at $p = 0.0085$. Finally, the interaction term in Region 6-7 was also not significant in the post-hoc analysis.

Region 2 (Charles)	Chi-square	*p*-value
Language x Type	6.0211	0.014136
Cognitive control x Stimulus order	9.5001	0.008651
Region 3 (shaved)		
Stimulus order	6.567	0.01039
Region 4 (the sheep)		
Subcategorisation	31.7447	<0.001
Region 5-6 (stopped)		
Type	7.4233	0.006438
Region 6-7 (grazing)		
Language x Type	7.0033	0.008136

Table 3. Significant model parameters for the different regions in the First Pass measure. The base model parameters fed into buildmer() were as follows: First Pass measure ((Language * Type + Subcategorisation) + (Cognitive control * Type + Stimulus order)) + (1|item) + (1|participant).

3.3.2. Regression path

The significant terms of the final models of the Regression path measure are shown in Table 4. The interaction term in Region 2 for Cognitive control x Stimulus order did not turn out to be significant in the post-hoc analysis. The same is true for the interaction term in Region 3. Subcategorisation in Region 4 was again significant, with a similar pattern to the one observed in the First pass measure. The Less probable category (M = 8.21, SE = 0.1002) was significantly higher than Improbable category (M = 7.96, SE = 0.0918) with a Beta estimate of 0.248 (SE = 0.0901), t-value = 2.749 at $p = 0.0238$ and Similar category (M = 7.77, SE = 0.0954) with a Beta estimate of 0.439 (SE = 0.0937) at $p < 0.001$. In Region 5-6, the main effect of Type was significant in the post-hoc test, with the estimated marginal means of the Control sentences (M = 3.08, SE = 0.0122) being lower than that of the Ambiguous sentences (M = 3.11, SE = 0.0121) with a Beta estimate of -0.0312, t-value of -2.344 at $p = 0.0224$. However, as this main effect is present in a significant interaction term, more weight should be assigned to the interaction. Post-hoc analysis of this interaction did also show a significant effect of Language across Type, with the native English group having a significantly higher reading time in the Target sentences at Chi-square = 8.2144, df = 1 at $p = 0.008312$, which is also visualised in Figure 4. The final region under consideration has a main effect for Stimulus order, in which Early stimuli (M = 20.7, SE = 0.524) have a significantly higher duration than the Late stimuli (M = 19.8, SE = 0.525) with a Beta estimate of 0.865 (SE = 0.21), t-value = 4.110 at $p < 0.001$. For the main effect of Subcategorisation, the Improbable category (M = 21.3, SE = 0.591) is significantly higher than both the Less probable category (M = 19.7, SE = 0.654) with a Beta estimate of 1.55 (SE = 0.612), t-value = 2.536 at $p = 0.0413$ and the Similar category (M = 19.8, SE = 0.621) with a Beta estimate of 1.50 (SE = 0.576), t-value = 2.604 at $p = 0.0346$. However, the latter main effect is also included in a significant interaction, which

Region 2 (Charles)	Chi-square	*p*-value
Cognitive control x Stimulus order	8.2363	0.01605
Region 3 (shaved)		
Cognitive control x Type	7.7046	0.02123
Region 4 (the sheep)		
Subcategorisation	21.966	<0.001
Region 5-6 (stopped)		
Type	4.3288	0.03747
Language x Type	3.9866	0.04586
Region 6-7 (grazing)		
Stimulus order	20.9241	0.008136
Subcategorisation	7.9092	0.01917
Language x Subcategorisation	6.6982	0.03512

Table 4. Significant model parameters for the different regions in the Regression path measure. The base model parameters fed into buildmer() were as follows: Regression Path measure ((Language * Type + Subcategorisation) + (Cognitive control * Type + Stimulus order)) + (1|item) + (1|participant).

may make it misleading. Post-hoc analysis of the interaction term of Language x Subcategorisation showed that the Native English language group had significant differences across Subcategorisation at Chi-square = 13.0897, df = 2, $p = 0.002875$. The Improbable category being the significantly different Subcategorisation as shown in Figure 5.

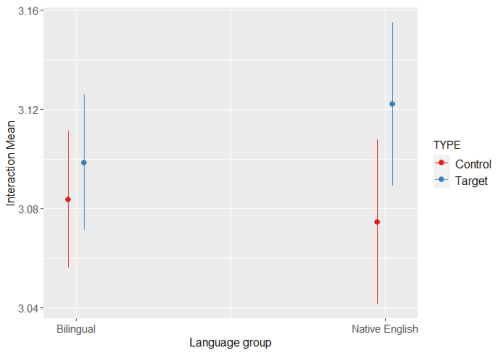

Figure 4. The significant interaction of Language x Type in Region 5-6 of the Regression path measure.

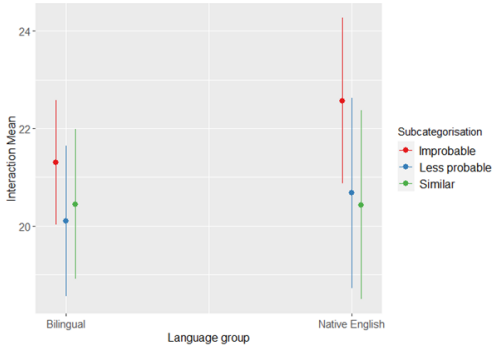

Figure 5. The significant interaction of Language x Subcategorisation in Region 6-7 of the Regression path measure.

3.3.3. Total reading time

Region 2 (Charles)	Chi-square	*p*-value
Language	5.6177	0.017780
Stimulus order	7.5075	0.006144
Cognitive control x Stimulus order	6.0290	0.049071
Region 3 (shaved)		
Type	5.4188	0.01992
Region 4 (the sheep)		
Subcategorisation	14.189	<0.001
Region 5-6 (stopped)		
Type	17.562	<0.001
Stimulus order	12.774	<0.001
Region 6-7 (grazing)		
Stimulus order	16.6704	<0.001
Language x Type	7.8542	0.005070
Cognitive control x Type	9.6663	0.007961
Cognitive control x Stimulus order	11.1376	0.003815

Table 5. Significant model parameters for the different regions in the Total Reading Time measure. The base model parameters fed into buildmer() were as follows: Total Reading Time measure ((Language * Type + Subcategorisation) + (Cognitive control * Type + Stimulus order)) + (1|item) + (1|participant).

The significant terms of the final models of the Total reading time measure are summarised in Table 5. In Region 2, the main effect of Language is significant, and the post-hoc shows that the Dutch-English bilingual group (M = 23.1, SE = 0.748) has significantly lower reading times than the native English group (M = 25.9, SE = 1.012) with a Beta estimate of -2.82 (SE = 1.19), t-value = -2.370 at $p = 0.0249$. Whilst the main effect for Stimulus order is also included in an interaction term in this region, a post-hoc was done for the sake of completion, which revealed that the Early stimuli (M = 24.9, SE = 0.678) had a significantly higher estimated mean reading time than the Late stimuli (M = 24.2, SE = 0.679) with a Beta estimate of 0.702 (SE = 0.295), t-value = 2.380 at $p = 0.0174$. The post-hoc analysis of the interaction effect in Region 2 revealed that the Mid cognitive control group was the only one with significant differences across Stimulus order at Chi-square = 10.3831, df = 1, $p = 0.003815$. This interaction effect is shown in Figure 6. Region 3 only has a significant main effect of Type, and post-hoc analysis reveals that the Control condition (M = 22.0, SE = 0.597) has a significantly lower estimated marginal mean as compared to the Ambiguous condition (M = 23.3, SE = 0.596) with a Beta estimate of -1.3 (SE = 0.557), t-value =

-2.328 at $p = 0.0234$.
In Region 4 we again see Subcategorisation as a retained main effect, with post-hoc analysis showing that, once again, it is the Less probable category (M = 28.9, SE = 0.849) that is significantly higher than both the Improbable category (M = 27.0, SE = 0.783) with a Beta estimate of 1.908 (SE = 0.737), t-value = 2.590 at $p = 0.0361$ and the Similar category (M = 26.0, SE = 0.811) with a Beta estimate of 2.857 (SE = 0.766), t-value = 3.728 at $p = 0.0013$. Region 5-6 has two significant main effects, and the post-hoc analysis of the first of these reveals that the estimated marginal mean of the Control sentences (M = 17.1, SE = 0.392) is significantly lower than that of the Target sentences (M = 18.9, SE = 0.390) with a Beta estimate of -1.79 (SE = 0.428), t-value = -4.191 at $p < 0.001$. Stimulus order was also significant in the post-hoc analysis, with the Early stimuli (M = 18.3, ES = 0.337) being higher than the Late stimuli (M = 17.7, SE = 0.337) with a Beta estimate of 0.582, t-value = 3.573 at $p < 0.001$.

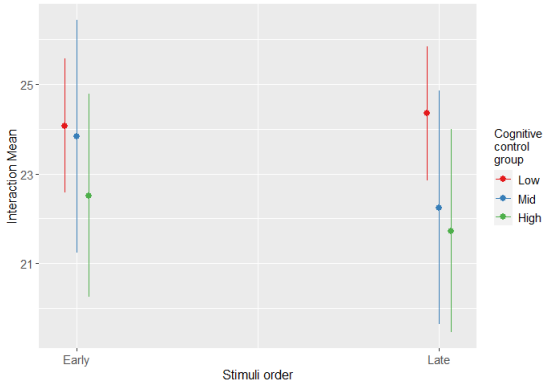

Figure 6. The significant interaction between Cognitive control and Stimulus order in Region 2 of the regression model for Total reading time.

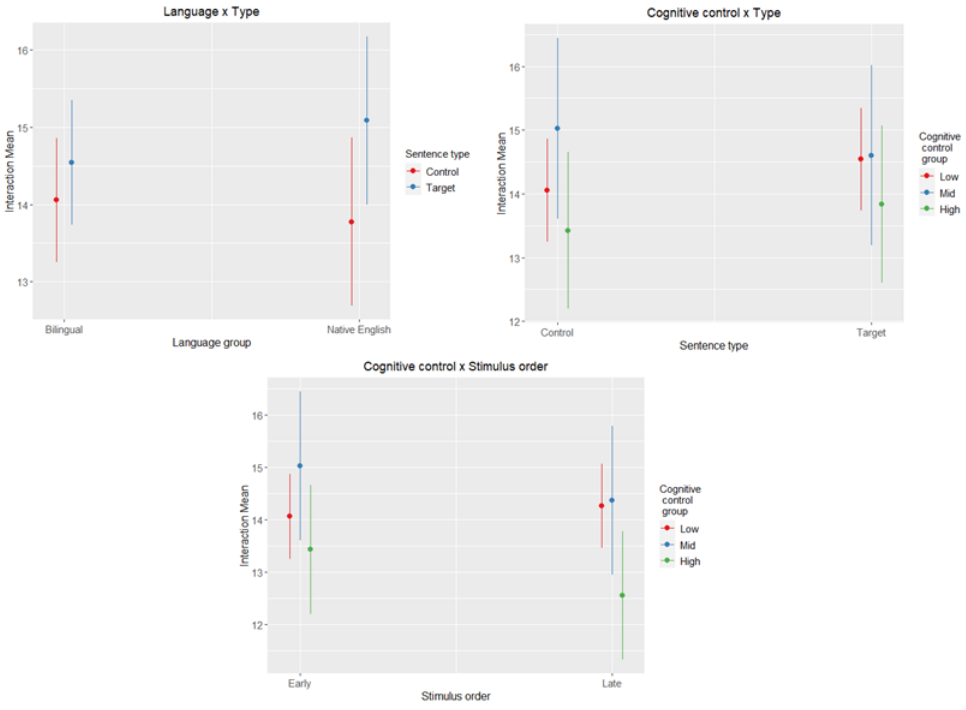

Figure 7. The significant interaction terms in the Total reading time measure for Region 6-7.

The final region has only one included main effect, which is also included in an interaction term, so should again be considered with caution. Nonetheless, the post-hoc analysis of the Stimulus order main effect showed that the marginal estimated mean of the Early stimuli (M = 14.3, SE = 0.333) was significantly higher than that of the Late stimuli (M = 13.9, SE = 0.333) with a Beta estimate of 0.444 (SE = 0.132), t-value = 3.367 at $p < 0.001$. The first interaction term, that of Language x Type, shows a post-hoc significant effect in the Native English language group with Chi-square = 8.2778, df = 1 at $p = 0.008026$. There were also significant effects within the interaction term for Cognitive control x Type, for both the Low group with Chi-square = 5.4165, df = 1 at $p = 0.03989$ and the High group with Chi-square = 6.5738, df = 1 at $p = 0.03105$. The final interaction term, that of Cognitive control x Stimulus order, had significant effects for the Mid group with Chi-square = 9.1382, df = 1 at $p = 0.005007$ and the High group with Chi-square = 18.0419, df = 1 at $p < 0.001$. These three interactions are visualised in Figure 7.

3.4. Comprehension accuracy

Similarly to the eye-tracking measures, some extreme outlier data points were also excluded from the analysis of accuracy. This was not based on the actual outcome measure (as it is hard to establish outliers on a binary variable), however, but rather on the reaction time of participants. Even though a longer reaction time on the comprehension question does not necessarily reflect on reliability, employing this outlier condition would allow for filtering out technical difficulties and make overall comparison more fair if some participants went with first impressions, whereas others may have deliberated for a longer amount of time, as the instructions were only explicit on attentively reading the sentences, rather than also the questions. We established thresholds based on the pilot study participants responses; the upper limit being based on the data where we asked pilot participants to carefully read and answer the questions, whereas for the lower limit we asked the pilot participants to answer as quickly as they could whilst still actually reading the question. This led to exclusion of all answers given with a reaction time of > 10000ms, as well as those that with a reaction time of < 800ms. In total, this meant we left approximately 3% of the accuracy data out of consideration.

3.4.1. Distribution and comparison of means

The overall accuracy of the participant groups was very high despite the complex nature of the sentences presented during the experiment, as can be seen in Table 6. As the accuracy scores were binary, we employed proportion tests to examine group-level differences, but none were significant. Within-group comparison of ambiguous versus control sentences did yield significant effects when compared across main clause and embedded clause accuracy. The Dutch-English bilinguals main clause comparison was significant at Chi-square = 20.42, df = 1, $p < 0.001$ and the embedded clause comparison was significant at Chi-square = 3.89, df = 1, $p = 0.0487$. For the native English speakers the main clause comparison was significant at Chi-square = 9.45, df = 1, $p = 0.002109$ and the embedded clause comparison was significant at Chi-square = 6.14, df = 1, $p = 0.01324$. We also examined main clause versus embedded clause accuracy across ambiguous and control conditions, which similarly yielded significant effects. The Dutch-English bilinguals ambiguous comparison was significant at Chi-square = 16.77, df = 1, $p < 0.001$ and the control comparison was significant at Chi-square = 6.26, df = 1, $p = 0.01236$. The native English speakers ambiguous comparison was significant at Chi-square = 8.00, df = 1, $p = 0.004686$ and the control comparison was significant at Chi-square = 7.58, df = 1, $p = 0.005901$.

	Native English speakers (N = 12)		Dutch-English bilingual speakers (N = 20)	
	Main clause	Embedded clause	Main clause	Embedded clause
Target	0.972 (0.166)	0.880 (0.326)	0.992 (0.0893)	0.910 (0.286)
Control	0.866 (0.342)	0.956 (0.205)	0.894 (0.308)	0.953 (0.212)

Table 6. The mean accuracy scores on the comprehension questions captured during the experiment, with the standard deviation in parentheses.

	Native English speakers (N = 12)		Dutch-English bilingual speakers (N = 20)	
	Main clause	Embedded clause	Main clause	Embedded clause
Target	1956 (1111)	2290 (1543)	1754 (787)	2071 (1333)
Control	2496 (1404)	1864 (1201)	2209 (1241)	1747 (873)

Table 7. The mean response time (in ms) to the comprehension questions captured during the experiment, with the standard deviation in parentheses.

Aside from accuracy we also examined reaction times on the comprehension questions, the mean results of which can be observed in Table 7. The reaction time data was also non-normally distributed, and seeing as this pertained to linear data we employed Mann-Whitney U-tests instead of proportion tests to ascertain whether there were significant between and within group differences. Similarly to the accuracy measure, there were no significant group-level differences between the Dutch-English bilinguals and the native English speakers. However, the within group analysis for the ambiguous versus control condition did yield significant results. The Dutch-English bilinguals main clause comparison was significant at $W = 22867$, $p < 0.001$ and the embedded clause comparison was significant at $W = 56048$, $p = 0.006065$. For the native English speakers the main clause comparison was significant at $W = 7847$, $p < 0.001$ and the embedded clause comparison was significant at $W = 19475$, $p = 0.006995$. Additionally, we examined main clause versus embedded clause reaction times across ambiguous and control conditions, which showed significant effects, but not across all conditions. The Dutch-English bilinguals ambiguous comparison was not significant, but the control comparison was significant at $W = 50706$, $p < 0.001$. The native English speakers ambiguous comparison was also not significant, whereas the control comparison was significant at $W = 17991$, $p < 0.001$.

3.4.2. Logistic regression model

Similarly to the eye-tracking measures, a regression model was also used to analyse accuracy further. Seeing as the accuracy variable was binary, we employed logistic regression instead of linear regression. Once again, we started with the baseline of a parsimonious maximal model (Bates et al. (2015)) as outlined section 2.6. Our first null model therefore included random intercepts for both item and participants and a random slope for item, but this model failed to converge. After further examination it turned out that random effects for item were causing this issue, hence these were left out, after which the model did converge. As with the eye-tracking models, we also included several interaction terms between different predictors for accuracy.

Most of these interaction terms were the same, and included with the same reasoning, as those for the eye-tracking measures. However, two novel interaction terms were also included based on Question type (i.e., whether the question targeted the content of the main clause or the embedded clause). The first of these is Type x Question type, which allows insights into the good-enough parsing aspects of our experiment, as Ferreira and Lowder (2016) proposed that main clause accuracy should be higher in garden-path sentences with an embedded matrix subclause. The second interaction term is Language x Question type, which instead examines a potential group-based difference in treatment of the different question types, which covers the potential for a bilingual (dis)advantage effect as well as specific Dutch-English language interaction effects.

	Chi-square	*p*-value
Stimulus order	8.3001	0.003964
Cognitive control	7.7597	0.020654
Subcategorisation	8.5115	0.014182
Question type x Type	36.5590	<0.001

Table 8. Significant model parameters for Accuracy. The base model parameters fed into buildmer() were as follows: Accuracy ((Language * Type + Subcategorisation + Question type) + (Cognitive control * Type + Stimulus order) + Question type * Type) + (1|item) + (1|participant).

Consequently the same procedure as that of the eye-tracking measures was used, in which we used the buildmer package in R (Voeten and Voeten (2021)) to establish minimal models in which all terms were significant for analysis. We fed the entire maximal parsimonious model into the buildmer function, which outputs a buildmer class object that contains the desired minimal model. After examining the final model we performed post-hoc analysis with the 'emmeans' package (Lenth (2022)) for analysis of significant main effects and the 'phia' package (De Rosario-Martinez (2015)) for analysis of significant interaction terms.

As can be seen in Table 8, the logistic regression model confirms that there were no group-level differences in accuracy, as already found in the mean and proportion comparisons above. Additionally, we find a similar learning effect as was found in some of the eye-tracking measures in the significant main effect of Stimulus order: the estimated marginal means of Early (M = 2.86, SE = 0.189) versus Late (M = 3.43, SE = 0.220) stimuli were significantly different in the post-hoc test with a Beta estimate of -0.567 (SE = 0.197), z-value = -2.881 at p = 0.0040. Similarly, Subcategorisation, which was also shown to be significant across eye-tracking measures, had a significant post-hoc difference between the Improbable (M = 2.82, SE = 0.199) and Similar (M = 3.51, SE = 0.250) categories with a Beta estimate of -0.695 (SE = 0.239), z-value = -2.903 at p = 0.0111. Cognitive control yielded an unexpected result in the post-hoc test as there was a significant difference with higher scores for the Low (M = 3.61, SE = 0.344) than the High (M = 2.66, SE = 0.209) group with a Beta estimate of 0.947 (SE = 0.369), z-value = 2.568 at p = 0.0304. The interaction

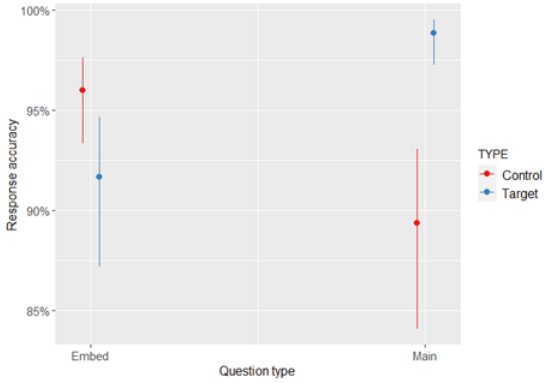

Figure 8. The significant interaction from the logistic regression model of Question type across Type (Ambiguous/Control).

of Question type x Type is also significant in the post-hoc analysis, both for the Embedded and Main clause question types (across the ambiguous/control condition). For the Embedded clause the interaction is significant at Chi-square = 11.800, df = 1, $p < 0.001$ and for the Main clause it is significant at Chi-square = 24.804, df = 1, $p < 0.001$. The direction of the effect can be seen in Figure 8.

4. Discussion

This study set out to investigate Dutch-English bilingual processing of English garden-path sentences, with a native English control group for comparison. We aimed to show that a bilingual advantage need not necessarily be present, but that specific language interaction pairings could lead to bilingual differences, surfacing as advantageous in some behavioural tasks. To this end we expected our Dutch-English bilingual participants to perform largely similarly to the native English speakers in the eye-tracking reading measures overall, but with potentially different reading patterns. Specifically, we expected that parallel activation, as advocated in Dijkstra et al. (2019), of typologically close languages might provide support in order to reduce the fixation duration in the regression path measure, which is often the most strongly affected by the temporary ambiguity in a garden-path sentence. In a similar vein, we expected the bilinguals to outperform the native English speakers in the accuracy of the accompanying comprehension questions due to the grammatical structure of Dutch not allowing for an ambiguity in the translation equivalents of our stimuli, which therefore allows Dutch to assist in the parsing of an English garden-path sentence. Regardless of these group-level differences, we expected all participants to perform better on comprehension questions regarding the second part of our stimuli sentences (the main clause). This expectation came to be in light of Ferreira and Lowder (2016)'s theory of good enough parsing with prediction and information structure incorporated, which stipulates that later information is more likely to be considered new information and therefore less likely to be subject to lingering misinterpretations. In our study, this means an expected lower performance on accuracy for questions targeting the embedded clause ('While John shaved...'), with a higher performance on questions targeting the main clause ('...the sheep stopped grazing.'). Additionally, the eye-tracking reading measures would be expected to reflect the increased processing power reserved for the latter part of a sentence by means of longer fixation durations in the later regions of the sentences, especially for the regression path measure as this concerns re-reading (which would be more likely to happen at information that is perceived to be important according to information structure principles).

4.1. General discussion

Our results with regards to the eye-tracking measures largely confirmed our second hypothesis that high-proficiency Dutch-English bilinguals would perform similarly to native English speakers, as they had very similar distributions of data (see Figure 3) and average reading times (see Table 2). Additionally, the regression models show that the predictor for Language is not significant in most of the regions, although it did show up in significant interactions in primarily the latter regions and, more relevantly, the regression path and total reading time measures, which are both important for measuring re-reading as triggered by encountering the unexpected verb in Region 5 of the ambiguous stimuli. As expected, we observed a robust effect of Type in Region 5-6, which is the disambiguating region in the Ambiguous sentences, across all eye-tracking measures. Post-hoc analyses of this main effect similarly yielded the expected results that the Control sentences universally had a lower fixation duration than the Ambiguous sentences, showing a successful replication

of the garden-path effect. Furthermore, Region 4 also contained a consistently significant main effect of Subcategorisation throughout all of the eye-tracking measures employed. We expected this to potentially play a role in the regions following the first (ambiguous) verb, as its subcategorisation bias could have influenced how expected certain parts of speech were to follow it. Interestingly, it was not the Improbable or Similar category that stood out, but rather the Less probable category that had increased fixation durations as compared to the other two categories. It is also unexpected that the main effect occurs regardless of condition, as we would have expected any effect of subcategorisation to mainly affect the ambiguous sentences. We also do not observe an interaction between Language and Subcategorisation for Region 4, which suggests that as opposed to the Spanish-English bilinguals in Dussias and Scaltz (2008), our Dutch-English bilinguals do not appear to employ their L1 subcategorisation bias to help resolve the temporary ambiguity. However, it must be noted that our Native English speakers were not uniformly monolingual, so some of the Native English speakers could also have used their L2 subcategorisation bias to help resolve ambiguities, hence this effect could have been obscured between the groups.

We had hoped to observe Language x Type interactions, in order to assess if there were systematically different reading patterns between our high-proficiency Dutch-English bilinguals and the native English speaker group with regards to the processing of English garden-path sentences. Our results show that these occur sporadically, as well as differently across varying eye-tracking measures. In the First pass measure the Language x Type interaction was present in Region 2 and Region 6-7. However, neither of the post-hoc analyses came up significant, so whilst this interaction significantly contributed to model fit, it did not yield further insights into differing reading patterns. This is not too surprising, as the First pass measure only concerns the very first time reading the particular regions, and usually garden-path effects surface through increased regressions to the temporarily ambiguous sentence parts. For this reason, we would expect to find Language x Type interactions in the Regression path measure, especially if we assume that Dutch-English bilinguals process English garden-path sentences differently. On the one hand, we only observe the Language x Type interaction in a single region, namely Region 5-6. On the other, it is the critical region with the disambiguating verb, and therefore does make it the most interesting region to observe this interaction effect in. The post-hoc analysis of this interaction term in the Regression path revealed that the Dutch-English bilinguals had relatively similar reading times regardless of condition, whereas the native English speakers had a significantly longer duration of the regression path measure in the ambiguous sentences. Similarly, for Total reading time there is also only a single instance of an interaction effect of Language x Type, this time in the spillover region of Region 6-7. The post-hoc also reveals the same interaction: only the native English speakers had significantly higher reading times for the Target condition.

This observation, together with the confirmation of our third hypothesis that there were no group-level differences in performance during the cognitive control battery at the pre-test, could suggest that the Dutch-English bilinguals do have a limited advantage in the parsing of English garden-path sentences. It is especially pertinent for the Regression path measure, the measure in which it was most likely to observe garden-path effects, and where it additionally was visible specifically in the critical disambiguating region, which is expected to be the most conflict-inducing region. This could be suggestive of dual processing in both language systems as proposed in Dijkstra et al. (2019), which surfaces as a language-specific interaction as the Dutch translation equivalent of the garden-path sentence in Table 1 ('*Nadat Charles* (ZICH) *scheerde het schaap stopte met grazen**') is not ambiguous: Dutch requires inversion of the direct object and the verb in a subclause to be grammatical. The sentence structure remains somewhat dubious in the translation provided, as the

main clause here would also require verb inversion ('*Nadat Charles* (ZICH) *scheerde* STOPTE HET SCHAAP *met grazen*') due to sentence as a whole starting with an (embedded) subclause. Nonetheless, this means that the sentence should never be perceived as Charles shaving the sheep with the English word order, so that the only possible direct object can be the (omittable) reflexive pronoun 'ZICH' (himself). This benefit would naturally not be present in the monolingual native English speakers, as they only have access to a single language system, nor is it available for those native English speakers that were bilinguals, as we confirmed that the translation equivalents of the garden-path sentences were still ambiguous in the respective other native tongues with the relevant participants after their eye-tracking sessions.

We expected that if this effect was observed in the Regression path, it would also be observable in Total reading time, which was not the case for the critical region. However, in this measure we did find the same effect in the spillover region after the critical region. Similarly to the effect in the Regression path, this can be explained by virtue of a specific Dutch-English language interaction, in combination with the expected processing power distribution from Ferreira and Lowder (2016): good-enough parsing with prediction assumes more processing power is allocated to the latter part of the sentence, as it is considered new information, and Dutch being simultaneously activated with English would mean a lesser processing cost as it provides a non-ambiguous underlying representation. As this supporting language system is not available to the native English group, it is more likely that they spend significantly longer on the target ambiguous sentences in the latter part, as more processing power is allotted and they do not have a supporting language system to help alleviate the processing cost.

With regards to accuracy, on the other hand, the same effect does not appear to be present. Whilst the Dutch-English bilingual group does have higher target sentence accuracy than the native English group as shown in Table 6, statistical comparison revealed that this was not significant. Our logistic regression model also showed neither a main effect for Language, nor an interaction for Language x Type. Therefore, our first hypothesis, in which we expected the Dutch-English language interaction to have a positive effect on comprehension accuracy (Section 1.8), has been refuted. However, overall accuracy was extremely high, which could have led to obscuring of any potential group-based differences. Additionally, we did manage to provide more evidence for the good-enough parsing model proposed in Ferreira and Lowder (2016). Not only did we see that the main clause questions in the ambiguous condition had consistently, and significantly, higher mean accuracy scores than those of the embedded clause questions (Table 6), we also see this effect replicated in the logistic regression with a significant interaction term for Question type x Type.

Post-hoc analysis of this interaction effect confirmed that Embedded clause question accuracy was significantly higher in the control condition as compared to the ambiguous condition. This falls perfectly in line with the updated good-enough parsing theory, as lower accuracy is observed with regards to the first part of the garden-path sentence, which is allocated lesser processing power on the basis of information structure leaving more chance at lingering misinterpretations. Conversely, questions related to the latter part of the sentence have near-perfect accuracy in both groups in the ambiguous condition, as this part does have higher processing power available and is thus not subject to good-enough parsing. The interaction term also revealed that the opposite effect was true for the control condition, which is rather unexpected. We assume this might have something to do with memory, as this was also not extensively assessed in our short cognitive control battery. It is conceivable that participants might not remember which of the two noun phrases in close proximity to one another in Region 4 and 5 of the control condition was which, due to the sentences being relatively long and complex. This is different from a lingering misinterpretation, as in this

case there would be no syntactic conflicts with regards to thematic role assignment, only a potential issue with recalling which noun phrase had which thematic role assigned to it.
Aside from these main results, some other interesting effects were also revealed by virtue of the regression models employed. The main effect of Stimulus order is relatively robustly present in all eye-tracking measures, most visibly so in the Total reading time. All but once, which is the one time Stimulus order is significant in the First pass measure, we see that there appears to be a learnability effect in the stimuli: the Late stimuli, representing the latter half of the experimental stimuli, generally have significantly lower reading times than the Early stimuli. An interaction with Cognitive control and Stimulus order is also often present, in fact, it is present in Region 2 across all measures, as well as in the spillover region in Total reading time. Post-hoc analyses of these interaction terms did not reveal further significance for the First pass measure.
On the other hand, in both the Regression path and the Total reading time, which are more likely to be influenced by cognitive control based on previous studies (e.g., Brothers et al. (2021)), the post-hoc analyses were significant. These showed that either the Mid-level performance, or both the Mid- and the High-level performance groups of the cognitive control tests performed significantly better in Late stimuli. This would suggest that cognitive control abilities play a part in whether or not the participants learn to recognise the crucial information in garden-path sentences leading to more efficient inhibition and therefore faster reading times. Alternatively, cognitive control could perhaps guide the attention mechanisms to learn how to process these sentences more efficiently the longer the experiment went on, especially as Region 2 and Region 6-7 are not very relevant for the temporary ambiguity of the garden-path sentences. Contrary to this seemingly beneficial effect of higher cognitive control abilities, the post-hoc of the main effect of Cognitive control in the accuracy logistic regression model suggests a negative effect of cognitive control: the Low group performed significantly better than the High group in comprehension accuracy. This result is unexpected, especially as in the eye-tracking measures Cognitive control did seem to have the expected effect. We propose that the adverse effect observed in the accuracy measure is therefore likely more related to a cognitive process that we did not test for in our short cognitive control battery, than it is related to inhibitory control, which was our main focus in the battery as well as of primary importance in garden-path processing.

4.2. Limitations

One major limitation of the current study was the difficulty of obtaining native English participants, especially truly monolingual ones, which has led to unequal groups in our experiment. Whilst balanced groups are not absolutely necessary for regression analysis, especially with the large number of stimuli employed leading to plenty of observations in each group regardless of the imbalance between the groups, a balanced experiment does usually yield higher statistical power. Additionally, our body of participants was not a representative sample of society as a whole due to the university environment the experiment was conducted in, which led to access to mainly students in our recruitment efforts. This has also been advantageous, as it did mean that there was relatively little within-group variation in social background and education, as well as ensuring a high proficiency in English for our bilingual group. However, it also means that our results might not be representative of the Dutch-English population as a whole. Finally, as this experiment was part of a Master's graduation project it was naturally not fully funded, which led to inclusion of only a small cognitive control battery in the pre-test. Seeing as robust effects of cognitive control as influencing garden-path processing have been found in, for example Novick et al. (2014), having a more thorough battery of cognitive control tests would have been preferable. It was, however,

not deemed feasible to include a standardised cognitive control battery for the purpose of this experiment, as those can span hours (e.g., the total time for a full session in Brothers et al. (2021) was about three hours) which were simply not available both time- and funding-wise.

4.3. Future research

In future, we propose that it would be beneficial for studies in bilingualism to clearly define their type of bilinguals. This naturally includes the specific language pairing involved, but also proficiency levels of both languages, as well as any potential confounding languages that may be present (e.g., third or more languages, especially if these are more dominant than the languages involved in the language pairing under investigation). This recommendation is in line with, and stems from, de Bruin et al. (2021) and Marian and Hayakawa (2021), which similarly advocate a higher degree of standardisation to ensure valid comparison between different studies on bilingualism. Furthermore, as discussed in Section 1.7, eye-tracking provides us with physiological data which can be very interesting when examining processing, but it does not provide insights into neuro-cognitive processes beyond being able to visualise difficulties in processing through, for example, increased fixation times or pupil dilation. EEG data, on the other hand, has proven to be insightful by allowing the collection of neuro-physiological data such as in Ye and Zhou (2008), but comes with the downside that it usually does not employ natural reading paradigms. A combination of eye-tracking and EEG, combining the best of both worlds, could therefore yield incredibly interesting insights into sentence processing. The use of this combination has already shown much promise in bilingual research, for example in Antúnez et al. (2021) which managed to provide further evidence for simultaneous bilingual semantic processing, and we expect that this could similarly be invaluable for further examination of bilingual processing of garden-path sentences.

5. Conclusion

Our study showed that highly proficient second-language learners of English can be considered functional bilinguals, as they performed similarly to native English speakers whilst reading difficult-to-parse garden-path sentences. Furthermore, we provided evidence for recent models of good-enough parsing that include information structure and prediction, as our accuracy results did show the expected significant disadvantage for embedded clause questions as compared to main clause questions. Finally, we showed how there is not necessarily a bilingual advantage either in general cognitive control ability, albeit through a limited test battery, nor in the processing of garden-path sentences. However, we did find a specific Language interaction effect with Type with our Dutch-English bilinguals, which was advantageous in processing critical disambiguating regions of the garden-paths. Crucially, this effect was only found in our Dutch-English bilinguals, and not also in the native English group, which did also contain some bilinguals but of different - non-beneficial for these types of garden-path sentences - language pairings. We hope this finding inspires future bilingual processing studies to carefully consider and describe the language pairings used, and what potential language-specific interactions could arise thereof.

Bibliography

Altmann, G.T., Kamide, Y., 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. Cognition 73, 247–264.

Antúnez, M., Mancini, S., Hernández-Cabrera, J., Hoversten, L., Barber, H., Carreiras, M., 2021. Cross-linguistic semantic preview benefit in basque-spanish bilingual readers: Evidence from fixation-related potentials. Brain and Language 214, 104905.

Baroni, M., 2020. Linguistic generalization and compositionality in modern artificial neural networks. Philosophical Transactions of the Royal Society B 375, 20190307.

Barr, D.J., Levy, R., Scheepers, C., Tily, H.J., 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of memory and language 68, 255–278.

Bates, D., Kliegl, R., Vasishth, S., Baayen, H., 2015. Parsimonious mixed models. arXiv preprint arXiv:1506.04967 .

Bialystok, E., Craik, F.I., Freedman, M., 2007. Bilingualism as a protection against the onset of symptoms of dementia. Neuropsychologia 45, 459–464.

Bialystok, E., et al., 2001. Bilingualism in development: Language, literacy, and cognition. Cambridge University Press.

Braver, T.S., 2012. The variable nature of cognitive control: a dual mechanisms framework. Trends in cognitive sciences 16, 106–113.

Brothers, T., Hoversten, L.J., Traxler, M.J., 2021. Bilinguals on the garden-path: Individual differences in syntactic ambiguity resolution. Bilingualism: Language and Cognition 24, 612–627.

de Bruin, A., Dick, A.S., Carreiras, M., 2021. Clear theories are needed to interpret differences: Perspectives on the bilingual advantage debate. Neurobiology of Language 2, 433–451.

Bylund, E., Hyltenstam, K., Abrahamsson, N., 2021. Age of acquisition–not bilingualism–is the primary determinant of less than nativelike l2 ultimate attainment. Bilingualism: Language and Cognition 24, 18–30.

Christianson, K., Luke, S.G., Hussey, E.K., Wochna, K.L., 2017. Why reread? evidence from garden-path and local coherence structures. The Quarterly Journal of Experimental Psychology 70, 1380–1405.

Clifton, C., Staub, A., Rayner, K., 2007. Chapter 15 - eye movements in reading words and sentences, in: Eye Movements: A Window on Mind and Brain. Elsevier Ltd, pp. 341–371.

Clifton Jr., C., Staub, A., 2011. Syntactic influences on eye movements during reading, in: Simon, L., Iain, G., Stefan, E. (Eds.), The Oxford Handbook of Eye Movements.. OUP Oxford. [Oxford Library of Psychology], pp. 895–909. URL: `https://login.ezproxy.leidenuniv.nl:2443/login?URL=https://search-ebscohost-com.ezproxy.leidenuniv.nl/login.aspx?direct=true&db=e000xww&AN=467510&site=ehost-live`.

Cop, U., Dirix, N., Drieghe, D., Duyck, W., 2017. Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. Behavior research methods 49, 602–615.

De Rosario-Martinez, H., 2015. phia: Post-Hoc Interaction Analysis. URL: `https://CRAN.R-project.org/package=phia`. r package version 0.2-1.

DeLuca, V., Rothman, J., Bialystok, E., Pliatsikas, C., 2019. Redefining bilingualism as a spectrum of experiences that differentially affects brain structure and function. Proceedings of the National Academy of Sciences 116, 7565–7574.

Dijkstra, T., Van Heuven, W.J., 2002. The architecture of the bilingual word recognition system: From identification to decision. Bilingualism: Language and cognition 5, 175–197.

Dijkstra, T., Wahl, A., Buytenhuijs, F., Van Halem, N., Al-Jibouri, Z., De Korte, M., Rekké, S., 2019. Multilink: a computational model for bilingual word recognition and word translation. Bilingualism: Language and Cognition 22, 657–679.

Duñabeitia, J.A., Hernández, J.A., Antón, E., Macizo, P., Estévez, A., Fuentes, L.J., Carreiras, M., 2014. The inhibitory advantage in bilingual children revisited. Experimental psychology .

Dussias, P.E., Scaltz, T.R.C., 2008. Spanish–english l2 speakers' use of subcategorization bias information in the resolution of temporary ambiguity during second language reading. Acta psychologica 128, 501–513.

Ferreira, F., Bailey, K.G., Ferraro, V., 2002. Good-enough representations in language comprehension. Current directions in psychological science 11, 11–15.

Ferreira, F., Henderson, J.M., 1990. Use of verb information in syntactic parsing: evidence from eye movements and word-by-word self-paced reading. Journal of Experimental Psychology: Learning, Memory, and Cognition 16, 555.

Ferreira, F., Henderson, J.M., 1991. Recovery from misanalyses of garden-path sentences. Journal of Memory and Language 30, 725–745.

Ferreira, F., Lowder, M.W., 2016. Chapter six - prediction, information structure, and good-enough language processing, in: Ross, B.H. (Ed.), Psychology of Learning and Motivation. Academic Press. volume 65, pp. 217–247. URL: `https://www.sciencedirect.com/science/article/pii/S0079742116300020`, doi:`https://doi.org/10.1016/bs.plm.2016.04.002`.

Ferreira, F., Patson, N.D., 2007. The 'good enough'approach to language comprehension. Language and linguistics compass 1, 71–83.

Frank, S.L., Fernandez Monsalve, I., Thompson, R.L., Vigliocco, G., 2013. Reading time data for evaluating broad-coverage models of english sentence processing. Behavior research methods 45, 1182–1190.

Frazier, L., 1987. Sentence processing: A tutorial review., in: Attention and performance 12: The psychology of reading. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US, pp. 559–586.

van Heuven, W.J., Dijkstra, T., 2010. Language comprehension in the bilingual brain: fmri and erp support for psycholinguistic models. Brain research reviews 64, 104–122.

Hsu, N.S., Kuchinsky, S.E., Novick, J.M., 2021. Direct impact of cognitive control on sentence processing and comprehension. Language, Cognition and Neuroscience 36, 211–239.

Hsu, N.S., Novick, J.M., 2016. Dynamic engagement of cognitive control modulates recovery from misinterpretation during real-time language processing. Psychological science 27, 572–582.

Jacob, G., Felser, C., 2016. Reanalysis and semantic persistence in native and non-native garden-path recovery. Quarterly Journal of Experimental Psychology 69, 907–925.

Johnson, M.L., Lowder, M.W., Gordon, P.C., 2011. The sentence-composition effect: processing of complex sentences depends on the configuration of common noun phrases versus unusual noun phrases. Journal of Experimental Psychology: General 140, 707.

Kaushanskaya, M., Blumenfeld, H.K., Marian, V., 2020. The language experience and proficiency questionnaire (leap-q): Ten years later. Bilingualism: Language and Cognition 23, 945–950.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V., 2014. The sketch engine: ten years on. Lexicography , 7–36URL: `http://www.sketchengine.eu/`.

Kremin, L.V., Byers-Heinlein, K., 2021. Why not both? rethinking categorical and continuous approaches to bilingualism. International Journal of Bilingualism 25, 1560–1575.

Lenth, R.V., 2022. emmeans: Estimated Marginal Means, aka Least-Squares Means. URL: `https://CRAN.R-project.org/package=emmeans`. r package version 1.8.3.

Luke, S.G., Christianson, K., 2018. The provo corpus: A large eye-tracking corpus with predictability norms. Behavior research methods 50, 826–833.

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

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, 940–967.

Marian, V., Hayakawa, S., 2021. Measuring bilingualism: The quest for a "bilingualism quotient". Applied Psycholinguistics 42, 527–548.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., Bates, D., 2017. Balancing type i error and power in linear mixed models. Journal of memory and language 94, 305–315.

Mayberry, R.I., Kluender, R., 2018. Rethinking the critical period for language: New insights into an old question from american sign language. Bilingualism: Language and Cognition 21, 886–905.

Millard, S.P., 2013. EnvStats: An R Package for Environmental Statistics. Springer, New York. URL: `https://www.springer.com`.

Novick, J.M., Hussey, E., Teubner-Rhodes, S., Harbison, J.I., Bunting, M.F., 2014. Clearing the garden-path: Improving sentence processing through cognitive control training. Language, Cognition and Neuroscience 29, 186–217.

Paap, K.R., Johnson, H.A., Sawi, O., 2015. Bilingual advantages in executive functioning either do not exist or are restricted to very specific and undetermined circumstances. Cortex 69, 265–278.

Paape, D., Hemforth, B., Vasishth, S., 2018. Processing of ellipsis with garden-path antecedents in french and german: Evidence from eye tracking. PloS one 13, e0198620.

R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Rayner, K., 2009. The 35th sir frederick bartlett lecture: Eye movements and attention in reading, scene perception, and visual search. Quarterly journal of experimental psychology 62, 1457–1506.

Roberts, L., Liszka, S.A., 2021. Grammatical aspect and l2 learners' online processing of temporarily ambiguous sentences in english: A self-paced reading study with german, dutch and french l2 learners. Second Language Research 37, 619–647.

Santorini, B., 1990. Part-of-speech tagging guidelines for the penn treebank project (3rd revision, 2nd printing). Ms., Department of Linguistics, UPenn. Philadelphia, PA .

Slattery, T.J., Sturt, P., Christianson, K., Yoshida, M., Ferreira, F., 2013. Lingering misinterpretations of garden path sentences arise from competing syntactic representations. Journal of Memory and Language 69, 104–120.

Snow, C.E., Hoefnagel-Höhle, M., 1978. The critical period for language acquisition: Evidence from second language learning. Child development , 1114–1128.

Suchomel, V., 2020. Better Web Corpora For Corpus Linguistics And NLP. Ph.D. thesis. PhD thesis, Masaryk University.

Teubner-Rhodes, S.E., Mishler, A., Corbett, R., Andreu, L., Sanz-Torrent, M., Trueswell, J.C., Novick, J.M., 2016. The effects of bilingualism on conflict monitoring, cognitive control, and garden-path recovery. Cognition 150, 213–231.

Traxler, M.J., 2014. Trends in syntactic parsing: Anticipation, bayesian estimation, and good-enough parsing. Trends in cognitive sciences 18, 605–611.

Voeten, C.C., Voeten, M.C.C., 2021. Package 'buildmer'.

Ye, Z., Zhou, X., 2008. Involvement of cognitive control in sentence comprehension: Evidence from erps. Brain Research 1203, 103–115.

Appendix A. Eye-tracking stimuli

The five practice items that preceded the actual experimental stimuli:

1. Rachel and Andy were walking the dog.
   (a) Was Andy walking the dog? YES
2. As the match ran late the fans were getting tired.
   (a) Did the match run late? YES
3. After the pizza arrived the kids cheered loudly.
   (a) Did the kids cheer for McDonalds? NO
4. The cat the man just bought was already feeling at home.
   (a) Did the man just buy a home? NO
5. The student who tried to enter the housing market was sorely disappointed.
   (a) Is the housing market inaccessible for students? YES

The stimuli for the experiment proper are divided as such: the first 60 items are the target (garden-path) sentences which appear in pairs, the odd-numbered sentences are the unambiguous control conditions and the even-numbered sentences are the temporarily ambiguous condition. The remaining 60 items are filler sentences, the first 30 of which being simple sentences which are based on stimuli from a previous eye-tracking study (Ferreira and Henderson (1990)), with some adaptations, based on pilot participant feedback, when they were deemed too informal or too vague. The latter 30 of the filler items are complex distractor sentences; the first 20 of which were made up by the experimenter and the latter 10 also being adapted from stimuli used in an earlier study (Johnson et al. (2011)).

Critical sentence pairs

1. While|the woman|ate|the ice cream|the polar icecaps|melted|quickly.
   (a) Did the woman eat the ice cream? YES
2. While|the woman|ate|the ice cream|melted|quickly.
   (a) Did the woman eat the ice cream? NO
3. When|the ministers|refused|the protesters|the police|responded|with violence.
   (a) Did the protesters respond? NO
4. When|the ministers|refused|the protesters|responded|with violence.
   (a) Did the protesters respond? YES
5. While|John|dressed|the child|his partner|read|a book.

   (a) Did John dress himself? NO

6. While|John|dressed|the child|read|a book.

   (a) Did John dress himself? YES

7. When|the patrol|followed|the suspects|the judge|hastened|their pace.

   (a) Did the suspects increase their speed? NO

8. When|the patrol|followed|the suspects|hastened|their pace.

   (a) Did the suspects increase their speed? YES

9. While|the firefighters|helped|the cat|its owner|followed|bystanders|around.

   (a) Did the firefighters help the cat? YES

10. While|the firefighters|helped|the cat|followed|bystanders|around.

    (a) Did the firefighters help the cat? NO

11. When|the librarian|studied|the book|an old lady|fell|on the floor.

    (a) Did the book fall on the floor? NO

12. When|the librarian|studied|the book|fell|on the floor.

    (a) Did the book fall on the floor? YES

13. After|the girl|dressed|her pet|the child|played|in the sandbox.

    (a) Did the girl dress herself? NO

14. After|the girl|dressed|her pet|played|in the sandbox.

    (a) Did the girl dress herself? YES

15. When|the guests|left|the employee|the shopkeeper|closed|the doors.

    (a) Did the guests leave the employee? YES

16. When|the guests|left|the employee|closed|the doors.

    (a) Did the guests leave the employee? NO

17. When|the driver|turned|the car|a cyclist|swerved|off the road.

    (a) Did the car stay on the road? YES

18. When|the driver|turned|the car|swerved|off the road.

    (a) Did the car stay on the road? NO

19. While|Jane|bathed|her baby|Charles|slept|on the sofa.

(a) Did Jane bathe her baby? YES

20. While|Jane|bathed|her baby|slept|on the sofa.

    (a) Did Jane bathe her baby? NO

21. When|the coach|chose|the team|the fans|laughed|mockingly.

    (a) Did the team laugh mockingly? NO

22. When|the coach|chose|the team|laughed|mockingly.

    (a) Did the team laugh mockingly? YES

23. While|the spectators|watched|the match|the excitement|escalated|quickly.

    (a) Did the match escalate quickly? NO

24. While|the spectators|watched|the match|escalated|quickly.

    (a) Did the match escalate quickly? YES

25. While|the daughter|wrote|the letter|a package|was|delivered.

    (a) Did the daughter write the letter? YES

26. While|the daughter|wrote|the letter|was|delivered.

    (a) Did the daughter write the letter? NO

27. While|the man|burned|the wood|the river|teemed|with life.

    (a) Was the wood teeming with life? NO

28. While|the man|burned|the wood|teemed|with life.

    (a) Was the wood teeming with life? YES

29. When|Matilde|hid|the valuables|her keys|were|stolen..

    (a) Did Matilde hide herself? NO

30. When|Matilde|hid|the valuables|were|stolen.

    (a) Did Matilde hide herself? YES

31. After|the players|trained|their partners|the bystanders|admired|their progress.

    (a) Did the partners admire the players' progress? NO

32. After|the players|trained|their partners|admired|their progress.

    (a) Did the partners admire the players' progress? YES

33. After|Eric|washed|the dishes|the napkins|were|collected.

(a) Were the dishes collected? NO

34. After|Eric|washed|the dishes|were|collected.

    (a) Were the dishes collected? YES

35. While|the artist|sketched|the dog|the cat|slept|soundly.

    (a) Did the dog sleep soundly? NO

36. While|the artist|sketched|the dog|slept|soundly.

    (a) Did the dog sleep soundly? YES

37. As|the driver|stopped|the bus|a pedestrian|crossed|the road.

    (a) Did the driver stop the bus? YES

38. As|the driver|stopped|the bus|crossed|the road.

    (a) Did the driver stop the bus? NO

39. As|Karen|lectured|the manager|an employee|made sure|the issue|was resolved.

    (a) Did the manager make sure to resolve the issue? NO

40. As|Karen|lectured|the manager|an employee|made sure|the issue|was resolved.

    (a) Did the manager make sure to resolve the issue? YES

41. As|Joe|bathed|the cat|the dog|stretched out|in a sunny spot.

    (a) Did Joe bathe himself? NO

42. As|Joe|bathed|the cat|stretched out|in a sunny spot.

    (a) Did Joe bathe himself? YES

43. As|the chef|cooked|the lobster|the crabs|looked|on in fear.

    (a) Did the lobster look on in fear? NO

44. As|the chef|cooked|the lobster|looked|on in fear.

    (a) Did the lobster look on in fear? YES

45. While|the journalist|wrote|the article|the newspaper|gained|in popularity.

    (a) Did the journalist write the article? YES

46. While|the journalist|wrote|the article|gained|in popularity.

    (a) Did the journalist write the article? NO

47. As|the dog|hid|his toys|the bedsheets|were|being cleaned.

(a) Did the dog hide himself? NO

48. As|the dog|hid|his toys|were|being cleaned.

    (a) Did the dog hide himself? YES

49. After|the student|paid|the debt|his despair|increased|accordingly.

    (a) Did the student pay the debt? NO

50. After|the student|paid|the debt|increased|accordingly.

    (a) Did the student pay the debt? YES

51. After|the owners|exercised|their dogs|their family|welcomed|them home.

    (a) Did the owners exercise themselves? NO

52. After|the owners|exercised|their dogs|welcomed|them home.

    (a) Did the owners exercise themselves? YES

53. After|the professor|lectured|the janitor|a company|cleaned|the building.

    (a) Did the professor lecture the janitor? YES

54. After|the professor|lectured|the janitor|cleaned|the building.

    (a) Did the professor lecture the janitor? NO

55. As|the athletes|taught|the toddlers|their parents|looked on|excitedly.

    (a) Were the toddlers watching? NO

56. As|the athletes|taught|the toddlers|looked on|excitedly.

    (a) Were the toddlers watching? YES

57. After|Charles|shaved|the sheep|the cow|stopped|grazing.

    (a) Did Charles shave himself? NO

58. After|Charles|shaved|the sheep|stopped|grazing.

    (a) Did Charles shave himself? YES

59. As|the priest|read|the leaflet|Communion bread|was|being distributed.

    (a) Did the priest read the leaflet? YES

60. As|the priest|read|the leaflet|was|being distributed.

    (a) Did the priest read the leaflet? NO

FILLER SENTENCES

1. Bill hoped that Jill arrived safely today.
   (a) Does Bill hope that Jill arrives safely? YES
2. He wrote that Sara fired her sister again.
   (a) Did Sara fire her sister before? YES
3. She dreamed that birds think like humans.
   (a) Did she dream humans think like birds? NO
4. Anne doubted that paper floats in hot water.
   (a) Did Anne have doubts about cold water? NO
5. Ted realised that cars travel on open road.
   (a) Does Ted understand cars travel on a road? YES
6. He recalled that truth follows honest folk.
   (a) Does he believe that truth follows honest folk? YES
7. Tom decided that history could be fun too.
   (a) Did Tom like history from the start? NO
8. She learned that air rises if warm.
   (a) Does air rise if it's cold? NO
9. John boasted that his luggage carries more now.
   (a) Did John boast about his luggage? YES
10. He claimed that men forget very little.
    (a) Did he claim that men have good memories? YES
11. He hinted that police officers drive in fast cars.
    (a) Does he think police officers have slow cars? NO
12. Jim warned that guns require a permit here.
    (a) Did Jim warn about fishing permits? NO
13. She confirmed that money would help him.
    (a) Will money help him? YES
14. Al promised that Mary would write a note.

(a) Did Al promise something for Mary? YES

15. He confessed that war scares little kids.

    (a) Are kids scared by bonfires? NO

16. Bob protested that dad cheats his friend too.

    (a) Did Bob like his dad's behaviour? NO

17. Don wished that kids liked reading books.

    (a) Does Don wish for kids to read more? YES

18. He forgot that Pam needed a ride with him.

    (a) Did Pam need a ride with him? YES

19. Sam insisted that teens drink light beer.

    (a) Did Sam allow the teens to drink strong liquor? NO

20. She admitted that fish like plankton too.

    (a) Did she think fish hate plankton? NO

21. She pretended that Jack owns credit cards.

    (a) Did she know that Jack does not own a credit card? YES

22. Ali suspected that policemen trained the guards as well.

    (a) Did Ali have suspicions about the guards' training? YES

23. Joe agreed that girls marry young nowadays.

    (a) Does Joe think girls wait too long for marriage? NO

24. He taught that leaves change in the fall.

    (a) Does he teach about animals? NO

25. Ed asserted that eggs cause heart problems.

    (a) Does Ed think eggs are bad for your health? YES

26. Jack revealed that drugs cause deaths too.

    (a) Does Jack think drugs are deadly? YES

27. He disputed that ideas affect the boy.

    (a) Does he believe ideas are helpful to the boy? NO

28. She observed that Mary walks to work now.

(a) Does Mary take the car to work? NO

29. Hank prayed that Jack would be bad now.

    (a) Does Hank want Jack to be bad? YES

30. She denied that the dog chased the car continuously.

    (a) Did she claim the dog did not chase the car? YES

31. That's the man who murdered his wife on the balcony on her birthday.

    (a) Did the man murder his wife on a boat? NO

32. The teacher laughed at the dog the owner who was playing with him threw a ball at.

    (a) Was the dog thrown a ball by the teacher? NO

33. The crowd the king on the stage at the festival waved to went wild.

    (a) Did the crowd get waved to by the king? YES

34. Here's the goat that got demoted at the party of the Queen by his superior.

    (a) Did the superior demote the goat at the Queen's party? YES

35. The athlete the friend the announcer knew believed in did well in the Olympics.

    (a) Did the athlete perform poorly? NO

36. There's the cowboy who shot the sheriff next to the deputy on the street.

    (a) Did the cowboy shoot the deputy? NO

37. The chef smiled at the lobster his friend who came to visit from work brought along.

    (a) Was the chef visited by a friend from work? YES

38. That's the bird that laid an egg in a nest in a tree in the park on that branch.

    (a) Was the bird's nest in the park? YES

39. The politician the reporter the policeman married trusts lied to the public.

    (a) Was the politician truthful? NO

40. The child waved at their mother the nurse had brought into the room.

    (a) Did the mother bring the nurse into the room? NO

41. Here's the artist who rocked the audience at the concert in an interview.

    (a) Was the artist being interviewed? YES

42. The country the businessman on vacation in a suit flew to was too warm for a suit.

(a) Was the country of destination warm? YES

43. The vaccine the funny aunt the woman has refused turned out to work quite well.

    (a) Did the funny aunt get a vaccine? NO

44. The tennis player denied entry at the Australian Open was disappointed.

    (a) Did the tennis player get to play at the Australian Open? NO

45. There's the cow the bull who charged the farmer in the pasture mated with.

    (a) Did the bull charge the farmer? YES

46. The song the band the girl loved released was well-received by their fans.

    (a) Was the song well-received? YES

47. That's the linguist who told his colleague in the lab about their new job over there.

    (a) Does the colleague have a new job? NO

48. The pilot wondered at the plane the co-pilot was asleep in barrelling towards him.

    (a) Was the co-pilot barrelling towards the pilot? NO

49. The heart pendant the girl the student liked stole was sold on the black market.

    (a) Did the student like the girl? YES

50. The bird the squirrel the fox chased upset started plotting its revenge.

    (a) Did the bird want to get revenge? YES

51. The doctor the student praised climbed the mountain outside of town when it snowed.

    (a) Did the student climb the mountain? NO

52. The teacher the officer phoned cooked pork chops in barbecue sauce on Christmas.

    (a) Did the teacher phone the officer? NO

53. The leader the husband liked dominated the conversation while the game was on TV.

    (a) Did the leader dominate the conversation? YES

54. The minister the brother despised drove the sports car home that evening.

    (a) Did the brother despise the minister? YES

55. The daughter that the president disliked clipped the coupons out with dull scissors.

    (a) Were the scissors very sharp? NO

56. The designer that the shopper frightened chuckled about the scare in retrospect.

(a) Was the shopper scared? NO

57. The detective that the foreigner flattered appreciated the exhibit at the museum.

    (a) Did the detective enjoy the museum? YES

58. The miser that the economist scolded blinked due to a sudden gust of dusty wind.

    (a) Was the gust of dusty wind unexpected? YES

59. The sergeant that the traveller pitied coached Little League baseball.

    (a) Did the traveller coach a baseball team? NO

60. The criminal that the painter tolerated poured syrup on the French toast.

    (a) Did the criminal pour syrup on the pancakes? NO