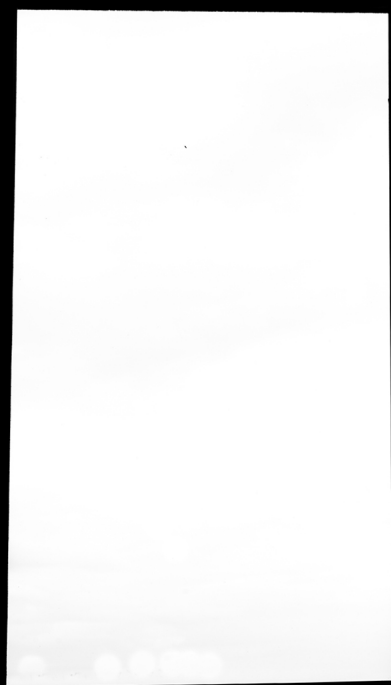


Syntax Processing



Edited by
Vicenç Torrens

Syntax Processing

Syntax Processing

Edited by

Vicenç Torrens

**Cambridge
Scholars
Publishing**



Syntax Processing

Edited by Vicenç Torrens

This book first published 2021

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

Copyright © 2021 by Vicenç Torrens and contributors

All rights for this book reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the copyright owner.

ISBN (10): 1-5275-7054-1

ISBN (13): 978-1-5275-7054-2

CONTENTS

Introduction	vi
The Processing of Temporal Concord in Sentences: An ERP Investigation into the Role of Adverb – Verb Distance	1
Nicoletta Biondo, Emma Bergamini & Francesco Vespignani	
Exploring the Syntax-Prosody Interface in Children with Developmental Dyslexia.....	48
Martina Caccia & Maria Luisa Lorusso	
All it Takes to Produce Passives in German.....	75
Yulia Esaulova, Sarah Dolscheid & Martina Penke	
Comprehension of Japanese Passives: An Eye-Tracking Study with 2-3-Year-Olds, 6-Year-Olds, and Adults.....	108
Miwa Isobe, Reiko Okabe, Yukino Kobayashi, Shigeto Kawahara, Tomoko Monou, Kazuhiro Abe, Rei Masuda, Saeka Miyahara & Yasuyo Minagawa	
Bilingual Judgments and Processing of Spanish Wh- Gap Constructions: An Exploratory Study of Cross-Linguistic Influence and Island Strength	127
Gita Martohardjono, Cass Lowry, Michael A. Johns, Ian Phillips, Christen M. Madsen II & Richard G. Schwartz	
The Timing of Interference Effects during Native and Non-Native Pronoun Resolution	171
Cecilia Puebla, Clare Patterson & Claudia Felser	
Effects of Language Dominance on L1 Relative Clause Processing.....	200
LeeAnn Stover, Michael C. Stern, Cass Lowry & Gita Martohardjono	
Remnants of the Delay of Principle B Effect in Adults: A New Approach to an Old Problem.....	228
Margreet Vogelzang, Regina Hert & Esther Ruigendijk	

INTRODUCTION

VICENÇ TORRENS¹

Language processing is a crucial part of cognition and studies conducted in this topic are particularly relevant to this field. The present volume deals with research on the processing of a native language, second language learning, bilingualism, typical and impaired syntax processing. The articles presented in this book cover a number of linguistic phenomena, including the following: passives, temporal concord, object pronouns, reflexives, embedded sentences, relative clauses, wh-movement and binding theory.

Syntactic processing has been described by processing models such as the garden path model, constraint-based models, the good-enough theory, serial and parallel processing, modular and interactive theories of language processing, and this volume covers some of these paradigms. The garden path model proposes that a single parse is constructed in a serial manner. When an initial parse of a sentence turns out to be incorrect due to syntactic ambiguity, a reanalysis of the syntactic parse takes place. During reanalysis, some principles take effect, such as late closure or minimal attachment; late closure states that words are attached to the current clause being processed; minimal attachment states that the parser builds the simplest syntactic structure possible in ambiguous sentences. Constraint-based theories are based on statistical learning: the parser takes into account the frequencies and distribution of linguistic structures; speakers apply probabilistic constraints when faced with ambiguous sentences. The good-enough models propose that listeners apply partial and superficial representations and that they do not apply a detailed syntactic analysis to sentences.

Papers in this collection apply various experimental methods, such as eye tracking, reaction times, Event-Related Potentials, picture selection tasks, sentence elicitation, pupillometry and picture matching tasks. The studies included in this book try to cover some of the most representative methods

¹ Corresponding author: Vicenç Torrens, Department of Psychology, National University of Distance Learning, C/ Juan del Rosal, 10, 28040, Madrid; email: vtorrens@psi.uned.es

used in language processing. Eye tracking is a method which measures where a participant looks, giving information about the point of gaze. The eye tracker measures the position of the eyes and their movements. Infrared light is directed to the pupils of the eyes, and the reflections are tracked by a camera. On the other hand, reaction time is a measure of the speed with which a participant reacts to an item: when a participant has to respond to a more difficult sentence, the reaction time is typically longer compared to an easier sentence. Event-related potentials (ERPs) measure the electrical activity in the scalp in reaction to an event. The event is a stimulus like a sound, a word or a sentence, sometimes followed by a question where the participants need to select an answer or remember an item. Pupillometry measures the pupil size and reactivity of participants to sentences. It provides a very accurate and precise data of pupil reactivity to stimulation. Pupillary responses can reflect activation of the brain in response to cognitive tasks: higher pupil dilation is associated with increased processing in the brain. Therefore, it has been found that pupil dilation increases when a participant needs to process a more difficult sentence. Finally, picture matching tasks study the preferential looking of participants, where they are exposed to different pictures or videos side-by-side, and the experimenter records participants' gaze data. Participants are exposed to an auditory or visual stimulus, and only one of the pictures or videos presented afterwards matches the situation depicted by the previous stimulus. The participants that understand the stimulus correctly look at the matching picture or video for a longer period of time. Measures of lexical frequency, familiarity, and imageability of target stimuli are relevant to understand the choice of participants.

The first paper of the book is the chapter “the processing of temporal concord in sentences: an ERP investigation on the role of adverb – verb distance”, where **Biondo et al.** focus on the temporal concord between the verb inflection expressing past or future information and a temporal adverb. In particular, the authors try to show how the sentence parser deals with different instances of agreement during online sentence comprehension. Temporal concord entails the coherence in time between a deictic temporal adverb and the main verb of a sentence. Deictic adverbs differ from other adverbs since they need to be anchored to the time of utterance. Biondo et al. studied thirty-nine Italian native speakers who participated in the experiment. Participants' accuracy and reaction times to the grammaticality judgement task were recorded, in addition to the EEG activity collected by 64 electrodes distributed over the scalp. With respect to the results from behavioural data, the authors did not find any significant effect of concord, distance, or concord per distance interaction. With respect to the EEG

activity, in early stages of processing no significant effects were found for distal and adjacent violations, although numerical trends are visible in the grand average ERPs; later in time, a significant effect of concord was instead found for distal and adjacent conditions, matching the properties of the P600 component. The authors conclude that distance is relevant during the processing of temporal concord, and this fact is more important during stages of reanalysis.

The next paper focuses on the syntactic properties in developmental dyslexia: in the paper “Exploring the Syntax-Prosody Interface in Children with Developmental Dyslexia”, **Martina Caccia & Maria Luisa Lorusso** study the causes and factors in the origin of reading disorders. These authors suggest that children with SLI and/or dyslexia aged 10–14 years are impaired at disambiguating linguistic structures through prosody. Typically developing (TD) children and children with developmental dyslexia (DD) who learned Italian as a native language took part in the experiment. Each sentence had an ambiguous syntactic structure which was disambiguated through prosody. The results were analyzed in terms of: a) the Response (target, alternative or distracter); b) the Type of linguistic structure; c) the Distance in the syntactic dependencies. Results showed significant main effects of type, distance and grammatical functions assignment. The presence of Dyslexia is thus associated with higher effects of syntactic complexity especially in terms of construction type and the need to restructure syntactic function, which seem not to be efficiently disambiguated by prosody.

Esaulova et al.’s paper assumes that in general, the production of passive forms is considered more effortful than that of active forms, although the use of passives may depend on both the preceding context and whether the referents are animate or patients. In addition, the differences in the outcome of these perceptual priming studies have led to the proposal that the formulation of utterances in the passive voice is dependent on the specific grammatical characteristics of a given language. These authors tested forty-five German-speaking participants in an eye-tracking study where they had to describe scenes depicting an agent and a patient character. In this study they examined whether explicit visual cueing affected the production rate of passive voice utterances in German. In half of the trials, the patient character was explicitly cued. Four lists of stimuli were created with patient Position (on the right or on the left of the scene) and Cueing (cued vs. non-cued patient) as within subjects and within items factors, and patient Animacy (animate vs. inanimate) as a within subjects and between items factor. Esaulova et al. found that explicit visual cueing of patients, as

opposed to implicit cueing, increased the probability of passive utterances produced by German speakers.

The goal of **Isobe et al.**'s paper is to investigate whether Japanese-speaking 2-3-year-olds can distinguish short passives from their active counterparts by employing the preferential looking paradigm using an eye-tracker. Previous studies found that children already have the ability to comprehend passives but have difficulty with interpreting the agent role in *ni*-phrases, i.e., 'by-phrases'. The authors tested sixty children ranging in age from 2;6 to 3;5 (mean = 2;10). They also tested twenty-five children of age 6 and ten adults for comparison. They tested a total of 16 sentences: 8 target sentences, including both active and passive sentences, along with 8 filler sentences. Passive sentences were all short passives, i.e., passives without *-ni* 'by' phrases. The results of the research showed that the 2-3-year-old children look longer at the congruent events for active sentences, whereas they tend to look at the incongruent ones longer for passive sentences; in contrast, they found that the 6-year-old children look longer at congruent scenes regardless of sentence types, a similar pattern compared to adults. Isobe et al. raise the possibility that 2-3-year-old children can notice the passive morpheme but cannot promptly revise the initial interpretation guided by the typical interpretive strategy of taking the first NP as the agent of the verb.

Martohardjono et al. assume that bilingual experience is determined by the relative exposure to the two languages, and that this exposure can vary greatly among speakers. These authors distinguish between heritage speakers, also commonly referred to as second-generation bilinguals; and late bilinguals, also commonly referred to as first-generation bilinguals. Heritage speakers are different in their bilingualism from late bilinguals, who have a more uniform and continuous experience of their first language, are schooled in that language, and acquire the other language only later in life. These authors try to disentangle whether syntactic structures found in the first-learned language that are equivalent to those found in the second-learned language might undergo weakening; also, they compare the judgment and processing between heritage speakers and late bilinguals of a *wh*-gap structure that differs in grammaticality between Spanish and English: Comp-trace interrogative sentences. This study applies pupillometry as an indicator of cognitive load in real time while participants listen to sentences, and an acceptability judgement task. Forty-five Spanish-English bilingual adults participated in this study and were categorized as either Spanish heritage speakers or adult late bilinguals. Based on these results, these authors conclude that heritage speakers don't show more influence

from English than late bilinguals, nor more weakening of grammatical principles in the first-learned language; however, even as the two groups exhibit similar grammars, their processing strategies show significant differences.

The contribution by **Puebla et al.** consists of research on the presence and timing of interference effects during the interpretation of personal pronouns, which are restricted by Condition B. Condition B states that a pronoun cannot take a c-commanding antecedent from its local domain. In pronoun resolution, there are two variables to take into account: the timing of constraint application and its interaction with other information sources. The research consists of an eye-movement monitoring task designed to investigate the presence and timing of interference effects during L1 and L2 pronoun resolution. Thirty-one native speakers of German and thirty Russian-native learners of L2 German took part in the experiment. These authors compared the L1 and L2 processing of German object pronouns in sentences that contained two c-commanding potential antecedents: a binding-accessible non-local antecedent, and an elaborated but structurally inaccessible local competitor antecedent. They found that L1 participants considered the inaccessible antecedent initially even in the presence of a gender-matching appropriate antecedent; however, the L2 participants only considered the inaccessible antecedent later on during processing, and only in the absence of a gender-matching accessible antecedent. The authors conclude that the cues to anaphor resolution are weighted differently for native and non-native populations.

The study by **Stover et al.** explores the relationship between L1 processing and language dominance in bilingual English-Spanish speakers. The authors treat language dominance as a continuous, relative measure rather than as a categorical or absolute measure. They studied the subject-object asymmetry in relative clause processing with gaze data and behavioral measures. With respect to eye-tracking data, these authors found that mean target fixation proportion is higher for Subject Relative Clauses than Object Relative Clauses in the Relative Clause Region but not in the Matrix Predicate Region. For Subject Relative Clause items, there is a pattern of increased target fixation proportions as Spanish dominance increases; however, Spanish dominance correlates negatively with target fixation proportion for Object Relative Clauses. With respect to behavioral data, Object Relative Clauses had greater response times than Subject Relative Clauses, and increased Spanish dominance also caused slower response times in Object Relative Clauses. Stover et al. found an effect of Object Relative Clauses

that affects the eye movements and response times of highly Spanish-dominant speakers compared to less Spanish-dominant speakers.

Finally, the paper by **Vogelzang et al.** deals with the processing of Principle B Effect in Dutch and German. The main experiment explored pupil dilation information, which was recorded during the presentation of stories. In German, the pupil dilation analysis shows that initially clauses with a reflexive elicit a larger pupil dilation, whereas at the end of the clause, clauses with a pronoun start eliciting a larger pupil dilation with respect to clauses with a reflexive. In contrast, the Dutch data shows that reflexives were not more effortful to process than pronouns at any point during the sentence, whereas pronouns were statistically more effortful to process than reflexives in the middle and at the end of the clause. Their results support the hypothesis that German adults in comparison to Dutch adults show less increased processing effort when resolving a pronoun compared to a reflexive. The authors argue that adults' processing effects in Dutch could be due to the Delay of Principle B Effect, and that differences in the pronominal system with respect to the use of pronouns and reflexives explain the acquisition difficulties found in children. It might be the case that if less relative processing effort is required for adults to resolve pronouns in German than in Dutch, cognitive resources could influence pronoun processing in these languages.

In conclusion, the present volume consists of a set of studies on syntax processing that were presented at the Experimental Psycholinguistics Conference in Palma de Mallorca (EPC) in June 2019. I would like to thank the plenary speakers of the Conference (Willem Mak, Esther Ruigendijk and Juan Uriagereka) and the members of the Scientific Committee (Sergi Balari, Lluís Barceló-Coblijn, Joe Barcroft, Antonio Benítez-Burraco, Denisa Bordag, Armanda Costa, Antoni Gomila, Pedro Guijarro, Aritz Irurtzun, Christer Johansson, Loes Koring, Evelina Leivada, Paulina Łęska, Paolo Lorusso, Manuela Pinto, Ankelien Schippers, and Maria del Mar Vanrell). I'd like to thank all the anonymous reviewers of this volume and also the contribution of the coordinators of this book series. Finally, I am also very grateful to Universidad Nacional de Educación a Distancia for their support on the organization of this Conference.

THE PROCESSING OF TEMPORAL CONCORD IN SENTENCES: AN ERP INVESTIGATION INTO THE ROLE OF ADVERB – VERB DISTANCE

NICOLETTA BIONDO¹, EMMA BERGAMINI
& FRANCESCO VESPIGNANI

Abstract

The processing of temporal concord anomalies (e.g., Last week/*Tomorrow I bought a car) has been rather understudied compared to the processing of other agreement relations. Moreover, previous event-related potential (ERP) studies investigating temporal concord violations reported quite heterogeneous findings. One crucial aspect that was not kept constant across studies was the linear distance between the verb and the adverb. After providing a review of the previous ERP literature investigating agreement relations in general and temporal concord in particular, we present and discuss new ERP data on the processing of temporal concord in Italian. In particular, we investigated whether adverb-verb linear distance affects the processing of temporal anomalies. Our results show that adverb-verb distance affects the amplitude of late components (P600, SFN) elicited by temporal concord violations in Italian. We conclude that distance plays a role during the processing of temporal concord, in particular during stages of reanalysis.

1. Introduction

One of the most interesting properties of human language is its redundancy. The speakers of a language can express the same formal

¹ Corresponding author: Nicoletta Biondo, Department of social, political and cognitive sciences, University of Siena, Via Roma 56 (Room 303), 53100 Siena, Italy, email: nicoletta.biondo@unisi.it

property on several words within the same sentence. For example, in a sentence like (1) in Italian, the reader comprehends that the entity performing the action of playing in a football team (the girls) refers to a plurality of individuals. We extract this information thanks to the plural number feature, which is expressed by both the determiner *le* and the noun inflection *ragazz-e* as well as by the verb inflection *gioc-ano*. Understanding the relation among constituents is pivotal to efficiently process sentences. Indeed, if we change the feature of the noun *ragazze* or the feature of the verb *giocano* as in (2) and (3) respectively, the sentences become ungrammatical.

(1) *Le ragazze che abbiamo visto all'aeroporto giocano in una famosa squadra di calcio.*

(The_{PL} girls_{PL} that (we_{PL}) have_{PL} seen at the airport play_{PL} in a famous football team)

(2) **Le ragazza che abbiamo visto all'aeroporto giocano in una famosa squadra di calcio.*

(The_{PL} girl_{SG} that (we_{PL}) have_{PL} seen at the airport play_{PL} in a famous football team)

(3) **Le ragazze che abbiamo visto all'aeroporto gioca in una famosa squadra di calcio.*

(The_{PL} girls_{PL} that (we_{PL}) have_{PL} seen at the airport plays_{SG} in a famous football team)

Linguists and psycholinguists have been using the term “agreement” or “concord” to define the consistent covariance of features between two or more words (Corbett, 2003), which is necessary in order to have grammatical and comprehensible sentences. One widely-studied instance of agreement is the one between the verb and the subject, or between the determiner and the noun. Another well-studied phenomenon is the feature covariance between a pronoun and its antecedent in sentences such as ‘*The lady is brushing herself*’.

In this chapter we will focus on a relatively understudied instance of concord, namely the temporal concord between the verb inflection expressing past or future information and a temporal adverb such as *last week*, in sentences as in (4).

(4) *Domani le ragazze giocheranno/*giocarono una partita.*

(Tomorrow_{FUT} the girls will play_{FUT}/*played_{PST} a match)

This investigation aims at widening the study of agreement phenomena and enriching the current debate on how the sentence parser deals with different instances of agreement during online sentence comprehension.

In the next sections, first we summarize the main findings on the electrophysiological correlates of agreement in sentences, then we provide a theoretical description and a review of previous experimental studies investigating temporal concord and, finally, we present new electrophysiological data on the effect of adverb-verb distance during the processing of a temporal violation within a sentence in Italian. The manipulation of linear distance can be extremely useful for the study of agreement phenomena. In the specific case of temporal concord, linear distance can affect the way the temporal information provided by an adverb is tracked before encountering a tensed verb, the way the information encoded in the verb is integrated with the information of the adverb, and the way an adverb-verb temporal inconsistency is dealt with. Our results show that distance plays a role during the processing of temporal violations, in particular during stages of reanalysis.

1.1. Event-related potential studies (ERPs) of agreement relations

ERPs is a derived measure of the electroencephalogram (EEG) which consists in the measurement of the electrical activity of the brain through electrodes placed on the scalp. One of the advantages of electrophysiological data (over many types of behavioral data) is that they reflect the neural activity continuously, without time delay. In particular, ERPs are computed by extracting and averaging the portions of EEG signal time-locked to a relevant stimulus (e.g., the verb of a sentence). A large number of trials per condition is required in order to gather a good signal to noise ratio. Psycholinguists often adopt the violation paradigm: the electrophysiological activity generated by a word containing a grammatical error (e.g., the verb in the sentence ‘*That girls dances*’) is generally compared with the activity generated by the same word in a sentence without errors (e.g., ‘*That girl dances*’). From the very first papers investigating the ERP generated by anomalous sentences (e.g., Kutas & Hillyard, 1980), researchers have been trying to characterize the ERP deflections specific to different linguistic manipulations in terms of *components*, which are defined by a functional interpretation.

For example, past ERP literature on agreement violations report that (adjective-noun, subject-verb, determiner-noun) agreement errors typically trigger a biphasic pattern, namely an early negative deflection in the 300-500ms interval followed by a positive deflection in the 500-900ms time window after the presentation of the violating word (see Molinaro, Barber & Carreiras, 2011 for an overview).

The presence, amplitude and topographic distribution of the early negativity was found to vary across studies, languages and manipulations, and its functional interpretation has been largely debated (e.g., Tanner & Van Hell, 2014; Molinaro, Barber, Caffarra & Carreiras, 2015; Tanner, 2015; Caffarra, Mendoza & Davidson, 2019). In some studies, there was no evidence of a negativity (e.g., Nevins, Dillon, Malhotra & Phillips, 2007; Frenck-Mestre, Osterhout, McLaughlin & Foucart, 2008; Alemán Bañón & Rothman, 2019). In other studies, the negativity was either distributed in the left-anterior part of the scalp or it was broadly distributed, spanning over central-posterior areas. When the negativity is anterior and left-lateralized, it is labeled as Left-Anterior Negativity (LAN). Researchers have argued that the LAN reflects automatic morpho-syntactic processing (e.g., Friederici, 1995; Friederici, Hahne & Mecklinger, 1996; De Vincenzi, Job, Di Matteo, Angrilli, Penolazzi, Ciccarelli & Vespignani, 2003; Mancini, Molinaro, Rizzi & Carreiras, 2011; Molinaro et al., 2011) or an increase in the use of working memory resources caused, for example, by the processing of long-distance dependencies (e.g., Kluender & Kutas, 1993; King & Kutas, 1995; Fiebach, Schlesewsky & Friederici, 2002; see also Martín-Loeches, Muñoz, Casado, Melcon & Fernández-Frías, 2005). Moreover, it has been argued that the LAN is more likely to appear more consistently in local within-phrase agreement such as in the determiner-noun relation than in across-phrase agreement such as the one between the subject noun phrase and the verb (Molinaro et al., 2011). When the negativity is more broadly distributed and its peak arises at 400ms after the stimulus onset, it is labeled as N400, a component that is also elicited by semantic integration problems (Kutas & Hillyard, 1980, 1984; Berkum, Haagort & Brown, 1999; see Lau, Phillips & Poeppel, 2008 for a review), and discourse-related processing (e.g., Mancini et al., 2011; see also Nieuwland & Van Berkum, 2006).

The following positive-going deflection was found more consistently across studies investigating syntactic anomalies. It typically emerges in the central-posterior areas of the scalp and it is often called P600, although sometimes in the past it was dubbed as Syntactic Positive Shift (SPS). The

P600 has been traditionally interpreted as an index of syntactic difficulty, in terms of repair, reanalysis and integration. Indeed, a P600 was found for morpho-syntactic violations (e.g., Hagoort, Brown & Groothusen, 1993; Osterhout & Mobley, 1995; Friederici et al., 1996), for garden-path sentences (e.g., Osterhout, Holcomb & Swinney, 1994; Kaan & Swab, 2003; Gouvea, Phillips, Kazanina & Poeppel, 2010), and for the processing of (grammatical) long-distance dependencies (e.g., Kaan, Harris, Gibson & Holcomb, 2000; Phillips, Kazanina, Abada, 2005). Molinaro et al. (2011), among others, proposed that the P600 could reflect two different processing stages. The early stage of the P600 (500-750ms) is more broadly distributed and it is related to the difficulty in the integration of the mismatching constituent with the previous portion of the sentence, while the later stage of the P600 (750-1000ms) is more posteriorly distributed and it represents reanalysis or sentence processing repair mechanisms (see also Kasparian, Vespignani & Steinhauer 2017 for a three-stages proposal). It should be noted, however, that the syntactic specificity of the P600 has been called into question. Indeed, a P600 emerged for non-syntactic violations, such as semantic, animacy, and thematic role violations (e.g., Kuperberg, Sitnikova, Caplan & Holcomb, 2003; Kim & Osterhout, 2005; Kuperberg, Caplan, Sitnikova, Eddy, & Holcomb, 2006; Van Petten & Luka, 2012 for a review) and for musical syntactic violations (e.g. Patel, Gibson, Ratner, Besson, & Holcomb, 1998; Patel, 2003). For this reason, several accounts have proposed that the P600 reflects an index of general reanalysis².

Many studies investigating the processing of agreement with the violation paradigm also reported a sustained negativity in the violation condition compared to the control condition (e.g. De Vincenzi et al., 2003; Hagoort et al., 1993; Osterhout & Mobley, 1995; Molinaro, Vespignani & Job, 2008). This negativity starts arising 300ms after the onset of the last word of the sentence, or as soon as the P600 response to the ungrammaticality ends, in central and posterior areas of the scalp. The functional interpretation of this negativity is still debated (see Stowe, Kaan, Sabourin & Taylor, 2018 for a recent review). One possible interpretation of this

² For example, the P600 has been related to a continued combinatorial analysis of mismatching morpho-syntactic and semantic-thematic constraints (e.g., Kuperberg, 2007), to mechanisms of well-formedness checking following a failed mapping between animacy/thematic roles and plausibility (e.g., Bornkessel-Schlesewsky & Schlewsky, 2008), to a conflict-monitoring mechanism triggered by the encountered input that does not match top-down expectations (e.g., Van De Meerendonk, Kolk, Vissers, & Chwilla, 2010).

late negativity is that it reflects the tension between the impossibility of fully analyzing and storing the sentence (because of the mismatch) and the impossibility of discarding the sentence completely (because new upcoming input could help solving the incongruities). Given that this effect often arises in presence of a metalinguistic task (i.e., grammaticality judgement task), it has also been proposed that this negativity arises when some information needs to be maintained to successfully perform the decision task.

1.2 Widening the study of agreement relations: the temporal concord

In the generative framework, the covariance of features between the subject and verb is guaranteed by a unique operation in which all features are transmitted/copied from the subject to the verb (e.g., Chomsky 1995, 2000). Similarly, in psycholinguistics, a unique mechanism for the processing of subject-verb agreement has been predicted in many mainstream models of sentence parsing (e.g., Friederici, 2002, 2011; Bornkessel & Schlesewsky, 2006; Hagoort, 2003, 2013). However, recent experimental evidence showed that the mechanism underlying agreement processing may not be unique, neither during the processing of different features (e.g., number, person) within the same relation (e.g., subject-verb), nor during the processing of the same feature (e.g., number) across different relations (e.g., determiner-noun, subject-verb, noun-reflexive). Evidence for a differentiation in the mechanisms underlying the processing of different agreement features comes from experimental studies showing longer reading time (e.g., Mancini, Postiglione, Laudanna & Rizzi, 2014; Biondo, Vespignani, Rizzi & Mancini, 2018) and different ERP responses (e.g., Mancini et al., 2011; Zawiszewski, Santesteban & Laka, 2015; Mancini, 2018) during the processing of person violations compared to number violations during in sentences. For example, Mancini et al. (2011) found the classic LAN-P600 pattern in response to number violations and an N400-like negativity followed by a P600 for person violations, compared to the control condition. Evidence for a differentiation in the processing of different concord relations comes from studies (e.g., Sturt, 2003; Phillips, Wagers & Lau, 2011; Dillon, Mishler, Sloggett & Phillips, 2013; Jäger, Engelmann & Vasishth, 2017; but see Jäger, Mertzen, Van Dyke & Vasishth, 2020) suggesting that the retrieval mechanisms implied in the resolution of subject-verb agreement and anaphora (e.g., *John likes himself*) could differ qualitatively, being the retrieved features equal (e.g., number). In particular, anaphora resolution

seems to be less prone to interference (i.e., in the retrieval of a syntactically illicit but feature matching antecedent) compared to subject-verb agreement, which shows clear interference effects from illicit antecedents.

This piece of evidence suggests that a more in-depth investigation of other concord phenomena is needed and pivotal for a more accurate formalization of the parsing mechanisms at play during sentence processing. To this end, we attempt to offer new insights about the computation of the adverb-verb temporal concord.

Past literature has been using terms such as “Tense agreement” (e.g., Sybesma, 2007; Sagarra & Han, 2008) or “temporal agreement” (Qiu & Zhou, 2012; Baggio, 2008) to label the adverb-verb relation, in line with a “broad” interpretation of the term agreement, which identifies the covariance of a semantic or formal property between two elements in a sentence. However, the term agreement has also been used in a “narrower” sense, so as to refer to the feature-checking mechanism specifically formalized to describe the feature sharing between the subject and the verb. In order to avoid terminological ambiguities, we will adopt the term “temporal concord” to refer to the adverb-verb relation. We think that this terminological choice is more appropriate. First, the term “concord” has been used to identify the adjective-noun relation (Chomsky, 2001), that is a modifier-phrase relation, such as the one between the adverb and the verb. Second, this terminological choice would also account for another property that makes adverbs and adjectives similar, namely their optionality (compared to the obligatoriness of elements such as the subject DP in the subject-verb relation).

Temporal concord entails the coherence in time between a deictic temporal adverb as *yesterday*, *next year* and the main verb of a sentence. Deictic adverbs differ from other adverbs such as clock-calendar adverbs (e.g., *on Monday*, *at 10 PM*), since they need to be “anchored” to the time of utterance (Smith, 1978; 1981). *Yesterday* is always interpreted as the 24-hour time interval preceding the time of utterance. As argued by Alexiadou (1997, 2000), deictic adverbs are marked [\pm PAST], and for this reason they can lead to ungrammaticality (‘*Yesterday Maurizio went/*will go home*’), while clock-calendar adverbs cannot (‘*At 10 PM Maurizio went/will go home*’).

From a generative perspective, the features expressed by verb inflection, namely tense and phi-features (i.e., number, person and gender), are all

encoded in a unique projection³ (TP). The co-existence of different features is easy to be seen in morphologically rich languages (e.g., Italian) where these features are all morphologically realized on the verb (e.g., *gioc-her*_{Tense-anno}_{Number/Person}), and where the verb can establish a concord relation with other phrases, be it a subject DP in subject-verb agreement or an adverbial phrase in the temporal concord.

One licit question at this point is whether the same mechanisms are at play when the verb is encountered, and subject-verb and adverb-verb relations need to be processed. These relations differ in two salient aspects at least. First, subject agreement with the inflected verb is an obligatory dependency while temporal concord is not. Second, the morphosyntactic features we are considering have an interpretive counterpart that intrinsically differ in nature.

As for the first, more formal aspect, both the subject position (EPP requirement of Government & Binding approaches, e.g., Chomsky, 1981) and the inflectional node are obligatory components of the clausal structure, and there is an obligatory requirement of match in person and number features when expressed in both constituents (in languages like Italian). In minimalist approaches, this requirement is expressed as an obligatory feature checking procedure that must take place both from the vantage point of the inflectional node, which must have its features valued by the subject, and of the subject DP, which must have its case licensed: the subject acquires nominative case through this local relation with the inflectional node⁴. Adverb concord with the temporal inflection on the verb is very different in this respect: tense is an obligatory position in the functional structure of (finite) clauses, but the adverb is an adjunct, an optional position⁵. Adverbial DPs such as *last week*, *next week* bear temporal features for past and future (Alexiadou, 1997; Enç, 1987), but the

³It should be noted that within the generative framework there are also accounts proposing and showing evidence for the existence of two distinct structural projections, AgrP and TenseP (Pollock, 1989 and subsequent cartographic work, e.g., Belletti, 1990; Shlonsky, 2010; Rizzi & Cinque, 2016).

⁴Indeed, in infinitives where the inflectional node is not active, the subject typically cannot be nominative.

⁵For many theories (e.g., Chomsky, 1986, 1995; Sportiche, 1988 among others), adverbs behave as satellites attached to the maximal projection of the phrase they modify, since they are not obligatory in a sentence. However, it is worth mentioning that there are other theories proposing that adverbs occupy the specifier position of hierarchically organized functional projections (e.g., Kayne, 1994; Alexiadou, 1997, 2000; Cinque, 1999, 2004).

temporal inflection on the verb does not need the presence of the adverbial, it has an inherent temporal value (in the minimalist terminology in Chomsky 1995, 2000 it carries an “interpretable feature”). Reciprocally, these adverbs do not need a structural case to be checked, they have an inherent case⁶ that does not need to be structurally licensed, differently from structural cases like nominative and accusative. In other words, differently from subject-verb agreement, no formal feature checking is assumed to take place for temporal concord.

The second salient difference between temporal concord and subject-verb number agreement is that tense and number have clearly different interpretive counterparts that impact different aspects of meaning construction. Number features relate to the cardinality of the set referred to by the subject argument (one entity, more entities). Tense features express the location of the event described by the verb on the temporal axis with respect to the time of utterance (past, present, future). According to the anchoring approach (e.g., Bianchi 2003, 2006; Sigurðsson 2004, 2016), we can say that the features are “anchored” to a distinct interpretive specification. Number is anchored to the cardinality of the entity referred to, a property expressed by the DP itself. Tense, on the other hand, must connect to the time of the speech (now), a specification that is expressed in the complementizer system (the left periphery of the sentence). For this reason, we can say that number features are “internally anchored” while tense features are “externally anchored”.

Do these different properties (non-/optionality, internal/external anchoring) play any role during online sentence comprehension? According to some accounts, they do. The Construal model by Frazier & Clifton (1996), for example, predicts different parsing routines at play during the processing of primary and non-primary relations. Primary relations entail obligatory phrases such as verbs and verb’s arguments. These phrases and relations are assumed to be processed through the attachment mechanisms described by classical syntax-first models (Frazier, 1987). Non-primary relations entail optional phrases such as adjuncts. The processing of these phrases and relations follows the construal principle: an upcoming non-primary phrase is associated to the current thematic domain (i.e., the last theta assigner) and it is interpreted by using both structural and non-structural information. It follows that the processing of primary relations should be more “resistant” and automatic while the processing of non-primary relations can be influenced by extra-syntactic (e.g., semantic,

⁶ As we see in languages with overt morphological case.

pragmatic) factors. Moreover, other accounts predict different mechanisms at play based on the (discourse-related) interpretive properties of different features. Following the anchoring approach proposed in linguistic theory, Mancini, Molinaro & Carreiras (2013) theorized and showed in several experimental studies that the anchoring properties of different features (i.e., number, person) can affect sentence processing. For example, in an ERP study Mancini et al. (2011) compared the processing of number and person subject-verb violations. While number violations gave rise to the classic LAN-P600 response compared to the control condition, person violations elicited an N400-like negativity followed by a P600. Moreover, the P600 was larger in amplitude and more frontally distributed for person than for number violations. The N400-like negativity for person violations was interpreted as an index of failure in the establishment of the interpretive relations among constituents (i.e., is the subject the speaker, the addressee or neither of the two, in the speech act?), which goes beyond the level of morpho-syntax (for which a LAN effect is expected). The frontally distributed P600 effect (compared to the more posterior topography of the P600 generated by number violations) was also linked to discourse-related integration difficulties, namely to the impossibility of integrating the (mismatching) participants' role expressed by the subject and the verb in the same discourse representation. More recently, in one eye-tracking study Biondo et al. (2018) compared the processing of number, person and tense violations in the same experimental paradigm. Number and person violations led to similar parsing costs in early stages of processing while the parsing costs for tense violations appeared only in later stages. In other words, the "primary" subject-verb violations gave rise to earlier and stronger parsing costs than the "non-primary" temporal violations. Moreover, person anomalies caused larger parsing costs than number anomalies in the same stage in which tense anomalies started showing a cost. In other words, the violation of person and tense, the two features that need a reference to discourse to be interpreted (speech participants, speech time), affected the same (late) stages of processing. These findings suggest that online sentence comprehension is differently affected by both the obligatoriness of the relation and by the discourse-related properties of the features under computation.

Despite the theoretical and experimental evidence suggesting that a deeper investigation of other concord phenomena is needed, the temporal concord has been rather understudied. One possible reason can be related to the

optionality of the temporal adverbs⁷. However, the processing and attachment preferences of optional phrases/adjuncts such as prepositional phrases or relative clauses has been quite a debated topic in psycholinguistics (Cuetos & Mitchel, 1988; De Vincenzi & Job, 1993; Carreiras & Clifton, 1993; Carreiras, Salillas & Barber, 2004). More importantly, the temporal concord has played a pivotal role in other research fields. For example, several studies investigating language impairment in aphasia reported that agrammatic patients show marked difficulties during the comprehension and production of temporal concord while subject-verb agreement is less impaired in the same patients. The source of this linguistic impairment has been related to a difficulty in the representation of the syntactic structure (TenseP) encoding temporal information (e.g., Friedmann & Grodzinsky, 1997), or to a difficulty in the representation of tense features encoded in the sentence structure (e.g., Nanousi, Masterson, Druks & Atkinson, 2006; Clahsen & Ali, 2009). Temporal concord and subject-verb agreement also seem to be acquired in different stages both in child first language acquisition (Weist, 2014; Belletti & Guasti, 2015) and in adult second language acquisition (Biondo & Mancini, 2019). We thus believe that a deeper investigation of the temporal concord processing, in unimpaired sentence comprehension is necessary to have a wider look at concord and agreement processing and to implement current mainstream formalizations of sentence parsing routines (e.g., Friederici, 2011; Haagort 2013; Bornkessel & Schlesewsky 2006). A richer knowledge of the mechanisms involved in “typical” sentence processing can help us to understand which mechanisms develop earlier/later during (first and second) language acquisition, and which routines are impaired in presence of a language disorder and why.

1.3 ERP literature on the processing of temporal concord violations

Compared to the extended ERP literature on the processing of subject-verb agreement, there are few ERP studies that investigated the processing of temporal concord anomalies in adult (unimpaired) sentence reading. Table 1 summarizes the studies where adult native speakers of a language (without language impairments) were asked to read sentences containing either a correct or a mismatching temporal relation between a deictic temporal adverb and a verb expressing a temporal value (either through tense or through temporal particles). Despite their intrinsic differences

⁷ Even when temporal adverbs are expressed in a sentence, a feature consistency is not always required, e.g., for calendar-clock adverbs.

(e.g., type of language, experimental material tested), these studies showed that temporal violations are detected quite early in sentence processing (300-500ms after the stimulus onset). In almost all the studies, the mismatching verb compared to its correct counterpart elicited an early negativity. The topography of this negativity, however, spans from a left anterior to a more central/posterior distribution. Conversely, a P600 component was reported consistently, in all the studies. Finally, some of the studies reported a widely distributed sustained negativity arising between 300ms and 800ms after the onset of the final word of the sentence when a temporal violation was encoded on the verb of the sentence.

One licit question is if/how the electrophysiological correlates of the temporal concord differ from the correlates of other agreement phenomena, and where the source of heterogeneity resides, especially in the early time window. In the following paragraphs, we first look at pure methodological variability⁸ (e.g., in the acquisition/analysis of the EEG data) which can explain the presence/absence/topographic distribution of the components involved in the processing of temporal violations. Then, we consider and test one of the linguistic factors that could have affected the outcome of previous studies, namely adverb-verb distance.

Table 1 (next page). Linguistic details, tasks and results of the ERP studies investigating the processing of temporal concord violations. The ERP components refer to the onset of the target words, which are underlined. LPN stands for left posterior negativity while RAN stands for right anterior negativity. The studies marked with an asterisk (see * in the Authors column) also reported a sustained SFN (Sentence Final Negativity) for tense violations.

⁸ Some studies (E1, E2) lacked methodological details so they could not be considered.

Code	Authors	Language	Experimental sentences (translated maintaining the same word order)	Task	0	300	500	750	1000 ms
E1	Fonteneau et al., 1998	French	Tomorrow the student <u>will</u> read/was reading the book.	Acceptability					
E2	Steinhauer & Ullman, 2002	English	Yesterday, I sailed/sail Diane's boat to Boston. Yesterday, we ate/en Peter's cake in the kitchen.	Acceptability		LAN		P600	
E3	Newman et al., 2007	English	Yesterday I <u>frowned/frown</u> at Billy. Yesterday I <u>ground/grind</u> up coffee.	Acceptability		LAN		P600	
E4	Baggio, 2008*	Dutch	Last Sunday painted/paints Vincent the window-frames of his country house.	Passive reading		LAN		P600	
E5	Qiu & Zhou, 2012*	Mandarin Chinese	Next month/*Last month United Nations <u>jiangzai</u> dispatch a special investigation team. Last month/*Next month United Nations <u>ceniging</u> dispatch a special investigation team.	Acceptability		N400	P600		
E6	Dillon et al., 2012	Hindi	Although last night that traveler stone upon fall- <u>aa</u> /*fall-e- <u>aaa</u> , but to him injures not happen.	Acceptability		Posterior negativity (N400-like)		P600	
E7	Dragoy et al., 2012*	Dutch	The waiter who now/*just before the pepper <u>grinds</u> gets no tip. The waiter who now the pepper <u>grinds</u> /* <u>ground</u> gets no tip.	Comprehension		P300		P600	
E8	Bos et al., 2013	Dutch	The grampa who a moment ago the coffee has <u>ground</u> /*will grind looks after his visitors.	Comprehension		P300		fP600	pP600
E9	De Vincenzi et al., unpublished*	Italian	The secretary long time ago called/*will call for a meeting.	Comprehension		RAN			P600

1.3.1 Methodological review

Let's first focus on the early component elicited in the 300-500ms time window. Some studies reported a left-lateralized negativity (E2, E3, E4), other studies reported a more distributed negativity (E1, E5, E6), one study reported a right-lateralized negativity (E9) and, finally, other studies reported a posterior positivity (E7, E8). It should be noted that the reference choice can be crucial for the detection of lateralized components (such as LAN). The activity detected by the reference (and the surrounding electrodes) is subtracted from the activity of the other electrodes. Consequently, if the reference is located on the same (e.g., left) side of the component under investigation, the amplitude of that effect could be reduced (Molinaro et al., 2011; 2015; but see Tanner, 2015). This could be the case of E9, where all scalp channels were referenced to the left mastoid and a right-lateralized negativity was found for temporal violations. The presence of a more broadly distributed negativity in E5, E6, on the other hand, should not be due to reference issues given that linked mastoids⁹ were used as reference. Some of these studies (E7, E8) also reported a very different component for the items containing a temporal violation, namely a sustained positive activity that started around 300ms after the stimulus onset. Crucially, both studies reported that less than/only a third of the sentences presented in each list contained violations (Dragoy et al., 2012:313; Bos et al., 2013:290). As pointed out by Molinaro and colleagues (2011), the P600 may be sensitive to the proportion of violations in the experimental set. Coulson, King & Kutas (1998) indeed showed that the rarest stimuli (i.e., the stimuli with the smallest proportion in the whole set) elicit a larger P3b, a positive going component with central-parietal maximum that can start around 300ms after the stimulus onset. It can thus be possible, that the early latency of the P600 in E7 and E8 is related to the rarity/small proportion of sentences containing a violation.

Differently from the early components, the P600 arising in the 600-900ms time window showed up quite clearly and constantly in all the studies on temporal concord violations. Whether this component is modulated by linguistic factors cannot be assessed. In the current study, we will be able to test whether the P600 is affected by the distance of the two constituents.

⁹ In particular, online linked mastoids were used as reference in E6 while offline linked mastoids were used in E3, E4, E5, E7, E8.

Finally, the presence of a broadly distributed sentence-final negativity (SFN) in response to temporal violations was partially reported. Some studies did report an SFN in response to temporal violations (E4, E5, E7, E9), while one study did not find any final negativity (E8). The remaining studies did not report any analysis of the ERPs triggered by the last word of the sentence (E1, E2, E3, E6). In other words, we cannot assess whether in these studies no SFN was present, or whether the SFN was present and not reported. One observation that should be considered is that the presence of the SFN does not seem to be related to a specific task, since it was found in passive reading (E4), in comprehension (E7, E9), and in acceptability (E5) tasks.

To sum up, the methodological review of the previous studies suggests that the presence of a right-lateralized early negativity or of an early positivity in response of a temporal violation may be related to specific methodological choices rather than to the processing of temporal concord per se. One issue that is still unsolved, however, is the left-anterior (LAN-like) or more central (N400-like) topographic distribution of the early negativity found in previous studies. Although we can exclude technical factors, there are still several linguistic factors that could account for this variability. The role of one of these factors, namely the distance between the adverb and the verb, is addressed in the current study. Moreover, the current study also allows to test/replicate the presence of the P600 and SFN in response to temporal violations, as well as their hypothetical modulation as a function of adverb-verb distance.

1.3.2 Does distance play any role in the processing of temporal concord violations?

A deeper look at the experimental material which was adopted in previous studies shows that the configuration between the adverb and the verb was not kept constant, both linearly and structurally. In some studies, the temporal adverb and the verb were adjacently located (E4, E9), while in others the two elements were separated by a pronoun (E2, E3), by a lexical subject (E5), by a lexical object (E7, E8), or by several phrases (E6). Some representative examples are reported in (5 - 7).

(5) From Baggio (2008):

*Afgelopen zondag*_{PST} *lakte*_{PST}/*lakt*_{PRS} Vincent de kozijnen van zijn landhuis. (Dutch)

(‘Last Sunday painted/*paints Vincent the window frames of his country house’)

(6) From Fonteneau et al. (1998):

*Demain_{FUT} l’étudiant lira_{FUT}/*lisait_{PST} le livre.* (French)

(‘Tomorrow the student will read/*read the book’)

(7) From Dillon et al. (2012):

*Haalaanki pichle shaam_{PST} vo raahgiir patthar ke-uupar gir-aa_{PERF}/*gir-e-gaa_{AGR-FUT}, lekin use choT nahiin aa-yii.* (Hindi)

(‘Although last night that traveler stone upon fell/*fall, but to him injures not happen’)

Interestingly, the studies where the adverb and the verb were adjacent, as in (5), reported a left anterior negativity, while the studies where the adverb and the mismatching verb were divided by one or more lexical constituents, as in (6) and (7), triggered a more central/posterior negativity. It thus seems that distance may play a role in the detection of the temporal violation. No ERP study has ever explicitly tested the effect of distance on temporal violations, but there is some behavioral evidence.

In an eye-tracking study, Biondo et al. (2018) manipulated both the grammaticality of the tensed verb and the distance between the verb and the adverb. Crucially, the authors found that the distance of the two constituents affected the processing of the concord violation. When the adverb and the verb were adjacent the effects of temporal mismatch showed up in late measures, while when the adverb and the verb were distally located the mismatch effect was found from early measures on. In other words, the larger the distance between the adverb and the verb, the earlier the detection/processing of the temporal violations. The authors interpreted these findings within a predictive framework. If phrases encoding a complex lexical content, such as nouns, require time to be semantically interpreted (e.g., Frazier & Clifton, 1998; Kreiner, Garrod, & Sturt, 2013; see also Chow, Lau, Wang & Phillips, 2018), some time may also be needed to anchor the temporal specification of the adverb to discourse (i.e., Speech Time in the left periphery of the sentence). Under the assumption that sentence comprehension proceeds in an incremental way and that anchoring requires time to be completed, one may expect that when the verb immediately follows the adverb, the temporal specification

retrieved from the lexical representation of a deictic temporal adverbial phrase could not be immediately available to syntactic processing. The system does not have enough time to correctly and fully activate a temporal representation provided by the adverb, leading to a more delayed detection of the violation. On the contrary, if some words intervene between the already parsed adverb and the verb, it can be more likely that there is more time to solidly anchor deictic information to discourse, allowing an earlier detection of the temporal mismatch. These findings suggest that distance may play a role in the unfolding of some mechanisms such as the extraction of discourse-related properties from the linguistic input, thus making the study proposed here specifically relevant to better understand the processing of temporal concord. Moreover, ERPs are particularly sensitive to qualitative changes in the processing of a relation, thus helping to qualify the nature of these extra costs found behaviorally.

2. The current study

The main aim of this study was to investigate how the parser deals with temporal concord violations, and whether the processing of temporal concord is affected by the distance between the two relevant constituents, the adverb and the verb. All previous ERP studies on temporal concord violations reported an early negativity, so we expect both adjacent and distal temporal violations to be detected early in processing (300-500ms post-verb). Moreover, the ERP literature on temporal concord, together with recent behavioral studies (e.g., Biondo et al. 2018), suggested that the distance between the adverb and the verb should affect early stages of processing, probably in terms of earliness and/or easiness of detection of the violation. We thus expected distance to affect the early ERP components. If the anchoring to discourse requires time (Biondo et al., 2018) and discourse-related processing can affect early ERP components (Mancini et al., 2011) we should expect an N400-like component for distal violations, where more time is given to the anchoring process to be completed, compared to the adjacent violations where a LAN effect could be found (i.e., a pure morphological mismatch detection). This pattern would also be in line with previous ERP findings reported in Table 1, showing LAN effects for adjacent temporal mismatches and N400-like effects for distal temporal mismatches.

Whether the adverb-verb distance also affects later stages of processing is hard to establish because all previous studies reported a P600 effect, and possible amplitude and topographic differences were not directly

documented, and thus hard to be quantified across studies. However, a modulation of the P600 amplitude as a function of the adverb-verb distance cannot be excluded, especially in light of the previous studies investigating the effect of distance during subject-verb agreement processing. Kaan (2002) tested whether intervening material between the subject and the verb affects feature tracking, integration, or the revision processes related to the detection of a subject-verb mismatch. The ERP components triggered by the mismatch (broad negativity followed by a P600 and a sustained end-of-sentence negativity) were not affected by distance, neither in early nor in late stages of processing. The effect of distance was only visible in the behavioral response: the readers were less accurate in judging the ungrammaticality of the sentences where the subject and the verb were distally located (compared to the sentences with an adjacent mismatch). Shen, Staub & Sanders (2013) also tested the effect of distance in an ERP paradigm. The ungrammatical sentences where the subject and the verb were adjacent (compared to their correct counterpart) showed an early anterior negativity followed by a P600. The ungrammatical sentences in the distal condition (compared to their control) triggered an early posterior negativity that the authors interpreted as an instance of N400. It should be noted, however, that the authors opted for a more naturalistic task in this study, namely a text describing a story in which semantic and pragmatic factors clearly played a stronger role and possibly induced a larger involvement of N400-related processes. It is thus hard to make a comparison with previous studies testing the same conditions in isolated sentences. More recently, Rispen & Amesti (2017) manipulated both the number (0, 1, 2) and the type of constituents (adverb, prepositional phrase containing agreement features) located between a mis/matching subject-verb relation in isolated sentences. They replicated the behavioral finding of Kaan (2002), since participants were less accurate in judging the distal violations than the adjacent ones. In the early time window (300-500ms) a posterior negativity was found for all the violations, independently from the number and type of interveners (the type of constituent did not affect any behavioral response or ERP component). Crucially, an effect of distance was found in the 500-1000ms time window: the distal condition elicited a larger P600 than the adjacent condition. The authors interpreted these data as evidence that two distally constituents are harder to integrate and more complex to process as compared to two adjacent constituents¹⁰. In sum, based on previous studies

¹⁰ This interpretation, however, is rather speculative since the authors did not find any significant grammaticality x distance interaction, but only two main effects of grammaticality and distance (Rispen & Amesti, 2017: 169)

on the effect of distance during the processing of subject-verb agreement, we also expected a modulation in the amplitude of the P600 component. Differently from previous studies, however, we considered two different phases in the P600 time window: the integration phase (earlyP600) and the revision phase (lateP600). If distance affects the integration of the verb and the adverb, we should expect an interaction between distance and grammaticality in the earlyP600 time window. If distance (also) affects the revision processes related to the violation, we should expect an interaction between distance and grammaticality in the lateP600 time window. In particular, based on the anchoring account (Mancini et al., 2013; Biondo et al., 2018) and other accounts on online incremental semantic interpretation (e.g., Frazier & Clifton, 1998), we expected both the integration and the revision process to be harder in the distal mismatch condition. The semantic interpretation of the reference time provided by the adverb should make the integration and the revision of the temporal information encoded in the verb harder to process. This integration and revision difficulties should be mirrored in a larger (early and late) P600 for distal compared to adjacent temporal violations.

The electrophysiological activity generated by the last word of the sentence was also analyzed to see whether temporal violations elicit a sustained SFN as in previous studies, and to explore whether this negativity is modulated by distance. To our knowledge, there are no previous studies suggesting that the SFN should be modulated by the distance between the two constituents. Based on the functional interpretation of this component, we hypothesized two scenarios. If the presence of an SFN is purely task-related, we should find this component both in the adjacent and in the distal violation conditions. Conversely, if the SFN is somehow related to the difficulty that the parser encounters in analyzing and storing the sentence, the SFN should be present (or boosted) in the adjacent condition, namely the condition in which the system has less time to fully process the adverb (and the adverb-verb mismatch).

3. Method

3.1. Participants

Thirty-nine Italian native speakers (11 men, 28 women) participated in the experiment as volunteers. Mean age of the participants was 22.6 years ($SD = 3.2$). All participants were right-handed, had normal or corrected to normal vision and had no history of neurological disorders. The research

was carried out fulfilling ethical requirements in accordance with the standard procedures at the University of Trento.

3.2 Experimental material

A set of 160 Italian experimental sentences was composed. All sentences consisted of an animate noun phrase in subject position, a temporal adverb, and a lexical verb followed by other constituents such as a direct/indirect object DP or a locative PP. The target word was never at the end of the sentence to avoid the overlap of wrap-up effects and the critical effects of interest (Osterhout & Holcomb, 1995). The temporal adverb was always deictic, that is it always defined a time interval that was unambiguously located in the past (e.g., *yesterday*) or in the future (e.g., *tomorrow*) with respect to the time of the speech (now). The tense features of the verb could either be past or future, and they were counterbalanced across lists.

Four different versions of each experimental sentence were created, as shown in Table 2. In half of the conditions, the adverb was located between the subject and the verb (adjacent condition) while in the other half the adverb was located at the beginning of the sentence (distal condition). Moreover, in half of the sentences the temporal adverb and the verb expressed the same temporal information (match condition) while on the other half the adverb and the verb mismatched in temporal features (mismatch condition). The experimental sentences were rotated across lists following a Latin square design such that each sentence was presented in one different condition across lists. A list contained four conditions of 40 items each, as well as 160 filler items. The filler items were sentences containing subject-verb mis/matches. Only 80 of these filler sentences presented agreement violations so that each participant saw an equal number of correct and incorrect sentences in the whole experiment. In total, each participant read 320 sentences.

Table 2. Sample of the experimental conditions. In this example, only the future conditions are displayed but past and future were counterbalanced across lists.

Distance	Concord	
	<i>Match</i>	<i>Mismatch</i>
<i>Adjacent</i>	Il poliziotto domani <u>testimonierà</u> di fronte al giudice. (‘The policeman tomorrow will testify in front of the judge’)	Il poliziotto ieri <u>testimonierà</u> di fronte al giudice. (‘The policeman yesterday will testify in front of the judge’)
<i>Distal</i>	Domani il poliziotto <u>testimonierà</u> di fronte al giudice. (‘Tomorrow the policeman will testify in front of the judge’)	Ieri il poliziotto <u>testimonierà</u> di fronte al giudice. (‘Yesterday the policeman will testify in front of the judge’)

3.3 Procedure

Each participant was seated in a dimly lit chamber. Sentences were displayed word by word in white letters on a monitor with dark-grey background, by using the E-Prime software. At the beginning of the session, 12 practice trials were presented to ensure that the participants understood the instructions. Each trial began with a fixation cross that disappeared when the participant pressed the space bar. Then, a blank screen of 300ms was followed by the first word of the sentence, which was displayed for 300ms, and again a 300ms blank interval followed. The alternation of words and blank screens continued with the same pace until the end of the sentence. The final word was displayed with a full stop. Participants were then asked to judge the correctness of the sentence just read, by pressing one of two buttons (C, M) on the keyboard. The button position was counterbalanced across participants, and the presentation of the trial was randomized for each participant. Participants were instructed to minimize eye movements, blinks and muscle activity while reading the

sentences. The experimental session was divided in two blocks with a short break between blocks. The entire session lasted one hour.

3.4. Data acquisition and analysis

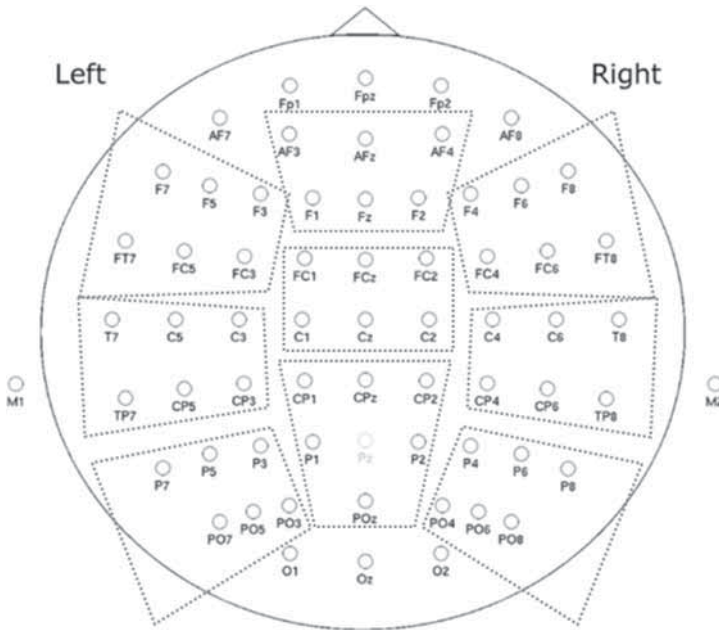
Participants' accuracy and reaction times to the grammaticality judgement task were recorded by using the E-prime software. One participant was excluded because of low accuracy (lower than 70%). Both accuracy and (log-transformed) reaction times were analyzed by using linear mixed-effect models while accuracy was analyzed by using logistic mixed-effect models (Jaeger, 2008), which are both implemented in R through the package lme4 (Bates, Maechler, Bolker & Walker, 2014). The models were built adding concord, distance and their interaction as fixed-effects factors. Crossed random intercepts and random slopes for all fixed-effect parameters both for subject and item grouping factors were also added (Barr, Levy, Scheepers & Tily, 2013). The selection of the best-fitting model was performed by using a parsimonious approach (Bates, Kliegl, Vasishth & Baayen, 2015). The starting model contained the most complex random effect structure (justified by the experimental design), and correlation parameters forced to zero (zero-correlation parameters model). The random effect structure of the model was then gradually reduced by performing a principal component analysis (PCA, RePsychLing package), which allowed to exclude the components that accounted for very small variance (around 0). Finally, the reduced model was re-extended by adding the (subject and item) correlation parameters, to see whether the goodness of fit could significantly increase. All the models were compared by using the *anova* function.

EEG was recorded from 64 electrodes placed in a shielded waveguard cap (eegoTMsports, ANT Neuro) and distributed over the scalp at standard positions (10–20 system). All sites were referenced to the left mastoid (M1). Impedance was kept below 5 k Ω for mastoid and scalp electrodes. An external bipolar channel was placed above and below the right eye (BIP1). Data were acquired at a sampling rate of 1000 Hz. The preprocessing of the EEG data was performed with the BrainVision Analyzer software. The EEG recordings were off-line resampled (200Hz) and re-referenced to the average activity of the two mastoids. The offline filtering consisted of a low cutoff filter of .01 Hz and a high cutoff of 30 Hz (Acunzo, MacKenzie, & van Rossum, 2012; Tanner, Morgan-Short & Luck, 2015). The ocular artifacts were corrected by performing a semiautomatic Independent Component Analysis (ICA) based artifact correction procedure, on each participant recording. The independent

components that explained artefactual activity were automatically identified by the Brain Vision Analyzer algorithm, but the selection and the final subtraction of the component that resembled the characteristic pattern of ocular artifacts (Plank, 2013; Jung et al., 1998) was made manually. After the ICA correction, if necessary, we interpolated the Pz electrode by considering the average activity of the 4 surrounding electrodes (P1, P2, CPz, POz). This step was made because in twenty EEG recordings, Pz either did not detect the scalp electrical activity or stopped working during the session. We segmented the whole EEG recording in epochs (-200 to 1200 ms) based on the trigger positions corresponding to the critical target word onset and to the sentence-final word onset. Artifacts due to muscle activity exceeding 100 μ V in amplitude (in a 100ms time window) were rejected automatically while other artifacts (e.g., physiological artifacts) were rejected through visual inspection. We then rejected the epochs that were marked as containing artifacts. Five participants were excluded from the analysis given the high number of rejected epochs (more than 35%). In the remaining participants, there were no differences in the number of epochs across conditions ($F_{(1,32)} = 0.05$, $p = 0.83$). The epochs of the remaining participants were then baseline-corrected to the average activity in the 200ms interval preceding the presentation of the target word and of the last word of the sentence. We first averaged the epochs separately for each participant and condition, and we then calculated the grand average for each of the four conditions.

The single channel ERP data were analyzed by considering 9 different groups of electrodes. Each cluster represented the averaged activity of 6 spatially close electrodes, as shown in Figure 1. In the lateral sites, we created six clusters: anterior-left (AL: F7, F5, F3, FT7, FC5, FC3), anterior-right (AR: F8, F6, F4, FT8, FC6, FC4), central-left (CL: T7, C5, C3, TP7, CP5, CP3), central-right (CR: T8, C6, C4, TP8, CP6, CP4), posterior-left (PL: P7, P5, P3, PO7, PO5, PO3), posterior-right (PR: P8, P6, P4, PO8, PO6, PO4). In the midline site, three clusters were considered: frontal (F: AFz, AF3, AF4, Fz, F1, F2), fronto-central (FC: FCz, FC1, FC2, C1, Cz, C2), central-posterior (CP: CPz, CP1, CP2, P1, P2, POz).

Figure 1. Topographic distribution of the electrodes and clusters.



The ERPs time-locked to the onset of the verb were analyzed through a 2 (concord) \times 2 (distance) \times 3 (anteriority) \times 2 (laterality) repeated measures ANOVA in the lateral sites, and a 2 (concord) \times 2 (distance) \times 3 (anteriority) repeated measures ANOVA in the midline sites. In other words, two different topographic factors were considered in the statistical analyses: *laterality* (right, left) and *anteriority* (anterior, central, posterior). The other (within-subject) experimental factors were *concord* (match, mismatch) and *distance* (adjacent, distal). Each analysis was performed on specific time windows that past studies have defined of interest for the observation of ERP components triggered by concord violations: 300–500ms (LAN), 500–800ms (earlyP600) and 800–1000ms (lateP600). We report significant main effects or interactions, in lateral and midline sites. The effect of the topographic factors is reported only if it interacted with the experimental factors. In presence of an interaction, also the post-hoc analyses that were performed on each cluster are reported. The Greenhouse-Geisser procedure was applied when the sphericity assumption was violated, and p-values were adjusted based on the “fdr”

method in the pairwise comparisons. The same analysis was also applied to the time intervals following the onset of the last word of the sentence to check the presence and/or modulation of the SFN.

4. Results

4.1. Results of behavioral data

Average accuracy and reaction times in the grammaticality judgment task are reported numerically in Table 3 and graphically in Figure 2.

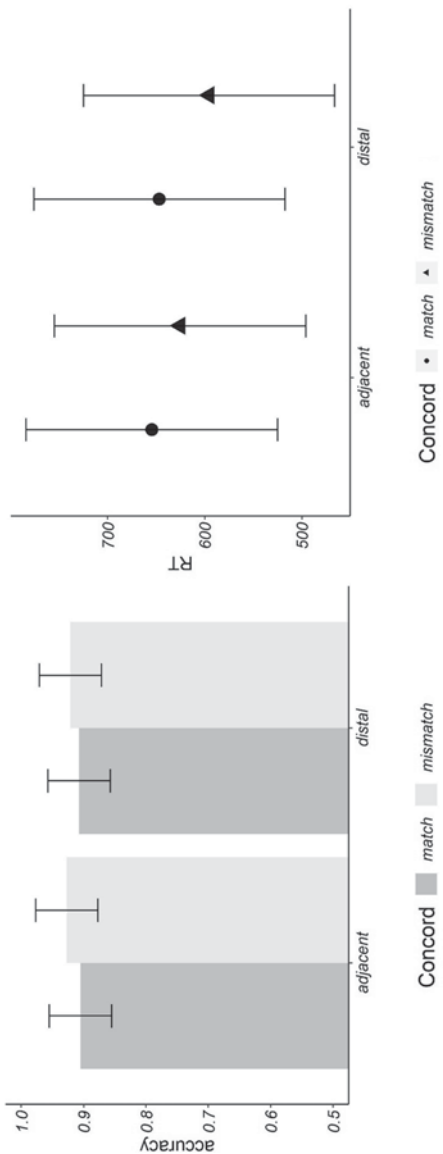
Table 3. Average accuracy and reaction times (standard deviations in brackets) for each condition, in the grammaticality judgment task.

Conditions	Accuracy	Reaction time
<i>Adjacent Match</i>	0.91 (0.29)	654.46 (533.95)
<i>Adjacent Mismatch</i>	0.93 (0.26)	625.46 (482.32)
<i>Distal Match</i>	0.91 (0.29)	646.75 (490.55)
<i>Distal Mismatch</i>	0.92 (0.27)	595.52 (457.21)

The analysis of accuracy did not show any significant effect of concord (Intercept: 2.75; Estimate: 0.18, SE: 0.26, Wald's $z = 0.69$, $p = 0.49$), distance (Intercept: 2.75; Estimate: 0.01, SE: 0.17, Wald's $z = 0.08$, $p = 0.94$), or concord x distance interaction (Estimate: -0.02, SE: 0.22, Wald's $z = -0.09$, $p = 0.93$).

The analysis of the log-transformed reaction time data showed a marginally significant concord x distance interaction (Estimate: -0.06, SE: 0.03, $t = -1.96$, $p = 0.05$). This interaction was driven by the fact that the mismatch distal condition has significantly faster reaction times than its control counterpart (Intercept: 6.25, Estimate: -0.09, SE: 0.03, $t = -3.03$, $p < 0.05$), while no effect of concord was found in the adjacent condition (Intercept: 6.23, Estimate: -0.02, SE: 0.03, $t = -0.69$, $p = 0.49$).

Figure 2. Average accuracy (plot on the left) and average reaction times (plot on the right) for each condition, in the grammaticality judgment task. The bars represent standard errors.



4.2 Results of EEG data

4.2.1 Target word

The average activity time-locked to the target word onset, for each cluster of electrodes and each experimental condition, is displayed in Figure 3.

In the 300-500ms time window, no significant main effects or interactions were found in the lateral sites (concord, distance, concord x distance $F < 1$). In the midline sites, only a main effect of distance was found [$F_{(1,32)} = 4.21$, $p < 0.05$] while concord [$F_{(1,32)} = 0.3$, $p > 0.05$] or the distance x concord interaction [$F_{(1,32)} = 0.3$, $p > 0.05$] did not reach significance.

In the 500-800ms time window, a concord x anteriority interaction [$F_{(2,64)} = 4.03$, $p < 0.05$] as well as a concord x distance x anteriority x laterality interaction [$F_{(2,64)} = 3.33$, $p < 0.05$] were found in the lateral sites. A concord x anteriority interaction [$F_{(2,64)} = 6.64$, $p < 0.05$] was also found in the midline sites. Post-hoc analyses showed that the concord violation effect was reliably present only in the mid-posterior cluster [$F_{(1,32)} = 4.70$, $p < 0.05$] for both violations, while no concord x distance interaction was found in any other cluster ($F < 1$).

In the 800-1000ms time interval, the analysis of the lateral sites revealed a main effect of concord [$F_{(1,32)} = 7.50$, $p < 0.05$], a concord x anteriority interaction [$F_{(2,64)} = 20.76$, $p < 0.05$] and a concord x distance x anteriority x laterality interaction [$F_{(2,64)} = 4.42$, $p < 0.05$]. In the midline sites, we found a main effect of concord, [$F_{(1,32)} = 6.45$, $p < 0.05$], a concord x anteriority interaction [$F_{(2,64)} = 20.73$, $p < 0.05$] and a concord x distance x anteriority x laterality interaction [$F_{(2,64)} = 5.94$, $p < 0.05$]. Post-hoc pairwise comparisons revealed that the violation effect was significant in both conditions in the left-posterior [*adjacent*: $t_{(32)} = 3.36$, $p < 0.05$; *distal*: $t_{(32)} = 3.31$, $p < 0.05$], mid-posterior [*adjacent*: $t_{(32)} = 4.29$, $p < 0.05$; *distal*: $t_{(32)} = 3.03$, $p < 0.05$], right-posterior [*adjacent*: $t_{(32)} = 3.26$, $p < 0.05$; *distal*: $t_{(32)} = 3.63$, $p < 0.05$], and right-central [*adjacent*: $t_{(32)} = 2.59$, $p < 0.05$; *distal*: $t_{(32)} = 2.49$, $p < 0.05$] clusters of electrodes. Crucially, in the mid-central site, the concord violation effect showed up only in the distal condition [*adjacent*: $t_{(32)} = 1.69$, $p > 0.05$; *distal*: $t_{(32)} = 2.37$, $p = 0.05$].

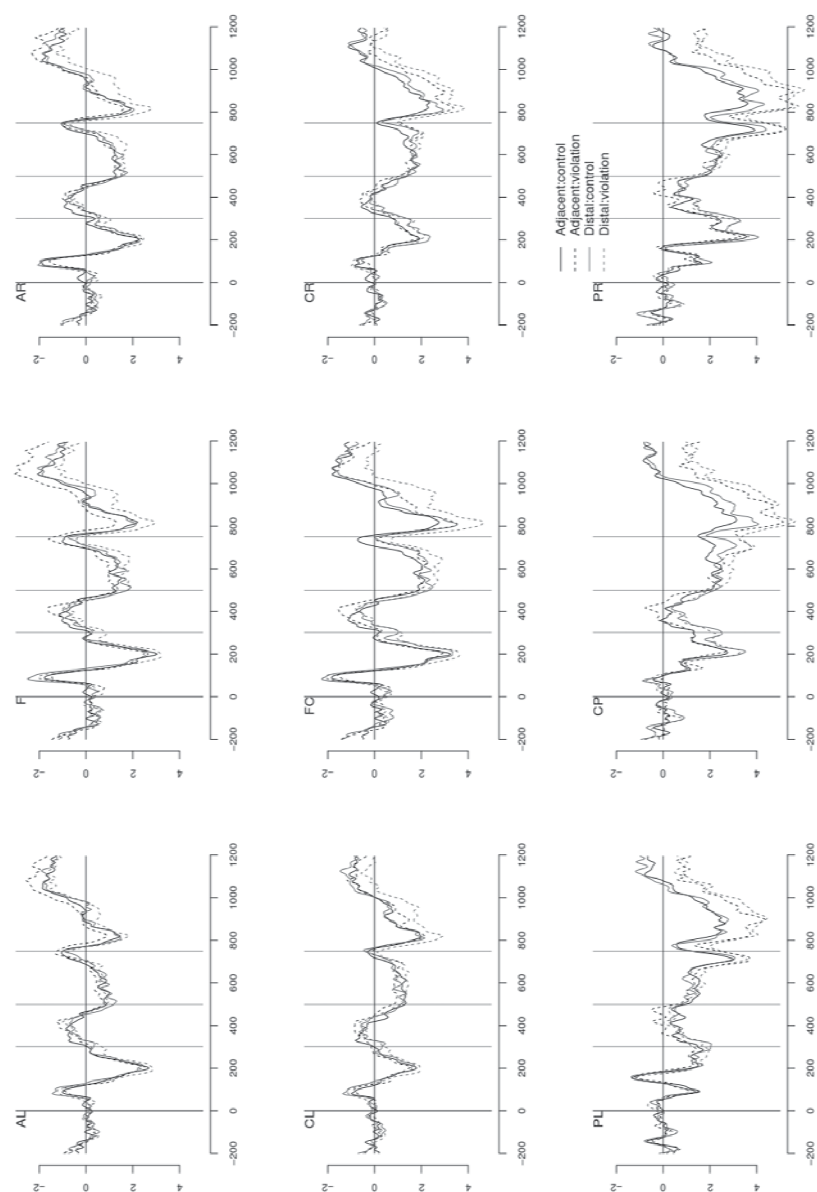


Figure 3 (previous page). Grand average ERPs time-locked to the onset of the target word in anterior-left (AL), frontal (F), anterior-right (AR), central-left (CL), fronto-central (FC), central-right (CR), posterior-left (PL), central-posterior (CP), posterior-right (PR) sites.

4.2.2 Sentence-final word

The average activity time-locked to the onset of the final word of the sentence, for each cluster of electrodes and each experimental condition, is represented in Figure 4.

In the 300-500ms time window, in the lateral sites, we found a main effect of concord [$F_{(1,32)} = 5.54, p < 0.05$] and distance [$F_{(1,32)} = 4.34, p < 0.05$], concord x distance [$F_{(1,32)} = 6, p < 0.05$], concord x anteriority [$F_{(2,64)} = 3.95, p < 0.05$], concord x anteriority x laterality [$F_{(2,64)} = 3.63, p < 0.05$] interactions and the 4-way interaction concord x distance x anteriority x laterality [$F_{(2,64)} = 8.64, p < 0.05$]. In the midline sites, we also found a main effect of concord [$F_{(1,32)} = 5, p < 0.05$] and distance [$F_{(1,32)} = 6, p < 0.05$], as well as the concord x distance [$F_{(1,32)} = 4.7, p < 0.05$] and concord x anteriority interaction [$F_{(2,64)} = 10.11, p < 0.05$]. The concord x distance x anteriority 3-way interaction failed to reach significance [$F_{(2,64)} = 1.76, p > 0.05$].

Post-hoc pairwise comparisons showed that the concord effect reached significance only in the adjacent conditions in left-anterior [*adjacent*: $t_{(32)} = 2.61, p < 0.05$; *distal*: $t_{(32)} = -0.66, p > 0.05$], left-central [*adjacent*: $t_{(32)} = 2.76, p < 0.05$; *distal*: $t_{(32)} = -0.08, p > 0.05$], left-posterior [*adjacent*: $t_{(32)} = 2.69, p < 0.05$; *distal*: $t_{(32)} = 0.38, p > 0.05$] sites, and in right-central [*adjacent*: $t_{(32)} = 2.93, p < 0.05$; *distal*: $t_{(32)} = -0.51, p > 0.05$] and right-posterior sites [*adjacent*: $t_{(32)} = 4.05, p < 0.05$; *distal*: $t_{(32)} = 0.98, p > 0.05$]. In the midline area, the effect of concord reached significance only in the adjacent condition [*adjacent*: $t_{(32)} = 2.75, p < 0.05$; *distal*: $t_{(32)} = -0.42, p > 0.05$], in posterior sites [$t_{(32)} = 4.01, p < 0.05$] and marginally in central sites [$t_{(32)} = 1.95, p = 0.06$].

In the 500-800ms time window, we found a main effect of concord [$F_{(1,32)} = 4.48, p < 0.05$] a concord x anteriority [$F_{(2,64)} = 5.17, p < 0.05$] and a concord x laterality interaction [$F_{(1,32)} = 4.43, p < 0.05$] in the lateral sites. Post-hoc t-tests showed that the effect of concord reached significance only in the lateral-central [$t_{(32)} = 2.35, p < 0.05$] and lateral-posterior sites [$t_{(32)} = 2.66, p < 0.05$], in particular in the right side of the scalp [$t_{(32)} = 2.52, p < 0.05$]. In the midline sites, only the concord x anteriority interaction reached significance [$F_{(2,64)} = 11.07, p < 0.05$]. Post-hoc t-tests

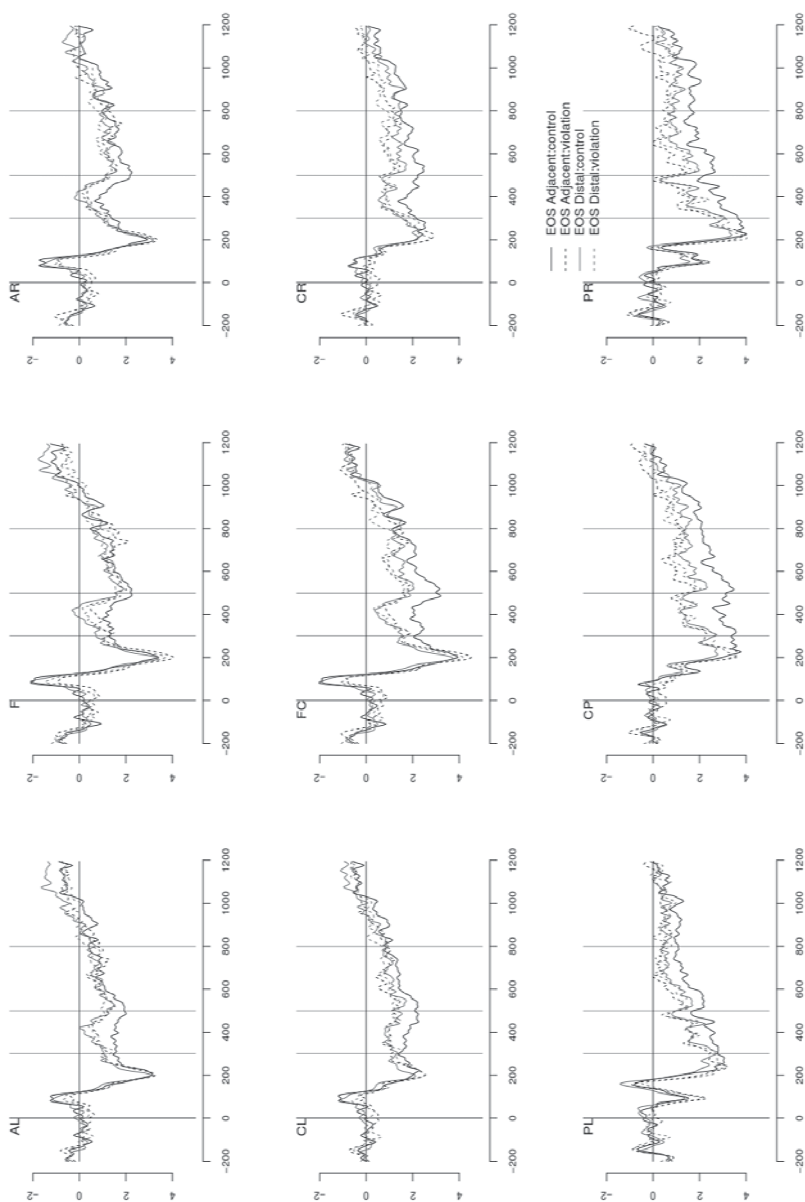


Figure 4 (previous page). Grand average ERPs time-locked to the onset of the sentence-final word in anterior-left (AL), frontal (F), anterior-right (AR), central-left (CL), fronto-central (FC), central-right (CR), posterior-left (PL), central-posterior (CP), posterior-right (PR) sites.

reported a significant effect of concord in posterior sites [$t_{(32)} = 3.15, p < 0.05$] and marginally in central sites [$t_{(32)} = 1.92, p = 0.06$].

In the 800-1000ms time window, the analysis in the lateral sites showed only a significant concord x laterality interaction [$F_{(1,32)} = 7.44, p < 0.05$]. Post-hoc t-tests showed that the effect of concord (for both violations) tends to be right-lateralized (right [$t_{(32)} = 1.91, p = 0.07$], left [$t_{(32)} = 0.73, p = 0.47$]). In the midline sites, we found a concord x anteriority interaction [$F_{(2,64)} = 5.17, p < 0.05$]. Post-hoc t-tests showed that the effect of concord is only present in mid-posterior sites [$t_{(32)} = 2.1, p < 0.05$].

5. Discussion

The main aim of this study was to enrich the current scarce ERP literature on the processing of temporal concord, and to test whether the heterogeneity in the ERP pattern found in previous studies was due to the distance between the adverb and the verb. Based on previous ERP literature and recent behavioral studies on the temporal concord, we expected an effect of distance to arise especially in early stages of processing. In particular, we expected the distal mismatch condition to give rise to an N400-like component, while we expected to find a LAN in the adjacent condition. We also considered the possibility that distance could affect later stages of processing, based on previous studies testing distance during the processing of other agreement relations. We expected a P600 to arise for both mismatch conditions and, based on previous findings (e.g., Rispens & Amesti, 2017), we hypothesized the distal mismatch condition to show a larger (early and late) P600 compared to the adjacent mismatch condition. Finally, we decided to analyze the average activity following the onset of the last word of the sentence, to test whether an SFN was present or not for both temporal violations. We expected to find an SFN for both violations in the case in which the presence of this component is strictly related to the judgement task we adopted. Conversely, we expected a more boosted SFN for the adjacent mismatch condition (compared to the distal mismatch condition) if this component mirrors the difficulty in the reanalysis/storage of a sentence where there is less time available to process the mismatch.

The analysis of the ERP data showed the following pattern of results. In early stages of processing no significant effects were found for either violations, although numerical trends are visible in the grand average ERPs (Figure 2). In particular, the plot showed a numerical trend towards a negativity arising around 380–450ms after the verb onset in central and left areas of the scalp, but only for adjacent violations. This null result (and even the direction of the numerical trend visible in the plot) was unexpected, for several reasons. First, all previous studies reported the presence of a negativity in the early time window¹¹, independently from the relative distance of the adverb and the verb. The lack of a significant effect here cannot be due to power issues. Our study has a sample size of 33, which is higher than the average sample size of 25 ($SD = 4.8$) used in the previous studies. A cross-linguistic explanation for this weak effect should be also excluded since the unique study testing adjacent temporal concord violations in Italian (De Vincenzi et al., manuscript) also reported an anterior (albeit right-lateralized) negativity. Second, the numerical trend visible for the adjacent violations seems to be more N400-like than LAN-like, based on its topographic distribution over central and posterior areas. However, previous studies where the adverb and the verb were adjacently located reported a LAN effect (Baggio, 2008; see also Steinhauer & Ullmann, 2002; Newman et al., 2007) or a RAN effect (De Vincenzi et al., manuscript), that is anterior negativities. The distal mismatch condition does not seem to show any effect or numerical trend in this study. Still, previous ERP studies reported N400-like responses (Qiu & Zhou, 2012; Dillon et al., 2012; see also Fonteneau et al., 1998), and previous eye-tracking studies (Biondo et al., 2018) reported early mismatch effects for this type of violations¹². The null result in the early time window does not allow us to make strong claims. We can only conclude that these findings

¹¹ Besides the two studies reporting a positivity (Dragoy et al., 2012, Bos et al., 2013) arguably for methodological issues, as discussed in the introduction.

¹² Clearly the discrepancy between the eye-tracking data from Biondo et al. (2018) and the ERP data could be related to other (methodological) factors. The Rapid Serial Visual Presentation method used in the ERP research, where each word is presented for a fixed amount of time, clearly differs from the more naturalistic sentence presentation of the eye-tracking while reading technique, where the participants are exposed to the whole sentence. Moreover, in order to match the length of the adverbial phrase and of the subject noun phrase, in the eye-tracking study the authors used a longer “intervener” between the adverb and the verb (a subject DP with an adjectival phrase e.g., ‘the tired travelers’). It is thus hard to define whether the readers had more time to process the adverb and to conclude the anchoring process in the eye-tracking study or in the ERP study, and thus in which of the two studies we should have expected “earlier” mismatch effects.

cannot provide evidence either for an early detection of temporal violations or for an early modulation of temporal concord processing as a function of distance.

Later in time, a significant effect of concord was found both in the adjacent and in the distal conditions. This effect took the form of a positive deflection arising 500ms after the mismatching verb onset, in mid-posterior areas. This effect matches the classic properties of the P600, and it is in line with previous ERP findings on temporal concord processing. Our results also showed that there is no reliable modulation of the early P600 as a function of distance. Conversely, distance plays a role in the late stage of the P600: a larger positivity was elicited by the distal violations compared to the adjacent ones. In general, these findings are in line with the study by Rispens & Amesti (2017) reporting larger P600 effects for distal compared to adjacent (subject-verb) agreement violations. In particular, these findings only partially match our predictions, since we expected larger P600 effects for the distal violations, both in the early and in the late stages. In Rispens and Amesti's study, the P600 effect was observed and analyzed in the 500-1000ms time window, so no distinction was made between the early and the late stage of the P600. In this study, we can add another piece of information about the effect of distance in the P600 modulation, since the effect was found to be significant only in the late stage of the P600. In sum, we can conclude that distance affects later stages of temporal concord processing, namely the revision processes (of discourse-related information) related to a violation.

The late P600 effect was maximal over the mid-central part of the scalp for distal violations while it was more posterior for the adjacent violations. The amplitude of the late P600 has been said to reflect the ease/difficulty with which the parser can revise/reanalyze the mismatch. In particular, violations that require the re-interpretation of representations that go beyond the level of morpho-syntax generally trigger larger and more frontally distributed P600 effects (e.g., see P600 effect for person violations in Mancini et al., 2011). In the introduction (and in previous work, i.e., Biondo et al., 2018), we proposed that the semantic/discourse-related interpretation of the adverb goes in an incremental way. In particular, we made this prediction: the more the time to process the adverb before encountering the verb, the richer/more complete the representation of the reference time of the sentence, the stronger the violation effect and the harder the re-interpretation processes related to the violation. The results coming from the late time windows in this study are in line with our predictions. In particular, the larger and more frontal P600

effect are here related to a more demanding revision process engaged in the distal mismatch condition (compared to the adjacent mismatch condition).

The analysis of the brain activity time-locked to the end of the sentence also showed interesting results. We found a main effect of concord in all the time windows. This effect was driven by the fact that the mismatch conditions, on average, showed a more negative activity than the control conditions. This result is in line with the idea that the SFN is present or “boosted” when a metalinguistic task has to be performed (Stowe et al., 2018). Interestingly, we also found an interaction between distance and concord in the early time window (300-500ms). The interaction was driven by the fact that only the two adjacent conditions differed (while the two distal conditions showed similar activity). Under the assumption that SFN reflects difficulties in the analysis and storage of the sentence (Stowe et al., 2018), we considered the possibility of finding a larger SFN in the adjacent condition, that is when the system has less time to fully process the adverb (and the adverb-verb mismatch). Our findings thus seem to be in line with our prediction, since data showed that the adjacent mismatch condition triggered a negativity compared to the adjacent control condition. Still, if we give a closer look to the activity of the four conditions, we can see that the condition that clearly detached from the others was the adjacent control condition, which showed a more positive activity compared to the other three conditions. In other words, the difference in the processing of the two adjacent conditions can be seen in two different ways. We can either say that the activity of the adjacent mismatch condition is more negative than the activity of the adjacent match condition, or that the activity of the adjacent match condition is more positive than the activity of the adjacent mismatch condition. Why should the adjacent control condition differ from the others and elicit positive activity at the end of the sentence? Sentence final positivities have been reported before and they have been mainly related to structural/syntactic wrap-up, while any relation with semantic integration costs has been discarded (see Stowe et al., 2018 for an overview). If this positivity was related to some kind of structural difficulty, we would have found it for both adjacent conditions, since they have exactly the same syntactic structure. The only measure in which the adjacent condition differed from the other conditions, besides the sentence-final ERPs, is the reaction times of the acceptability judgement task. In this task, the adjacent condition shows numerically longer RTs compared to the other conditions, as if the identification of the un/grammaticality of this condition is harder to be defined compared to the other conditions. If the

behavioral results and the sentence-final ERPs are in some way related, the positivity could mirror a difficulty in the final “grammatical evaluation” of this condition.

These findings are related to sentence processing routines performed by speakers with unimpaired linguistic abilities, but they can be extremely relevant for the investigation of temporal concord processing in clinical populations. For example, in previous studies aphasic patients were asked to perform grammaticality judgement tasks on either adjacent temporal concord violations in German (e.g., ‘**Tomorrow stood many topics on the agenda*’; in Wenzlaff & Clahsen, 2004) or in Greek (e.g., ‘**The parents yesterday will leave early from the house*’; in Nanousi et al., 2006) or on distal temporal concord violations in English (e.g., ‘**Next year, my sister lived in New Hampshire*’; in Farooqi-Shah & Dickey, 2009) or in Greek (e.g., ‘**Yesterday Popi will watch TV*’; in Varlokosta et al., 2006). Our data show that adverb – verb distance is a factor that should be taken into account when comparing patterns of results coming from different experimental studies, since adjacent and distal temporal concord processing differ in online sentence reading.

Moreover, although our findings refer to reading for (comprehension and) grammatical evaluation, we cannot exclude that the effect of distance also applies to production tasks. For example, in previous sentence completion studies, Friedmann & Grodzinsky (1997), Nanousi et al. (2006) and Varlokosta et al. (2006) tested temporal concord in sentences where the adverb and the verb were distally located in Hebrew and Greek respectively (e.g., ‘*Tomorrow the boy ...*’; ‘*Tomorrow we ...*’; ‘*Yesterday Popi ...*’), while Wenzlaff & Clahsen (2004), Burchert, Swoboda-Moll & De Bleser (2005) and Fyndanis et al. (2018) tested adjacent temporal concord respectively in German, in Greek and in Italian (e.g., ‘*Last month ... his plan*’; ‘*Tomorrow ... I the director*’; ‘*The gardener tomorrow ...*’). This manipulation was arguably related to the type of language under investigation: while in null-subject languages such as Italian, Greek or Spanish, or in V2 languages such as German and Dutch, adjacent adverb-verb configurations are allowed, in languages such as English and French only distal temporal concord configurations can be tested. Further research in languages with a relatively free word order, such as Italian, should thus be conducted in order to test how aphasic patients are affected by different adverb – verb distance patterns during the processing of temporal concord, in a strictly controlled paradigm as in this study.

Recent accounts claim that aphasic patients and healthy speakers adopt similar processing mechanisms; aphasia only exacerbates patterns or trends observed in healthy individuals (e.g., Fyndanys et al. 2018; see also Garraffa, 2007; Dick et al., 2001; Miyake, Carpenter & Just, 1994). Under this assumption, our findings can be extremely informative to predict how aphasic patients would be affected by distance. The reanalysis process of (distal) temporal concord violations should be even harder to be handled by aphasic patients.

Still, there are several questions that we could not address in this study and that should be addressed in the future. These questions mainly deal with the linguistic factors that could explain why the presence and topography of early ERP components is so heterogeneous across studies. As claimed in the introduction, the processing of non-primary relations such as the temporal concord can be influenced by numerous syntactic, semantic and pragmatic factors. In this study, we tried to address the issue by investigating the role of adverb-verb distance during the processing of temporal violations, and we can add a small piece to the puzzle: our data did not provide evidence for a modulation of early ERP components as a function of distance; yet, distance was responsible for the differences in the amplitude of later ERP components (P600, SFN).

Other relevant factors that should be taken into account and tested in future research are: the intrinsic differences in the processing of past and non-past information, the morphological realization of tense features, and cross-linguistic variability. In many studies, as in ours, past and non-past tense were collapsed across conditions. However, there is evidence suggesting that past and non-past processing may differ. For example, studies showed that violations of present tensed verb trigger a LAN followed by a P600 (Baggio, 2008) or just a P600 (Bos et al., 2013) compared to the control condition. Conversely, the detection of violations on a past tensed verb takes longer and does not trigger any reliable, different pattern compared to its correct counterpart, apart from an SFN (Dragoy et al., 2012)¹³. It could thus be possible that only half of the sentences (i.e., violations of the present verb) led to an early effect and when past and non-past items were averaged, the mismatch effect became

¹³ Overall, these findings are in line with the so-called Past Discourse Linking Hypothesis (Bastiaanse, Bamyaci, Hsu, Lee, Duman & Thompson, 2011; Bastiaanse, 2013), which stipulates that the processing of past time reference is more difficult compared to the processing of non-past time reference since only in the first case a link to the discourse needs to be established, while in the second case the time of the event is locally bound so it is easier to be interpreted.

weaker. In this study we did not use present tense, but future tense. However, we cannot exclude that future verbs behave as present verbs in this respect. More recently, Biondo, Soilemeizidi & Mancini (in press) tested the processing of past and future adverb-verb violations in an eye-tracking study in Spanish. Their data showed that future verbs mismatching a past temporal adverb gave rise to early mismatch effects while past verbs mismatching a future temporal adverb gave rise to mismatch effects only in later measures, in line with previous ERP research testing past/present violations.

Another issue to be taken into account is that tense features can be expressed both through regular and irregular verb forms. Both Steinhauer & Ullmann (2002) and Newman et al. (2007) manipulated the morphological realization of tense features in English, by testing past/present temporal violations in regular and irregular verb forms (as shown in Table 2). While the first study reported similar results in the two conditions (LAN-P600), Newman and colleagues did find a difference in the processing of regular and irregular verbs. In particular, the violation of regular tensed verbs triggered a LAN while the violation of irregular tensed verbs triggered a more posterior negativity. Whether this pattern applies only to English verbs or cross-linguistically is unclear. For example, Baggio (2008) found a LAN effect for tense violations in (collapsed) regular and irregular verb forms in Dutch. In the current study, almost all regular verb forms were used (only the 18% of the items contained irregular verb forms and they were equally presented in the four different conditions, following a Latin square design).

The last factor we want to discuss is cross-linguistic variability. Barber & Carreiras (2005) showed that the topographic distribution of the early negativity for agreement mismatches in Spanish word pairs can vary depending on the grammatical class of the words involved. The more the lexical-semantic information processed, the more the N400-like effects observed in response to an agreement error. This could be the case of Qiu & Zhou's (2012) study in Mandarin Chinese, where temporal violations were triggered by lexical entities (*jiangyao/cengjing*) preceding the verb and encoding past/future information. However, this would not explain why N400-like responses were found in response to tense violations triggered by verb suffixes in Dillon et al.'s study (2012) in Hindi. Future studies in languages different from English are needed to test whether the difference in the processing of temporal concord is related to cross-linguistic differences. In particular, future research could aim at disentangling whether the ERP components related to past and future

verbs significantly differ and whether other factors, such as verb form regularity, play any role, cross-linguistically.

Authors contribution

N.B. wrote the entire manuscript and was responsible for the conception and the design of the work, as well as for the acquisition, preprocessing, analysis and interpretation of the behavioral/EEG data. E.B. was responsible for the creation of the experimental material and for the acquisition of the behavioral/EEG data. F.V. was responsible for the design of the work, as well as for the acquisition and interpretation of the behavioral/EEG data.

References

- Acunzo, D. J., MacKenzie, G., & van Rossum, M. C. (2012). Systematic biases in early ERP and ERF components as a result of high-pass filtering. *Journal of neuroscience methods*, 209(1), 212-218.
- Alemán Bañón, J., & Rothman, J. (2019). Being a Participant Matters: Event-Related Potentials Show That Markedness Modulates Person Agreement in Spanish. *Frontiers in psychology*, 10, 746.
- Alexiadou, A. (1997). *Adverb placement: A case study in antisymmetric syntax* (Vol. 18). Amsterdam: John Benjamins.
- Alexiadou, A. (2000). On the syntax of temporal adverbs and the nature of Spec, TP: 1888. *Rivista di linguística*, 12(1), 55-76.
- Baggio, G. (2008). Processing temporal constraints: An ERP study. *Language Learning*, 58(s1), 35-55.
- Barber, H., & Carreiras, M. (2005). Grammatical gender and number agreement in Spanish: An ERP comparison. *Journal of Cognitive Neuroscience*, 17(1), 137-153.
- 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(3), 255-278.
- Bastiaanse, R. (2013). Why reference to the past is difficult for agrammatic speakers. *Clinical linguistics & phonetics*, 27(4), 244-263.
- Bastiaanse, R., Bamyaci, E., Hsu, C. J., Lee, J., Duman, T. Y., & Thompson, C. K. (2011). Time reference in agrammatic aphasia: A cross-linguistic study. *Journal of Neurolinguistics*, 24(6), 652-673.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

- Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.
- Belletti, A. (1990). *Generalized verb movement: Aspects of verb syntax*. Torino: Rosenberg & Sellier.
- Belletti, A., Guasti M.T., (2015). *The acquisition of Italian*. Amsterdam: John Benjamins.
- Berkum, J. J. V., Hagoort, P., & Brown, C. M. (1999). Semantic integration in sentences and discourse: Evidence from the N400. *Journal of cognitive neuroscience*, 11(6), 657-671.
- Bianchi, V. (2003). On finiteness as logophoric anchoring. *Temps et point de vue/Tense and point of view*, 213-246.
- Bianchi, V. (2006). On the syntax of personal arguments. *Lingua*, 116(12), 2023-2067.
- Biondo, N. & Mancini, S. (2019). The grammaticalization of different relations during adult second language (L2) acquisition. Poster presented at the 11th Annual Meeting of the Society for the Neurobiology of Language (SNL), Helsinki, Finland.
- Biondo, N., Soilemezidi, M. & Mancini, S. (in press). Yesterday is history, tomorrow is a mystery: an eye-tracking investigation of the processing of past and future time reference during sentence reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. DOI: 10.1037/xlm0001053
- Biondo, N., Vespignani, F., Rizzi, L., & Mancini, S. (2018). Widening agreement processing: a matter of time, features and distance. *Language, Cognition and Neuroscience*, 33(7), 890-911.
- Bornkessel, I., & Schlesewsky, M. (2006). The extended argument dependency model: a neurocognitive approach to sentence comprehension across languages. *Psychological review*, 113(4), 787.
- Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2008). An alternative perspective on “semantic P600” effects in language comprehension. *Brain research reviews*, 59(1), 55-73.
- Bos, L. S., Dragoy, O., Stowe, L. A., & Bastiaanse, R. (2013). Time reference teased apart from tense: Thinking beyond the present. *Journal of Neurolinguistics*, 26(2), 283-297.
- Burchert, F., Swoboda-Moll, M., & De Bleser, R. (2005). Tense and agreement dissociations in German agrammatic speakers: Underspecification vs. hierarchy. *Brain and Language*, 94(2), 188-199.
- Caffarra, S., Mendoza, M., & Davidson, D. (2019). Is the LAN effect in morphosyntactic processing an ERP artifact?. *Brain and Language*, 191, 9-16.

- Carreiras, M., & Clifton, C. (1993). Relative clause interpretation preferences in Spanish and English. *Language and Speech*, 36(4), 353-372.
- Carreiras, M., Salillas, E., & Barber, H. (2004). Event-related potentials elicited during parsing of ambiguous relative clauses in Spanish. *Cognitive Brain Research*, 20(1), 98-105.
- Chomsky, N. (1981). *Lectures on Government and Binding: The Pisa Lectures*. Dordrecht: Foris.
- Chomsky, N. (1986). *Barriers* (Vol. 13). Cambridge, MA: MIT press.
- Chomsky, N. (1995). *The minimalist program*. Cambridge, MA: MIT press.
- Chomsky, N. (2000). Minimalist inquiries. In *Step by step: Essays on minimalism in honor of Howard Lasnik*. Cambridge, MA: MIT press, 83-155.
- Chomsky, N. (2001) 'Derivation by Phase.' In M. Kenstowicz (ed.) *Ken Hale: A Life in Language*. Cambridge, MA: MIT Press, 1-53.
- Chow, W. Y., Lau, E., Wang, S., & Phillips, C. (2018). Wait a second! Delayed impact of argument roles on on-line verb prediction. *Language, Cognition and Neuroscience*, 33(7), 803-828.
- Cinque, G. (1999). *Adverbs and functional heads: A cross-linguistic perspective*. Oxford: Oxford University Press.
- Cinque, G. (2004). Issues in adverbial syntax. *Lingua*, 114(6), 683-710.
- Clahsen, H., & Ali, M. (2009). Formal features in aphasia: tense, agreement, and mood in English agrammatism. *Journal of Neurolinguistics*, 22(5), 436-450.
- Corbett, G. G. (2003). Agreement: terms and boundaries. In *The Role of Agreement in Natural Language. Proceedings of the 2001 Texas Linguistic Society Conference*, Austin, TX. (pp. 109-122).
- Coulson, S., King, J. W., & Kutas, M. (1998). Expect the unexpected: Event related brain response to morphosyntactic violations. *Language and Cognitive Processes*, 13, 21-58.
- Cuetos, F., & Mitchell, D. C. (1988). Cross-linguistic differences in parsing: Restrictions on the use of the Late Closure strategy in Spanish. *Cognition*, 30(1), 73-105.
- De Vincenzi, M., & Job, R. (1993). Some observations on the universality of the late-closure strategy. *Journal of Psycholinguistic Research*, 22(2), 189-206.
- De Vincenzi, M., Job, R., Di Matteo, R., Angrilli, A., Penolazzi, B., Ciccarelli, L., & Vespignani, F. (2003). Differences in the perception and time course of syntactic and semantic violations. *Brain and Language*, 85(2), 280-296.

- De Vincenzi, M., Rizzi, L., Portolan, D., Di Matteo, R., Spitoni, G., & Di Russo, F. (unpublished). Mapping the language: A reading time and topographic ERP study on tense, agreement, and Aux-V violations. Manuscript, University of Chieti.
- Dick, F., Bates, E., Wulfeck, B., Utman, J. A., Dronkers, N., & Gernsbacher, M. A. (2001). Language deficits, localization, and grammar: evidence for a distributive model of language breakdown in aphasic patients and neurologically intact individuals. *Psychological review*, 108(4), 759.
- Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, 69(2), 85-103.
- Dillon, B., Nevins, A., Austin, A. C., & Phillips, C. (2012). Syntactic and semantic predictors of tense in Hindi: An ERP investigation. *Language and Cognitive Processes*, 27(3), 313-344.
- Dragoy, O., Stowe, L. A., Bos, L. S., & Bastiaanse, R. (2012). From time to time: Processing time reference violations in Dutch. *Journal of Memory and Language*, 66(1), 307-325.
- Enç, M. (1987). Anchoring conditions for tense. *Linguistic inquiry*, 633-657.
- Farooqi-Shah, Y., & Dickey, M. W. (2009). On-line processing of tense and temporality in agrammatic aphasia. *Brain and Language*, 108(2), 97-111.
- Fiebach, C. J., Schlesewsky, M., & Friederici, A. D. (2002). Separating syntactic memory costs and syntactic integration costs during parsing: The processing of German wh-questions. *Journal of Memory and Language*, 47(2), 250-272.
- Fonteneau, E., Frauenfelder, U. H., & Rizzi, L. (1998). On the contribution of ERPs to the study of language comprehension. *Bulletin suisse de linguistique appliquée*, (68), 111-124.
- Frazier, L. (1987). Sentence processing. A tutorial review. In M. Coltheart (Ed.), *Attention and performance XII. The psychology of reading*, pp. 559-586. Mahwah, NJ: Lawrence Erlbaum.
- Frazier, L., & Clifton, C. (1996). *Construal*. Cambridge, MA: MIT Press.
- Frazier, L., & Clifton, Jr. (1998). Sentence reanalysis and visibility. In J. D. Fodor & F. Ferreira F. (Eds.) *Reanalysis in sentence processing*. pp. 143-177. Dordrecht: Kluwer.
- Frenck-Mestre, C., Osterhout, L., McLaughlin, J., & Foucart, A. (2008). The effect of phonological realization of inflectional morphology on verbal agreement in French: Evidence from ERPs. *Acta Psychologica*, 128(3), 528-536.

- Friederici, A. D. (1995). The time course of syntactic activation during language processing: A model based on neuropsychological and neurophysiological data. *Brain and Language*, 50, 259–281.
- Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in cognitive sciences*, 6(2), 78–84.
- Friederici, A. D. (2011). The brain basis of language processing: from structure to function. *Physiological reviews*, 91(4), 1357–1392.
- Friederici, A. D., Hahne, A., & Mecklinger, A. (1996). Temporal structure of syntactic parsing: early and late event-related brain potential effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(5), 1219.
- Friedmann, N. A., & Grodzinsky, Y. (1997). Tense and agreement in agrammatic production: Pruning the syntactic tree. *Brain and Language*, 56(3), 397–425.
- Fyndanis, V., Arcara, G., Capasso, R., Christidou, P., De Pellegrin, S., Gandolfi, M. & Miceli, C. (2018). Time reference in nonfluent and fluent aphasia: A cross-linguistic test of the PAST DIscourse LINKing Hypothesis. *Clinical Linguistics & Phonetics*, 32(9), 823–843.
- Garraffa, M. (2007). Impoverishment of grammatical features in a non fluent aphasic speaker: the grammatical nature of minimal structures. PhD. Thesis, University of Siena.
- Gouvea, A.C., Phillips, C., Kazanina, N. & Poeppel, D. (2010). The linguistic processes underlying the P600. *Language and Cognitive Processes*, 25(2), 149–188.
- Hagoort, P. (2003). How the brain solves the binding problem for language: a neurocomputational model of syntactic processing. *Neuroimage*, 20, S18–S29.
- Hagoort, P. (2013). MUC (memory, unification, control) and beyond. *Frontiers in Psychology*, 4, 416.
- Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and Cognitive Processes*, 8(4), 439–483.
- Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4), 434–446.
- Jäger, L. A., Engelmann, F., & Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language*, 94, 316–339.
- Jäger, L. A., Mertzen, D., Van Dyke, J. A., & Vasishth, S. (2020). Interference patterns in subject-verb agreement and reflexives

- revisited: A large-sample study. *Journal of Memory and Language*, 111, 104063.
- Kaan, E. (2002). Investigating the Effects of Distance and Number Interference in Processing Subject-Verb Dependencies: An ERP Study. *Journal of Psycholinguistic Research*, 31(2), 165-193.
- Kaan, E., Swaab, T.Y.Y., (2003). Repair, revision, and complexity in syntactic analysis: an electrophysiological differentiation. *Journal Cognitive Neuroscience*, 15, 98–110.
- Kaan, E., Harris, A., Gibson, E., & Holcomb, P. (2000). The P600 as an index of syntactic integration difficulty. *Language and Cognitive Processes*, 15(2), 159-201.
- Kasparian, K., Vespignani, F., & Steinhauer, K. (2017). First Language Attrition Induces Changes in Online Morphosyntactic Processing and Re-Analysis: An ERP Study of Number Agreement in Complex Italian Sentences. *Cognitive Science*, 41(7), 1760-1803.
- Kayne, R. S. (1994). *The antisymmetry of syntax* (No. 25). Cambridge, MA: MIT Press.
- Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language*, 52(2), 205-225.
- King, J. W., & Kutas, M. (1995). Who did what and when? Using word and clause-level ERPs to monitor working memory usage in reading. *Journal of Cognitive Neuroscience*, 7, 376–395.
- Kluender, R., & Kutas, M. (1993a). Bridging the gap: Evidence from ERPs on the processing of unbounded dependencies. *Journal of Cognitive Neuroscience*, 5, 196–214.
- Kreiner, H., Garrod, S., & Sturt, P. (2013). Number agreement in sentence comprehension: The relationship between grammatical and conceptual factors. *Language and Cognitive Processes*, 28(6), 829–874.
- Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, 1146, 23-49.
- Kuperberg, G. R., Sitnikova, T., Caplan, D., & Holcomb, P. J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cognitive Brain Research*, 17(1), 117-129.
- Kuperberg, G. R., Caplan, D., Sitnikova, T., Eddy, M., & Holcomb, P. J. (2006). Neural correlates of processing syntactic, semantic, and thematic relationships in sentences. *Language and Cognitive Processes*, 21(5), 489-530.
- Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427), 203-205.

- Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947), 161.
- Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics:(de) constructing the N400. *Nature Reviews Neuroscience*, 9(12), 920.
- Mancini, S. (2018). *Features and Processing in Agreement*. Newcastle: Cambridge Scholars Publishing.
- Mancini, S., Molinaro, N., & Carreiras, M. (2013). Anchoring agreement in comprehension. *Language and Linguistics Compass*, 7(1), 1–21.
- Mancini, S., Molinaro, N., Rizzi, L., & Carreiras, M. (2011). A person is not a number: Discourse involvement in subject–verb agreement computation. *Brain Research*, 1410, 64–76.
- Mancini, S., Postiglione, F., Laudanna, A., & Rizzi, L. (2014). On the person-number distinction: Subject-verb agreement processing in Italian. *Lingua*, 146, 28–38.
- Martín-Loeches, M., Muñoz, F., Casado, P., Melcon, A., & Fernández-Frías, C. (2005). Are the anterior negativities to grammatical violations indexing working memory?. *Psychophysiology*, 42(5), 508–519.
- Miyake, A., Carpenter, P. A., & Just, M. A. (1994). A capacity approach to syntactic comprehension disorders: Making normal adults perform like aphasic patients. *Cognitive neuropsychology*, 11(6), 671–717.
- Molinaro, N., Barber, H. A., & Carreiras, M. (2011). Grammatical agreement processing in reading: ERP findings and future directions. *Cortex*, 47(8), 908–930.
- Molinaro, N., Barber, H. A., Caffarra, S., & Carreiras, M. (2015). On the left anterior negativity (LAN): The case of morphosyntactic agreement. *Cortex*, 66(156–159).
- Molinaro, N., Vespignani, F., & Job, R. (2008). A deeper reanalysis of a superficial feature: An ERP study on agreement violations. *Brain Research*, 1228, 161–176.
- Nanousi, V., Masterson, J., Druks, J., & Atkinson, M. (2006). Interpretable vs. uninterpretable features: Evidence from six Greek-speaking agrammatic patients. *Journal of Neurolinguistics*, 19(3), 209–238.
- Nevins, A., Dillon, B., Malhotra, S., & Phillips, C. (2007). The role of feature-number and feature-type in processing Hindi verb agreement violations. *Brain Research*, 1164, 81–94.
- Newman, A. J., Ullman, M. T., Pancheva, R., Waligura, D. L., & Neville, H. J. (2007). An ERP study of regular and irregular English past tense inflection. *NeuroImage*, 34(1), 435–445.

- Nieuwland, M. S., & Van Berkum, J. J. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, 18(7), 1098-1111.
- O'Rourke, P. L., & Van Petten, C. (2011). *Brain research*, 1392, 62-79.
- Osterhout, L., & Holcomb, P. J. (1995). Event-related potentials and language comprehension. *Electrophysiology of Mind*, 171-215.
- Osterhout, L., Holcomb, P. J., & Swinney, D. A. (1994). Brain potentials elicited by garden-path sentences: evidence of the application of verb information during parsing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(4), 786.
- Osterhout, L., & Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. *Journal of Memory and Language*, 34(6), 739-773.
- Patel, A. D. (2003). Language, music, syntax and the brain. *Nature Neuroscience*, 6(7), 674.
- Patel, A. D., Gibson, E., Ratner, J., Besson, M., & Holcomb, P. J. (1998). Processing syntactic relations in language and music: An event-related potential study. *Journal of Cognitive Neuroscience*, 10(6), 717-733.
- Phillips, C., Kazanina, N., & Abada, S. H. (2005). ERP effects of the processing of syntactic long-distance dependencies. *Cognitive Brain Research*, 22(3), 407-428.
- Phillips, C., Wagers, M. W., & Lau, E. F. (2011). Grammatical illusions and selective fallibility in real-time language comprehension. *Experiments at the Interfaces*, 37, 147-180.
- Pollock, J. Y. (1989). Verb movement, universal grammar, and the structure of IP. *Linguistic Inquiry*, 20(3), 365-424.
- Qiu, Y., & Zhou, X. (2012). Processing temporal agreement in a tenseless language: An ERP study of Mandarin Chinese. *Brain Research*, 1446, 91-108.
- Rispens, J., & de Amesti, V. S. (2017). What makes syntactic processing of subject-verb agreement complex? The effects of distance and additional agreement features. *Language Sciences*, 60, 160-172.
- Rizzi, L., & Cinque, G. (2016). Functional categories and syntactic theory. *Annual Review of Linguistics*, 2, 139-163.
- Sagarra, N., & Han, Z. (2008). Working memory and L2 processing of redundant grammatical forms. *Understanding Second Language Process*, 25, 133-147.
- Shen, E. Y., Staub, A., & Sanders, L. D. (2013). Event-related brain potential evidence that local nouns affect subject-verb agreement processing. *Language and Cognitive Processes*, 28(4), 498-524.

- Shlonsky, U. (2010). The cartographic enterprise in syntax. *Language and Linguistics Compass*, 4(6), 417-429.
- Sigurðsson, H. Á. (2004). The syntax of Person, Tense, and speech features. *Rivista di Linguistica-Italian Journal of Linguistics*, 16(1), 219-251.
- Sigurðsson, H. Á. (2016). The split T analysis. *Finiteness matters: on finiteness-related phenomena in natural languages*, 231, 79-92.
- Smith, C. S. (1978). The syntax and interpretation of temporal expressions in English. *Linguistics and Philosophy*, 2(1), 43-99.
- Smith, C. S. (1981). Semantic and Syntactic Constraints on Temporal Interpretation. In P. Tedeschi and A. Zaenen (eds.) *Syntax and Semantics. Tense and Aspect*. New York, NY: Academic Press, 14, 213-237.
- Sportiche, D. (1988). A theory of floating quantifiers and its corollaries for constituent structure. *Linguistic Inquiry*, 19(3), 425-449.
- Steinhauer, K., & Ullman, M. T. (2002, October). Consecutive ERP effects of morpho-phonology and morpho-syntax. *Brain and Language*, 83(1), 62-65.
- Stowe, L. A., Kaan, E., Sabourin, L., & Taylor, R. C. (2018). The sentence wrap-up dogma. *Cognition*, 176, 232-247.
- Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, 48(3), 542-562.
- Sybesma, R. (2007). Whether we tense-agree overtly or not. *Linguistic Inquiry*, 38(3), 580-587.
- Tanner, D. (2015). On the left anterior negativity (LAN) in electrophysiological studies of morphosyntactic agreement: a commentary on "Grammatical agreement processing in reading: ERP findings and future directions" by Molinaro et al., 2014. *Cortex*, 66, 149.
- Tanner, D., Morgan-Short, K., & Luck, S. J. (2015). How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition. *Psychophysiology*, 52(8), 997-1009.
- Tanner, D., & Van Hell, J. G. (2014). ERPs reveal individual differences in morphosyntactic processing. *Neuropsychologia*, 56, 289-301.
- Van De Meerendonk, N., Kolk, H. H., Vissers, C. T. W., & Chwilla, D. J. (2010). Monitoring in language perception: mild and strong conflicts elicit different ERP patterns. *Journal of Cognitive Neuroscience*, 22(1), 67-82.

- Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, 83(2), 176-190.
- Varlokosta, S., Valeonti, N., Kakavoulia, M., Lazaridou, M., Economou, A., & Protopapas, A. (2006). The breakdown of functional categories in Greek aphasia: Evidence from agreement, tense, and aspect. *Aphasiology*, 20(8), 723-743.
- Weist, R.M., (2014). Future temporal reference in child language. In P. De Brabanter, M. Kissine and S. Sharifzadeh (Eds.) *Future Times, future Tenses* (pp.87-113). Oxford: Oxford University Press.
- Wenzlaff, M., & Clahsen, H. (2004). Tense and agreement in German agrammatism. *Brain and Language*, 89(1), 57-68.
- Zawiszewski, A., Santesteban, M., & Laka, I. (2016). Phi-features reloaded: An event-related potential study on person and number agreement processing. *Applied Psycholinguistics*, 37(3), 601-626.

EXPLORING THE SYNTAX-PROSODY INTERFACE IN CHILDREN WITH DEVELOPMENTAL DYSLEXIA

MARTINA CACCIA & MARIA LUISA LORUSSO¹

Abstract

The study investigates the relationship between prosody and syntax in a group of children with developmental dyslexia (DD) and typical development (TD), age 11-14 years. The children performed a picture matching task, after listening to a recorded sentence with an ambiguous syntactic structure disambiguated through prosody. Data show significant differences between groups in accuracy, depending on the type of sentence. Specifically, sentences where most elements can be assigned to different grammatical function depending on prosody are more difficult to process for children with DD than sentences with other types of syntactic ambiguity. DD children chose distractors more often than TD children in such complex psycholinguistic structures and prosody seems not to help them disambiguate the sentence. Since prosody in other sentence types is effectively used to disambiguate their meaning, the problem appears to be non-phonological and to relate to the syntax-prosody interface.

1. Introduction

The Report of the Task Force on Dyslexia (2001) suggests the following definition: “Dyslexia is manifested in a continuum of specific learning difficulties related to the acquisition of basic skills in reading, spelling and/or writing, such difficulties being unexplained in relation to an individual’s other abilities and educational experiences” (p.31). Moreover, dyslexia is a neurobiological condition with a genetic basis (Peterson &

¹ Corresponding author: Maria Luisa Lorusso, Scientific Institute IRCCS E. Medea, Unit of Child Psychopathology - Neurodevelopmental Disorders of Language and Learning, Bosisio Parini, Italy.

Pennington, 2012; Siegel, 2006). There are a number of hypotheses for the origin of dyslexia but many researchers now converge on the idea that several causes and factors (sometimes mutually related, sometimes only co-occurring) result in the emergence of reading disorders (Pennington, 2006). Among these factors, deficits in auditory processing of incoming stimuli deserve special attention because their influence is exerted at very early stages of language development (starting from Tallal, 1980 onward, Ramus & Szenkovits, 2008, etc.). Furthermore, several recent studies (Corriveau & Goswami, 2009; Goswami, 2011; Peterson & Pennington, 2012; Thomson, Leong, & Goswami, 2013) stress the importance of rhythm perception for language development.

Speech and music make use of structured patterns of pitch, duration, and intensity and these elements contribute to create the supra-segmental aspects of speech, known as prosody. Prosody is necessary to convey a number of different things, for example, lexical stress, focus, some aspects of meaning and emotion (Nootenboom, 1997; Ladd, 2008; Peppé, 2009 etc.).

Words, just like musical notes, are grouped together into phrases by their rhythmic and durational properties (Patel, 2003, 2008; Frazier, Carlson, & Clifton, 2006). These properties, called ‘prosodic phrasing’ and temporal grouping, seem to affect comprehension of sentences (Frazier et al 2006; Geiser, Kjelgaard, Christodoulou, Cyr, & Gabrieli, 2014). Moreover, recent studies suggest that the global pattern of prosodic phrasing is central in sentence comprehension (Frazier et al., 2006).

In auditory presentation, prosodic structure can guide syntactic parsing. Prosodic boundaries can also disambiguate the interpretation of many constituent structure ambiguities in spoken utterances, such as *I met the wife of the doctor who was in the room*: this sentence can be parsed as either “[I met the wife] of [the doctor who (= the doctor) was in the room]” or “[I met the wife of the doctor] [who (= the wife) was in the room]” and the correct parsing is suggested by prosodic contour (Carroll & Slowiczek, 1987; Price, Ostendorf, Shattuck-Hufnagel, & Fong, 1991; Warren, Grabe, & Nolan, 1995, Kjelgaard & Speer, 1999; Roncaglia-Denissen, Schmidt-Kassow, & Kotz, 2013).

There is evidence of a systematic relationships between prosody and syntax in terms of both processing (Price, Ostendorf, Shattuck-Hufnagel, & Fong, 1991; Patel, 2003; Heffner & Slevc, 2015) and brain activation

(Patel, Gibson, Ratner, Besson, & Holcomb, 1998; Sammler, Kotz, Eckstein, Ott, & Friederici, 2010; Kreiner & Eviatar, 2014).

As far as prosody in language acquisition is concerned, toddlers and young pre-schoolers show a clear understanding of the syntactic function of prosody (Snow 1994; Trueswell, Sekerina, Hill, & Logrip, 1999; Snedeker & Yuan 2008). Furthermore, children with developmental dyslexia have been found to be impaired in auditory timing perception, especially in temporal grouping of prosodic phrase boundaries, even if they correctly identify the prosodic phrase boundaries (Goswami, Gerson, & Astruc, 2010; Huss, Verney, Fosker, Mead, & Goswami 2011; Geiser et al., 2014).

There are just a few studies on the relationship between prosody and syntax in children with a reading and/or language impairment. Marshall, Harcourt-Brown, Ramus, and van der Lely (2009) suggest that children with Developmental Language Disorders (DLD) and/or dyslexia aged 10–14 years are impaired at disambiguating linguistic structures through prosody. However, these children can discriminate and imitate the prosodic structures as such, if no processing of their linguistic meaning is required. The authors thus conclude that it is the interaction between prosody and other components of language (such as syntax and pragmatics) to be problematic for children with DLD and/or dyslexia, whereas prosody itself does not seem to be a core impairment. Geiser et al. (2014) have further shown that identification of prosodic phrase boundaries for speech processing should not be impaired in children with dyslexia.

The present study will focus on how the syntax-prosody interface is processed by children with Developmental Dyslexia (DD); in particular the role of prosody in resolving syntactic ambiguities will be investigated. Special attention has been devoted to the analysis of the syntactic structures proposed. In fact, studies on syntactic ambiguity (Ferreira and Clifton, 1986, MacDonald, Just & Carpenter, 1992; Marslen-Wilson, Tyler, Warren, Grenier, & Lee, 1992; Trueswell, 1996; Kjølgaard & Speer, 1999; Ferreira & Dell, 2000; Carlson, Clifton, & Frazier, 2001, among others) mostly used sentences in which ambiguity results from the attachment of the PP (e.g. “The man hit the girl with a hat”: the PP “with a hat” can be attached to the VP “hit” specifying its instrumental vehicle, or to the NP “the girl”, acting as a modifier) or of the relative clause, whereas in this study different types of syntactic ambiguity are tested and compared. Several principles govern the parsing process of such structures, relating to lexical, syntactic, semantic and pragmatic factors. In

spoken sentences, prosody strongly contributes to disambiguation: a PP could be attached to the VP (the verbal phrase) or to the NP (the object phrase), depending on the pauses due to the F0 movement (Silverman, 1986). If a pause is placed after the verbal phrase, low attachment (in this case, attachment to the NP) is preferred, while a pause after the object phrase supports high attachment (to the VP).

Another type of ambiguity investigated in this study is the so-called “garden path effect”, which leads the parser initially to an incorrect interpretation, usually due to the components having multiple meanings, with the grammatical parse being significantly less frequent than the misinterpretation (Pritchett, 1988). For instance, when hearing or reading the sentence “Because Bill drinks wine is never kept in the house”, there normally is a first interpretation of the sentence with “wine” as the direct object of the verb “drink”, followed by a reinterpretation of the sentence, triggered by the second verb “is”, with “wine” as the subject of the second clause. This phenomenon depends on the ambiguous function of the word wine, but also on the possibility of the verb drink to act in a transitive (more frequent) or intransitive manner (Ferreira & Henderson, 1991). There are various studies on ambiguity in garden path sentences but they concern reading comprehension and mostly involve adults (for example Rayner, Carlson, & Frazier, 1983; Ferreira & Clifton, 1986; Frazier, 1987; Spivey & Tanenhaus, 1998, Christianson, Hollingworth, Halliwell & Ferreira, 2001; Meseguer, Carreiras, & Clifton, 2002). As prosody guides syntactic parsing in auditory presentation, the syntactic “garden path” effects found in reading studies should be resolved in sentences spoken with appropriate prosody. Prosody (as well as implicit prosody during reading) may play different roles in sentence parsing. With relative clause (RC) garden-path sentences, for instance, the absence of a prosodic break before a short RC follows optimal phrase-length principles but may also be simply signaling local (low) attachment. By contrast, the presence of a prosodic break before a long RC may reflect the need of a break at the left edge of a clause, but it may also be marking non-local (high) attachment.

Finally, ambiguity can also result from the “closure” of the syntactic parser, i.e. the process of terminating a clause. Usually, an intransitive verb will determine an early termination (e.g., “While the boy slept (end of clause) the dog yawned”), a so-called early closure. By contrast, a transitive verb will determine a late closure (e.g., “While the boy scratched the dog (end of clause) the girl yawned”). In English sentence processing, new words tend to be associated with the phrase or clause currently being processed (Frazier, 1978, 1987). This is known as the “principle of late-

closure”, or as recency principle (Frazier & Fodor, 1978). In fact, for early closure sentences (e.g. “While the boy scratched the dog yawned” or “Because Bill drinks wine is never kept in the house”), a relatively longer processing time is required due to an initial misanalysis of the sentence as a late closure structure followed by restructuring (Frazier & Rayner, 1982; Ferreira & Henderson, 1991; Frazier & Clifton, 1996), or to lexically associated frequency information, favouring verbal argument structure analysis (Tanenhaus & Carlson, 1989; MacDonald, Perlmutter, & Seidenberg, 1995). The experimental questions are a) whether children with Dyslexia are able to use prosody to disambiguate potentially ambiguous sentences or whether they are impaired in this domain; b) whether difficulties, if any are found, depend on the type of syntactic structure involved.

2. Method

2.1. Participants

Eighteen children with typical development (TD) and fifteen children with Developmental Dyslexia (DD), age ranging from 11 to 14 years, took part in the experiment. TD children were recruited from local primary and secondary schools, while children with DD were selected among those diagnosed at Scientific Institute “IRCCS E. Medea” in Lombardy, Northern Italy, as having Specific Reading Disorders according to standard ICD-10 criteria (WHO, 1992).

All participants were native Italian speakers and they were regularly attending school. Standard exclusion criteria had been applied (ICD-10, WHO 1992); moreover, children with additional diagnosis of Attention Deficit Disorders (ADD-ADHD) were excluded, and so were children who had been previously treated with music-based intervention programs.

Children with developmental dyslexia were included in the clinical sample if they had a score at least 2SD below the mean in at least two phonological and reading tests, IQ scores ≥ 85 and no concomitant language impairment diagnosis. TD children, before being included in the control group, were administered a battery test to evaluate their general intellectual and linguistics abilities². All parents signed informed consent.

² The tests administered were: CPM (Raven, 1947), a test of morphosyntactic comprehension and production (CoSiMo, Milani, Soddu, Cattaneo, Peverelli, Cataldo & Lorusso, 2005 described in Cantiani, Lorusso, Perego, Molteni &

The study was approved by the Ethics Committee of the University of Pavia according to standards of the Helsinki Declaration (1964).

2.2 Materials

Eighteen sentences involving ambiguous syntactic structures were created. The choice of lexical items was controlled, using a corpus of primary school - related Italian language (Marconi et al. 1994). Three types of sentences were tested:

1. Six sentences with a temporary syntactic closure ambiguity (TSCA). E.g.:

- (1)
 - a. [*Quando Marta guida*] [*la macchina fuma*]
 [‘When Mary drives’] [‘the car smokes’]
 - b. [*Quando Marta guida la macchina*] [*fuma*]
 [‘When Mary drives the car’] [‘(she) smokes’]

In the sentences in (1), the F0 guides the syntactic closure during the parsing; the issue is whether *la macchina* (the car) is the subject of the main verb (1a) or the object of the embedded verb (1b). A prosodic break before or after *la macchina* resolves the ambiguity.

2. Six pseudo-Garden Path Sentences (pGP). E.g.:

- (2)
 - a. [*La giovane*] [*fotografa la pianta*]
 [‘The young lady’] [‘photographs the plant’]
 - b. [*La giovane fotografa*] [*la pianta*]
 [‘The young photographer’] [‘plants it’].

In these kinds of Italian sentences, each phrase has a double syntactic function, with its own prosody. In the example (2), *la giovane* (‘the young’), in Italian, could be either a noun phrase (equivalent to ‘the young

Guasti, 2015), a test of sentence repetition (Ferrari, De Renzi, Faglioni & Barbieri, 1981) and tests of meta-phonological abilities (phonemic analysis and rhyme) (CMF, Marotta et al. 2002, VAU-MeLF, Bertelli e Bilancia, 2006).

lady') or an adjective phrase; *fotografa* ('photographer/photographs') could be either a noun phrase or a verbal phrase and *la pianta* could be a determiner + noun ('the plant') or a clitic + verb (equivalent to 'plants it'). A prosodic break after "la giovane" would favour an interpretation of the preceding word as a noun phrase and of "fotografa" as a VP followed by its object NP "la pianta". By contrast, a prosodic break after "fotografa" would favour the interpretation of the two preceding words as an adjective + noun NP, so that the following "la pianta" would be interpreted as a VP (object clitic + verb). Also pitch (F0) variations differ between the two versions of the sentence.

3. Six sentences in which ambiguity results from the Prepositional Phrase (PP) attachment, which can be either "high", as in (3b), or "low", as in (3a). E.g.:

- (3)
 - a. [*Gianni saluta*] [*la ragazza con il cappello*]
 ['John greets'] ['the girl with the hat']
 - b. [[*Gianni saluta la ragazza*] *con il cappello*]
 [['John greets the girl'] 'with the hat']

In this case, a break after the VP favours low attachment of the PP, while a break after the first NP (the object NP "la ragazza") favours high attachment of the PP. The full list of sentences is reported in the Appendix. Three grammatical variables were taken into account:

(a) Type (the three types of sentence described above: TSCA, pseudo-Garden Path, PP-attachment).; (b) GFA (Grammatical Function re-Assignment) expressing the need to re-assign grammatical function after processing of the prosodic cue. Specifically, the grammatical class of some words in these sentences can vary, depending on prosody and syntax. Clearly, such polysemous words are quite uncommon (Italian is a morphologically rich language where verbs are generally inflected and very different from nouns) and their combination in a single sentence is rather unusual; nonetheless, they are perfectly grammatical and meaningful in both versions, as shown in (2) above. GFA is a characteristic of Type 2 sentences only; (c) Distance, i.e. the distance in the syntactic dependencies. Specifically, it refers to the distance between two phrases (highlighted in bold) after a final syntactic structure has been assigned based on the prosodic cues. The target phrases could be within the same clause (Short Distance Dependency), as in (4), repeated here:

- (4) [*Quando Marta guida*] [*la macchina fuma*]
 ['When Mary drives'] ['the car smokes']

Otherwise, the target phrases could be in separate clauses (Long Distance Dependency), e.g.:

- (5) [*Quando **Marta** guida la macchina*] [*fuma*]
 ['When Mary drives the car'] ['(she) smokes']

See table 1 for a summary of the materials and manipulations.

Table 1. Types of sentences presented in this research

Sentence	Type	GFA	Distance
1a. [<i>Quando Marta guida</i>] [<i>la macchina fuma</i>] ['When Mary drives'] ['the car smokes']	Closure ambiguity	No	short
1b. [<i>Quando Marta guida la macchina</i>] [<i>fuma</i>] ['When Mary drives the car'] ['(she) smokes']	Closure ambiguity	No	long
2a. [<i>La giovane</i>] [<i>fotografa la pianta</i>] ['The young lady'] ['photographs the plant']	pseudo-Garden-Path	Yes	short
2b. [<i>La giovane fotografa</i>] [<i>la pianta</i>] ['The young photographer'] ['plants it']	pseudo-Garden-Path	Yes	long
3a. [<i>Gianni saluta</i>] [<i>la ragazza con il cappello</i>] ['John greets'] ['the girl with the hat']	PP attachment	No	short
3b. [<i>Gianni saluta la ragazza con il cappello</i>] ['John greets the girl'] ['with the hat']	PP attachment	No	long

There were six sentences for each sentence type. A complete list of the sentences with their characteristics is reported in the Appendix A. For each

sentence, the two possible versions, disambiguated with different prosodic patterns, were spoken by a female trained speaker and recorded with Audacity software. Background noise was filtered out through the software. Three pictures were created for each of the sentences: the target one, a picture of the alternative interpretation and a distractor figure, where the main semantic elements (for instance, the girl, the man and the hat) were represented but the crucial mutual relationship or action was missing. The colour pictures were drawn by a professional graphic designer with Paint X Lite and Adobe Photoshop CC 2015 through the Wacom Bamboo graphic tablet, with a high resolution. The correspondence of the spoken sentence with the correct picture was tested on a group of 27 adult participants: the recordings and the pictures were progressively refined until accuracy rates reached at least 80 %.

2.3. Procedure

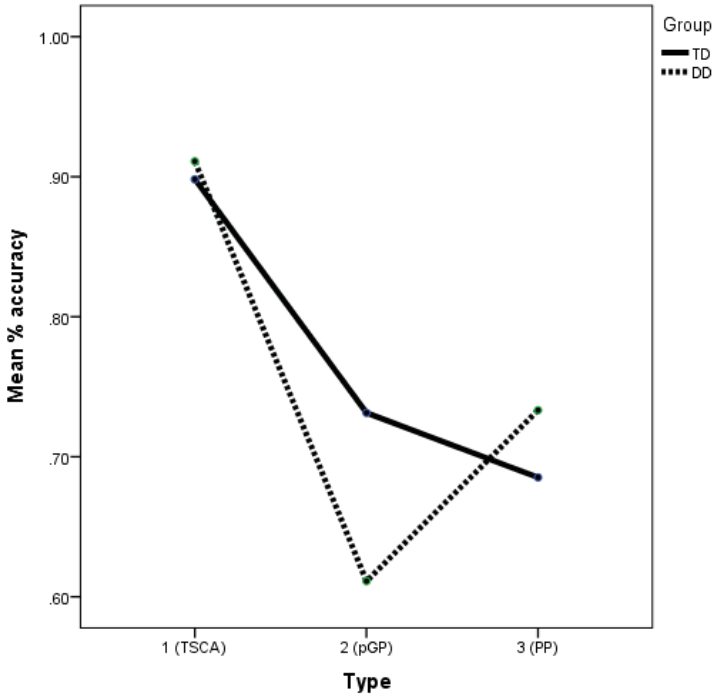
Both auditory and visual stimuli were delivered by a laptop Dell Inspiron 7347 x64, 13,3" screen, through an algorithm that was designed using Psychopy software (Peirce, 2007). All participants were individually tested in a quiet room, seated next to the experimenter, in front of the PC screen. They listened to the stimuli through XB550AP Sony Extra Bass headphones. Three aligned pictures appeared on the PC screen simultaneously to the recorded sentence and the children had to choose the corresponding picture, by pressing "1", "2" or "3" with the index finger of the dominant hand on the keyboard. The pictures appeared on the screen until the child pressed on the keyboard. When the child pressed the key to respond, the pictures disappeared and a new sentence was presented through the headphones. The pictures representing the target answer, the alternative interpretation and the distractor were always presented simultaneously, horizontally aligned, but their order differed for each sentence (see Appendix 1 for an example). The sentences were presented in two fixed, internally randomized sequences. In each sequence, only one of the two audio-recordings was presented for each sentence, corresponding to one of the two possible interpretations (based on prosody), so as to ensure that the two alternative structures for each sentence were equally represented in each of the sequences. Half of the children received the first sequence, the other half received the second sequence, balancing across and within groups. The two parallel lists were assigned in alternated fashion to participants in order to test all possible syntactic structures without inducing response bias effects due to the repetition of a same syntactic structure. No feedback was given regarding response accuracy.

Response and responding time were automatically recorded by the program. No time limits were given for responding and the recorded files could not play again. Children started with a warm-up session in order to make sure they correctly understood the task.

3. Results

The main aim of the study was to investigate the effects of the experimental variables on performance. In order to highlight such effects, within-subjects differences were computed for all the variables of interest. Specifically, all differences in scores between conditions were computed regarding types of sentence (3 types of sentence), Grammatical Function re-Assignment (difference between performances when GFA was required versus when it was not required) and short or long distance (difference between performances on sentences with short versus long dependency distance). The difference-scores did not show any large deviations from the normal distribution, hence a series of ANOVAs was performed. Two different types of information were considered: accuracy (percentage of correct responses) and type of errors (proportion of the two types of errors in incorrect responses). Since only two types of errors are possible (Distractors or Alternative choices), and they are thus complementary with respect to each other (their sum always representing 100% of errors) only Distractor choices were analysed. It is implicit that, any time the proportion of Distractor choices produces significant differences in a comparison, the proportion for choice of Alternative responses is also significantly different, albeit in the opposite direction. First of all, Accuracy scores were compared between the two groups with respect to Type. A General Linear Model (GLM) was run on Accuracy scores for type 1, type 2 and type 3 sentences with Group as a between factor. No significant effects emerged for Group nor for its interaction with Type, but a significant main effect emerged for Type: $F(2, 54.572) = 12.033$, $p < .001$. This effect is described in Figure 1.

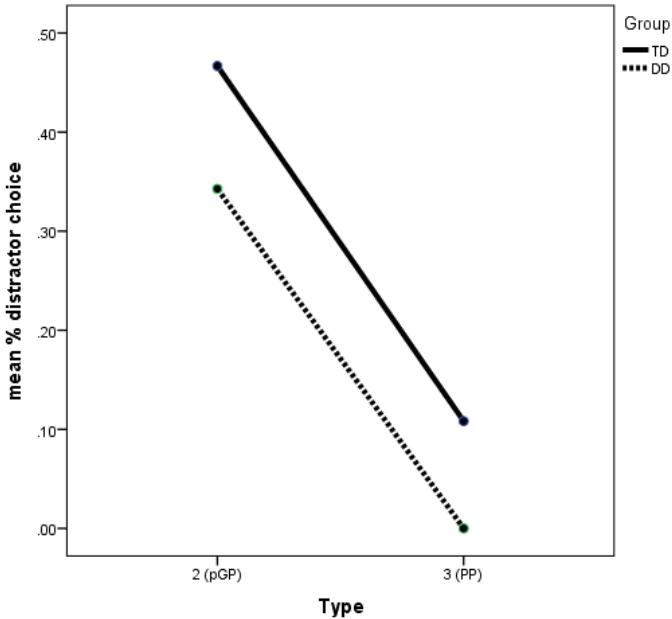
Figure 1. Significant Effect of Type in the two Groups.



It can be seen that Type 2 (pGP) and 3 (PP-attachment) sentences are more difficult than Type 1 (TSCA) sentences, with no significant differences between Groups. Post-hoc tests show that significant differences exist between Types 1 and both Type 2 ($F(1,31) = 31.86, p < .001$) and Type 3 ($F(1,31) = 13.861, p = .001$), but not between Type 2 and Type 3. Next, the effect of Type on Distractor frequency was analysed, i.e. how often, in an error, the distractor was chosen depending on sentence Type. Group was introduced in the GLM as between-subject factor. Since the number of Distractor choices for Type 1 sentences was too low, only Types 2 and 3 were included in the analysis. Again, Group did not produce any significant main effect nor interactions with Type, while Type showed a significant main effect $F(1,17) = 13.038, p = .002$. The effect of Type on Distractor choices is represented in Figure 2. As clearly shown in the figure, and for both groups, Type 2 sentences (pGP) produced more distractor choices than Type 3 sentences (PP) (Fig. 2). This also implies

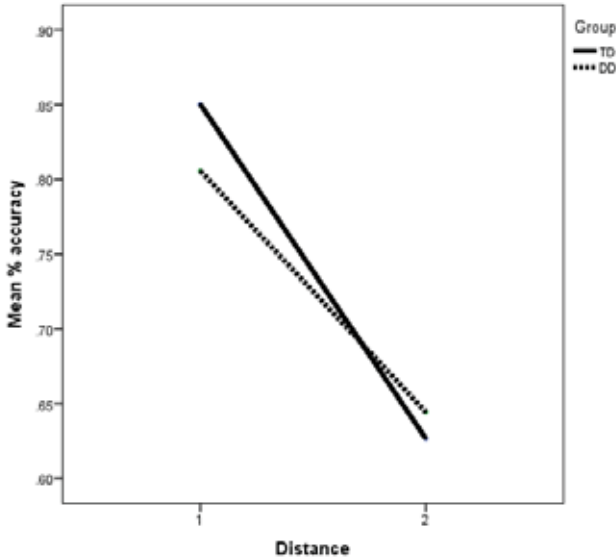
that in Type 3 sentences, errors were more often represented by Alternative choices than they were in Type 2 sentences.

Figure 2. Distractor Choices in Errors, in Type 2 and Type 3 Sentences.



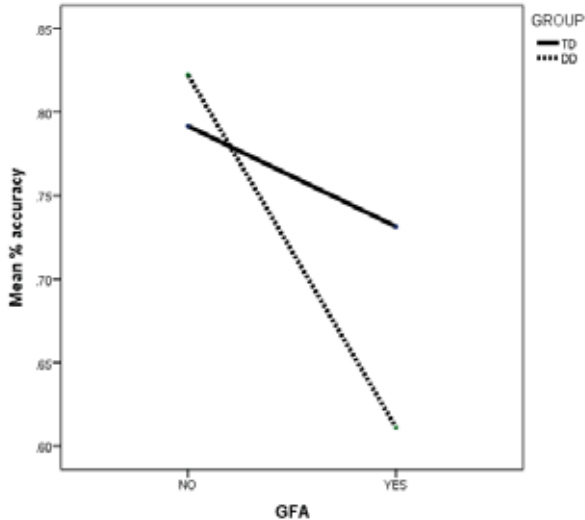
The second variable taken into consideration was Distance. A GLM on accuracy scores with Distance as within-subject factor and Group as between-subject factor revealed a main effect of Distance, $F(1, 31) = 13.770$, $p = .001$. Sentences with short distance dependencies produced more accurate responses than those with long-distance dependencies (see Fig. 3). Group did not yield any significant main effect nor significant interactions with Distance; similarly, no differential effects emerged for the choice of the various types of wrong response (Distractor versus Alternative response).

Figure 3. Distance Effects in the Two Groups (1 = short distance; 2 = long distance).



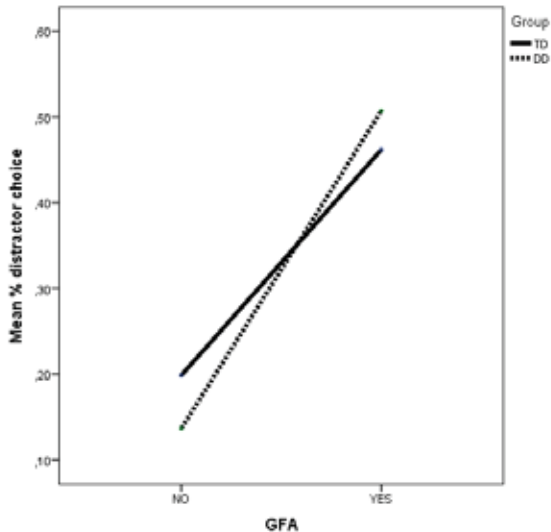
The last variable taken into consideration was GFA. A GLM analysis was performed on Accuracy scores with GFA (presence vs absence of the need to reassign grammatical function) as within-subject factor and Group as between-subject factor. A significant main effect of GFA emerged, $F(1, 31) = 9.992$, $p = .004$. The interaction of GFA with Group approached significance: $F(1, 31) = 3.085$, $p = .089$. Taking into account that the predicted effect is unidirectional (dyslexic children being expected to have stronger interference from the process of re-assigning grammatical function as compared to typically reading children), it was argued that one-tailed significance levels could be considered (in $p = .045$), and post-hoc analyses (t-tests for paired samples) were run on the single groups, showing that no difference is present in the TD group between the two conditions (presence-absence of GFA), whereas this difference is highly significant for the group of children with Dyslexia: $t(14) = .105$, $p = .008$ (Fig. 4).

Figure 4. The Effect of GFA on Accuracy Scores (NO = GFA absent; YES = GFA present).



Finally, the effect of GFA on the type of non-correct response was taken into consideration. A GLM analysis on the percentage of Distractor choice with GFA as within-subject factor and Group as between-subject factor showed a significant effect of GFA, $F(1,23) = 11.448$, $p = .003$, but no significant interaction with Group. This means that, when making an error, the distractor is chosen more often (and the alternative choice less often) if GFA is required than if it is not, and this is true for both dyslexic and non-dyslexic children (Fig. 5).

Figure 5. The Effect of GFA on Distractor Choices (NO = GFA absent; YES = GFA present).



4. Discussion

The aim of the present study was to investigate processing of the syntax-prosody interface in children with and without reading impairments. Since the first days of life, infants are able to exploit rhythmic and sound organization of their native language in order to extract the underlying structure of utterances (Jusczyk, 2002). Moreover, following the prosodic bootstrapping, hypothesis infants rely on the prosodic characteristics of their mother tongue to infer its syntactic properties (Wanner & Gleitman, 1982; Mazuka, 1996; Morgan & Demuth, 1996, Christophe, Nespore, Guasti, & Van Ooyen, 2003). Interestingly, no previous studies focused, to our knowledge, on the processing of the syntax-prosody interface by children with Developmental Dyslexia.

Our experimental questions were focused on the possibility that children with Dyslexia are impaired at using prosody to disambiguate potentially ambiguous sentences and that such difficulties, if any, may depend on the type of syntactic structure. According to our data, identification of prosodic phrase boundaries for speech processing seems not to be impaired in children with dyslexia (in line with previous literature, see for

example Geiser et al., 2014). In fact, children with DD use prosody to disambiguate sentences and their response accuracy does not differ from that of TD children. Nonetheless, significant differences between groups emerge when distractor choices are taken into consideration. Distractor choices differ from both target and alternative choices because children who choose distractors are not simply unable to use prosody to disambiguate sentence structure (as the choice of the alternative structure would suggest), but they seem to be confused about sentence structure at all. Taking into account Distractor choices, difficulties in syntactic processing can be highlighted in DD in the GFA condition, which seems to be a particularly complex operation. It should be remembered that children with DD do not show reduced accuracy, and that the confusion suggested by choice of distractors does not generalize to all syntactic structures: rather, confusion arises only when syntactically complex sentences must be processed and disambiguated. It can be concluded that the processing of sentence prosody per se seems to be unaffected in dyslexia, and that the problem appears to relate to syntax and, possibly, to its interface with prosody. This is a rather unexpected finding, considering that the children had reading impairments and did not have a concomitant diagnosis of developmental language disorder (DLD). It should be acknowledged, though, that no systematic search had been conducted on the children's clinical records to exclude that they had had language difficulties in their pre-school years. Due to the high rate of comorbidity between language disorders and reading impairments (e.g., Bishop and Snowling, 2004; Catts et al. 2005), it could be hypothesized that at least some of the children in the DD group had partially resolved previously existing language difficulties, and that such difficulties still emerged when sentence complexity is particularly high, as in the case of GFA. This is also what appears to result from current work comparing children with "pure" DD and children with DLD on the same type of task (see Caccia and Lorusso, 2019).

More generally, different rates of target answers are found in both groups, depending on the different sentence types. Precisely, both pseudo-Garden Path sentences and sentences with an ambiguous PP-attachment are more difficult to disambiguate based on prosody than sentences with a temporary syntactic closure ambiguity. Indeed, sentences with a more complex psycholinguistic structure, as in type 2 ("pseudo-Garden-Path" sentences), in which the prosodic cue defines also the grammatical function of each syntactic phrase, are more complex to manage by all children. This is true even if type 2 sentences have fewer constituents than other sentence types: the crucial aspect, then, seems to be not the number

of constituents but rather their complexity in terms of grammatical function. This seems to be confirmed by the higher proportion of distractor choices for this type of sentence than for sentences with PP-attachment. As to the latter, by contrast, it is possible to hypothesize that length contributes to making them more complex than the other sentence types. For this group of sentences, the largest proportion of incorrect choices derives from choosing the alternative structure (i.e., not using prosody efficiently for disambiguation). This suggests the possibility that prosodic patterns for the two versions of sentences with PP-attachment are less distinctive (less different, or less saliently different) than those of sentences with temporary syntactic closure ambiguity, so that disambiguation based on prosody turns out to be more difficult for this type of sentence structure. This hypothesis should be confirmed with in-depth analyses of the prosodic phrase profiles.

As to the two grammatical effects taken into account in this study, both Distance in syntactic dependencies and GFA have a significant effect on both groups' response accuracy. Moreover, GFA (but not Distance) produces a higher percentage of Distractor choices in incorrect responses. Thus, long-distance dependencies produce less accurate responses than short-distance ones, and the presence of GFA induces more errors deriving from the choice of distractors (in other terms, more confusion) than structures where the grammatical function and category of each term is non-ambiguously determined.

The present study has shown that the use of more specific language evaluation tools and more fine-grained linguistic analyses allows us to identify differences and difficulties that do not emerge at the surface level. In light of the above, focusing on fine-grained linguistic aspects and on different linguistic levels, such as prosody and syntax, would be useful for a more accurate diagnosis. Indeed, sophisticated linguistic tools could provide more specific linguistic profiles of the children, helping to identify more effective rehabilitation strategies.

Nevertheless, the present study has some limitations: first of all, the number of the items per experimental condition was small (the duration of the experiment had been kept to a minimum in consideration of the reduced attentional span of the children, especially those with DD) resulting in limited statistical power. Moreover, the sample size was rather limited and should be enlarged in order to have a clearer picture, especially aiming at a better characterization of the specific linguistic profile in Developmental

Dyslexia in comparison with other neurodevelopmental disorders such as Developmental Language Disorders.

Acknowledgments

We would like to thank Carolina Caccia for her valuable contribution in the drawing of the pictures and all the children and the families who took part in the experiment. This work was supported by the Italian Ministry of Health [grant number RC2018-2019]

References

- Bertelli, B., Bilancia, G. (2006). *VAUMeLF Batterie per la valutazione dell'attenzione uditiva e della memoria di lavoro fonologica nell'età evolutiva*. [Batteries for the assessment of auditory attention and phonological working memory in the developmental age] Giunti O.S, Firenze.
- Bishop, D.V.M., Snowling, M.J., 2004. Developmental dyslexia and specific language impairment: same or different? *Psychological Bulletin*, 130 (6), 858-886.
- Caccia, M., & Lorusso M.L. (2019). When prosody meets syntax: The processing of the syntax-prosody interface in children with developmental dyslexia and developmental language disorder. *Lingua*, 224, 16-33.
- Cantiani, C., Lorusso, M.L., Perego, P., Molteni, M., & Guasti, M.T. (2015). Developmental Dyslexia With and Without Language Impairment: ERPs Reveal Qualitative Differences In Morphosyntactic Processing. *Developmental Neuropsychology*, 40:5, 291-312, DOI: 10.1080/87565641.2015.1072536.
- Carlson, K., Clifton, C., & Frazier, L. (2001). Prosodic boundaries in adjunct attachment. *Journal of Memory and Language*, 45(1), 58-81.
- Carroll, P. J., & Slowiaczek, M. L. (1987). Modes and modules: Multiple pathways to the language processor, J. Garfield (Ed.), *Modularity in knowledge representation and natural language understanding* (pp.221-247) New York: Academic Press.
- Catts, H. W., Adlof, S. M., Hogan, T. P., & Weismer, S. E. (2005). Are specific language impairment and dyslexia distinct disorders?. *Journal of Speech, Language, and Hearing Research*, 48(6), 1378-1396.
- Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive psychology*, 42(4), 368-407.

- Christophe, A., Nespor, M., Teresa Guasti, M., & Van Ooyen, B. (2003). Prosodic structure and syntactic acquisition: The case of the head-direction parameter. *Developmental Science*, 6(2), 211–220. <http://doi.org/10.1111/1467-7687.00273>
- Corriveau, K. H., & Goswami, U. (2009). Rhythmic motor entrainment in children with speech and language impairments: tapping to the beat. *Cortex*, 45(1), 119–130. <http://dx.doi.org/10.1016/j.cortex.2007.09.008>
- Ferrari, E., De Renzi, E., Faglioni, P., & Barbieri, E. (1981). Standardizzazione di una batteria per la valutazione dei disturbi del linguaggio nell'età scolare [Standardization of a battery for language assessment in school-age]. *Neuropsichiatria Infantile*, 235, 148–158.
- Ferreira, F., & Clifton, C. (1986). The independence of syntactic processing. *Journal of memory and language*, 25(3), 348–368.
- Ferreira, V. S., & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive psychology*, 40(4), 296–340.
- Ferreira, F., & Henderson, J. M. (1991). Recovery from misanalyses of garden-path sentences. *Journal of Memory and Language*, 30(6), 725–745.
- Frazier, L. (1987). Syntactic processing: evidence from Dutch. *Natural Language & Linguistic Theory*, 5(4), 519–559.
- Frazier, L., Carlson, K., & Clifton, C. (2006). Prosodic phrasing is central to language comprehension. *Trends in Cognitive Sciences*, 10(6), 244–249. <http://doi.org/10.1016/j.tics.2006.04.002>
- Frazier, L., & Clifton, C. (1996). *Construal*. MIT Press.
- Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, 6(4), 291–325.
- Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive psychology*, 14(2), 178–210.
- Geiser, E., Kjelgaard, M., Christodoulou, J. A., Cyr, A., & Gabrieli, J. D. E. (2014). Auditory temporal structure processing in dyslexia: Processing of prosodic phrase boundaries is not impaired in children with dyslexia. *Annals of Dyslexia*, 64(1), 77–90. <http://doi.org/10.1007/s11881-013-0087-7>
- Goswami, U., Gerson, D., & Astruc, L. (2010). Amplitude envelope perception, phonology and prosodic sensitivity in children with developmental dyslexia. *Reading and Writing*, 23(8), 995–1019. <http://doi.org/10.1007/s11145-009-9186-6>
- Heffner, C. C., & Slevc, L. R. (2015). Prosodic structure as a parallel to musical structure. *Frontiers in Psychology*, 6(DEC), 1–14.

- <http://doi.org/10.3389/fpsyg.2015.01962>
- Huss, M., Verney, J. P., Fosker, T., Mead, N., & Goswami, U. (2011). Music, rhythm, rise time perception and developmental dyslexia: perception of musical meter predicts reading and phonology. *Cortex*, 47(6), 674–689.
- Jusczyk, P. W. (2002). How infants adapt speech-processing capacities to native-language structure. *Current Directions in Psychological Science*, 11(1), 15–18. <http://doi.org/10.1111/1467-8721.00159>
- Kjelgaard, M. M., & Speer, S. R. (1999). Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity. *Journal of Memory and Language*, 40(2), 153–194.
- Kraljic, T., & Brennan, S. E. (2005). Prosodic disambiguation of syntactic structure: For the speaker or for the addressee? *Cognitive Psychology*, 50(2), 194–231. <http://doi.org/10.1016/j.cogpsych.2004.08.002>
- Kreiner, H., & Eviatar, Z. (2014). The missing link in the embodiment of syntax: Prosody. *Brain and Language*, 137, 91–102. <http://doi.org/10.1016/j.bandl.2014.08.004>
- Ladd, D. R. (2008). *Intonational phonology*. Cambridge University Press.
- Mazuka, R. (1996). Can a grammatical parameter be set before the first word? Prosodic contributions to early setting of a grammatical parameter. *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*, 313–330.
- MacDonald, M. C., Just, M. A., & Carpenter, P. A. (1992). Working memory constraints on the processing of syntactic ambiguity. *Cognitive psychology*, 24(1), 56–98.
- MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101, 676–703.
- Marconi, L., Ott, M., Pesenti, E., Ratti, D., Tavella, M. (1993). *Lessico Elementare. Dati statistici sull'italiano letto e scritto dai bambini delle elementari*. Bologna: Zanichelli.
- Marotta, L. (2008). *Test CMF. Valutazione delle competenze metafonologiche. Con CD-ROM*. [Evaluation of metaphonological skills with CD-ROM]. Edizioni Erickson.
- Marshall, C. R., Harcourt-Brown, S., Ramus, F., & van der Lely, H. K. J. (2009). The link between prosody and language skills in children with specific language impairment (SLI) and/or dyslexia. *International Journal of Language & Communication Disorders / Royal College of Speech & Language Therapists*, 44(4), 466–488. <http://doi.org/10.1080/13682820802591643>
- Marslen-Wilson, W. D., Tyler, L. K., Warren, P., Grenier, P., & Lee, C. S.

- (1992). Prosodic effects in minimal attachment. *The Quarterly Journal of experimental psychology*, 45(1), 73-87.
- Meseguer, E., Carreiras, M., & Clifton, C. (2002). Overt reanalysis strategies and eye movements during the reading of mild garden path sentences. *Memory & Cognition*, 30(4), 551-561.
- Morgan, J. & Demuth, K. (1996). The prosodic structure of early words. *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*, 171, 184.
- Nooteboom, S. (1997). The prosody of speech: melody and rhythm. *The handbook of phonetic sciences*, 5, 640-673.
- Nordquist, R. (2020, February 11). Late Closure (Sentence Processing). Retrieved from <https://www.thoughtco.com/late-closure-sentence-processing-1691101>
- Patel, A. D. (2003). Language, music, syntax and the brain. *Nature Neuroscience*, 6(7), 674-682. <http://doi.org/10.1038/nn1082>
- Patel, A. D. (2008). *Music, Language, and the Brain*. Oxford University Press, Inc.
- Patel, A. D., Gibson, E., Ratner, J., Besson, M., & Holcomb, P. J. (1998). Processing syntactic relations in language and music: An event-related potential study. *Journal of cognitive neuroscience*, 10(6), 717-733. <http://dx.doi.org/10.1162/089892998563121>
- Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of neuroscience methods*, 162(1), 8-13
- Pennington, B. F. (2006). From single to multiple deficit models of developmental disorders. *Cognition*, 101(2), 385-413. <http://dx.doi.org/10.1016/j.cognition.2006.04.008>
- Peppé, S. J. (2009). Why is prosody in speech-language pathology so difficult?. *International Journal of Speech-Language Pathology*, 11(4), 258-271. <http://dx.doi.org/10.1080/17549500902906339>
- Peterson, R.L., & Pennington B.F. (2012). Developmental dyslexia. *Lancet*, 379 (9830), 1997-2007. [http://dx.doi.org/10.1016/S0140-6736\(12\)60198-6](http://dx.doi.org/10.1016/S0140-6736(12)60198-6)
- Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *The Journal of the Acoustical Society of America*, 90(6), 2956-2970. <http://doi.org/10.1121/1.401770>
- Pritchett, B. L. (1988). Garden path phenomena and the grammatical basis of language processing. *Language*, 539-576.
- Ramus, F., & Szenkovits, G. (2008). What phonological deficit?. *The Quarterly Journal of Experimental Psychology*, 61(1), 129-141. <http://dx.doi.org/10.1080/17470210701508822>

- Raven, J. C. (1947). *Coloured progressive matrices*. London: H. K. Lewis.
- Rayner, K., Carlson, M., & Frazier, L. (1983). The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. *Journal of verbal learning and verbal behavior*, 22(3), 358-374.
- Report of the Task Force on Dyslexia (2001).
https://www.sess.ie/sites/default/files/Dyslexia_Task_Force_Report_0.pdf
- Roncaglia-Denissen, M. P., Schmidt-Kassow, M., & Kotz, S. A. (2013). Speech Rhythm Facilitates Syntactic Ambiguity Resolution: ERP Evidence. *PLoS ONE*, 8(2), 1-9.
<http://doi.org/10.1371/journal.pone.0056000>
- Sammler, D., Kotz, S. A., Eckstein, K., Ott, D. V. M., & Friederici, A. D. (2010). Prosody meets syntax: The role of the corpus callosum. *Brain*, 133(9), 2643-2655. <http://doi.org/10.1093/brain/awq231>
- Siegel, L. S. (2006). Perspectives on dyslexia. *Paediatrics & child health*, 11(9), 581-587. <http://dx.doi.org/10.1093/pch/11.9.581>
- Silverman, K. (1986). F₀ Segmental Cues Depend on Intonation: The Case of the Rise after Voiced Stops. *Phonetica*, 43(1-3), 76-91.
- Snedeker, J., & Yuan, S. (2008). Effects of prosodic and lexical constraints on parsing in young children (and adults). *Journal of memory and language*, 58(2), 574-608.
- Snow, D. (1994). Phrase-final syllable lengthening and intonation in early child speech. *Journal of Speech, Language, and Hearing Research*, 37(4), 831-840.
- Spivey-Knowlton, M., & Sedivy, J. C. (1995). Resolving attachment ambiguities with multiple constraints. *Cognition*, 55(3), 227-267.
- Spivey, M. J., & Tanenhaus, M. K. (1998). Syntactic ambiguity resolution in discourse: modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(6), 1521.
- Tallal, P. (1980). Language and reading: Some perceptual prerequisites. *Bulletin of the Orton Society*, 30(1), 170-178.
- Tanenhaus, M. K., Carlson, G., & Trueswell, J. C. (1989). The role of thematic structures in interpretation and parsing. *Language and cognitive processes*, 4(3-4), SI211-SI234.
- Thomson, J. M., Leong, V., & Goswami, U. (2013). Auditory processing interventions and developmental dyslexia: a comparison of phonemic and rhythmic approaches. *Reading and Writing*, 26(2), 139-161.
- Trueswell, J. C. (1996). The role of lexical frequency in syntactic ambiguity resolution. *Journal of memory and language*, 35(4), 566-585.

- Trueswell, J. C., Sekerina, I., Hill, N. M., & Logrip, M. L. (1999). The kindergarten-path effect: Studying on-line sentence processing in young children. *Cognition*, 73(2), 89-134.
- Wanner, E., & Gleitman, L. R. (Eds.). (1982). *Language acquisition: The state of the art*. CUP Archive.
- Warren, P., Grabe, E., & Nolan, F. (1995). Prosody, phonology and parsing in closure ambiguities. *Language and cognitive processes*, 10(5), 457-486.
- World Health Organization (1992). *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines* (Vol. 1). World Health Organization.

Appendix A

Full list of sentences with their properties

TYPE	VARIANTS	GFA	DISTANCE
TYPE 1: Temporary Syntactic Closure Ambiguity (TSCA)			
1) Mentre la mamma chiama il figlio piange	A) [While Mum calls][the child cries]	N	S
	B) [While Mum calls the child][(she) cries]	N	L
2) Intanto che Mario mangia la minestra si riscalda	A) [While Mario eats][the soup gets warm]	N	S
	B) [While Mario eats the soup][(he) warms himself]	N	L
3) Mentre la nonna cuce il maglione cade	A) [While Grandma saws][the sweater falls]	N	S
	B) [While Grandma saws the sweater][(she) falls]	N	L
4) Quando Marta guida la macchina fuma	A) [while Marta drives][the car smokes]	N	S
	B) [while Marta drives the car][(she) smokes]	N	L
5) Ogni volta che Luca saluta Giulia arrossisce	A) [Every time Luca greets][Giulia blushes]	N	S
	B) [Every time Luca greets Giulia][(he) blushes]	N	L
6) Mentre Marta sogna Claudio ride	A) [while Marta dreams][Claudio laughs]	N	S
	B) [while Marta dreams about Claudio] [(she) laughs]	N	L
TYPE 2: Pseudo-Garden-Path			
7) La giovane fotografa la pianta	A) [the young photographer][plants it]	Y	S
	B) [the young lady][photographs the plant]	Y	L
8) La vecchia ruota la sveglia	A) [the old wheel][wakes her up]	Y	S
	B) [the old woman][rotates the clock]	Y	L
9) La cattiva spia la pesca	A) [the evil spy][fishes it]	Y	S
	B) [the evil woman][spies the fishing]	Y	L

10) il cattivo pilota la scala	A) [the bad pilot][climbs it]	Y	S
	B) [the evil man] [pilots the broom]	Y	L
11) La bella sposa la spazzola	A) [the beautiful bride][brushes it]	Y	S
	B) [the beautiful girl][marries the brush]	Y	L
12) La vecchia porta la macchia	A) [the old door][stains it]	Y	S
	B) [the old woman][carries the stain]	Y	L
TYPE 3: PP-attachment			
13) Gianni saluta la ragazza con il cappello	A) [Gianni greets][the girl with the hat]	N	S
	B) [Gianni greets the girl] [with the hat]	N	L
14) La ragazza colpisce l'uomo con l'ombrello	A) [The girl hits][the man with the umbrella]	N	S
	B) [The girl hits the man][with the umbrella]	N	L
15) Il soldato tira la moto distrutta con il carrarmato	A) [The soldier carries] [the motorbike destroyed with the tank]	N	S
	B) [The soldier carries the destroyed motorbike] [with the tank]	N	L
16) Il poliziotto osserva il ladro nascosto dietro al cespuglio	A) [The policeman observed][the thief hiding behind the bush]	N	S
	B) [The policeman observed the thief] [while hiding behind the bush]	N	L
17) Il dottore visita il paziente in maglietta	A) [The doctor visits][the patient in a T-shirt]	N	S
	B) [The doctor visits the patient][in a T-shirt]	N	L
18) La mamma canta la canzone suonata con il papà	A) [Mum sings the song][(she) played with Dad]	N	S
	B) [Mum sings the played song][with Dad]	N	L

Appendix B

Examples of stimulus presentation


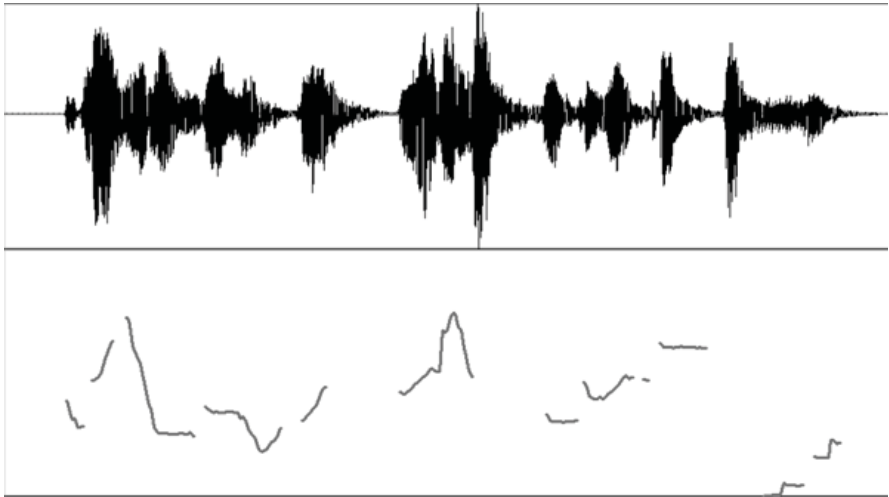
1.  [[Gianni saluta] la ragazza con il cappello].
[[*John greets*] *the girl with the hat*].

Figure 6. Sound Spectrum and Pitch Contour (F0) of Sentence 1.




2.  [[Gianni saluta la ragazza] con il cappello].
[[*John greets the girl*] *with the hat*].

Figure 7. Sound Spectrum and Pitch Contour (F0) of Sentence 2.

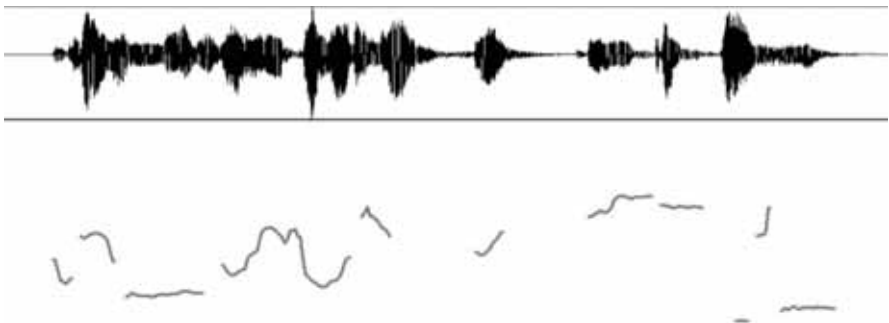


Figure 8. Three Pictures for The Audio Stimulus “John Greets_ The Girl With The Hat”.



The First Picture is the Distractor (No Actions), the Second One Shows the Alternative Interpretation (John Is Using A Hat To Greet A Girl) And The Third Picture Represents The Target Answer (John Is Greeting A Girl Who Is Wearing a Hat).

ALL IT TAKES TO PRODUCE PASSIVES IN GERMAN

YULIA ESAULOVA, SARAH DOLSCHEID
& MARTINA PENKE¹

Abstract

When do speakers produce passive sentences in order to describe an event? Previous research has shown that in English speakers the use of passives depends on drawing visual attention to a patient character depicted in event scenes. In case-marking languages, in contrast, this has not been found to coerce speakers to produce passives. The present study employed eye-tracking methodology in a picture description task to investigate passive voice production in speakers of German, a case-marking language. The results show that explicit cueing of patients, as well as their animacy, significantly increased the production of passive utterances. Crucially, speech onset and gaze pattern data suggest two different formulation strategies applied by speakers when passives are produced. Drawing speakers' attention to patients via cueing appears to result in a linear incremental planning and relatively low costs associated with passive production. When both agents and patients are animate, a more costly structure-driven strategy is applied. The findings indicate that depending on a given context speakers may employ different formulation strategies in sentence planning. Moreover, the data reveal that producing passives can be easier than producing active voice sentences.

1. Introduction

Are you reading this sentence? Or is this sentence being read by you? While both questions convey the same message, they differ in their syntactic

¹ Corresponding author: Martina Penke, Department of Special Education and Rehabilitation, Herbert-Lewin-Str. 10, Cologne 50931, Germany. Tel.: +49 221 4706373; e-mail: martina.penke@uni-koeln.de

structure. The first question is formulated in active voice, whereas the second one contains a passive. Unlike in active voice structures where the agent is realized as the subject, in passive voice the agent role is demoted and the patient is realized as the subject instead. For decades, research in psycholinguistics has claimed that the resulting non-canonical linking of argument roles to syntactic functions is associated with a higher cognitive load. Indeed, from early on psycholinguistic research has provided abundant evidence suggesting that passive clauses are more difficult than active ones when it comes to their comprehension, acquisition, production and memorization (e.g., Borer & Wexler, 1987; Brown & Hanlon, 1970; Mehler, 1963; Miller, 1962; Savin & Perchonock, 1965). Later research provided yet more evidence for passive sentences being more difficult to process than active ones by healthy adults (Mack, Meltzer-Asscher, Barbieri, & Thompson, 2013), as well as by aphasic patients (Dickey & Thompson, 2009; Grodzinsky, 2000; Grodzinsky, 1990; Penke, 2015) and by children with developmental language disorders (Penke, 2015; Ring & Clahsen, 2005; van der Lely, 1996). A number of current theories propose plausible explanations for the assumption that passive forms are more difficult than active ones. For example, passives might be syntactically more complex than actives as they require additional operations (see Kiparsky, 2013, for a more recent view on this approach). With respect to sentence processing, for instance, the *Good Enough* theory (Christianson, Hollingworth, Halliwell, & Ferreira, 2001; Ferreira, 2003) suggests that comprehending a linguistic input is driven by a set of heuristics rather than engaging in detailed syntactic processing. When it comes to complex structures, the correct interpretation may require a revision of commonly applied heuristics. Unless such a revision is made (whenever the heuristic is considered “good enough”), applying the agent-first strategy to interpret the NP mentioned first in English would lead to misinterpretations of a passive sentence (i.e. interpreting *the clown* as an agent in *The clown is filmed by the fisherman*). Another theoretical approach views the relative frequency of a particular structure in a language as the main predictor of the processing difficulty related to that structure (Johns & Jones, 2015). Within this approach, the greater difficulty of passive structures is due to their less frequent occurrence compared to active ones. In English language production corpora, for instance, passives only occur in about 6% of the cases (Roland, Dick, & Elman, 2007). Such theories, supporting the idea that passives are more difficult for comprehension than actives, are easily generalized to production. The general assumption in this case is that the production of passive forms is considered more effortful than that of active forms (Ferreira, 1994; Tannenbaum & Williams, 1968).

However, psycholinguistic research has also shown that a number of factors might facilitate the production of passives. Providing speakers with passive voice sentences as primes, for instance, results in more frequent choices of passive voice transitive sentences over active voice alternatives – a phenomenon known as structural or syntactic priming (Bock & Griffin, 2000; Bock, Dell, Chang, & Onishi, 2007; Segaert, Menenti, Weber, & Hagoort, 2011). Similarly, the animacy status of a referent seems to modulate the number of produced passives. Having animate rather than inanimate entities as thematic patients, for instance, has been reported to reduce preferences for active compared to passive voice across languages (e.g., in English McDonald, Bock, & Kelly, 1993; in Spanish Prat-Sala, 1997; in German van Nice & Dietrich, 2003). In addition to animacy, perceptual priming has been shown to modify speakers' structural choice preferences. In Tomlin's "Fish film" (Tomlin, 1995), for instance, speakers saw a cartoon of one fish eating another and had to describe it. While priming the agent fish by means of a visually presented arrow resulted in active voice descriptions, priming the patient fish led to an increase in passive voice descriptions. Taken together, these results suggest that speakers' choices to use passives may depend on both the preceding context (presence or absence of passive sentence primes) and the characteristics of the referents (animate or primed patients).

1.1 Visual Attention and Passive Production across Languages

In a seminal paper Gleitman, January, Nappa, and Trueswell (2007) investigated whether and how manipulating visual attention would affect the formulation of utterances. In their study, participants described event scenes containing two characters, one of which was cued by a subliminally presented (60-75 ms) black square appearing at this character's position before the scene onset. Among a number of constructions, Gleitman and colleagues focused on the production of utterances in active or passive voice in English speakers by presenting scenes where an agent was acting on a patient (e.g., a man kicking a boy, Experiment 2). The results showed that participants not only made more initial fixations on the cued character but cueing the patient led to a significant increase in the proportion of utterances produced in passive voice. When the patient was cued, 26% of the produced utterances were passives (e.g., *The boy was kicked by the man*) as opposed to 15% when the agent was cued, indicating that initial looks to a character influenced the structural choices between active or passive voice (Gleitman et al., 2007).

For English speakers, this finding of Gleitman and colleagues was replicated by later research (Myachykov, Garrod, & Scheepers, 2011, 2018; Myachykov, Thompson, Garrod, & Scheepers, 2012). However, studies employing a similar experimental paradigm with speakers of other languages failed to find an increase of passive voice utterances after patient cueing. For instance, the Finnish speakers investigated by Myachykov et al. (2011) produced active utterances irrespective of whether the agent (100% actives) or the patient (99% actives) in the scene was cued. Likewise, Hwang and Kaiser (2015) found no influence of subliminal cueing on produced utterance structures in Korean. Their participants produced active voice utterances in 93% of the cases where the agent of a visual event scene was cued and in 94% of the cases where the patient was cued. In a study with Russian speakers, Myachykov and Tomlin (2008) found that participants produced active utterances independent of whether the agent (100% actives) or the patient (98%) was cued in a “Fish film” (as described above for Tomlin, 1995). Importantly, in the above-mentioned studies subliminal perceptual priming of the patient was successful in attracting visual attention as evidenced by initial fixations on the patient. However, in contrast to English, drawing speaker’s attention to the patient character did not affect their preference to describe depicted events using active voice structures.

The differences in the outcome of these perceptual priming studies have led to the proposal that the formulation of utterances in passive voice is dependent on specific grammatical characteristics of a given language (Myachykov et al., 2011; Norcliffe & Konopka, 2015). Myachykov et al. (2011) have suggested that the very low percentages of passives produced by Russian and Finnish speakers are due to the fact that passives are highly marked, dispreferred, and very infrequent in these languages. Moreover Finnish, Russian, and Korean, are all case-marking languages with relatively free word order in contrast to English: in these languages subject and object of a sentence display different case markings. In order to produce an utterance starting with a noun phrase a speaker, first has to determine its syntactic function for appropriate case-marking. It might be that the necessity to case-mark a noun phrase before initiating the utterance counteracts the potential effects of perceptual priming. Rather than linking the primed patient to the subject position, which would result in a passive, speakers of case-marking languages may instead rely on the canonical linking of the agent role to subject function resulting in an active voice sentence.

1.2 Sentence Planning Strategies

The difference in the propensity to produce passive utterances in perceptual priming experiments observed for English on the one hand and case-marking languages such as Russian, Finnish or Korean on the other hand might be related to two different mechanisms of sentence formulation that are discussed in current psycholinguistic research (Norcliffe & Konopka, 2015). According to one formulation strategy, linguistic encoding of a message might proceed in a linear incremental manner. Utterance formulation starts with the first lemma that is retrieved in the mental lexicon. This lemma then constrains the structure of the sentence to be produced and sentence formulation proceeds incrementally as more and more lemmas are retrieved (Gleitman et al., 2007; Norcliffe & Konopka, 2015). An umbrella term *conceptual accessibility* has been suggested by Bock and Warren (1985) to refer to all factors influencing the ease with which lemmas can be retrieved. Prat-Sala & Branigan (2000) further specified the notion of accessibility by distinguishing inherent (word-related properties, such as concreteness and animacy) and derived accessibility (context-related properties, such as priming or focus). Both word-related and context-related factors may affect the accessibility of a lemma and, subsequently, the structure of an utterance. According to this view, perceptual priming of an agent or a patient in a visually depicted event scene should result in the retrieval of a corresponding lemma, constraining the choice of voice of the produced utterance: agent priming will lead to agent-first active voice structures, whereas patient priming will result in a patient-initial structure such as passive voice. Gleitman et al. (2007) have argued that their results provide evidence for linear incrementality in sentence formulation, suggesting that speakers do not engage in pre-planning of an utterance structure but rather encode the accessible word which then constraints their selection of voice. Their English-speaking participants did not only make more initial fixations on the cued character but also mentioned it when first describing the scenes.

An alternative approach to utterance formulation assumes that speakers first apprehend the relational structure of a depicted event before planning an utterance. This view holds that the whole message is first planned conceptually rather than that an individual lemma is retrieved. A structural representation of a sentence is subsequently generated which then guides lemma retrieval (e.g., Bock & Griffin, 2000; Lee, Brown-Schmidt & Watson, 2013). Sauppe et al. (2013) provide evidence for this theoretical approach based on gaze patterns during picture description in Tagalog, an Austronesian verb-initial language that requires marking the privileged

syntactic argument (depending on the voice, agent or patient) on the verb. Speakers fixated the privileged syntactic argument within the first 600 ms and later fixated characters in the order of mention. The authors interpret the early fixations as related to the generation of the event structure and later fixations as reflecting lexical retrieval, supporting the notion of structural pre-planning. Similar evidence is presented by Norcliffe and Konopka (2015) on Tzeltal, a verb-initial Mayan language. The finding that in case-marking languages such as Finnish, Korean, or Russian, perceptual priming does not result in an increase of passive utterances might be another point in case where utterance formulation requires initial structural planning. As presented above, speakers of these languages do not start their event descriptions with the visually cued patient (i.e. the first lemma that may become accessible), presumably due to affordances by case marking. Since case marking of the depicted characters necessarily involves determining their syntactic functions first (as subjects and objects are marked by different case), this requires structural planning before starting the utterance with the first noun phrase.

1.3. Evidence from perceptual priming experiments in German

The aforementioned studies suggest that whether speakers adopt a formulation strategy assigning syntactic functions and generating the utterance structure first or the alternative strategy relying on the accessibility of an argument's lemma and producing utterances in a linear incremental manner strongly depends on the properties of a given language. German constitutes an interesting case in this respect. Like Russian, Finnish or Korean, it is a case-marking language, which allows relative flexibility in word order as compared to English. Agents may be placed first as subjects of active sentences (e.g., *Der_[NOM] Angler filmt den_[ACC] Clown* 'The fisherman_{AGENT} is filming the clown_{PATIENT}') and patients may be placed first as either subjects of passive sentences (e.g., *Der_[NOM] Clown wird vom_[DAT] Angler gefilmt* 'The clown_{PATIENT} is filmed by the fisherman_{AGENT}') or as topicalized objects (e.g., *Den_[ACC] Clown filmt der_[NOM] Angler* 'The clown_{PATIENT}, the fisherman_{AGENT} is filming'). Active voice sentences with a canonical subject-verb-object word order are more frequent than passives or sentences with topicalized objects. However, in contrast to Russian and Finnish, passives in German appear with a frequency comparable to English in language production corpora: in English, passives amount to 5-6% (Roland et al., 2007) and in German to 7-9% (Bader & Häussler, 2010).

Esaulova, Penke, and Dolscheid (2019) examined whether directing speakers' attention to patients by means of implicit 60 ms cueing modulates

their preferences to produce a passive. However, patient cueing resulted in only 6% passive voice utterances, comparable to the frequency of passives in German corpus data. Critically, the produced number of passives did not increase due to cueing. This finding presents an interesting contrast with production data in English reported by Gleitman et al. (2007) and Myachykov et al. (2011, 2018). While the frequency of passive voice occurrence in English corpora is comparable to that in German, cueing patients in English led to passives being produced over 20% of the time. Whereas perceptual priming with a subliminal cue increased the likelihood of produced passive voice structures in English, it did not affect structural choice preferences in German, despite a similar frequency of occurrence in language corpora. The observed difference between English and German provides counter-evidence to the suggestion made by Myachykov et al. (2011) that the insensitivity of structural choice preferences, observed in implicit perceptual priming experiments in Finnish or Russian, might be due to the fact that passives are highly marked, dispreferred, and very infrequent in these languages. This explanation cannot account for the insensitivity of German-speakers to perceptual priming as passives occur with a similar frequency in language production corpora in both English and German.

1.4 The Influence of Cue Types

In their perceptual priming experiment with English speakers, Myachykov et al. (2018) also manipulated the duration of the visual cue. Besides a short implicit cue (70 ms) that typically goes unnoticed by participants (e.g., Gleitman et al., 2007), the authors also presented an explicit visual cue (a red dot) of longer duration, i.e. for 700 ms. Critically, explicit cueing of the patient appeared to be more effective than implicit cueing in eliciting passives. In particular, explicit patient cueing led to an increase in the number of produced passive utterances of about 8% when compared to implicit cueing of the patient (cf. Myachykov et al., 2018). While these findings suggest that explicit cueing is a particularly effective manipulation for eliciting passive structures, this evidence is exclusively based on English. So far, to our knowledge, explicit cueing has not been employed with speakers of case-marking languages such as German. Hence it is unclear whether explicit cueing in these languages will lead to a comparable increase in passive voice utterances or whether the ineffectiveness of implicit cueing will also hold for explicit cueing in a case-marking language like German. In the present study we addressed this question.

2. Method

2.1 Goals and Hypotheses

One central objective of the present study was to examine whether manipulating attention beyond a subliminal level would have an effect on speakers' production of passive voice utterances in German. Answering this question can provide important insights into the factors determining structural choices. It is possible that German speakers are unaffected by explicit cueing of the patient (just like they are 'immune' to implicit cueing), resulting in no increase of produced passives. This would entail that German speakers' sentence production strategies differ even more substantially from those of English speakers (who are affected by implicit cueing but show even stronger effects of explicit cueing). German is an especially promising language to test this manipulation, since (1) it is a case-marking language, like Russian, Finnish, or Korean, but unlike English; and (2) passive voice production is not particularly infrequent, unlike Russian, Finnish, or Korean, but similar to English. These features allow us to clarify whether case-marking is indeed a factor generally preventing the effectiveness of visual cueing on structural choice. Alternatively, the observed differences of implicit cueing in English and German may be a matter of degree, rather than kind: given both the necessary case encoding and the larger choice of syntactic options in German, it is possible that attention manipulation at a subliminal level may not suffice for speakers to commit to passive voice formulations whereas a more explicit cueing of patients might result in more passives. Thus, German speakers may in principle be susceptible to effects of visual cueing but they may need more explicit cues in order to produce a greater number of passives.

To delineate between these two options, we tested German-speaking participants in an eye-tracking study where they had to describe scenes depicting an agent and a patient character. In half of the trials, the patient character was explicitly cued by means of a red dot. The present study is designed after Esaulova et al.'s (2019) experiment with the difference that we are using an explicit (700 ms) rather than an implicit (60 ms) cue. For a better comparability, two other factors – referent position and animacy – are also manipulated as in Esaulova et al. (2019). Keeping the design equivalent to Esaulova and colleagues' (2019) should allow us to establish whether it is indeed the duration of visual cueing that affects the production of passives in German, rather than alternative factors. If the salience of perceptual priming influences speakers' structural choices in German, we should observe an increase in the produced passive utterances when patients are

explicitly cued compared to when they are not cued. No increase in passive production would indicate that sentence planning is not susceptible to visual cueing in a case-marking language like German. Such an outcome would suggest that sentence planning strategies are affected by typological differences between languages.

As to the formulation of passive voice, we expect a higher difficulty (reflected in longer speech onset times) of passive compared to active voice utterances in accordance with the psycholinguistic evidence discussed above that claims passives to be more complex than actives. Moreover, provided that explicit priming results in a substantially higher number of passive forms produced by participants, we aim at taking a closer look at the planning process of passive utterances. While previous studies on sentence production mainly focused on manipulating factors that modify preferences for active or passive voice (e.g., animacy or cueing of agents/patients), they often left the question about speakers' formulation strategies open. In this study, we relate speech data to gaze behavior in order to better understand whether speakers may apply linear incremental or structural formulation strategies in response to the manipulated factors. Employing the eye-tracking technique is ideally suited for this purpose, since it provides information with high temporal resolution about the time-course of sentence planning as it unfolds in real time. In this respect, we expect to observe differences in gaze patterns associated with the production of active and passive voice utterances. Fixations on a patient rather than an agent before the initiation of a passive sentence, for instance, would be indicative of a linear incremental rather than structural formulation strategy. An opposite gaze pattern with fixations on an agent rather than a patient before the onset of the sentence could be expected for active utterances.

Note that Esaulova et al. (2019) tested the influence of two additional factors (besides implicit cueing) on sentence production: (i) the position of patient and agent vis-à-vis each other and (ii) the animacy of the patient. They found that the position of actants in a transitive event scene is the main factor influencing the speech onset of active sentences. Positioning patients to the left rather than to the right of agents led to a delayed initiation of utterances. Due to their low number, it was not possible for Esaulova et al. (2019) to determine whether the production of passive utterances may also be influenced by the position of actants. Should the present study indeed result in a higher number of passive sentences, it will then also be possible to clarify if the position of actants affects the production of passives. If this is the case, the speech onset times for both types of utterances should be

affected by the position of actants rather than by other manipulated factors, i.e. cueing and animacy.

2.2 Participants

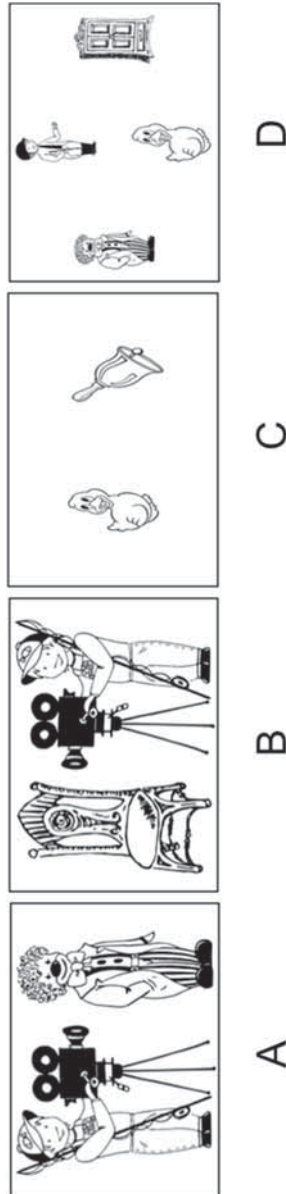
Forty-five native speakers of German (9 male and thirty-six female, mean age 23.83, $SD = 3.34$), all students at the University of Cologne, were either paid or received a course credit for taking part in the study. None of the participants reported the knowledge of another language at a native level. Participants were naive with respect to the hypotheses of the experiment, reported no language or attention related medical conditions and had normal or corrected-to-normal vision. Ethical approval was granted by the Ethics Commission of Cologne University's Faculty of Medicine.

2.3 Materials

A set of 143 black-and-white stimuli pictures were prepared for the experiment and included 56 experimental items displaying transitive event scenes, 56 fillers, 28 familiarization items and three practice items. Experimental items depicted transitive events with an animate agent and either an animate (28 items, Figure 1, A) or an inanimate (28 items, Figure 1, B) patient, where patients were displayed either to the right (28 items, Figure 1, A) or to the left (28 items, Figure 1, B) of agents. A blank screen (in the no cueing condition) or a cue (a red point centered in the right or left half of the screen where the patient would appear next) was presented for 700 ms preceding each experimental item. All agents and patients represented grammatically masculine nouns in German that did not differ in word length or in lemma frequencies collected using the archive of written language corpora provided by the Mannheim Institute for German Language (COSMAS II). The depicted agents and patients were comparable in terms of their size, visual complexity, and spatial distance between them.²

² For more details on event scenes used in this experiment and their pretesting, see Esaulova et al. (2019).

Figure 1. Examples of materials used in the present study: A, B – experimental items, C – filler item, D – familiarization item.



Fillers consisted of images representing two nouns that were displayed one above another (28 items) or next to each other (28 items, Figure 1, C). The latter were preceded by a cue appearing for 700 ms in the center of the right (14 items) or the left (14 items) side of the screen. The selected nouns were balanced with respect to animacy (animals and inanimate objects), grammatical gender (masculine and feminine nouns), and where their depictions appeared on the screen (left, right, top or bottom). Compared to experimental items, describing filler items did not involve transitive events and required the use of both feminine and masculine determiners, as well as dative and not only accusative and nominative case. Increasing the variation in produced descriptions minimized the likelihood for participants to adhere to a strategy of producing syntactically identical sentences throughout the experiment.

Familiarization items consisted of four depictions of nouns (Figure 1, D) used in experimental and filler items that were named during the familiarization phase (see Design and Procedure). These items were used to ensure that participants could easily recognize and name the depicted figures and objects during the picture description task.

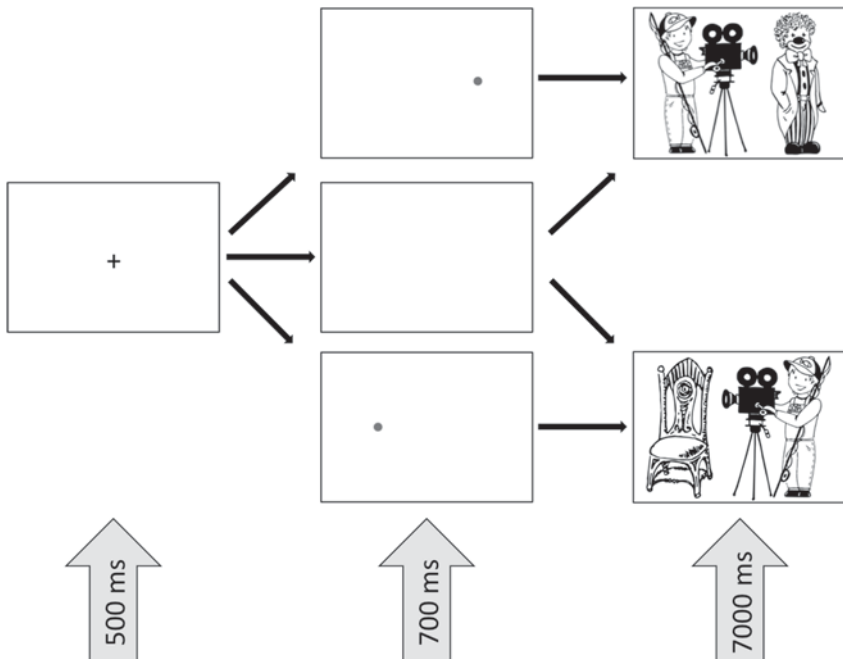
2.4 Design and Procedure

Four lists of stimuli were created with patient POSITION (on the right or on the left of the scene) and CUEING (cued vs. non-cued patient) as within subjects and within items factors, and patient ANIMACY (animate vs. inanimate) as a within subjects and between items factor. Each of the four lists contained all eight experimental conditions: animate cued patient on the left or on the right side of the agent, animate non-cued patient on the left or on the right side of the agent, inanimate cued patient on the left or on the right side of the agent, and inanimate non-cued patient on the left or on the right side of the patient. Each participant saw seven items per each of the eight conditions in a randomized order and each experimental item only once. Agent-patient combinations were never presented repeatedly within the same participant.

Participants were seated 60 cm from an LCD monitor equipped with an Eyelink 1000 Plus eye tracker (SR Research Ltd.). Eye movements were recorded with the 500 Hz sampling rate from the dominant eye after a nine-point calibration which was repeated whenever necessary to ensure the quality of recording. Throughout the experimental session participants wore a PC-headset Hama “Fire Starter” with a stereo headphone and a boom microphone with a frequency range of 50–5000 Hz. To make sure the

instructions are not a potential source of discrepancies in results, they were identical to those in Esaulova et al. (2019). Participants were asked to describe scenes that they see on the screen in one sentence using one of the three syntactic forms available in German (active voice, passive voice or object topicalization). The instructions were given via headphones and illustrated by examples and images on the screen. The stimuli list was presented as seven consecutive blocks, each block beginning with four familiarization items that contained figures and objects from eight experimental and eight filler items presented later in this block. As familiarization items appeared on the screen, participants had to answer questions that they heard (e.g., *Where is the clown?*) using corresponding arrow keys on the keyboard. The 7000-ms presentation of experimental and filler items was preceded by a 500 ms presentation of a fixation cross in the center of the screen and a 700 ms presentation of either a blank screen (in no cueing conditions) or a red circle subtending approximately 1° of visual angle as a cue centered in the left or the right half of the screen where the patient was then displayed (Figure 2).

Figure 2. The experimental paradigm used in the study.



2.5 Data Analysis

Data types collected in the experiment include language production and gaze data. Language production data were analyzed with respect to utterance types produced by speakers and their speech onset times (identified using the Praat software, Boersma & Weenink, 2017). Utterance types produced in response to the experimental items were active subject-verb-object clauses and full passive structures. No other structures were produced, so that the complete dataset could be used to analyze the probability of passivizations. As to speech onset times, 0.2% of recorded trials contained interruptions irrelevant to the task before utterances were produced and were excluded from the analyses. The analyzed gaze data were collected for two rectangular interest areas of the same size drawn around agents and patients and did not require corrections or exclusions.

Statistical analyses were carried out in R (*RStudio*, 2017) and included mixed-effects linear and logistic regression modeling, as well as *t*-tests. The *lme4* package (Bates, Mächler, Bolker, & Walker, 2014) was used to analyze binomial data (the probability of passive utterances) using *glmer* function, as well as continuous data (speech onset times and gaze fixation times) using *lmer* function. Sum-coded contrasts were assigned to POSITION, CUEING, and ANIMACY factors as categorical predictors (e.g., Barr et al. 2013; Levy, 2014) and included in the models with interactions between them as fixed effects. Participants and items were included in the models as random effects (e.g., Baayen, Davidson & Bates, 2008). Models were built systematically reducing the maximal structure and then compared to each other. The best-fitting model was selected based on the lowest AIC value. To account for the non-normality of the distribution, speech onset times and gaze fixation times data were transformed as determined by the Box-Cox procedure (Osborne, 2010). Welch's *t*-test was carried out to compare samples of different sizes with unequal variances (e.g., speech onsets for active vs. passive utterances). The assumptions of all statistical tests reported in the results were otherwise met. Study materials, raw data and analytic methods will be made available by the authors to any qualified researcher upon request.

3. Results

3.1 The Occurrence of Passive Voice

The production of passive voice structures across conditions occurred 278 times, which is 11.03% of the total number of produced 2520 utterances,

the rest of which constituted active voice structures with canonical subject-verb-object order (see Table 1 for proportion of passive and active utterances produced in each experimental condition).

Table 1. The number (and relative percentage per condition) of passive and active utterances produced in each of the 8 experimental conditions: animate and inanimate patients on the left and on the right of the agent after cueing and no cueing.

	<i>animate patients</i>				<i>inanimate patients</i>			
	<i>left</i>		<i>right</i>		<i>Left</i>		<i>right</i>	
	<i>cued</i>	<i>non-cued</i>	<i>cued</i>	<i>non-cued</i>	<i>cued</i>	<i>non-cued</i>	<i>cued</i>	<i>non-cued</i>
passive	43 (13.65)	29 (9.21)	45 (14.29)	36 (11.43)	41 (13.02)	23 (7.30)	39 (12.38)	22 (6.98)
active	272 (86.35)	286 (90.79)	270 (85.71)	279 (88.57)	274 (86.98)	292 (92.70)	276 (87.62)	293 (93.02)

The mixed-effects logistic regression model on the probability of passive voice descriptions revealed two main effects: a main effect of CUEING and a main effect of ANIMACY (Table 2). The main effect of CUEING was due to a higher probability of passive voice descriptions after cued ($M = 0.13$, $SD = 0.34$, $SE = 0.01$) than non-cued patients ($M = 0.09$, $SD = 0.28$, $SE = 0.01$). The main effect of ANIMACY showed that the probability of passive voice production was significantly higher after scenes with animate ($M = 0.12$, $SD = 0.33$, $SE = 0.01$) rather than inanimate patients ($M = 0.10$, $SD = 0.30$, $SE = 0.01$).

Table 2. Main effects and interactions from the mixed-effects logistic regression model on the probability of passive utterances.

	<i>b</i>	<i>SE</i>	<i>z</i>	<i>p</i>
Intercept (estimated grand mean)	-6.31140	1.48552	-4.25	<.001***
Animacy	0.26764	0.09913	2.70	.007**
Position	-0.04861	0.08914	-0.55	.586
Cueing	-0.45473	0.09168	-4.96	<.001***
Animacy X Position	-0.06625	0.09288	-0.71	.476
Animacy X Cueing	0.13549	0.09648	1.40	.160
Position X Cueing	-0.04337	0.08949	-0.49	.628
Animacy X Position X Cueing	-0.06123	0.09249	-0.66	.508

3.2 Passive Vs. Active Voice: Speech Onset Times and Gaze Behavior

When averaged across conditions, the initiation of passive voice utterances was significantly faster ($M = 1530.76$, $SD = 488.50$) than that of active voice utterances ($M = 1690.77$, $SD = 599.88$), $t(388.86) = -5.01$, $p < .001$.³ Moreover, before the average onset time of passive utterances the probability of fixations on patients ($M = 0.53$, $SD = 0.21$, $SE = 0.01$) was higher than that on agents ($M = 0.27$, $SD = 0.18$, $SE = 0.01$), $t(277) = -15.62$, $p < .001$. The reverse pattern was observed for active utterances: the probability of fixations on agents before the average onset time of active utterances ($M = 0.49$, $SD = 0.18$, $SE = 0.004$) was higher than that on patients ($M = 0.27$, $SD = 0.17$, $SE = 0.004$), $t(2241) = 43.489$, $p < .001$. In terms of gaze patterns, the average start times of first fixations on agents relative to scene onsets were later for passive utterances ($M = 2062.22$, $SD = 455.30$, $SE = 29.57$) than for active ones ($M = 1671.67$, $SD = 211.12$, $SE = 4.47$), $t(246.87) = 13.06$, $p < .001$. Complementing this pattern, the start times of first fixations on patients were earlier for passive utterances ($M = 1529.82$, $SD = 305.73$, $SE = 18.44$) than for active ones ($M = 1706.35$, $SD = 520.14$, $SE = 11.77$), $t(530.63) = -8.07$, $p < .001$. Figure 3 reflects these differences in speech onsets and gaze when active and passive utterance types were produced.

The mixed-effects linear regression model on speech onset times of produced passive structures (Table 3) revealed a main effect of CUEING. Speech onset times of passive voice descriptions after cued patients were shorter ($M = 1421.83$, $SD = 440.21$, $SE = 33.96$) than when no cueing occurred ($M = 1697.14$, $SD = 513.14$, $SE = 48.93$). In addition, the model showed a main effect of ANIMACY with passive utterances initiated faster after inanimate ($M = 1426.96$, $SD = 394.98$, $SE = 35.33$) rather than animate ($M = 1615.57$, $SD = 539.91$, $SE = 43.65$) patients. Taking a closer look at the speech onsets of passive utterances across conditions (Figure 3) indicates that longer speech onsets after animate patients are mainly due to non-cued ($M = 1823.00$) rather than cued ($M = 1455.02$) animate patients. At the same time, the analysis of active voice speech onsets (Table 4) revealed that the initiation of active utterances remained unaffected by both ANIMACY and CUEING of patients but instead showed a main effect of

³ The same difference is observed when speech onset times of passive utterances ($M = 1530.76$, $SD = 488.50$) are compared to those of active utterances ($M = 1717.79$, $SD = 555.27$, $SE = 19.83$) produced by the same speakers, $t(548.25) = -5.29$, $p < .001$.

POSITION. Active voice descriptions had shorter onsets when patients appeared to the right ($M = 1666.58$, $SD = 590.29$, $SE = 17.69$) than to the left ($M = 1714.76$, $SD = 608.56$, $SE = 18.16$) of agents.

Figure 3. Probability of looks to agents and patients in all conditions when passive and active utterances were produced.

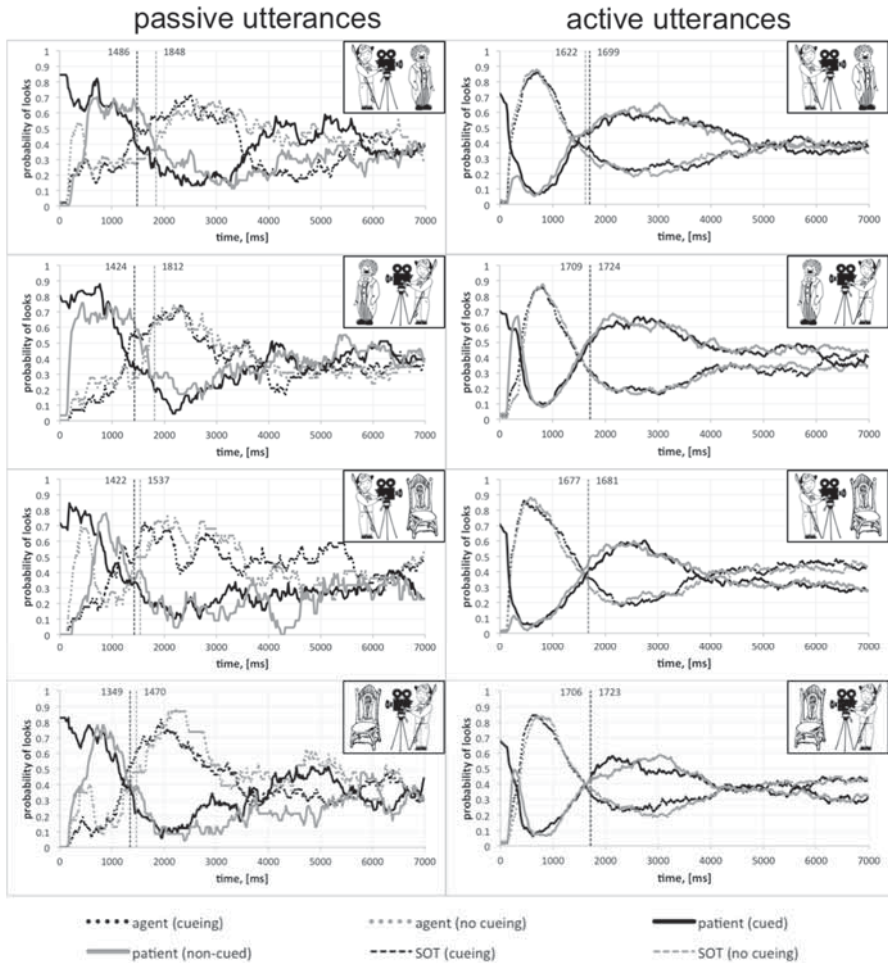


Table 3. Main effects and interactions from the mixed-effects linear regression model on speech onset times of passive utterances.

	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Intercept (estimated grand mean)	0.02521	0.00053	47.90	<.001***
Animacy	-0.00053	0.00020	-2.59	.010*
Position	0.00005	0.00018	0.29	.771
Cueing	-0.00083	0.00019	-4.35	<.001***
Animacy X Position	0.00009	0.00019	0.47	.637
Animacy X Cueing	-0.00011	0.00020	-0.56	.575
Position X Cueing	-0.00015	0.00018	-0.81	.417
Animacy X Position X Cueing	0.00004	0.00019	0.19	.851

* $p < .05$, ** $p < .01$, *** $p < .001$.

Table 4. Main effects and interactions from the mixed-effects linear regression model on speech onset times of active utterances.

	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Intercept (estimated grand mean)	0.02518	0.00036	70.23	<.001***
Animacy	0.00004	0.00007	0.49	.627
Position	-0.00016	0.00006	-2.64	.008**
Cueing	0.00006	0.00006	1.04	.298
Animacy X Position	0.00000	0.00006	-0.08	.939
Animacy X Cueing	0.00000	0.00007	0.03	.977
Position X Cueing	-0.00005	0.00006	-0.80	.424
Animacy X Position X Cueing	-0.00007	0.00006	-1.10	.270

* $p < .05$, ** $p < .01$, *** $p < .001$.

Gaze patterns before the initiation of both passive (Table 5) and active (Table 6) utterances were affected by all three factors: CUEING, ANIMACY, and POSITION of patients. In both cases, total fixation times were longer on cued than non-cued, animate than inanimate, and left- than right-positioned patients (see Table 7). In addition, the analysis of gaze before the onset of speech revealed an interaction between the factors CUEING and POSITION for both passive and active utterances. Before passive utterances (Table 8), total fixation times on left-positioned patients were significantly longer than on

right-positioned ones in the no cueing condition but did not differ significantly when patients were cued. Fixations on cued and non-cued patients also did not differ depending on patients' position. Before active utterances, total fixation times on left-positioned patients were significantly longer than on right-positioned ones in both cueing conditions. At the same time, total fixation times after cueing were longer on both left- and right-positioned patients than when no cueing occurred.

Table 5. Main effects and interactions from the mixed-effects linear regression model on total fixation times of patients before the onset of passive utterances.

	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Intercept (estimated grand mean)	29.76100	1.22300	24.34	<.001***
Animacy	-2.85300	1.26300	-2.26	0.026*
Position	-5.74900	1.28400	-4.48	<0.001***
Cueing	-3.72000	1.67600	-2.22	0.034*
Animacy X Position	2.45700	1.85300	1.33	0.186
Animacy X Cueing	2.25200	1.81600	1.24	0.216
Position X Cueing	6.62600	1.81500	3.65	<0.001***
Animacy X Position X Cueing	-3.60800	2.58200	-1.40	0.163

* $p < .05$, ** $p < .01$, *** $p < .001$.

Table 6. Main effects and interactions from the mixed-effects linear regression model on total fixation times of patients before the onset of active utterances.

	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Intercept (estimated grand mean)	4.09907	0.11065	37.05	<.001***
Animacy	-0.42985	0.14164	-3.04	0.003**
Position	-0.76404	0.13425	-5.69	<0.001***
Cueing	0.31698	0.14059	2.26	0.025*
Animacy X Position	-0.07472	0.17638	-0.42	0.672
Animacy X Cueing	0.11519	0.18186	0.63	0.527
Position X Cueing	0.44067	0.17005	2.59	0.010**
Animacy X Position X Cueing	0.04816	0.24491	0.20	0.844

* $p < .05$, ** $p < .01$, *** $p < .001$.

Table 7. Means (with standard deviations and standard errors) of total fixation times before the onset of passive and active utterances for each factor level.

	Animacy		Position		Cueing	
	animate	inanimate	left	right	cued	non-cued
passive	827.75	737.50	833.38	742.90	816.71	755.42
	(300.95; 24.33)	(295.75; 26.45)	(276.76; 23.73)	(318.07; 26.69)	(301.40; 25.12)	(299.40; 25.86)
active	388.82	303.85	383.18	308.23	396.06	298.08
	(252.64; 7.59)	(223.60; 6.64)	(243.44; 7.26)	(234.91; 7.03)	(241.41; 7.31)	(232.99; 6.87)

Table 8. Pairwise comparisons for POSITION*cueing interaction in total fixation times on patients before the onset of passive and active utterances.

		<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>
<i>passive utterances</i>					
cued	(Intercept)	25.58245	1.37146	18.65	<0.001***
	right	0.30676	1.04858	0.29	0.780
non-cued	(Intercept)	28.43382	0.74365	38.24	<0.001***
	right	-5.58301	1.93134	-2.89	0.014*
left	(Intercept)	28.50843	0.76624	37.21	<0.001***
	cued	-1.97139	1.68981	-1.17	0.270
right	(Intercept)	22.50739	1.80001	12.50	<0.001***
	cued	3.96982	1.94220	2.04	0.065
<i>active utterances</i>					
cued	(Intercept)	19.54694	0.54161	36.09	<0.001***
	right	-2.09494	0.54025	-3.88	<0.001***
non-cued	(Intercept)	16.89142	0.55421	30.48	<0.001***
	right	-3.75089	0.55112	-6.81	<0.001***
left	(Intercept)	16.92161	0.54926	30.81	<0.001***
	cued	2.69227	0.49430	5.45	<0.001***
right	(Intercept)	13.14585	0.58707	22.39	<0.001***
	cued	4.23745	0.61032	6.94	<0.001***

4. Discussion

In this study we examined whether explicit visual cueing affected the production rate of passive voice utterances in German. In contrast to research in English (e.g., Gleitman et al., 2007; Myachykov et al., 2011, 2018) where implicit visual cueing (60-75 ms) of a patient in an event scene led to an increase in the production of passive voice descriptions of the scene, no such effect was found in a previous study by Esaulova and colleagues (2019) for German. In the experiment reported in this paper, we increased the cue duration to 700 ms and observed a greater number of produced passive utterances, as well as some important differences in speech onsets and gaze behavior related to passive and active voice production. These observations shed light on factors that influence speakers' choice to produce passive voice descriptions with regard to transitive event scenes and suggest qualitative differences in the planning of passives in German.

4.1 Explicit Cueing Affects the Choice of Passive Voice

Given that the occurrence of passive utterances has nearly doubled (from 6% reported in Esaulova et al., 2019, to 11% in the present study) when patients in transitive event scenes were cued for 700 ms compared to 60 ms, the production of passive voice indeed seems to be susceptible to attention manipulations. The observed increase in passive utterances is comparable to findings by Myachykov and colleagues (2018) who reported an increase of 8% in the proportion of passive utterances when a long (700 ms) cue was employed instead of a short (70 ms) cue in a perceptual priming experiment conducted with English speakers. Importantly, Esaulova and colleagues (2019) did not find any evidence that implicit 60 ms cueing had any effect at all on the production of passive voice descriptions in German speakers, despite the fact that such cueing effectively led to more initial looks to patients. Instead, a closer analysis of the 6% passive utterances that were produced in Esaulova and colleagues' (2019) study suggested that it was the patients' position and animacy that were decisive for speakers' choice of passive structures: more passives were produced to describe scenes with left- than right-positioned and animate than inanimate patients. In contrast, in the present study explicit cueing did not only lead to more initial looks to patients but also to more passive voice structures produced after cueing than when no cueing occurred. These findings suggest that case-marking does not prevent effects of visual cueing altogether but that German speakers simply need more explicit cues than speakers of English in order to produce

a greater number of passives. Given the necessary case encoding and the larger choice of syntactic options in German as opposed to English, attention manipulation at a subliminal level does not seem to suffice for Germans to commit to passive voice formulations but more explicit cueing of patients results in an increase of produced passives.

Although the increase in the production of passive voice utterances in German observed in the present study is in line with previous findings in English (Gleitman et al., 2007; Myachykov et al., 2011, 2018, 2012) in that orienting attention via cueing appears to effectively modulate speakers' syntactic choices. Nevertheless, English and German speakers differ with respect to the production of passives in perceptual priming experiments at least in two ways. For one, implicit cueing (60 ms) does not result in an increase of passive utterances in German as opposed to English, but it takes an explicit cue to raise German speakers' propensity to produce an utterance in passive voice. Secondly, a difference in production rates for passives in English (28% after explicit cueing in Myachykov et al., 2018) and in German speakers (11%) remains. These differences suggest that speakers of German are not as sensitive to perceptual salience in sentence planning as speakers of English. Since German differs from English in terms of case-marking but not in terms of frequency of passive occurrence, case-marking seems a likely candidate accounting for the observed differences (see also Myachykov et al. (2011) for similar arguments). Alternatively, differences in experimental designs may have contributed to the diverging observations in speakers of English vs. German. While previous studies with English speakers employed cueing of both agents and patients, in our present study cueing manipulations were restricted to the patient (also see Esaulova et al., 2019). However, it should be noted that cross-linguistic differences between English and a case-marking language like Korean were still observed even when the experimental design was equivalent and involved implicit cueing of both agents and patients (e.g. Hwang & Kaiser, 2015). These findings suggest that differences in the experimental design are rather unlikely to explain the observed cross-linguistic patterns between speakers of German and English. While future work is necessary to directly compare the effects of explicit cueing in typologically different languages, our results demonstrate that factors like case-marking can modulate the degree to which speakers are susceptible to effects of visual cueing.

At the same time, our current findings depart from our previous observations regarding the effectiveness of patient position. Unlike in Esaulova et al. (2019), the position of a patient relative to an agent was no longer decisive for the choice of active or passive voice. Since the design of the present

study and Esaulova et al. (2019) was identical apart from the duration of the cue, the observed difference can only be attributed to cue duration. Consequently, our findings seem to indicate that speakers' structural choices may depend on how easily (and early) they can access certain types of information. Being unaware of the 60 ms cue, speakers accessed the patients' position first and used it to accommodate their choice of utterance type. Instead, when the cue was explicit and therefore accessible earlier than any other characteristics of the event, speakers' choice of passive voice was primarily guided by cueing. Taken together, our findings suggest that various visual properties of a scene – such as patient position or visual cueing – can impinge on sentence production but that their relative weighting seems to change, depending on the particular type of manipulation. Once a particular property becomes more salient (e.g. the visual cue due to its increased duration), this can override the importance of other factors (in this case character position), pointing to an intricate interrelation between various properties of an event scene.

4.2 Different Factors Influence the Production of Active and Passive Utterances

Considering a long tradition of experiments that started by Tomlin's "Fish film" (Tomlin, 1995) and related attention orienting to the production of particular sentence structures, it is worth noting that orienting attention to a particular character participating in an event may not necessarily result in modulations of speakers' syntactic choices. As our study demonstrates, cueing does not seem to be the only factor affecting visual attention when it comes to transitive event scenes: gaze patterns before speech onsets of both passive and active utterances show clear visual preferences based on the patients' position and its animacy in addition to cueing. Left-positioned patients, as well as animate patients, seem to effectively attract visual attention compared to right-positioned or inanimate patients. Moreover, more passive descriptions were elicited in response to scenes with animate compared to inanimate patients, confirming previous findings that animate entities are more likely to be assigned subject functions or to occupy initial sentence positions (Bock & Levelt, 1994; Branigan, Pickering, & Tanaka, 2008), as is the case in a German passive clause.

While factors such as cueing, position and animacy of a patient influence scene description, a finer differentiation might be fruitful when it comes to the influence of these factors on the production of utterances in active or passive voice. However, a comparison between speech onsets for active and

passive utterances is often not undertaken, due to a much lower number of passives compared to actives (Esaulova et al., 2019; Ferreira, 1994; Konopka, 2018). This may result in loss of information that is potentially critical for the understanding of sentence planning and production. The results of the current study indeed suggest that the initiation of active and passive utterances appears to be influenced by different factors: whereas both cueing and animacy affect speech onsets of passive utterances, it is the position of participants in the scene that mainly influences the initiation of active structures. Speech onset times were significantly longer when a patient was depicted to the left of the agent character than when it appeared to the right of the agent, reflecting additional costs in sentence formulation. Our finding that positioning of actants and objects in event scenes influences onset times of active voice utterances confirms a similar result in a previous study by Esaulova and colleagues (2019). The strong effect exerted by the position of a patient vis-à-vis an agent on the initiation of active voice utterances reflects the speaker's tendency to expect the agent to the left of the patient and to assign the subject function to a left-positioned rather than a right-positioned character, an expectation possibly shaped by reading habits (see Esaulova, Dolscheid, Reuters, & Penke, 2020).

4.3 Planning a Passive Utterance is not Always Costly

Our results show that passive voice scene descriptions were initiated on average about 160 ms faster than active voice descriptions. This finding is in stark contrast with the assumption expressed in much psycholinguistic research that passives should be associated with more processing costs and thus take longer to be produced due to their deviance from canonical linking and to their lower frequency. However, previous evidence supporting the claim of longer speech-onset times for passive utterances is not as unequivocal as might be thought. Myachykov and colleagues (2018), for instance, report longer times for sentences produced after patient-cueing compared to agent-cueing and conclude from this that planning and producing of canonical active sentences is easier and quicker compared to passive utterances. However, a comparison of speech onset latencies of active and passive utterances is not provided in this study. Sometimes findings suggesting a greater difficulty of passives seem to overshadow findings that indicate the contrary. Consider the study by Ferreira (1994), where she finds no difference in the formulation times for active and passive sentences from word triples (e.g., from *DIETER – HYPNOTIST – TRUSTED*) in Experiment 2 but longer formulation times for passives in Experiments 3 and 4. While the study is often referenced as demonstrating

how verb type and animacy may modulate the production of passives (Bader, Ellsiepen, Koukouloti, & Portele, 2017; Myachykov et al., 2011), the result of Experiment 2 does not seem to be as widely discussed in the literature even though it is certainly not accidental. In the same vein, the comprehension of passives may not always be more effortful than the comprehension of canonical sentences. Paolazzi et al. (2019), for instance, found that passive forms are only more difficult to comprehend than active when it comes to stative but not eventive passives. Such fine differentiations, however, are rarely made or followed-up on.

Our study confirms that a finer-grained investigation of the factors influencing the production of passive utterances results in a more differentiated picture regarding the processing costs involved. The initiation of passive utterances was influenced by the factors CUEING and ANIMACY. Cueing patients led to faster speech onsets compared to non-cued ones, whereas animate patients slowed down the utterance onset compared to inanimate patients. Indeed, as indicated in Figure 3, the only conditions where the production of passives was actually initiated later than the production of actives were the two conditions that displayed a non-cued animate patient besides the agent. In all other experimental conditions, passives were initiated faster than actives. This discrepancy in speech onsets of passive utterances suggests that different mechanisms and processes are involved in passive production across experimental conditions. In particular, we propose that German speakers employ both utterance planning strategies in formulating passive utterances, linear incremental planning as well as structural planning (see Sentence Planning Strategies section), depending on the specific context of experimental conditions tested.

4.4 Linear Incremental Planning of German Passives

A central finding of our study is that cueing of patients results in more passives that are initiated faster than utterances describing events with non-cued patients. Why are passive utterances produced so quickly, i.e. faster than utterances in active voice, in these experimental conditions? Following Gleitman and colleagues' (2007) lemma-activation account, we assume that visual cueing of a patient draws a speaker's attention to it, resulting in an increased activation of the corresponding lemma and, thereby, of its phonological form (i.e. lexeme). The increase in activation makes the lexical entry more likely to be produced first, starting the sentence with the patient. In contrast to English, however, German nouns have a grammatical gender, i.e. masculine, feminine or neuter. Grammatical gender is

considered to be an arbitrary, lexically stored, syntactic feature of the noun (Schriefers, Hantsch & Jescheniak, 2002). Gender in German is expressed on the determiner preceding the noun. According to current psycholinguistic models, activation of a noun's lemma activates this gender information and a determiner is selected accordingly (Schiller & Caramazza, 2003; Schriefers et al., 2002). As German determiners fuse gender, number and case information, the activated determiner also necessarily expresses case information. In the default case, this is nominative case (Emonds, 1985). Thus, cueing of the patient in a scene where a fisherman is filming a clown (e.g., Figure 1, A), is likely to result in activation and production of the noun phrase *der Clown* where the determiner *der* expresses the default nominative case besides masculine gender. Given the activation of this noun phrase, the easiest way to proceed with the utterance might be to go on with a passive clause where the patient is realized as the nominative case-marked subject (i.e. *Der Clown wird vom Angler gefilmt*. 'The clown is being filmed by the fisherman'). The suggested process seems to be well-supported by participants' gaze behavior. Looks to patients are being initiated earlier and are only directed to agents shortly before the onset of passive utterances. The proposed formulation strategy is also in line with the 'Principle of Immediate Mention' proposed by Ferreira & Dell (2000, p. 299): "Production proceeds more efficiently if syntactic structures are used that permit quickly selected lemmas to be mentioned as soon as possible". The suggested process conforms to the proposal of Gleitman and colleagues (2007) that linguistic encoding of a message proceeds in a linear incremental manner, building-up from the first lemma retrieved in the mental lexicon. This lemma constrains the structure of the sentence to be produced, resulting in a passive clause in both English and German since the activation of a nominative noun phrase in German leaves the passive as the easiest and thus least costly option for the grammatical continuation of the utterance.

4.5 Structural Planning of German Passives

A different process might be involved in the formulation of passives in experimental contexts where scenes depicted animate patients and no cueing occurred. In these conditions, subjects took particularly long to initiate a passive utterance ($M = 1823.00$) compared to the other experimental conditions. It seems as if the presence of two animate characters in the scene and the lack of cueing as an attention directing device requires the speakers to first apprehend the depicted scene and to identify the roles these animate characters play in the depicted event. Rather than following a linear

incremental formulation strategy driven by lemma activation, as observed for cued patients, the strategy to produce passive in these cases is driven by first apprehending the relation of the characters depicted in the scene before a decision as to the structure of the utterance to be produced is made. This initial ascertainment of an animate character as a patient and not an agent in the depicted scene during utterance planning is not only reflected in longer speech-onset times. Gaze patterns show correspondingly longer-lasting fixations on these patients before passive utterances are produced. This performance is in line with the proposal that utterance formulation proceeds after speakers have identified the relational structure of a depicted event (Griffin & Bock, 2000; Lee et al., 2013). It also suggests that speakers attempt to rely on semantic features of characters in the scene to aid the decision about their syntactic function (see Kirby, 2010, for evidence from language acquisition). When the character possesses proto-agent features like sentience (Dowty, 1991) that are also more prevalent in animacy (e.g. only animates but not inanimates are able to perceive), matching an inanimate entity with the patient role results in additional costs.

Summarizing, our data suggest that speakers of German employ two different strategies when they formulate passives, namely linear-incremental planning and structural planning, depending on the particular circumstances defined by experimental manipulations. While future studies are necessary to further illuminate the conditions under which a particular strategy is preferred and whether or not one of the strategies can be considered the default, our findings emphasize the flexible use of different strategies within a single language. That speakers may make use of different formulation strategies depending on the experimental context is in line with Kuchinsky & Bock (2010), who found that speakers mentioned the cued character early when its role in the scene was easily interpretable but found no such effect when events were difficult to interpret (also see van de Velde, Meyer & Konopka, 2014, for converging evidence). Our findings add to a body of evidence indicating that the choice of formulation strategies made by speakers may be quite flexible and differ not only between languages but also within the same language depending on the context (Norcliffe & Konopka, 2015).

5. Conclusions

The present study examined whether the visual cueing of patients in transitive event scenes increases the probability of produced passive utterances in German. Our findings demonstrate that explicit visual cueing of patients, as opposed to implicit cueing (Esaulova et al., 2019), efficiently

increased the probability of passive utterances produced by German speakers. As to the planning and production of passive utterances, our data indicate two distinct strategies applied by speakers. Based on the observed speech onsets and gaze patterns, we argue that speakers proceed in a linear incremental manner when formulating a passive voice in response to patient cueing. When the manipulation of animacy resulted in the production of passives, i.e. in conditions where two animate (non-cued) characters were depicted, speakers applied a structure-driven strategy. Identifying and matching an animate patient with the patient role and subject function was associated with more costs than retrieving the first lemma made accessible by cueing. Consequently, rather than merely identifying animacy and cueing as factors triggering passive voice, our study encourages to distinguish specific mechanisms underlying each of these factors as a promising approach that accounts best for the flexible strategies speakers have at their disposal in producing passives.

Acknowledgments

This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 281511265 – SFB 1252. We are thankful to our fellow SFB 1252 colleagues, as well as colleagues from SFB 1102 “Information Density and Linguistic Encoding” and SFB 1287 “Limits of Variability in Language” for the feedback and support we received at the SFB-Networking Workshop Language Processing (Potsdam, 2019). We are also grateful for the helpful comments from the participants at the Experimental Psycholinguistics Conference (Palma de Mallorca, 2019).

References

- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412.
<https://doi.org/10.1016/j.jml.2007.12.005>
- Bader, M., & Häußler, J. (2010). Word order in German: A corpus study. *Lingua*, 120(3), 717–762. <https://doi.org/10.1016/j.lingua.2009.05.007>
- Bader, M. J., Ellsiepen, E., Koukouliti, V., & Portele, Y. (2017). Filling the prefield: Findings and challenges. In C. Freitag, O. Bott, & F. Schlotterbeck (Eds.), *Two perspectives on V2: The invited talks of the DGfS 2016 workshop “V2 in grammar and processing: Its causes and its consequences”* (pp. 27–49). Konstanz: University of Konstanz.

- 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(3), 255–278.
<https://doi.org/10.1016/j.jml.2012.11.001>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4. *ArXiv:1406.5823 [Stat]*. Retrieved from <http://arxiv.org/abs/1406.5823>
- Bock, J. K., & Warren, R. K. (1985). Conceptual accessibility and syntactic structure in sentence formulation. *Cognition*, 21(1), 47–67.
- Bock, K., & Griffin, Z. M. (2000). The persistence of structural priming: transient activation or implicit learning? *Journal of Experimental Psychology: General*, 129(2), 177–192.
- Bock, K. J., & Levelt, W. J. (1994). Language Production: Grammatical Encoding. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 945–984). New York: Academic Press.
- Bock, K., Dell, G. S., Chang, F., & Onishi, K. H. (2007). Persistent structural priming from language comprehension to language production. *Cognition*, 104(3), 437–458.
<https://doi.org/10.1016/j.cognition.2006.07.003>
- Boersma, P., & Weenink, D. (2017). Praat: doing phonetics by computer (Version 6.0.32). Retrieved from <http://www.praat.org/>
- Borer, H., & Wexler, K. (1987). The Maturation of Syntax. In T. Roeper & E. Williams (Eds.), *Parameter Setting* (pp. 123–172).
https://doi.org/10.1007/978-94-009-3727-7_6
- Branigan, H. P., Pickering, M. J., & Tanaka, M. (2008). Contributions of animacy to grammatical function assignment and word order during production. *Lingua*, 118(2), 172–189.
<https://doi.org/10.1016/j.lingua.2007.02.003>
- Brown, R., & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In J. R. Hayes (Ed.), *Cognition and the development of language*. New York: Wiley.
- Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic Roles Assigned along the Garden Path Linger. *Cognitive Psychology*, 42(4), 368–407. <https://doi.org/10.1006/cogp.2001.0752>
- Dickey, M. W., & Thompson, C. K. (2009). Automatic processing of wh- and NP-movement in agrammatic aphasia: Evidence from eyetracking. *Journal of Neurolinguistics*, 22(6), 563–583.
<https://doi.org/10.1016/j.jneuroling.2009.06.004>
- Dowty, D. (1991). Thematic Proto-Roles and Argument Selection. *Language*, 67(3), 547–619. <https://doi.org/10.2307/415037>

- Emonds, J. E. (1985). *A unified theory of syntactic categories*. Dordrecht: Foris Publications.
- Esaulova, Y., Penke, M., & Dolscheid, S. (2019). Describing events: Changes in eye movements and language production due to visual and conceptual properties of scenes. *Frontiers in Psychology*, 10:835. <https://doi.org/10.3389/fpsyg.2019.00835>
- Esaulova, Y., Dolscheid, S., Reuters, S., & Penke, M. (2020). *The alignment of agent-first preferences with visual event representations in German vs. Arabic speakers*. Manuscript submitted for publication.
- Ferreira, F. (1994). Choice of Passive Voice is Affected by Verb Type and Animacy. *Journal of Memory and Language*, 33(6), 715–736. <https://doi.org/10.1006/jmla.1994.1034>
- Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, 47(2), 164–203.
- Ferreira, V. S., & Dell, G. S. (2000). Effect of Ambiguity and Lexical Availability on Syntactic and Lexical Production. *Cognitive Psychology*, 40(4), 296–340. <https://doi.org/10.1006/cogp.1999.0730>
- Gleitman, L. R., January, D., Nappa, R., & Trueswell, J. C. (2007). On the give and take between event apprehension and utterance formulation. *Journal of Memory and Language*, 57(4), 544–569. <https://doi.org/10.1016/j.jml.2007.01.007>
- Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, 11(4), 274–279. <https://doi.org/10.1111/1467-9280.00255>
- Grodzinsky, Y. (2000). The neurology of syntax: language use without Broca's area. *The Behavioral and Brain Sciences*, 23(1), 1–71.
- Grodzinsky, Y. (1990). *Theoretical Perspectives on Language Deficits*. Cambridge, MA: M.I.T. Press.
- Hwang, H., & Kaiser, E. (2015). Accessibility effects on production vary cross-linguistically: Evidence from English and Korean. *Journal of Memory and Language*, 84, 190–204. <https://doi.org/10.1016/j.jml.2015.06.004>
- Johns, B. T., & Jones, M. N. (2015). Generating structure from experience: A retrieval-based model of language processing. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 69(3), 233–251. <https://doi.org/10.1037/cep0000053>
- Kiparsky, P. (2013). Towards a null theory of the passive. *Lingua*, 125, 7–33. <https://doi.org/10.1016/j.lingua.2012.09.003>
- Kirby, S. (2010). Passives in first language acquisition: What causes the delay? *University of Pennsylvania Working Papers in Linguistics*, 16(1): Article 13. Retrieved from

- <https://repository.upenn.edu/pwpl/vol16/iss1/13>
- Konopka, A. E. (2018). Encoding actions and verbs: Tracking the time-course of relational encoding during message and sentence formulation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. <https://doi.org/10.1037/xlm0000650>
- Kuchinsky, S. E., & Bock, K. (2010). *From seeing to saying: Perceiving, planning, producing*. Presented at the 23rd meeting of the CUNY Human Sentence Processing Conference, New York, NY.
- Lee, E.-K., Brown-Schmidt, S., & Watson, D. G. (2013). Ways of looking ahead: Hierarchical planning in language production. *Cognition*, 129(3), 544–562. <https://doi.org/10.1016/j.cognition.2013.08.007>
- Levy, R. (2014). Using R formulae to test for main effects in the presence of higher-order interactions. *ArXiv:1405.2094 [Stat]*. Retrieved from <http://arxiv.org/abs/1405.2094>
- Mack, J. E., Meltzer-Asscher, A., Barbieri, E., & Thompson, C. K. (2013). Neural Correlates of Processing Passive Sentences. *Brain Sciences*, 13, 1198–1214. <https://doi.org/10.3390/brainsci3031198>
- McDonald, J. L., Bock, K., & Kelly, M. H. (1993). Word and world order: Semantic, phonological, and metrical determinants of serial position. *Cognitive Psychology*, 25(2), 188–230. <https://doi.org/10.1006/cogp.1993.1005>
- Mehler, J. (1963). Some effects of grammatical transformation on the recall of English sentences. *Journal of Verbal Learning & Verbal Behavior*, 2(4), 346–351. [https://doi.org/10.1016/S0022-5371\(63\)80103-6](https://doi.org/10.1016/S0022-5371(63)80103-6)
- Miller, G. A. (1962). Some psychological studies of grammar. *American Psychologist*, 17(11), 748–762. <https://doi.org/10.1037/h0044708>
- Myachykov, A., Garrod, S., & Scheepers, C. (2011). Perceptual priming of structural choice during English and Finnish sentence production. In R. K. Mishra & N. Srinivasan (Eds.), *Language-cognition interface: state of the art* (pp. 53–71). München: Lincom Europa.
- Myachykov, A., Garrod, S., & Scheepers, C. (2018). Attention and Memory Play Different Roles in Syntactic Choice During Sentence Production. *Discourse Processes*, 55(2), 218–229. <https://doi.org/10.1080/0163853X.2017.1330044>
- Myachykov, A., Thompson, D., Garrod, S., & Scheepers, C. (2012). Referential and visual cues to structural choice in visually situated sentence production. *Frontiers in Psychology*, 2:396. <https://doi.org/10.3389/fpsyg.2011.00396>
- Myachykov, A., & Tomlin, R. (2008). Perceptual priming and structural choice in Russian sentence production. *Journal of Cognitive Science*, 9(1), 31–48. <https://doi.org/10.17791/jcs.2008.9.1.31>

- Norcliffe, E., & Konopka, A. E. (2015). Vision and language in cross-linguistic research on sentence production. In R. K. Mishra, N. Srinivasan, & F. Huettig (Eds.), *Attention and Vision in Language Processing* (pp. 77-96). Berlin: Springer. <https://doi.org/10.1007/978-81-322-2443-3>
- Osborne, J. W. (2010). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research & Evaluation*, 15(12), 1-9. Retrieved from <https://pareonline.net/getvn.asp?v=15&n=12>
- Paolazzi, C. L., Grillo, N., Alexiadou, A., & Santi, A. (2019). Passives are not hard to interpret but hard to remember: evidence from online and offline studies. *Language, Cognition and Neuroscience*, 1–25. <https://doi.org/10.1080/23273798.2019.1602733>
- Penke, M. (2015). Syntax and language disorders. In M. Penke (ed.) *Handbooks of Linguistics and Communication Science. Syntax: Theory and analysis: Vol. 42/3* (pp. 1833–1874). Retrieved from <http://hdl.handle.net/1854/LU-885253>
- Prat-Sala, M. (1997). *The production of different word orders: A psycholinguistic and developmental approach* (PhD thesis). University of Edinburgh.
- Prat-Sala, M., & Branigan, H. P. (2000). Discourse constraints on syntactic processing in language production: A cross-linguistic study in English and Spanish. *Journal of Memory and Language*, 42(2), 168–182. <https://doi.org/10.1006/jmla.1999.2668>
- Ring, M., & Clahsen, H. (2005). Distinct patterns of language impairment in Down's syndrome and Williams syndrome: The case of syntactic chains. *Journal of Neurolinguistics*, 18(6), 479–501. <https://doi.org/10.1016/j.jneuroling.2005.06.002>
- Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of Basic English Grammatical Structures: A Corpus Analysis. *Journal of Memory and Language*, 57(3), 348–379. <https://doi.org/10.1016/j.jml.2007.03.002>
- RStudio (Version 1.0.143). (2017). Boston. Retrieved from <http://www.rstudio.org/>
- Sauppe, S., Norcliffe, E., Konopka, A. E., Van Valin Jr., R. D., & Levinson, S. C. (2013). Dependencies first: Eye tracking evidence from sentence production in Tagalog. *Proceedings of the 35th annual meeting of the Cognitive Science Society*, 1265–1270. Retrieved from <http://mindmodeling.org/cogsci2013/papers/0242/index.html>
- Savin, H. B., & Perchonock, E. (1965). Grammatical structure and the immediate recall of English sentences. *Journal of Verbal Learning and Verbal behavior* 4, 348–353.

- Schiller, N. O., & Caramazza, A. (2003). Grammatical feature selection in noun phrase production: Evidence from German and Dutch. *Journal of Memory and Language*, 48(1), 169–194. [https://doi.org/10.1016/S0749-596X\(02\)00508-9](https://doi.org/10.1016/S0749-596X(02)00508-9)
- Schriefers, H., Hantsch, A., & Jescheniak, J. D. (2002). Determiner selection in noun phrase production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(5), 941–950.
- Segaert, K., Menenti, L., Weber, K., & Hagoort, P. (2011). A Paradox of Syntactic Priming: Why Response Tendencies Show Priming for Passives, and Response Latencies Show Priming for Actives. *PLoS ONE*, 6(10): e24209. <https://doi.org/10.1371/journal.pone.0024209>
- Tomlin, R. S. (1995). Focal attention, voice, and word order: an experimental, cross-linguistic study. In P. A. Downing & M. Noonan (Eds.), *Typological Studies in Language* (Vol. 30, p. 517). <https://doi.org/10.1075/tsl.30.18tom>
- van de Velde, M., Meyer, A. S., & Konopka, A. E. (2014). Message formulation and structural assembly: Describing “easy” and “hard ” events with preferred and dispreferred syntactic structures. *Journal of Memory and Language*, 71(1), 124–144. <https://doi.org/10.1016/j.jml.2013.11.001>
- van der Lely, H. K. (1996). Specifically language impaired and normally developing children: Verbal passive vs adjectival passive sentence interpretation. *Lingua*, 98, 243–272.
- van Nice, K. Y., & Dietrich, R. (2003). Task sensitivity of animacy effects: evidence from German picture descriptions. *Linguistics*, 41(5), 825-849. <https://doi.org/10.1515/ling.2003.027>

COMPREHENSION OF JAPANESE PASSIVES: AN EYE-TRACKING STUDY WITH 2-3-YEAR-OLDS, 6-YEAR-OLDS, AND ADULTS

MIWA ISOBE,* REIKO OKABE,
YUKINO KOBAYASHI, SHIGETO KAWAHARA,
TOMOKO MONOU, KAZUHIRO ABE,
REI MASUDA, SAEKA MIYAHARA
& YASUYO MINAGAWA

Abstract

It has been reported that Japanese-speaking 4-6-year-olds can correctly interpret passives without the *ni*-phrase (short passives), and that even 2-3-year-olds can produce short passives, albeit with a limited number of verbs. The present study thus investigates whether 2-3-year-olds can distinguish short passives from their active counterparts by employing the preferential looking paradigm using an eye-tracker. Our experiment revealed that the eye-gaze behaviors of 2-3-year-old children were different from those of 6-year-olds and adults. The fixation time data suggest that 2-3-year-olds were able to comprehend active sentences but had difficulty in dealing with passives in the same way as older children. The results of the experiment also raise the possibility that younger children can notice the passive morpheme, which is not included in an active sentence, but cannot promptly revise the initial interpretation guided by the typical interpretive strategy of taking the first NP as the agent of the verb.

* Corresponding author: Miwa Isobe, Faculty of Education, University of Yamanashi, 4-4-37 Takeda, Kofu, Yamanashi, 400-8510, Japan, email: isobe.miwa@gmail.com

1. Introduction

Previous acquisition studies within the framework of generative grammar have shown that, while children with different language backgrounds acquire many properties of their native grammars by around age 4, they often have difficulty with comprehending passive sentences cross-linguistically even between the ages of 5 and 6. Several experimental studies on passives in child Japanese have demonstrated that children as old as 5 or 6 exhibit non-adult-like interpretations of passives with a *ni*-phrase, equivalent to a *by*-phrase in English (e.g., Sugisaki 1999; Minai 2000; Sano et al. 2001). These studies report that children interpret passive sentences as if they were active sentences. On the other hand, it has also been reported that Japanese 4-6-year-olds can correctly interpret passives without the *ni*-phrase (e.g., Okabe & Sano 2002; Ishikawa et al. 2018), and that even children between the ages of 2 and 3 can produce short passives, albeit with a limited number of verbs (Harada & Furuta 1999). These previous studies raise the possibility that children as young as 2 or 3 have already acquired basic knowledge of Japanese passives but that experimental methods prevent researchers from extracting such young children's knowledge of passives. To our knowledge, no experimental studies have been attempted to explore the comprehension of Japanese passives by children around the age of 3 since few reliable methods were available to test young children's comprehension of this construction.

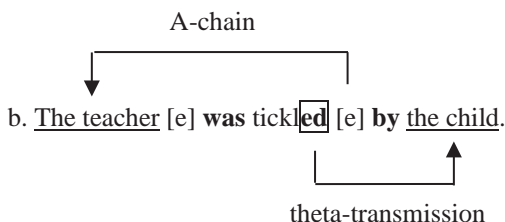
In light of this situation, the present study investigates whether Japanese-speaking 2-3-year-olds can distinguish short passives from their active counterparts by employing the preferential looking paradigm using an eye-tracker, which allows children's sentence comprehension to be examined in a non-intrusive manner. We compare young children's eye-gaze data with those of 6-year-olds and adults. Although the previous acquisition studies on Japanese passives have reported that 6-year-olds are perfectly adult-like, no studies have ever reported gaze data of 2-3-year-olds or of 6-year-olds and adults. Our experiment revealed that the eye-gaze behaviors of young children were different from those of 6-year-olds and adults. The fixation time data suggest that children aged 2 to 3 were able to comprehend active sentences but had difficulty in dealing with passives in the same way as older children. The results of the experiment also raise the possibility that younger children can notice the passive morpheme, which is not included in an active sentence, but cannot promptly revise the initial interpretation guided by the typical interpretive

strategy of taking the first NP as the agent of the verb.

2. Passives in English and Japanese

This section briefly surveys the syntactic analyses of passive constructions in both English and Japanese, on which we base our experimental study concerning the acquisition of passives. An example of the English passive is given in (1a), and its underlying structure is shown in (1b) along with illustrations of an A-chain and theta-transmission, following Jaeggli (1986).

(1) a. The teacher was tickled by the child.



In the modern generative grammar, the English passive sentence is considered to include promotion of the object to subject position through A-movement. Sentence (1a) indicates that the object NP of the verb *tickle* moves to the subject position on the surface structure, and the external theta-role of the verb which is absorbed by the passive morpheme *-ed* is optionally discharged to the NP in the *by*-phrase. Passive sentences with a *by*-phrase are widely called *long passives* and those without it are called *short passives*, and we will adopt these terminologies in our study.

We will now see that a similar analysis can be applied to Japanese passives as well. It is well-known that passives in Japanese are largely divided into two types: direct passives and indirect passives (e.g., Shibatani 1978). Although both types are similar in that they involve the passive morpheme *-(r)are*, it is argued that the former corresponds to the English passive, while the latter, which is formed by adding an extra argument to the active sentence, does not have an English counterpart. Since we will discuss the acquisition of Japanese passives in parallel with English passives and our experiment deals with only direct passives, we will focus only on that type of passive henceforth.

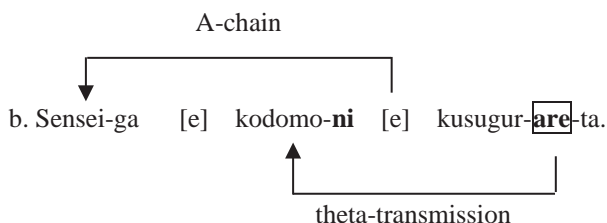
An example of a Japanese active sentence is given in (2) with its passive counterpart shown in (3).

(2) *Active*

Kodomo-ga	sensei-o	kusugut-ta.
child-Nom	teacher-Acc	tickle-Past ¹
‘The child tickled the teacher.’		

(3) a. *Passive*

Sensei-ga	kodomo-ni	kusugur-are-ta.
teacher-Nom	child-by	tickle-Pass-Past
‘The teacher was tickled by the child.’		



The Japanese passive sentence in (3a) is posited to have the underlying structure in (3b) (e.g., Kubo 1992). Like English passives, Japanese passives are assumed to involve A-chain formation and theta-transmission. The passive suffix *-(r)are* is attached to the verb stem *kusugur-* ‘tickle,’ and the object of the corresponding active sentence, *sensei* ‘teacher,’ appears as the subject with nominative case *-ga*. The suppressed theta-role of the subject in the active sentence, *kodomo* ‘child,’ can optionally appear with the dative marker *-ni* ‘by,’ just as in the English *by*-phrase.

3. Child Passives

3.1. Acquisition of passives in English and other languages

Previous acquisition studies have shown that children exhibit non-adult-like interpretations of long passive sentences cross-linguistically

¹ The abbreviations used in the glosses throughout this study are: NOM = nominative case, ACC = accusative case, Past = past tense, and Pass = passive suffix.

until around the age of 5 or 6. In particular, several sources have been proposed for the delay in understanding English passives. Borer & Wexler (1987) propose that children's ability to form an A-chain undergoes maturation and until the ability matures children treat passives as adjectival ones. Fox & Grodzinsky (1998), on the other hand, propose that children have difficulty in discharging the external theta role to the *by*-phrase, not in forming an A-chain. Their claim is based on the results of their experiment, which revealed that children performed well on passives with actional verbs such as *chase* and *touch*, and they made mistakes only on long passives with non-actional verbs such as 'The boy is seen by the horse.' Although both Borer and Wexler (1987) and Fox and Grodzinsky (1998) attribute children's difficulty with passive comprehension to their non-adult-like grammatical knowledge, some recent studies argue against the grammatical accounts for this difficulty. O'Brien et al. (2006), for example, showed experimentally that comprehension of English passives by children as young as ages 3-4 was adult-like when the construction was used in a pragmatically felicitous context, suggesting that the children had already acquired adult-like knowledge of passives. Recently, an incremental processing account for children's difficulty in comprehending English passives has been proposed by Deen et al. (2018), who claim that children's processing mechanisms are still in the process of developing while they may already have developed adult-like knowledge of passives.

Moreover, many acquisition studies on passives in a wide variety of languages other than English have reported that passives are difficult for children cross-linguistically (e.g., Spanish (Pierce 1992), Greek (Terzi & Wexler 2002), and Russian (Babyonyshev & Brun 2004), although there is evidence that children acquiring Sesotho have adult-like knowledge of passives by around the age of 3 (Demuth et al. 2010). Recently, Armon-Lotem et al. (2016) conducted experiments with 5-year-olds across different European languages and found that they performed well on short passives but showed difficulty with long passives in languages such as Catalan and Hebrew.

3.2. Acquisition of passives in Japanese

It has also been reported that Japanese-speaking children have difficulty with passives and tend to interpret passive sentences like (3a) 'The teacher

was tickled by the child' as if they were active, i.e., 'The teacher tickled the child' (e.g., Sugisaki 1999; Minai 2000; Sano et al. 2001; Okabe & Sano 2002; Sano 2013). Some experimental studies, such as Sugisaki (1999) and Minai (2000), claim that children's difficulty in understanding Japanese passives is caused by their immature grammatical ability to form an A-chain following Borer & Wexler (1987). On the other hand, studies such as Sano et al. (2001) and Okabe & Sano (2002) have shown experimentally that Japanese-speaking children around the age of 4 can comprehend short passives but have trouble interpreting long passives. This suggests that children already have the ability to form an A-chain: Otherwise they could never interpret short passives which involve the movement of an object NP to the subject position, i.e. an A-chain. These studies claim that children already have the ability to comprehend passives but have difficulty with interpreting the agent role in *ni*-phrases, i.e., 'by-phrases.' Their experimental results are compatible with Fox & Grodzinsky's (1998) claim. Recently, Ishikawa et al. (2018) reported a similar observation: If agent *ni*-phrases were omitted, the comprehension of passives by 4-year-olds and older children improved compared to comprehension of long passives with an agent role. In addition, analyses of spontaneous speech corpora have reported that utterances by children at age 2 contain short passive sentences, although they are limited in the variety of verbs used in their speech (Harada & Furuta 1999).

Considering the findings from these previous studies, it might be possible that children younger than 4 years of age can comprehend short passives. Yet, comprehension of Japanese passives by children as young as ages 2-3 has rarely been experimentally tested, partly because the tasks adopted for experiments on child passives are generally not suitable for children younger than 4 years old. The previous experimental studies on the acquisition of Japanese passives adopted methods such as the act-out task (e.g., Sano 1977; Hakuta 1982; Murasugi & Kawamura 2005), the picture-selection task (e.g., Sugisaki 1999; Minai 2000), or the truth-value judgment task (e.g., Sano et al. 2001; Okabe & Sano 2002). These tasks place an extra cognitive burden on young children, which may hinder a precise understanding of their comprehension of passives (see Goodluck 1996: 152 for an example of the discussion on the cognitive complexity of the act-out task). The fact that they can produce some passive sentences in their spontaneous speech raises the possibility that they may be able to understand passive sentences but that technological limitations may have prevented us from revealing their grammatical knowledge. To address this

gap, we conducted an eye-tracking study on the acquisition of Japanese passives, which we believe sheds new light on the behavior of younger children.

Given the findings of previous acquisition studies, the research question given in (4) arises concerning young Japanese-speaking children's comprehension of passives.

(4) *Research question:*

Do Japanese-speaking children as young as ages 2-3 show eye-gaze behavior similar to older children and adults when presented with passive sentences along with visual stimuli depicting matching and mismatching scenes?

4. Experiment

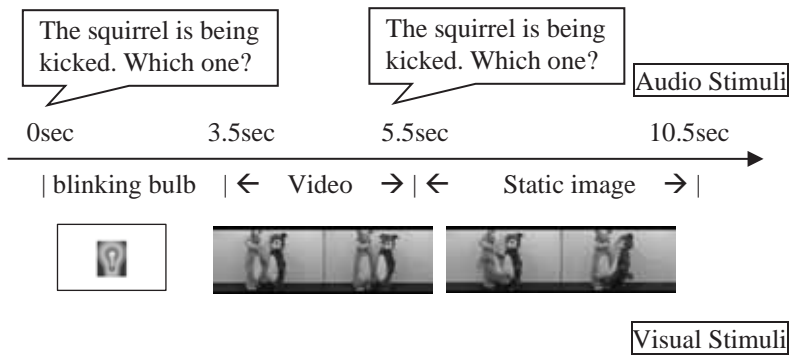
4.1. Participants

We tested 60 children ranging in age from 2;6 to 3;5 (mean = 2;10). We also tested 25 children of age 6 (mean=6;7) and 10 adults for comparison. We selected 6-year-olds in reference to Ishikawa et al. (2018): In their experiment, the participants of age 6 correctly comprehended Japanese short passives 94.0% of the time. All of the participants were from Tokyo or surrounding areas, and all the participants of ages 2-3 and age 6 were female because this experiment was conducted as part of a larger project on cognitive development which required female participants.

4.2. Procedure

The task we adopted was a preferential looking task in which two short videos followed by static images of their respective final scenes were presented side-by-side to each participant. We recorded participants' gaze data using a Tobii X120 eye-tracker. Each participant was asked to sit in front of a monitor and watch the screen very carefully. The timeline of a trial with sample auditory and visual stimuli is shown in Figure 1.

Figure 1: Timeline of a trial



Each trial lasts 10.5 seconds and begins with an alerting sound along with a blinking light bulb intended to attract the child's gaze to the center of the monitor. After the alerting sound, a test sentence, such as "the squirrel is being kicked," followed by "which one?" is played for 2.5 seconds with an animation of a light bulb. After the test sentence is auditorily presented, two animations with two animals begin playing side-by-side on the split-monitor. One of the animations matches the situation depicted by the audio stimulus, while the other mismatching animation shows a scene where the agent and the patient are reversed. The animations are played for 2 seconds, after which the test sentence is played for the second time. The animation becomes a static image showing the outcome of the action, which enables children to remember what has happened to the two animals in the scene. The static images are shown on the monitor for a total of 5 seconds.

We decided to present the stimulus sentences auditorily before any visual stimuli appeared on the monitor for the following reason: It was often the case in previous experimental studies that child participants looked at pictures or stories before the test sentences were presented and so were likely to choose a picture or judge the truth-value of a sentence even before they finished listening to the whole sentence. Note also that the passive morpheme *-(r)are* appears at the very end of the sentence following the verb stem, which makes the passive morpheme less conspicuous and easily ignored. Therefore, in order to ensure that child participants would choose a picture after encountering the passive

morpheme, we decided on the ordering of auditory and visual stimuli shown in Figure 1.

Additionally, each test sentence was played twice in order to help the child review her initial interpretation, which was made just after the first presentation of the sentence. This is in line with the results of an experiment on child passives by Deen et al. (2018): Their experiment revealed that English-speaking children's comprehension of passives improved when test sentences were given twice, suggesting that repeating test sentences might enable children to review an initial interpretation while listening to the test sentence for the second time.

Before the test session, each of the children between ages 2 and 3 was shown pictures of four animals which appeared in the animations and asked to name the animals to be used in the test sentences. If a child was unable to give the correct names of the four animals, an experimenter told her the names and asked her to repeat them, and then asked her to say the names of the animals several times while showing her each picture, like a flash card game. The child then went through a pretest in which she familiarized herself with the procedure and all the transitive verbs used in the experiment, i.e. *kick* and *push* for test sentences and *hug* and *pat* for fillers. The pretest sentences consist of an overt verb in a progressive form with a null object, such as “the squirrel is kicking (*pro*),” followed by “which one?” Each trial in the pretest session lasts for 7.5 seconds. After the alerting sound, the pretest sentence is played for 1.5 seconds with an animation of a light bulb. Then two animations with two animals begin playing side-by-side on the split-monitor. One of the animations matches the situation depicted by the audio stimulus, while the other mismatching animation shows a scene where the agent's action toward the patient is completely different from the target action. The animations are played for 2 seconds before becoming a static image showing the outcome of the action. The static images are shown on the monitor for 3 seconds.

4.3. Test Sentences

We tested a total of 16 sentences: 8 target sentences, including both active and passive sentences, along with 8 filler sentences.² Among the 8 test

² Since the attention span of 2-3-year-old children is generally short, we judged that 16 stimuli in total was the maximum for the young children.

items, 4 contained the verb *keru* ‘kick’ and the other 4 contained the verb *osu* ‘push.’ Examples of both active and passive test sentences are given in (5) and (6) respectively. Note that active test sentences were given without overt direct object NPs, which is entirely acceptable in Japanese, a null argument language. Passive sentences were all short passives, i.e., passives without *-ni* ‘by’ phrases. This allowed both active and passive test sentences to consist of the same number of words, making them minimal pairs that differed only in the presence/absence of the passive morpheme.

(5) *Test sentence with verb keru* ‘kick’:

- | | | | | | |
|----|---|---------------|------|--------|-----------|
| a. | Risusan-ga | ket-teru | yo. | Dotti? | (Active) |
| | squirrel-Nom | kick-ing | excl | which | |
| | ‘The squirrel is kicking <i>pro</i> . Which one?’ | | | | |
| b. | Risusan-ga | ke-rare-teru | yo. | Dotti? | (Passive) |
| | squirrel-Nom | kick-Pass-ing | excl | which | |
| | ‘The squirrel is being kicked. Which one?’ | | | | |

(6) *Test sentence with verb osu* ‘push’:

- | | | | | | |
|----|--|---------------|------|--------|-----------|
| a. | Inusan-ga | osi-teru | yo. | Dotti? | (Active) |
| | dog-Nom | push-ing | excl | which | |
| | ‘The dog is pushing <i>pro</i> . Which one?’ | | | | |
| b. | Inusan-ga | os-are-teru | yo. | Dotti? | (Passive) |
| | dog-Nom | push-Pass-ing | excl | which | |
| | ‘The dog is being pushed. Which one?’ | | | | |

By adding a question phrase *dotti?* ‘which one?’ at the end of each stimulus, we ensured that the child knew that she was expected to look at one or the other of the panels during the sessions. The order of the 16 trials was randomized per participant. The congruent events appeared on either side of the display in pseudo-random orders.

In addition, parents of the 2-3-year-olds were asked to fill out the Japanese MacArthur Communicative Development Inventory (Watamaki & Ogura 2004), which contains questions as to whether a child has already acquired certain words: The two target verbs are included in that list.

4.4. Results and analysis

We focused on the total fixation time on the two static images showing the

outcome of the event, which appear during the last 5 seconds of each trial, starting with the second auditory stimulus and lasting until the end of the trial. By looking at the total fixation time data, we are able to infer how participants reach their interpretations after hearing the stimulus sentences twice. Previous off-line studies using the picture selection task, for example, can only provide us with information as to which picture each child has eventually chosen.

We excluded from further analysis the data of those who had not acquired the relevant verbs and children with few gaze samples or whose total fixation time on the two panels for at least one trial was zero, which left 32 2-3-year-olds (mean = 2;11) and 19 6-year-olds (mean = 6;4). We also had to exclude the data of one adult since she told the experimenter just after the experiment that she had consistently looked at the incongruent panels throughout the trials on purpose, questioning whether the task had some hidden tricks.

To analyze the data statistically, we first fit a linear mixed effects model with fixation time as the dependent variable and congruency, the active/passive distinction and age group as three fixed predictor variables. All the interaction terms between the three fixed factors, as well as a random intercept for participants, were included in the model. As we can see in Figures 2, 3, and 5, participants from different age groups appear to exhibit different patterns. This observation manifests itself as a significant three-way interaction term ($\beta = -0.74$, *s.e.* = 0.30, $t = -2.49$, $p < .05$). Since this complex model is difficult to interpret, we fit a simpler model to the data in each age group.³

Figure 2 shows the total fixation time by adults in the test session. In this graph, the vertical axis indicates total fixation times in the last 5 seconds. The gray boxplots indicate “Congruent,” which represents the duration of time when the participants were looking at the correct panels, while the white “Incongruent” boxplots show how long the participants were looking at the incorrect ones. We can see that the adults looked at the congruent scenes longer than the incongruent ones for both active and passive sentences. A linear mixed effects model with congruency and active/passive as the fixed factors, together with a random intercept for speakers, shows that congruency had a significant impact on total fixation

³ An R markdown file with complete model details for all the analyses reported below is available upon request.

time ($\beta = 3.59$, $s.e. = 0.14$, $t = 26.3$, $p < .001$), whereas the active/passive distinction and its two-way interaction with congruency did not ($\beta = 0.07$, $s.e. = 0.14$, $t = 0.54$, $n.s.$; $\beta = 0.27$, $s.e. = 0.27$, $t = 1.00$, $n.s.$).

Figure 2: Total fixation time by adults⁴

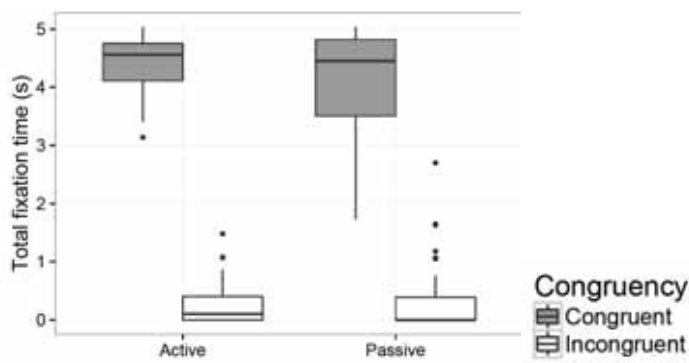
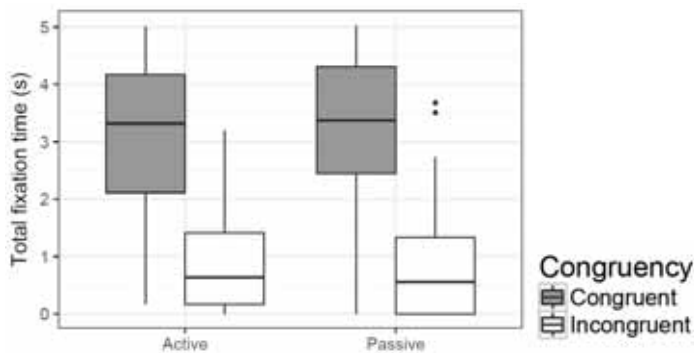


Figure 3: Total fixation time by 6-year-olds



⁴ The thick black lines represent the medians. The upper edge of the box represents the 75th percentile, the lower edge the 25th percentile. The length of the whiskers is defined as 1.5 times interquartile range; dots were data points outside that range.

Figure 4: Total fixation time by 2-3-year-olds in pretest

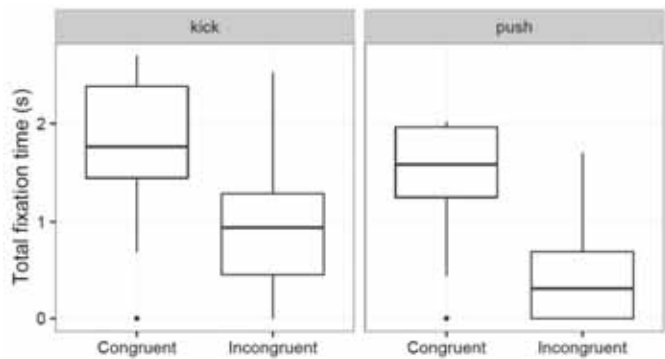
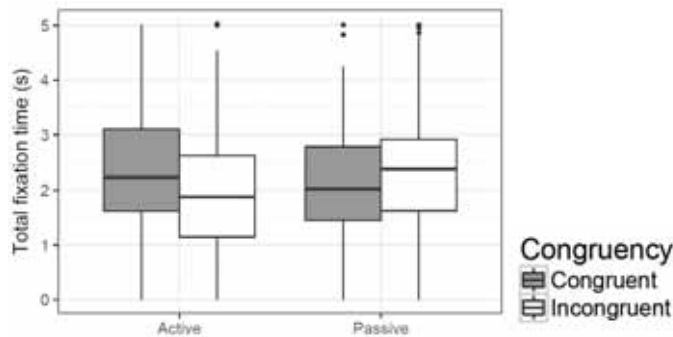


Figure 5: Total fixation time by 2-3-year-olds



The results for the 6-year-olds are given in Figure 3, where we see a pattern similar to the results for the adults, although the trend is less clear-cut. The 6-year-olds looked longer at the matching scenes for both active and passive sentences. For the 6-year-olds’ data, a linear mixed effects model with congruency and active/passive as fixed factors revealed that the only significant factor was the congruent/incongruent distinction ($\beta = 2.3$, $s.e. = 0.12$, $t = 19.48$, $p < .001$). Neither the active/passive distinction nor the interaction term was significant ($\beta = 0.01$, $s.e. = 0.12$, $t = 0.07$, $n.s.$; $\beta = -0.12$, $s.e. = 0.24$, $t = -0.49$, $n.s.$).

As for the children aged 2-3, let us consider first the results of the pretest section. Figure 4 shows that the children looked at the congruent image longer than the incongruent one to a statistically significant degree ($\beta = -0.81$, $s.e. = 0.15$, $t = -5.57$, $p < .001$). The verb *kick* has a slight advantage over the verb *push* ($\beta = -0.30$, $s.e. = 0.15$, $t = -2.04$, $p < .05$), but the interaction was non-significant ($\beta = -0.24$, $s.e. = 0.21$, $t = -1.16$, $n.s.$). This means that they were able to connect the verbs *kick* and *push* with their matching scenes. In contrast, the total fixation times in the test sessions, as seen in Figure 5, indicate that passive sentences have slightly longer incongruent times, whereas active sentences have longer congruent times. A linear mixed model with congruency and active/passive as fixed factors with a random intercept of speakers revealed a significant interaction between congruency and the active/passive distinction ($\beta = 0.62$, $s.e. = 0.19$, $t = 3.33$, $p < .001$). Neither congruency nor the active/passive distinction had a significant impact on total fixation time ($\beta = 0.06$, $s.e. = 0.09$, $t = 0.69$, $n.s.$; $\beta = -0.08$, $s.e. = 0.09$, $t = -0.87$, $n.s.$). Separate models exploring the effects of congruency in each of the active and passive conditions show that the 2-3-year-olds looked at congruent scenes longer than incongruent scenes for the active sentences ($\beta = 0.38$, $s.e. = 0.13$, $t = 2.79$, $p < .01$), while no significant differences were identified in the passive condition ($\beta = -0.24$, $s.e. = 0.13$, $t = -1.91$, $n.s.$). These results present a stark contrast from those of the 6-year-olds and the adults.

5. Discussion

The results of the current experiment revealed that the 2-3-year-old children looked longer at the congruent events for active sentences, whereas they tended to look at the incongruent ones longer for passive sentences. This asymmetry suggests that they comprehended active sentences but had difficulty with passive sentences. We also found that the 6-year-old children showed different behaviors from the 2-3-year-olds: They consistently looked longer at congruent scenes regardless of sentence types, performing almost as well as the adults did.

There are several possible causes for the non-adult-like behavior of the 2-3-year-old children. First, it is possible that the young children did not understand that they were supposed to look at the scene which they thought matched the stimulus sentences, even though we added the

guiding phrase ‘which one?’ at the end of each stimulus. We can deny this possibility when we look at their performance during the pretest. We have already confirmed in the previous section (Figure 4) that they correctly looked at the congruent images much longer than the incongruent images in the pretest. This result suggests that their poor performance is not attributable to an inability to understand the task itself.

Another possible cause of the non-adult-like behavior of the 2-3-year-old children for passive sentences may be Lexical-ordering Strategy or the agent-first strategy (Bever 1970; Suzuki 1977). It has been suggested that this strategy is typically employed by young children who have trouble interpreting sentences with non-canonical word order, such as passive sentences or sentences with scrambling. Those children tend to regard the first NP as the actor. It is possible that the 2-3-year-old children in our experiment also used this strategy, which led them to interpret the first NPs of the passive sentences as actors rather than as actees. However, this explanation is not fully tenable when we compare their performances on passives and actives. If the children regarded the first NPs of passive sentences as actors, treating them just as they would in active sentences, the results for passives and actives would have been perfect mirror images. On the contrary, as was pointed out in the previous section (Figure 5), the distinction between congruent and incongruent total fixation times for passives is not statistically significant, but it is for actives. Thus, the agent-first strategy alone does not seem to tell the full story.

It is reasonable then to conjecture that the young children recognized the passive morpheme and thought that the passive sentences they had heard were somehow different from the active sentences, although they could not react promptly and confidently to the passive sentences within a few seconds. This account is along the same lines as the idea proposed by Omaki & Lidz (2015) based on many previous studies on child processing. Huang et al. (2013), for example, examined how Mandarin-speaking 5-year-olds interpret passive sentences with a referential subject, such as *Seal BEI it eat* ‘The seal is eaten by it’ (where *BEI* is a passive morpheme and is followed by an Agent NP and a VP (Huang 1999)), and with a pronominal subject, such as *It BEI seal eat* ‘It is eaten by the seal.’ The children interpreted both types of passive at around chance levels, but they comprehended sentences with a pronominal subject better than those with a referential subject. Huang et al. (2013) argue that the agent-first interpretation does not allow pronominal subjects of passives, and

children's difficulties in comprehending passives were thus mitigated. Conversely, they had difficulty in revising the agent-first interpretation bias when the subjects of passives were expressed nouns. Omaki & Lidz (2015) argue that children's difficulties in comprehending certain constructions may be partly due to their immature sentence revision mechanisms. We can apply this assumption to the results of our experiment: The sentence revision mechanisms of the youngest participants in our experiment are not mature enough to succeed in revising the initial misinterpretations induced by the Lexical-ordering Strategy even after they have realized that the sentence they heard had a passive morpheme. If they solely relied on the Lexical-ordering Strategy, completely ignoring the passive morpheme and treating the sentences as if they were actives, they would have looked statistically longer at the incongruent scenes. However, the data showed otherwise. We thus conjecture that the child participants, whose sentence revision mechanisms are still not fully matured, could not decide confidently on which scene to watch and whether to cancel their first interpretation during the limited time window in the experimental condition. This is the most reasonable explanation for the observed behaviors of the 2-3-year-old children with passive sentences based on the results we have at hand.

6. Conclusion

The present study on comprehension of Japanese passives revealed that the participants of ages 2-3 looked longer at the congruent events when given active sentences, while they looked at the incongruent ones only slightly but not significantly longer when they heard passive sentences. We argued that their non-adult-like behavior for passive sentences cannot be accounted for only in terms of the difficulty of the experimental task or by the Lexical-ordering Strategy. We claimed that although young children could not revise their initial interpretations guided by the agent-first strategy as quickly as 6-year-olds or adults, they were aware of the existence of passive morphemes and that those sentences they heard were different from the active ones. Our current account is compatible with the idea that young children have difficulty with sentence revision, as has been observed with various constructions.

Acknowledgements

This study was supported by JSPS KAKENHI Grant Number JP17K02711.

References

- Armon-Lotem, S., Haman, E., Jensen de López, K., Smoczynska, M., Yatsushiro, K., Szczerbinski, M., van Hout, A., Dabašinskienė, I., Gavarró, A., Hobbs, E., Kamandulytė-Merfeldienė, L., Katsos, N., Kunnari, S., Nitsiou, C., Sundahl Olsen, L., Parramon, X., Sauerland, U., Torn-Leesik, R. and van der Lely, H. (2016) A large-scale cross-linguistic investigation of the acquisition of passive, *Language Acquisition*, 23:1, 27-56.
- Babyonyshev, M. and Brun, D. (2004) The acquisition of perfective and imperfective passive constructions in Russian, *University of Pennsylvania Working Papers in Linguistics* 10, 17-31.
- Bever, T. G. (1970) The cognitive basis for linguistic structures, *Cognition and Language Development*, ed. by J. R. Hayes, 279-352, NY: Wiley.
- Borer, H. and Wexler, K. (1987) The maturation of syntax, *Parameter Setting*, ed. by T. Roeper and E. Williams, 123-172, Reidel: Dordrecht.
- Deen, K. U., Bondoc, I., Camp, A., Estioca, S., Hwang, H., Shin G.-H., Takahashi, M., Zenker, F., and Zhong, J. C. (2018) Repetition brings success: Revealing knowledge of the passive voice, *Proceedings of the 42nd Annual Boston University Conference on Language Development*, ed. by A. B. Bertolini and M. J. Kaplan, 200-213, Somerville, MA: Cascadia Press.
- Demuth, K., Moloi, F. and Machobane, M. (2010) 3-year-olds' comprehension, production, and generalization of Sesotho passives, *Cognition* 115, 238-251.
- Fox, D. & Grodzinsky, Y. (1998) Children's passive: A view from the by-phrase, *Linguistic Inquiry* 29, 311-332.
- Goodluck, H. (1996) The act out task, *Methods for Assessing Children's Syntax*, ed. by D. McDaniel, C. McKee, and H. S. Cairns, 147-162, Cambridge, MA: MIT Press.
- Hakuta, K. (1982) Interaction between particles and word order in the comprehension and production of simple sentences in Japanese children, *Developmental Psychology* 18, 62-76.
- Harada, K. & Furuta, T. (1999) On the maturation of A-chains: A view

- from Japanese passives, *COE Research Report* 3, Kanda University of International Studies, 397-426.
- Huang, C.-T. J. (1999) Chinese passives in comparative perspective, *Tsing Hua Journal of Chinese Studies* 29, 423-509.
- Huang, Y. T., Zheng, X., Meng, X. and Snedeker, J. (2013) Children's assignment of grammatical roles in the online processing of Mandarin passive sentences, *Journal of Memory and Language* 69, 589-606.
- Ishikawa, M., Goro, T. and Ito, T. (2018) Nihongo-zi-ni okeru tadosi-ukemibun-no rikai [Comprehension of transitive passives in Japanese children], Paper presented at the 156th Meeting of the Linguistic Society of Japan.
- Jaeggli, O. (1986) Passive, *Linguistic Inquiry* 17, 587-622.
- Kubo, M. (1992) Japanese passives. *Gengo Bunka Kiyoo* 23, 231-302, Hokkaido University.
- Minai, U. (2000) The acquisition of Japanese passives, *Japanese/Korean Linguistics* 9, 339-350.
- Murasugi, K. and Kawamura, T. (2005) On the acquisition of scrambling in Japanese, *The Free Word Order Phenomenon: Its Syntactic Sources and Diversity*, ed. by J. Sabel and M. Saito, 335-376, Berlin: Mouton de Gruyter.
- O'Brien, K., Grolla, E. and Lillo-Martin, D. (2006) Long passives are understood by young children, *Proceedings of the 30th Annual Boston Conference on Language Development*, ed. by D. Bamman, T. Magnitskaia, and C. Zaller, 441-451, Somerville, MA: Cascadilla Press.
- Okabe, R. and Sano, T. (2002) The acquisition of implicit arguments in Japanese and related matters, *Proceedings of the 26th Annual Boston Conference on Language Development*, ed. by B. Skarabela, S. Fish, and A. H.-J. Do, 485-499, Somerville, MA: Cascadilla Press.
- Omaki, A. and Lidz, J. (2015) Linking parser development to acquisition of syntactic knowledge, *Language Acquisition* 22, 158-192.
- Pierce, A. (1992) The acquisition of passives in Spanish and the question of A-chain maturation, *Language Acquisition* 2, 55-81.
- Sano, K. (1977) An experimental study on the acquisition of Japanese simple sentences and cleft sentences, *Descriptive and Applied Linguistics* 10, 213-233, International Christian University, Tokyo.
- Sano, T. (2013) Remarks on theoretical accounts of Japanese children's passive acquisition, *Generative Linguistics and Acquisition: Studies in Honor of Nina M. Hyams*, ed. by M. Becker, J. Grinstead, and J. Rothman, 35-64, Amsterdam: John Benjamins.

- Sano, T., Endo, M. and Yamakoshi, K. (2001) Developmental issues in the acquisition of Japanese unaccusatives and passives, *Proceedings of the 25th Annual Boston Conference on Language Development*, ed. by A. H.-J. Do, L. Domínguez, and A. Johansen, 668-683, Somerville, MA: Cascadilla Press.
- Shibatani, M. (1978) *Nihongo no Bunseki* [Analysis of Japanese], Tokyo: Taishukan.
- Sugisaki, K. (1999) Japanese passives in acquisition, *UConn Working Papers in Linguistics* 10, 145-156.
- Suzuki, S. (1977) Nihon no yozi ni okeru gozyun horyaku [Lexical-ordering strategy in early Japanese children], *The Japanese Journal of Educational Psychology* 25:3, 56-61.
- Terzi, A. and Wexler, K. (2002) A-chains and s-homophones in children's grammar: Evidence from Greek passives. *Proceedings of NELS* 32, 519-537.
- Watamaki, T. and Ogura, T. (2004) *Technical Manual of the Japanese MacArthur Communicative Development Inventory: Words and Grammar*, Kyoto International Social Welfare Exchange Center.

BILINGUAL JUDGMENTS AND PROCESSING OF SPANISH WH- GAP CONSTRUCTIONS: AN EXPLORATORY STUDY OF CROSS- LINGUISTIC INFLUENCE AND ISLAND STRENGTH

GITA MARTOHARDJONO,¹ CASS LOWRY,
MICHAEL A. JOHNS, IAN PHILLIPS,
CHRISTEN N MADSEN II
& RICHARD G. SCHWARTZ

Abstract

This study investigated whether two groups of fluent bilinguals—who vary in their lifetime exposure to their first-learned language, Spanish—display variable sensitivity to grammatical and ungrammatical *wh*- gap constructions in their first language. The participants were heritage speakers ($n = 32$)—whose home language was Spanish but who were raised in the anglophone US—and first-generation late bilinguals ($n = 24$)—who moved to the anglophone US from a Spanish speaking region in adulthood. *Wh*- gap constructions were selected to test for cross-linguistic influence (Comp-trace) and sensitivity to grammaticality gradience (strong and weak island violations). Both groups demonstrated crosslinguistic influence and gradient sensitivity in their behavioral judgments (acceptability judgment task) of the gap constructions. However, implicit, online measures (pupillometry) revealed group differences in the processing of three out of the five gap constructions. Nonetheless, both groups demonstrated increased sensitivity to strong island conditions as compared to weak

¹ Corresponding author: Gita Martohardjono, MA/PhD Program in Linguistics, CUNY Graduate Center, 365 Fifth Ave., New York, NY 11016; email: gmartohardjono@gc.cuny.edu

violations in their processing. Our results emphasize that while these two bilingual groups are similar in their ratings of *wh*- gap constructions, their processing of these structures show nuanced differences as measured by pupillometry. Potential sources of this difference in processing are discussed.

1. Introduction

An axiomatic assumption in the study of bilingualism is that the bilingual mind is not the sum total of two monolingual minds (de Houwer & Ortega, 2018; Grosjean & Li, 2013) as coexisting language systems can exert significant influence on each other. While differences between monolinguals and multilinguals have long been established empirically (e.g., Kroll & Bialystok, 2013), researchers have also begun to compare different groups of bilinguals (e.g., Montrul, 2016; Serratrice et al., 2009). This follows from the observation that the bilingual experience is largely determined by the relative exposure to the two languages, and that this exposure can vary greatly from one speaker to the next. Hence the literature recognizes different bilingual “types”, whose experience of the two languages varies systematically along age of first exposure, linguistic environment, and amount of use.

This study investigates two such types of Spanish-English bilinguals: heritage speakers, also commonly referred to as second-generation bilinguals; and late bilinguals, also commonly referred to as first-generation bilinguals. Heritage speakers have recently garnered much attention in the literature on bilingualism (e.g., Benmamoun et al., 2013; Montrul, 2016; Polinsky, 2018; Polinsky & Scontras, 2020; Rothman, 2007, 2009). They are typically children of immigrants in a particular situation of first language acquisition, involving majority vs. minority language settings. As such, they are raised in the home language, which is the societal minority language, until they reach school age, when they begin education in the majority language. Many, though not all, heritage speakers become dominant in the societal majority language. Nonetheless, heritage speakers often retain fluency in the home language, depending on their particular linguistic environment—for example if they live in a community where maintenance of the minority language is prevalent, leading to sustained use. This is often the case in Hispanic communities in the US (Otheguy & Zentella, 2011).

Heritage speakers are thought to be qualitatively distinct in their bilingualism from late bilinguals, who have a more uniform and continuous experience of their first language, are schooled in that language, and acquire the other

language only later in life. For example, it was originally argued that heritage speakers are distinct from child first language learners, and that the particular conditions under which they learn the home language often leads to interrupted, “incomplete acquisition” of that language (see for example Montrul, 2008). In recent years, this deficit-framing of heritage speakers’ acquisition of their home language has faded in the literature, being replaced with more neutral terms such as “differential acquisition” (Kupisch & Rothman, 2018) and “divergent attainment” (Polinsky & Scontras, 2020). In contrast, late bilinguals are reasonably assumed to have completely acquired their first-learned language, having been exposed to it continuously beyond childhood. Heritage speakers are also thought to be more susceptible than late bilinguals to first-language attrition and crosslinguistic transfer—two phenomena common to bilinguals in cases of language contact—since in the process of becoming dominant in the later-learned language, the mental representation and processing of the first-learned language can weaken (e.g., Schmid, 2011; Tsimpli et al., 2004).

We compare Spanish heritage speakers to a Spanish-English late bilingual group from the same community. The difference between these two groups rests primarily in the age of first exposure to and lifetime experience of English. Our aim is to see whether, and to what degree, these differences affect the representation and processing of Spanish, the first-learned language of both bilingual groups.

Two main questions are of interest to us: The first is degree of crosslinguistic influence, or how the later-learned English might affect the first-learned Spanish. Specifically, we explore the two groups’ judgment and processing of a *wh*-gap structure that differs in grammaticality between Spanish and English: Comp-trace interrogative sentences. The second question asks whether syntactic structures found in the first-learned language that are equivalent to those found in the second-learned language might undergo weakening. That is, will heritage speakers’ representation or processing of *wh*-islands differ from that of the late bilinguals, and to what degree. Both questions are explored from the perspective of grammatical representation, as measured with acceptability judgments, and that of syntactic on-line processing, which we measure with pupillometry, a highly sensitive method of cognitive load recently introduced in experimental studies of bilingualism (e.g., Fernández, 2003; Schmidtke, 2018).

This chapter is organized as follows: in section 2.1 we discuss the first issue of concern to this study, crosslinguistic influence, and describe the particular construction we use to test for it. Section 2.2 describes the

constructions we use to investigate potential first language weakening, namely “syntactic islands”. We then briefly motivate the use of pupillometry, in section 2.3, as our choice of method to investigate processing, followed by an outline of our research questions and predictions in section 2.4. The acceptability judgment task (AJT) is reported in section 3, followed by the pupillometry task in section 4. The chapter ends with a discussion of the results of both experiments according to our research questions and a conclusion.

2. Theoretical background

2.1 Crosslinguistic influence: Comp-trace

We begin with a brief description of crosslinguistic influence (CLI) which is hypothesized to affect both the mental representation and the processing strategies in bilingual speakers and is therefore an important issue to examine in a comparison of different bilingual types. Crosslinguistic influence is defined as the way languages in the mind of the bilingual influence each other, and has been studied in children (e.g., Serratrice, 2013) as well as adults (e.g., Dussias & Sagarra, 2007). The effect of crosslinguistic influence can vary according to level of immersion and amount of exposure to each language. In a wide-ranging set of studies on bilinguals of various language pairs, Sorace and her colleagues found that bilinguals immersed in the later-learned language (e.g., Italian-English bilinguals in the U.K.) differ in their representations of overt pronoun reference in Italian, the first-learned language, when compared to Italian-English bilinguals in Italy (Sorace et al., 2009), an effect they ascribe to different amounts of exposure and use. CLI has also been found to affect the processing of the first-learned language. In investigating high vs. low relative clause attachment in Spanish/English bilinguals, Dussias and Sagarra (2007) found a tendency to override the preferred native language strategy of low attachment, and adopt instead those of the second language, English, of high attachment. This has led them to the hypothesis that the parsing mechanisms of the first language are “permeable” and that both level of proficiency in, and amount of exposure to the later-learned language are thought to be important factors in determining whether parsing mechanisms of the first language will be influenced by those of the second.

One way of investigating CLI is to examine the treatment of constructions that contrast in grammaticality in bilinguals’ two languages, and one such construction for our language pair is the so-called Comp-trace construction. In English (1a), the overt complementizer *that* cannot be present before a

wh- gap. Only when there is no complementizer is the interrogative sentence grammatical (1b). In Spanish, the opposite is the case. The overt complementizer *que* is needed for a grammatical sentence (2a); omitting it (2b) results in ungrammaticality (Torrego, 1984).

- (1) a. *Who did John say **that** ____ knew the answer?
English Comp-trace minimal pair
 b. Who did John say **Ø** ____ knew the answer?
- (2) a. ¿Quién dijo Juan **que** ____ sabía la respuesta?
Equivalent structures in Spanish
 b. *¿Quién dijo Juan **Ø** ____ sabía la respuesta?
 who said Juan **that/Ø** knew the response

Investigating the treatment of Spanish Comp-trace constructions by heritage speakers and late bilinguals will tell us whether English has affected the mental representation and/or processing of these structures in the bilinguals' first-learned language, Spanish and whether it has done so in similar ways in the two groups.

2.2 Weakening of the first-learned language: Syntactic islands

Our second research question involves a set of structures that have been of considerable interest in both the syntactic and psycholinguistic literature since the 1960s, namely syntactic island constructions. At the center of a long-standing debate, these constructions have been claimed by some to be the result of processing mechanisms (e.g., Hofmeister et al., 2013), while others contend their origin to be grammatical (e.g., Phillips, 2013).

Our goal in this chapter is not to provide evidence for or against a grammatical or a processing account for the so-called island effect, but rather to examine their treatment in the Spanish varieties of the two bilingual groups we tested, and to see whether differential experience with the first-learned language affects either the grammatical representation or the processing of these constructions.

Island constructions have been documented in a variety of languages and have been argued to be governed by universal principles (e.g., *subjacency*, Chomsky, 1986). At the same time, linguists have noted variability both across languages and across the various sentence-types constituting islands (e.g., Perlmutter, 1968). Typically, these constructions involve a filler-gap

dependency, that is, a relationship between a noun phrase and its argument position. This is illustrated in the set of sentences in (3).

- (3) a. Susan read *War and Peace* in one sitting.
 b. What did Susan read ___ in one sitting? *War and Peace*.

In (3b), the word *what* is related thematically to the argument of the verb *read*, i.e., *War and Peace*. *Wh-* gap dependencies can be unbounded in the sense that they can hold across several clauses (and arguably across an unlimited number of clauses) as seen in (4).

- (4) a. What did Leigh think _{CP}[that Susan read ___ in one sitting]?
 b. What did John claim _{CP}[that Leigh thought _{CP}[that Susan read ___ in one sitting]]?

Nonetheless, there are certain clause-types where the filler-gap dependency becomes degraded rendering the sentence unacceptable. An example given in (5), where (5b) illustrates a filler-gap dependency across a relative clause boundary.

- (5) a. Jill likes the bookstore [that sells *War and Peace* in fifty languages].
 b. ???/*What did Jill like the bookstore [that sells ___ in fifty languages]?

Constructions like (5b) are degraded because relative clauses constitute “syntactic islands” that prevent the *wh-* word to be related to its argument position following the verb *sell*.

A number of syntactic islands have been identified across a variety of languages (for a summary, see Sprouse & Hornstein, 2013). It has also been noted that syntactic islands vary in their effect, resulting in a gradient scale of (un)acceptability. Some, like *wh-* islands, only marginally degrade the sentence, as seen in (6).

- (6) ?What did John wonder whether Susan read ___?

Others, as seen in the relative clause island in (5b), result in high unacceptability. These observations have led syntacticians to propose a grouping of “weak” and “strong” islands (for a summary, see Szabolcsi, 2006).

Our study examines both weak and strong island effects in the first-learned Spanish of our bilingual populations and whether there are significant differences between the two groups. In particular, if heritage speakers are subject to attrition and/or divergent attainment, it is reasonable to assume that patterns in their first-learned language such as island effects, might diminish or disappear altogether. Furthermore, the diminished effect might occur across the board or incrementally, affecting weak islands more than strong islands. And finally, this effect might occur in the grammar and/or in the processing of the first-learned language.

Note that while this question is related to the previous one it does not involve CLI, since island effects occur in both Spanish and English. Rather, the issue under investigation here is the potential *weakening of a pattern* in the first-learned language, due to experiential factors affecting one group but not the other, i.e. divergent attainment, attrition, or dominance shift. Having motivated our research questions we now turn to a brief description of pupillometry, given that it is a novel methodology for the field of linguistics and bilingualism research.

2.3 Pupillometry in psycholinguistics

The present study makes use of a methodology relatively new to the language sciences: pupillometry. Psychological and neurological work over the past several decades have demonstrated that changes in pupil size are linked not only to changes in ambient luminance, but also to aspects of the sympathetic nervous system (e.g., Goldwater, 1972) and the locus coeruleus and norepinephrine system (LC-NE; Aston-Jones & Cohen, 2005; Samuels & Szabadi, 2008). In particular, increases in pupil size have been linked to increased cognitive load, attentional allocation, and arousal, among other processes (see Einhäuser, 2017; Koelewijn et al., 2014, 2015; Krejtz et al., 2018; Schmidtke, 2018; Sirois & Brisson, 2014). Recently, pupillometry has been applied to the study of various linguistic processes, including effortful speech processing (Kuchinsky et al., 2013), sentence processing in older adults (Häuser et al., 2019), lexical retrieval in bilinguals (Schmidtke, 2014), bilingual cognate facilitation (Guasch et al., 2017), and the processing of language-mixed speech (Byers-Heinlein et al., 2017). Generally, an increase in pupil size with respect to a particular stimulus is assumed to be indicative of greater cognitive load resulting from an increase in the allocation of attentional resources (Alnæs et al., 2014; Gabay et al.,

2011). Given this, the present study uses pupillometry as an indicator of cognitive load in real time while participants listened to sentences.

Pupillometry was chosen as this study's processing measure over other on-line methods for three reasons: first, it allows the stimuli to be presented aurally as naturally timed audio clips. Presenting naturalistic audio clips of language is necessary for experiments with heritage speakers who often have limited home language literacy due to not being schooled in the minority language (Montrul, 2016, ch. 3). Second, it is less invasive than methods such as event-related potentials, which require lengthy set-up, increasing the duration of the experiment, participant fatigue, and potentially increasing linguistic insecurity in heritage speakers. Finally, the pupillary response is slower and less sensitive to small modulations than other on-line methods (such as the EEG signal). Pupillometry therefore permits some flexibility in the alignment of epochs across conditions, which was necessary for comparing the grammatical and ungrammatical conditions of some of our stimuli (see stimuli for relative clause and temporal adverbial islands in sections 3.1.2 and 4.1.2).

2.4 Research questions

The main issue we are addressing in this study is the differential experience of the first-learned language between types of bilingual speakers and how this might affect the grammatical representation and processing of that language. We compare two Spanish-English bilingual groups who differ in their experience with the first-learned language, Spanish. Heritage speakers, who are claimed in the literature to be prone to divergent attainment, attrition and a shift in dominance to the later-learned language; and late bilinguals, who are hypothesized to be more stable both in their representation as well as their processing of their first language—due to having completely acquired the first-learned language through consistent and continued use of that language beyond childhood and by having less life-time experience with the later-learned language, English. This study addresses two questions:

1. Does crosslinguistic influence alter the representation and processing of a structure (Comp-trace) that contrasts in grammaticality between Spanish and English? If so, does it have the same effect in heritage speakers and late bilinguals?

We compare the grammatical and ungrammatical forms of the Spanish Comp-trace structure (described in sections 2.1 and 3.1.2). The Spanish

Comp-trace structure, which requires the overt complementizer *que* ‘that’, stands in direct contrast with the equivalent English structure, where including the overt complementizer *that* results in ungrammaticality. We can formulate two hypotheses for how this crosslinguistic contrast in structure will affect our groups’ judgments and processing. The first hypothesis is that there is no crosslinguistic influence and the grammatical and ungrammatical conditions show a clear contrast. The second hypothesis is that crosslinguistic influence from English has made the standard ungrammatical structure (with no complementizer) acceptable, and the standard grammatical structure (with complementizer) unacceptable, as this is the pattern found in English. Further, we may expect the groups to show different patterns. Given that heritage speakers have greater lifetime exposure to English, they may show more influence from English.

2. Do bilinguals show sensitivity to island effects of different strengths in their first-learned language? If so, is this sensitivity different in heritage speakers and late bilinguals?

Here we use two types of islands: weak islands are exemplified with *wh*-islands and NP complements, while strong islands are exemplified with temporal adverbial and relative clause islands (see section 3.1.2 for sample sentences).

We ask whether heritage speakers and late bilinguals will show the same or different sensitivity to island violations of different strengths. First, given the experiential factors affecting heritage speakers, the prediction would be to see an overall weakening of island effects in this group compared to the late bilinguals. Second, since the effect is more pronounced in the strong than the weak islands, we might expect to see gradient sensitivity, such that strong islands (temporal adverbial, relative clause) will be treated differently from weak islands (*wh*-islands, noun complement). In the AJT this would result in higher acceptability of weak compared to strong islands. The pupillometry results will be more exploratory, given that this is (to our knowledge) the first measurement of island violation sensitivity using this method. Since pupillometry indexes cognitive load and processing difficulty, and under the assumption that unacceptability in grammar aligns with increased cognitive load in processing, we might expect greater pupil dilation in the strong island conditions than in the weak conditions.

3. Acceptability Judgment Experiment

3.1 Methods

3.1.1 Participants

Forty-five Spanish-English bilingual adults (aged 18-45, $M = 27.89$, $SD = 7.51$) participated in this study in New York City. All participants were screened using an extensive language background questionnaire (see Appendix 1) and self-rated their Spanish fluency as four or higher on a five-point scale of Spanish comprehension ($M = 4.87$, $SD = 0.33$). Participants were categorized as either Spanish heritage speakers ($n = 28$) or adult late bilinguals ($n = 17$) based on age of arrival (AoA) to the anglophone United States. Eighteen of the heritage speakers were born in the continental United States and the other ten moved to the United States before age eight ($M_{AoA} = 4.3$, $SD_{AoA} = 1.77$). Heritage speakers were raised speaking primarily Spanish until at least age 10 by Spanish-speaking immigrant parents originally from a Spanish-dominant country/region. Late bilinguals were raised in a Spanish-dominant country/region and moved to the anglophone US at the age of 15 or older ($M_{AoA} = 25.6$, $SD_{AoA} = 5.23$).

3.1.2 Stimuli

Stimuli for the experiment were designed as Spanish sentence-pairs. Each target stimulus was a *wh*- interrogative question preceded by a declarative context sentence. Different items were written for each of the five structural conditions (1-5): the Comp-trace construction ($n = 7$), *wh*- islands ($n = 10$), noun complement islands ($n = 10$), temporal adverbial islands ($n = 10$), and relative clause islands ($n = 15$). Each item appeared in two grammaticality conditions, grammatical (*b* sentences) and ungrammatical (*c* sentences). The ungrammatical conditions were formed by extracting a noun phrase from the position following a null complementizer (\emptyset in 7c) or from a syntactic island phrase (8c, 9c, 10c, 11c). The grammatical conditions were minimally changed from the ungrammatical by adding an overt complementizer (7b, 8b), removing a noun phrase (9b) or changing the extracted noun phrase (10b, 11b). Participants heard both the grammatical and ungrammatical versions of all target stimuli across the five conditions ($N = 104$). The “||” symbols in the stimuli below mark points of measurement for the pupillometry experiment, which used similar stimuli, described in section 4.1.2.

- (7) Comp-trace
- a. El buzo exclamaba que un tiburón había
the diver shout.IMP.3SG COMP a shark have.IMP.3SG
mordido su tanque de oxígeno.
bite.PART his tank of oxygen
'The diver shouted that a shark had bitten his oxygen tank.'
 - b. ¿Qué tiburón exclamaba el buzo || que
what shark shout.IMP.3SG the diver COMP
había mordido su tanque de oxígeno?
have.IMP.3SG bite.PART his tank of oxygen
'What shark did the diver shout that had bitten his oxygen tank?'
 - c. *¿Qué tiburón exclamaba el buzo Ø || había
what shark shout.IMP.3SG the diver have.IMP.3SG
mordido su tanque de oxígeno?
bite-PART his tank of oxygen
'What shark did the diver shout had bitten his oxygen tank?'
- (8) *Wh*- island
- a. Ignacio confirmó por qué la enfermera
Ignacio confirm.PRET.3SG why the nurse
había llevado la medicina.
have.IMP.3SG bring.PART the medicine
'Ignacio confirmed why the nurse had brought the medicine.'
 - b. ¿Qué enfermera confirmó Ignacio || que
what nurse confirm.PRET.3SG Ignacio COMP
había llevado la medicina?
have.IMP.3SG bring.PART the medicine
'What nurse did Ignacio confirm had brought the medicine?'
 - c. *¿Qué enfermera confirm Ignacio || por qué
what nurse confirm.PRET.3SG Ignacio why
había llevado la medicina?
have.IMP.3SG bring.PART the medicine
'What nurse did Ignacio confirm why had brought the medicine?'

- (9) Noun complement island
- a. Juan contó el chisme que el vecino
 Juan tell.PRET.3SG the gossip COMP the neighbor
 robó el carro anoche.
 rob.PRET.3SG the car last.night
 'Juan told the gossip that the neighbor stole the car last night.'
 - b. ¿Qué vecino contó Juan || que robó
 what neighbor tell.PRET.3SG Juan COMP rob.PRET.3SG
 el carro anoche?
 the car last.night
 'What neighbor did Juan tell that stole the car last night?'
 - c. *¿Qué vecino contó Juan || el chisme que
 what neighbor tell.PRET.3SG Juan the gossip COMP
 robó el carro anoche?
 rob.PRET.3SG the car last.night
 'What neighbor did Juan tell the gossip that stole the car last night?'
- (10) Temporal adverbial island
- a. El niño comió el dulce mientras que su
 the child eat.PRET.3SG the candy while COMP his
 tía buscaba la comida.
 aunt search.IMP.3SG the food
 'The child ate the candy while his aunt looked for food.'
 - b. ¿Qué niño comió el dulce || mientras que su
 what child eat.PRET.3SG the candy while COMP his
 tía buscaba la comida?
 aunt search.IMP.3SG the food
 'What child ate the candy while his aunt looked for food?'
 - c. *¿Qué tía || el niño comió el dulce mientras
 what aunt the child eat.PRET.3SG the candy while
 que buscaba la comida?
 COMP search.IMP.3SG the food
 'What aunt did the child eat the candy while looked for food?'

- (11) Relative clause island
- a. Paola hizo el gesto que causó
 Paola make.PRET.3SG the joke COMP cause.PRET.3SG
 la controversia.
 the controversy
 ‘Paola made the joke COMP caused the controversy.’
 - b. ¿Qué gesto hizo Paola || que causó
 what joke make.PRET.3SG Paola COMP cause.PRET.3SG
 la controversia?
 the controversy
 ‘What joke did Paola make that caused the controversy?’
 - c. *¿Qué controversia hizo Paola || el gesto
 what controversy make.PRET.3SG Paola the joke
 que causó?
 COMP cause.PRET.3SG
 ‘What controversy did Paola make the joke that caused?’

3.1.3 Procedure

The experiments described in this study (the AJT, described here, and the pupillometry experiment, described in section 4) were conducted in two separate testing sessions, the pupillometry experiment in the first, and the AJT experiment in a second session at least 10-14 days later.

For the AJT task, the stimuli were presented aurally in sentence dyads: a declarative context sentence followed by the target sentence. After each dyad, participants rated the naturalness of the target sentence, the *wh*-question, using a 5-point Likert scale with the ends and midpoints labeled in Spanish (1 = *es natural* “natural”, 3 = *es posible* “maybe”, and 5 = *no es natural* “not natural”). Items were pseudorandomized (no conditions were repeated back-to-back, neither structural nor both grammatical conditions of the same item) over two blocks, each lasting approximately 20 minutes with a short break in-between. Participants were given instructions in Spanish or English—their preference—and completed a practice block before beginning the experiment.

3.1.4 Analysis

Participant-average rating was plotted and calculated in R to explore trends in the data. Cumulative link mixed-effects models (CLMMs) were used to analyze the acceptability judgements as these models are well-suited for dealing with ordinal data such as those extracted from a Likert scale (e.g.,

Douven, 2018 and references therein). Models were created using the `clmm` function in the `ordinal` library (Christensen, 2019). Models predicted the acceptability judgements by group (late bilingual, heritage speaker), grammaticality (grammatical, ungrammatical), and their interaction. Random by-subject and by-item intercepts were included, with the threshold (cut-points) specified to be equidistant. First, each structural condition (Comp-trace, *wh*- island, noun complement, temporal adverbial, and relative clause) was analyzed separately; in the case of a significant two-way interaction between group and grammaticality, follow-up models were run analyzing the grammatical and ungrammatical items separately to examine between-group differences. Next, a follow-up analysis was conducted on just the four island types' ungrammatical items to compare each groups' ratings to look for effects of island type (i.e., strong and weak islands).

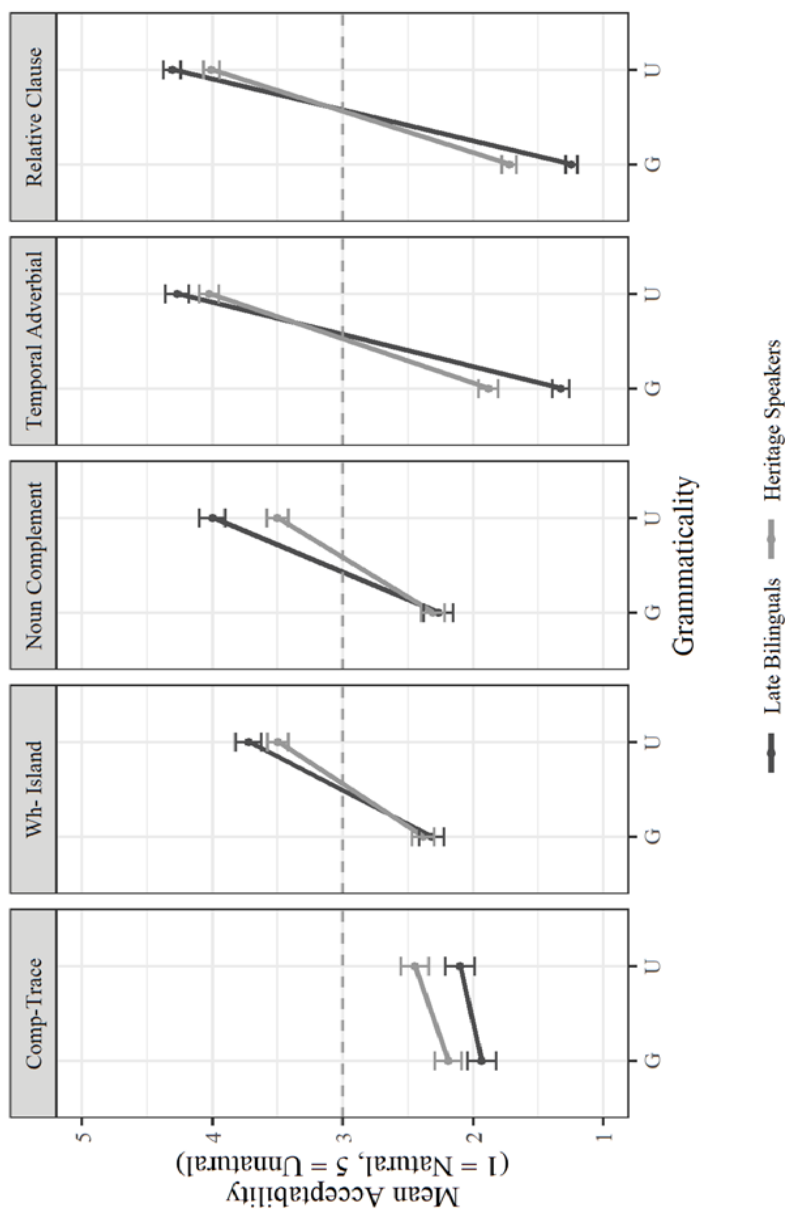
3.2 AJT results

Average ratings are summarized in Table 1 and visualized in Figure 1 below.

Table 1. Average ratings for acceptability judgment task (AJT); standard deviations in parentheses.

	Comp- Trace	Wh- Island	Noun Complement	Temporal Adverbial	Relative Clause
Late Bilinguals					
<i>Grammatical</i>	1.93 (1.20)	2.32 (1.25)	2.26 (1.44)	1.32 (0.83)	1.24 (0.71)
<i>Ungrammatical</i>	2.10 (1.22)	3.72 (1.30)	4.00 (1.29)	4.27 (1.16)	4.31 (1.04)
Heritage Speakers					
<i>Grammatical</i>	2.19 (1.44)	2.38 (1.43)	2.31 (1.48)	1.88 (1.24)	1.72 (1.14)
<i>Ungrammatical</i>	2.44 (1.49)	3.50 (1.35)	3.50 (1.36)	4.03 (1.27)	4.01 (1.23)

Figure 1. Average ratings for acceptability judgment task (AJT).



For the Comp-trace condition, there were no significant main effects or interactions (all p 's $> .2$), suggesting that late bilinguals and heritage speakers rated both grammatical and ungrammatical conditions equally. For *wh*- islands, both groups rated ungrammatical items as less natural than grammatical items ($Z = 6.71, p < .001$). For noun complement islands, both groups rated ungrammatical items as less natural than grammatical items ($Z = 6.70, p < .001$). A significant interaction between group and grammaticality ($Z = -3.30, p < .001$) showed that while both late bilinguals and heritage speakers rated grammatical items similarly ($Z = 0.06, p = .95$), late bilinguals rated ungrammatical items as significantly less natural than heritage speakers ($Z = -2.79, p < .01$). For temporal adverbial islands, both groups rated ungrammatical items as less natural than grammatical items ($Z = 13.53, p < .001$), and heritage speakers rated all items in this condition as less natural than late bilinguals ($Z = 4.05, p < .001$). A significant interaction between group and grammaticality ($Z = -5.94, p < .001$) showed that while late bilinguals rated grammatical items as significantly more natural than heritage speakers ($Z = 3.23, p < .01$), both groups rated ungrammatical items similarly ($Z = -1.56, p = .12$). Lastly, for relative clause islands, both groups rated ungrammatical items as less natural than grammatical items ($Z = 17.51, p < .001$), and heritage speakers rated all items in this condition as less natural than late bilinguals ($Z = 4.57, p < .001$). A significant interaction between group and grammaticality ($Z = -7.06, p < .001$) showed that while late bilinguals rated grammatical items as significantly more natural than heritage speakers ($Z = 3.21, p < .01$), they rated ungrammatical items as only marginally more unnatural than heritage speakers ($Z = -1.71, p = .087$).

When comparing just the ungrammatical items across the four island types, each group showed slightly different effects. Overall, ungrammatical *wh*-island and noun complement items were rated as only marginally different from one another ($Z = 1.77, p = .076$). Ungrammatical *wh*- island items were rated as significantly more natural than ungrammatical temporal adverbial items ($Z = 3.48, p = .001$), but there was only a marginal difference between ungrammatical noun complement and temporal adverbial items ($Z = 1.70, p = .089$). Lastly, ungrammatical relative clause items did not differ from ungrammatical temporal adverbial items ($Z = 0.15, p = .88$), but they were rated as marginally less natural than ungrammatical noun complement items ($Z = -1.72, p = .086$) and significantly less natural than ungrammatical *wh*-island items ($Z = -3.70, p < .001$). When comparing across the two groups, we see that while heritage speakers showed no differences between the ungrammatical *wh*- island and noun complement items, late bilinguals did ($Z = -1.99, p = .047$). Given this, the following two hierarchies can be deduced for each group:

Late Bilinguals: *wh*- island < Noun Complement \lesssim Temporal Adverbial
= Relative Clause

Heritage Speakers: *wh*- island = Noun Complement < Temporal
Adverbial = Relative Clause

3.3 Summary of AJT Results

For the Comp-trace items, both groups rated the grammatical and ungrammatical questions as natural, with no difference between grammaticality conditions. For the four island structures, both groups rated the ungrammatical violations as significantly less natural than their grammatical counterparts. For both groups, the stronger islands (relative clause, temporal adverbial) showed a greater difference in ratings between the grammaticality conditions than the weaker islands (*wh*- island, noun complement). Late bilinguals rated ungrammatical violations as less natural than the heritage speakers in the noun complement and relative clause island conditions. For the temporal adverbial condition, late bilinguals rated grammatical items as more natural than the heritage speakers.

4. Pupillometry Experiment

4.1 Methods

4.1.1 Participants

Fifty-six participants completed the pupillometry experiment (aged 18-45, $M = 28.02$, $SD = 7.41$); these participants were a superset of the participants for the AJT experiment. All participants were screened using an extensive language background questionnaire (see Appendix 1) and self-rated their Spanish fluency as four or higher on a five-point scale of Spanish comprehension ($M = 4.88$, $SD = 0.32$). Late bilinguals ($n = 24$) arrived in the anglophone United States after age 15 ($M_{AoA} = 25.6$, $SD_{AoA} = 5.17$) and heritage speakers ($n = 32$) were either born in the United States ($n = 19$) or arrived before age eight ($n = 13$, $M_{AoA} = 3.69$, $SD_{AoA} = 2.10$).

4.1.2 Stimuli

The stimuli for the pupillometry task were the same as those used in the AJT task, described in section 3.1.2, except that each syntactic condition contained more items. The noun complement, temporal adverbial, and *wh*-island condition each had 30 items. The relative clause island condition had

45 items, and the comp-trace condition had 20 items. Items were presented in each of the two grammaticality conditions for a total of 310 target stimuli. Within each stimulus, a timestamp was marked for the epoch of interest, indicated in the stimuli in (7-11) with “||” (for sake of space, we do not repeat the stimuli here). For the ungrammatical stimuli, the beginning of the epoch was the point at which the sentence becomes ungrammatical. For the grammatical stimuli, it was the beginning of the structure of interest.

4.1.3 Procedure

Target stimuli were presented aurally, preceded by a context sentence (same design as for the AJT, see section 3.1.3), and pseudorandomized over five blocks. The experiment took approximately 1.5 hours, including set-up, with breaks every 10 minutes. Participants were not asked to make any metalinguistic judgments about the target items in the pupillometry experiment. A white fixation marker “+” centered on a black screen was provided throughout the auditory blocks. Following 40% of the auditorily-presented trials, the participants were prompted on screen to answer a yes/no comprehension question about the preceding sentence pair. The comprehension question was unrelated to the research questions in this study and served the purpose of ensuring participants’ continued attention. The auditory block resumed after participants answered the comprehension question. Participants were given instructions in Spanish or English—their preference—and completed a practice block before beginning the experiment. Pupil diameter and gaze location were recorded with Tobii TX300 infrared cameras for each eye separately. Data were recorded for the whole trial (including the context and target) and 1 second (1000ms) before and after the trial at 60Hz (one sample every 16.67 milliseconds).

4.1.4 Analysis

Before analyzing the pupillary response, data were first pre-processed. For each trial, any samples that were marked as invalid during recording (a Tobii validity code of 1, 2, 3, or 4) were excluded; this includes the pupil diameter and x- and y- gaze positions for both the left and right eyes. Missing samples were not interpolated as interpolation can increase autocorrelation in the residuals leading to anticonservative models (see van Rij et al., 2019, p. 5). Next, the pupil diameter and x- and y- gaze positions were averaged for the left and right eyes. Data were time-locked to the point of ungrammaticality (and the corresponding position in each grammatical counterpart). The average pupil size was calculated during the 200ms (12 sample) period before the onset of this epoch, and baseline subtraction was performed to

account for non-stimulus-related changes in pupil size during the course of the experiment. Trials where more than 35% of all samples were marked for exclusion were removed, resulting in 37% of all trials being removed.² Participants with an insufficient number of trials within each structural condition were likewise excluded from the analysis for that particular condition only (Comp-trace: 17 participants; *wh-* island: 9 participants; noun complement: 12 participants; temporal adverbial: 4 participants; relative clause: 4 participants).

Data were analyzed using generalized additive mixed-effects models (GAMMs) using the `bam` function in the `mgcv` package (Wood, 2011); model criticism, plotting, and testing was carried out using the `itsadug` package (van Rij et al., 2020). GAMMs are ideal for analyzing time-series data, like the pupillary response, as they allow the fitting of non-linear smooths and include options to account for autocorrelation—an inherent issue in time-series data. The modelling procedure partially followed the recommendations laid out by van Rij and colleagues (2019). To facilitate analyses, a four-level factor *GramGen* was created representing the grammatical and ungrammatical conditions for both the heritage speakers and late bilinguals. For all structural conditions, an initial maximal model was specified, estimating the baseline subtracted pupil size by: 1) the parametric coefficients *GramGen* and *Session* (used to account for possible differences between the first and second halves of the experiment); 2) a smooth term by *GramGen*; 3) a smooth term capturing the x- and y- gaze position (used to account for changes in pupil size related to gaze position; see Gagl et al., 2011); and 4) by-subject and by-item random factor smooths. To account for autocorrelation in the residuals characteristic of time-series data, a nested AR1 model was included in each GAMM, with the correlation coefficient ρ selected by testing different values and converging on a value that sufficiently reduced autocorrelation in the residuals without inducing anticorrelation (using the `acf_resid` function in the `itsadug` package). Model criticism was performed with the help of the `gam.check` function; fitted smooths were plotted using the `plot_smooth` function, and difference smooths were calculated and visualized using the `plot_diff` function, all in the `itsadug` package.

² While this amount of excluded data may seem high compared to behavioral methods, where more than 10% of data excluded would be rare, this is not the case for pupillometry data. For a discussion, see Schmidtke (2018, pp. 542-543).

4.2 Pupillometry results

All results described below are based on the models containing the nested AR1 model to account for autocorrelation. Full models are given in Appendix 2. For all figures presented below, solid lines represent the grammatical conditions while dashed lines represent the ungrammatical conditions. For each structural condition, the grammatical and ungrammatical conditions were compared within each bilingual group, allowing us to examine grammaticality effects within each group. The fitted smooths for the Comp-trace condition are given below in Figure 2. Difference smooths indicated that there were no significant differences between the grammatical and ungrammatical Comp-trace conditions for late bilinguals. For heritage speakers, the ungrammatical Comp-trace condition elicited a significantly larger pupillary response from approximately 555ms to 1778ms (samples 33.33-106.67) post-target onset.

Fitted smooths for the *wh*- island condition are given below in Figure 3. For late bilinguals, difference smooths indicated that the grammatical *wh*- island condition elicited a significantly larger pupillary response from approximately 1533ms post-target onset until the end of the target epoch (samples 92-120). For heritage speakers, the grammatical *wh*- island condition elicited a significantly larger pupillary response from approximately 466ms post-target onset until the end of the target epoch (samples 28-120).

Figure 2. Fitted and difference smooths for Comp-trace condition with AR1.

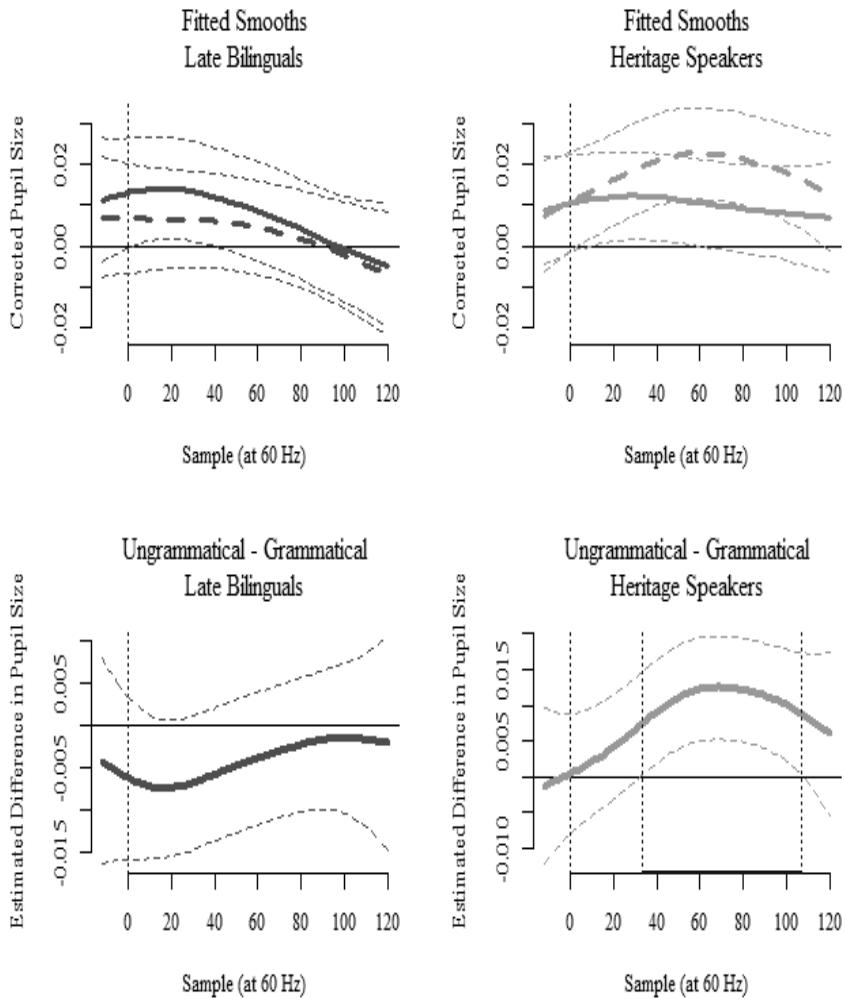
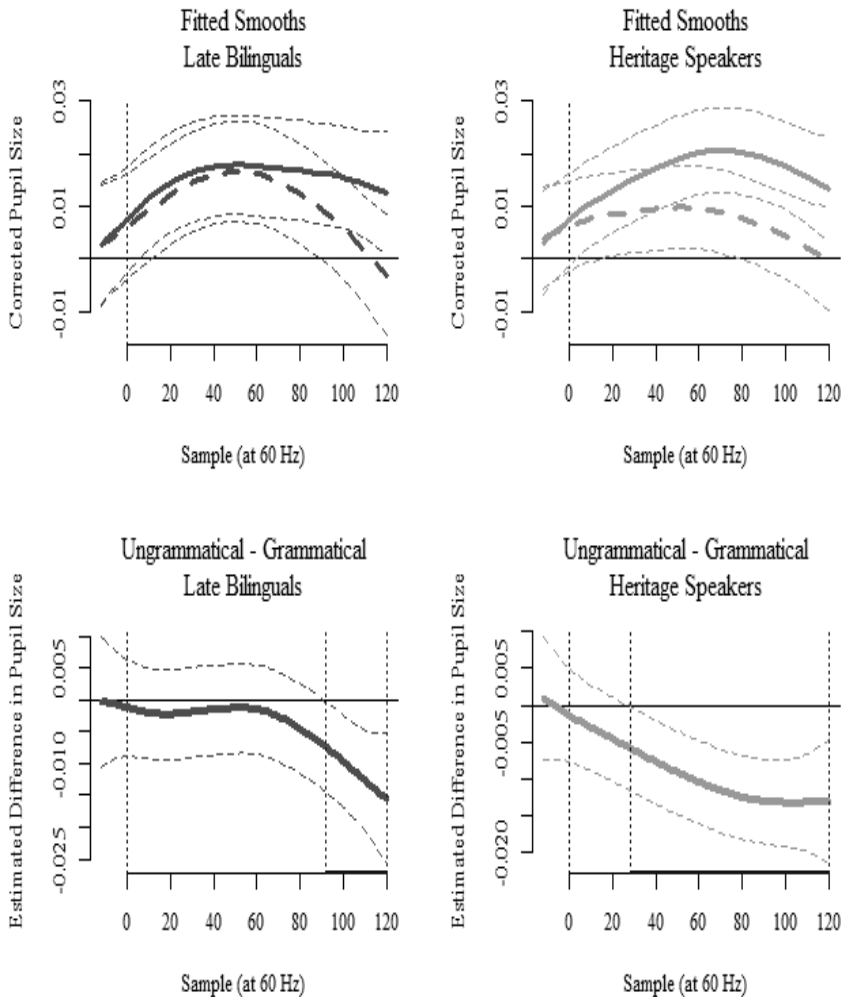


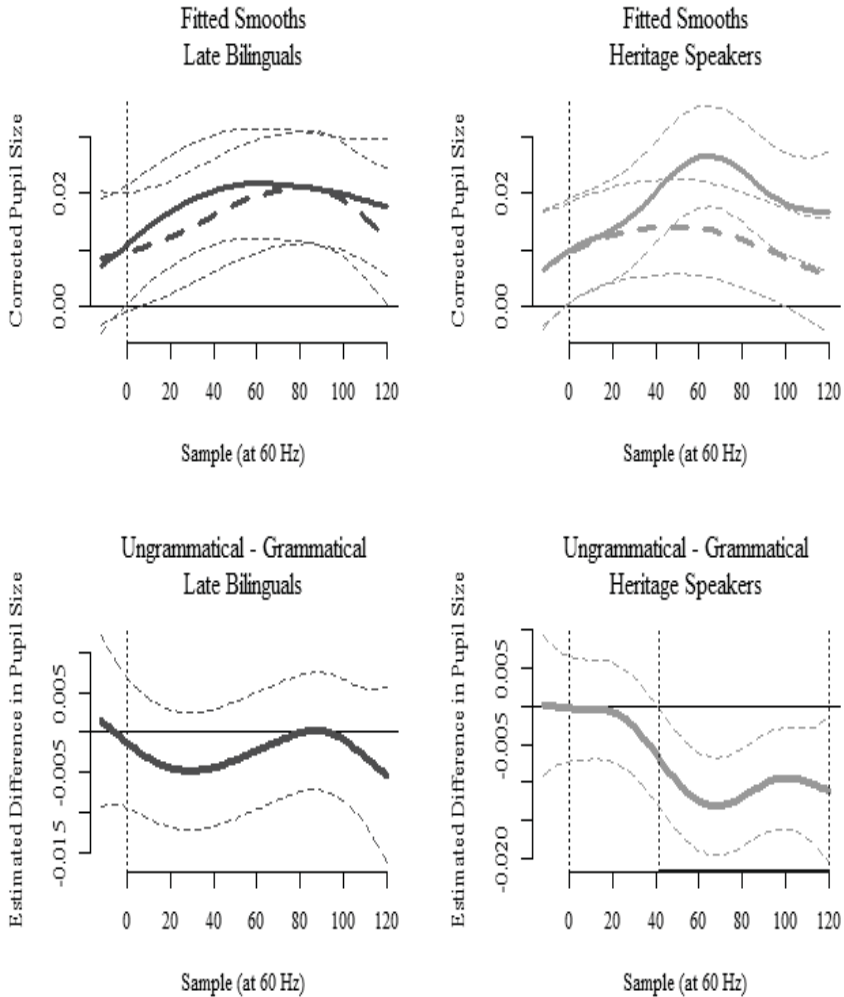
Figure 3. Fitted and difference smooths for *wh*- island condition with AR1.



Fitted smooths for the noun complement island condition are given below in Figure 4. For late bilinguals, difference smooths suggested that there were no significant differences between the grammatical and ungrammatical conditions. For heritage speakers, the grammatical condition elicited a significantly larger pupillary response than the ungrammatical condition

from approximately 689ms post-target onset until the end of the target epoch (samples 41.33-120).

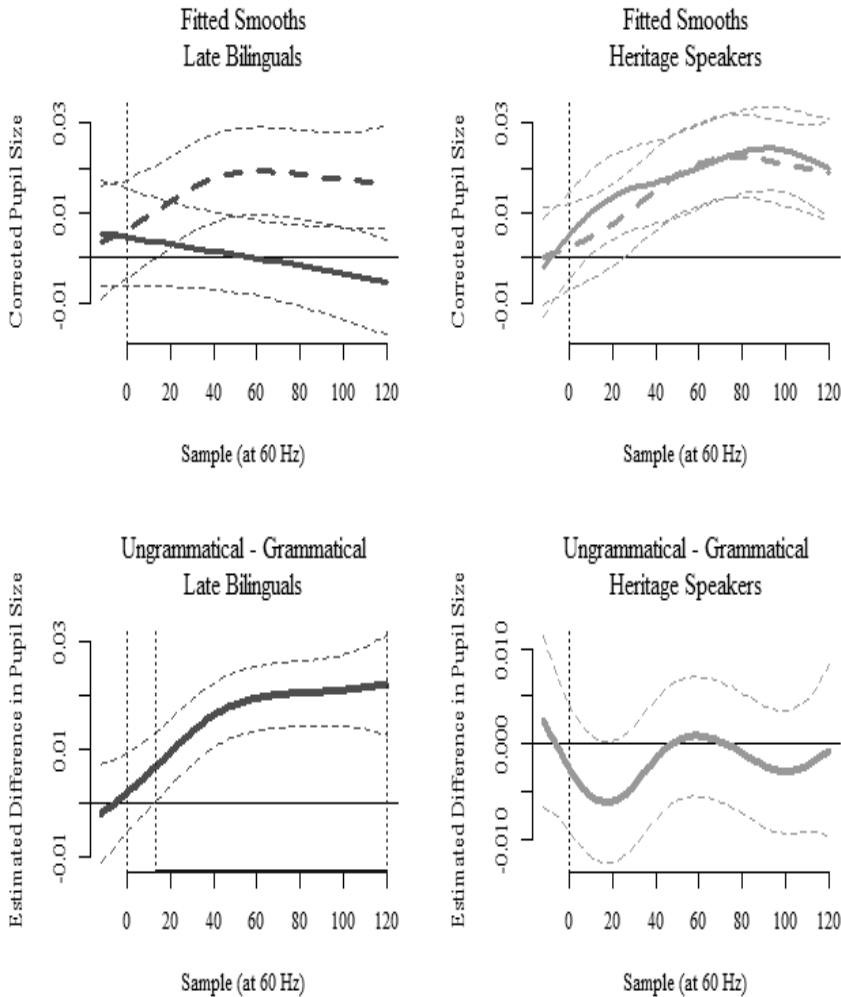
Figure 4. Fitted and difference smooths for noun complement condition with AR1.



Fitted smooths for the temporal adverbial island condition are given below in Figure 5. For late bilinguals, difference smooths indicated that the ungrammatical condition elicited a significantly larger pupillary response

from approximately 222ms post-target onset until the end of the target epoch (samples 13.33-120). For heritage speakers, there were no significant differences between the grammatical and ungrammatical conditions.

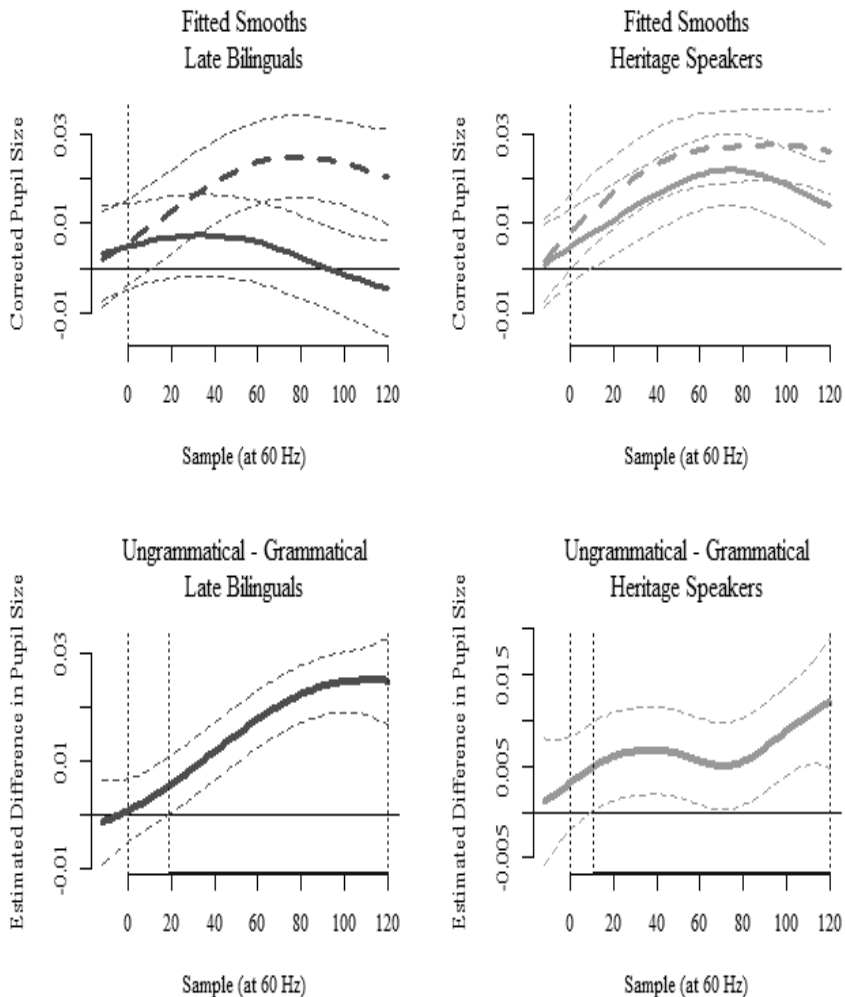
Figure 5. Fitted and difference smooths for temporal adverbial condition with AR1.



Fitted smooths for the relative clause island condition are given below in Figure 6. For late bilinguals, difference smooths indicated that the

ungrammatical condition elicited a significantly larger pupillary response from approximately 311ms post-target onset until the end of the target epoch (samples 18.67-120). For heritage speakers, difference smooths indicated that the ungrammatical condition elicited a significantly larger pupillary response from approximately 117ms post-target onset until the end of the target epoch (samples 10.67-120).

Figure 6. Fitted and difference smooths for relative clause condition with AR1.



4.3 Summary of pupillometry results

For the Comp-trace violation, heritage speakers showed greater dilation for the ungrammatical question, but late bilinguals did not. For the syntactic island structures, both bilingual groups showed an ungrammaticality effect only in the strong island conditions: late bilinguals showed a consistent increase in dilation for temporal adverbial and relative clause island violations, while heritage speakers only showed an effect in the relative clause condition. Heritage speakers showed a reverse pattern for the weaker island effects. In the *wh*- island and noun complement conditions, heritage speakers exhibited increased dilation for the grammatical, as compared to the ungrammatical, conditions. Late bilinguals showed a similar response to grammaticality in the *wh*- island condition, though their response was later, about 1.6 seconds after the start of the epoch.

5. Discussion

We discuss the results from the AJT and the pupillometry experiments in light of the two research questions posed in section 2.4 focusing on the similarities and differences we found between the two groups and what these might tell us about the grammatical representation on the one hand and the processing strategies on the other, in the Spanish varieties of these two groups. Our first question addresses crosslinguistic influence and compares a contrast in acceptability in Spanish and English. Our second question looks at island effects which hold in both Spanish and English, but which may have diminished in strength in the first-learned language for the heritage speakers, who are susceptible to a variety of factors such as divergent attainment (e.g., Polinsky & Scontras, 2020), attrition (e.g., Schmid, 2011), or shift in dominance (e.g., Rothman, 2009). In considering the extent to which metalinguistic judgements do or do not align with psychophysiological measures we make some speculative remarks regarding the relationship between grammatical representation and cognitive load.

Research question 1: Do we see crosslinguistic influence in heritage speakers and late bilinguals?

The relevant construction here was the Comp-trace sentence, where the absence or presence of the complementizer, *que* ‘that’, is reported to result in contrasting grammaticality in standard Spanish (Torrego, 1984) and standard English (Coward, 1997; Sobin, 2002). We will discuss the AJT and pupillometry results in turn.

Comp-trace: Acceptability judgment task

Results from the AJT show some degree of crosslinguistic influence in both groups. As discussed in section 2.1, in Spanish the complementizer *que* is considered obligatory in Comp-trace constructions, so that sentences without *que* should have received lower naturalness ratings than sentences with *que*. Instead, both groups rated sentences with and without *que* as natural (see Table 1), showing crosslinguistic influence from the equivalent structures in English, where the complementizer *that* is absent.³ However, influence from English was not complete since both groups also accepted the standard Spanish pattern with *que*, indicating optionality rather than a complete shift to the English pattern. It seems that the grammar for these constructions in the Spanish varieties of the bilinguals we tested is undergoing change, allowing both English and Spanish patterns. Furthermore, the similarity of responses between the two groups does not support the conventionally expected difference between heritage speakers and late bilinguals. As mentioned before, heritage speakers are characterized as being more susceptible to crosslinguistic influence than late bilinguals, having been exposed to the influencing language at an earlier age. Our results go against this characterization, suggesting that earlier age of arrival in the influencing language did not affect responses to this construction.

Comp-trace: Pupillometry

Unlike the AJT, results from the pupillometry experiment showed different patterns in the two groups (cf. Dussias & Sagarra, 2007). Here we expected greater pupil dilation for sentences without *que*, as these go against standard Spanish. In addition, standard assumptions of greater crosslinguistic influence in heritage speakers led to the expectation that only late bilinguals would show this pattern. Instead, we saw the opposite: Late bilinguals had descriptively greater mean pupil dilation for sentences with *que* than without *que*, although this difference did not reach significance. Heritage speakers, on the other hand showed significantly greater dilation for the sentences without *que* than for sentences with *que*, in line with expectations for standard Spanish, indicating absence of influence from English. One way to interpret these results is to say that the optionality seen in the AJT was evidenced in the pupillometry results, but only in the late bilingual group. The heritage speakers on the other hand, showed a pattern that

³ This is true for standard American English, although there is some evidence of optionality in certain dialects (see Sobin, 1987).

conforms more to what would be expected for the standard Spanish pattern showing again, that conventional assumptions about heritage language are unsupported here. If earlier exposure to the later-learned language makes HS more susceptible to crosslinguistic influence than late bilinguals, we should have seen the reverse pattern in pupil dilation. The results also suggest that alignment between judgments and processing load was only found for the late bilingual group, while in the heritage speaker group we saw non-alignment between grammatical and processing measures.

Research question 2: Do heritage speakers and late bilinguals show sensitivity to island effects of different strengths in their first-learned language?

Four island conditions were tested: *wh*- islands and noun complement represented weak island violations, while temporal adverbial islands and relative clause islands were included as examples of strong island violations (Chomsky, 1986; Szabolcsi, 2006).

Syntactic islands: Acceptability judgment task

Results indicate that overall, island effects still hold in the Spanish varieties of the two groups, as both clearly differentiated between grammatical and ungrammatical sentences across conditions. Secondly, both groups assigned higher unnaturalness ratings to strong violations (temporal adverbial, relative clause) than to weak violations (*wh*- islands, noun complements). Together, this suggests 1) similarity of representation of these structures across heritage speakers and late bilinguals and 2) psychological reality of gradience across weak and strong constraints.⁴

Since the results largely indicate similar, if not equal grammatical representation in the two bilingual groups, we can conclude that for the heritage speakers we tested island effects were not affected by divergent attainment, attrition or a shift in dominance. Group differences consisted primarily of a larger differential between grammatical and ungrammatical sentences for the late bilinguals than for the heritage speakers. In both weak and strong island conditions there was a tendency for heritage speakers to rate ungrammatical sentences as more natural compared to the late bilinguals, although this difference was only statistically significant for noun complement islands in the weak and relative clause islands in the

⁴ We leave open the question whether this gradience is the result of grammatical constraints (Chomsky, 1986) or processing constraints (Hofmeister & Sag, 2010).

strong condition. Nonetheless, this tendency is reminiscent of the indeterminacy often found in L2 learners' grammaticality ratings. While we intentionally selected fluent heritage speakers for this study, a certain level of indeterminacy would still be expected for this group, since Spanish, although first-learned, is more likely the less dominant language for this population, having been immersed in English at earlier ages. Late bilinguals, on the other hand, showed a wider spread between grammatical and ungrammatical sentences (see Figure 1), suggesting a greater degree of certainty in the judgement task, similar to what is found for monolinguals.

Syntactic islands: Pupillometry

Here results were more complex than in the AJT experiment, and, as with the findings for the Comp-trace construction, were largely unresponsive of an alignment between grammatical representation and processing. Response patterns to weak and strong islands showed some similarity across the two groups, with weak islands producing indeterminate pupil dilation and strong islands producing greater dilation for ungrammatical items. But differences were also evident. The pupillometry results are summarized in Table 2.

Table 2. Summary of the significant differences in pupillary dilation between ungrammatical (UNG) and grammatical (GR) conditions, by island type by group.

	WEAK		STRONG	
	<i>Wh-</i> island	Noun Complement	Temporal Adverbial	Relative Clause
<i>Heritage speakers</i>	GR > UNG	GR > UNG	GR = UNG	UNG > GR
<i>Late bilinguals</i>	GR > UNG	GR = UNG	UNG > GR	UNG > GR

In the *wh-* island condition, both groups had greater pupil dilation for grammatical than for ungrammatical items. This goes against the hypothesis that ungrammatical items would be harder to process than their grammatical counterparts, again suggesting a fundamental difference between measures of grammatical representation and measures of cognitive load. Given the novelty of the pupillometry measure in psycholinguistic studies, this result is difficult to interpret, but a possible linguistic explanation suggests itself when we look at the actual sentences. The grammatical and ungrammatical

items from the *wh*- island condition, given fully in (8), are repeated here in (12).

- (12) *Wh*- island
- b. ¿Qué enfermera confirmó Ignacio que había llevado la medicina?
'What nurse did Ignacio confirm had brought the medicine?'
 - c. *¿Qué enfermera confirmó Ignacio por qué había llevado la medicina?
'What nurse did Ignacio confirm why had brought the medicine?'

Although (12c) is clearly degraded, we see that it still has some level of interpretability. In English, such sentences are often "repaired" with resumptive pronouns (Ross, 1967), which has the effect of reducing unacceptability without much restructuration. In fact, when we look at the results from the AJT, we see that these sentence types also received relatively low unnaturalness ratings (see Table 1). It is possible that this reduction in the complexity differential between grammatical and ungrammatical items in the *wh*- island condition resulted in inconsistent dilation patterns. If so, we could surmise that near-interpretability might mitigate increase in processing load, thus not producing the greater dilation we expected. Considering that the ungrammatical *wh*- island items received greater naturalness ratings than those in the other conditions, it is possible that near-interpretability had an effect on both AJT and pupillometry.

Similar arguments could be made for the noun complement condition where the ungrammatical version (13b) is also still interpretable, although perhaps to a lesser degree than for *wh*- islands.

- (13) Noun complement island
- b. ¿Qué vecino contó Juan que robó el carro anoche?
'What neighbor did Juan tell that stole the car last night?'
 - c. *¿Qué vecino contó Juan el chisme que robó el carro anoche?
'What neighbor did Juan tell the gossip that stole the car last night?'

In sum, comparing AJT to pupillometry in the weak conditions, we do not find strong evidence for alignment between grammaticality and cognitive

load, since the pupillometry results failed to show the clear differentiation between grammatical and ungrammatical sentences found in the AJT. Only in the strong conditions (temporal adverbial and relative clause islands) was there a tendency for increased pupil size in the ungrammatical conditions. Late bilinguals showed this result in both the strong conditions, while the heritage speakers showed this result only for relative clause islands, thus pointing to a difference in sensitivity in the processing of the two types of sentences. This difference might partly be explained by factors of experience with and exposure to Spanish, which for heritage speakers tends to be the less dominant language. The high complexity of both grammatical and ungrammatical sentences in the strong condition might result in pupil size fluctuations for the heritage speaker group reflecting inter-item variability and/or reduced knowledge of transitional probabilities, effectively canceling the differential in the temporal adverbial condition. If this explanation is on the right track, we would have to assume a critical difference between the temporal adverbial and relative clause structures that would cause the expected reaction in the heritage speakers to the latter but not the former.

To summarize, while the AJT experiment provided clear evidence for the recognition of syntactic islands as well as for gradience across weak and strong conditions in both groups, pupil dilation revealed the expected patterns only in the strong conditions, and more consistently so for the late bilinguals.

6. Conclusion

Although this study was exploratory in nature, results point to several potential conclusions. Firstly, heritage speakers and late bilinguals did not show significant differences in the way they judged any of the structures tested, indicating that their grammatical representation for these structures is largely, if not wholly, the same. With regard to crosslinguistic influence from English, both groups showed a shift towards optionality of *que*-deletion in Comp-trace sentences, away from standard Spanish. Syntactic islands were recognized as such by both groups, suggesting, contra standard assumptions, that there was no weakening effect in the grammars of the heritage speakers. This strongly suggests that the assumptions made in the literature about heritage speakers do not hold for the population we tested, as they did not show more influence from English than we saw in the late bilinguals, nor more weakening of grammatical principles in the first-learned language. Thus, we can conclude that the grammatical representation

of *wh*- gaps, such as the Comp-trace construction and syntactic islands, were not different in the two groups.

Our pupillometry results paint a more complex picture: the similarity we found in the AJT experiment across heritage speakers and late bilinguals largely disappeared. If this result holds with larger numbers of participants, it suggests a dichotomy between the grammatical representation and the processing patterns utilized by the two bilingual types. In particular, it is an indication that even as the two groups exhibit similar grammars, their processing strategies show significant differences. In fact, only in the relative clause island condition did we find similar patterns across the two groups, both in judgment and pupil dilation. While this must remain speculative, a possible explanation could be that there is a threshold of complexity at which grammar and processing, at least as measured by pupillometry, converge. We have suggested that one of the factors mitigating complexity and unacceptability could be interpretability, as shown in the weak island conditions.

As research comparing different types of bilinguals continues, probing both grammar and processing in these populations, more discoveries will surely emerge about the complexities of the bilingual mind. Our study is only one small step in that direction.

Acknowledgements

We would like to thank all of the participants who participated in this study. We also thank the Second Language Acquisition Lab research assistants Matthew G. Stuck, Pamela Franciotti, João Pedro Marinotti, Lianye Zhu, Reid Vancelette, Jennifer Chard, Kevin P. Guzzo, Armando Tapia, Michael Stern, Anthony Vicario, LeeAnn Stover, Andrea Monge, and Benjamin Shavitz, who helped to make this study possible. Many thanks also to Vincent Torrens for organizing the 2019 Experimental Psycholinguistics Conference in Mallorca, Spain, where an earlier version of this study was presented. This work was partially funded by a New York State Department of Education Grant (#016-042) to the first author.

References

- Alnæs, D., Sneve, M. H., Espeseth, T., Endestad, T., van de Pavert, S. H. P., & Laeng, B. (2014). Pupil size signals mental effort deployed during multiple object tracking and predicts brain activity in the dorsal attention network and the locus coeruleus. *Journal of Vision*, 14(4), 1–1.
- Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience*, 28, 403–450.
- Benmamoun, E., Montrul, S., & Polinsky, M. (2013). Heritage languages and their speakers: Opportunities and challenges for linguistics. *Theoretical Linguistics*, 39(3–4), 129–181.
- Byers-Heinlein, K., Morin-Lessard, E., & Lew-Williams, C. (2017). Bilingual infants control their languages as they listen. *Proceedings of the National Academy of Sciences*, 114(34), 9032–9037.
- Chomsky, N. (1986). *Barriers*. Cambridge, MA: The M.I.T. Press.
- Christensen, R. H. B. (2019). *Ordinal—Regression models for ordinal data*. R package version 2019.12-10.
<https://CRAN.R-project.org/package=ordinal>
- Cowart, W. (1997). *Experimental Syntax*. London: Sage.
- de Houwer, A., & Ortega, L. (2018). *The Cambridge handbook of bilingualism*. Cambridge: Cambridge University Press.
- Douven, I. (2018). A Bayesian perspective on Likert scales and central tendency. *Psychonomic Bulletin & Review*, 25(3), 1203–1211.
- Dussias, P. E., & Sagarra, N. (2007). The effect of exposure on syntactic parsing in Spanish-English bilinguals. *Bilingualism*, 10(1), 101.
- Einhäuser, W. (2017). The pupil as marker of cognitive processes. In Q. Zhao (ed.) *Computational and cognitive neuroscience of vision* (pp. 141–169). Berlin: Springer.
- Fernández, E. M. (2003). *Bilingual sentence processing: Relative clause attachment in English and Spanish* (Vol. 29). Amsterdam: John Benjamins.
- Gabay, S., Pertzov, Y., & Henik, A. (2011). Orienting of attention, pupil size, and the norepinephrine system. *Attention, Perception, & Psychophysics*, 73(1), 123–129.
- Gagl, B., Hawelka, S., & Hutzler, F. (2011). Systematic influence of gaze position on pupil size measurement: Analysis and correction. *Behavior Research Methods*, 43(4), 1171–1181.
- Goldwater, B. C. (1972). Psychological significance of pupillary movements. *Psychological Bulletin*, 77(5), 340.
- Grosjean, F., & Li, P. (2013). *The psycholinguistics of bilingualism*. New

- York, NY: John Wiley & Sons.
- Guasch, M., Ferre, P., & Haro, J. (2017). Pupil dilation is sensitive to the cognate status of words: Further evidence for non-selectivity in bilingual lexical access. *Bilingualism*, 20(1), 49.
- Häuser, K. I., Demberg, V., & Kray, J. (2019). Effects of aging and dual-task demands on the comprehension of less expected sentence continuations: Evidence from pupillometry. *Frontiers in Psychology*, 10, 709.
- Hofmeister, P., Casasanto, L. S., & Sag, I. A. (2013). Islands in the grammar? Standards of evidence. In J. Sprouse & N. Hornstein (Eds.), *Experimental syntax and island effects*. Cambridge: Cambridge University Press.
- Hofmeister, P., & Sag, I. A. (2010). Cognitive constraints and island effects. *Language*, 86(2), 366.
- Koelewijn, T., de Kluiver, H., Shinn-Cunningham, B. G., Zekveld, A. A., & Kramer, S. E. (2015). The pupil response reveals increased listening effort when it is difficult to focus attention. *Hearing Research*, 323, 81–90.
- Koelewijn, T., Shinn-Cunningham, B. G., Zekveld, A. A., & Kramer, S. E. (2014). The pupil response is sensitive to divided attention during speech processing. *Hearing Research*, 312, 114–120.
- Krejtz, K., Duchowski, A. T., Niedzielska, A., Biele, C., & Krejtz, I. (2018). Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PloS One*, 13(9), e0203629.
- Kroll, J. F., & Bialystok, E. (2013). Understanding the consequences of bilingualism for language processing and cognition. *Journal of Cognitive Psychology*, 25(5), 497–514.
- Kuchinsky, S. E., Ahlstrom, J. B., Vaden Jr, K. I., Cute, S. L., Humes, L. E., Dubno, J. R., & Eckert, M. A. (2013). Pupil size varies with word listening and response selection difficulty in older adults with hearing loss. *Psychophysiology*, 50(1), 23–34.
- Kupisch, T., & Rothman, J. (2018). Terminology matters! Why difference is not incompleteness and how early child bilinguals are heritage speakers. *International Journal of Bilingualism*, 22(5), 564–582.
- Montrul, S. (2008). *Incomplete Acquisition in Bilingualism: Re-examining the Age Factor*. Amsterdam: John Benjamins.
- Montrul, S. (2016). *The acquisition of Heritage Languages*. Cambridge: Cambridge University Press.
- Otheguy, R., & Zentella, A. C. (2011). *Spanish in New York: Language contact, dialectal leveling, and structural continuity*. Oxford: Oxford University Press.

- Perlmutter, D. M. (1968). *Deep and surface structure constraints in syntax*. Ph.D. Dissertation. Cambridge, MA: Massachusetts Institute of Technology.
- Phillips, C. (2013). On the nature of island constraints II: Language learning and innateness. In J. Sprouse & N. Hornstein (Eds.), *Experimental syntax and island effects* (pp. 132–157). Cambridge: Cambridge University Press.
- Polinsky, M. (2018). *Heritage Languages and their Speakers* by Maria Polinsky. Cambridge: Cambridge University Press.
- Polinsky, M., & Scontras, G. (2020). Understanding heritage languages. *Bilingualism: Language and Cognition*, 23(1), 4–20.
- Ross, J. R. (1967). *Constraints on Variables in Syntax*. Unpublished Ph.D. dissertation. Cambridge, MA: Massachusetts Institute of Technology.
- Rothman, J. (2007). Heritage speaker competence differences, language change, and input type: Inflected infinitives in Heritage Brazilian Portuguese. *International Journal of Bilingualism*, 11(4), 359–389.
- Rothman, J. (2009). Understanding the nature and outcomes of early bilingualism: Romance languages as heritage languages. *International Journal of Bilingualism*, 13(2), 155–163.
- Samuels, E. R., & Szabadi, E. (2008). Functional neuroanatomy of the noradrenergic locus coeruleus: Its roles in the regulation of arousal and autonomic function part I: principles of functional organisation. *Current Neuropharmacology*, 6(3), 235–253.
- Schmid, M. (2011). *Language Attrition*. Cambridge University Press.
- Schmidtke, J. (2014). Second language experience modulates word retrieval effort in bilinguals: Evidence from pupillometry. *Frontiers in Psychology*, 5, 137.
- Schmidtke, J. (2018). Pupillometry in linguistic research: An introduction and review for second language researchers. *Studies in Second Language Acquisition*, 40(3), 529–549.
- Serratrice, L. (2013). Cross-linguistic influence in bilingual development: Determinants and mechanisms. *Linguistic Approaches to Bilingualism*, 3(1), 3–25.
- Serratrice, L., Sorace, A., Filiaci, F., & Baldo, M. (2009). Bilingual children's sensitivity to specificity and genericity: Evidence from metalinguistic awareness. *Bilingualism: Language and Cognition*, 12(2), 239–257.
- Sirois, S., & Brisson, J. (2014). Pupillometry. *Wiley Interdisciplinary Reviews: Cognitive Science*, 5(6), 679–692.
- Sobin, N. (1987). The variable status of Comp-trace phenomena. *Natural Language & Linguistic Theory*, 5(1), 33–60.

- Sobin, N. (2002). The Comp-trace effect, the adverb effect and minimal CP. *Journal of Linguistics*, 38(3), 527–560.
- Sorace, A., Serratrice, L., Filiaci, F., & Baldo, M. (2009). Discourse conditions on subject pronoun realization: Testing the linguistic intuitions of older bilingual children. *Lingua*, 119(3), 460–477.
- Sprouse, J., & Hornstein, N. (2013). Experimental syntax and island effects: Toward a comprehensive theory of islands. In J. Sprouse & N. Hornstein (Eds.), *Experimental syntax and island effects* (pp. 1–17). Cambridge: Cambridge University Press.
- Szabolcsi, A. (2006). Strong vs. Weak islands. In M. Everaert & H. van Riemsdijk (Eds.), *The Blackwell companion to syntax, volume 1* (pp. 479–531). New York, NY: Blackwell Publishing.
- Torrego, E. (1984). On Inversion in Spanish and Some of Its Effects. *Linguistic Inquiry*, 15(1), 103–129.
- Tsimpli, I., Sorace, A., Heycock, C., & Filiaci, F. (2004). First Language Attrition and Syntactic Subjects: A Study of Greek and Italian Near-Native Speakers of English. *International Journal of Bilingualism*, 8(3), 257–277.
- van Rij, J., Hendriks, P., van Rijn, H., Baayen, R. H., & Wood, S. N. (2019). Analyzing the time course of pupillometric data. *Trends in Hearing*, 23, 2331216519832483.
- van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, H. (2020). *Itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs. R package version 2.4*.
<https://cran.r-project.org/web/packages/itsadug/index.html>
- Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1), 3–36.

Appendix 1 – Language background questionnaire

Screening questions:

- 1) Where were you born?
- 2) When did you arrive in the US?
- 3) How old are you?
- 4) What is the highest level of formal schooling you have completed?
- 5) Do you live alone, or with a partner?
 - a) Where was your partner born?
 - b) When did your partner arrive in the US?
 - c) How old is your partner?
 - d) What is the highest level of formal schooling your partner has completed?
- 6) What language do you speak at home?
- 7) Who were your primary caregiver(s) from birth to age 10?
 - a) What country were your primary caregiver(s) born in?
 - b) How old were your primary caregiver(s) when they arrived in the US?
 - c) What year did your primary caregiver(s) arrive in the US?
- 8) What language did you speak with your primary caregiver(s) from birth to age 10?
- 9) How well do you understand Spanish:
 - 1 = little to nothing of what I hear
 - 2 = some of what I hear
 - 3 = about half of what I hear
 - 4 = most of what I hear
 - 5 = everything I hear

Administered before first experimental session:

- 10) What do you consider to be your native language?
- 11) Please list all the languages that you speak (DO NOT include languages that you can read but do not speak):

For *level*:

- 1 = I have limited knowledge of the language
- 2 = I have some ability to use the language
- 3 = I have good ability to use the language*
- 4 = I am a fluent speaker/user of the language
- 5 = I am a native speaker/user of the language

**If you select “3 = I have good ability to use the language”, please write “YES” if you are able to give an opinion and defend it in that language.*

Language ____, level 1 2 3 4 5 , when did you start learning? _ years old

Language ____, level 1 2 3 4 5 , when did you start learning? _ years old

Language ____, level 1 2 3 4 5 , when did you start learning? _ years old

Language ____, level 1 2 3 4 5 , when did you start learning? _ years old

- 12) For each of the above-listed languages, please describe where and how you learned it:

Example: Language French: was taught in school from 1st-5th grade

Language Guarani: picked it up from friends

Language _____:

Language _____:

Language _____:

Language _____:

- 13) What was the first language you learned?

- 14) What languages were spoken in your house growing up?

- 15) Which of the languages from (14.) were used most often?

- 16) Who spoke each of the languages in (14.) to each other in your house growing up?

Example: Language Spanish: everyone spoke Spanish to each other

Language Nahuatl: grandparents spoke Nahuatl to each other and no one else

Language _____:

Language _____:

Language _____:

Language _____:

17) Please complete the following table:

Age	What country did you live in?	What was the primary language spoken in your local community?	Did you attend school?	What was the language of instruction?
5-6				
6-7				
7-8				
8-9				
9-10				
10-11				
11-12				
12-13				
13-14				
14-15				
15-16				
16-17				
17-18				

Administered after second experimental session:

- 18) Participant's sex: _____
 19) Participant's profession in U.S.: _____
 20) Participant's social class (choose one):
 working ____ middle ____ upper ____
 21) Which languages do you read/write? At what level? When did you start?

For level: 1 = I have limited reading/writing ability in the language
 2 = I have some ability to read/write in the language
 3 = I have good ability to read/write in the language*
 4 = I am a fluent reader/writer of the language
 5 = I am a native reader/writer of the language

**If you select "3 = I have good ability to read/write in the language", please write "YES" if you are able to defend an opinion in writing in that language.*

Language ____, level 1 2 3 4 5 , when did you start learning? _ years old
 Language ____, level 1 2 3 4 5 , when did you start learning? _ years old

Language ____, level 1 2 3 4 5 , when did you start learning? _ years old

Language ____, level 1 2 3 4 5 , when did you start learning? _ years old

22) Which language(s) do you use to speak with **your**:

- a. **father**
English / Spanish / both / N/A
- b. **mother**
English / Spanish / both / N/A
- c. **sisters/brothers**
English / Spanish / both / N/A
- d. **children** (older)
English / Spanish / both / N/A
- e. **children** (younger)
English / Spanish / both / N/A
- f. **friends**
English / Spanish / both / N/A
- g. **boss**
English / Spanish / both / N/A
- h. **co-workers**
English / Spanish / both / N/A
- i. **classmates**
English / Spanish / both / N/A
- j. **significant other**
English / Spanish / both / N/A

23) How much Spanish do you use in/at:

- k. **home**
mostly / little / none / N/A
- l. **school**
mostly / little / none / N/A
- m. **work**
mostly / little / none / N/A
- n. **social activities**
mostly / little / none / N/A
- o. **reading**
mostly / little / none / N/A
- p. **listening to the radio/music**
mostly / little / none / N/A
- q. **watching TV**
mostly / little / none / N/A

24) In a typical day, how much do you interact with the following (please give answers as relative percentages, e.g. Spanish-speakers 75%, English-speakers 25%):

Spanish-speakers _____ English-speakers _____

25) Where do the interactions in (24.) occur?

Spanish-speakers: _____

English-speakers: _____

26) How often do you travel to Spanish-speaking countries?

27) How long is/are your typical stay(s) in (26.) _____

28) Do you plan on living in a Spanish-speaking country?

29) Which language do you prefer (choose one):

English ____ Spanish ____ no preference ____

30) What is/are the reason(s) for your preference in (29.)?

Appendix 2 – Pupillometry models by structural condition

Generalized additive mixed models reporting parametric coefficients and smooths terms

Comp-Trace Condition with AR1

Parametric	Estimate	SE	<i>t</i>-value	<i>p</i>-value
Intercept	-0.018	0.006	-3.205	< .01
HS, Grammatical	0.003	0.007	0.371	.71
LB, Ungrammatical	-0.004	0.003	-1.455	.15
HS, Grammatical	0.010	0.007	1.489	.14
Session (Second)	-0.004	0.003	-1.648	.10

Smooth Terms	edf	Ref.df	<i>F</i>-value	<i>p</i>-value
LB, Grammatical	2.117	2.619	1.765	.12
HS, Grammatical	1.803	2.183	0.273	.64
LB Ungrammatical	1.465	1.734	1.088	.21
HS, Ungrammatical	2.519	3.130	2.246	.08
X- and Y-gaze Position	38.142	38.955	132.965	< .001
Random effect for Subjects	137.134	389.000	1.815	< .001
Random effect for Items	43.433	199.000	0.701	< .001

Adjusted $R^2 = 0.0315$, Deviance explained = 3.52%

Wh- island Condition with AR1

Parametric	Estimate	SE	<i>t</i>-value	<i>p</i>-value
Intercept	-0.001	0.004	-0.157	.88
HS, Grammatical	0.001	0.005	0.217	.83
LB, Ungrammatical	-0.004	0.002	-1.870	.06
HS, Grammatical	-0.007	0.005	-1.516	.13
Session (Second)	-0.005	0.002	-2.618	< .01

Smooth Terms	edf	Ref.df	<i>F</i>-value	<i>p</i>-value
LB, Grammatical	2.827	3.505	2.270	0.08
HS, Grammatical	2.880	3.576	3.374	0.01
LB Ungrammatical	3.057	3.798	3.770	< .01
HS, Ungrammatical	2.370	2.915	1.878	0.14
X- and Y-gaze Position	37.269	38.818	95.233	< .001
Random effect for Subjects	175.259	460.000	1.786	< .001
Random effect for Items	39.233	300.000	0.343	< .001

Adjusted $R^2 = 0.0276$, Deviance explained = 2.58%

Complex NP Condition with AR1

Parametric	Estimate	SE	t-value	p-value
Intercept	-0.006	0.004	-1.383	.17
HS, Grammatical	-0.002	0.005	-0.004	.99
LB, Ungrammatical	-0.002	0.002	-0.949	.34
HS, Grammatical	-0.007	0.005	-1.394	.16
Session (Second)	0.003	0.002	1.243	.21

Smooth Terms	edf	Ref.df	F-value	p-value
LB, Grammatical	2.372	2.939	2.042	.09
HS, Grammatical	4.165	5.193	3.207	< .01
LB Ungrammatical	3.126	3.905	1.711	.16
HS, Ungrammatical	2.467	3.049	1.688	.17
X- and Y-gaze Position	35.532	38.278	116.116	< .001
Random effect for Subjects	143.696	439.000	1.463	< .001
Random effect for Items	60.318	299.000	0.822	< .001

Adjusted $R^2 = 0.0315$, Deviance explained = 3.52%

Temporal Adverbial Condition with AR1

Parametric	Estimate	SE	t-value	p-value
Intercept	-0.026	0.004	-6.021	< .001
HS, Grammatical	0.017	0.005	3.231	< .01
LB, Ungrammatical	0.015	0.002	6.510	< .001
HS, Grammatical	0.015	0.005	2.832	< .01
Session (Second)	-0.007	0.002	-3.636	< .001

Smooth Terms	edf	Ref.df	F-value	p-value
LB, Grammatical	1.007	1.012	1.366	.22
HS, Grammatical	3.548	4.396	4.205	< .01
LB Ungrammatical	2.482	3.098	2.362	.06
HS, Ungrammatical	3.259	4.029	4.096	< .01
X- and Y-gaze Position	38.485	38.982	200.734	< .001
Random effect for Subjects	199.363	520.000	2.077	< .001
Random effect for Items	82.082	300.000	1.008	< .001

Adjusted $R^2 = 0.0281$, Deviance explained = 3.14%

Relative Clause Condition with AR1

Parametric	Estimate	SE	<i>t</i>-value	<i>p</i>-value
Intercept	-0.030	0.004	-7.113	< .001
HS, Grammatical	0.012	0.005	2.320	.02
LB, Ungrammatical	0.014	0.002	8.246	< .001
HS, Grammatical	0.018	0.005	3.594	< .001
Session (Second)	-0.004	0.002	-2.546	.01

Smooth Terms	edf	Ref.df	<i>F</i>-value	<i>p</i>-value
LB, Grammatical	2.520	3.097	2.047	.11
HS, Grammatical	3.288	4.071	5.727	< .001
LB Ungrammatical	2.891	3.575	5.425	< .001
HS, Ungrammatical	3.192	3.943	7.918	< .001
X- and Y-gaze Position	38.203	38.954	366.502	< .001
Random effect for Subjects	194.463	519.000	2.629	< .001
Random effect for Items	133.474	449.000	1.275	< .001

Adjusted $R^2 = 0.0274$, Deviance explained = 3.11%

THE TIMING OF INTERFERENCE EFFECTS DURING NATIVE AND NON-NATIVE PRONOUN RESOLUTION

CECILIA PUEBLA,¹ CLARE PATTERSON
& CLAUDIA FELSER

Abstract

The real-time interpretation of pronouns is affected by both structure-sensitive and non-grammatical constraints, and models of anaphor resolution make different claims about the relative weighting and/or timing of structure-sensitive constraints such as Condition B of the binding theory. Here we used an interference paradigm to examine whether, and when during processing, Condition B may be violated. Using eye-movement monitoring during reading, we investigated and compared the resolution of object pronouns in native and non-native speakers of German. The gender of both a binding-accessible and a structurally inaccessible competitor antecedent was manipulated in a gender-mismatch paradigm, with the inaccessible antecedent's saliency increased through elaboration. Our results show that both participant groups experienced interference from an inaccessible antecedent but at different points during processing. While native German speakers experienced interference early on even in the presence of a matching accessible antecedent, the non-native group showed later interference effects, and only when no gender-matching accessible antecedent was available. Based on these and previous findings, we argue that the cues to anaphor resolution are weighted differently for native and non-native populations, a possibility that models of anaphor resolution should be able to accommodate.

¹ Corresponding author: Cecilia Puebla, Potsdam Research of Multilingualism, University of Potsdam, Campus Golm, Haus 2, Karl-Liebknecht-Str. 24-25, 14476 Potsdam, e-mail: cecilia.puebla.antunes@uni-potsdam.de

1. Introduction

Successful language comprehension requires the ability to construct complex meaning representations from the linguistic input quickly and efficiently. During reading or listening, grammatical structure, semantic and discourse-level information must be encoded and integrated. At the same time, referential dependencies need to be resolved by establishing links between anaphoric expressions such as pronouns and their antecedents, which may be separated by variable amounts of linguistic material. The linking process is affected by a range of information sources, including phrase structure-sensitive constraints and non-structural (e.g., discourse-level) cues, that combine and interact during processing in guiding the comprehender towards a suitable antecedent. Anaphor resolution is a highly automatized process, and it is central to both native (L1) and non-native (L2) sentence and discourse comprehension. However, it is not yet fully clear what happens under the hood during real-time processing, and a number of empirical questions remain open regarding the availability and relative weighting of information sources and constraints over time, and how these are utilised during the resolution of referential dependencies during L2 as compared to L1 comprehension. Here we report the results of an eye-tracking-during-reading experiment aimed at investigating the presence and timing of interference effects during the resolution of German object pronouns whose interpretation is constrained by binding Condition B.

Binding Condition B is a structure-sensitive constraint on intrasentential referential dependencies which is traditionally subsumed under the Binding Theory (Chomsky, 1981). Binding relations are syntactically mediated and typically involve c-command, a relationship between phrase constituents based on structural hierarchy and dominance (Reinhart, 1983). The interpretation of personal pronouns is restricted by Condition B, which states that a pronoun cannot take a c-commanding antecedent from its local domain (i.e., a pronoun cannot be locally bound). In (1), the embedded subject *Nick* is a grammatically inappropriate (or 'inaccessible') antecedent for the object pronoun *him* because both items are coarguments of the same predicate. The pronoun can refer to the matrix subject *Martin* instead, or to a non-commanding antecedent outside the current sentence.

- (1) Martin_i completely forgot that Nick_j had invited him_{i/*j} for dinner.

Within Reuland and colleagues' Primitives of Binding (PoB) framework (e.g., Grodzinsky & Reinhart, 1993; Reuland, 2001), Condition B is

considered a semantic constraint preventing personal pronouns from being assigned an unwanted reflexive interpretation. Note that Condition B is a negative constraint in that it only prohibits local coreference but it does not point to the correct antecedent. This implies that Condition B alone does not provide enough information for the unambiguous identification of a referent, and other factors such as semantic information or the relative discourse-prominence of antecedent candidates need to be considered in order to fully understand the computation of pronoun-antecedent relations.

1.1 Condition B in real-time processing

Theories of sentence processing commonly assume that the identification of an antecedent engages a memory search and retrieval mechanism; pronoun resolution therefore constitutes an ideal context to test psychological theories of memory encoding and retrieval, and to explore the memory systems that underlie language comprehension (Lewis & Vasishth, 2005; Lewis, Vasishth, & Van Dyke, 2006). The question of how the memory search and retrieval mechanism makes use of the different information sources relevant for referential dependency formation has been widely discussed in the L1 literature, especially with respect to the processing role of the binding conditions (see Nicol & Swinney, 2003, and Sturt, 2013, for reviews). The real-time application of Condition B has been investigated using a variety of psycholinguistic techniques, but results vary with respect to the question of whether (and how) binding-inaccessible candidates affect antecedent search and retrieval (compare e.g., Badecker & Straub, 2002; Chow, Lewis, & Phillips, 2014; Clifton, Kennison, & Albrecht, 1997; Kennison, 2003; Kim, Montrul, & Yoon, 2015; Nicol & Swinney, 1989; Patterson, Trompelt, & Felser, 2014). These mixed results have led to various psycholinguistic approaches to pronoun resolution that make different predictions regarding two central aspects: the timing of constraint application (i.e., when during processing a given constraint is applied) and its interaction with other information sources (e.g., whether pronoun resolution is susceptible to interference from binding-inaccessible antecedents). The 'Binding as Initial Filter' (BAIF; Nicol & Swinney, 1989) hypothesis assumes that Condition B applies as an initial filter on candidates. This means that only candidates compatible with binding conditions are considered, and binding-inaccessible candidates never influence processing. A less strict version of this account, as put forward by Sturt (2003) for reflexives, maintains binding as an initial filter but allows for interference from inaccessible antecedents later during processing under certain conditions, for example if the latter are especially prominent in the

discourse. Later approaches have been more explicitly tied to more general models of memory retrieval during language processing. Multiple-constraint models (e.g., Badecker & Straub, 2002) and cue-based retrieval models (e.g., Lewis & Vasishth, 2005; Lewis, Vasishth, & Van Dyke, 2006; Parker, Shvartsman, & Van Dyke, 2017) do not assume an *a priori* primacy of structure-sensitive constraints and argue instead for the use of all relevant information sources in parallel during the linking process. The reactivation of antecedent candidates exploits a set of retrieval cues compatible with the anaphor, which can include gender cues but also cues that distinguish binding-accessible from inaccessible antecedents.² Crucially, a candidate antecedent's degree of activation depends on how well it matches the set of retrieval cues, so that an item that partially matches the cues can be temporarily mis-retrieved, resulting in interference. For example, an item with matching gender or number cues may be retrieved even if the locality cue does not match the pronoun's requirements, thus leading to interference from items that should be excluded by binding constraints.

Chow et al. (2014) put forward a multiple-constraint model in the style of cue-based retrieval that does predict the exclusion of binding-inaccessible antecedents via simultaneous application of agreement and structural cues. Under this view, inaccessible antecedents influence processing only as an emergency repair process when no matching accessible antecedent is available.

1.2 Constraints on anaphoric binding in L2 processing

Current models of anaphor resolution have been proposed on the basis of L1 data. However, a growing body of experimental research investigating bilingual sentence processing has revealed differences in the strategies that L1 and L2 speakers use for resolving certain types of intrasentential dependencies, such as long filler-gap dependencies (e.g., Felser & Roberts, 2007; Marinis, Roberts, Felser, & Clahsen, 2005) and backwards-looking anaphora, which require re-accessing previously constructed sentence or discourse representations (see Felser, 2016, and Cunnings, 2017, for reviews). Findings from reading-time studies investigating the processing of reflexives (e.g., *himself*, *herself*) suggest that L2 learners may be more susceptible than L1 speakers to interference from binding-inaccessible antecedents that are rendered highly discourse-prominent (e.g., Felser &

² For different proposals on how structural notions such as c-command or locality can be integrated in cue-based retrieval models compare e.g., Alcocer and Phillips (2012); Kush, Lidz and Phillips (2015). For a discussion, see e.g., Kush (2013).

Cunnings, 2012; Felser, Sato, & Bertenshaw, 2009). While there is not yet a clear consensus on how to explain the observed L1/L2 differences, claims such as the Shallow Structure Hypothesis (SSH; Clahsen & Felser, 2006, 2018) and Cunnings' (2017) Memory-interference Hypothesis have argued for a different weighting of information sources across populations. According to the SSH, even highly proficient non-native comprehenders may show a reduced sensitivity to morphosyntactic and phrase-structure information during processing compared to native comprehenders, and as a consequence may rely more strongly on non-grammatical cues to comprehension. A more recent alternative proposal put forward by Cunnings (2017) in the context of cue-based retrieval models of language comprehension maintains that even highly advanced L2 learners may implement the syntactic and discourse-level cues that guide memory retrieval differently from L1 speakers. According to Cunnings' account, differences between L1 and L2 processing arise from L2 comprehenders being more vulnerable than L1 comprehenders to memory interference when retrieving information previously constructed during online parsing.

Non-native speakers have sometimes been reported to have difficulty applying binding Condition B when interpreting pronouns (e.g., Slabakova, White, & Guzzo, 2017), but the L2 processing of pronouns in Condition B configurations has been investigated only in a few studies. In an eye-tracking-during-reading experiment, Patterson et al. (2014, Exp.2) presented English-native and proficient German-speaking L2 learners of English with sentences such as (2a-c) that contained the object pronoun *him* and two c-commanding antecedent candidates. The authors sought to examine whether (and if so, when) during processing local antecedents, which are ruled out by Condition B, were considered for dependency formation.

- (2)
 - a. DOUBLE MATCH CONDITION
John remembered that Mark had taught him a new song on the guitar.
 - b. LOCAL MISMATCH CONDITION
John remembered that Jane had taught him a new song on the guitar.
 - c. NON-LOCAL MISMATCH CONDITION
Jane remembered that John had taught him a new song on the guitar.

The authors used a gender-mismatch paradigm, with either *John* or *Mark* being replaced by a female name so as to create selective gender-

mismatches with the pronoun. Both participant groups showed significantly longer reading times at and following the pronoun in (2c), where the non-local (i.e., binding-accessible) antecedent did not match in gender with the pronoun, compared to the other two conditions (2a,b). No gender effects of the local antecedent were found. This reading-time pattern indicates that upon processing the pronoun, only structurally appropriate antecedents were retrieved, which is consistent with the application of Condition B. L1/L2 processing differences were however observed in a second eye-tracking experiment (Exp.3) with ambiguous 'short-distance' pronouns in configurations exempt from Condition B, as in *Barry saw Gavin place a gun near him on the ground with great care*. Even though local coreference is allowed here, the L2 (unlike the L1) readers were exclusively drawn towards the non-local antecedent despite showing awareness of the pronoun's referential ambiguity in an offline task (Exp.1). Given that in both of the two sentence types tested, the accessible antecedent was highly discourse-prominent by virtue of being the matrix subject, the authors raised the question of whether the accessible-antecedent effects observed in the L2 data did indeed reflect the application of Condition B or a general preference for discourse-prominent antecedents (compare e.g., Felser & Cunnings, 2012, for eye-movement evidence of L2 readers adopting a discourse-based strategy in the processing of reflexives).

L1/L2 differences in the interpretation of English object pronouns in Condition B environments have also been reported by Kim et al. (2015). In an eye-tracking-during-listening experiment, the authors observed that Korean L2 learners of English, but not English native speakers, attempted coreference with a binding-inaccessible antecedent in contexts such as *Look at Goofy. Have Mickey touch him*, where the local, inaccessible noun phrase (NP) *Mickey* was a subject/action character and, as such, possibly more salient than the Condition B accessible antecedent *Goofy*, which was a non-commanding object and a sentence-external NP.

In summary, there is evidence that L2 comprehenders are drawn more strongly than L1 comprehenders towards salient antecedents even if these are binding-inaccessible. Models of anaphor resolution need to be able to accommodate any such L1/L2 processing differences and allow for the possibility of cross-population or inter-individual differences in cue weightings.

Exactly what might render an antecedent sufficiently salient for it to be (mis-)retrieved during L2 processing is still not fully clear, however. One possibility is that salience is defined at the discourse-representational level,

with (e.g.) topical antecedents being more likely to be retrieved than others in L2 processing. Another factor that has not yet been systematically examined in L2 comprehension is the role of a potential antecedent's syntactic or semantic 'heaviness'. The results from previous L1 studies suggest that a higher degree of elaboration for referents correlates with increased retrieval facilitation. More descriptive and structured expressions can result in stronger memory representations and higher activation levels, and as a consequence, they may be more likely accessed and retrieved from memory during online comprehension than less 'heavy' antecedents (e.g., Hofmeister, 2011; Troyer, Hofmeister, & Kutas, 2016).

For the present experiment, we used the same methodology (i.e., eye-tracking during reading) and a similar design as in Patterson et al. (2014). Our aim was to answer the following research questions:

- a) Are structurally inaccessible but 'heavy' competitor antecedents considered for referential dependency formation during L1 and/or L2 processing?
- b) If yes, at which point during processing do interference effects occur?

2. Method

The current experiment consisted of an eye-movement monitoring task designed to investigate the presence and timing of interference effects during L1 and L2 pronoun resolution.

2.1 Participants

Thirty-one native speakers of German (21 female) and 30 Russian-native learners of L2 German (26 female) took part in the experiment. The L1 group had a mean age of 27.6 years (SD 5.5, age range 20-42 years). The L2 speakers' mean age was 27.2 years (SD 4.3, age range 19-37 years). Participants were mostly students at the University of Potsdam who were recruited via e-mail invitations, through online advertisements and by flyers distributed in the Potsdam/Berlin area. None of the L2 speakers had started learning German before the age of five. The L2 group's mean age of acquisition (AoA) of German was 14 years (SD 6.1, range 5-29 years), and their German proficiency was at or above B2 level as measured by the paper-and-pencil version of the Goethe placement test (courtesy of the Goethe Institute, 2011). The group's mean Goethe test score was 25.7 points

out of a total of 30, corresponding to a C1 level of the Common European Framework of Reference for Languages (range 18-30, B2-C2, SD 3.2).

All participants had normal or corrected-to-normal vision and confirmed that they could read the sentences presented on the screen without effort. None of the participants reported any diagnosed language disorder at the time of data collection. After completing the experiment each participant either received course credit or a small fee as a reward for their contribution.

2.2 Materials

The stimulus materials consisted of 24 sets of short texts introduced by a context sentence and followed by a critical sentence that contained the third-person object pronoun *ihn* ('him') and two c-commanding potential antecedent NPs. The gender features of both NPs were manipulated in a gender-mismatch paradigm (Sturt, 2003). This manipulation yielded four experimental conditions, as shown in (3a-d) below.

- (3) Context: *Die Firma hatte gute Kontakte im Ausland.*
the company had good contacts abroad
'The company had good contacts abroad.'

a. DOUBLE MATCH

Florian glaubte, dass der Kollege aus Frankreich
F. thought that the colleague_{masc} from France
ihn schon bald vorstellen würde.
him very soon introduce would

b. NP1 MATCH & NP2 MISMATCH

Florian glaubte, dass die Kollegin aus Frankreich
F. thought that the colleague_{fem} from France
ihn schon bald vorstellen würde.
him very soon introduce would

c. NP1 MISMATCH & NP2 MATCH

Marlena glaubte, dass der Kollege aus Frankreich
M. thought that the colleague_{masc} from France
ihn schon bald vorstellen würde.
him very soon introduce would

d. DOUBLE MISMATCH

Marlena glaubte, dass die Kollegin aus Frankreich
 M. thought that the colleague_{fem} from France

ihn schon bald vorstellen würde.
 him very soon introduce would

‘{Florian/Marlena} thought that the colleague_{masc/fem}
 from France would introduce him very soon.’

In all critical sentences, the matrix subject NP1 (*Florian/Marlena*) was a proper name, whereas the embedded subject NP2 (*der Kollege/die Kollegin* ‘the colleague_{masc/fem}’) was a definite role name modified by a prepositional phrase (PP) of similar length across items (e.g., *aus Frankreich* ‘from France’).³ Condition B prohibits coreference between the pronoun and NP2, as both are coarguments of the same predicate. NP2, but not NP1, was thus a structurally inaccessible antecedent for the pronoun in terms of binding Condition B. Note that adding a modifying PP to the local antecedent increased its syntactic and semantic ‘heaviness’ relative to NP1, the binding-accessible antecedent.

The materials were arranged in a 2x2 Latin-square design with the factors NP1 (match, mismatch) and NP2 (match, mismatch), and distributed across four presentation lists. The experimental items were mixed and randomized with 48 filler-items, resulting in a total of 72 items (plus five practice trials) per list.

The fillers were created to distract participants from the research aims of the experiment. Four fillers were short texts of comparable length and structurally similar to the experimental items but containing a third-person feminine object pronoun *sie* (‘her’) and two feminine potential antecedents. The remaining fillers displayed a variety of syntactic structures and were generally shorter and/or structurally less complex than our critical items. Most of them contained feminine singular or plural personal pronouns and

³ All proper names used in this experiment were frequent and gender-typical German names. They had been previously tested for gender prototypicality in a rating task administered as a web-based questionnaire to a group of 20 Russian-native L2 learners of German. Participants were asked to rate the gender of 80 proper names in a 5-point scale (1 = *typical male name*, 5 = *typical female name*). Only names that obtained extreme rating values were selected for the current experiment.

other assorted pronouns (e.g., reflexives, possessives, indefinites), and seven filler items contained no pronoun.

Half of the items were followed by a *yes/no* comprehension question. In order to encourage full processing of the pronouns, 18 questions asked for the referent of a pronoun. Eight of these questions followed an experimental item, and asked about the relationship between one of the potential antecedents and the critical pronoun.

2.3 Predictions

Readers' sensitivity to an antecedent's gender features, reflected in a difference in reading times between the gender match and the mismatch conditions, was taken as indication that the antecedent was considered for dependency formation. In our experiment, NP1 was a binding-accessible antecedent and NP2 an inaccessible one. The processing hypotheses considered above make the following predictions.

According to the BAIF hypothesis (Nicol & Swinney, 1989), only structurally accessible NPs are included in the initial candidate set; binding-inaccessible antecedents are excluded. This hypothesis predicts gender effects for NP1 but not NP2, which will be reflected in a reading-time difference between NP1 match conditions (3a,b) and NP1 mismatch conditions (3c,d).

The presence of interference effects at any point would be incompatible with the BAIF hypothesis but would be consistent with other approaches, depending on the timing of the effect. According to the defeasible filter hypothesis (Sturt, 2003) and Chow et al.'s (2014) 'repair strategy' account, inaccessible antecedents are not included in the initial candidate set but could affect processing with some delay. Evidence for this account would be reflected in NP1 gender effects followed by NP2 gender effects, which could manifest as either main effects or as an NP1*NP2 interaction. Main effects of NP2 gender would be reflected in a difference in reading times between NP2 match conditions (3a,c) and NP2 mismatch conditions (3b,d). An NP1*NP2 interaction would reflect a difference between conditions (3c) and (3d), with NP2 considered only if NP1 mismatches the pronoun's gender.

Finally, according to Badecker and Straub's (2002) multiple-constraint approach and most implementations of cue-based retrieval models, Condition B does not act as an early filter, so feature-matching but structurally inaccessible antecedents may sometimes be retrieved. If this is

the case, we expect to find competition between the two potential antecedents when they both match in gender with the pronoun immediately or shortly after it is encountered and before main effects of NP1 gender start emerging. An early competition (or similarity-based interference) pattern would become observable in our experiment as an NP1*NP2 interaction, with the double match condition (3a) yielding longer reading times relative to (3b) during initial processing.

As for L2 processing, we may expect a greater likelihood of finding interference effects than for the L1 group, in line with Cummings' (2017) Memory-interference Hypothesis. This prediction is based on the assumption that elaborated antecedent NPs are more salient in memory than, for example, simple names such as those used in Patterson et al. (2014). Conversely, if L2 comprehenders are guided by information-structure or discourse-level information more strongly than L1 comprehenders (e.g., Clahsen & Felser, 2018), then our L2 participants may be drawn towards NP1 instead, which functions as the discourse topic.

2.4 Procedure

All participants were tested individually in a quiet laboratory in Berlin or in Potsdam.⁴ Participants first filled in an online demographic questionnaire. At the beginning of the testing session, each participant was given instructions in German regarding the procedure and asked to sign an informed consent form. For the eye-tracking task, participants sat at a distance of 80 cm from the monitor, with their chin resting on a cushioned chin rest and their forehead slightly pressed against a bar. Eye movements were recorded using a desktop eye-tracker (EyeLink 1000, SR Research). The camera was located below the screen approximately 50 cm from the participant's eyes. Reading was binocular but only the right eye was tracked. After successful calibration and validation, a screen with detailed instructions was presented. The experiment began with five practice items included to familiarise the participant with the task. Half of the trials (including two of the practice items) were followed by a comprehension question. All items were presented one by one in black text (Courier New font, 18 pt) on a white background. The experimental items occupied two lines of text, with the critical pronoun falling always approximately in the middle of the second line. This was done to reduce the skipping rate in the pronoun region by avoiding inaccurate fixations travelling from the

⁴ The present experiment was approved by the ethics committee of the University of Potsdam (application 37/2011).

previous line. Participants were asked to read each text for comprehension, silently at their normal reading pace, and to answer the questions by pressing the correct button on a gamepad. At the beginning of each trial, a black fixation dot controlled for drift. In order to make the text appear, participants were instructed to look at the dot while pressing the corresponding button of the gamepad. Recalibration in the course of the experiment was necessary for most of the readers in order to guarantee good tracking. The presentation of items was divided into two blocks of 36 items each, allowing participants to take a short break in between blocks. Depending on the length of the break, the eye-tracking task could last up to 30 min.

After the eye-tracking task, L2 participants completed a short computer-based vocabulary checklist containing 48 critical items from the experimental sentences. Participants were asked to mark any words or phrases they were not familiar with. A testing session lasted between 40-60 minutes in total.

2.5 Data analysis

The eye-tracking-during-reading method allows for the examination of multiple reading-time measures for different regions of text (i.e., interest regions) that potentially indicate different points in the time-course of processing (e.g., Conklin & Pellicer-Sánchez, 2016; Rayner, 1998). Our interest regions were the *pronoun region*, which contained the target pronoun plus the following word (e.g., *ihn schon* in [3a-d]), and the *spillover region*, containing the next word after the pronoun region (e.g., *bald* in [3a-d]). Reading times in these regions are assumed to be affected by how easy it is for readers to establish a link between the pronoun and an antecedent. For each region, we will report five continuous measures for which the unit of measurement is time in milliseconds, and three binomial *yes/no* measures. *First fixation duration* refers to the length of readers' initial fixation within a given region; *first-pass reading times* is the summed duration of all fixations in a region until it is first exited to the left or right; *regression-path duration* is the sum of all fixations on a region until it is exited to the right (this measure may also include regressive eye movements); *rereading time* corresponds to the total duration of all fixations within a region after it was first exited to the left or right; *total reading times* is the summed duration of all fixations inside a given region. *Regressions out* indicates whether a regression was made from a given interest region into earlier ones before a later region is fixated; *regressions in* indicates whether the region received any regression from later parts of the text; and *rereading probability* is the likelihood of refixating a given region.

Although a mapping between individual measures and processing stages is not necessarily straightforward, we can broadly distinguish between ‘early’ measures such as first fixation duration or first-pass reading times, which indicate initial referential decisions, and ‘late’ measures such as rereading time, which can be informative of reanalysis processes.

Prior to statistical analysis the complete eye-movement dataset was checked and experimental trials were individually cleaned. Datasets of two L2 speakers were excluded due to track loss. Fixations with vertical drift were manually adjusted (L1 group: 3.9% of total fixations; L2 group: 7.14% of total fixations). Individual fixations that clearly did not belong to the reading pattern (e.g., travelling fixations from the previous line, fixations falling between two lines of text, participant blinks) were manually deleted (L1 group: 0.06% of experimental data; L2 group: 0.08% of experimental data). Fixations shorter than 80 ms within one degree of visual angle of a neighbouring fixation were merged together. All other short (< 80 ms) and extremely long (> 800 ms) fixations were automatically removed and excluded from analysis.

Regions skipped in first-pass were excluded from analysis. Skipping rates for the pronoun and spillover regions were 3.5 and 24.7% in the L1 group and 1.9 and 14% in the L2 group, respectively. Individual trials for which L2 speakers reported unknown vocabulary were removed from statistical analysis. One participant who had unknown vocabulary in every trial had to be additionally excluded from analysis. Removed trials on the basis of unknown vocabulary accounted for a total of 1.57% of the L2 experimental data. Trials for which participants incorrectly answered a comprehension question were included in statistical analyses.

The data were analysed in R (R Core Team, 2016) using mixed modelling with the package *lme4* (Bates, Mächler, Bolker, & Walker, 2015). Linear-mixed effects models were fitted for the continuous measures and mixed-effects logistic regressions were calculated on the binomial variables. Given that the distribution of fixation durations is often right-skewed, a non-linear transformation (log) was applied to each reading-time continuous measure to satisfy the assumptions of normality (Vasishth & Nicenboim, 2016). To confirm that the transformation was appropriate in each case, the Box-Cox procedure (Box & Cox, 1964) was used. The statistical analyses were performed on the log-transformed data.

For the purpose of finding out whether the reading patterns across conditions differed significantly between both groups of participants, we

carried out a between-groups analysis. The models contained the sum-coded fixed two-level factors NP1 (match, mismatch), NP2 (match, mismatch), Group (L1, L2), and their interactions. Full random-effects models included random intercepts and slopes for both subjects and items as long as convergence was achieved. In case of non-convergence, the random structure was gradually simplified by dropping elements one by one as suggested in Barr, Levy, Scheepers and Tily (2013) until convergence was achieved.

3. Results

Participants of both groups answered with high accuracy to the end-of-trial comprehension questions. The overall response accuracy was 93.5% (SD 0.05, range 80-100%) for the L1 group and 91% (SD 0.05, range 77-99%) for the L2 group. These results demonstrate that both participant groups paid attention to the task and understood the texts. In the following, we first report the eye-tracking results from an omnibus between-group analysis, followed by the results for each participant group separately.

3.1 Between-group analysis

The between-group analysis revealed significant ($p < .05$) main effects of the factor Group across all continuous measures in both regions of interest as well as in rereading probability at the pronoun region. Additionally, a number of main effects of NP1 and NP2 gender were found in several measures at the pronoun and spillover regions. Along with the reported main effects, we observed two- and three-way interactions with Group. There was an early, yet marginal NP2*Group interaction in first fixation duration at the spillover region ($t = 1.73, p < .09$) and a late and significant NP1*Group interaction in rereading probability at the pronoun region ($z = -2.15, p < .03$). Three-way NP1*NP2*Group interactions at the spillover region were found significant for total reading times ($t = -2.19, p < .03$) and marginally significant for rereading probability ($z = -1.85, p < .06$).

No further main effects nor interactions arose from the between-groups analysis. However, as the above results suggest different L1/L2 reading patterns, the groups were subsequently analysed separately.⁵ In what

⁵ For the individual per-group analyses, trial index was included as a covariate in all models to account for possible habituation effects as the experiment progressed.

follows, the outcomes of each of the group's analysis will be presented, beginning with the L1 group.

3.2 L1 results

An overview of the L1 group's reading times (continuous measures) and proportions (binomial data) for our regions of interest is provided in Table 1 below. The results of the statistical analysis are shown in Table 2.

Pronoun region. At the pronoun region we observed significant main effects of NP2 gender in first-pass reading times, with the NP2 mismatch conditions (3b,d) yielding longer reading times than the NP2 match conditions (3a,c). Although statistically marginal, this pattern was replicated in regression-path duration. A marginal main effect of NP2 gender was also found for rereading time, but in this measure longer reading times were registered for the NP2 match conditions (3a,c) compared to the NP2 mismatch conditions (3b,d).

Effects of NP1 gender started arising later during processing. Significant main effects of NP1 gender were observed in total reading times, regressions in and rereading probability at the region containing the pronoun. All three measures showed a gender-mismatch effect, i.e., higher reading times for mismatching than for matching non-local antecedents.

Spillover region. The analysis of the spillover region again revealed early main effects of NP2 gender, already visible in first fixation duration. However, unlike the early NP2 effects seen at the pronoun region, this was a gender-match effect, with longer reading times for matching than for mismatching local antecedents. As was also the case for the pronoun region, NP2 gender effects were followed by effects of NP1 gender later during processing. Significant main effects of NP1 gender (mismatch effects) were seen for regression-path duration and regressions out.

There was a marginal NP1*NP2 interaction for regressions out. However, subsequent pairwise comparisons revealed no significant difference between conditions (NP2 gender within NP1 match conditions: $z = -1.09$, $p = .28$; NP2 gender within NP1 mismatch conditions: $z = 1.46$, $p = .14$).

Table 1. L1 Group: Means in milliseconds and proportions (standard deviations in parenthesis) for eight eye-movement measures at the pronoun and spillover regions.

		Pronoun region	Spillover region
First fixation duration	(a) DOUBLE MATCH	221 (71)	207 (69)
	(b) NP2 MISMATCH	221 (68)	195 (53)
	(c) NP1 MISMATCH	222 (72)	214 (70)
	(d) DOUBLE MISMATCH	233 (77)	204 (57)
First-pass reading times	(a) DOUBLE MATCH	327 (140)	221 (94)
	(b) NP2 MISMATCH	338 (162)	213 (85)
	(c) NP1 MISMATCH	334 (188)	230 (97)
	(d) DOUBLE MISMATCH	374 (234)	221 (78)
Regression-path duration	(a) DOUBLE MATCH	331 (143)	288 (324)
	(b) NP2 MISMATCH	344 (164)	263 (247)
	(c) NP1 MISMATCH	350 (213)	283 (180)
	(d) DOUBLE MISMATCH	388 (256)	311 (256)
Rereading time	(a) DOUBLE MATCH	393 (198)	264 (133)
	(b) NP2 MISMATCH	346 (205)	290 (139)
	(c) NP1 MISMATCH	515 (428)	303 (225)
	(d) DOUBLE MISMATCH	441 (372)	268 (156)
Total reading times	(a) DOUBLE MATCH	451 (274)	296 (173)
	(b) NP2 MISMATCH	450 (259)	306 (184)
	(c) NP1 MISMATCH	558 (418)	336 (235)
	(d) DOUBLE MISMATCH	582 (403)	310 (181)
Regressions out	(a) DOUBLE MATCH	.01 (0.11)	.13 (0.34)
	(b) NP2 MISMATCH	.02 (0.13)	.09 (0.29)
	(c) NP1 MISMATCH	.03 (0.18)	.15 (0.36)
	(d) DOUBLE MISMATCH	.02 (0.13)	.22 (0.41)
Regressions in	(a) DOUBLE MATCH	.22 (0.41)	.13 (0.33)
	(b) NP2 MISMATCH	.24 (0.43)	.18 (0.39)
	(c) NP1 MISMATCH	.31 (0.46)	.15 (0.36)
	(d) DOUBLE MISMATCH	.33 (0.47)	.13 (0.34)
Rereading probability	(a) DOUBLE MATCH	.31 (0.47)	.28 (0.45)
	(b) NP2 MISMATCH	.33 (0.47)	.32 (0.47)
	(c) NP1 MISMATCH	.43 (0.50)	.35 (0.48)
	(d) DOUBLE MISMATCH	.47 (0.50)	.33 (0.47)

Table 2. L1 group: Summary of statistical analyses for eight eye-movement measures at the pronoun and spillover regions.

	Pronoun region				Spillover region			
	Est.	SE	<i>t</i> (<i>z</i>)	<i>p</i>	Est.	SE	<i>t</i> (<i>z</i>)	<i>p</i>
First fixation duration								
ME NP1	-0.016	0.011	-1.443	.164	-0.020	0.013	-1.492	.147
ME NP2	-0.011	0.011	-1.040	.304	0.023	0.013	2.106	.036 *
NP1*NP2	0.011	0.010	1.071	.285	0.003	0.011	0.303	.762
First-pass reading times								
ME NP1	-0.012	0.014	-0.808	.428	-0.022	0.017	-1.254	.220
ME NP2	-0.031	0.015	-2.131	.047 *	0.016	0.015	1.072	.296
NP1*NP2	0.020	0.013	1.511	.131	0.000	0.014	0.002	.998
Regression-path duration								
ME NP1	-0.021	0.015	-1.353	.193	-0.059	0.021	-2.830	.006 *
ME NP2	-0.030	0.017	-1.752	.093 †	0.005	0.023	0.211	.835
NP1*NP2	0.015	0.017	0.871	.392	0.023	0.020	1.131	.263
Rereading time								
ME NP1	-0.022	0.049	-0.442	.662	0.012	0.041	0.293	.772
ME NP2	0.070	0.037	1.883	.069 †	0.005	0.042	0.111	.913
NP1*NP2	0.002	0.034	0.067	.946	-0.045	0.039	-1.165	.252
Total reading times								
ME NP1	-0.095	0.025	-3.798	.001 **	-0.037	0.028	-1.341	.192
ME NP2	-0.028	0.024	-1.187	.248	0.007	0.021	0.332	.742
NP1*NP2	0.013	0.024	0.560	.580	-0.024	0.021	-1.158	.255
Regressions out								
ME NP1	-0.274	0.294	-0.932	.351	-0.325	0.127	-2.551	.011 *
ME NP2	0.080	0.294	0.271	.787	-0.017	0.126	-0.134	.893
NP1*NP2	-0.263	0.294	-0.893	.372	0.216	0.126	1.723	.085 †
Regressions in								
ME NP1	-0.282	0.092	-3.077	.002 *	0.056	0.123	0.458	.647
ME NP2	-0.089	0.092	-0.973	.330	-0.074	0.123	-0.600	.549
NP1*NP2	-0.004	0.091	-0.046	.963	-0.154	0.123	-1.249	.212
Rereading probability								
ME NP1	-0.382	0.089	-4.288	.000 **	-0.104	0.097	-1.074	.283
ME NP2	-0.079	0.088	-0.898	.369	-0.039	0.097	-0.403	.687
NP1*NP2	0.038	0.087	0.440	.660	-0.085	0.097	-0.870	.384

ME = main effects; Est. = estimate; SE = standard error; † $p < .10$, * $p < .05$, ** $p \leq .001$

3.3 L2 results

Table 3 provides an overview on the L2 group's mean reading times (continuous measures) and proportions (binomial data). The results of the statistical analysis are shown in Table 4.

Pronoun region. At the region containing the pronoun, significant main effects of NP1 gender emerged in three measures: rereading time, total reading times and regressions in. In all three measures we observed a gender-mismatch effect.

Spillover region. At the spillover region we found significant main effects of NP1 gender (mismatch effects) in regression-path duration and in regressions out. Additionally, there were two NP1*NP2 gender interactions: a significant interaction for total reading times, and a marginally significant interaction for rereading probability. For the first interaction in total reading times, follow-up pairwise comparisons revealed a marginally significant difference within the NP1-mismatch conditions, where the double mismatch condition was read more slowly than the NP2-match condition ($t = 1.98, p < .06$). The same pattern was observed for the second interaction in rereading probability ($z = 1.83, p < .07$).

3.4 Summary

Our eye-tracking data revealed different patterns of results for the two participant groups. For our L1 group, early main effects of NP2 gender emerged at the pronoun and spillover regions, along with main effects of NP1 gender later during processing. For the L2 group, we observed main effects of NP1 gender at the pronoun and spillover regions in later measures. No early effects, or main effects of NP2 gender, were found for this group. Additionally for the L2 group, a significant NP1*NP2 interaction was found in total reading times at the spillover region. This reflected a marginally significant difference within the NP1-mismatch conditions, where the presence of a gender-matching inaccessible antecedent facilitated processing compared to a gender-mismatching one, but only in cases where no gender-matching accessible antecedent could be found within the sentence.

Table 3. L2 Group: Means in milliseconds and proportions (standard deviations in parenthesis) for eight eye-movement measures at the pronoun and spillover regions.

		Pronoun region	Spillover region
First fixation duration	(a) DOUBLE MATCH	245 (81)	224 (66)
	(b) NP2 MISMATCH	237 (74)	217 (58)
	(c) NP1 MISMATCH	238 (75)	218 (76)
	(d) DOUBLE MISMATCH	248 (99)	226 (74)
First-pass reading times	(a) DOUBLE MATCH	431 (205)	253 (106)
	(b) NP2 MISMATCH	449 (220)	251 (111)
	(c) NP1 MISMATCH	434 (210)	258 (129)
	(d) DOUBLE MISMATCH	471 (275)	266 (121)
Regression-path duration	(a) DOUBLE MATCH	466 (261)	296 (198)
	(b) NP2 MISMATCH	497 (416)	296 (187)
	(c) NP1 MISMATCH	463 (297)	350 (312)
	(d) DOUBLE MISMATCH	498 (329)	362 (289)
Rereading time	(a) DOUBLE MATCH	581 (482)	415 (419)
	(b) NP2 MISMATCH	570 (526)	388 (342)
	(c) NP1 MISMATCH	819 (726)	443 (410)
	(d) DOUBLE MISMATCH	737 (607)	417 (326)
Total reading times	(a) DOUBLE MATCH	743 (509)	413 (378)
	(b) NP2 MISMATCH	739 (504)	385 (298)
	(c) NP1 MISMATCH	903 (716)	420 (357)
	(d) DOUBLE MISMATCH	885 (656)	449 (328)
Regressions out	(a) DOUBLE MATCH	.04 (0.20)	.09 (0.29)
	(b) NP2 MISMATCH	.03 (0.17)	.11 (0.32)
	(c) NP1 MISMATCH	.02 (0.15)	.16 (0.37)
	(d) DOUBLE MISMATCH	.04 (0.18)	.16 (0.37)
Regressions in	(a) DOUBLE MATCH	.22 (0.42)	.16 (0.36)
	(b) NP2 MISMATCH	.25 (0.43)	.15 (0.36)
	(c) NP1 MISMATCH	.36 (0.48)	.17 (0.37)
	(d) DOUBLE MISMATCH	.29 (0.46)	.17 (0.38)
Rereading probability	(a) DOUBLE MATCH	.54 (0.50)	.39 (0.49)
	(b) NP2 MISMATCH	.51 (0.50)	.35 (0.48)
	(c) NP1 MISMATCH	.57 (0.50)	.37 (0.48)
	(d) DOUBLE MISMATCH	.56 (0.50)	.44 (0.50)

Table 4. L2 group: Summary of statistical analyses for eight eye-movement measures at the pronoun and spillover regions.

	Pronoun region				Spillover region			
	Est.	SE	<i>t</i> (<i>z</i>)	<i>p</i>	Est.	SE	<i>t</i> (<i>z</i>)	<i>p</i>
First fixation duration								
ME NP1	0.003	0.011	0.260	.795	0.000	0.014	-0.032	.975
ME NP2	0.001	0.014	0.059	.954	-0.005	0.013	-0.410	.686
NP1*NP2	0.015	0.010	1.414	.158	0.018	0.012	1.564	.120
First-pass reading times								
ME NP1	-0.008	0.016	-0.513	.612	-0.014	0.017	-0.800	.430
ME NP2	-0.020	0.016	-1.284	.209	-0.010	0.015	-0.677	.503
NP1*NP2	0.005	0.015	0.353	.726	0.017	0.015	1.101	.283
Regression-path duration								
ME NP1	0.001	0.018	0.083	.935	-0.055	0.021	-2.657	.016 *
ME NP2	-0.017	0.017	-1.041	.305	-0.015	0.020	-0.753	.463
NP1*NP2	0.008	0.015	0.503	.615	0.021	0.019	1.113	.267
Rereading time								
ME NP1	-0.145	0.033	-4.366	.000 **	-0.061	0.051	-1.208	.240
ME NP2	0.054	0.033	1.651	.101	-0.013	0.043	-0.304	.762
NP1*NP2	-0.035	0.032	-1.088	.277	0.022	0.040	0.543	.588
Total reading times								
ME NP1	-0.071	0.020	-3.473	.002 *	-0.042	0.030	-1.407	.172
ME NP2	0.010	0.018	0.569	.570	-0.020	0.022	-0.930	.362
NP1*NP2	-0.002	0.019	-0.108	.915	0.039	0.019	2.022	.047 *
Regressions out								
ME NP1	0.094	0.240	0.393	.694	-0.264	0.128	-2.053	.040 *
ME NP2	-0.060	0.244	-0.245	.806	-0.079	0.128	-0.617	.537
NP1*NP2	0.220	0.240	0.916	.359	-0.035	0.128	-0.270	.787
Regressions in								
ME NP1	-0.264	0.096	-2.760	.006 *	-0.053	0.116	-0.454	.650
ME NP2	0.043	0.095	0.455	.649	-0.004	0.116	-0.038	.970
NP1*NP2	-0.148	0.095	-1.549	.121	0.039	0.116	0.334	.739
Rereading probability								
ME NP1	-0.141	0.095	-1.489	.136	-0.098	0.098	-0.996	.319
ME NP2	0.083	0.095	0.875	.382	-0.053	0.098	-0.539	.590
NP1*NP2	0.029	0.094	0.306	.760	0.181	0.098	1.846	.065 †

ME = main effects; Est. = estimate; SE = standard error; † $p < .10$, * $p < .05$, ** $p \leq .001$

4. Discussion

The purpose of the current study was to compare the time-course of L1 and L2 processing of German object pronouns in sentences that contained two c-commanding potential antecedents: a discourse-prominent, binding-accessible non-local antecedent, and an elaborated but structurally inaccessible local competitor antecedent. More specifically, we sought to examine whether, and at what point during processing, an antecedent ruled out by binding Condition B might be considered for dependency formation. The inaccessible antecedent was rendered more salient (in comparison to Patterson et al.'s, 2014, stimulus materials) through modification.

The eye-movement data revealed different patterns of interference across our two groups. While our native readers considered the inaccessible antecedent initially even in the presence of a gender-matching appropriate antecedent, the L2 speakers only considered the inaccessible antecedent later on during processing, and only in the absence of a gender-matching accessible antecedent. Our findings are discussed in more detail below.

4.1 Interference from binding-inaccessible antecedents

Effects of the local antecedent (NP2) were seen in the eye-movement records of both participant groups, indicating that Condition B does not prevent binding-inaccessible antecedents from being considered. However, the inaccessible antecedent affected our native and non-native comprehenders differently. For the L1 group we observed an NP2-mismatch effect during participants' initial reading of the pronoun region, that is, at a point in time *before* the accessible antecedent was considered. This effect suggests that our native speakers tried to establish local coreference when the pronoun was first encountered, and even in the presence of a matching accessible antecedent. The change in the direction of the effect (from gender-mismatch to gender-match) observed immediately after, in first fixation duration at the region following the pronoun, may index the processing costs of re-evaluating the dependency when Condition B was applied, with a disruption caused by the presence of a gender-matching NP in a structurally inaccessible position.

Our L2 participants, by contrast, did not show any early NP2 effects. Instead, the two antecedents' gender features interacted during later processing stages, reflecting the fact that NP2 effects were restricted to cases where no gender-matching accessible antecedent was available. In those cases, the presence of an inaccessible antecedent that matched in

gender with the pronoun facilitated processing compared to when neither antecedent matched. Although the pattern of the interaction reached only marginal significance, it implies that our L2 readers attempted coreference with the local antecedent only as a last-resort strategy, as was previously argued by Chow et al. (2014) for L1 English speakers. Notwithstanding the observed between-group differences, the presence of NP2 gender effects in our experiment suggests that binding-inaccessible antecedents were evaluated for dependency formation by both L1 and L2 speakers.

Our findings contrast with the general lack of interference effects reported by Patterson et al. (2014, Exp.2). However, this discrepancy can be explained by two crucial differences in the experimental design and materials. First, the interference pattern seen in our L2 data could only possibly be found thanks to the double mismatch condition, which Patterson et al. did not include in their experimental design. Second, note that the local antecedents in Patterson et al.'s stimulus sentences were plain proper names (e.g., *Mark*, *Jane*), whereas in our study, inaccessible antecedents were definite descriptions modified by a PP (e.g., *der Kollege aus Frankreich* 'the colleague from France'). A higher degree of elaboration for referents has been found to correlate with increased retrieval facilitation, at least in native speakers. The comparatively greater 'heaviness' of the local antecedent in our stimulus materials might have made them more salient in memory, and thus more likely to be retrieved compared to the simple names used by Patterson et al. (2014). In short, antecedent elaboration may increase the likelihood of interference effects becoming measurable.

4.2 Effects of the accessible antecedent

The NP1 gender effects observed for both participant groups show that L1 and L2 readers were sensitive to the gender features of the binding-accessible antecedent during processing. We found NP1-mismatch effects across different measures and regions, indicating higher processing costs when the accessible antecedent mismatched the pronoun's gender. Moreover, both participant groups patterned alike with respect to the timing of NP1 effects, which were not visible in early eye-movement measures but only in later measures (compare also Patterson et al., 2014, Exp.2, for English object pronouns). This indicates that pronouns were not linked to the accessible antecedent immediately.

Finding effects of the accessible antecedent to be restricted to later eye-movement measures contrasts with what has been reported for the resolution of reflexives. In L1 comprehension, reflexives tend to be linked to

accessible antecedents immediately during processing (e.g., Sturt, 2003), although this is not necessarily also the case in L2 comprehension (Felser & Cunnings, 2012). A comparatively delayed resolution of personal pronouns is consistent with the divergent nature of the referential dependencies involved. Personal pronouns differ from standard reflexives in that they do not need to be bound and can be linked to an antecedent via discourse-based coreference assignment instead. Unlike binding Condition A, which helps identify the correct antecedent of reflexives, Condition B is a negative constraint on pronoun interpretation that merely blocks local coreference. Within the PoB framework (e.g., Grodzinsky & Reinhart, 1993; Reuland, 2001), the application of Condition B requires the computation and comparison of two alternative semantic representations, and then rejecting the one that yields an interpretation of the pronoun that is indistinguishable from that of a reflexive. Building antecedent-pronoun relations is therefore argued to be computationally more costly (and thus may take longer than) reflexive resolution, which is assumed to involve purely syntactic operations.

From a PoB point of view, the eye-movement pattern we saw in our L1 participants might in fact be expected: Following the initial consideration of the local antecedent and its evaluation in the light of binding Condition B, our L1 speakers then identified the non-local antecedent as a more suitable one. It is possible that processing reflexes of the earlier evaluation stage were missed in some previous studies either because less time-course sensitive experimental techniques were used or due to differences in experimental designs or materials. If this interpretation of our L1 results is along the right lines, then this highlights the need for refining current models of anaphor resolution to incorporate different antecedent search procedures for different types of anaphoric expression.

4.3 L1/L2 processing differences

Given claims to the effect that L2 comprehenders should be more prone to memory interference than L1 comprehenders (Cunnings, 2017), our finding that only the L1 group –but not the L2 group– were initially drawn towards binding-inaccessible local antecedents may seem surprising. Like Patterson et al. (2014), but unlike Kim et al. (2015), we found no evidence of our L2 comprehenders trying to link an object pronoun to the local antecedent in violation of Condition B when the pronoun was encountered. Are we to conclude from this that non-native speakers apply Condition B more efficiently than native speakers? It is indeed conceivable that Condition B was applied in the same way by both participant groups, but that for some

reason no processing reflex of the initial evaluation of NP2 was observed in the L2 data. Recall that significant effects of the accessible antecedent emerged relatively late in both groups, suggesting that neither group established a link to NP1 immediately after the pronoun was encountered.

As an alternative explanation for our results, it is possible then that our non-native readers' focus on the non-local antecedent reflected a discourse-based interpretation strategy, with pronouns being preferentially linked to first-mentioned or topical antecedents. As noted above, pronouns can be resolved without establishing a binding relationship (via discourse-based coreference assignment), and in our experiment, as well as in Patterson et al. (2014, Exp.2), binding-accessible antecedents were also sentence or discourse topics. Our L2 speakers seemed to consider the inaccessible antecedent only as a last resort, in the absence of a matching accessible one. Rather than reflecting the application of binding Condition B, this relatively late, selective interference effect is likely to reflect the attempt to find a potential referent in the absence of a matching discourse-prominent antecedent. This interpretation is more in line with previous L2 processing findings, which have shown L2 comprehenders being drawn towards discourse-prominent antecedents during the processing of both reflexives (Felser & Cunnings, 2012) and pronouns exempt from Condition B (Patterson et al., 2014, Exp.3). Note also that Kim et al.'s (2015) Korean L2 speakers of English allowed discourse-based coreference between object pronouns and local NPs, thus ignoring the structural restriction imposed by Condition B, in contexts where the local NP was comparatively more salient than the accessible antecedent. While the authors attributed this pattern to L2 learners' difficulties in integrating discourse and grammatical information, they conceded that a general preference of learners for pragmatically/contextually salient antecedents could also explain their results.

Contrary to our predictions, providing a comparatively 'heavy' inaccessible antecedent did not lead to the L2 speakers considering this antecedent at any point prior to the consideration of the accessible antecedent. The observation that our L2 speakers were more resistant to similarity-based interference than the L1 group indicates that for non-native comprehenders, elaboration was not sufficient to increase the likelihood of the inaccessible antecedent being initially (mis)-retrieved instead of the accessible antecedent. This, in turn, suggests that discourse factors such as first-mention or topic-hood are more heavily weighted during L2 pronoun processing than elaboration. Note that elaboration is not a retrieval cue but it plays a role during structure-building and encoding processes that may influence memory retrieval. It is possible, alternatively, that the semantic

and syntactic complexity of the inaccessible antecedent could not be encoded properly and/or quickly enough during L2 comprehension, so that its mental representation was not active or strong enough for it to be retrieved initially.

4.4 Theoretical implications

The presence of NP2 gender effects in our participant groups suggests that binding-inaccessible antecedents may be considered during both native and non-native pronoun resolution. These effects are inconsistent with the BAIF hypothesis (Nicol & Swinney, 1989), according to which inaccessible antecedents are not considered for dependency formation at all. Our results are compatible with a multiple-constraint approach (Badecker & Straub, 2002) and with cue-based retrieval models that allow for similarity-based interference. Using comparatively 'heavy' inaccessible antecedents may have facilitated the detection of interference effects which were not detected in previous studies, at least in our L1 group.

As we discussed above, what look like interference effects from a processing point of view could in fact be reflexes of an early evaluation stage as has been proposed within the theoretical linguistic literature. If applying Condition B involves a comparison of alternative semantic representations as has been claimed by proponents of the PoB framework (e.g., Grodzinsky & Reinhart, 1993; Reuland, 2001), then the early NP2 effects we saw in our L1 data may have been a processing reflex of this semantic evaluation. Future research may want to compare the time-course of processing reflexives vs. pronouns more directly to test the PoB model's claims regarding differences in the computational effort required to resolve different types of anaphoric element.

The observed L1/L2 differences regarding the nature and timing of interference effects have implications for hypotheses about L2 processing. Our findings add further support to the claim that real-time L2 anaphor resolution relies on a discourse-based strategy, where discourse-level cues such as topic-hood are strongly weighted (e.g., Clahsen & Felser, 2018; Cummings, 2017, Felser, 2016), leading L2 comprehenders to try and resolve anaphor-antecedent dependencies in favour of a discourse-prominent antecedent initially. Our findings also emphasize the need for processing models to be able to accommodate cross-population differences in the relative weighting and/or timing of information sources during anaphor resolution. Cummings' (2017) memory interference account may be a step towards this goal, but as the current results show, interference effects are

not necessarily more likely to occur in L2 than in L1 processing. There are also some challenges in implementing notions such as discourse prominence or topic-hood in a cue-based framework (Jacob, Lago & Patterson, 2017).

Acknowledgements

This research has been supported by an Alexander-von-Humboldt Professorship awarded to Harald Clahsen (Potsdam Research Institute for Multilingualism) and by the German Research Foundation (DFG) through Grant No. FE 1138/1-1 awarded to Claudia Felser.

We thank our colleagues of the Potsdam Research Institute for Multilingualism for helpful discussion, and the audiences at the ACLC-Workshop “Doing Experiments with Theoretical Linguistics” and at the Experimental Psycholinguistics Conference 2019 for their insightful comments. We also thank Janna Drummer and Thea Villinger for their help with the preparation of materials, participant recruitment and data collection. We are grateful to an anonymous reviewer for her/his valuable comments and helpful suggestions for improvement on a previous version of this paper.

References

- Alcocer, P., & Phillips, C. (2012). Using relational syntactic constraints in content-addressable memory architectures for sentence processing. Unpublished manuscript. www.colinphillips.net/wpcontent/uploads/2014/08/alcocer_phillips2012_v2.pdf
- Badecker, W., & Straub, K. (2002). The processing role of structural constraints on the interpretation of pronouns and anaphors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(4), 748 – 769.
- Barr, D., Levy, R., Scheepers, C., & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3).
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1 – 48.
- Box, G., & Cox, D., (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B (Methodological)*, 26(2), 211 – 252.

- Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht: Foris.
- Chow, W-Y., Lewis, S., & Phillips, C. (2014). Immediate sensitivity to structural constraints in pronoun resolution. *Frontiers in Psychology*, 5(630). doi: 10.3389/fpsyg.2014.00630
- Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics*, 27(1), 3 – 42.
- Clahsen, H., & Felser, C. (2018). Some notes on the Shallow Structure Hypothesis. *Studies in Second Language Acquisition*, 40(3), 693 – 706.
- Clifton, C., Kennison, S., & Albrecht, J. (1997). Reading the words *her*, *his*, *him*: Implications for parsing principles based on frequency and on structure. *Journal of Memory and Language*, 36, 276 – 292.
- Conklin, K., & Pellicer-Sánchez, A. (2016). Using eye-tracking in applied linguistics and second language research. *Second Language Research*, 32(3), 453 – 467.
- Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, 28(3), 369 – 382.
- Cunnings, I. (2017). Parsing and working memory in bilingual sentence processing. *Bilingualism: Language and Cognition*, 20(4), 659 – 678.
- Felser, C. (2016). Binding and coreference in non-native language processing. In A. Holler & K. Suckow (Eds.), *Empirical perspectives on anaphora resolution* (pp. 241 – 266). Berlin: De Gruyter.
- Felser, C. (2019). Structure-sensitive constraints in non-native sentence processing. *Journal of the European Second Language Association*, 3(1), 12–22.
- Felser, C., & Cunnings, I. (2012). Processing reflexives in English as a second language: The role of structural and discourse-level constraints. *Applied Psycholinguistics*, 33(3), 571 – 603.
- Felser, C., & Roberts L. (2007). Processing wh-dependencies in a second language: A cross-modal priming study. *Second Language Research*, 23(1), 9 – 36.
- Felser, C., Sato, M., & Bertenshaw, N. (2009). The on-line application of binding Principle A in English as a second language. *Bilingualism: Language and Cognition*, 12(4), 485 – 502.
- Grodzinsky, Y., & Reinhart, T. (1993). The innateness of binding and of coreference. *Linguistic Inquiry*, 24, 69 – 101.
- Hofmeister, P. (2011). Representational complexity and memory retrieval in language comprehension, *Language and Cognitive Processes*, 26(3), 376 – 405.

- Jacob, G., Lago, S., & Patterson, C. (2017). L2 processing and memory retrieval: some empirical and conceptual challenges. *Bilingualism: Language and Cognition*, 20(4), 691 – 693.
- Kennison, S. (2003). Comprehending the pronouns her, him, and his: Implications for theories of referential processing. *Journal of Memory and Language*, 49(3), 335 – 352.
- Kim, E., Montrul, S., & Yoon, J. (2015). The on-line processing of binding principles in second language acquisition: Evidence from eye tracking. *Applied Psycholinguistics*, 36(6), 1317 – 1374.
- Kush, D. (2013). *Respecting relations: Memory access and antecedent retrieval in incremental sentence processing*. PhD. dissertation. College Park: University of Maryland.
- Kush, D., Lidz, J., & Phillips, C. (2015). Configuration-sensitive retrieval: Resisting interference in processing bound variable pronouns. *Journal of Memory & Language*, 82, 18 – 40.
- Lewis, R., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3), 375 – 419.
- Lewis, R., Vasishth, S., & Van Dyke, J. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, 10(10), 447 – 454.
- Marinis, T., Roberts, L., Felser, C., & Clahsen, H. (2005). Gaps in second language sentence processing. *Studies in Second Language Acquisition*, 27(1), 53 – 78.
- Nicol, J., & Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *Journal of Psycholinguistic Research*, 18(1), 5 – 19.
- Nicol, J., & Swinney, D. (2003). The psycholinguistics of anaphora. In A. Barss (Ed.), *Anaphora: A reference guide* (pp. 72 – 104). Oxford: Blackwell.
- Parker, D., Shvartsman, M., & Van Dyke, J. A. (2017). The cue-based retrieval theory of sentence comprehension: New findings and new challenges. In L. Escobar, V. Torrens, & T. Parodi (Eds.), *Language Processing and Disorders* (pp. 121 – 144). Newcastle: Cambridge Scholars Publishing.
- Patterson, C., Trompelt, H., & Felser, C. (2014). The online application of binding condition B in native and non-native pronoun resolution. *Frontiers in Psychology*, 5(147). doi: 10.3389/fpsyg.2014.00147
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3), 372 – 422.

- R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org.
- Reinhart, T. (1983). *Anaphora and semantic interpretation*. London: Croom Helm.
- Reuland, E. (2001). Primitives of binding. *Linguistic Inquiry*, 32(3), 439 – 492.
- Slabakova, R., White, L., & Guzzo, N. B. (2017). Pronoun interpretation in the second language: effects of computational complexity. *Frontiers in Psychology*, 8(1236). doi: 10.3389/fpsyg.2017.01236
- Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, 48(3), 542 – 562.
- Sturt, P. (2013). Syntactic constraints on referential processing. In R. P. G. van Gompel (Ed.), *Sentence processing* (pp. 136 – 159). Hove: Psychology Press.
- Troyer, M., Hofmeister, P., & Kutas, M. (2016). Elaboration over a discourse facilitates retrieval in sentence processing. *Frontiers in Psychology*, 7(374). doi: 10.3389/fpsyg.2016.00374
- Vasishth, S., & Nicenboim, B. (2016). Statistical methods for linguistic research: Foundational ideas – Part I. *Language and Linguistics Compass*, 10, 349 – 369.

EFFECTS OF LANGUAGE DOMINANCE ON L1 RELATIVE CLAUSE PROCESSING

LEEANN STOVER,¹ MICHAEL C. STERN,
CASS LOWRY & GITA MARTOHARDJONO

Abstract

The present study investigated the effects of language dominance during bilingual comprehension of relative clauses. We asked whether language dominance, operationalized as a continuous variable, modulates whether/how Spanish-English bilinguals exhibit a relative clause subject-object processing asymmetry in their first-learned language, Spanish. Highly proficient bilinguals with varying ages of arrival to the US completed a language dominance questionnaire and a visual world eye-tracking experiment with auditorily presented Spanish relative clauses. Results revealed that higher Spanish dominance led to a larger processing asymmetry while listening to both the relative clause and the matrix predicate. However, rather than a facilitatory effect in subject relatives, this asymmetry was primarily driven by a late negative effect in object relative constructions. To account for these results, we propose that increased dominance in the first-learned language leads to more active online syntactic structure building, leading to a higher integration cost when an expected parse fails.

1. Introduction

In recent literature on bilingualism, attempts have been made to formalize a theoretical notion of language dominance as a super-construct which subsumes a bilingual's proficiency, use, and exposure to both languages (e.g., Montrul, 2015; Treffers-Daller, 2015). Although there is still no

¹ Corresponding author: LeeAnn Stover, Department of Linguistics, CUNY Graduate Center, 365 Fifth Ave., New York, NY 10016; email: lstover@gradcenter.cuny.edu

consensus on a precise definition in the literature (see Meisel, 2007; Silva-Corvalán & Treffers-Daller, 2015), most recent research agrees that language dominance is (1) relative to both of a bilingual's languages, (2) multidimensional, and (3) continuous in nature. However, despite the general acknowledgement that language dominance is a continuous construct, studies of bilingual processing tend to treat it as a categorical variable (e.g., Fernández, 2003; Puig-Mayenco et al., 2018). While some researchers have studied the effects of language dominance as a continuous measure on linguistic *knowledge* (e.g., Bedore et al., 2012; Dunn & Tree, 2009; Gollan et al., 2012), the link between continuous dominance and language *processing* is much less understood (Robinson & Blumenfeld, 2018).

The present study explores the relationship between L1 processing and language dominance. Focusing on the subject-object asymmetry in relative clause processing, we conducted an eye-tracking experiment using the Visual World Paradigm on Spanish-English bilinguals who share Spanish as the L1 but live and work in an L2-dominant society. To look at a continuum of language dominance and a variety of L1 experience, we included participants with varying ages of arrival to the anglophone US. At one end of the spectrum, Spanish heritage speakers learned Spanish as their home language but grew up in an English-speaking society and are generally dominant in their L2. On the other end of the spectrum lie first-generation late arrivals who immigrated to the anglophone US after adolescence. These speakers grew up and were educated in a Spanish-speaking society and are generally L1-dominant. However, as language dominance is gradient and dynamic (e.g., Birdsong, 2015), we do not presume categorical dominance of any individuals in this study. Rather, we use a relative dominance index to determine each participant's language dominance on the English-Spanish spectrum at the time of testing. This allows us to operationalize dominance as a continuous variable and measure its potential effects on L1 Spanish language processing with higher resolution.

2. Theoretical Background

2.1 Language dominance

Despite having long been present in the literature on bilingualism (Lambert, 1955), the multifaceted and complex construct of language dominance is difficult to define and measure (for a review, see Treffers-Daller & Silva-Corvalán, 2016). At the core of the concept of language dominance is the idea that isolating one of a speaker's languages is insufficient in depicting the bilingual/multilingual experience. Language dominance is inherently

relative between all of an individual's languages. One's 'dominant' language is said to be the language with higher proficiency, use, exposure, or a combination of any or all of these dimensions (e.g., Silva-Corvalán & Treffers-Daller, 2015). However, language dominance is not categorical or static, but rather a continuous measure that can shift over a lifetime. Domain of use also plays a crucial role, as an individual can have different dominance in oral vs. written language and in different situations, e.g. when talking to their parents vs. talking to their friends, when talking about art vs. talking about soccer, etc. (e.g., Grosjean, 2015). Dominance is clearly multidimensional and not absolute, and as of yet there is no consensus as to how different dimensions and domains interact with linguistic outcomes.

Similar to the difficulty in defining language dominance, its operationalization and measurement are equally complex (see Montrul, 2015; Birdsong, 2015). Most studies that measure language dominance rely on biographical measures or self-reports, which are easily collected with a language background questionnaire. This method can capture some of the multidimensionality of language dominance in a way that objective measures cannot by probing topics such as lifetime exposure and proficiency across multiple domains (e.g., reading, writing, speaking, listening). However, self-reported measurements vary in their reliability, with some studies supporting their effectiveness (e.g., Luk & Bialystok, 2013) and others finding them misleading or unreliable (e.g., Dunn & Tree, 2009) as they may reflect individual language attitudes more than language facility. More recently, there has been a call to use objective measures in operationalizing language dominance (Montrul, 2015). These measures vary from body-part naming tasks (e.g., O'Grady et al., 2009), other picture-naming tasks (e.g., Gollan et al., 2015), sentence repetition (Flege et al., 2002), self-paced reading (Fernández, 2003), and speech rate (e.g., Stevens, 2019), among others. However, these tasks are more time-consuming, may not straightforwardly operationalize the intended construct, and are inherently limited to one or two domains of language dominance.

Despite these difficulties in definition and measurement, language dominance plays an important role in the bilingual experience. Research on bilingualism has shown that lifetime and current language exposure, proficiency, and other factors subsumed by the super-construct of language dominance correlate with processes such as lexical access (e.g., O'Grady et al., 2009) and spoken fluency (e.g., Dunn & Tree, 2009). Obtaining a composite measure of dominance through a continuous, relative index that weighs both languages allows us to operationalize language dominance as a continuous

variable, which we can then correlate with online sentence processing measures.

2.2 Relative clause processing

To examine the effects of language dominance on bilingual sentence processing, we chose relative clauses as the syntactic structure of interest. The relative clause has played a large role in both theoretical syntax and psycholinguistic studies of language comprehension due to its suitability in testing knowledge of recursion, competence vs performance distinctions, and working memory limitations. Relative clauses are argued to be syntactically unambiguous (Babyonyshev & Gibson, 1999), manipulatable with simple word order changes (e.g., Grodner & Gibson, 2005), early-acquired (e.g., McKee et al., 1998), and cross-linguistically universal (Comrie & Fernández, 2012), reducing confounds of ambiguity resolution, lexical factors, and language attrition. While some argue that there are other complex influences on relative clause processing (e.g., MacDonald, 2013), a wealth of experimental research attests that the relative clause is an informative and useful structure for studying syntactic processing.

Many studies across a variety of methodologies and languages have shown that object relative clauses, such as (1b), are more costly or difficult to process than matched subject relative clauses, such as (1a) (see O'Grady, 2011 for review).

- (1) a. Subject relative clause (SRC):
the student [that ___ met the teacher]
- b. Object relative clause (ORC):
the student [that the teacher met ___]

Various syntactic, semantic, and psycholinguistic accounts attempt to explain this asymmetry. Some syntactic/semantic theories posit that a mismatch in thematic roles between the extracted noun in the matrix clause and its trace in the relative clause incurs a processing cost, making ORCs more difficult than SRCs (Sheldon, 1974). Others argue that intervening feature-matched elements (in this case, the subject of the embedded clause) disrupt the local relation between an extracted element and its trace (Relativized Minimality: Rizzi, 1990). Influential psycholinguistic theories attribute the asymmetry to either prominence factors, working memory constraints, or filler-gap dependencies. For example, the active filler hypothesis (Clifton & Frazier, 1989) posits that the parser will try to fill potential gaps at the earliest point allowed by the grammar, which is

unproblematic for SRCs but causes a failed parse in ORCs because the first possible gap is filled by the embedded noun phrase, necessitating a reanalysis.

Regardless of the explanation, the presence of this SRC-ORC asymmetry is well-documented in the literature.² Monolinguals of English (e.g., Traxler et al., 2002), Spanish (Betancort et al., 2009), and many other languages from various typological families have shown this SRC preference. This asymmetry has also been found in bilinguals in their L1 (e.g., Madsen, 2018) as well as highly proficient second-language learners in their L2 (Juffs & Rodríguez, 2014).

In a recent study, Stern and colleagues (2019) investigated L1 relative clause processing in Spanish heritage speakers as compared to first-generation late bilinguals. The study found that the late bilingual group demonstrated the expected SRC processing advantage (i.e. significantly higher fixation proportions on the target image while hearing a SRC compared to an ORC), while the heritage speaker group showed little evidence of a processing asymmetry despite self-rating as highly proficient in Spanish and performing similarly to the late bilinguals in behavioral measures of accuracy and response times. Since the groups did not differ in proficiency, the question arises as to what other factors could have driven the observed difference. The current study looks at individual differences among bilinguals with varying language experience in their L1 to probe the effect of language dominance on subject and object relative clause processing.

2.3 Bilingual processing in the Visual World

Based on Cooper's (1974) mind-eye theory that posits a closely time-locked relationship between spoken language processing and eye fixations on a visual scene, the Visual World Paradigm (VWP; Allopenna et al., 1998) has become an increasingly-utilized methodology for examining online language processing. VWP studies track the real-time location of participant eye gaze on a visual display while they listen to spoken language (for a review, see Huettig et al., 2011), allowing researchers to examine language processing without confounds caused by varying literacy levels or metalinguistic judgments (Tanenhaus et al., 1995). Research in the VWP has shown that listeners use semantic and morphosyntactic cues to predict upcoming linguistic information, showing anticipatory eye movements

² For an alternative account, see MacDonald, 2013.

towards images before the spoken word (e.g., Altmann & Kamide, 1999). This methodology has also provided insight into how spoken input is integrated with information retrieved from the visual environment (e.g., Grodner et al., 2010).

Studies in the VWP have also enhanced our understanding of bilingual processing. Bilingual children have been shown to exhibit semantic prediction in their L2 at a comparable level to their monolingual peers, sometimes with more rapid prediction which suggests a possible bilingual advantage over monolingual participants (Brouwer et al., 2017). Highly-proficient bilinguals take advantage of morphosyntactic cues in the L2 as well, utilizing gender marking (Dussias et al., 2013) to predict upcoming linguistic information. However, the VWP has also revealed limitations in L2 processing and prediction based on proficiency, L1 similarity, and productive accuracy (e.g., Dussias et al., 2013; Lew-Williams & Fernald, 2010).

The modulating effects of L2 proficiency on L1 processing have also been noted using the VWP, though to a limited degree. This has been evidenced with Chinese-English bilinguals with high English proficiency, who showed increased eye fixations when presented with smaller Chinese phonological units that are not major processing units for spoken word recognition in Chinese (Brouwer et al., 2017). An effect of L2 exposure on L1 processing was also found by Stern et al. (2019), where Spanish heritage speakers exhibited different processing patterns than late-US arrival Spanish-English bilinguals while parsing subject and object relative clauses despite showing comparable levels of comprehension. The methodology of Stern et al. (2019) was replicated by the current study.

2.4 Present study

In this study, we sought to explore the effect of individual language dominance on subject and object relative clause processing among Spanish speakers with different L1 experiences. This study was an extension of Stern et al. (2019), but rather than manipulating a group comparison we focused on the effect of individual language dominance. To operationalize language dominance, we chose the index provided by the Bilingual Language Profile (Birdsong et al., 2012). Gaze fixation served as a proxy measurement of language processing during a picture selection task in the VWP. In addition to the effect of language dominance on target fixation, we also looked for an effect of relative clause type to further probe the subject-object processing asymmetry. Thus, our main research question was whether

language dominance, operationalized as a continuous variable, modulates gaze movements in L1-Spanish relative clause processing.

We additionally analyzed participant accuracy and response time during the picture selection task to further examine effects of language dominance and sentence type. We made two predictions: that increased Spanish dominance would 1) lead to increased evidence of a subject-object asymmetry in gaze data measures, with significantly higher fixation proportions on the target image during SRCs compared to ORCs, and 2) have minimal effect on behavioral measures of comprehension accuracy and response time. These predictions are based on findings from Stern et al. (2019), which found greater evidence of the subject-object asymmetry in the gaze data of late bilinguals compared to heritage speakers, while showing no group differences in behavioral measures. We will first present the results of the language dominance questionnaire, followed by the results of the eye-tracking experiment including both gaze data and behavioral measures.

3. Language dominance

3.1 Methods

3.1.1 Participants

Fifty-nine Spanish-English bilingual adults with normal or corrected-to-normal hearing and vision (aged 19-55: $M = 27.25$, $SD = 8.37$) participated in this study. Data from forty-one of these participants was reported in the previous group-level analysis in Stern et al. (2019). All participants self-rated as fluent in both Spanish and English, spoke primarily Spanish with their caregivers until at least age 10, and resided in New York City at the time of testing. In other words, all participants were highly proficient bilingual adults whose childhood home language was Spanish but who live and work in a society where English is the majority language.

To represent a continuum of language dominance, participants were chosen to range widely in their age of arrival to the anglophone US. In our sample, 21 participants were born in the continental US ($n = 11$) or arrived before age 9 ($n = 10$). These participants were classified as heritage speakers in Stern et al. (2019), as they spent their entire childhood in an environment where their home language was the societal minority language. They had

minimal formal education in their L1 (Spanish),³ and as a group tended to be English dominant. Another 20 participants were raised in regions where Spanish is the societal majority language, were educated in Spanish, and did not arrive to the anglophone US until age 17 or older. These participants were classified as late bilinguals by Stern et al. (2019), as they did not live in an English-majority society until adulthood and are mostly dominant in Spanish. Finally, to bridge the continuum of language dominance, a third L2-English learner group was tested ('Middle' group: $n = 18$) consisting of participants who moved to an English-dominant society between ages 10 and 16. These participants vary more in their language exposure and dominance, as they have received education in both Spanish and English and became immersed in an English-speaking society during adolescence. Together, these participants represent a wide spectrum of relative language dominance between Spanish and English.

3.1.2 Procedure and analysis

For the dominance measurement task, participants completed the Bilingual Language Profile (BLP; Birdsong et al., 2012).⁴ This is a four-module, 19-item questionnaire which probes language history, use, proficiency, and attitudes. Weighted answers for the four modules generate a subtractive score of relative language dominance, with positive scores indicating increased Spanish dominance and negative scores indicating increased English dominance. This dominance score is not intended to be absolute or categorical but rather relative to an individual's experience in both languages. Scores close to zero are more reflective of a balanced bilingual than of a bilingual who is clearly dominant in one language over the other. It is crucial to interpret this score as a continuous and relative index of language dominance.

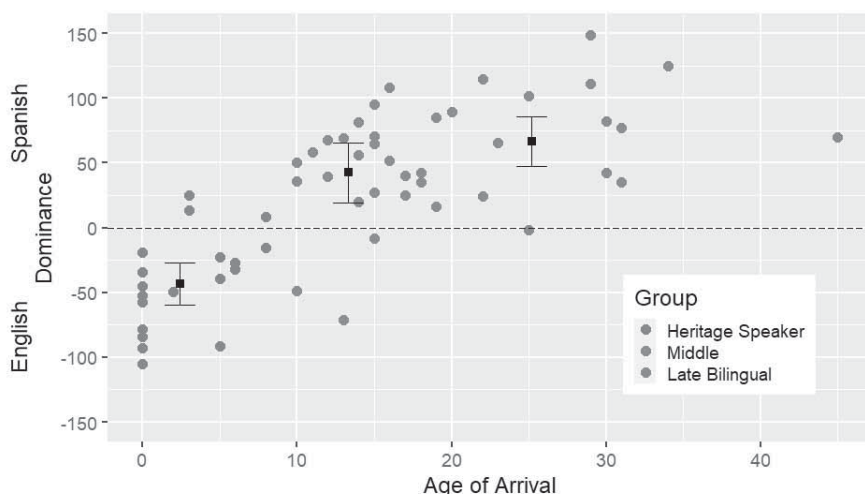
³ Spanish is referred to throughout this paper as the L1 of all participants. However, four participants did report having exposure to English from birth, and could be classified as simultaneous bilinguals with two L1s. The labeling of Spanish as the L1 is not intended to diminish the distinction between simultaneous and sequential learners, but rather for simplicity's sake as this is not relevant to the research question of this study.

⁴ The BLP was administered at different times. Most participants completed the BLP the same day as the eye-tracking experiment. However, some participants had already filled out the BLP during a previous experiment (within 12 months of the current study) and were not asked to re-complete it.

3.2. Results

To illustrate the continuum of dominance represented in this study, each participant's relative language dominance score by age of arrival is plotted in Figure 1. We include groups in the visualization not because it is an experimental variable, but rather to show the range of dominance scores across the traditional group categories. Heritage speakers are represented with green dots, late bilinguals with purple dots, and the late childhood arrivals with orange dots. Means and 95% confidence intervals are also shown for each group.

Figure 1: Age of arrival and relative language dominance



Overall, Spanish dominance increases as age of arrival to the anglophone US increases. That is, participants who were born in the US or arrived in early childhood tend to be more English dominant, while participants who immigrated during adolescence or adulthood tend to be more Spanish dominant. Additionally, despite traditional definitions of heritage speakers in the literature which assume that bilinguals raised with early exposure to the societal majority language will have stronger dominance in that language, there are three early arrivals who have positive scores, indicating Spanish dominance. There are also many participants with scores close to zero in this dataset, indicating a balance between Spanish and English rather than a clear, categorically-dominant language. These results demonstrate the utility of the continuous, relative measure of language dominance over

a categorical or absolute one. This measure is sensitive to small variation, thus making it a suitable index for language dominance to assess its effects on relative clause processing.

4. Eye-tracking experiment

4.1 Method

4.1.1 Participants

The fifty-nine participants in this experiment are the same ones previously described in Section 3.1.1, as all participants who were administered the BLP also completed the eye-tracking experiment.

4.1.2 Stimuli

Participants were presented with 40 complex Spanish sentences: 10 items per experimental condition (subject relative, SRC; and object relative clauses, ORC) and 20 fillers. All items contained varying combinations of the same five noun phrases, which were anthropomorphic animals with masculine gender in Spanish. An example of an SRC stimulus is shown in (2), and an ORC in (3).⁵ Relative and matrix verbs were all transitive, and all relative clauses were subject embedded across both stimuli conditions.

Subject relative clause (SRC)

- (2) El conejo, que ____ abraza al perro, cepilla al oso.
 the.M rabbit that hug.3SG to-the.M dog brush.3SG to-the.M bear
 ‘The rabbit, that ____ hugs the dog, brushes the bear.’

Object relative clause (ORC)

- (3) El perro, que el conejo abraza ____, cepilla al oso.
 the.M dog that the.M rabbit hug.3SG brush.3SG to-the.M bear
 ‘The dog, that the rabbit hugs ____, brushes the bear.’

⁵ As a reviewer aptly pointed out, there is another construction for expressing ORCs in Spanish where the word order does not differ from SRCs by manipulating the DP/PP introducing the embedded noun phrase. For example, ‘El perro, que abraza el conejo, cepilla al oso’ shares the same meaning as the ORC in (4). In fact, this manipulation has been adopted in eye-tracking studies (e.g., Betancort et al., 2009). This alternation primarily manipulates differential object marking, which has been shown to be a difficult grammatical feature for Spanish heritage speakers (e.g., Montrul & Bowles, 2009). For this reason, we chose the ORC manipulation that alters word order to reduce possible misinterpretation.

Sentences were presented auditorily with simultaneous presentation of a three-image picture array. The use of auditorily-presented stimuli reduces confounds that may arise from varying L1 literacy levels among these participants with different language experiences in Spanish. The positions of the image types were counterbalanced across trials. One image corresponded to the linguistic stimulus, one distractor image always corresponded to the stimulus until the matrix verb ('Consistent' distractor), and the 'Other RC' distractor always corresponded to the reverse interpretation of the relative clause (e.g., if the stimulus was an SRC, then the distractor would depict the ORC interpretation of the matrix subject). Figure 2 shows an example of a three-image visual display that is consistent with example (2) above.

Figure 2. Sample visual display during experimental trial



4.1.3 Procedure

Each trial began with a black cross fixation marker appearing at the top-center of the screen. When participants clicked on the cross, they were shown three images and asked to familiarize themselves with the images. After a two-second delay the cross reappeared, and when ready participants clicked again to hear the auditory stimulus. Participants then selected the image that best represented the aurally-presented stimulus with a mouse click. After a short practice session, the experiment began with stimuli presented in a pseudorandomized order.⁶ Gaze fixations were recorded throughout each trial at a sampling rate of 60 Hz using a Tobii TX300 eye-


⁶ Participants first completed a portion of the experiment where relative clauses were embedded in intransitive matrix sentences, but these results are not reported in the present paper. After a short break, participants' eye movements were recalibrated and the second half of the eye-tracking experiment was completed with transitive matrix clauses, which is the focus of this study.

tracker, and the experiment was presented with E-Prime 2.0 (Schneider et al., 2002).

4.1.4 Analysis

Following the analysis of Stern et al. (2019), gaze data was divided into four temporal regions of the stimulus. This is illustrated in Figure 3.

Figure 3. Division of auditory stimuli into temporal regions

	1	2	3	4*
SRC:	El conejo que	abrazo al perro	cepilla al oso	
ORC:	El perro que	el conejo abraza	cepilla al oso	

Region 1 was coded from the onset of the sentence until the onset of the first word after the relativizer *que*. Linguistically, this information is equivalent across both conditions. No information is provided to the participant that would allow them to eliminate any of the three images in this region. During Region 2 (the Relative Clause Region), participants hear the onset of the first word after the relativizer *que* through the onset of the matrix verb. For SRCs the first word of this region is the subordinate verb, while for ORCs the first word is the determiner *el* (which begins the subordinate noun phrase). Information provided during the Relative Clause Region would allow participants to eliminate the ‘Other RC’ distractor. Region 3 (the Matrix Predicate Region) provides the remaining information required to eliminate the ‘Consistent’ Distractor and converge upon the target image, and for both conditions this region extends from the onset of the matrix verb to sentence offset. Finally, Region 4 is the time window from spoken sentence offset until participants click on an image. No linguistic information is provided in the final region. The two regions of interest for this study are the Relative Clause Region and the Matrix Predicate Region.

To analyze gaze data, only observations for which the eye tracker was at its highest validity level were retained (0 on the 0-4 scale output by the eye tracker: 8.9% of the data was removed). Furthermore, two participants with average total gaze fixation proportions of less than 30% on any image across all regions and stimuli were removed entirely from gaze data analysis. Proportion of fixation on the target image within each region from accurate trials only was then used as the dependent variable for gaze models. Two behavioral measures were also analyzed in addition to gaze data: comprehension accuracy and response time. Participant-average accuracy

was calculated by relative clause type, and comprehension accuracy was operationalized as dichotomous accuracy (0,1) for modeling. One highly Spanish-dominant participant was removed as an outlier from analyses because their average accuracy across conditions was more than three standard deviations lower than the mean. Response time, measured from sentence offset until participants clicked on an image, was log-transformed to address the non-Gaussian distribution of the data, and outlying response times of less than 50ms or greater than 20,000ms were removed. All analyses, visualizations, and models were run using R (R Core Team, 2019).

4.2 Results

Our results are presented in two sections: gaze data results and behavioral results. We begin each section by providing relevant descriptive statistics. Then, we present inferential statistics for the effects of relative clause type and language dominance on the dependent measures described above. Finally, we provide a brief summary of all findings.

4.2.1 Gaze data

A course-grained analysis of target gaze fixation proportion by region is plotted in Figure 4. This shows the proportion of fixation on the target image by condition for each of the four temporal regions, including means and within-subject adjusted 95% confidence intervals (see Morey, 2008). The dashed line represents chance, assuming that participants are looking at one of the three images on the screen. As expected, target fixation increases over time as participants gain more information allowing them to converge on the target image. Figure 5 shows target fixation proportions for SRCs and ORCs in the two regions of interest; the Relative Clause Region and the Matrix Predicate Region. Mean target fixation proportion is higher for SRCs than ORCs in the Relative Clause Region (SRC: $M = 0.421$, $SD = 0.521$; ORC: $M = 0.374$, $SD = 0.527$) but not in the Matrix Predicate Region (SRC: $M = 0.549$, $SD = 0.525$; ORC: $M = 0.544$, $SD = 0.557$).

Figure 4. Target proportion fixation by condition for each temporal region

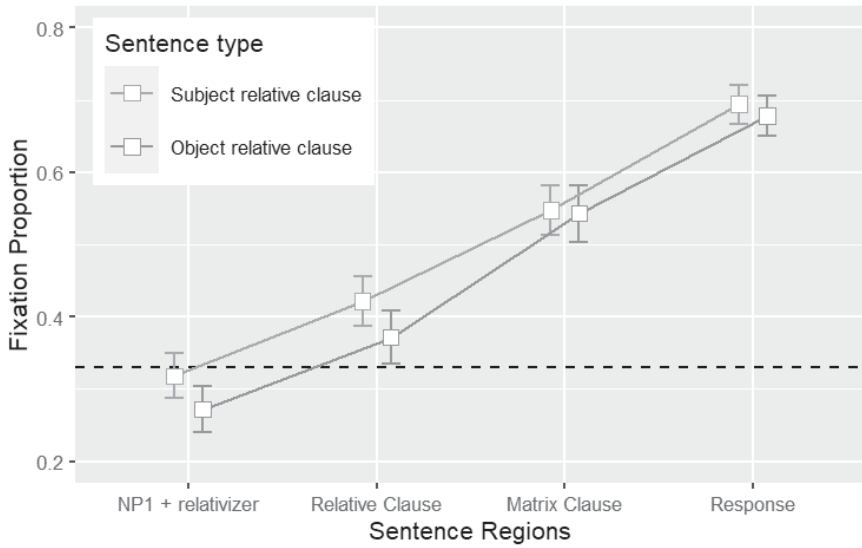
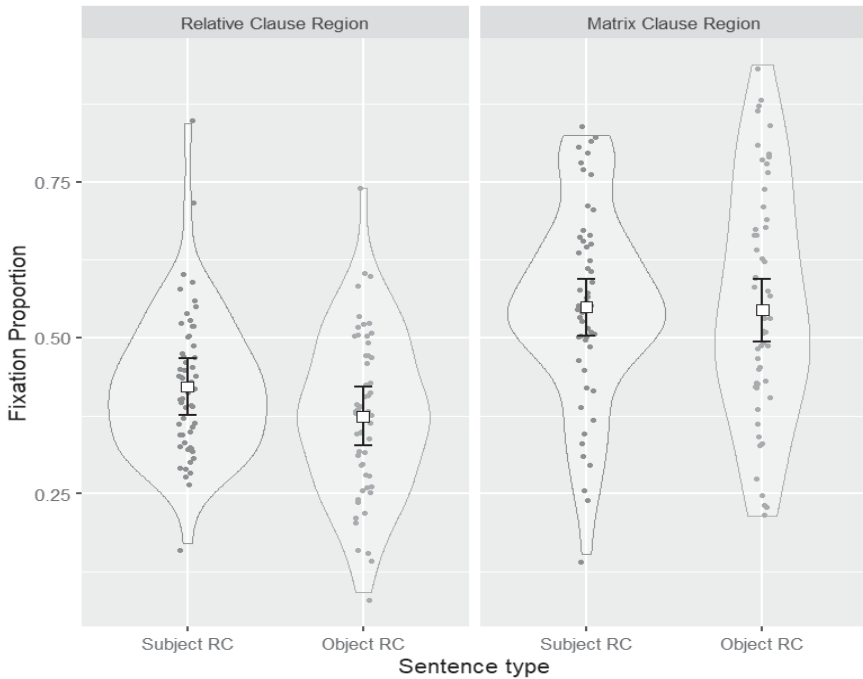


Figure 5. Target fixation proportion in the Relative Clause Region (left) and Matrix Predicate Region (right)



Proportion of gaze fixation on the target image during the Relative Clause and Matrix Predicate Regions is modeled using mixed effects beta regression models in R using the *gamlss* function (*gamlss* R package; Rigby & Stasinopoulos, 2005). Beta regression models are argued to be flexible for modeling limited range data such as proportion data (Johnson et al., 1995). Fixation proportions are adjusted for beta regression by adding 10^{-17} to all values of 0 and subtracting 10^{-17} from all values of 1. Model predictors include sentence type (SRC, ORC) which was sum contrast coded with SRC as “-1” and ORC as “1”, language dominance which was scaled and centered, and the interaction between the two factors. The *gamlss* function affords the estimations of both means (μ ; μ) and standard deviations (σ ; σ), as well as the inclusion of random intercepts and slopes for random effects (i.e., participants and items). We compared four models for the two regions of interest using a generalized Akaike information criterion (GAIC).

These models contrasted the benefit of the inclusion of a sentence type random slope for items,⁷ and the inclusion of fixed effects to model sigma. The model with the lowest AIC for the Relative Clause Region included by-participant and by-item random intercepts, as well as a sentence type random slope for items, with no fixed effect for sigma. For the Matrix Predicate Region, the model with the lowest AIC included fixed effects, by-participant and by-item random intercepts, and sentence type random slopes for items for both mu “ μ ” and sigma “ σ ”. While the sum-contrast coding described above allowed us to model the effect of each predictor across levels of the other, we also conducted follow-up models using treatment contrasts to further probe the effect of language dominance at each level of the sentence type predictor.

Relative Clause Region

Table 1. Effects of RC type and language dominance on target fixation proportions- Relative Clause Region

Parameters	Fixed effect	<i>Estimate</i>	<i>SE</i>	<i>t</i>	<i>P</i>
μ link = logit	(Intercept)	-0.008	0.043	-0.182	0.856
	Sentence type (ST)	-0.084	0.042	-1.975	0.049
	Language dominance (LD)	0.005	0.042	0.117	0.907
	ST:LD	-0.099	0.042	-2.368	0.018
	LD effect on SRCs	0.104	0.058	1.792	0.074
	LD effect on ORCs	-0.094	0.060	-1.564	0.118
σ (link = logit)	(Intercept)	2.641	0.032	83.680	0.000

The output of the Relative Clause Region model is shown in Table 1, and a visualization of the relationship between language dominance and target fixation is shown in Figure 6 (panel on the left). For SRC items, there is a pattern of increased target fixation proportions as Spanish dominance increases. The opposite pattern is found for ORCs: Spanish dominance correlates negatively with target fixation proportion. Modeling shows no

⁷ Models with more complex random structure, particularly with dominance random slopes or any random slope for participants did not converge.

main effect of language dominance on target fixation proportions in the Relative Clause Region, while sentence type does show a main effect such that target fixation proportions are higher for SRCs than ORCs. The interaction between these two factors is also a significant predictor. Follow up models show that the effect of language dominance has opposite directionality for the SRC and ORC conditions, with neither significantly driving the interaction more than the other, although the effect is marginal for SRC sentences. Thus, the interaction appears to be driven by both the increase in Spanish dominance as target fixation increases during SRCs as well as the decrease in Spanish dominance as target fixation increases during ORCs.

Matrix Predicate Region

Table 2. Effects of RC type and language dominance on target fixation proportions- Matrix Predicate Region

Parameters	Fixed effect	<i>Estimate</i>	<i>SE</i>	<i>t</i>	<i>p</i>
μ link = logit	(Intercept)	0.061	0.043	1.405	0.160
	Sentence type (ST)	-0.017	0.043	-0.391	0.696
	Language dominance (LD)	-0.118	0.043	-2.758	0.006
	ST:LD	-0.117	0.043	-2.738	0.006
	LD effect on SRCs	-0.001	0.058	-0.015	0.988
	LD effect on ORCs	-0.236	0.063	-3.755	0.000
σ link = logit	(Intercept)	2.858	0.031	91.344	0.000
	Sentence type (ST)	0.112	0.031	3.580	0.000
	Language dominance (LD)	0.033	0.030	1.101	0.271
	ST:LD	0.056	0.030	1.841	0.066
	LD effect on SRCs	-0.022	0.041	-0.552	0.581
	LD effect on ORCs	0.089	0.045	1.983	0.048

Effects of relative clause type and language dominance on target fixation proportions for the Matrix Predicate Region are modeled in Table 2 and

visualized in Figure 6 (panel on the right). Similar to the Relative Clause Region, the Matrix Predicate Region shows a positive correlation trend between target fixation proportions and Spanish dominance for SRCs and a negative correlation for ORCs. While relative clause type shows no main effect on target fixation, language dominance is a significant predictor in this region such that as Spanish dominance increases, proportion fixation to the target image decreases. Conversely, increased English dominance predicts increased target fixation proportions across conditions. The model also revealed a significant interaction between language dominance and sentence type, which follow up models show to be largely driven by the negative effect of increased Spanish dominance on gaze fixation proportions to the target image while parsing a sentence with an ORC. The model of sigma revealed a significant effect of sentence type and a marginally significant interaction between relative clause type and language dominance. Target fixations were significantly more variable for ORCs than SRCs, and increased Spanish dominance increased variance during ORC processing but not SRC processing. In the Matrix Predicate Region, increased Spanish dominance is detrimental to target fixation during ORCs.

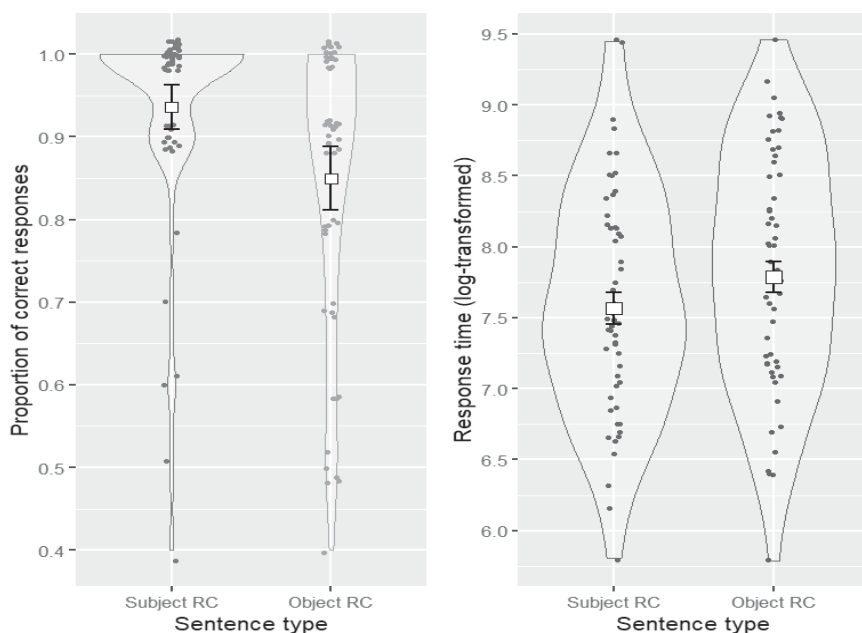
Figure 6. Target fixation proportions by Spanish dominance for Regions 2 and 3, where positive scores indicate greater Spanish dominance.



4.2.2 Behavioral data

Accuracy and log-transformed response time results by condition are visualized in Figure 7, with violin plots including means and 95% confidence intervals (adjusted for within-subject designs). Overall, participants were highly accurate in both conditions, though participants were more accurate on items with SRCs ($M = 0.936$, $SD = 0.32$) than those with ORCs ($M = 0.850$, $SD = 0.47$). In turn, mean response time was lower for the SRCs ($M = 7.568$, $SD = 1.27$) than for ORCs ($M = 7.787$, $SD = 1.17$).

Figure 7. Comprehension accuracy and log-transformed response time by condition



Response accuracy is modeled through a generalized linear mixed effects regression using the `glmer` function (`lme4` package; Bates et al., 2014) with a binary dependent variable coded for correct and incorrect responses. Predictors in the model include sentence type (SRC vs. ORC), participants' relative language dominance, and the interaction between the two. Sentence type is sum contrast coded, while the continuous predictor of language dominance is scaled and centered. We opted for models with maximal random structure justified by the design and followed recommendations given (Barr et al., 2013) whenever confronted with non-convergence. The

model is reported in Table 3, which includes by-item and by-participant random intercepts as well as their corresponding random slope for sentence type.

Table 3. Effects of RC type and language dominance on comprehension accuracy

	<i>Estimate</i>	<i>SE</i>	x^2	$p(x^2)$
(Intercept)	2.973	0.261		
Sentence type (ST)	-0.653	0.204	9.009	0.003
Language dominance (LD)	0.106	0.203	0.275	0.600
ST:LD	-0.239	0.182	1.714	0.191

Based on significance reporting from a chi-square statistic using a model comparison approach, sentence type is the only predictor with a main effect on accuracy. Participants are significantly more accurate on SRCs than ORCs. Language dominance and the interaction between the two predictors do not show a main effect on accuracy.

Log-transformed response times are modeled using a linear mixed effects regression also with the lme4 R package (Bates et al., 2014). The results from the model are presented in Table 4. We follow the same approach as for the accuracy data, and the final model includes predictors of sentence type and language dominance (both as described above) as well as the interaction between the two, with by-item and by-participant random intercepts as well as their corresponding random slope for sentence type.

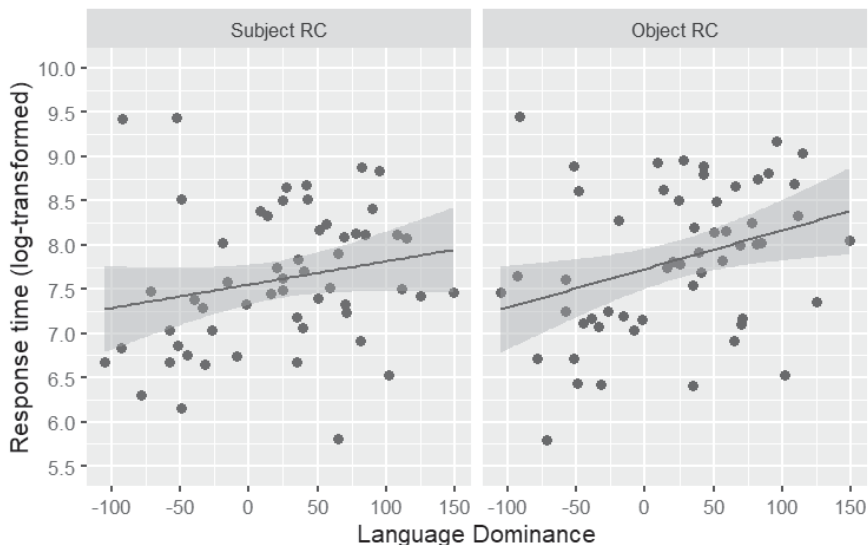
Table 4: Effects of RC type and language dominance on log-transformed response time

	<i>Estimate</i>	<i>SE</i>	x^2	$p(x^2)$
(Intercept)	7.711	0.115		
Sentence type (ST)	0.110	0.044	4.845	0.028
Language dominance (LD)	0.225	0.097	5.089	0.024
ST:LD	0.038	0.028	1.791	0.181

Significance scores from chi-square model comparisons reveal a main effect of sentence type such that ORCs have longer response times than SRCs, and

a main effect of language dominance such that increased Spanish dominance leads to increased response times.

Figure 8. Log-transformed response time by language dominance for SRCs and ORCs



4.3 Summary of findings

In the gaze data, increased Spanish dominance led to greater evidence of a subject-object asymmetry in relative clause processing, driven primarily by a late negative effect of increased Spanish dominance on ORC processing (in the Matrix Predicate Region). Language dominance did not have a main effect on target gaze fixation proportions in the Relative Clause Region, but a significant interaction in this region showed that increased Spanish dominance led to higher target fixation proportions for SRCs and to lower target fixation proportions for ORCs. Relative clause type also had a main effect in the Relative Clause Region, such that target fixation proportions were greater for SRCs than ORCs. In the Matrix Clause Region, language dominance predicted target fixation proportions while sentence type in itself did not. A significant interaction in the Matrix Predicate Region showed that as Spanish dominance increased, target fixation proportions increased for SRCs but decreased for ORCs, and the interaction was driven mainly by the negative effect on ORCs. Sentence type was the only significant

predictor of comprehension accuracy, such that SRCs had significantly higher accuracy than ORCs. Finally, both sentence type and language dominance had significant effects on log-transformed response times. ORCs led to greater response times than SRCs, and increased Spanish dominance also led to slower response times.

5. Discussion

Overall, the prediction that increased Spanish dominance leads to greater evidence of a subject-object processing asymmetry is supported. The processing asymmetry is present to a higher degree among participants with increased Spanish dominance, as evidenced by the interaction between language dominance and relative clause type in both regions. Interestingly, the increased SRC/ORC asymmetry among participants with greater Spanish dominance is driven more by a late ORC processing cost than an early SRC benefit. In the Relative Clause Region (Region 2), English-dominant participants are looking at the target at almost the same proportion for SRCs and ORCs (see Figure 6). Conversely, Spanish-dominant participants diverge in their target fixations for SRCs and ORCs by the Relative Clause Region. There is minimal evidence that this reflects an SRC advantage in the Relative Clause Region though, as higher Spanish dominance only leads to slightly higher target fixation proportions. This pattern holds for SRCs in the Matrix Predicate Region (Region 3), where increased Spanish dominance offers only the slightest increase in target gaze fixation. On the other hand, there is a steep decline in target fixation proportions in this region for ORCs as Spanish dominance increases.

We posit that the detrimental effect of Spanish dominance on ORC processing in the matrix predicate region is due to increased integration costs for participants with higher Spanish dominance when an SRC is predicted (along the lines of the active filler hypothesis) and the missed parse forces a reanalysis. While this study did not explicitly measure this prediction, results are compatible with an interpretation that participants who are not dominant in the testing language are not utilizing active prediction and online resources to the extent that test-language-dominant participants are. This was suggested in Stern et al. (2019) as a group-level effect of differential online resources, such that the heritage speakers (who were generally less Spanish-dominant) predicted less actively than the late bilinguals (who were generally more Spanish-dominant). The current study indicates that this group-level difference was in fact driven by an effect of L1 language dominance on L1 processing. This interpretation would explain

the slight advantage on SRCs with increased Spanish dominance, and the more substantial detrimental effect on ORCs.

The behavioral results were slightly unexpected. Comprehension accuracy was not affected by language dominance, which is consistent with previous findings that accuracy is a less sensitive measure for highly proficient bilingual populations with subtle differences in relative language strength (e.g., O'Grady et al., 2009). However, participants with higher Spanish dominance responded slower to Spanish stimuli than participants with higher English dominance. While no significant interaction was found, numerical trends suggest that the processing demand of ORCs in Spanish may affect response times among Spanish-dominant bilinguals more than English-dominant participants.

This study demonstrates that operationalizing language dominance as a continuous, relative measure has the potential to reveal insights into bilingual language processing beyond those afforded by categorical or absolute measures. While many studies have found that language dominance correlates with different measures of language proficiency, performance, and comprehension (see Section 2), most previous studies have still treated language dominance categorically. That is, groups have been defined as dominant in Language A or dominant in Language B. However, as the theoretical literature on language dominance emphasizes, the construct is inherently gradient with different individual degrees of language dominance. Operationalizing it as such allows us to more carefully explore the effects of language dominance on other measures, which could benefit the field of bilingual processing immensely.

6. Conclusion

This study probed the effect of language dominance on the online processing and offline comprehension of subject and object relative clause constructions. Spanish-English bilinguals along a continuum of language dominance who share Spanish as the L1 but live in an English-dominant society completed an eye-tracking study with aurally-presented Spanish stimuli during a picture selection task. Our results demonstrate that language dominance impacts bilinguals' eye movements as they are presented aural stimuli in their first-learned language while looking at corresponding images. This influence extends beyond group differences between heritage speakers and late bilinguals as found in Stern et al. (2019) and shows that individual language dominance plays an important role in bilingual relative clause processing. Participants with higher Spanish

dominance seem to integrate aural information quickly during stimuli with subject relatives and have a delayed convergence on the correct image during object relative constructions. On the other hand, participants with lower Spanish dominance (i.e., more English dominant) do not seem to demonstrate the expected processing asymmetry between subject and object relative clauses despite comparable levels of comprehension to Spanish-dominant counterparts.

This study revealed a detrimental effect of ORCs that disproportionately affects the eye movements and response times of highly Spanish-dominant individuals as compared to less Spanish-dominant participants. This effect of language dominance could possibly be due to differential processing strategies in the dominant and non-dominant languages, as more active online prediction among individuals with greater Spanish dominance could possibly be leading to higher integration costs. However, further probing of this topic is crucial to understanding the interaction between language dominance and bilingual processing. Importantly, this study demonstrates the utility of continuous, relative measures of language dominance in the investigation of bilingual language processing.

Acknowledgements

We are grateful to the vibrant Spanish-speaking community in NYC which has inspired and allowed us countless ways to explore the bilingual experience, and to those members who participated in this study. Many thanks to Ernesto Guerra for his significant contribution to data analysis. We would also like to thank Christen N. Madsen II, Richard G. Schwartz, and Second Language Acquisition Lab research assistants Daniela Castillo, Daniel Choconta, Christina Dadurian, Andrea Monge, Omar Ortiz, Matthew Stuck, and Armando Tapia.

References

- Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, 38(4), 419–439.
- Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3), 247–264.

- Babyonyshev, M., & Gibson, E. (1999). The complexity of nested structures in Japanese. *Language*, 423–450.
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure in mixed-effects models: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *ArXiv Preprint ArXiv: 1406.5823*.
- Bedore, L. M., Peña, E. D., Summers, C. L., Boerger, K. M., Resendiz, M. D., Greene, K., Bohman, T. M., & Gillam, R. B. (2012). The measure matters: Language dominance profiles across measures in Spanish–English bilingual children. *Bilingualism: Language and Cognition*, 15(3), 616–629.
- Betancort, M., Carreiras, M., & Sturt, P. (2009). The processing of subject and object relative clauses in Spanish: An eye-tracking study. *Quarterly Journal of Experimental Psychology*, 62(10), 1915–1929.
- Birdsong, D. (2015). Dominance in bilingualism: Foundations of measurement, with insights from the study of handedness. In C. Silva-Corvalán & J. Treffers-Daller (eds.) *Language Dominance in Bilinguals: Issues of Measurement and Operationalization* (pp. 85–105). Cambridge: Cambridge University Press.
- Birdsong, D., Gertken, L. M., & Amengual, M. (2012). Bilingual language profile: An easy-to-use instrument to assess bilingualism. *COERLL, University of Texas at Austin*.
- Brouwer, S., Özkan, D., & Küntay, A. C. (2017). Semantic prediction in monolingual and bilingual children. In E. Blom, L. Cornips & J. Schaeffer (eds.) *Cross-linguistic Influence in Bilingualism* (pp. 49–74). Amsterdam: John Benjamins.
- Clifton, C., & Frazier, L. (1989). Comprehending Sentences with Long-Distance Dependencies. In G.N. Carlson, M.K. Tanenhaus (eds.) *Linguistic Structure in Language Processing. Studies in Theoretical Psycholinguistics*, vol 7. Dordrecht: Springer.
- Comrie, B., & Fernández, Z. E. (2012). *Relative Clauses in Languages of the Americas: A Typological Overview*. Amsterdam: John Benjamins.
- Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, 6(1), pp. 84–107.
- Dunn, A. L., & Tree, J. E. F. (2009). A quick, gradient Bilingual Dominance Scale. *Bilingualism: Language and Cognition*, 12(03), 273–289.
- Dussias, P. E., Valdés Kroff, J. R., Guzzardo Tamargo, R. E., & Gerfen, C. (2013). When Gender and Looking Go Hand in Hand: Grammatical

- Gender Processing In L2 Spanish. *Studies in Second Language Acquisition*, 35(2), 353–387.
- Fernández, E. M. (2003). Bilingual Sentence Processing: Relative Clause Attachment in English and Spanish. Amsterdam: John Benjamins.
- Flege, J. E., Mackay, I. R. A., & Piske, T. (2002). Assessing bilingual dominance. *Applied Psycholinguistics*, 23(04), 567–598.
- Gollan, T. H., Starr, J., & Ferreira, V. S. (2015). More than use it or lose it: The number-of-speakers effect on heritage language proficiency. *Psychonomic Bulletin & Review*, 22(1), 147–155.
- Gollan, T. H., Weissberger, G. H., Runnqvist, E., Montoya, R. I., & Cera, C. M. (2012). Self-ratings of spoken language dominance: A multilingual naming test (MINT) and preliminary norms for young and aging Spanish-English bilinguals. *Bilingualism: Language and Cognition*, 15(3), 594.
- Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science*, 29(2), 261–290.
- Grodner, D. J., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010). “Some,” and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition*, 116(1), 42–55.
- Grosjean, F. (2015). The Complementarity Principle and its impact on processing, acquisition, and dominance. In C. Silva-Corvalán & J. Treffers-Daller (eds.) *Language Dominance in Bilinguals: Issues of Measurement and Operationalization* (pp. 66–84). Cambridge: Cambridge University Press.
- Huetting, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*, 137(2), 151–171.
- Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions*. New York: John Wiley.
- Juffs, A., & Rodríguez, G. A. (2014). *Second language sentence processing*. London: Routledge.
- Lambert, W. E. (1955). Measurement of the linguistic dominance of bilinguals. *Journal of Abnormal Psychology*, 50(2), 197–200.
- Lew-Williams, C., & Fernald, A. (2010). Real-time processing of gender-marked articles by native and non-native Spanish speakers. *Journal of Memory and Language*, 63(4), 447–464.
- Luk, G., & Bialystok, E. (2013). Bilingualism is not a categorical variable: Interaction between language proficiency and usage. *Journal of Cognitive Psychology*, 25(5), 605–621.

- MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, 4, 226.
- Madsen II, C. N. (2018). *De-centering the monolingual: A psychophysiological study of heritage speaker language processing*. Ph.D Dissertaion. The Graduate Center, CUNY.
- McKee, C., McDaniel, D., & Snedeker, J. (1998). Relatives children say. *Journal of Psycholinguistic Research*, 27(5), 573–596.
- Meisel, J. M. (2007). The weaker language in early child bilingualism: Acquiring a first language as a second language? *Applied Psycholinguistics*, 28(3), 495.
- Montrul, S. (2015). Dominance and proficiency in early and late bilingualism. In C. Silva-Corvalán & J. Treffers-Daller (eds.) *Language Dominance in Bilinguals: Issues of Measurement and Operationalization* (pp. 15–35). Cambridge: Cambridge University Press.
- Montrul, S., & Bowles, M. (2009). Back to basics: Incomplete knowledge of Differential Object Marking in Spanish heritage speakers. *Bilingualism: Language and Cognition*, 12(3), 363–383.
- Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Reason*, 4(2), 61–64.
- O’Grady, W. (2011). Relative clauses: Processing and acquisition. In E. Kid (ed.) *The Acquisition of Relative Clauses: Processing, Typology and Function*. Amsterdam: John Benjamins 13–38.
- O’Grady, W., Schafer, A. J., Perla, J., Lee, O.-S., & Wieting, J. (2009). A psycholinguistic tool for the assessment of language loss: The HALA project. *Language Documentation & Conservation*, 3(1), pp. 100-112.
- Puig-Mayenco, E., Cunnings, I., Bayram, F., Miller, D., Tubau, S., & Rothman, J. (2018). Language dominance affects bilingual performance and processing outcomes in adulthood. *Frontiers in Psychology*, 9, 1199.
- R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape, *Applied Statistics*, 3(54), 507–554.
- Rizzi, L. (1990). *Relativized minimality*. Cambridge, MA: The MIT Press.
- Robinson, J. J., & Blumenfeld, H. K. (2018). Language dominance predicts cognate effects and inhibitory control in young adult bilinguals. *Bilingualism: Language and Cognition*, 22(5), 1–17.
- Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime: User’s guide*. Pittsburg, PA: Psychology Software Incorporated.

- Sheldon, A. (1974). The role of parallel function in the acquisition of relative clauses in English. *Journal of Verbal Learning and Verbal Behavior*, 13(3), 272–281.
- Silva-Corvalán, C., & Treffers-Daller, J. (2015). Digging into dominance: A closer look at language dominance in bilinguals. In C. Silva-Corvalán & J. Treffers-Daller (eds.) *Language Dominance in Bilinguals: Issues of Measurement and Operationalization* (pp. 1–14). Cambridge: Cambridge University Press.
- Stern, M. C., Madsen II, C. N., Stover, L. M., Lowry, C., & Martohardjono, G. (2019). Language history attenuates syntactic prediction in L1 processing. *Journal of Cultural Cognitive Science*, 3(2), 235–255.
- Stevens, L. S. (2019). *Heritage Speaker and Late Bilingual L2 Relative Clause Processing and Language Dominance Effects*. Master's Thesis. The Graduate Center, CUNY.
- Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217), 1632–1634.
- Traxler, M. J., Morris, R. K., & Seely, R. E. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language*, 47(1), 69–90.
- Treffers-Daller, J. (2011). Operationalizing and measuring language dominance. *International Journal of Bilingualism*, 15(2), 147–163.
- Treffers-Daller, J. (2015). Language dominance: The construct, its measurement, and operationalization. In C. Silva-Corvalán & J. Treffers-Daller (eds.) *Language Dominance in Bilinguals: Issues of Measurement and Operationalization* (pp. 235–265). Cambridge: Cambridge University Press.
- Treffers-Daller, J., & Silva-Corvalán, C. (2016). *Language dominance in bilinguals: Issues of measurement and operationalization*. Cambridge: Cambridge University Press.

REMNANTS OF THE DELAY OF PRINCIPLE B EFFECT IN ADULTS: A NEW APPROACH TO AN OLD PROBLEM

MARGREET VOGELZANG,¹ REGINA HERT
& ESTHER RUIGENDIJK

Abstract

It is known that German children generally show correct interpretation of object pronouns like *him* in German sentences equivalent to ‘*John saw him*’ by age 4. In contrast, Dutch (and English) children often incorrectly allow such object pronouns to refer to the sentential subject up to age 7, an effect that is known as the Delay of Principle B Effect. Such language acquisition difficulties are most often investigated through observations or experiments with children. Here we present a novel approach and examine whether remnants of the Delay of Principle B Effect that children show can be found in adults’ processing using pupil dilation measurements. Our results indicate a difference between Dutch and German adults’ processing of object pronouns: for Dutch adults, pronouns seem to be more effortful to process than reflexives, whereas for German adults this does not seem to be the case. We argue that looking at processing effort using pupil dilation measurements in sentences for which adults show ceiling level performance in offline comprehension may reveal new insights into children’s acquisition difficulties in the case of the Delay of Principle B Effect. This is a promising approach with potential applications in (cross-linguistic) language acquisition research in general, and in pronoun processing research in particular.

¹ Corresponding author: Margreet Vogelzang, Department of Theoretical and Applied Linguistics, University of Cambridge, 9 West Road, Cambridge, CB3 9DP, United Kingdom, email: mv498@cam.ac.uk

1. Introduction

Referring expressions can occur in a number of forms, ranging from a simple pronoun to a short phrase. For the interpretation of a referring expression, a listener needs to determine what the expression refers to. An example of specific referring expressions are object pronouns such as *him* in *John saw him*. Children up to age seven are known to show problems with the interpretation of these types of pronouns in some languages (e.g., Chien & Wexler, 1990; Grimshaw & Rosen, 1990; Thornton & Wexler, 1999; Wexler & Chien, 1985 (English), Koster, 1993; Philip & Coopmans, 1996; Spenader, Smits, & Hendriks, 2009 (Dutch), Sigurjónsdóttir & Hyams, 1992 (Icelandic)), but not in others (see among many others Ruigendijk, 2008; Ruigendijk, Friedmann, Novogrodsky, & Balaban, 2010 (German), McKee, 1992 (Italian), Hestvik & Philip, 2000 (Norwegian)).

Such difficulties in children's language acquisition are most often investigated, naturally, through observations or experiments with children. One of the reasons for not investigating object pronoun comprehension in more detail with non-brain-damaged adults is that their performance is generally at ceiling (Baauw & Cuetos, 2003; Hendriks, Banga, Van Rij, Cannizzaro, & Hoeks, 2011; Van Rij, Hollebrandse, & Hendriks, 2016) and it is generally assumed that non-brain-damaged adults don't have any difficulties interpreting object pronouns in simple transitive sentences such as *John saw him*². In this paper, we take an alternative approach and examine whether remnants of the object pronoun interpretation difficulties that children show can be found in adults' processing with more sensitive methods, in our case pupil dilation measurements. If so, this would create the possibility for a new line of investigations into object pronouns, namely through examining adults' processing costs even when their offline comprehension performance is at ceiling level.

1.1. Pronoun interpretation: The problem

Chomsky (1981) formulated a set of rules that describe reference assignment of pronouns and reflexives, Principle A and B. These Principles of Binding Theory suggest that pronouns and reflexives have different functions and can be paraphrased in the following way:

² Note that people with agrammatic aphasia have been found to show a similar comprehension pattern for object pronouns as found in children (see, e.g., Grodzinsky et al., 1993 on English, Ruigendijk et al., 2006 on Dutch).

Principle A: a reflexive must be bound in the local domain.³

Principle B: a pronoun must be free (not bound) in the local domain.

English-speaking children show good comprehension of reflexives (Principle A, see (1)) around age 3;0 well before they acquire good comprehension of pronouns (Principle B) around age 6;6 (Chien & Wexler, 1990). Similar effects have been found for Dutch (Koster, 1993). Before acquiring correct, adult-like comprehension of pronouns, children often allow *him* in (2) to refer to the local subject *John*. In the literature, this is referred to as the Delay of Principle B Effect (DPBE)⁴.

(1) John_i saw himself_i.

(2) John_i saw him_j.

A lingering question in language acquisition is why children in some languages like Dutch and English show a DPBE, whereas children in other, very similar, languages like German and several Romance languages don't. Several explanations for the DPBE have been put forward, some of which account for the observed cross-linguistic differences based on pragmatic rules (e.g., Grodzinsky & Reinhart, 1993; Reinhart, 2004) or grammatical constraints in combination with discourse properties (e.g., Hendriks, 2014; Hendriks & Spenader, 2006).

It has been suggested that the difference between children's errors in different languages may be explained by the different referential behavior of pronouns in these languages (Ruigendijk, 2008; Ruigendijk et al., 2010). For example, in Dutch and English both a reflexive *zich/himself* and a pronoun *hem/him* can be used in a locative PP like (3) to refer back to the sentential subject *de man/the man*. Thus, whereas the Principles of Binding suggest that reflexives and pronouns should show complementary

³ A reflexive must be bound by a c-commanding antecedent in the local domain. In the experiment described in this paper, we can take the local domain as being the clause. For a more precise definition, see Chomsky (1981).

⁴ It is debated whether the errors that children make are really due to problems with Principle B (Grodzinsky & Reinhart, 1993). A further discussion of this is outside the scope of the current paper (but see, e.g., Hamann, 2011), so we will use the term Delay of Principle B Effect to refer to the described object pronoun interpretation problems that children show.

distribution⁵ (Reuland & Everaert, 2001, p. 641), there are examples in Dutch and English in which there is no complementarity. In German, however, the object pronoun *ihn* in (3) cannot refer back to the sentential subject *der Mann*. Only the reflexive *sich* can be used in this function, and therefore complementarity of reflexives and pronouns is observed⁶.

- | | |
|---|-----------|
| (3) De man _i legt het boek naast zich _i /hem _i neer. | (Dutch) |
| Der Mann _i legt das Buch neben sich _i /*ihn _i . | (German) |
| <i>The man_i puts the book next to himself/him_i.</i> | (English) |

Thus, Dutch and English on the one hand and German on the other seem to differ with respect to the complementarity of pronouns and reflexives. The result of this partial lack of complementarity is that object pronouns in Dutch and English are functionally more ambiguous than in German. What is meant with this is that the reference assignment of Dutch and English pronouns is not only based on a structural rule (i.e. ‘pronouns cannot bind locally’, Principle B), but also on discourse rules (e.g., Rule I, which, slightly simplified, states that coreference for a pronoun is not allowed when it would yield an interpretation indistinguishable from that of a reflexive, Grodzinsky & Reinhart, 1993; see for further discussion Ruigendijk & Schumacher, 2020; Ruigendijk, Vogelzang, Schouwenaars, & Hendriks, submitted). Indeed, discourse manipulations have been found to influence children’s processing and interpretation of object pronouns in Dutch (Spenader et al., 2009; Van Rij et al., 2016). In contrast, reference assignment of German object pronouns should be more reliably based on the structural rule of Principle B. This could then also be the explanation of cross-linguistic differences observed in language acquisition: a language with a pronominal system that is functionally more ambiguous in the sense that the distribution of pronouns and reflexives is not always complementary will lead to slower acquisition of object pronoun comprehension and hence the observed pattern of DPBE. Given 1) the observation that young children make errors in the interpretation of object pronouns, i.e. show the DPBE, in

⁵ Note that it was already clear during the time that the Binding Principles were formulated by Chomsky that a very strict complementarity between reflexives and pronouns would not hold, see for instance Ross’ (1982) discussion of sentences like ‘James Bond noticed the gun near him/himself’.

⁶ For a more precise analysis of these dependencies, we refer to Reuland (2001, 2011). In brief, his argumentation is that, in Dutch and English, in these sentences Principle B does not apply, since the pronoun is not in a coargument relationship with the subject. For the anaphor *zich* he assumes a long distance dependency.

Dutch and English, but not in German, and 2) the observation that object pronouns in Dutch and English are functionally more ambiguous than in German in some contexts, learning when a pronoun cannot bind locally and under what discourse restriction it may bind locally, i.e. deciding when there is complementarity and when not, may be difficult in Dutch and English, whereas German children arguably have less confusion with respect to complementarity and hence are not led to make errors in this sense.

Note that we are not arguing that German pronouns as such cannot be ambiguous, they can indeed. Take a sentence like (4). In (4a), the pronoun *ihn* can in principle refer to both *der Junge* and *den Opa*. With neutral intonation, based on parallelism, the preferred interpretation of the pronoun is the object of the first clause: *den Opa*. With contrastive stress, the interpretation can shift to the subject though, as is seen in other languages like English when object pronouns are stressed (Arnold, 1998; Reinhart, 1981). Importantly, in both interpretations the local interpretation of the object pronoun *ihn* is ruled out. In (4a) this is based on Principle B and also due to a mismatch in gender features (*ihn* being masculine and the local subject being neutral). In contrast, in (4b) the object pronoun is not ambiguous anymore, since there is only one non-local antecedent that *ihn* can refer to (again, the local interpretation being ruled out by Principle B).

(4) a. Der Junge wäscht den Opa, und danach wäscht das Mädchen ihn.
The boy washed the grandfather and after that the girl washed him.

b. Das Mädchen wäscht den Opa, und danach wäscht der Junge ihn.
The girl washed the grandfather and after that the boy washed him.

So, our point is that the structural rule of Principle B in German can more reliably rule out a local interpretation, and this is why discourse operations are not needed at this point.

1.2 Pronoun interpretation: A new approach

In research on pronoun acquisition, children's interpretations are traditionally compared to adults' interpretations. When the children's interpretations are found to be non-adult-like, it is then investigated at what age they do acquire adult-like interpretations. In addition, different sentence-, context-, or visual conditions can be used to investigate aspects that make the interpretation easier for children (e.g., slowed-down speech, Van Rij, Van Rij, & Hendriks, 2010 or a more coherent discourse, Spenader et al., 2009). An approach that is to our knowledge rarely used to investigate pronoun

processing and more specifically the DPBE, however, is to examine processing in adults, whose interpretation is at ceiling level. This approach has been successfully used in investigations of other phenomena. An example is the well-known subject-object asymmetry that refers to the fact that object-first sentences (object questions, object relatives, and topicalized sentences) are acquired later than subject-first sentences (e.g., Biran & Ruigendijk, 2015; Friedmann, Belletti, & Rizzi, 2009). The same asymmetry has been found in adult processing in the sense that more errors are made on object-first sentences than on subject-first sentences, and that processing times are longer for the object-first-sentences (e.g., Frazier & Clifton, 1989; Vogelzang, Thiel, Rosemann, Rieger, & Ruigendijk, 2020; Wingfield, Peelle, & Grossman, 2003).

A notable exception to the examination of adults' processing in relation to the DPBE is the study of Hendriks et al. (2011), which did explore adults' pronoun processing in order to better understand children's difficulties. They measured response times and gaze data to investigate object pronoun processing in Dutch. Their results indicate that adults' response times and gaze data show remnants of the DPBE, namely longer response times and slower effects in the gaze data for pronouns than for reflexives in classical DPBE sentences.

In a different study, Vogelzang, Hendriks, and Van Rijn (2016) investigated the processing of object pronouns compared to object reflexives in Dutch adults by means of pupil dilation. They too found potential remnants of the DPBE (although they do not interpret their effects as such), namely larger pupil dilation, reflecting more processing effort, for pronouns than for reflexives. Overall, Hendriks et al.'s (2011) and Vogelzang et al.'s (2016) studies show the potential of examining adults' processing to learn more about the DPBE in language acquisition and its cross-linguistically different patterns.

In this paper, we will use pupil dilation to examine whether this measure can also provide useful information about adults' pronoun processing with respect to the presence or absence of the DPBE in a language. Pupil dilation has been found to be a valuable tool in measuring processing effort during sentence processing (e.g., Beatty & Lucero-Wagoner, 2000; Engelhardt, Ferreira, & Patsenko, 2010; Hyönä, Tommola, & Alaja, 1995; Just & Carpenter, 1993; Scheepers & Crocker, 2004; Schmidtke, 2014; Zellin, Pannekamp, Toepel, & Van der Meer, 2011, for and overview see Zekveld, Koelewijn, & Kramer, 2018). Using this measure, we can examine *whether*

and *when* during sentence processing pronouns are more or less effortful to process than reflexives.

As a starting point we take the aforementioned work of Vogelzang et al. (2016), who found that pronouns were more effortful to process than reflexives in Dutch. We will replicate this study with German adults, in order to test whether German object pronouns, as compared to reflexives, cause no or less processing costs than has been found for Dutch. In essence, we ask whether German pronoun processing differs from Dutch pronoun processing. This follows from the previously discussed assumption that establishing object pronoun reference in German is argued to be more straightforward and therefore less ambiguous than in Dutch, which should make their processing less effortful. That is, if indeed the German pronominal system shows more complementarity, in a sense the division of labor between pronouns and reflexives is clear. When the parser encounters a pronoun in German, it unlikely refers to a local subject, and hence a referent outside the local domain is to be found. For a reflexive, the situation is clear as well: a reflexive needs to refer locally and cannot refer outside the local domain. This basically reflects Principle A and B. In languages like Dutch and English, however, a structural rule (Principle B) is not sufficient for finding the correct antecedent for an object pronoun. Instead, discourse always comes into play (for instance in the form of Rule I), and it has to be processed whether in the use of a certain object pronoun there is a context that would have allowed for a local interpretation. If not, then the local interpretation is ruled out, and an antecedent outside the local domain is needed. One can see that, in this case, on top of the structural rule, also discourse comes into play, and this comes at a cost, as is supported by processing data from Dutch (Hendriks et al., 2011).

Our main hypothesis is thus based on the fact that when pronouns and reflexives in principle only need a structural rule, German adults will show little or no increased processing effort when resolving a pronoun compared to a reflexive. Note that the purpose of this study is not to uncover the cause of the DPBE. Rather, it is a methodological endeavor to examine the differences between Dutch and German adults' processing of object pronouns, and to gather additional support for the presence of remnants of the DPBE in adults' online processing. Note also that we restrict ourselves to typical 'Principle B' contexts, that is, our sentences always contain one local antecedent and one non-local antecedent, and both could be an antecedent of the pronoun or reflexive based on their gender and number features. Of course, in sentences with more than one non-local antecedent,

discourse always plays a role in finding the correct antecedent for a pronoun, but this is not the focus of the current study.

2. Method

2.1 Participants

41 students (mean age 23, 12 men) from the University of Oldenburg (Germany) participated in the study. All participants were monolingual native speakers of German and had normal or corrected-to-normal vision and hearing. The ethics committee of the University of Oldenburg approved the study (approval number Drs. EK/2019/013) and written informed consent was obtained from all participants. Participants received monetary compensation for their participation.

2.2 Design and Materials

We replicated the study of Vogelzang et al. (2016) in German, presenting German adults with auditory mini-stories⁷ (see the example in 5) of which the last clause contained either a **full NP subject** or a **pronominal subject** and either a **reflexive object** or a **pronominal object**. The subject pronoun was included to keep the design parallel to that of the Dutch study, the comparison of an object pronoun to a reflexive is our comparison of interest (i.e. sentences with a subject pronoun can be considered fillers).

(5) a. Der Igel hat ein Baumhaus gebaut.

The hedgehog has built a tree house.

b. Letzte Woche Freitag lief der Igel mit dem Panda durch den Wald nach Hause,

Last Friday the hedgehog walked home with the panda through the forest,

c. Während der Igel **ihn** über den dunklen Pfad verfolgt hat. [Pronoun]
*while the hedgehog followed **him** along a dark trail.*

d. Während der Igel **sich** über den dunklen Pfad beeilt hat. [Reflexive]
*while the hedgehog hurried **himself** along a dark trail.*

⁷ The full list of German stimuli can be found at <https://uol.de/speech-music-lab/projekte>

We used German translations of the original Dutch stimuli of Vogelzang et al. (2016). Only masculine characters were used throughout the experiment, as the Dutch experiment used masculine object pronouns and German feminine and neuter object pronouns are indistinguishable from subject pronouns (i.e. the pronouns *sie* and *es* can be both subject and object). As some verbs that are transitive in Dutch take a dative object in German (e.g., to help/*helfen/helfen*) and only accusative objects were suitable for the experiment (as accusative and dative object pronouns have different forms in German, i.e. *ihn* vs. *ihm*), we used 8 different verbs per object type in the German experiment compared to 10 in the original Dutch experiment. 96 mini-stories were created in total. Different verbs per object type were used so that the verbs used in the reflexive and transitive constructions were typically used in these constructions and therefore no surprisal effects would occur. The stories were recorded with a female native German speaker and all pronouns were unstressed. The mini-stories were distributed over 8 different lists, so that each participant heard each story only once. After each story, a question was asked about one of the characters in the story (e.g., *who was being followed?*) to record participants' interpretations of the object pronouns and to ensure that participants kept paying attention throughout the experiment. Questions could ask for the interpretation of the subject NP or the object in sentences with a pronominal object, and of the subject NP in sentences with a reflexive object (e.g., *who hurried himself?*). The questions were recoded with a male native German speaker, to make the distinction between the mini-stories and the questions clear.

2.3 Procedure

Participants were tested in a sound-proof booth using an Eyelink Portable Duo eye-tracker, a headrest, a computer monitor, headphones, and a gamepad. Stories and questions were presented auditorily. The two possible responses to the questions (e.g., *the hedgehog* and *the panda* in (5)) were presented as pictures on the screen from before the onset of the story until after the question. Participants were instructed before the experiment to listen carefully and select their response to the questions using the left and right trigger button on the gamepad, which corresponded directly to the left and right picture on the screen. In between the trials, participants were explicitly instructed to blink, so that blinking would be kept to a minimum during the trials. Pupil size was recorded continuously during the stories.

Participants received a practice block of six trials before starting the main experiment. The main experiment consisted of 4 blocks of 24 stories each. In between these blocks, participants could take a short break. Before every

block the eye-tracker was (re)calibrated using a 5-point calibration. The complete experiment, including the instructions and practice block, took around 60 minutes.

2.4 Analysis

Pupil dilation information (area) was recorded during the stories at 2000 Hz. All data processing and analysis was done using the statistical analysis software R (version 3.6.2, R Core Team, 2019). Preprocessing consisted of removing blinks and other missing data from the pupil dilation data. Trials with more than 25% missing data in the time window ranging from 300 ms before the onset of the subject in the last clause of the last sentence (from now on: subject) until the question were excluded (5.8% of the trials). 3706 trials remained for analysis. Any missing data in the remaining trials were linearly interpolated. The data were subsequently downsampled to 20 Hz. The data were aligned to the onset of the subject, and baselined per participant per trial to a time window ranging from 100 ms before the onset of the subject until the onset of the subject. This is the same alignment as was used in the Dutch experiment of Vogelzang and colleagues (2016), which also investigated the subject manipulation. As the current experiment is only interested in the object manipulations, we evaluated whether the same qualitative effects were found when using alignment on the onset of the subject and alignment on the onset of the object. Since this was indeed the case, the data with alignment to the onset of the subject are reported here.

The dilation data were analyzed with a Generalized Additive Model (Wood, 2006) from the onset of the subject until the offset of the clause. Fixed effects for subject, object, and the interaction between subject and object were included in the model, as well as random slopes and intercepts for participants and items, an effect of trial order (centered), and an interaction between time within a trial and trial order. The full model output is presented in Appendix A. In addition, the average pupil dilation in the time window from the onset of the subject until the offset of the clause was calculated on a by-participant basis for both clauses with a pronoun and clauses with a reflexive. It was then tested whether the average pupil dilations following a pronoun and a reflexive differed with a two-sided paired-samples t-test.

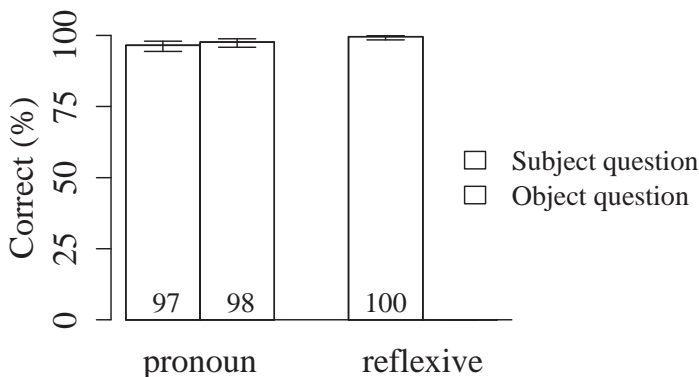
A comparison between the German data collected in the current study and the Dutch data of Vogelzang et al. (2016) was made based on the relative effects of the within-experiment comparisons between the processing of

pronouns and reflexives. As pupil dilation measures are easily influenced by lighting conditions and other factors unrelated to the content of the experiment, it would be inappropriate to perform a direct statistical comparison of data that were collected in different labs.

3. Results

Our adult participants answered > 95% correct in all critical conditions, and can thus be considered to be at or near ceiling-level comprehension for both clauses with pronouns and clauses with reflexives (see Figure 1). Following Van Rij et al. (2016), these ceiling-level responses were not analyzed any further.

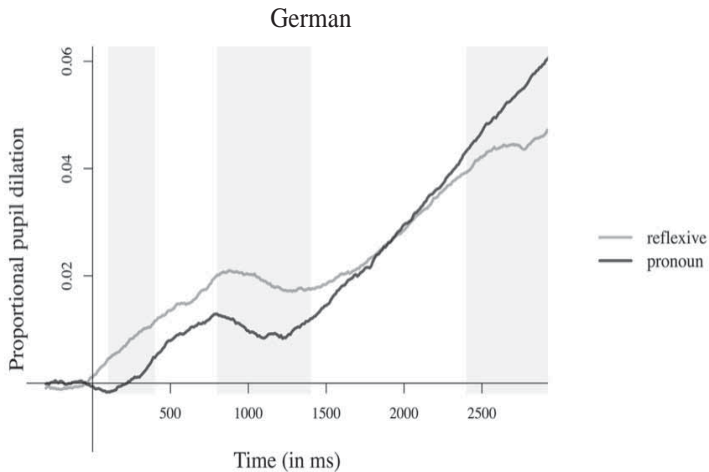
Figure 1. Percentages of correct responses to the comprehension questions. As expected, German adults show ceiling-level performance for both clauses with a pronominal object and with a reflexive object.



The pupil dilation analysis, however, reveals differences between the processing of pronouns and reflexives following a full NP ($\text{Edf} = 16.34$; $F = 19.77$; $p < 0.001$). Figure 2 shows the proportional pupil dilation in clauses with object pronouns and reflexives following a full NP in German. The graph shows that, initially (until around 1400 ms), clauses with a reflexive elicit a larger pupil dilation, whereas at the end of the clause (from around 2400 ms onwards), clauses with a pronoun start eliciting a larger pupil dilation than clauses with a reflexive. In the time window from 0 to 3000 ms, the average proportional pupil dilation in response to clauses with pronouns is 0.023 and the average proportional pupil dilation in response to clauses with reflexives is 0.025. These results indicate that when averaged over the whole clause, a pronoun does not elicit a larger pupil dilation than

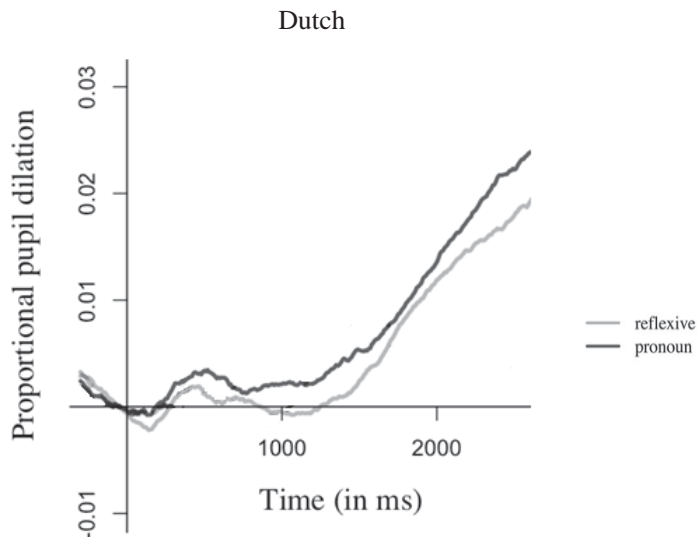
a reflexive in German ($t(38) = 0.52$; $p = 0.60$). Only at the end of the clause, more effort is required for processing a pronoun than processing a reflexive.

Figure 2. Proportional pupil dilation in response to clauses with a pronominal object and with a reflexive object for the German data of Vogelzang et al. (2016). Shaded areas indicate significant differences based on the Generalized Additive Model. Time is aligned to the onset of the subject in the last clause of the last sentence.



In contrast, the Dutch data of Vogelzang and colleagues (2016), which can be seen in Figure 3, indicate that pronouns are more effortful to process than reflexives throughout the final clause. Indeed, Vogelzang et al. (2016) found that reflexives were not more effortful to process than pronouns at any point during the sentence, whereas pronouns were statistically more effortful to process than reflexives from around 1100 ms onwards. For a detailed statistical analysis of these differences between the processing of pronouns and reflexives in Dutch adults we refer the reader to the original 2016 paper.

Figure 3. Proportional pupil dilation in response to clauses with a pronominal object and with a reflexive object for the the Dutch data of Vogelzang et al. (2016). Time is aligned to the onset of the subject in the last clause of the last sentence.



4. Discussion

In this study, we investigated the processing effort involved in resolving a pronoun compared to a reflexive in German adults. Our results show that especially in the first half of the clause, reflexives require more processing effort than pronouns in German. In the study of Vogelzang et al. (2016) in Dutch, notably, there was no point in the clause at which a reflexive was more effortful to process than a pronoun, as can be seen in Figure 3. Since reflexives can be integrated immediately, this initial effect in German could simply reflect normal syntactic processing, which may be postponed for pronouns (as pure syntactic processing cannot resolve the pronoun, it can only exclude the local reference assignment, in line with Reuland, 2001). Moreover, in the Dutch study, pronouns were more effortful to process than reflexives throughout the sentence, an effect that occurs much later, at around 2400 ms, in German. In Dutch, compared to German, it thus seems that the resolution of pronouns is more effortful early on, in line with the idea that the resolution of a Dutch object pronoun requires more operations as it is functionally more ambiguous. During sentence processing, this can

be seen as a larger increase in processing effort for pronouns compared to reflexives in Dutch compared to German.

Although increased processing effort for pronouns compared to reflexives in Dutch could be explained by the pronoun requiring retrieval of an antecedent that was mentioned in the previous clause, whereas the reflexive requires retrieval of the immediately preceding word, this explanation cannot account for the effects found in German. When averaged over the entire clause, the average proportional pupil dilation was not larger when hearing a pronoun than when hearing a reflexive in German. Thus, our results support the main hypothesis that German adults in comparison to Dutch adults show less increased processing effort when resolving a pronoun compared to a reflexive.

In light of the problems that we know that Dutch (but not German) children have with object pronoun interpretation, these findings are highly relevant. Although the findings do not (and did not aim to) resolve the question regarding the cause of the DPBE, they do show that children's errors and adults' processing difficulties may stem from the same source: adults' processing effects in Dutch could be a remnant of the DPBE. The results support the idea that German pronouns are functionally less ambiguous, in the sense that for a German sentence like (2) it is immediately clear that the pronoun cannot refer to the subject of the sentence and hence less additional processing costs arise. These findings could be a step towards explaining why Dutch and English children have more problems interpreting object pronouns than German children. As supported by the results obtained in this study, differences in the pronominal system with respect to the functional use of pronouns and reflexives (i.e. when can what type of pronominal be used to refer to a certain antecedent) could explain the acquisition difficulties in children (cf. Ruigendijk et al., 2010).

One prediction that follows from the study is that if less relative processing effort is required for adults to resolve pronouns in German than in Dutch, cognitive resources could influence pronoun processing. The idea that cognitive resources influence pronoun processing has indeed been discussed in the literature (e.g., Burkhardt, 2005; Hendriks, Koster, & Hoeks, 2014; Vogelzang, 2017). Limited cognitive resources have been argued to cause the comprehension problems with object pronouns in children (see, e.g., Avrutin, 1999) and people with agrammatic aphasia (see Grodzinsky, Wexler, Chien, Marakovitz, & Solomon, 1993; Ruigendijk, Vasić, & Avrutin, 2006). If we follow the assumption that discourse plays a more important role in Dutch pronoun resolution than in German, and that

discourse processing loads cognitive resources (Van Rij, Van Rijn, & Hendriks, 2013; Vogels, Krahmer, & Maes, 2015), that would further support the idea that cognitive resources could play a role in the DPBE.

Some suggestions for future research can be derived from this study. Firstly, it would be interesting to examine whether manipulations of context that have been found to decrease the DPBE for children in languages such as Dutch (e.g., Spenader et al., 2009) also decrease processing effort in adults as reflected in pupil dilation effects. Such manipulations of context were examined in the study of Hendriks et al. (2011), which did not find reflections of context type in the gaze data. Secondly, the current approach could be applied in many more languages in which children either do or don't show a DPBE. Interestingly, in a study with an unrelated research question, it was shown that in Italian (a language without a DPBE; McKee, 1992), the processing effort involved in processing object pronouns vs. reflexives is very similar to that found in the current German study (see Vogelzang, Foppolo, Guasti, Van Rijn, & Hendriks, 2019). Finally, the functional use of pronouns and reflexives and their complementarity or non-complementarity can be investigated more closely in future studies. Specifically, it would be very interesting to examine adults' processing of pronouns and reflexives in contexts in which complementarity is either observed or not, to determine whether and how this influences their processing.

Overall, the results indicate that examining adults' pronoun processing for sentences or structures that children have difficulties with is a promising approach, even when these adults' offline comprehension performance is at ceiling level. This offers many new possibilities for research in language acquisition in general, and for pronoun processing specifically: since adults are generally easier to come by and can perform lengthier experiments than children, it could make for more efficient research.

References

- Arnold, J. E. (1998). *Reference form and discourse patterns*. Ph.D dissertation. Stanford, CA: Stanford University.
- Avrutin, S. (1999). *Development of the Syntax-Discourse Interface*. Dordrecht: Kluwer Academic Publishers.
- Baauw, S., & Cuetos, F. (2003). The Interpretation of Pronouns in Spanish Language Acquisition and Breakdown: Evidence for the “Principle B Delay” as a Non-Unitary Phenomenon. *Language Acquisition*, 11(4), 219–275. https://doi.org/10.1207/s15327817la1104_2
- Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (pp. 142–162). Cambridge: Cambridge University Press.
- Biran, M., & Ruigendijk, E. (2015). Do case and gender information assist sentence comprehension and repetition for German- and Hebrew-speaking children? *Lingua*, 164, 215–238. <https://doi.org/10.1016/j.lingua.2015.06.012>
- Burkhardt, P. (2005). *The syntax-discourse interface: representing and interpreting dependency*. Amsterdam: John Benjamins.
- Chien, Y.-C., & Wexler, K. (1990). Children’s Knowledge of Locality Conditions in Binding as Evidence for the Modularity of Syntax and Pragmatics. *Language Acquisition*, 1(3), 225–295. https://doi.org/10.1207/s15327817la0103_2
- Chomsky, N. (1981). *Lectures on government and binding: the Pisa lectures*. Dordrecht: Foris Publications.
- Engelhardt, P. E., Ferreira, F., & Patsenko, E. G. (2010). Pupillometry reveals processing load during spoken language comprehension. *The Quarterly Journal of Experimental Psychology*, 63(4), 639–645. <https://doi.org/10.1080/17470210903469864>
- Frazier, L., & Clifton, C. (1989). Successive Cyclicity in the Grammar and the Parser. *Language and Cognitive Processes*, 4(2), 93–126. <https://doi.org/10.1080/01690968908406359>
- Friedmann, N., Belletti, A., & Rizzi, L. (2009). Relativized relatives: Types of intervention in the acquisition of A-bar dependencies. *Lingua*, 119(1), 67–88. <https://doi.org/10.1016/j.lingua.2008.09.002>
- Grimshaw, J., & Rosen, S. T. (1990). Obeying the Binding Theory (pp. 357–367). New York: Springer. https://doi.org/10.1007/978-94-011-3808-6_15
- Grodzinsky, Y., & Reinhart, T. (1993). The innateness of binding and the development of coreference. *Linguistic Inquiry*, 24(1), 69–101.

- Grodzinsky, Y., Wexler, K., Chien, Y. C., Marakovitz, S., & Solomon, J. (1993). The Breakdown of Binding Relations. *Brain and Language*, 45, 396–422. <https://doi.org/10.1006/brln.1993.1052>
- Hamann, C. (2011). Binding and Coreference: Views from Child Language. In J. De Villiers & T. Roeper (Eds.), *Handbook of Generative Approaches to Language Acquisition* (pp. 247–290). New York: Springer.
- Hendriks, P. (2014). *Asymmetries between Language Production and Comprehension*. New York: Springer. <https://doi.org/10.1007/978-94-007-6901-4>
- Hendriks, P., Banga, A., Van Rij, J., Cannizzaro, G., & Hoeks, J. (2011). Adults' on-line comprehension of object pronouns in discourse. In A. Grimm, A. Müller, C. Hamann, & E. Ruigendijk (Eds.), *Production-Comprehension Asymmetries in Child Language* (Vol. 31, pp. 193–216). Berlin: de Gruyter. <https://doi.org/10.1515/9783110259179.193>
- Hendriks, P., Koster, C., & Hoeks, J. (2014). Referential choice across the lifespan: Why children and elderly adults produce ambiguous pronouns. *Language and Cognitive Processes*, 29(4), 391–407. <https://doi.org/10.1080/01690965.2013.766356>
- Hendriks, P., & Spenader, J. (2006). When production precedes comprehension: An optimization approach to the acquisition of pronouns. *Language Acquisition*, 13(4), 319–348. https://doi.org/10.1207/s15327817la1304_3
- Hestvik, A., & Philip, W. (2000). Binding and Coreference in Norwegian Child Language. *Language Acquisition*, 8(3), 171–235. https://doi.org/10.1207/S15327817LA0803_1
- Hyönä, J., Tommola, J., & Alaja, A. M. (1995). Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *The Quarterly Journal of Experimental Psychology*, 48(3), 598–612.
- Just, M. A., & Carpenter, P. A. (1993). The intensity dimension of thought: Pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology*, 47(2), 310–339.
- Koster, C. (1993). *Errors in anaphora acquisition*. Ph.D dissertation, University of Groningen, the Netherlands.
- McKee, C. (1992). A Comparison of Pronouns and Anaphors in Italian and English Acquisition. *Language Acquisition*, 2(1), 21–54.
- Philip, W., & Coopmans, P. (1996). The Role of Lexical Feature Acquisition in the Development of Pronominal Anaphora. In W. Philip & F. Wijnen (Eds.), *Amsterdam Series on Child Language Development* (Vol. 5, pp. 73–106). Amsterdam: University of Amsterdam.

- R Core Team. (2019). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.r-project.org/>
- Reinhart, T. (1981). Pragmatics and linguistics: an analysis of sentence topics. *Philosophica*, 1(1), 53–94.
- Reinhart, T. (2004). The Processing Cost of Reference Set Computation: Acquisition of Stress Shift and Focus. *Language Acquisition*, 12(2), 109–155.
- Reuland, E. (2001). Primitives of binding. *Linguistic Inquiry*, 32(3), 439–492. <https://doi.org/10.1162/002438901750372522>
- Reuland, E. (2011). *Anaphora and language design*. Cambridge, MA: M.I.T. Press.
- Reuland, E., & Everaert, M. (2001). Deconstructing Binding. In M. Baltin & C. Collins (Eds.), *The Handbook of contemporary syntactic theory* (pp. 634–669). Oxford: Blackwell.
- Ross, J. R. (1982). Pronoun Deleting Processes in German. *Paper presented at the annual meeting of the Linguistic Society of America*. San Diego, CA.
- Ruigendijk, E. (2008). Pronoun interpretation in German kindergarten children. In A. Gavarró & M. J. Freitas (Eds.), *Proceedings of GALA* (pp. 370–380). Newcastle upon Tyne: Cambridge Scholars Publishing.
- Ruigendijk, E., Friedmann, N., Novogrodsky, R., & Balaban, N. (2010). Symmetry in comprehension and production of pronouns: A comparison of German and Hebrew. *Lingua*, 120(8), 1991–2005. <https://doi.org/10.1016/j.lingua.2010.02.009>
- Ruigendijk, E., & Schumacher, P. B. (2020). Variation in reference assignment processes: psycholinguistic evidence from Germanic languages. *The Journal of Comparative Germanic Linguistics*. <https://doi.org/10.1007/s10828-019-09112-x>
- Ruigendijk, E., Vasić, N., & Avrutin, S. (2006). Reference assignment: Using language breakdown to choose between theoretical approaches. *Brain and Language*, 96(3), 302–317. <https://doi.org/10.1016/j.bandl.2005.06.005>
- Ruigendijk, E., Vogelzang, M., Schouwenaars, A., & Hendriks, P. (submitted). Cross-linguistic differences in the effects of discourse on pronoun processing.
- Scheepers, C., & Crocker, M. W. (2004). Constituent order priming from reading to listening: A visual-world study. In M. Carreiras & C. Clifton (Eds.), *The On-line Study of Sentence Comprehension: Eye tracking, ERPs, and Beyond* (pp. 167–186). New York: Psychology Press.
- Schmidtke, J. (2014). Second language experience modulates word retrieval

- effort in bilinguals: Evidence from pupillometry. *Frontiers in Psychology*, 5(137). <https://doi.org/10.3389/fpsyg.2014.00137>
- Sigurjónsdóttir, S., & Hyams, N. (1992). Reflexivization and Logophoricity: Evidence From the Acquisition of Icelandic. *Language Acquisition*, 2(4), 359–413. https://doi.org/10.1207/s15327817la0204_5
- Spenader, J., Smits, E.-J., & Hendriks, P. (2009). Coherent discourse solves the pronoun interpretation problem. *Journal of Child Language*, 36(1), 23–52. <https://doi.org/10.1017/S0305000908008854>
- Thornton, R., & Wexler, K. (1999). *Principle B, VP Ellipsis, and Interpretation in Child Grammar*. Cambridge, MA: M.I.T. Press.
- Van Rij, J., Hollebrandse, B., & Hendriks, P. (2016). Children’s eye gaze reveals their use of discourse context in object pronoun resolution. In A. Holler & K. Suckow (Eds.), *Empirical Perspectives on Anaphora Resolution* (pp. 267–293). Berlin: De Gruyter.
- Van Rij, J., Van Rijn, H., & Hendriks, P. (2010). Cognitive architectures and language acquisition: A case study in pronoun comprehension. *Journal of Child Language*, 37(3), 731–766. <https://doi.org/10.1017/S0305000909990560>
- Van Rij, J., Van Rijn, H., & Hendriks, P. (2013). How WM load influences linguistic processing in adults: a computational model of pronoun interpretation in discourse. *Topics in Cognitive Science*, 5(3), 564–580. <https://doi.org/10.1111/tops.12029>
- Vogels, J., Krahmer, E., & Maes, A. (2015). How cognitive load influences speakers’ choice of referring expressions. *Cognitive Science*, 39(6), 1396–1418. <https://doi.org/10.1111/cogs.12205>
- Vogelzang, M. (2017). *Reference and cognition: Experimental and computational cognitive modeling studies on reference processing in Dutch and Italian*. Ph.D dissertation. Groningen: University of Groningen.
- Vogelzang, M., Foppolo, F., Guasti, M. T., Van Rijn, H., & Hendriks, P. (2020). Reasoning about alternative forms is costly: Comparing the processing of null and overt pronouns in Italian using pupillary responses. *Discourse Processes*, 57(2), 158–183. <https://doi.org/10.1080/0163853X.2019.1591127>
- Vogelzang, M., Hendriks, P., & Van Rijn, H. (2016). Pupillary responses reflect ambiguity resolution in pronoun processing. *Language, Cognition and Neuroscience*, 31(7), 876–885. <https://doi.org/10.1080/23273798.2016.1155718>
- Vogelzang, M., Thiel, C. M., Rosemann, S., Rieger, J. W., & Ruigendijk, E. (2020). Neural mechanisms underlying the processing of complex sentences: An fMRI study. *Neurobiology of Language*, 1(2), 226–248.

- https://doi.org/10.1162/nol_a_00011
- Wexler, K., & Chien, Y.-C. (1985). The development of lexical anaphoras and pronouns. *Papers and Reports on Child Language Development*, 24, 138-49.
- Wingfield, A., Peelle, J. E., & Grossman, M. (2003). Speech Rate and Syntactic Complexity as Multiplicative Factors in Speech Comprehension by Young and Older Adults. *Aging, Neuropsychology, and Cognition*, 10(4), 310–322.
- <https://doi.org/10.1076/anec.10.4.310.28974>
- Wood, S. N. (2006). *Generalized Additive Models: an introduction with R*. Boca Raton, FL: Chapman and Hall/CRC Press.
- <https://doi.org/10.18637/jss.v016.b03>
- Zekveld, A. A., Koelewijn, T., & Kramer, S. E. (2018). The pupil dilation response to auditory stimuli: Current state of knowledge. *Trends in Hearing*, 22, 1–25. <https://doi.org/10.1177/2331216518777174>
- Zellin, M., Pannekamp, A., Toepel, U., & Van der Meer, E. (2011). In the eye of the listener: pupil dilation elucidates discourse processing. *International Journal of Psychophysiology*, 81(3), 133–141.
- <https://doi.org/10.1016/j.ijpsycho.2011.05.009>

Acknowledgments

We would like to thank Andreas Hiemstra for his help running the experiment. The first author (MV) was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2177/1 - Project ID 390895286.

Appendix A

Table A1. Results of the Generalized Additive Model analysis of the pupil dilation data, with a baseline of full NP subjects and reflexive objects.

Parametric coefficients:				
Predictor	Estimate (β)	SE	<i>t</i>	<i>p</i>
Smooth term	0.050	0.005	9.261	< 0.001
Approximate significance of smooth terms:				
Smooth term	Edf	Ref.df	<i>F</i>	<i>p</i>
NP-reflexive baseline	17.801	18.652	57.34	< 0.001
Subject form: pronoun	16.885	18.851	12.75	< 0.001
Object form: pronoun	16.340	18.525	19.77	< 0.001
Subject form: pronoun x object form: pronoun	17.908	19.275	14.46	< 0.001
s(Trial)	8.776	8.983	101.44	< 0.001
s(Participant, Time)	39.676	40.000	374.07	< 0.001
s(Item, Time)	93.207	95.000	66.42	< 0.001
ti(Time, Trial)	72.778	79.506	31.86	< 0.001