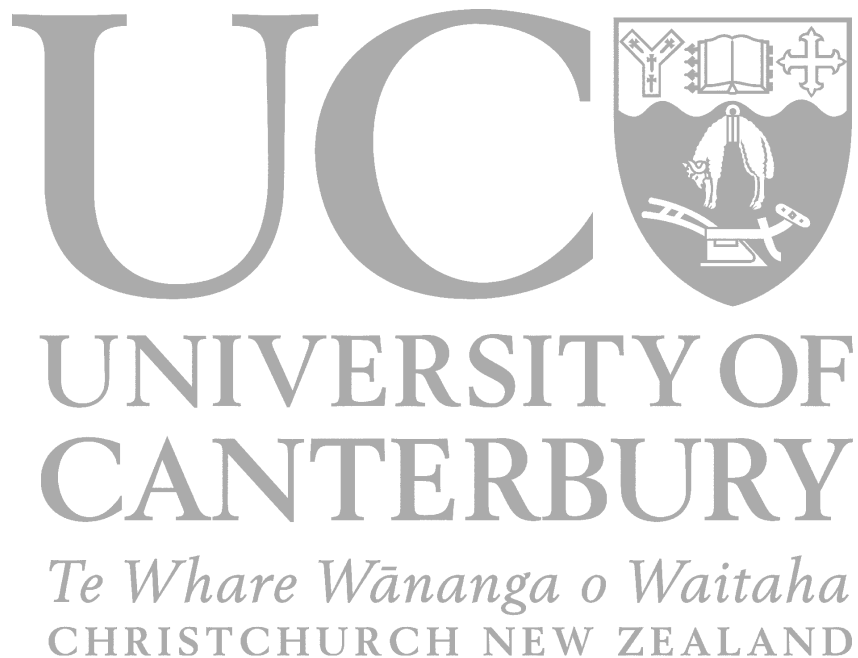

The acquisition of sentence alternations

How children understand and use the English dative alternation

Daniel Matthias Bürkle



A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Linguistics
in the Department of Linguistics, University of Canterbury

2015

Contents

Acknowledgements	1
Abstract	3
Abbreviations	4
List of Tables	5
List of Figures	6
1 Introduction	8
2 Background	10
2.1 The dative alternation in English	10
2.2 Other alternations and reordering phenomena in English	18
2.3 Animacy	26
2.4 Number	35
2.5 Length	37
2.6 Other features	47
2.7 First language acquisition	55
3 Research questions	63
3.1 Is there any evidence for an animacy ordering preference?	63
3.2 Is there any evidence for a number ordering preference?	63
3.3 Is there any evidence for a length ordering preference with a minimal length difference?	64
3.4 Do these preferences emerge in a particular order?	64
3.5 Does touchscreen input reflect attention?	65
4 Experiment 1: Categorization	66
4.1 Methodology	66
4.1.1 Participants	66
4.1.2 Procedure	67
4.1.3 Limitations	68
4.2 Results	69
4.3 Summary	70
5 Experiment 2: Reaction and completion	72
5.1 Methodology	72
5.1.1 Participants	72
5.1.2 Procedure	72
5.1.3 Stimuli	75
5.1.4 Limitations	76
5.1.4.1 Pilot study	82
5.2 Analysis	84
5.2.1 Touchscreens	84

5.2.2	Eyetracker setups	85
5.2.3	Correcting systematic errors in eye gaze data	88
5.2.4	Analyzing eye gaze data	90
5.2.4.1	Time windows	93
5.2.4.2	Growth curve analysis (GCA)	94
5.2.4.3	Smoothing spline analysis of variance (SSANOVA)	100
5.2.4.4	Comparing these methods	104
5.2.4.5	Summary	118
5.2.5	Regression modelling	119
5.2.5.1	Numeric independent variable	119
5.2.5.2	Categorical independent variable	123
5.2.5.3	Multiple variables and their interactions	125
5.2.5.4	Cross-validation	127
5.2.5.5	A word on mixed-effects modelling	130
5.2.6	The multiple comparisons problem	130
5.3	Results	133
5.3.1	Reaction times	134
5.3.2	Choices	135
5.3.3	Touch input	147
5.3.4	Eye gaze	151
5.3.5	Gaze and touch	168
5.4	Summary	171
6	Experiment 3: Elicitation	174
6.1	Methodology	174
6.1.1	Participants	174
6.1.2	Procedure	174
6.1.3	Limitations	177
6.2	Results	178
6.2.1	Construction	179
6.2.2	Accurate reproduction	183
6.2.3	Reproduction initiation time	189
6.2.4	Disfluencies	191
6.2.5	Oddness reactions	196
6.3	Summary	197
7	Discussion	200
7.1	Is there any evidence for an animacy ordering preference?	200
7.2	Is there any evidence for a number ordering preference?	202
7.3	Is there any evidence for a length ordering preference with a minimal length difference?	204
7.4	Do these preferences emerge in a particular order?	205
7.5	Does touchscreen input reflect attention?	206
8	Conclusion	207
8.1	Further research	210

9	References	211
	Appendices	229
A	Human ethics	229
B	Blocks and trials in experiment 2	236
C	Images used in experiment 2	238
D	Sound files used in experiment 2	255
E	Sentences and drawings used in experiment 3	256
F	Information sheets and consent forms	265
G	Statistical computation in R	276

Acknowledgements

Heidi Quinn, Susan Foster-Cohen, and Jen Hay supervised this thesis. I don't think I could overstate the extent and quality of their support, even in the "gosh isn't everyone awesome" language that is typical of Acknowledgements. They were very receptive to my initial (bad) thesis proposal and did everything in their power to get me to Christchurch. Not content with achieving that, they treated my thinking as on their level (even though it wasn't, and isn't) and engaged as much with my ideas as I did. That made quick work of my silly preconceptions about academia generally as well as about linguistic concepts specifically, and I was left with a well-founded and feasible research project. Heidi, my Senior Supervisor, helped me with just about everything, even the smallest details, like she does for all her students. I was particularly lucky—with her office just across the hall from mine, she always was (or seemed) happy to give me 'a few' minutes or to read yet another draft. Jen was in charge of the big picture and, leading by example, showed me how to structure writing as well as time (and found the first research participant for the project!). Susan made connections: how does this finding relate to that paper? and have I heard of these books? Heidi, Susan, Jen: thank you for making it interesting.

The Department of Linguistics at the University of Canterbury and the New Zealand Institute of Language, Brain and Behaviour (the continuation of the Department by others' means) are a great place anyway. Jacqui Nokes, Beth Hume, Lynn Clark, Kevin Watson, Vica Papp, Kota Hattori, Simon Todd, Peter Racz, Donald Derrick, Tom de Rybel, Megan McAuliffe, Thomas Klee, Stephanie Stokes, and Lucy Johnston listened to my ideas, helped me improve them, and kindly shared their own ideas with me. This and all the other things they do to make the Department and Institute a community of researchers have left their mark on this thesis and on me. Morgana Mountfort-Davies, Scott Lloyd, and Emma Parnell deserve special thanks for all their practical help on top of that, as does my research assistant, the incredibly patient, ever-diligent, and generally wonderful Ailsa Walker. The School of Language, Social and Political Sciences at Canterbury forms a community as well, not least thanks to Beth Hume's efforts—and Ghislaine Lewis, Khin-Wee Chen, Martina Wengenmeir, Ana Yuchshenko, and Viviana Cedeño Bustos are also awesome. These communities also helped make the many and illustrious visitors feel at home here—I am particularly grateful for the opportunity to meet and talk to Joan Bresnan, Andy Wedel, Kathleen Currie Hall, Chigusa Kurumada, Graeme Trousdale, Florian Jaeger, Jeremy Needle, Peter Culicover, Richard Kayne, Dave Kleinschmidt, Cathi Best, and Victor Kuperman. Janet Grijzenhout and Josef Bayer encouraged me to pursue further study after an MA, and I must thank them for that. I acknowledge the financial and equipment-al support of the University of Canterbury as well as the School of Languages, Cultures, and Linguistics (later Social and Political Sciences, later still Language, Social and Political Sciences) and the New Zealand Institute of Language, Brain and Behaviour. This includes generous travel funding, which allowed me to attend overseas conferences in 2014 and 2015 as well as visit the University of California, Merced (enjoying Stephanie Shih and Bodo Winter's hospitality and input there) and the University of Arizona (where I was welcomed particularly by Andy Wedel, Ryan Smith, Maureen Hoffmann, Jorge Muriel, and Louise St. Amour). I am so grateful to all participants in the experiments that make up this thesis for their time and patience, and to Dunsandel School, Team Tamariki, the University of Canterbury Early Childhood Learning Centre, the Ilam and Montana Early Learning Centres, students in LING101 in 2014 and 2015 as well as LING103

in 2014, Jihyun Lee, Marrenna Berry, CASPA, Riccarton Park Montessori Preschool, Oaklands School, St Martins Primary, Anthony Rimmell of Riccarton Baptist Church, Sam Jarman, and the Canterbury Home Educators network for helping me find research participants in the first place. Elena Moltchanova and Daniel Gerhard of the University of Canterbury's Department of Mathematics and Statistics were a great help with the background on some of the statistical methods I use in this thesis. Much of the work for this thesis was done with free software. Jon Peirce and Sol Simpson of the PsychoPy project graciously helped get my experiment script to run, and deserve my sincere thanks as well as more users (seriously, if you suffer from E-Prime, go to psychopy.org today). I am also grateful to the teams behind R, L^AT_EX, and Python as well as the StackExchange geniuses that speak both Programmer and User and never seem to tire of translating. To all of the above: thank you for making it possible.

Ksenia Gnevsheva, Darcy Rose, Ahmad Haider, Matthias Heyne, Xuan Wang, and Ryan Podlubny are the best people you will ever meet. My friends: thank you for everything.

Abstract

Many English verbs expressing transfer can be used in two different constructions, one with no preposition (*Rick gave Kate a coffee*) and one with the preposition *to* (*Rick gave a coffee to Kate*). Whenever speakers use such a verb, they have to choose between these two constructions. This choice is determined in part by some features of the two objects: all other things being equal, speakers are more likely to use whichever construction places a shorter object before a longer one (*Rick gave a coffee to the tall and well-dressed woman standing next to the desk at the southern side of the room*), an animate object before an inanimate one (*Rick gave Kate a coffee*), a plural object before a singular one (*Rick gave Kate and Roy an espresso machine*), and so on. This system of feature-based choices is established very well for adult language using language corpora and experiments, but there are fewer corpora and experimental studies of child language. Because of this dearth of data, it is unknown how children acquire this choice-making system: do they start making choices determined by only one of these features and add the others piecemeal, or do they learn the system wholesale and only tweak which features win out over others?

The three experiments in this thesis are a first step in answering this question. They are designed to map out the effects of length, animacy, and grammatical number on these choices over the course of typical first language acquisition. Because animacy is less stable a concept than length and number, the first experiment measures children's and adults' conceptions of animacy more indirectly. The second experiment presents the same participants with sentences using *give* where one of the two objects has been replaced by noise, and measures which of a constrained set of options they gaze at and which they choose to fill the noise gap. This provides measures of their expectations and preferences for the length, animacy, and number of the objects in these gaps. The third experiment has participants reproduce *give* sentences with different combinations of animacy, number, and construction. Participants reproduce sentences that conform to their choice-making system more easily.

The results of these three experiments show that children as young as four years already prefer the animate-before-inanimate order. The shorter-before-longer preference is not found in any age group when the difference in lengths is just one syllable. This evidence adds to a growing body of literature converging on the finding that choices in ordering phenomena are affected by the same features wherever these phenomena occur, throughout language acquisition as well as across languages. Data from the second experiment also substantiates the common assumption that touchscreen input and eye gaze are both closely linked to attention. This will allow researchers in the cognitive sciences to use touchscreens as an alternative to eyetracking more confidently.

Abbreviations

ANOVA analysis of variance

Appl applicative (Bruening 2010:289)

AppIP applicative phrase

ARCHER A Representative Corpus of Historical English Registers
(manchester.ac.uk/archer)

CHILDES Child Language Data Exchange System (childes.psy.cmu.edu)

CELEX Centre for Lexical Information database (celex.mpi.nl)

cm centimeter(s)

CP complementizer phrase (*because Kate said so*)

DQ dative question (*Whom did you send the woman?*; Langendoen et al. 1973)

EIC Early Immediate Constituents principle (Hawkins 1994)

GCA growth curve analysis (Mirman et al. 2008)

HNPS heavy noun phrase shift (*I saw on the ship* [_{NP} *some angry people who all needed a shower and a stiff drink*].)

Hz Hertz

ICE International Corpus of English (ice-corpora.net/ice)

LFG Lexical functional grammar

ms millisecond(s)

MSE mean square error ($\frac{\sum_{i=1}^n (\hat{Y}_i - Y_i)^2}{n}$)

NP noun phrase (*Rick, a coffee, me*)

NZD New Zealand Dollar

ONZE Origins of New Zealand English corpus project
(www.nzilbb.canterbury.ac.nz/onze.shtml)

OT Optimality Theory

P preposition

PP prepositional phrase (*to Rick, into the coffee, for me*)

Prt particle (*up in give up*)

s second(s)

SSANOVA smoothing spline ANOVA (Gu 2013)

SVO subject-verb-object (used to describe sentences with that order, like *Kate drank the coffee*, or languages that use mostly that order, like English)

UG Universal grammar

UR Universal Representation (Bod 2006)

V verb

VP verb phrase

List of Tables

4.1	Animacy categorizations by age group	70
4.2	Answers to motion question	70
4.3	Answers to play question	70
5.1	Combinations of animacy and length used in trial design	76
5.2	Training trials in the pilot study	83
5.3	Linear models based on orthogonal polynomials	99
5.4	Regression model for writing corpus frequency on morphological family size	119
5.5	Regression model of Dutch verb frequencies	122
5.6	Regression model of Dutch verb frequency and regularity	125
5.7	Regression model of Dutch verb frequency, regularity, and family size	127
5.8	Regression model of reaction times	135
5.9	Regression model of length of choices	143
5.10	Regression model of animacy of choices	145
5.11	Regression model of number of choices	146
5.12	Regression model of dragging path duration	150
5.13	Regression model of gaze position	170
6.1	Features of objects in target sentences with construction-changing reproduction	180
6.2	Regression model of construction used by participants	181
6.3	Regression model of accurate reproductions	187
6.4	Regression model of reproduction initiation times	190
6.5	Regression model of disfluencies	194
6.6	Regression model of oddness reactions	197
B.1	Trials in block 1 of experiment 2	236
B.2	Trials in block 2 of experiment 2	236
B.3	Trials in block 3 of experiment 2	237
B.4	Trials in block 4 of experiment 2	237

List of Figures

4.1	Animacy categorizations by age group	69
5.1	Initial state of trials	73
5.2	State of trials during instruction sentence	74
5.3	Interactive state of trials	74
5.4	Different eyetracker setups	86
5.5	Different eyetracker orientations	86
5.6	Participant and eyetracker view	87
5.7	Locations of pictures at beginning of trial in experiment 2	89
5.8	Density of gaze data from one participant, by block type	89
5.9	Correction of gaze data	91
5.10	Un-correct-able gaze data	92
5.11	First-, second-, and third-order polynomials	95
5.12	Linear combination of first- to third-order polynomials	96
5.13	The same first-, second-, and third-order polynomials for larger x	96
5.14	Orthogonal polynomials for two different ranges	97
5.15	Linear models based on orthogonal polynomials	99
5.16	Proportions of gazes	100
5.17	Proportion of gazes on one image only	101
5.18	Proportions of gazes with different smoothing splines	102
5.19	SSANOVA of gaze proportions	104
5.20	Subset of eyetrack data	105
5.21	Third-order GCA model of eyetrack subset	106
5.22	Seventh-order GCA model of eyetrack subset	107
5.23	SSANOVA model of eyetrack subset	108
5.24	Data generation for first set of simulations	110
5.25	First set of simulations	110
5.26	GCA models of first set of simulations	111
5.27	SSANOVA models of first set of simulations	111
5.28	SSANOVA false and true positives in first set of simulations	112
5.29	Second set of simulations	114
5.30	GCA models of second set of simulations	115
5.31	SSANOVA models of second set of simulations	115
5.32	SSANOVA models of second set of simulations, without extreme values	116
5.33	Cosine diagnostics of SSANOVA models of second set of simulations	117
5.34	Frequency and family size of Dutch verbs	120
5.35	Regression model of Dutch verb frequencies	123
5.36	Regression model of Dutch verb frequency and regularity	125
5.37	Regression model of Dutch verb frequency, regularity, and family size	128
5.38	Reaction times by age group	135
5.39	Choices by length	136
5.40	Choices by animacy	137
5.41	Choices by number	138
5.42	Choices by number	139
5.43	Choices by number	140

5.44	Choices by animacy	141
5.45	CELEX frequency of choices	147
5.46	Sinuosity of drag paths by age groups	148
5.47	Durations of drag paths by age groups	149
5.48	Proportional durations of drag paths by age groups	149
5.49	Durations of drag paths by age groups and experiment 1 answers	150
5.50	Percentages of gaze by response	153
5.51	Percentages of gaze by matching feature	154
5.52	Percentages of gaze by animacy, in animate-explicit trials	156
5.53	Percentages of gaze by length, in monosyllabic-explicit trials	157
5.54	Percentages of gaze by length, in bisyllabic-explicit trials	158
5.55	Age-of-acquisition norms	160
5.56	Percentages of gaze by number, in singular-explicit trials	161
5.57	Percentages of gaze by number, in plural-explicit trials	162
5.58	Percentages of gaze by matching feature, in plural-explicit trials	163
5.59	Percentages of gaze by animacy, in theme-gap blocks	165
5.60	Percentages of gaze by number, in theme-gap blocks	166
5.61	Percentages of gaze by length, in theme-gap blocks	167
5.62	Distances between touch and gaze, by age group	169
5.63	Position of gaze by age group	170
6.1	Stuffed toy used as ‘addressee’ in experiment 3	175
6.2	Drawing used to illustrate <i>Mom gave the cushions Anne.</i>	175
6.3	Number of sentences by dative constructions	179
6.4	Percentages of accurate reproductions of target sentences by age group	183
6.5	Accurate reproductions by age group and construction	184
6.6	Accurate reproductions by age group and number	185
6.7	Accurate reproductions by age group and animacy	185
6.8	Accurate reproductions by age group and animacy order	186
6.9	Reproduction initiation times by age group	189
6.10	Reproduction initiation time by age group and accuracy	191
6.11	Reproduction initiation time by age group and animacy	192
6.12	Percentages of disfluencies by age group	193
6.13	Disfluencies by age group and animacy	195
6.14	Disfluencies by age group and animacy pattern	195
6.15	Oddness reactions by age group	196
6.16	Oddness reactions by age group and number	198
6.17	Oddness reactions by age group and animacy	198

1 Introduction

Linguistic phenomena are rarely pure and never simple—even the most straightforward and elegant explanations have exceptions, unexplained underlying assumptions, and trouble accounting for their phenomena in variation and diachrony. While striving for elegance is not futile, dealing with untidy aspects can not be avoided completely. Moreover, the mere fact that there is so much untidiness suggests that it is worth studying: if the human language faculty is able to handle highly complex and varying systems, the complexity and variation will need to be understood before a full account of the language faculty is possible.

Gradient phenomena make good subjects for these studies. One of them, the English dative alternation (or dative shift), has been studied much in the past 15 years. It is the curious availability of two distinct constructions for many ditransitive verbs that express transfer of some sort. One construction, which I will call the **double object construction** here, consists of the recipient object, describing the goal or receiver of the transfer event, expressed as a noun phrase (NP)¹ followed by the theme object, describing the thing transferred, also expressed as an NP. An example of the double object construction is given in (1.1a). The other construction is the **prepositional construction**, where the theme object NP is followed by the recipient object expressed as a prepositional phrase (PP) headed by the preposition *to*. An example of the prepositional construction is given in (1.1b).

- (1.1) a. Rick gave Kate a coffee.
b. Rick gave a coffee to Kate.

Recent work on this construction suggests that speakers make highly complex choices between these two constructions whenever one or the other is used, and that the specific factors and how they are weighted differ across different varieties of English. However, the acquisition of this gradient phenomenon is severely understudied, and most of the few existing studies used corpus data. Due to the inherent limitations of corpora, these studies have not been successful in determining whether all the factors that seem to affect adult speakers' choices between the two construction also affect the choice in child language. There is no consensus on whether the animacy and grammatical number of the two objects, both known to affect the dative alternation choice in adults, are also at work in child language. The interactions between these different features as well as between any animacy effect and the most common pattern of animacy, namely an inanimate theme and an animate recipient, have also not been established. The best-established effect in adult and child language is that of the relative lengths of the two

¹For convenience, I will refer to phrases like *he* and *Kate* and *a coffee* as noun phrases rather than determiner phrases throughout this thesis.

objects: when one object is longer than the other, speakers appear to prefer the construction that places the shorter object before the longer one. However, the smallest length difference that causes this effect and the most appropriate measure of length are not known. Psycholinguistic experiments are necessary to answer these specific questions.

As technologies develop, the toolbox of psycholinguistics grows. Desktop computers allow reaction-time measurements with millisecond accuracy as well as very involved interactive or reactive designs that adapt to each participant's responses. Eye- and mouse-tracking allow psycholinguists to investigate participants' attention and processing. Touchscreen technology has overcome initial difficulties and is now very intuitive to use. This makes it an interesting option for psycholinguistics, as touchscreen interactions are more natural than mouse-based interactions and thus may make a better attention-measure, even approaching the usefulness of eyetracking.

Chapter 2 of this thesis discusses previous research on the English dative alternation as well as the acquisition of the cognitive concepts behind animacy, grammatical number, and length. The three experiments used here are described in Chapters 4, 5, and 6, and Section 5.2 discusses statistical and methodological challenges and how they were approached. The results of these three experiments, presented in Sections 4.2, 5.3, and 6.2, show that animacy does affect children's dative alternation choices, but number does not. A length difference of one syllable appears to be too little to trigger the length effect. Touchscreens as such emerge as a highly useful hardware for psycholinguistic research, but the experimental task has to be designed with users' likely behavior in mind. The results of experiment 3 show that even four-year-olds prefer the more common animacy patterns, although the expression of this preference changes with age. These results and their implications for our understanding of the dative alternation as well as for future research are discussed in Chapter 7.

2 Background

2.1 The dative alternation in English

As the term ‘dative’ itself suggests,² the dative alternation is observed with verbs of transfer or giving. The recipient in such a transfer event is often marked with the dative case cross-linguistically. The English object pronouns (*her/him/them*) are well-known examples of dative-marked forms, and some linguists have even defined ‘dative’ as the typical marking of such a recipient object (for example Lambert 2010). The verbs that participate in the dative alternation can be used in several distinct constructions, so the sentences that contain these verbs alternate between the different constructions. While there are dative alternations in languages all over the world,³ the phenomenon is “not very frequent in the world’s languages” (Malchukov et al. 2010a:18). The English dative alternation involves two distinct syntactic structures that order the two objects (theme and recipient) differently, but also differ in (at least) one other respect: the recipient–theme ordering, as in (2.1a) (repeated below for convenience), has both objects being realized as NPs, whereas the theme–recipient ordering in (2.1b) has the recipient as a PP with the preposition *to*.⁴ Because of this difference, the former is called the double object (dative) construction, and the latter is called the prepositional (dative) construction.

²Though *dative* itself is obviously Latinate, the term can be traced back to Greek writers: the grammar of Dionysios Thrax (170–90 BCE) speaks of the “πρωσεις . . . δοτική” or “ἐπισταλτική” (Davidson 1874:10 and tinyurl.com/greekThrax), that is the ‘case of giving’ or ‘sending (of letters)’. The former term is also used by Strabo (64 BCE–24 CE) (Jones 1950:254–255). Aulus Gellius quotes Nigidius Figulus (ca. 98–45 BCE) on the “*dandi casus*”, the ‘giving-case’, in Latin (Rolfe 1927:502); and Marcus Terentius Varro (116–27 BCE) uses the same term (Kent 1951:400–401, 548–549, and 582–583). The term *dativus* is used in the second century CE by the grammarians Velius Longus (Keil 1961:57) and Quintus Terentius Scaurus (Keil 1961:24; though see Law 1987 on the authorship and date of this grammar).

³Alternations like the English one discussed here have been claimed to exist in Greek (Cuervo 2003b), Spanish (Cuervo 2003b and (2003a)), Chinese (Liu 2006), Croatian (Zovko Dinković 2007), Dutch (Colleman and de Clerck 2009; see also van der Ziel 2012’s account of scope interactions in the Dutch and English dative alternations), Basque (Ormazabal and Romero 2010; though compare Oyharçabal 2010), and all Scandinavian languages (Anderssen et al. 2012, Kizach and Balling 2013); see Malchukov et al. (2010b) for more.

⁴Emonds (1976:81) argues that the double object construction also includes a preposition (heading the theme), as evidenced by *credit/furnish/supply/provide with*-constructions. This makes the dative alternation fit into his transformational theory quite elegantly, but it does not add any explanatory power. However, Emonds is right to argue that these verbs should be considered as part of the dative alternation—web searches with the pattern “(verb)ed her/him/them the” show that they **are** used in the ‘canonical’ double object construction:

- i it . . . credited him the point for finding it (tinyurl.com/creditedpoint)
- ii a bank error mistakenly credited him the money (tinyurl.com/creditedmoney)
- iii being away . . . furnished her the luxury of focusing totally on the music (tinyurl.com/furnishDO)
- iv Meet the man who killed Jim Morrison . . . or at least supplied him the heroin (tinyurl.com/supplyDO)
- v makeup artistry provided her the opportunity for something far greater than creative expression (tinyurl.com/provideDO)

- (2.1) a. Rick gave Kate a coffee.
 b. Rick gave a coffee to Kate.

In these simple examples, it would seem that both sentences have exactly the same meaning. On this basis, one could argue that the dative alternation is a purely syntactic phenomenon, since involves two different structures but no difference in meaning. However, an explicit interruption of the transfer event is more felicitous with the prepositional construction. *Give* in particular obscures this: its meaning specifies the transfer of possession, and therefore, as (2.2) shows, an interruption is infelicitous with *give* no matter which construction is used (see also Rappaport Hovav and Levin 2008). Other verbs make the possibility of an interruption in the prepositional construction more obvious: so-called verbs of ballistic transfer demonstrate this particularly well, as in (2.3). A typical *throw* event, for example, certainly specifies the thrower’s intention to hit a certain goal, but it also includes the possibility for the thrown entity to never get to that goal (the throw may be badly executed, a gust of wind may interfere, and so on).

- (2.2) a. * Rick gave Kate a coffee, but it didn’t reach her.
 b. * Rick gave a coffee to Kate, but it didn’t reach her.
 (2.3) a. ? Rick threw Kate a bagel, but it didn’t reach her.
 b. Rick threw a bagel to Kate, but it didn’t reach her.

It has been argued that (2.3b) is more acceptable than (2.3a) because of a subtle meaning difference between the two constructions: the double object construction encodes caused possession of an object by a recipient, whereas the prepositional construction encodes caused motion of an object towards a goal or recipient or an event which we know to be “normally sufficient to bring [the theme] into the sphere of [the recipient]’s physical control” (Oehrle 1976:129). The caused possession includes a caused motion event, at least with metaphorical motion, but it does specify the possession at the end of the motion path. Therefore, if the motion is interrupted and changed, the possession at its end is no longer possible, while the motion itself still exists, albeit in a changed way. Thus, the double object construction is ruled out for interrupted transfer events (see Gropen et al. 1989:241–242, Shimpi et al. 2007:1335, or Ogawa 2008).

However, the difference in acceptability between (2.3a) and (2.3b) that this analysis is partially based on is at best a tendency. There are ‘interrupted’ double object sentences in the literature that are claimed to be acceptable, such as (2.4).⁵

- (2.4) Max handed her a cigarette, but she wouldn’t take it. (Oehrle 1977:206)

⁵I presented sentences like (2.3a) to a few friends of mine, native speakers of New Zealand English all, who accepted them as ‘good English’ and understood the intended meaning perfectly.

The presence or absence of caused possession therefore does not necessarily determine which construction is used. Neither does the idiomatic form in idioms using transfer verbs: Bresnan et al. (2007:72), for example, list examples of *give the creeps/a headache to X*, even though the idiomatic form for these is the double object construction, of course. Bruening (2010) argues that these examples are not true prepositional constructions, but rather double object constructions with the recipient phrase moved rightward (and *to* inserted or made overt). He bases this account on two phenomena: the locative inversion and scope interactions. Prepositional constructions that are not based on idioms can be passivized, as in (2.5a), and the recipient PP can then be fronted by what Bruening calls locative inversion, as in (2.5b). Double object idioms themselves can also be passivized, as (2.6a) shows, but the locative inversion pattern with the prepositional construction, as in (2.6b), is not possible according to Bruening (2010).

- (2.5) a. Helicopters were given to the generals that lost the battle.
 b. To the generals that lost the battle were given helicopters. (Bruening 2010:297)
- (2.6) a. The generals that lost the battle were given hell.
 b. * To the generals that lost the battle were/was given hell.
 (both Bruening 2010:297, original judgments)

Ormazabal and Romero (2012) argue against this position by pointing out that the passive construction which (2.6b) would be derived from is not (2.6a), but rather (2.7).

- (2.7) * Hell was given to the generals that lost the battle.
 (Ormazabal and Romero 2012:458, original judgment)

This passive, which mirrors (2.5a) neatly, is ungrammatical according to Ormazabal and Romero (as well as Abeillé 1990:294) and can therefore not serve as a “source” (Ormazabal and Romero 2012:459) for the locative inversion. Since the passive for the idiomatic double object construction (2.6a) is grammatical, Bruening (2010:298)’s claim that the prepositional version of the idiom “patterns like the double object construction” is false in Ormazabal and Romero’s view.

However, the examples in (2.8)⁶ and (2.9) show that the non-inverted passives for (some of) these are in fact available.⁷

⁶Admittedly, (2.8b) is the only example I could find for *give the sack* that clearly intended the idiomatic meaning and was clearly written by a native speaker.

⁷At this point, I should explain why Kilgarriff (2007)’s criticism of using Google search results as linguistic data, though valid, does not apply in this case: Kilgarriff argues that web search engines should not be used as tools of corpus linguistics “[i]f the goal is to find frequencies or probabilities” (Kilgarriff 2007:147); my goal here is just to show that these (odd) constructions are in fact used, and are therefore available.

- (2.8) a. * The sack was given to him. (Ormazabal and Romero 2012:458, original judgment)
b. . . . if the sack was given to him by carrier pigeon. . .

(tinyurl.com/sackwasgiven)

- (2.9) a. * The moon was promised to him. (Ormazabal and Romero 2012:458, original judgment)
b. I didn't leave thinking the moon was promised to me

(tinyurl.com/moonwaspromised)

Ormazabal and Romero's argument on this point is therefore invalid. Their comments on Bruening's argument from scope stand, however: Bruening (2010:292–293) shows that the prepositional dative forms of double-object idioms like *give the sack* and *give a headache* behave unlike other prepositional constructions, but exactly like other double object constructions in that both (2.10a) and (2.10b) only allow the reading with the recipient having scope over the theme, meaning that one specific person (who is different from a previously discussed person) gets all the headaches. The other reading, where Charlie gets a cluster headache, Mitch gets a migraine, Tina gets a tension headache, and so on, is unavailable with both of these sentences.

- (2.10) a. This lighting gives a different person every kind of headache.
b. This lighting gives every kind of headache to a different person.

(both after Bruening 2010:294)

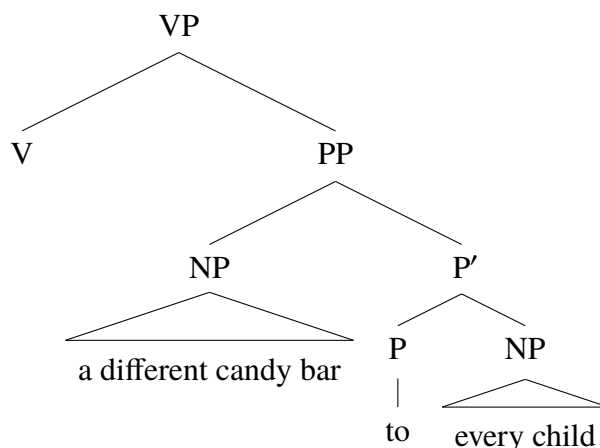
For the double object construction in (2.10a), this is as expected, since similar non-idiomatic sentences with the double object construction also allow only the one reading: (2.11a) can only mean that one specific child gets all the candy bars. The scope restriction for the prepositional construction in (2.10b), on the other hand, is surprising, since similar non-idiomatic sentences like (2.11b) do allow both the reading where one child gets all candy bars and the reading where Charlie gets a Curly-Wurly, Mitch gets a Mars bar, and Tina gets a Toblerone.⁸

- (2.11) a. I gave a different child every candy bar. (Bruening 2010:292)
b. I gave every candy bar to a different child. (Bruening 2010:293)

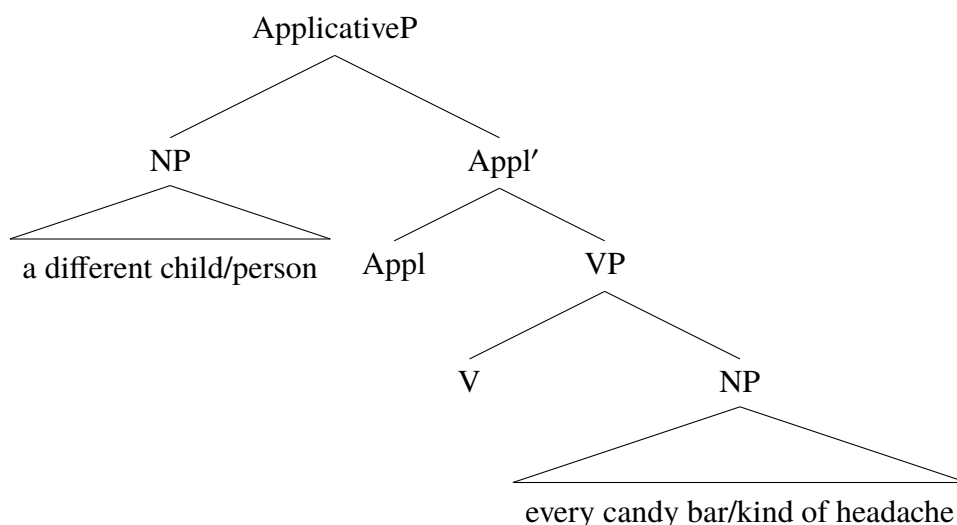
Since it behaves like other double object constructions in this respect, Bruening (2010:292–294) argues that the rare 'prepositional' idiom as in (2.10b) has the same underlying structure as other double object constructions. His account of the structure of the (non-idiomatic) prepositional construction is repeated in (2.12) below, and the structure of double object constructions (both idiomatic and non-idiomatic) is as in (2.13) (both Bruening 2010:289).

⁸It is interesting that children do not interpret scope in these examples like adults do at all (see van der Ziel 2012). This is not discussed further here because the present study does not touch on quantifier scope.

(2.12)



(2.13)

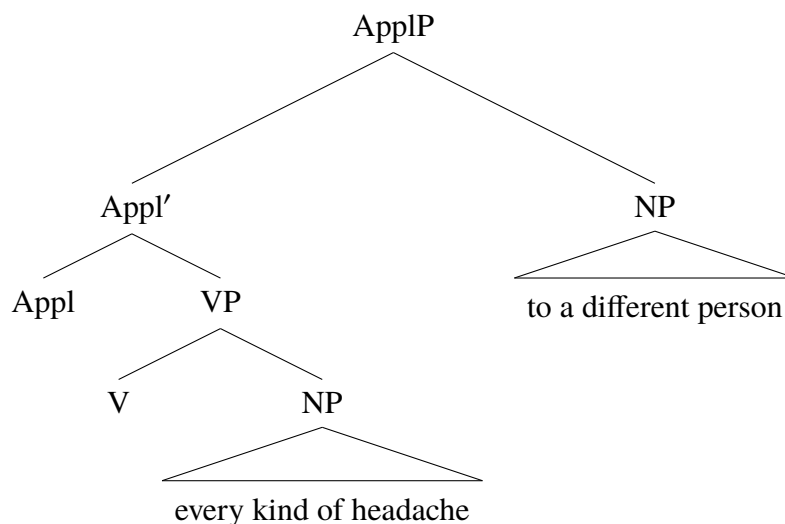


In the prepositional structure given in (2.12), both the recipient P' and the theme NP are at the same level and are thus both equally able to take scope. In the double-object (2.13), however, the recipient NP is at a higher level and thus always takes scope. Bruening then uses this structural difference to explain the availability of two readings in the prepositional construction, but of only one (the one where the indefinite on the recipient has scope over the the theme) in the double object construction.

Since the idiomatic 'prepositional' construction (2.10b) also allows only that one reading, it follows that this sentence has the same structure as in (2.13), but with the order of the immediate daughters of ApplP reversed to give the required theme-before-recipient order (Bruening 2010:294):⁹

⁹Despite the presence of the preposition *to*, Bruening (2010:294) specifically labels the recipient phrase (*to a different person* in (2.14)) as an NP in this structure he posits for the prepositional construction of double object idioms. Indeed, he has to: If this recipient phrase were a PP headed by *to*, the recipient would not c-command the theme; the point of Bruening's analysis of the scope phenomena (as he sees them) is precisely that the recipient **always** (and asymmetrically) c-commands the theme.

(2.14)



Ormazabal and Romero (2012:459) show that this account is “unnecessarily complex and ad-hoc”: the constraint that Bruening (2010) uses to limit overgeneralizations of the structure shown in (2.14) is specific to this kind of structure and apparently does not explain or simplify anything else; the preposition *to* has to be added to the recipient phrase by some unspecified process (a fact that Bruening 2010:291 acknowledges without addressing it) while keeping the recipient phrase an NP (so the recipient *c*-commands the theme; see Bruening 2010:294); and finally, even with the constraints used, the account is not sufficiently constrained and therefore generates both the non-idiomatic prepositional structure as in (2.12) and the idiom-derived structure as in (2.14) for many prepositional sentences, thus (wrongly) predicting the reading shown in (2.15), with *his* bound by *every worker*, to be available.

(2.15) ?? I gave his_i paycheck to every worker_i that came in on Friday.

(after Ormazabal and Romero 2012:460, original judgment)

Thus, while he points out an interesting scope asymmetry, Bruening (2010)’s account of double object idioms appearing in the prepositional construction is seriously flawed. There is no compelling reason to consider the prepositional versions of double object idioms to be anything other than prepositional dative constructions. As with the incomplete or interrupted transfer event sentences in (2.2) and (2.3) above, these scope interactions are interesting, but they do not support a conclusive transformational analysis of the dative alternation. There is little doubt that there are two possible readings of (2.11b) but only one of (2.11a). However, other differences seem to suggest a different analysis than the one given in Bruening (2010). In the absence of a comprehensive account of the meaning differences and derivational paths, I can only assume that there is no systematic difference in meaning between the two constructions. Properties of the verb and the two objects appear to have stronger and more systematic effects, as will be shown below. To avoid possible problems caused by any of the alleged meaning differences discussed

here, this study only uses *give*, which does not allow interruption of the transfer event (see (2.2) above), and there will be no quantifiers which could prompt different readings for adults and children (see van der Ziel 2012).

According to Wasow (2002:8), the meaning difference discussed above was the central focus of the work on the dative alternation until Hawkins (1994). The latter's principle of Early Immediate Constituents (EIC; Hawkins 1994:77) says that the heads of a phrase's immediate constituents will be kept as close as possible to the head of that larger phrase. In an SVO language with prepositions and right-branching phrases such as English, this leads to the well-known law of increasing constituents: placing a long head-initial phrase at the end means all the material after the head of this phrase does not come between the head and subordinate phrase heads (Hawkins 1994:211–212). For example, the subordinate heads *her* and *to* in (2.16a) are as close as possible to the phrase head *brought*, with only one word (*dog*) intervening between the three of them. The alternative ordering in (2.16b) sees as many as three¹⁰ words between the three heads. Therefore, the EIC principle predicts (2.16a) to be the preferred ordering for this combination of verb and postverbal elements.

- (2.16) a. Liz [**brought** [**her dog**] [**to** the school reunion]].
b. Liz [**brought** [**to** the school reunion] [**her dog**]]. (both Bürkle 2011:52)

In a left-branching, verb-last language such as Korean, the same principle explains the preference for longer phrases to be placed at the beginning: long-before-short ordering stops the non-head material of a head-final phrase from coming between that head and other heads (Hawkins 1994:211–212). While the EIC principle does explain a lot of crosslinguistic data, Wasow (2002:45–46) points out that it implicitly relies on utterances being planned out completely before ordering decisions are made and that this is not supported by the data. The corpus study in Bürkle (2011:34–50) shows that native speakers do in fact produce sentences violating this principle, and the explicit grammaticality judgment experiment in Bürkle (2011:51–59) shows that non-native speakers do not judge them as ungrammatical: the sentence in (2.17), for example, was judged as grammatical by 83.3% of participants.

- (2.17) Liz [brought [her trusty but old dog] [to the school]]. (Bürkle 2011:52)

Furthermore, unintuitive transformational accounts are required to allow the principle to account for certain structures (Wasow 2002:44). While he does mention several derivational possibilities, Hawkins (1994:212) merely assumes that the double object construction is derived from the

¹⁰The argument might be made that *school reunion* is one word, and there are thus only two non-head words between the heads on (2.16b). This does not affect the EIC principle's predictions here, as two intervening non-head words (*the* and *school-reunion*) is still more than the uncontroversially single intervening word *dog* in (2.16a).

prepositional construction in the English dative alternation. The opposite derivation has also been proposed: in saying that the prepositional construction is used “when the indirect object [the recipient, in the terminology used here] is significantly longer than the direct object [the theme]”, Sinclair et al. (1990:160) are implying that the prepositional construction is applied to a sentence where the ‘default’ double object construction is bad in some way (see also Berwick 1985:275–291). However, neither of these accounts is supported by conclusive evidence (Bürkle 2011:12–18), and including either derivational approach introduces inconsistencies into the theory of the EIC principle: Hawkins (1994:181) argues that the EIC principle explains the existence of particle verb constructions with the verb’s object before the particle (and thus a long element before a very short one; see (2.18a) for an example) because this Verb–NP_{object}–Particle ordering has become the basic construction for particle verbs, and the V–Prt–NP ordering (as in (2.18b)) is derived from it (licensed by the EIC principle).

(2.18) a. Rick [**turned** [**the** lights in his apartment] [**on**]].

b. Rick [**turned** [**on**] [**the** lights in his apartment]].

This derivation obviously is consistent with the EIC principle: the verbal head *turned* and the constituent heads *on* and *the* are much closer to each other in (2.18b) than in (2.18a). Hawkins shows that the V–Prt–NP ordering is more likely with longer NPs, which would support this derivational account since larger differences in constituent lengths (presumably) make the effect of the EIC principle ‘stronger’ (against the unchanging ‘strength’ of the basic construction). However, Wasow (2002:44) dismisses the idea that the basic (non-derived) ordering is automatically the preferred one and points out that Hawkins (1994) also explains the **basic**, non-derived, ordering of other constructions with the EIC principle. Thus, this derivational explanation using the EIC principle is internally inconsistent. **Not** including the derivational specifics, however, makes the principle much less predictive: for two phrases of equal length (30% of items in Bürkle 2011:36’s weight effect data from CHILDES corpora), neither ordering would be predicted to be preferred; yet we do find ordering preferences in such cases. Thus, while length differences do explain a lot of the dative alternation and similar weight effect phenomena (Wasow 2002:2 found that 80 to 90% of his corpus data observe the rule of “nondecreasing length”), they are not the only factor.

The particular verb used has been argued to be one of the other factors: most verbs that participate in the dative alternation are much more commonly used with one of the constructions than with the other. For example, it has been found that the majority of *give* sentences are double object constructions (76% in Gries and Stefanowitsch 2004:106 and Bürkle 2011:39), while the majority of sentences with directional *take* are prepositional constructions (84% in Gries and Stefanowitsch 2004:106, 58% in Bürkle 2011:39). These verb biases could be idiosyncratic

(Wasow 1997:101–102), or they could systematically reflect the meaning difference discussed above (transfer of possession versus spatial transfer) (Gries and Stefanowitsch 2004:104–107). Furthermore, it has been argued that the dative alternation (and other weight effects) are epiphenomena of prosodic preferences (Zec and Inkelas 1990). Previous (recent) use of one construction, even with different objects, has also been shown to increase the likelihood of that construction. Apart from these, the factors that have been claimed to affect the dative alternation are all features of the two objects: not only their length, as mentioned above, but also the animacy of their referents, their grammatical number, grammatical person, pronominality, and ‘givenness’ in the context (Wasow 2002, Bresnan et al. 2007, Bresnan and Hay 2008, Bresnan and Ford 2010, de Marneffe et al. 2012). These features are not independent from each other, of course: short words tend to be more frequent and less morphologically complex, pronouns generally refer to ‘given’ material and tend to be very short (McDonald et al. 1993), and animates are more likely to be represented with pronouns. However, Bresnan et al. (2007) showed that these correlations do not reduce to fewer features, which means that all of these individual features are taken into account: speakers prefer the construction that allows them to place a shorter object before a longer one, an animate object before an inanimate one, a plural object before a singular one, a pronoun before a non-pronominal object, and a discourse-given object before a discourse-nongiven one. I will discuss the effects of animacy, number, and length in separate sections below and also give a brief overview of the other features mentioned here. Before that, however, it is necessary to clarify how the dative alternation is defined in this study, particularly in relation to other alternations and ‘weight’-related ordering phenomena in English.

2.2 Other alternations and reordering phenomena in English

One reordering phenomenon that is often considered together with the dative alternation as defined above is the benefactive alternation (see for example Sinclair et al. 1990:160–161 and 407–408, Hawkins 1994:212, Whong-Barr and Schwartz 2002, Gries 2003, or Otsuka 2006; Szabóné Papp 2003:2 also considers them together, but notices several systematic differences (27–28, 32, and 65–68) and even splits up what she calls “TO-verbs” (87–131) and “FOR-verbs” (131–148) for detailed discussion). This is intuitively appealing: except for the verb and the preposition, the examples in (2.20) perfectly mirror those in (2.1) above, and the events described are quite similar. As (2.21) shows, there is even a preference for short-before-long order in the benefactive alternation, although it may be weaker than the preference in the dative alternation shown in (2.19) (see also Theijssen et al. 2009).

- (2.19) a. Rick gave Kate a grande skim latte with two pumps of sugar-free vanilla.
 b. [?] Rick gave a grande skim latte with two pumps of sugar-free vanilla to Kate.

- (2.20) a. Rick made Kate a coffee.
 b. Rick made a coffee for Kate.
- (2.21) a. Rick made Kate a grande skim latte with two pumps of sugar-free vanilla.
 b. ? Rick made a grande skim latte with two pumps of sugar-free vanilla for Kate.

However, the dative and benefactive alternations cannot be considered to be totally equivalent, because recipients (but not beneficiaries) are objects and because beneficiaries (but not recipients) can be added to either dative alternation construction. As for the first of these points, beneficiaries are not objects, at least not in the same way that dative recipients are (Nisbet 2005): beneficiaries can be easily omitted, but recipients cannot.

- (2.22) a. Rick made a coffee for Kate.
 b. Rick made a coffee.
- (2.23) a. Rick gave a coffee to Kate.
 b. * Rick gave a coffee.

Beneficiaries can also be easily added; the only restriction is that the event described include something being made available, created (in a broad meaning of the term), selected, or prepared (Oehrle 1976:110 and Kurniasih 2009:580). A beneficiary *for-PP*¹¹ can be added to intransitive (2.24), monotransitive (2.25), and ditransitive sentences (2.26).

- (2.24) a. Rick died.
 b. Rick died for Kate.
- (2.25) a. Rick made a coffee.
 b. Rick made a coffee for Kate.
- (2.26) a. Rick gave Kevin a coffee.
 b. Rick gave Kevin a coffee for Kate.

The ditransitive with additional beneficiary in (2.26b) illustrates the second point against equivalence for the dative and benefactive alternations: benefactives and datives can be combined

¹¹I am aware that the beneficiary NP is much more restricted: it is really only acceptable when it is added to a monotransitive, as in (ii). This is easily explained though, since the intransitive with beneficiary NP, as in (i), could be misinterpreted as a (mono-)transitive; and the ditransitive with beneficiary NP, as in (iii), presumably violates a rule against ‘tritransitivity’. The monotransitive thus is an exception, because the NP–NP frame that is compatible with the benefactive meaning happens to exist.

- i * Rick died Kate.
 ii Rick made Kate a coffee.
 iii * Rick gave Kate Kevin a coffee.

in one and the same clause. This combination is limited in that it is only possible with the beneficiary realized as a *for*-PP.¹² However, the two dative constructions can never be ‘combined’ like that at all: while the sentences in (2.27) could potentially be understood to have a meaning similar to that of (2.26b), they are clearly unacceptable.

- (2.27) a. * Rick gave Kevin a coffee to Kate.
b. * Rick gave a coffee to Kevin to Kate.

While a benefactive construction can be combined with a dative construction, a dative construction can never be combined with another (see also Szabóné Papp 2003:47 and 52–53).

Thus, in some crucial respects, the benefactive constructions are not like the dative constructions, and it follows that an in-depth study should not consider the benefactive alternation to be the same as the dative alternation. In this study, I will therefore use the term ‘dative alternation’ only for the alternation that uses *to* in the prepositional alternative and that is exemplified by the verb *give*.

While the dative alternation is thus defined more narrowly here than elsewhere, I do not approach it as an isolated phenomenon. Some of the evidence used throughout this study comes from work on other weight effect phenomena, such as the benefactive alternation or heavy noun phrase shift (HNPS). ‘Heaviness’ or length serves as a good example: it is evident that the length of the two objects has a noticeable effect on the choice of construction (see Section 2.5 for details). The well-described length effect in the dative alternation could be seen as nothing more than an instance of HNPS (Siewierska and Hollmann 2007), the effect of ‘newness’ of information as nothing more than an instance of information structure constraints that specify a given-before-new order (see Section 2.6 for details), and so on; the choice of construction would then be separate from these ordering decisions. The existence of the ‘reverse’ prepositional and double object constructions, as in (2.28), at first glance seems to support this idea: the dative alternation is often seen as an ordering decision between two alternatives, one of which (the double object construction) having recipient–theme order and the other (the prepositional construction) having theme–recipient order (and a preposition that has to be explained somehow). If, however, there are **two** recipient–theme orderings, one with the preposition (the ‘reversed’ prepositional as in (2.28a) below) and one without it (the canonical double object construction), the choice between these two cannot be explained as an ordering decision alone.

¹²Of course, *a pancake for Kate* in (2.25b) and *a coffee for Kate* in (2.26b) can also be one constituent (each), which makes them nothing more than complex theme arguments. In (2.26b), this reading means that the coffee itself was intended for *Kate* and that *Kevin* is to pass it on to her. However, *a coffee* and *for Kate* can also be constituents by themselves, and that reading of the sentence (to me) allows both the coffee being intended for *Kate* and the action of giving the coffee to *Kevin* being for her benefit or on her behalf or request (see also Oehrle 1976:108 for the distinction between the latter two readings). Thus, the *for*-PPs in (2.25b) and (2.26b) can be read as benefactives exactly like the *for*-PP in (2.24b).

- (2.28) a. . . . gave to Anastasia’s face a humor she herself did not possess.
(Biber et al. 1999:929)
- b. Rick gave to Kate a coffee [?](and to Kevin a a bagel).
- c. Rick gave it her. (after Haspelmath 2007:79)

However, these ‘reversed’ constructions appear to be much more restricted and therefore much rarer than the canonical prepositional and double object constructions (see also Siewierska and Hollmann 2007). The example in (2.28a) illustrates Biber et al. (1999:929)’s point that this construction is “very rare” and “characteristic of more formal writing” like academic prose or some styles of fiction. The same example also suggests a length effect (like HNPS) that is distinct from the (partially length-based) construction choice: the theme phrase *a humor she herself did not possess* is longer (and more complex) than the recipient phrase, which may well have allowed the rare reverse prepositional construction to occur here. In (2.28b), the parenthesis demonstrates how the reverse prepositional construction is more felicitous with a contrastive reading or list than without it. As for the reverse double object construction in (2.28c), Haspelmath (2007) says it is only possible for (some) speakers of British English, and only when both objects are pronouns (see also Sinclair et al. 1990:160 and Biber et al. 1999:929–930). Haspelmath explains this with harmonic alignment of the recipient role with a pronoun and of the theme role with a non-pronominal ‘full’ NP: recipients will typically be animate (or human) and given in the context, meaning they can typically be referred to using pronouns; themes will typically be inanimate and less specific, which makes full NPs necessary. The simple construction (which is the double object construction, in Haspelmath’s analysis) is only possible with a range of harmonic or near-harmonic alignments. For many varieties of English, this includes only the fully harmonic alignment and the near-harmonic full-NP recipient and full-NP theme combination; for (some varieties of) British English, the range includes these as well as the combination of pronominal recipient and pronominal theme. This applies to the double object construction in both the canonical and the ‘reversed’ order: the entire paper has “nothing to say about word order” (Haspelmath 2007:79, fn. 1). However, the examples he gives (repeated in (2.29) below) show that he only considers two different orderings to be possible for the double object construction with a pronominal theme—the reversed prepositional construction as in (2.28b) above is not mentioned as either a grammatical or an ungrammatical example. Interestingly, including it would not affect Haspelmath (2007)’s argument at all: the reordering of a possible construction is not connected to the possibility of that construction.

- (2.29) a. She gave me the book.
 b. She gave the book to me.
 c. She gave Kim the book.

- d. She gave the book to Kim.
- e. % She gave me it./She gave it me.¹³
- f. She gave it to me.
- g. * She gave Kim it./She gave it Kim.
- h. She gave it to Kim. (all Haspelmath 2007:79, emphasis removed)

This argument relies on the impossibility of the double object construction with a full-NP recipient and a pronominal theme (see also Wasow 1997:84). However, a few Google searches show that this combination is actually used (emphasis added in all examples).

- (2.30) a. you want me to draw this and give Emma it from you? (tinyurl.com/giveEmmait)
- b. I did give George it back(honest, I did) (tinyurl.com/giveGeorgeit)
- c. I gave my husband me. (tinyurl.com/gavemyhusband1)
- d. . . . and that is why God gave my husband me as his wife. (tinyurl.com/gavemyhusband2)
- e. “I’m going to give John me for Christmas!”¹⁴ (tinyurl.com/giveJohnme)
- f. . . . and give your wife yourself. . . ¹⁵ (tinyurl.com/giveyourwife)
- g. . . . but I want to give the lady it so my mom can have it. . . . (tinyurl.com/givetheladyit)

It is interesting to note that all but one of these sentences have the pronominal theme as the second postverbal element, but not the utterance-final one. Wasow 2002:8’s “prohibition against unstressed pronouns as the second object in the double object construction” could be amended to hold only if the second object is also the utterance-final element, so that an unstressed utterance-final element (the theme object pronoun) can be avoided by introducing further material (presumably with stress at the end). The one example above where the pronoun is utterance-final, (2.30c), would even support this: the pronoun *me* arguably is stressed there, which obviously means that this sentence could still occur even in the presence of a strict constraint against

¹³Haspelmath (2007) uses the percent symbol in (2.29e) to mean that these sentences are acceptable to some speakers and quite unacceptable to others.

¹⁴It is clear from the context of (2.30e) that a sexual meaning is intended, and the same probably goes for (2.30c), though there is less context there. This sexual undertone is not universal, but it could well mean that a sense of impropriety gives rise to the rare construction here. In any case, these counterexamples to the claims of Haspelmath (2007) and others evidently exist and must be accounted for. (Note also that Koster 1994 claims dative alternation sentences with non-reflexive pronouns to be categorically ungrammatical, which cannot be maintained in light of examples like (2.30c).)

¹⁵It appears that the reflexive theme pronoun is more common than the non-reflexive, particularly with the prepositional construction. Either way, double object sentences with NP themes and pronominal recipients are attested.

unstressed utterance-final elements. Furthermore, all of the examples in (2.30) are certainly somewhat rare, and many speakers may find them odd. However, none of these points affects the fact that these examples do show that a full-NP recipient and a pronominal theme can be used in the double object construction. Therefore, the argument of Haspelmath (2007) would have to be modified: he argues that the unusualness of a full-NP recipient together with a pronominal theme makes the double object construction with that alignment ungrammatical; all that can be maintained in the face of the examples in (2.30) is that this unusualness makes this combination of construction and alignment very rare and unusual.

Oehrle (1976:168)'s "surface filter" arguably has to be abandoned entirely in the face of (2.30). Briefly, it says that a dative construction can only have two preposition-less NP objects if the first one is not more 'prominent' than the second one, with the following scale of prominence: cliticized/reduced pronouns < *me* and *it* < *us* and *you* < other pronouns < "everything else". Thus, (2.31a) is permitted because the first object, *'im* ('him'), is a reduced pronoun and therefore less prominent than the second object, *it*; and (2.31b) is ungrammatical because the order is reversed and thus a more prominent object precedes a less prominent one. Conversely, (2.32a) is ungrammatical because the full pronoun *him* comes under "other pronouns" in the scale and is thus more prominent than *it*; therefore, the theme-recipient order as in (2.32b) is grammatical here.

- (2.31) a. I gave 'im it.
b. * I gave it 'im.

- (2.32) a. * I gave him it.
b. I gave it him.

(judgments after Oehrle 1976)

Because not every example of a reduced form will be 'marked' as *'im*, *'er*, and so on, it is hard to tell whether a pronoun is the full or the reduced form in writing or written speech, and those are the modes of languages that a web search will overwhelmingly find. In other words, examples like *I gave him it* would not be conclusive evidence against Oehrle (1976), since the *him* might well be a reduced form. However, the "everything else" level on his scale of prominence explicitly captures non-pronominal objects. This means that Oehrle is saying that (2.33b) is grammatical (or at least acceptable to his surface filter) and (2.33a) is not. However, (2.33a) is structurally equivalent to (2.30g) above, which is attested to occur.

- (2.33) a. * I gave the lady it.
b. I gave it the lady.

(judgments after Oehrle 1976)

There can be no confusion about reduced forms here—the predictions of Oehrle (1976) do not match up with the data.

Of course, Haspelmath's and Oehrle's approaches remain internally consistent, and they appear to explain other (crosslinguistic) data; my point here is that neither explains the English dative alternation well. No categorical restriction based on pronominality is apparent for the dative alternation. Also, (2.28a) and (2.28b) suggest that the reverse prepositional construction is much more likely in very specific situations where the prepositional construction is chosen, but other considerations (length and focus, respectively) strongly favor the theme phrase to be in the final position. It is therefore reasonable to go back to seeing the dative alternation as a choice between two orderings (the canonical constructions). What must then be explained is why the reverse constructions are found at all, but much more rarely than their canonical counterparts.

A general explanation for this is readily available: assuming the dative alternation is actually just an alternation between the two canonical orderings, and that HNPS (defined broadly, such that focus and properties other than length could also make a phrase 'heavy' and thus cause it to shift) is distinct from the dative alternation, HNPS could simply occur after the dative alternation choice is made.¹⁶ The two phenomena would follow the same general principles, though obviously with some differences in features or their relative strengths. Most of the dative alternation choices would then already be 'optimal' from the point of view of HNPS, so they would be produced in the canonical orderings. Only in a few cases, HNPS would still reorder the chosen dative construction, thus resulting in the rare reverse dative constructions (which Branigan et al. 2008:183 call "a 'heavy-shifted' prepositional object form", implying an analysis much like this). Van der Beek (2004)'s analysis of the four-way dative alternation found in Dutch divides that phenomenon into two binary alternations and thus assumes a theory like this one (though she argues that the two alternations are governed by different features), and her results show this approach to be fruitful.

This is a very rough explanation,¹⁷ obviously, but it is useful in explaining the approach to the dative alternation phenomenon taken here. A lot of effort has been expended on determining the exact rules that govern the grammaticality of the two dative alternation constructions, but most of these rules have been overturned. The syntactic rule shown to be questionable with (2.30) above demonstrates this, as do the phono-etymological categorization done in Berwick (1985:275–291) or the morphological rule proposed by Oehrle (1976:124). The former is an attempt to extend the well-known but inaccurate rule that English verbs of Germanic origin can

¹⁶Note that this is not an implicitly transformational account of the dative alternation—I am **not** saying that there is some rule that transforms the basic dative alternation construction into the nonbasic one and that that rule applies before the HNPS rule (which would go against Anderson 1977:158–160, for example). My proposal is that whatever process decides the dative alternation choice, its output is subject to HNPS.

¹⁷Future studies that investigate this this rough dative-alternation-plus-HNPS analysis are easily imagined. That analysis predicts that the reverse dative constructions are more likely to have a second object heavier than the first, and that the reverse constructions would allow (less heavy) non-object phrases to occur between the first object and the shifted second objects. These predictions could be tested with appropriate corpora.

be used with both dative alternation constructions, while Latinate verbs can only be used in the prepositional construction. Berwick (1985:289) adds Latinate verbs with initial stress to the group that can be used with both, using *promise* and *donate* as examples. While he notes that there are exceptions to this rule (*say*, which is not Latinate and so should be acceptable with both constructions, but is not; *design*, *assign*, and *award*, which are Latinate with final stress and so should not be acceptable with both constructions, but are), Berwick’s account does not address these exceptions in any way. If this handful of verbs were the only exception, that might be tolerable; they are not, however. Latinate verbs with non-initial stress like *convey*, *return*, *reveal*, and so forth are in fact used in the supposedly ungrammatical double object construction, as the examples in (2.34) show. If these (and all the other verbs like these, for which examples are easily found with web searches along the lines of “(verb)ed her the”) are all exceptions to the rule, there is hardly any rule left. Berwick (1985)’s theory, while elegant, is thus not compatible with the data.

- (2.34) a. . . . and that Appellant conveyed her the property in satisfaction thereof. [\(tinyurl.com/conveyDO\)](http://tinyurl.com/conveyDO)
- b. . . . so I returned them the item [\(tinyurl.com/returnDO\)](http://tinyurl.com/returnDO)
- c. . . . when I revealed him the truth [\(tinyurl.com/revealDO\)](http://tinyurl.com/revealDO)

Oehrle (1976)’s morphological rule fares no better. According to that, the double object construction is ungrammatical with verbs that contain Chomsky and Halle (1968:371)’s boundary =, which is “introduced by special rules”.¹⁸ Among these are *convey*, *return*, and *reveal*, for which double object examples are given in (2.34) above. Similar examples of the supposedly unavailable double object construction can be found easily for the rest of Oehrle (1976:137)’s ten alleged non-double object verbs as well. These verbs are probably much more frequently used with the prepositional construction than the double object alternative, but the double object alternative does not appear to be entirely ungrammatical. Even with the (erroneous) rule, “[m]any cases remain which we have nothing to say about” (Oehrle 1976:138). Any generalization about the ungrammaticality of one or the other dative construction with certain patterns is likely to either fall prey to that problem of limited applicability, or to be easily disproven (as above). Moreover, allegedly ungrammatical sentences may be perfectly parseable to speakers, as Langendoen et al. (1973) showed: they studied (supposedly) ungrammatical ‘dative questions’ (DQs) that ask after the recipient using the double object construction, such as (2.35).

¹⁸In a strong parallel to Berwick (1985), all the examples of the = boundary given by Chomsky and Halle (1968:94–96 and 371) as well as Oehrle (1976:137) are Latinate verbs. Oehrle (1976:126) even introduces a specific caveat to his constraint to “restrict the occurrence of the ‘=’ boundary to Latinate words with a prefix + stem structure”.

(2.35) Who(m) did you send the woman?

(Langendoen et al. 1973:463)

Pilot studies showed them that “subjects had no objections at all to DQs of any sort” (Langendoen et al. 1973:462), and only five of the 269 participants (1.9%) who saw the sentence in (2.35) in their questionnaire studies did not understand it at all (Langendoen et al. 1973:465).

Thus, the approach exemplified by Bresnan et al. (2007) and also taken here is to not categorically rule out any pattern and focus on the differences in usage instead. In that approach, it seems most fruitful to view the dative alternation as one of several ‘weight’-based (re-)ordering phenomena (like HNPS, particle verb ordering effects, and the like; see Hawkins 1994:214, Wasow 2002:4–12, and Anttila et al. 2010:971–972). The same general factors (length, givenness, animacy, and so on) affect all of them. This does suggest that there is a more basic underlying aspect of language that the various phenomena are really epiphenomena of. However, since there is little clear empirical evidence for that, this study will focus on the dative alternation involving the prototypical dative verb *give* and recipient phrases headed by *to* in the prepositional construction.

2.3 Animacy

It is common sense that the concept of animacy is important in linguistics as well as in learning generally: the sentence *The rock ran to the coffee shop* is odd because rocks, as inanimates, cannot run; and learning the names of different animals and telling them apart from other objects is a common theme of early childhood picture-books. This section discusses the concept of animacy and its linguistic effects. However, since the ‘animates’ and ‘inanimates’ categories are not simple ones, it is necessary to present and discuss different theories of categorization first.

In the classical view of categorization, each entity belongs to a category by virtue of its features matching those that define the category—a tree is a tree because it has a trunk, leafy branches, and a certain height, which agrees with the **concept** of trees also having a trunk, leafy branches, and a certain (minimum) height. A bush is not a tree because it will never grow to the height of trees, a lamppost is not a tree because it does not have branches or leaves, and so on. Categories can be hierarchical and have subcategories: the ‘animals’ category, for example, has the ‘birds’ and ‘humans’ categories among its subcategories. This follows from the Platonic method of division (Skemp 1952), which divides entities into categories by (recursively) grouping them according to their ‘natural’ differences: for example, animals live in the water or on land, the land-dwellers are quadru- or bipedal, and bipeds are feathered (birds) or featherless (humans). It was of course noted that even the most ‘natural’ categories defined by division only last until

a counterexample is found and thus are not actually natural or underlyingly true: the story of Diogenes the Cynic's plucked chicken running among the other featherless bipeds of the Platonic academy illustrates the problem of defining categories by a list of necessary features (Yonge 1853, Desmond 2008). Nevertheless, this classical view of a category as being defined exactly in terms of the common features of all its members was accepted as truth for centuries (Gries 2003:1–3, Qi et al. 2006:776, and Gabora et al. 2008).

One highly influential approach in another direction was Eleanor Rosch's theory of categories centered around prototypes (see for example Rosch et al. 1976), which has been applied in many fields (like geography, see Qi et al. 2006). Considering that the original experimental materials were lists of words for categories like furniture and vehicles (Rosch and Mervis 1975), it is unsurprising that the idea of prototypes has had a strong influence in linguistics particularly.¹⁹ Prototype theory brings with it the idea of degrees or scales of category membership, which shows what role category features play in this theory: the ability to fly is surely one feature of the 'birds' category, and kakapo, penguins, and all other flightless birds are less prototypical birds precisely because they are flightless. This example also shows that explicitly learned ('scientific') knowledge can affect categorization: the fact that some bird species cannot fly cannot possibly be a perceptually simple feature, but once it is learned, it undoubtedly affects category membership. Thus, both categories and features are developing in children's minds (Taylor 1995:241).

Prototype theory is not easy to reconcile with the idea of complex categories though: as Fodor (1998:93–108) points out, the prototypes for (most, if not all) complex or combined categories are not simply the combined prototypes of their respective constituent categories, and some complex categories (like 'non-cats') have no prototype at all. Proponents of prototype theory have argued that combinatorial prototypes can in fact make sense, given a 'syntax' of prototypes that takes all the features of the constituent categories, specifies relevant features of the complex prototype, and defaults to the feature specification of the (constituent) simplex prototypes for all non-specified features, but this does not work theoretically or practically (Connolly et al. 2007): at least in some contexts, a complex category has features which none of its constituent categories is specified for. Further iterations of prototype- or stereotype-based theories continue to address problems like these (Gabora et al. 2008), but the idea that categories (or prototypes, or stereotypes, or whatever terminology is used) have **features** is common to them all. Therefore, the above discussion of malleable categories and features is not affected by these more recent advances, and the fact that categories and their features can develop over time remains. The question then is this: how exactly **do** categories develop?²⁰

¹⁹In linguistics, prototypicality has been applied not only to words, but also to causatives (Gilquin 2008) and transitivity (Ibbotson and Tomasello 2009), for example.

²⁰Lupyan and Rakison (2006)'s finding that a neural network simulation **can** learn to categorize animacy even

Children’s non-adult-like use of prototype words is of particular interest to that question: “*doggie* is typically used by the young child to refer . . . to all small four-legged animals . . . his word roughly corresponds, in fact, to the superordinate term *animal* in adult language” (Taylor 1995:252). As Taylor argues, this suggests that children build (at least some) categories from prototypes. Not all features of the prototype will matter for the category—even at the age when cats are members of a ‘doggie’ category, young children will not expect them to bark. However, it does mean that entities that resemble the prototype in many features will be easier for children to categorize. Interestingly, this would be true for all prototype-based categories: while a large part of the literature on prototype theory is centered on ‘noun-y’ object categories, and it has been argued that infants’ first categories are of that type (Waxman 1999), it is crucial to note that prototypes have not been posited for these object category concepts only. Constructions may also have prototypes: to use an example that is relevant to this study, “the ditransitive construction prototypically involves transfer of possession . . . [and] prototypically has the form of NP1 + VERBditrans. + NP2 + NP3” (Ibbotson and Tomasello 2009:63).

However, it appears that children learn object/category labels of a certain level before they learn the sub- and superordinate category terms (Hammer et al. 2009). Although this finding is not absolutely inconsistent with prototype theory, some additional assumptions would be necessary to reconcile the two—after all, in the above example, the child’s term for small mammals could just as well be *poodle* or the name of the family dog.²¹ *Doggie* and other terms of this basic level are in all likelihood more frequent in child-directed speech than *poodle*, *small mammal*, and other sub- and superordinate terms, of course, but frequency and prototypicality do not necessarily correlate (Ibbotson and Tomasello 2009:65). Exemplar theory, on the other hand, readily explains this fact: considering that categorization is easier for children when they are presented with many similar examples of a category than with a comprehensive variety (Gentner and Namy 2006), the important within-category similarities for *doggie* and the category differences between *doggie* and *car* or *tree* (such as animacy, legs, barking, absence of wheels, or absence of leaves) are generally perceptual (visual, mostly) and thus more salient and easier to learn than those for *poodle* (versus *dachshund* and *poplar*) would be (Hammer et al. 2009).

The categorization experiment of Fisher (2009) supports this view: in that experiment, children aged three to five were shown a picture of a common object and two pictures of ‘similar’ objects, one of which being similar in appearance to the target but belonging to a different category and the other looking different but belonging to the same category. One trial, for example, had a white wall clock as the ‘pivot’ and a white dinner plate (similar appearance but different

with abstract, ‘meaningless’ features means that feat is computationally possible, but not necessarily that actual human cognition **does** develop like that.

²¹The fact that children’s understanding of these basic-level categories also precedes their use of the correct name for these categories is unproblematic (Taylor 1995:253).

category) and a grandfather clock (different appearance but same category) as the ‘similar’ pictures (Fisher 2009:1332). Half of the participants were asked to pick the object that was similar to the pivot in appearance, the other half the one that belonged to the same category. Halfway through the experiment, the instructions were changed: participants who had been asked to pick the similar-looking object were now asked to pick the category-matching one, and vice-versa. Participants who had to switch from perceptual (visual) matching to category matching did significantly worse in picking the correct choice after the switch than they had been before. Four- and five-year-old participants who switched from category to perceptual matching, however, did not show this drop in performance (Fisher 2009:1333). This suggests that perceptual information is easier to attend to and process, at least in this relatively confusing task—in other words, “perceptual information is more salient than conceptual information” (Fisher 2009:1334). Based on this result, Sloutsky and Fisher (2011) argue that conceptualizing and categorizing successfully means ignoring irrelevant features for more relevant ones, and that this process of intelligent ignoring has to develop for adult-like categories to arise.

This capacity for ignoring or filtering is a prime candidate for Hammer et al. (2009:117)’s cognitive limitations that prevent young children from learning subordinate categories: the relevant features for the *poodle* category, for example, must surely be all relevant features of *dog(gie)* **plus some more**, which makes *poodle* harder to learn when there is no filter to pick out the important features from all the irrelevant ones. This can be extended to subordinate categories in general. Lack of a filter would also explain why children under ten seem to learn categories better from within-category similarities while children over ten (and adults) learn better from between-category differences (Hammer et al. 2009:113), as well as why human categorization seems to use rules in situations of low confusability and exemplars in high confusability (Rouder and Ratcliff 2006). A given category has fewer possibilities for within-category similarities than for differences to other categories: ‘has four legs’ and ‘breathes’ are easier to pick out from other *poodle* similarities than ‘does not have leaves’ and ‘cannot fly’ are to pick out from the myriad features that differentiate *poodle* from other categories, though all four of these are arguably relevant for the category. Relying on within-category similarities thus is the economical thing to do when categorization is difficult, regardless of the particular reason for that difficulty.

These radical differences between (young) children and adults might be argued to make research into children’s concepts and particularly comparison with adults practically impossible—if children do not even have the capabilities required for certain kinds of conceptualization, the concepts that they do have could be very far removed from the corresponding adult concepts. However, the existing literature suggests that at least the concept of animacy does not undergo major restructuring from age four onward (Opfer and Gelman 2011:221–226). As four years is the youngest age group in the present study, this means that the difference between young

children's and adults' concepts of animacy is not a problem here.

One further possible cause for problems with research is the fact that children do not always respond like adults would to explicit categorization questions, particularly to questions regarding categories that have two possible applications: dolls, for example, can be used in “a play and non-play mode” (Gelman et al. 1983:108), and children realize that they are agentive and animate-like in the former but inanimate in the latter. What they do not always realize is which of these modes the experimenter ‘wants’, whereas adults seem more aware of the different possible interpretations (Subrahmanyam et al. 2003:365–366). Moreover, it is not always clear from just the data in which mode a particular answer was given. If such ambiguous items are to be used in research, the methodology must be designed and the results interpreted with particular care. Not doing so can invalidate a whole study—or even a program of research, as has been argued for Piagetian child psychology.

The Piagetian tradition of research into children's cognitive development focused on the idea of stages of development. In this view of development, a child is at one particular stage for a time, until further development advances them to the next stage. The ages of child participants who were found to be in a certain stage are then used to give an approximate age range for that stage in typical development. The ages for the last stage according to Piagetian research can be surprisingly high: for example, it has been claimed that children as old as nine fail to understand domino chains (Piaget 1978:22) or the difference between animals and inanimate objects (Laurendeau and Pinard 1962). This theory is relevant to linguistic studies since it has been argued that children are unable to say what they cannot conceptualize (Johnston 1985:Section 2)—in other words, if nine-year-olds cannot distinguish between animates and inanimates, their language use can hardly be affected by that distinction. However, Piagetian research consists in part of asking children quite abstract questions about experimental set-ups and comparing their answers to the physical facts: for example, Piaget (1978:196–210)'s “Mirrors” task asks children to manipulate the position of a pencil representing a beam of light reflected from a mirror, based on changes to a pencil representing the beam impinging on the mirror. Children aged four to eight are not competent at this task. Piaget is certainly right in concluding that these children do not know about the optical mechanism of reflection, but he goes too far in assuming that a simple mental model of angles of reflection is the only cognitive mechanism that allows intelligent use of mirrors (see Brown and DesForges 1979:51–52)—for example, thinking of mirrors as ‘reversers’ would also allow for some competence in using mirrors (wiping one's right cheek to remove a stain seen on the left cheek of one's mirror image, moving to the right to shift one's view in a stationary mirror to the left, and so on). It may be that competent (adult-like) use of mirrors relies on a cognitive mechanism of calculating, estimating, or visualizing angles of reflection, but it is not necessarily so. Piaget (1978), however, assumes

that it is and therefore tests to it, which means he cannot find competence based on another mechanism. Moreover, his method of testing for explicit knowledge means he may even miss competence in accordance with his preconception: asking children where a beam of light comes from, or asking them to use pencils as representations for beams of light, requires not only the angles-of-reflection model of mirrors, but also the ability to use abstractions. Even a child who knows about angles of reflection may not be able to express that knowledge in the confines of a Piagetian methodology, and may thus incorrectly be seen to be ignorant.

Thus, the experimental support for Piaget's theory relies on unvalidated preconceptions, and other interpretations of it may be possible (Brown and DesForges 1979:Chapters 3 and 4). It has even been argued "that **no** support for a stage view of cognitive development has been derived from the Neo-Piagetian training literature" (Brainerd 1973:366, emphasis added). More recent Neo-Piagetian theories have improved on certain aspects, based on new methods and data, but they have not gone unchallenged either (for an overview of recent Neo-Piagetians see Morra et al. 2008).

In conclusion, the literature suggests that studies of children's concept of animacy as well as comparisons with adult data **are** possible and possibly insightful, provided that methods are carefully chosen to be appropriate to the possibilities that the state of scientific knowledge leaves open as well as to participants' cognitive and meta-cognitive skills.

What the above also shows is that all theories of concepts or categories, in one way or another, incorporate the idea of concepts having features. It is therefore necessary to find the features of the concept at hand here, animacy, regardless of which theory of categorization one subscribes to. The features of the *tree* concept are fairly obvious, but animacy is less easily defined. Nevertheless, a few features emerge from the literature on this subject quite clearly. Firstly, animacy is of course connected to agentivity. Children's speech is likely to have animate entities as subjects or actors (Maratsos 1983). Infants as young as five weeks are more likely to produce "neutral" and "negative vocalizations" when faced with a person who smiles but remains passive and does not react to them, than when faced with the same person talking and reacting to them (Legerstee et al. 1987:88). This negativity towards passivity disappears when infants are faced with a doll: Legerstee et al. found no difference in vocalizations towards the doll depending on whether the doll was passive or appeared to talk and react to the child. Furthermore, infants smiled or vocalized less when faced with the doll than when faced with a person. This strongly suggests that people as such appear more active or agentive and are therefore more 'interesting' perceptually than everything else. According to Gentner and Boroditsky (2001), this perceptual bias (or ease of individuation in the perceptual stream) would also explain children's early lexicon being mostly made up of animate-referent nouns. One reason for humans and other animates being easier to pick out from the visual input is that they move in a special way: their

motion is goal-directed, autonomous, and irregular, while inanimate objects do not exhibit this kind of movement. Motion has long been recognized as a central feature of animacy (Maratsos 1983, Rakison and Poulin-Dubois 2001), and children’s developing awareness of motion may parallel their developing concept of animacy (Schwartz 1980). Computational work has shown that motion and having legs are crucial for categorization according to animacy (Lupyan and Rakison 2006:526). Using this simple metric allows for a crucial distinction: “[i]nfants seem to know that animals are not inanimate (they can move themselves) and neither human (one cannot communicate with them)” (Legerstee 2001:200). Communication or ‘talking’ (as opposed to barking and the like) is thus another important feature for animacy, as it distinguishes two important animate categories from each other (Gelman et al. 1983). However, both children and adults will ascribe the ability to talk or communicate more often to non-human animates than to inanimates (Subrahmanyam et al. 2003), which shows that communication is also used to define animacy. Aspects of higher cognition (thought and memory, but also emotions) have also been shown to be related to the concept of animacy (Gelman et al. 1983, Subrahmanyam et al. 2003), and three- and four-year-old children seem to grasp this: they are almost adult-like in judging (as ‘OK’ or ‘silly’) sentences ascribing emotions and personality to animates or inanimates (Becker 2007). Finally, the scientific or biological features of animacy (such as breathing, having internal organs, and so on) are the latest features to become reliable, probably as explicitly learned knowledge (Gelman et al. 1983, Subrahmanyam et al. 2003, Leddon et al. 2009).

While the concept of animacy is multidimensional, as shown above, it is also gradient. This gradience is expressed by the well-established animacy hierarchy: expressions fall on it according to the animacy or sentience of their referents, and languages can reflect these hierarchical differences in word order. Some Bantu languages, for example, order the objects of certain verbs according to their position in the animacy hierarchy (Demuth et al. 2005): higher-ranked, ‘more animate’ objects must be placed before less animate ones, and thematic roles are apparently assigned using information from the context or world knowledge. When both objects are on the same level of the animacy hierarchy, both orderings are possible, and both allow both meanings (first object as theme and second as beneficiary, or vice-versa, for example). Demuth et al. (2005) showed that even four-year-olds observe this animacy ordering rule, at least for a three-tiered hierarchy of humans > animals > inanimates.²² In Jóla Banjal, the recipient and theme objects of ditransitive verbs can be ordered freely only when the referent of the recipient is higher on the animacy hierarchy than the referent of the theme. When the recipient is of the same grade of animacy as the theme, or of a lower one, the recipient has to come after the theme (Bassène 2010). Arguably, Malchukov et al. (2010a)’s argument is in the same vein: they view animacy as part of the prominence of an object and argue that highly animate objects are generally more

²²More fine-grained hierarchies of animacy or potentiality-of-agency have been devised (Dixon 1979, Bowerman 2011), only the binary animate/inanimate distinction will be discussed and used here.

prominent in speakers' minds. Since it is based on common knowledge and human perception and has been shown to have many different effects in many different languages, the animacy hierarchy as such may very well be universal; for a study of animacy in a particular language, however, it is crucial to bear in mind that languages can organize their hierarchies differently (Gentner and Boroditsky 2001:229).

For the animacy hierarchy of English, it has been argued that there are the same three rungs as above: humans > animals > inanimates (Ransom 1977). While Ransom (1977:421–423) also shows that dative alternation choices tend to be made to ensure the object that is higher on that hierarchy comes first, there is also some evidence for a simple binary hierarchy of [humans & animals] > inanimates: this binary distinction has been shown to be active in adults (Paczynski and Kuperberg 2011) and children (Dewart 1979).

At the same time, results of more recent research indicate that children have a more fine-grained appreciation of animacy: even three-year-olds ascribe some features of animacy (such as thinking, breathing, and even using language) more to humans than to other animals, but also more to mammals (like dogs and elephants) than to worms (Saylor et al. 2010:837–838) or insects (Subrahmanyam et al. 2003:357). In the same studies, these 'lower' animals were in turn clearly differentiated from inanimates (like rocks and spoons), to whom children almost never assigned features of animacy. Subrahmanyam et al.'s study also shows motion to be relevant here: practically all participants described insects, dogs, and humans alike as being able to move. However, motion was also ascribed to vehicles and other wheeled things (see also Byrne and Davidson 1985). This was done by adults as well as children, and it is arguably correct: a car can move, after all. It is crucial to note how other features show quite clearly that children on the whole do not categorize moving machines with animals. While anthropomorphic robots are granted some features of (human) animacy (such as thought and memory, Saylor et al. 2010:836–838), they are clearly distinguished from animals: young children know that robots differ from animals in terms of "eating and growing" (Saylor et al. 2010:837), and they "are less inclined than adults to attribute animate predicates to robots" (Subrahmanyam et al. 2003:365). Children however also ascribe more machine features to these robots than to clearly inanimate machines (Saylor et al. 2010:842), meaning the robots could simply be more interesting or exciting to children and thus evoke more description overall. Thus, even three-year-old children do not think of wheeled and otherwise mobile machines as animals, and focusing on just one feature of animacy means obscuring the finely-graded animacy distinctions that children are capable of.

Of course, not every animacy distinction is gradient: when presented with a novel word for an (explicitly or implicitly) animate novel entity, children as young as three will extend that word only to entities that match the exemplar in both shape and texture. When the exemplar is

inanimate, the word is extended to all entities of the same shape, regardless of their texture or other physical features (Booth and Waxman 2002, Colunga 2006). These experiments show a clear animacy dichotomy: animals of a kind are expected to have the same shape and texture, objects of a kind are only expected to have the same shape. There is no gradience or ‘middle ground’ here. This is perfectly in line with the suggestion that different features of animacy matter in different contexts (Gelman and Koenig 2001:699–700): when asked explicitly and specifically about motion, to re-use the above examples, children as well as adults will treat some inanimates (vehicles and the like) like most animals, and possibly some animals like the majority of inanimates (such as rocks). To tease out a subcategory, it can be necessary to draw on several features—again, “[i]nfants seem to know that animals are not inanimate (they can move themselves) and neither human (one cannot communicate with them)” (Legerstee 2001:200). When there is an implicit (and ‘general’) distinction to be made, however, the various features collapse onto a one-dimensional hierarchy with a clear division between animates and inanimates.

This still leaves one question unanswered: **when** does child language start to show effects of an animacy hierarchy? On the one hand, English-speaking six-year-olds seem to be sensitive to it: Dewart (1979) showed that six- to eight-year-old children are more likely to change a passive (monotransitive) sentence to an active one at recall when the agent is animate and the patient inanimate, i.e. when the active sentence has animate-before-inanimate order, than when the agent is inanimate and the patient animate. On the other hand, research in the Piagetian tradition places the median age of arriving at a proper concept of animacy in the tenth year of life (Laurendeau and Pinard 1962:141–159). The Piagetian methods are however unnecessarily complex (see the discussion above; for another example, compare Laurendeau and Pinard 1962:67 and 265–266 with the criticism of Brainerd 1973) and only capture explicit knowledge about biology in any case. This knowledge is presumably taught in science classes, and it is thus no surprise that children as old as eight do not exhibit it (Schwartz 1980, Okita and Schwartz 2006, Leddon et al. 2009). Because of these methodological and theoretical shortcomings, as well as the lack of independent support for it, the figure of ten years can therefore be ruled out as the earliest age for animacy effects. In fact, the true figure is most likely much lower than the initial figure of six to eight years mentioned above: generally, other research (in the Piagetian tradition) has shown that the attention of newborns is drawn to animates or humans (see for example Legerstee 2001:195–197). More specifically, linguistic tasks that use animacy implicitly show that children between two and a half years and four years are sensitive to the animacy of referents (Lempert 1989, Au and Romo 1999, Rakison and Poulin-Dubois 2001, Thal and Flores 2001, Becker 2007, 2009, Leddon et al. 2009). Children begin to use truly transitive sentences at roughly the same age (Ibbotson and Tomasello 2009:66–68), which means that there is no time when children will reliably use transitives without also attending to the animacy of the objects—in

other words, “animacy seems to be a prelinguistic concept that is appreciated by children at a very young age” (Gelman and Koenig 2001:700).

That, however, raises another question: when exactly do children have the (cognitive, prelinguistic) **concept** of animacy? This question again ties into the discussion about the structure of that concept (see Rakison and Poulin-Dubois 2001 for review). As shown above, it is uncontroversial that the concept of animacy is a complex one, meaning that it consists of several features. The concept is surely also more removed from direct perceptual data than other concepts are; this could mean that its constituent features themselves are also complex. It follows that “features themselves might undergo development” (Taylor 1995:241). To recapitulate, there is evidence that animacy depends on motion (Legerstee 2001) and communication (Subrahmanyam et al. 2003): infants expect animates to move irregularly and by themselves (Legerstee 2001). Considering that “infants **initially** associate animate properties with people rather than with animate entities in general” (Rakison and Poulin-Dubois 2001:222, emphasis added), this is a good expectation for them to have: the most salient property of humans in visual perception must be their particular kind of motion, and animals share (this or a similar kind of) irregular motion. This makes animates easier to individuate, which may be why the first words of children generally refer to animates (Gentner and Boroditsky 2001:231). The (early) importance of motion also explains why legs seem to be particularly important for animacy categorization (Lupyan and Rakison 2006:526): if the concept of animacy at that stage is little more than ‘moves by itself’, then the parts that are associated with movement (or their functional cues, Rakison and Poulin-Dubois 2001) are the best indicators of whether something is animate.

Language use, on the other hand, is not associated with easily identifiable functional cues in the physical shape itself. The fact that it apparently begins to affect children’s concept of animacy later than motion does is in all likelihood not only tied to this absence of a functional cue, but also crucially depends on children having at least a simple understanding of others’ attention in the first place, which may arise as late as 12 months (Tomasello 2003:25). In Subrahmanyam et al. (2003:357)’s experiments, children aged four were roughly equally likely to ascribe the ability to talk to dogs as to elephants, which suggests that these children by and large understand what ‘talking’ means but are connecting it to animacy in general rather than to a (human) faculty for language in particular. It is true that adults apparently are more ready than children to grant ‘talking’ to animals, which might be seen to suggest that not even adults know that (non-human) animals do not use language; however, the “incorrect answers” (Subrahmanyam et al. 2003:358) in particular suggest that child participants did indeed not know, while the adults were overthinking the question and thus allowing any intake of air as ‘(sort of) breathing’, any change of shape as ‘moving’, and any noise as ‘talking’. One of the correct answers cited there indicates that adults (or at least the adult participant(s) who gave that answer) do know that ‘talking’ in

the narrower sense relies on a faculty for language that animals do not have: one participant said a bug cannot talk because it has “no language” (Subrahmanyam et al. 2003:358). ‘Talking’ seems to be connected to the concept of animacy, but it may be ascribed to (language-less) animals for several reasons: the subject may genuinely think that the animal in question uses language. They may, however, also be taking the question in a ‘play/pretend mode’ (see Gelman et al. 1983)—animal characters in picture books and the like do talk, after all. Finally, they may assume that the term ‘talking’ is being used in a loose sense of the term, according to which some animals do ‘talk’: they communicate (apes grunt and hoot, dolphins whistle) or at least make noises purposefully (dogs can be trained and commanded to ‘speak’, whales have ‘songs’). On the other hand, a subject may also ascribe talking only to humans, since only humans use language. Thus, talking can be understood in several quite different ways, which means a study of animacy that uses ‘talking’ without very specific explanation or control of what exactly is meant will yield inconsistent and unreliable results. Experimental tasks for use with children must be designed with children’s cognitive limitations in mind anyway. Even though dolls and puppets are designed to resemble highly animate and agentive entities, children as young as three will (correctly) treat them as inanimate if the study is designed carefully to ensure that the children do not assume it to be in ‘play mode’ (Gelman et al. 1983). This even works for “potentially ambiguous entities, such as computers and robots” (Opfer and Gelman 2011:229), provided the task does not overwhelm children. Therefore, all that is necessary is a capability of animals that distinguishes them from inanimates clearly and will not be misunderstood or over-thought.

Luckily, there are some candidates for such a capability in the literature. Gelman et al. (1983:300) used various “reciprocal actions”, asking participants whether they could, for example, hug or play with a rock, and whether the rock could hug or play with them in return. Participants as young as three years performed very well on these types of questions (Gelman et al. 1983:303–304, 310, and 315). The crucial questions were the ones asking after the objects’ capabilities for these actions: creative children might play with a rock just as well as with a cat, but Gelman et al.’s participants seemed perfectly aware that only the cat could play with them. It is thus the latter type of question (‘Can a (noun) play with you?’) that provides the most insight into participants’ concept of animacy, which is why this question (along with one about motion) is used in the present study (see Section 4.1).

2.4 Number

The case of children’s understanding of the plural is much clearer. Again, children may not be able to produce markers for every concept they possess (see Zapf and Smith 2009); but studies

of comprehension and production of plural markers (both canonical, like *-s*, and non-canonical, like *two* with no plural morpheme on the noun) have shown that two-year-old children do understand the idea of a plural (Clark and Nikitina 2009, Zapf and Smith 2009, Barner et al. 2012). These early concepts of plurality may indeed be as simple as ‘two or more’ (Clark and Nikitina 2009:135), but at least for English this is unproblematic—that is the exact number distinction of adult-like English, after all. There is evidence that two-year-olds need a complete sentence (with redundant plural marking) to process plural nouns correctly, while three-year-olds need only the plural marker on the noun itself (Kouider et al. 2006, Wood et al. 2009). Children acquiring other languages may of course exhibit understanding and use of plural marking later than this, as the language in question may have additional number categories (Kovačević et al. 2009) or mark the plural less obviously (compare the early acquisition of the clearly marked Brazilian Portuguese plural shown in Corrêa et al. 2005 with the later acquisition and more variable plural marking in Chilean Spanish shown in Miller and Schmitt 2012). Nevertheless, the English data show that children in general can process and produce simple plural marking by the age of three, if not earlier. Studies using simple plural markers can therefore elicit valid results from quite young children, although data from two-year-olds will need to be treated with caution when there is no redundant plural marking. The present study arguably lacks redundant marking: where plurals are used, only the noun itself carries a plural marker (see Sections 5.1 and 6.1)—the respective pictures showing more than one referent is not **linguistic** plural marking. Therefore, the limited understanding of the plural that Kouider et al. (2006) found in two-year-old children could affect their performance in this study. Since the participants in the present study were four- and eight-year-olds, however, they can be assumed to understand the plural marking as intended.

Children having some concept of grammatical number is necessary for number to have an effect in child speech, but it does not guarantee such an effect. In the case of the dative alternation, it is not entirely clear whether number affects the choice of construction even in the relatively well-studied English of adults: in all of Bresnan et al. (2007:81, 83, and 87)’s models, there is an effect of number (which can be summarized as a plural-before-singular preference), but it always is one of the weakest. Subsequent studies did not include (or report on) the effect of number (de Marneffe et al. 2012, Wolk et al. 2013). However, it is reasonable to assume that number may affect the dative alternation—after all, the dative alternation can be viewed as decision between the two orderings of objects (as I introduced it in Section 2.1 above, and as it is commonly viewed in the Wasow/Bresnan et al. literature), and it is uncontroversial that word order and number can interact. Lorimor (2007:137–152) shows that word order has an effect on the processing of number in English: when the subject NP consists of two coordinated singular nouns, polar questions are more likely than declaratives to be used with a (strictly speaking incorrectly) singular-marked verb—for example, (2.36a) is more likely than (2.36b).

- (2.36) a. Was the bee and ant red?
 b. The bee and ant was red. (both after Lorimor 2007:140)

It has been noted that singular-marked verbs are more likely with coordinations of singulars than with plural-marked subjects (see the discussion in Lorimor Chapter 2 of 2007). However, Lorimor (2007:143–144) took particular care to ensure that participants used the same words in both types of sentences, and in the exact construction as in the examples in (2.36). This means that the striking difference in singular agreement rates between the two types cannot be due to the peculiar type of subject, since they shared that. Therefore, word (or constituent) order can be seen as the relevant difference between questions and declaratives here.²³ In this view, the (incorrect) singular marking on the verb is more likely with postverbal subjects than preverbal ones—in other words, word order affects number agreement (in English).

To return to ditransitives, Duranti (1979) argues that number is one of the features that determine which orders of object clitics in two Bantu languages are grammatical at all. It is true that his work focuses on topicality, and that topicality (or ‘givenness’) is certainly included in many accounts of the English dative alternation. Nevertheless, it is possible that number has a distinct effect: while Bresnan et al. make no strong claim regarding the independence of its effect, the fact that removing number (and two other factors) slightly reduced the classification accuracy of their models B and C (Bresnan et al. 2007:89) is suggestive. Of course, there are counterexamples of phenomena where language processing is affected by features other than number, but not by number: for example, reading time and comprehension in Basque are affected by NP case, but not number (Laka and Erdocia 2012, Santesteban et al. 2013). However, this does not mean that number categorically cannot affect word order. Thus, the effect of grammatical number on the choice of construction in the dative alternation is worthy of further investigation, and it is possible to do such work in the realm of child language.

²³The literature on these types of agreement ‘errors’ abounds with theories on it (see for example den Dikken 2001 or Lorimor 2007 for an overview). Kayne (2003)’s approach to similar phenomena might be extended to cover coordinations: he argues that an unpronounced auxiliary verb that agrees with the subject of a relative clause, as in (i), explains noncanonical agreement patterns found with that construction.

- i the people that John_i Aux_i think should be invited (after Kayne 2003:262 and 264)

It may be possible to extend this to an ellipsis-based account of coordination, such that an unpronounced copula (each) agrees with every conjunct but one, and the overt verbal element only agrees with that last conjunct:

- ii The bee_i Aux_i and the fly_j Aux_j and the ant_k was_k red.

However, this would not explain why singular agreement is more common with questions than with declaratives, which is what Lorimor (2007) found.

2.5 Length

The length²⁴ of the two objects of a dative alternation verb has long been seen to affect the choice of construction. The more general law of increasing constituents (Cooper and Ross 1975, Behaghel 1928) or “principle of end-weight” (Biber et al. 1999:898, also Wasow 2002) says that longer constituents are to be placed after shorter ones, at least in languages like Sanskrit, English, and German. Structures adhering to this law are very conspicuous as their purely structural description can violate other rules of the language in question: compare the ungrammatical (2.37a) with (2.37b), which is more acceptable only because it places a very long object after a shorter adverbial.

- (2.37) a. * I like very much [object NP apples]. (after Downing and Locke 2006)
b. ? I like very much [object NP apples that have been stored in a cool cellar and cut into cubes with care].

To apply the law of increasing constituents to the English dative alternation is straightforward even though ungrammaticality is not generally involved: apart from the preposition *to* that is included in the prepositional construction but not the double object construction, these two constructions are nothing more than the two possible orderings for two (immediately adjacent) object phrases. The choice between constructions can thus also be understood as a choice between orders, and (all else being equal, since other features also seem to have an effect) the order that places a shorter object before a much longer one is preferred. To extend the above example, while both (2.38a) and (2.38b) below are quite acceptable, (2.39b) is strongly preferred to (2.39a).²⁵ The only difference between (2.38) and (2.39) is the length of the theme object, and so it must be this length that causes a construction and ordering preference in (2.39) where there is none (or not a strong one) in (2.38).

- (2.38) a. Kate handed an apple to Rick.
b. Kate handed Rick an apple.
- (2.39) a. ? Kate handed an apple that had been stored in a cool cellar and cut into quarters with care to Rick.
b. Kate handed Rick an apple that had been stored in a cool cellar and cut into quarters with care.

²⁴The term ‘weight’ is also used in the literature to describe this difference of quantity or complexity between objects. However, it could also be seen to include other factors (see Wasow 1997:88), and for the sake of clarity, the less confusing term ‘length’ will be used here.

²⁵Biber et al. (1999:928) found the length effect to be much stronger in the double object construction than in the prepositional construction; this could however just be an artifact of their corpus data and does not change the basic fact that there is a length effect.

These examples show that the effect of this law (or an effect very much like it) in the dative alternation is also one of the more easily demonstrable ones, which explains why it has been included in most (if not all) studies of the features that affect the English dative alternation. Interestingly, this length effect emerged as statistically significant and strong in all studies (that I know of) which included it. This strong and independent preference for the construction that places shorter objects before longer ones has been found data from the (Canadian English) Aligned-Hansard corpus (Arnold et al. 2000), the (American English) Switchboard corpus (Bresnan et al. 2007 and Snider 2011), the (New Zealand English) ONZE corpus (Bresnan and Hay 2008), the British English parts of the ICE corpus (Theijssen 2009), African American English (Kendall et al. 2011), the historical ARCHER corpus (Wolk et al. 2013), the English-speaking children in CHILDES (de Marneffe et al. 2012), the (Indian English) Kolhapur corpus (de Cuypere and Verbeke 2013), and six South Asian varieties of English (Bernaisch et al. 2014).²⁶ Clearly, it is established as firmly as the existence of the English dative alternation itself.

The only point of contention is how to measure the lengths of the two objects, or the length difference (see Wasow 2002:28–32 for a more detailed discussion and evaluation): following the seminal Bresnan et al. (2007) study, the log-transformed number of graphemic words by which the two objects differ in length appears to be the generally accepted measure (see for example Bresnan and Hay 2008 or Jaeger and Snider 2013), though the non-transformed number of (graphemic) words has also been used with reasonable success (see for example Bürkle 2011 or de Marneffe et al. 2012). In a production experiment, Stallings and MacDonald (2011) showed that it is the difference in length, not the length of one of the phrases by itself, that affects reordering in heavy NP shift. As long as the length measure is consistent, it does not seem to matter much which one is used: Hundt and Szmrecsányi (2012:248) report an informative model of the genitive alternation (the choice between *Kate's desk* and *the desk of Kate*) that measured length by the number of orthographic characters. De Marneffe et al. (2012) used the number of syllables as an alternative to the number of words, which takes up Akasaka and Tateishi (2001)'s point regarding object phrases consisting of one long word: for example, both *Rick* and *watermelons* in (2.40) have a length of 1 when length is counted by words, whereas a syllable-counting method would value *Rick* at a length of 1 and *watermelons* at 4.

- (2.40) a. Kate handed watermelons to Rick.
b. Kate handed Rick watermelons.

It is unsurprising that the de Marneffe et al. (2012:34) model was not significantly changed by using syllable counts rather than word counts. Their study was based on corpus data, which is

²⁶See Arnold et al. (2000) and Wasow (2002) for yet more references.

not systematically varied. The (or a) general length effect emerges even from an uncontrolled dataset such as that; the question of whether longer words are effectively ‘heavier’ can only be answered by a carefully designed, controlled study. What de Marneffe et al. (2012) do show is that both word and syllable counts are useful measures of length in the dative alternation, and it is obvious that (graphemic) words are easier to count in corpus data. These points are supported throughout the literature: in a study of small corpora of written and spoken English, Szmrecsányi (2004:1034–1035) found that number of words, number of syntactic nodes, and a very detailed “Index of Syntactic Complexity” correlate very strongly, particularly in speech data. He also points out that number of words is by far the easiest of these to use. Similarly, Wasow (1997:94) concludes that node and word counts, when applied to English weight phenomena, are “essentially indistinguishable”. Anttila et al. (2010:974–976) use the log-transformed number of primary stresses, since it emerges as the best measure in their dataset; however, they acknowledge that only data with prosodic annotation even allows this measure to be used. Furthermore, the log-transformed number of words is not very much worse in their data (Anttila et al. 2010:974), and that measure can be used with almost any data. In conclusion, the number of graphemic words can be considered a reasonable and useful count of object length.²⁷ Wasow (2002:38–41) shows that different measures of length are strongly correlated, and the strong effect of length in the models of Bresnan et al. (2007), de Marneffe et al. (2012), and others demonstrates quite clearly that object phrase length as measured by graphemic words does have an effect in the dative alternation. The possible effect of word length (measured in syllables) will be one of the objects of investigation in the present study (see Sections 3.3, 5.1 and 6.1).

There is some speculation as to the (psycholinguistic) cause of the preference for short objects before long ones (see also the overview focused on information structure in Stephens 2010:34–39). It might ease processing, as Hawkins (1994) argued for the dative alternation. Of course, children’s cognitive capabilities are limited when compared to those of adults, as Bloom (1990) shows using subject-less utterances produced by English-speaking children.²⁸ For the dative alternation, this might lead us to expect that the effect of object length on choice of construction is at least as strong in children as it is in adults, but probably stronger.²⁹ The short-before-long preference might also be explained by a preference of the language faculty for committing to a choice as late as possible: Wasow (2002:55–56) points out that some verbs that participate in

²⁷I must stress that this discussion is only valid for English. The Danish dative alternation, for example, is arguably better modeled using node counts: Kizach and Balling (2013) point out that graphemic word counting would consider the indefinite *en bil* (‘a car’) to be twice as long as the definite *bil-en* (‘the car’), whereas node counting assigns the same length to both. Having no obvious asymmetries like this, English data can be analyzed using the simpler graphemic word count.

²⁸Even those that disagree with Bloom (1990) on the interpretation of his data agree that there are limitations to children’s cognitive resources when compared to those of adults, either in production (Hyams and Wexler 1993:452; see also Bloom 1993’s reply) or in learning (Freudenthal et al. 2007:86, with the apparent production limitations being an epiphenomenon of this limited learning).

²⁹Data that allow direct and valid comparison do not yet exist.

the dative alternation can also be used in a double object-like construction with a CP theme, as in (2.41b). He argues that speakers are more likely to use the double object construction with these verbs because it gives them more time to decide whether to use an NP or a CP theme: (2.42a) and (2.42b) diverge after *Rick showed. . .*, meaning the decision is made by the time *his* or *what* are pronounced. Both (2.41a) and (2.41b), on the other hand, begin with *Rick showed Kate. . .*, so the time taken up by *Kate* is extra time for decision-making.

- (2.41) a. Rick showed Kate his book.
 b. Rick showed Kate what he had written.
- (2.42) a. Rick showed his book to Kate.
 b. Rick showed [what he had written] [to Kate].

This “late commitment” (Wasow 2002:49) explanation thus assumes that the dative alternation really results from a difference between two sub-classes of dative alternation verbs: the verbs that also allow a CP theme are more likely to be used in the double object construction (because it allows late commitment), and those that do not allow a CP theme are (roughly) equally likely to be used in either construction. One might then argue that CP themes are typically longer than NP themes, and that the average theme of a verb of the former type is therefore bound to be longer. The apparent short-before-long preference could then be argued to be an artifact of this late commitment preference and typical phrase lengths. However, this account in no way explains the clear preferences for short-before-long ordering with verbs that only allow NP themes, and Wasow (2002:56, fn. 32) concedes that it does not explain the apparent preference for long-before-short orderings in languages like Korean.

An intriguing explanation from a different angle is that short-before-long combines well with the prosody of the sentence as a whole. Zec and Inkelas (1990) describe and explain several phenomena of syntactic ‘weight’ by claiming that the syntactic operations in question can only apply to prosodically branching constituents. In the case of English heavy NP shift,³⁰ they present this phono-syntactic constraint as an explanation of the difference in grammaticality between (2.43a) and (2.43b) as well as between (2.44a) and (2.44b).

- (2.43) a. * Mark showed to Jane (some letters).³¹

³⁰Zec and Inkelas (1990) apparently intend the narrow meaning of ‘heavy NP shift’, which excludes the dative alternation. Their examples with *show*, a clear dative alternation verb, are carefully worded to just shift the theme NP ‘past’ the recipient PP without changing the sentence to a double object construction. However, since (narrowly-defined) heavy NP shift, the dative alternation, and particle verb ordering phenomena are affected by the same features of objects or phrases, they can reasonably be viewed as examples of one and the same basic phenomenon (see Section 2.2). Zec and Inkelas (1990)’s explanation of one of these weight effects would then have to explain the others as well.

³¹The parentheses in (2.43) and (2.44) indicate prosodic phrases (following Zec and Inkelas 1990’s algorithm). Everywhere else in this study, brackets indicate syntactic constituents.

- b. Mark showed to Jane (some letters) (from Paris).

(both after Zec and Inkelas 1990:377, original judgments)

- (2.44) a. * Mark showed to Jane (that report on him).

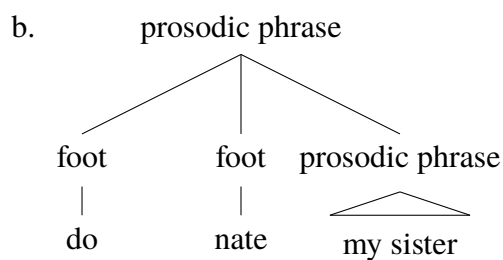
- b. Mark showed to Jane (that report) (on Dukakis).

(prosodic phrasing for both after Inkelas 1990:16, judgments for both following Zec and Inkelas 1990:377)

Enlarging the theme NP even by just a little (two words, three syllables, one stress) makes the HNPS construction much more acceptable. It appears likely that most speakers of English would agree with this judgment. Since Inkelas (1990) argues that English pronouns are not prosodic words, the same difference in acceptability should be apparent in (2.44): the theme NP in (2.44a) contains a pronoun, but no additional noun other than *report*, so the whole NP is parsed as one phonological phrase under Zec and Inkelas (1990)'s Phonological Phrase Algorithm. The theme in (2.44b), on the other hand, does contain another noun and thus another prosodic word. Because of this, the same algorithm parses it as two phonological phrases, meaning the NP is heavy. However, an informal survey of native speakers suggests that both (2.44a) and (2.44b) are acceptable. While not a rigorously scientific finding, this directly contradicts a prediction made by Inkelas (1990) and Zec and Inkelas (1990): the grammaticality difference in (2.44) does not exist, at least not for all speakers of English. Furthermore, Zec and Inkelas (1990:377) claim that a heavy shifted NP is “preceded by pause and associated with special intonational effects”. The informal survey of native speakers mentioned above also revealed the same pattern of intonation and pauses for (2.44a) as for (2.44b) (and this pattern was different from the one found in a similar non-shifted sentence). Moreover, Selkirk (1995:440) argues that personal pronouns can be phonological words (see also Selkirk 1980:32 and 130–133), which would allow for *on him* to form a phonological phrase by itself under her analysis (Selkirk 1995:452). As Inkelas (1990:16) also points out, the analysis as in (2.44a) above therefore only applies to unstressed or weak pronouns. Thus, this prosodic approach to length in English reordering phenomena does not seem to be able to account for all attested patterns.

More recently, Anttila et al. (2010) have proposed a more detailed prosodic theory of the dative alternation and other weight effects. Their approach is founded on Optimality Theory (OT): it tries to explain dative alternation choices and availabilities using constraints on the formation of prosodic phrases and the location of stresses. Anttila et al. argue that some verbs with two prosodic feet do not take the double object construction because the verb has to form a prosodic phrase with the recipient in that construction. Since the recipient forms a sub-phrase in that phrase, any verb of two feet (like for example *donate*, *deliver*, or *transfer*) makes this verb-recipient phrase ternary, as shown in (2.45). This violates *TERNARY, the constraint that dictates “avoidance of ternary prosodic phrases” (Anttila et al. 2010:957) and thus rules out the double object construction for these verbs.

(2.45) a. * donate my sister the book (after Anttila et al. 2010:957, original judgment)



This OT constraint against ternary prosodic phrases can be violated, of course, which Anttila et al. (2010:956–957) mention in connection with two-foot verbs that do alternate (like *catapult* or *radio*). Tableau (I)³² from Anttila et al. 2010:964 illustrates the violability of OT constraints: the double object candidate (Ia) violates the constraint against stress clashes (the lexical stress of *give* and the phrasal stress of *my sister* clash in the verb-recipient phrase) and the constraint that demands that the recipient is placed in the utterance-final salient position. The prepositional candidate (Ib) does not violate these two, but it does violate the constraint for focusing the theme and the constraint against using the preposition *to*. Both fo them violate *STRESS-TO-STRESS, the constraint against lexical stresses outside the prosodic phrase that carries sentence stress (see Anttila et al. 2010:954–955 for details and justification), and *P-PHRASE, the constraint against having prosodic phrases (intended to keep prosodic phrasing as simple as possible). Since all the constraints given in (I) are unranked with respect to each other, neither of these two candidates incurs a violation at a higher rank than the other, and both candidates are optimal.

	*TERNARY	*CLASH	STRESS-TO-STRESS	*P-PHRASE	Focus(Theme)	Focus(Recipient)	*to
<i>give, my sister, the book</i>							
☞ a. (give my sister) (the book)		*	**	**		*	
☞ b. (give) (the book) (to my sister)			**	***	*		*

Tableau I: The dative alternation in OT (Anttila et al. 2010:964) (note that all constraints given here are not ranked)

Although they do not give the equivalent tableau for *donate* themselves, Anttila et al. (2010:962)’s constraints can be used to construct it as shown in (II). Since the *TERNARY constraint is introduced specifically to rule out the double object construction with two-foot verbs (see Anttila

³²In the interest of clarity, I have left out irrelevant candidates as well as the three top-stratum constraints that rule out the candidates with ill-formed phrases (see Anttila et al. 2010:964). The parentheses in the OT candidates indicate prosodic phrases.

et al. 2010:957), their intent must be for the violation of this constraint to rule out (IIa) and make (IIb) the optimal candidate.³³


<i>donate, my sister, the book</i>	*TERNARY	*CLASH	STRESS-TO-STRESS	*P-PHRASE	Focus(Theme)	FocusRecipient	*to
a. (donate my sister) (the book)	*	*	**	**		*	
 b. (donate) (the book) (to my sister)			**	***	*		*

Tableau II: Non-alternating dative verbs in OT (following Anttila et al. 2010)

As with previously discussed constructions that were claimed to be unacceptable or ungrammatical in the literature, sentences like the non-optimal (IIa) are actually attested:

- (2.46) a. . . . a farmer living near their university donated them the trunk of a banana tree
tinyurl.com/donatedthem
- b. Ross also donated the church the large brass bell that still calls people to worship.
tinyurl.com/donatedchurch

Again, these might be rarer than the corresponding prepositional construction, but it is easy to see that they are used.³⁴ Anttila et al. (2010:968–969) suggests that their analysis would allow for less optimal candidates to be rare rather than impossible (particularly if the fatal constraint violations is one in an unordered stratum, see Anttila et al. 2010:964). The constraints and ranking as in (II) might allow for this gradient acceptability—the (supposedly rarer) double object candidate violates more constraints than the (supposedly more frequent) prepositional candidate, after all, and the number of violations of unranked constraints seems to be an obvious candidate for a predictor of relative frequency in cases where several candidates are optimal. Even so, *TERNARY does nothing to explain the alleged difference between alternating and non-alternating two-foot verbs (see Anttila et al. 2010:956–957): since both types of verbs have two feet, they should be identical with respect to prosodic phrasing, and thus be equally acceptable and frequent with the double object construction. Anttila et al. (2010) do not address this critical point.

³³It is apparent from Anttila et al. (2010:963–964) that the number of constraints violated is more important in their analysis than the total number of violations, so the four violations of (IIb) make it optimal compared to (IIa) with its five violations.

³⁴The relevant search strings here are “conveyed/entrusted/dictated her/him/them the”.

Furthermore, even the less controversial facts of the dative alternation are not explained well by Anttila et al. (2010)'s theory. Their STRESS-TO-STRESS constraint is violated by every lexically stressed word that is not part of the rightmost phonological phrase. In many cases, this makes predictions that are much the same as those made by other length-based accounts: all other things being equal, the construction that places a long object at the end will be optimal because it has the most words in the rightmost phrase. For example, (2.47a) has more violations of STRESS-TO-STRESS than (2.47b), and the latter is therefore preferred.

- (2.47) a. ? not to give [children who have a peach allergy] [peaches]
 b. not to give [peaches] to [children who have a peach allergy]

However, “the theory predicts that the weight effect should disappear if nuclear stress falls outside the VP” (Anttila et al. 2010:955). The example sentences in (2.48a) illustrate this: as there is another phrase after the VP, neither object of the verb can be the rightmost phrase. This makes the constraint on stresses in the rightmost phrase irrelevant for the relative optimality. Therefore, Anttila et al. (2010) predict that there should be no weight effect here. For one-word objects, this appears to be perfectly true: (2.48b) does not seem to be any worse than (2.48a).³⁵

- (2.48) a. not to give [children] [peaches] to avoid possible allergic reactions
 (after Anttila et al. 2010:955)
 b. not to give [peaches] to [children] to avoid possible allergic reactions
- (2.49) (People with a latex allergy may also be allergic to peaches.³⁶ Parents are advised. . .)
 a. ? . . . not to give [children who have a latex allergy] [peaches] to avoid possible allergic reactions.
 b. . . . not to give [peaches] to [children who have a latex allergy] to avoid possible allergic reactions.

Weight effects do seem to appear when one object is very long, however, even when there is material after the objects: in an informal survey of native speakers, (2.49a) was consistently described as less comprehensible and less likely to be said than (2.49b). While not a rigorously scientific finding, this preference goes against the predictions of Anttila et al. (2010). Therefore, that theory makes correct but commonplace and unique but unsound predictions, and it is far from offering a full account of the dative alternation. Also, based on several experiments, Bock and Levelt (1994:969) argue that syntactic structure cannot be affected by prosodic information,

³⁵I have replaced the theme *it* with *peaches* here in order to avoid any sentences being ruled out by the constraint against lexically unstressed non-branching prosodic phrases (Anttila et al. 2010:950–953). *It* would form such a phrase; *peaches*, being lexically stressed, does not.

³⁶According to <http://www.foodallergy.org/allergens/other-allergens>, latex and peach allergies sometimes cross-react.

and the dative alternation is (at least partially) a syntactic phenomenon. In conclusion, there is little to support the claim that the effect of phrase length on the dative alternation is really an effect of prosody. This does not mean that prosody is ruled out, of course; rather, it means that this approach to length in the dative alternation can only be continued fruitfully once a carefully designed study has found some unambiguous initial support for prosodic effects.

Accepting the idea that several ‘weight’ phenomena share the same cause means that this shared cause should be investigated further in any case. The effect of phrase length in the dative alternation and other phenomena is clearly established, and it is not unreasonable to assume that this effect manifests itself in all length distinctions. There are some (albeit weak) indications of such an effect for word differences as small as one syllable: in her study of the effect of information structure on dative alternation choices, Stephens (2010:98) finds that children tend to place given objects before new ones, but “cannot rule out the possibility that givenness influenced construction choice indirectly, via pronominality (short-first)”. This is due to the fact that virtually all given objects were also pronouns in her data from children participants. Stephens is right to also consider length: her study 2B, which had adults as participants, showed a much weaker correlation of givenness and pronominality (while all given recipients were pronouns, about half of the given themes were ‘full’ NPs such as *a/the hat*, Stephens see 2010:112–117). A strong givenness effect was evident only for *give*, *show*, and *throw*. The fourth verb used in this study, *read*, was the only one that was used with a female recipient in one of the two trials and a male recipient in the other, while the other verbs all had two female or two male recipients (Stephens 2010:84 and 105): in the trials with *show*, for example, participants saw videos of a man showing a shirt to a woman or a man showing a coat to a woman (see (2.50)³⁷ and (2.51) below), whereas the *read* trials had a woman reading a book to a man and a man reading a book to a woman (see (2.52) and (2.53) below). The examples given by Stephens indicate that participants in this study used *a/the man/kid* for full-NP male recipients and *a/the woman* for female ones. Thus, the female full NPs were one syllable longer than the male ones. For *give*, *show*, and *throw*, this is presumably unproblematic, as the recipients in both trials for each of these verbs had the same gender and thus the same length. Analysis within these verbs (as in Stephens 2010:107–109) is therefore unaffected by the length difference between *man/kid* and *woman*: in the given-theme condition, the preference for (2.50a) over (2.50b) is presumably just as strong as the preference for (2.51a) over (2.51b), since the prepositional construction in both cases fulfills the given-before-new and short-before-long preferences. The prepositional construction would therefore be predicted to be favored more strongly in the given-theme condition than in

³⁷For the sake of clarity, the examples here all use full NPs only. The actual responses that Stephens elicited had pronominal agents, themes, and recipients as well (see Stephens 2010:113–114 for details). The crucial length difference between *woman* and everything else could be preserved under the assumption that *a/an* and *the* do not count for length (Anttila et al. 2010:974–975), which would mean that *it*, *him*, *a/the book/shirt/...*, and *a/the man* all effectively have one syllable and that *a/the woman* has two.

the control condition (where the agent was given, and both theme and recipient therefore new; see Stephens 2010:106) across the *show* data, and the results (Stephens 2010:109) are consistent with this prediction.

(2.50) (video stimulus: man shows shirt to woman)

- a. The man showed the shirt to the woman.
- b. The man showed the woman the shirt.

(2.51) (video stimulus: man shows coat to woman)

- a. The man showed the coat to the woman.
- b. The man showed the woman the coat.

For *read*, however, the length difference may pose a problem. In the given-theme condition, the preference for (2.52a) over (2.52b) might be weaker than the preference for (2.53a) over (2.53b): since the theme and recipients are of equal length in (2.52), the preference for (2.52a) would be driven only by given-before-new, while the preference for (2.53a) would be driven by both given-before-new and short-before-long. Assuming the strength of these preferences affects relative frequencies, (2.53a) will therefore be relatively more frequent than (2.52a) (and, conversely, (2.52b) will be relatively more frequent than (2.53b)). When lumping both trials together and thereby averaging the frequencies of (2.52a) and (2.53a), the prepositional construction would be predicted to be only slightly more frequent overall in the given-theme condition than in the control condition. The results shown in Stephens (2010:109) are consistent with this prediction as well.

(2.52) (scene: woman reads book to man)

- a. The woman read the book to the man.
- b. The woman read the man the book.

(2.53) (scene: man reads book to woman)

- a. The man read the book to the woman.
- b. The man read the woman the book.

This alternative analysis is obviously more than a little speculative. “[S]parsity of data” and “semantic/pragmatic and or usage factors” (Stephens 2010:118 and fn. 13) are valid explanations as well, but they are not inherently better: the effect of word length claimed above would be a very small one, and therefore difficult to spot. The length difference in textbook examples of the dative alternation (like (2.38) and (2.39) on page 38, for instance) illustrates the established phrase length effect so well because it is immediately apparent, even when there are other effects at work simultaneously, and the various studies of this length effect have been able to establish

it so strongly because it is so apparent even in large and messy datasets. As mentioned above, de Marneffe et al. (2012:34) found no significant improvement by switching from number of words to number of syllables, but that only shows that the number of syllables is no better and no worse than the number of words in their data. Corpus data does have many advantages, a lot of them because of the large size of corpora, but the one major disadvantage of corpus data is that it is messy: transcriptions are often somewhat idealized, which is a particular problem when the transcripts are based on the more variable speech of children. This is further compounded by the fact that different transcribers will use different standards, and it is practically impossible to check inter-transcriber accuracy with large collaborative corpora (like CHILDES). Moreover, the context of a corpus item is hard or even impossible to establish in some cases. Thus, if there was an effect of word length, it might easily be buried by the noise inherent in corpus data. A controlled study is therefore necessary to establish or disprove this possible word length effect.

2.6 Other features

The previous sections have laid out the three features of objects in the dative alternation that this study focuses on: animacy, grammatical number, and length. There is good evidence that other features of the objects, the verbs, and the context also affect the choice of construction in the dative alternation, but these will not be investigated in this study. This section will very briefly discuss the well-supported ones, namely prosody, ‘givenness’, pronominality, grammatical person, definiteness, priming or structural persistence, and the strong tendencies of certain verbs to occur in one construction much more frequently than the other. The terms used here are mostly those used by Bresnan et al. (2007) rather than, for example, those of Malchukov et al. (2010a), but the overlapping terms will be mentioned where appropriate.

The prosodic approaches of Zec and Inkelas (1990) and Anttila et al. (2010) have already been discussed in Section 2.5 above, and the problems mentioned there and in Wasow (2002:20–21) apply: these accounts fail to explain a lot of the data, and there is some evidence against them. Metrical prosody affecting syntactic structure or word order is not generally ruled out, of course (see Erteschik-Shir and Rochman 2010 and Section 2.5 above), but (to the best of my knowledge) there is no explicit and substantiated theory of the English dative alternation as an operation that generates optimal meter. Therefore, meter is not included as one of the factors to be tested in this study.

In the context of the dative alternation, ‘givenness’ and ‘newness’ are used as terms for whether a particular object has been mentioned previously in the same utterance context, meaning whether it can be assumed to be given or new. Malchukov et al. (2010a:20) combine givenness

with animacy and pronominality in their measure of “prominence”. Givenness has repeatedly been found to be among the predictive factors for the choice of dative alternation construction: generally speaking, constructions that place given objects before new ones are more frequent (see for example Wasow 2002:68-82, Bresnan et al. 2007, or Bresnan and Hay 2008). Brown et al. (2012) found that new-before-given sentences like (2.54a) are read more slowly than given-before-new ones like (2.54b), but only when both use the double object construction. No significant reading time difference was observed between the corresponding prepositional construction sentences like (2.54c) and (2.54d), and the prepositional sentences were also read faster than the double object sentences overall (Brown et al. 2012:202–203).

(2.54) (A biologist discovered an owl that seemed to belong to a new species.)

- a. The biologist showed [a rat] [the owl] to gauge their reactions.
- b. The biologist showed [the owl] [a rat] to gauge their reactions.
- c. The biologist showed [a rat] to [the owl] to gauge their reactions.
- d. The biologist showed [the owl] to [a rat] to gauge their reactions.

(all Brown et al. 2012:208)

They argue that this shows the supposed given-before-new preference to be a restriction on the use of the marked double object construction; however, as shown in Sections 2.1 and 2.2 above, there is no conclusive evidence for the hypothesis that one construction is more marked than, or derived from, the other.³⁸ The only justified assumption is therefore that givenness has some effect in the dative alternation and other (re-)ordering phenomena. Though he discusses it in the context of length and stress, Oehrle (1976:259–260)’s HNPS example (xxii), repeated here as (2.55), also shows an effect of newness.

(2.55) We found in the barn, a stick.

(Oehrle 1976:260)

As Oehrle points out, the ordering can not be due to a length effect here. The sentence seems to be most felicitous when *the barn* has been previously mentioned (or can be inferred from mention of *the farm* or the like) and *a stick* is new. Thus, this example supports the preference for ordering given phrases before new ones. Malchukov et al. (2010a) include this feature in their term ‘prominence’: animate, given, pronominal objects are more prominent, and (in the English dative alternation) the construction that places the more prominent object first is preferred.

³⁸Kizach and Balling (2013)’s replication of the Brown et al. (2012) study does not provide additional support here, since it concerns only Danish. The Danish dative alternation is different at least in regard to length: as mentioned in Section 2.5, Danish definite NPs would be counted as one word shorter than their indefinite counterparts by certain length measures (definite *bilen* versus indefinite *en bil*). As there are relatively few studies of the Danish dative alternation, we cannot assume that it corresponds well to the English dative alternation, and evidence from the Danish one is therefore not good support of any arguments regarding the English one just yet.

For child language, the effect of givenness appears to be much weaker, although the cognitive prerequisites seem to be present very early (de Cat 2011). Stephens (2010:92–98) showed experimentally that children tend to choose the dative alternation construction that places a previously mentioned object first, and Bürkle (2011:42) found statistically significant effects for givenness of both objects that are in line with the given-before-new preference: to use examples from the CHILDES data used there, (2.56a) has the given recipient *the – this prince guy* before the new theme *the – the claw*, and (2.56b) has the given theme *my – my cold* before the new recipient *Marcia*.

- (2.56) a. he was trying to give [the – this prince guy] [the – the claw]
(Eng-US-MOR/Carterette/fifth.cha, line 9181)
- b. I won't give [my – my cold] to [Marcia] (Eng-UK-MOR/Wells/olivia08.cha, line 420)

De Marneffe et al. (2012:38 and 42–43), on the other hand, did not find significant givenness effects in a similar corpus study, but they did find an effect approaching significance in child-**directed** speech (de Marneffe et al. 2012:48). This led them to conclude that children do mirror the givenness effect that they see in their linguistic input but are initially unable to assign the effect strength that produces adult-like output.³⁹ Stephens (2010:148–154) found that children produce fewer *uhs*, *ums*, and pauses in all given objects as well as in whole dative alternation utterances that have given-before-new order. This finding arguably supports that conclusion: given objects and given-before-new utterances being easier for children to produce is good evidence that givenness does have some effect in children's speech after all. Moreover, children as young as three years tend to act out both given-before-new and new-before-given instruction sentences in the given-before-new order of actions, suggesting that given-before-new is the preferred ordering (Junge et al. 2015). Bringing givenness to bear on the dative alternation decision and setting the target effect strength may require additional learning, as de Marneffe et al. (2012) argue. The present study was designed to ensure that givenness would not have an effect (see Sections 5.1 and 6.1 for details).

The discussion in Section 2.2 shows how important it is in the dative alternation whether an object is a pronoun or not. This binary distinction has been called that object's pronominality. Haspelmath (2004 and 2007) claims that the double object construction(s) can only be used with certain combinations of pronominality; as shown above, this strong claim is not borne out by speaker data. The weaker version, however, can be maintained: there is little doubt that the double object construction is much less acceptable and frequent with nonpronominal recipient and pronominal theme than with other combinations of pronominality (compare (2.57a) with

³⁹This difficulty, de Marneffe et al. (2012) argue, stems from the strong correlation of givenness with pronominality, which itself has an effect in the dative alternation (see below).

the other examples below). Therefore, it is reasonable to expect that pronominality has an effect on the choice of dative alternation construction.

- (2.57) a. Rick gave Kate it.
b. Rick gave her the coffee.
c. Rick gave her it.
d. Rick gave Kate the coffee. (after Haspelmath 2007)

Scholes (1981) found that children are significantly above chance in comprehending sentences with two personal pronouns only by the age of five or six. For sentences with one personal pronoun, other studies (such as Charney 1980, Loveland 1984, Valian 1986, Girouard et al. 1997, and Tomasello 2000) show accurate understanding and production as early as the third or even second year of life. It is therefore also reasonable to expect a pronominality effect in children's choice of dative alternation constructions, at least when only one object is pronominal. Such effects are significant in de Marneffe et al. (2012:39–40)'s model, which shows that children choose the construction that orders pronouns before full NPs. Pronominality is obviously connected to length and givenness—pronouns are short and typically require a given or at least accessible antecedent—but the connection is far from being a perfect correlation, since short 'full' NPs can have given referents as well. Accordingly, Gries (2003), Bresnan et al. (2007), and Snider (2011) find similar effects of the pronominality of theme and recipient in adult language, along with, but independent of, effects of length and givenness. As this was found in several studies, pronominality can be argued to have a distinct and independent effect in the dative alternation, at least in adult language. The pronominality effect is also considered as part of the objects' prominence by Malchukov et al. (2010a): pronouns are more prominent than full NPs (see also Aissen 2003 and Branigan et al. 2008). The experiments in the present thesis will not use pronominal objects, meaning the feature of pronominality is kept constant.

Bresnan and Nikitina (2003) and Bresnan et al. (2007) found a minor effect of grammatical person: recipients of nonlocal person (meaning third person as opposed to first or second person, following Aissen 1999:679) somewhat favor the prepositional construction. Both papers use this to imply a preference for a local-before-nonlocal ordering preference. Bresnan et al. (2007:78) removed the person of the theme from consideration because of data sparsity: it is intuitive that themes of dative alternation verbs would mostly and indeed prototypically be nonlocal, since "local person pronouns are . . . animate" (Bresnan and Nikitina 2003:25) and the prototypical theme is inanimate. The prototypical double object construction can therefore be expected to observe local-before-nonlocal ordering: it has the prototypically local recipient before the prototypically nonlocal theme, after all. The prototypical prepositional construction violates this supposed ordering preference: (2.58) has a nonlocal object before a local one, yet it is not

unacceptable.

(2.58) Give that cup to me, will you?

(2.59) Give that cup to Kate when you see her.

However, not all constructions are prototypical—(2.59) has two nonlocal objects and thus is not inconsistent with any locality- or person-based ordering preference. Furthermore, assuming that the double object construction is more frequent than the prepositional construction (see for example Cook 1976:436, Gropen et al. 1989:219, Bresnan et al. 2007:79, or Bürkle 2011:44), the prototypical double object construction is more frequent overall than the prototypical prepositional construction. Since the more frequent one happens to have local-before-nonlocal order, a spurious locality effect could arise in a model of the dative alternation. As with pronominality, Bresnan et al. (2007) argue that the effect of (the recipient's) person is independent of the effects of other features. Be that as it may, this person effect is statistically weak and therefore of lesser concern than the effects of the features discussed in detail above. To rule out any person effect, the stimuli used in this study were designed so that all objects are third person (and thus nonlocal) throughout.

The grammatical definiteness was investigated as a potential factor of influence by Wasow (2002:65–67). He found no effect of definiteness in HNPS, but highlighted the fact that his methodology was not ideally suited to finding it. Indeed, later studies (Gries 2003, Bresnan et al. 2007, and Snider 2011) do show effects of definiteness in the dative alternation. As with other features discussed here, these effects can be summed up in a simple ordering preference: definite-before-indefinite order, as in the examples in (2.60), seems to be preferred.

(2.60) a. the man gives the little boy a green balloon

(Eng-UK-MOR/Fletcher/7/hnclap.cha, line 821)

b. giving the eggs to someone

(Eng-UK-MOR/Fletcher/5/hnmich.cha, line 602)

These studies are aware that definite objects are often, but crucially not always, given (see for example Gries 2003:12–13 or Bresnan et al. 2007:77–78). What is generally done is accounting for definiteness by including it in statistical models of the data, which are supposed to ensure no two strongly correlated variables both emerge as significant. The fact that both definiteness and givenness are found to be significant is then taken to mean that they cannot be strongly correlated. Stephens (2010:168) found “some [direct] evidence that givenness has an independent influence on the choice of construction”, but the connection between the two in the context of the dative alternation still has not been investigated fully.

As for child language, the literature shows that two- and three-year-olds are above chance in

tasks regarding definiteness and that the understanding of definiteness improves quickly with age (Maratsos 1974, Warden 1976, Zehler and Brewer 1980, Rozendaal and Baker 2008, and Stephens 2010:21–22). Malchukov et al. (2010a) also mentions the objects' definiteness as one of the factors influencing the dative alternation choice. As with grammatical person, the effects of definiteness appear to be less important than others. Nevertheless, the present study keeps all objects definite.

In the context of the dative alternation, the terms “syntactic persistence” (Bock 1986:359), “syntactic priming” (Gries 2005:365), or “structural parallelism” (Bresnan et al. 2007:77) all mean the same: it appears that the double object construction is a little more likely following a recent double object construction, and the prepositional construction is a little more likely after a prepositional construction. Gries (2005), Bresnan et al. (2007), and Snider (2011) found such effects in their adult corpus data; de Marneffe et al. (2012) found them in child speech and child-directed speech; and Shimpi et al. (2007) showed experimentally that three- and four-year-old children are more likely to describe depictions of transfer events⁴⁰ using the primed dative construction than with the non-primed alternative. Bock (1986) discusses priming in general as the activation of syntactic structures, whereas Wasow (2002:105–106) speculates that the preposition *to* alone may be enough to prime the prepositional dative construction. Jaeger and Snider (2013) show that speakers are primed more strongly towards a given structure if the utterance that primed them for this structure was unexpected, which they explain by speakers constantly trying to reduce prediction error in their own production. Whatever its exact cause, the effect of syntactic priming in the dative alternation is certainly established. This study therefore does not investigate it in detail—it is not manipulated, though it is controlled (see Sections 5.1 and 6.1 for details) so that a priming effect can be investigated if it arises unexpectedly.

Following Stallings et al. (1998)'s work on what they call the disposition of individual verbs to use the HNPS order (or structure) over the canonical one, Biber et al. (1999:928), Wasow (2002), Gries and Stefanowitsch (2004), Bresnan et al. (2007), and others have observed that some verbs are used with one dative alternation construction much more often than with the other. While these individual verb biases can be quantified very exactly (see for example Gries and Stefanowitsch 2004:106 or Bürkle 2011:40 and 42), the proposed explanations for these biases are much less exact. Gries and Stefanowitsch (2004) argue that the verbs with a bias for the double object construction prototypically encode primarily transfer of possession, while the

⁴⁰Strictly speaking, two of the ten verbs used by Shimpi et al. (2007), namely *buy* and *bake*, participate in the benefactive alternation rather than the dative alternation as defined in Section 2.2 above. (The other eight verbs are dative alternation verbs in the stricter sense; Shimpi et al. see 2007:1346.) However, I see no reason to assume that priming has no effect in the benefactive alternation or that benefactive priming is substantially different from dative priming. Shimpi et al. (2007) do not report different patterns for *buy* and *bake* versus the other verbs, most likely because there were no significant differences. Therefore, it seems reasonable to treat the results of Shimpi et al. (2007) as relevant here.

verbs biased towards the prepositional construction prototypically encode physical transfer or movement without a necessary change of possession (following Gropen et al. 1989; compare the discussion regarding interruptibility in Section 2.1 above). This does explain the biases of the two most distinctive verbs (Gries and Stefanowitsch 2004:106): *give*, which is strongly biased towards the double object construction, is the prototypical transfer-of-possession verb; *bring*, strongly biased towards the prepositional construction, does specify the (accompanied) motion without specifying possession at all. However, Gries and Stefanowitsch (2004:107) also point out the shortcomings of their proposed explanation: *sell*, *supply*, and *pay* intrinsically specify transfer of possession, yet they are biased towards the prepositional construction. Gries and Stefanowitsch’s proposed explanation is that these verbs prototypically describe physical transfer of goods or money. This allows them to be consistent with their example of the sale of a house as well as with the example in (2.61), but the fact remains that these examples and all other selling, supplying, and paying events always include transfer of possession.

(2.61) I sold him my car two weeks ago, and he still hasn’t come by to pick it up.

(2.62) These Girl Scout cookies cost me only three dollars!

The above examples show another problems with this proposed explanation: the assumption that *cost* “never involves motion” (Gries and Stefanowitsch 2004:107) is invalidated by any specifically cash-based transaction like (2.62), where money literally moves from buyer to seller. Gries and Stefanowitsch (2004)’s account of verb bias in the dative alternation thus fails to explain all the data and describes at best a tendency.

Szabóné Papp (2003) proposes a pragmatics-based approach: she says that two different construals of transfer events can account for the acceptability patterns of the two dative alternation constructions with different verbs. In this view, the double object construction represents a construal that places the pragmatic focus on the theme and thus emphasizes the transfer of possession of that theme (Szabóné Papp 2003:93, Fig. 11). The prepositional construction, on the other hand, represents the construal that focuses on the recipient, thus making it a “Transfer-Caused-Motion Construction” (Szabóné Papp 2003:91, Fig. 10). According to Szabóné Papp (2003), these two construals explain, for example, why *whisper* is not used in the double object construction and why *deny* is not used in the prepositional construction: as for *whisper*, it focuses on the way in which the message is conveyed⁴¹ and thus on the ‘motion’ of the message from speaker to hearer. The substance of the message is less important in a sentence using *whisper*, and it may even be lost: according to Szabóné Papp (2003:151), manner-of-communication verbs lexically “imply the existence of some kind of obstacle on the communication channel

⁴¹I should point out here that this is my interpretation of the phrase “the exact formulating” (Szabóné Papp 2003:106), but it is consistent with Szabóné Papp’s argument as a whole and therefore reasonable.

... , which puts the success of the transfer of message and the affectedness of the recipient at risk”. As for *deny* as well as *refuse* and *spare*, the two construals explain why they are not used in the prepositional construction: a theme of any of these three verbs never actually moves in the event described, so the motion construal and thus the prepositional construction are ruled out; and using the double object construction means that the theme and recipient are (or can be) placed immediately adjacent without an intervening preposition, which highlights “the personal loss of opportunity suffered” by the recipient (Szabóné Papp 2003:127). However, this pragmatic account fundamentally disagrees with data: the examples in (2.63) show that *whisper* and other manner-of-communication verbs are in fact used in the double object construction.⁴² The examples in (2.64) show that *deny*, *refuse*, and *spare* are in fact used in the prepositional construction.

- (2.63) a. ... a kind few ... whispered me the answer.
 b. The shepherd-dogs-dogs ... barked him a welcome
 c. You just mumble him an answer. (all Bresnan and Nikitina 2003:7–8, emphasis removed)
- (2.64) a. ... why it previously denied entry to Mr. Deripaska (tinyurl.com/deniedentry)
 b. the worker or workers who denied food to these kids should be fired
 (tinyurl.com/deniedfood)
 c. ... the Venetians, who ... refused the support to the army on land which they had given to their naval expedition (Oliphant 1893:240)
 d. Henry Gladwin ... refused the help to the Indians (tinyurl.com/refusedhelp)
 e. she spared that to her watchers. (tinyurl.com/sparedthat)
 f. and she would have spared all the suffering to her younger cousin Lillie
 (tinyurl.com/sparedsuffering)

Thus, as with the proposed pronominal patterns in Section 2.2, all that can be concluded is that these verbs are rare or dispreferred with one construction (Ford and Bresnan 2013). The meaning-based explanations for verb bias make predictions that turn out to be insufficient by themselves, and thus they fail to explain verb bias. It may be that the double object construction **tends to** be used in a sentence that emphasizes transfer of possession and that the prepositional construction **tends to** be used “to focus on the indirect object [meaning the recipient]” (Sinclair et al. 1990:160), but these tendencies do not appear enough to form a fully explanative theory. Therefore, individual verb bias is to be accepted as an empirical fact. This study only uses the verb *give* for independent reasons (see the discussion in Section 5.1.4), and thus verb bias is kept constant.⁴³

⁴²Note that the five double object examples of Bresnan and Nikitina (2003:7–8) all use verbs listed as prepositional-only in Szabóné Papp (2003:151).

⁴³Interestingly, a verb with a strong bias apparently “resists priming” (Gries 2005:379). In the case of *give*, this

2.7 First language acquisition

The aim of the present study is to document how speakers of English process and produce the two dative alternation constructions during their acquisition of English and to discover any patterns or paths of acquisition in that processing and production. This section contrasts the nativist, generative approach to first language acquisition with the frequentist or empiricist usage-based framework. After discussing these theories of first language acquisition, this section will also lay out how they relate to the dative alternation in particular.

Nativist theories argue from the “poverty of the stimulus” (Chomsky 1980:34): in their view, the sum of all the language input received by a child is logically insufficient for the linguistic generalizations that the child will later make; it follows that there must be another component to language acquisition beside the input, and nativists argue that some innate biological endowment is the obvious candidate. Furthermore, since all adult speakers of one language community have (more or less) the same grammar, language acquisition must stop at some point and thus yield a steady adult grammar. The innate component or “initial state of the mind might be regarded as a function . . . that maps experience into the steady state” (Chomsky 1980:187; see also Pinker 2004). This process of mapping is then specified in the Principles and Parameters approach (Chomsky and Lasnik 1993): all stages of language acquisition, including the final steady one, adhere to certain principles, and the differences between stages and between different adult steady stages, or languages, are entirely due to parameters being set to one value or another (see Hyams 1998 for crosslinguistic evidence).

Although a comprehensive and uncontroversial list of parameters has not yet been produced, it is generally agreed (and indeed obvious) that setting a parameter value requires a trigger or cue in the input. If the triggers for one parameter value are too infrequent (or absent) in the input, that value is not set (in a diachronic view, the value is lost; see Niyogi 2004). In effect, Principles and Parameters theory assumes that there is a cognitive system that uses the currently best grammar (or combination of parameter values) while also evaluating this and other grammars against the input. Parameters are mostly seen as binary; the value that has fewer formal features can be seen as the unmarked, ‘default’ one (Roberts 2007). Like values, entire parameters can also be more marked or complex than others. This is used to explain why parameters are apparently set successively, even though a simple sentence can arguably contain triggers for many different parameter values: at such a stage, it may be the case that “the overall system has not matured sufficiently . . . to permit certain parameters or parameter values to be attained”

means that the double object construction is more likely even after a prepositional prime (Gries 2005:376). It should be noted, however, that Gries (2005) is a corpus-based study, whereas the priming in this study is systematically varied. The present study is therefore better suited for investigating the existence of a priming effect and any resistance of that from verb bias.

(Roberts 2007:211).

The same point is also recognized in Lightfoot (2010)'s cue-based variant of nativist acquisition theory. This approach does not use parameters, but instead restricts the learner's cognitive operations on linguistic representations to those that are innate and those that the input contains enough evidence or cues for. In this view, complex operations will often operate on the output of simpler ones. Therefore, the cues for the simple operations have to be recognized before the more complex cues can even be found (Lightfoot 2003:7), moving from simpler to more complex operations and from core to periphery. This theory also offers an account of how language can change in a strongly nativist approach: the nativist argument is an answer to the question how language acquisition can so consistently yield a system that is the best model for the input. A priori, this assumption would also lead us to expect that the output of that system would in effect mirror the input (for an infinitely large sample) and thus that there is no language change; yet there is. If, however, the input changes (for whatever reason), then the cues expressed by it would of course also change, as would the system learned from them—in other words, “grammatical shifts are to be explained **only** by a prior change in the trigger experience” (Lightfoot 2003:18, original emphasis). To summarize, nativist theories of language acquisition posit innate language-specific cognitive mechanisms in order to explain the apparent contradiction between the limited input and the unlimited generative capacity and changeability.

Frequentist or empiricist theories, on the other hand, argue that it is pointless to find explanations for such apparent contradictions because the nativist generative approach is inherently flawed in several respects: it assumes psychologically real generalizations about (or rules of) grammar without having any evidence for them (Bowerman and Croft 2008:287–292), it fails to explain children's use of learned expressions as wholes (Abbot-Smith and Tomasello 2006), and it ignores the flexibility of adults' grammars (Bybee 2010:112–114).

The nativist argument from the poverty of the stimulus can be simplified into the following form: innate domain-specific language learning mechanisms are necessary to infer general grammatical rules from limited input. This formulation makes more apparent the fact that this necessity would not exist if there were no categorical rules at the end of the acquisition process: if there are no psychologically real grammatical generalizations, then no specialist cognitive mechanism is needed to infer them. Bowerman and Croft (2008) argue that the absence of evidence for psychologically real rules of grammar is enough to reject the premise that such rules exist. This is perfectly parsimonious, of course, but stronger evidence for this absence of categorical grammatical rules is naturally difficult to find. For example, Bowerman (1994) shows that it is not the case that all children use one and the same innate set of motion categories, even at a very early age, and argues that this is evidence against innate categories at least in the domain of motion. However, it may be that all newborn children have the same set of motion

categories (and other linguistic mechanisms) and that they have subsequently learned enough to deviate according to language by age two.⁴⁴ Therefore, Bowerman (1994)'s findings do not constitute unequivocal evidence against grammatical rules.

The case against psychologically real grammatical rules is also made by Paul (1880). It is axiomatic in linguistics that speakers form some categories of, or associations between, stored forms in their minds. Paul points out that this fact does not mean that these associations are identical with grammatical concepts, "even though they usually coincide with them" (Paul 1880:31, my translation).⁴⁵ In other words, his argument is that the fact that a language can be described using concepts like 'verb', 'case', and 'collocation' does not automatically mean that these concepts have psychological reality.⁴⁶ Thus, Bowerman and Croft (2008)'s argument against the poverty of the stimulus is sound—if one accepts the premise that there are no psychologically real grammatical rules.

Generative nativist theories assume that learners analyze input and (re-)generate output based on what their innate language processing system has been able to extract from the input. Empiricist theories recognize the importance of direct imitation in early language acquisition: in the view of Abbot-Smith and Tomasello (2006), generalizations are formed slowly out of many single items that 'overlap' in one way or another. The elements that are shared by all these items (for example, *do* in second position in *When do you want to eat?*, *Why do we think that?*, *Where do they come from?*, and so on) are retained in the generalization, whereas the differences 'blur' and thus leave only their common features behind (wh-word before the *do* and VP after it, in the previous example). This process, however, requires many examples and is therefore slow. Abbot-Smith and Tomasello (2006:283) argue that much of children's early speech must therefore be made up of exemplars that are very frequent as a whole in the input (*What's that?* or *What do you want?*, to continue the example). Support for this comes from Dąbrowska and Lieven (2005)'s study of questions in early speech, which shows that most of them are in fact imitations of previously heard questions (as much as 75% in one case; Dąbrowska and Lieven 2005:450), and from Lieven et al. (2009)'s corpus study of early child language in general, which shows that the majority of two-year-olds' utterances (as much as 92% in one case; Lieven et al. 2009:492–493) are either exact imitations of a previous utterance or can be reduced to a previous utterance with no more than one simple substitution or addition. These findings thus

⁴⁴This is of course one of the difficulties in showing the absence of innate grammatical rules: people who have only the innate mechanisms without the benefit of any subsequent learning—newborn children—do not make good subjects of linguistic studies, and therefore older children are often chosen as subjects.

⁴⁵In Paul (1880:31)'s original words: "... *wenn sie sich auch gewöhnlich mit diesen decken*".

⁴⁶Psychological reality of linguistic structures, processes, or rules is frequently assumed tacitly. There are some explicit claims of empirical evidence for it, for example with CPs in some Semitic languages (Friedmann 2001), NP traces (McElree and Bever 1989), and rules concerning theta attachment and center-embedding (Sadeh-Leicht 2007). It is important to note that such evidence is a matter of interpretation, at least to some degree. For one argument against psychological reality of linguists' constructs, see Barrios (2012).

provide empirical support for the above argument against grammatical rules: at least for young children's speech production (if not their internal representations), generative rules seem largely unnecessary.

Finally, some empiricist approaches like Yang (2004) and Bybee (2010) discuss the notion of 'grammatical rules' more intensively. The original argument from the poverty of the stimulus assumes that innate language-specific mechanisms or knowledge are needed to explain how all speakers of one language at one time converge on the same grammar when there logically are many grammars that can analyze and generate the same data equally well (Berwick 1985:235–238). These empiricists argue that there is no evidence for all speakers in a language community actually sharing the very same grammar (see Kayne 2012:Section 2). Yang (2004:51), for example, argues that “adults speakers, at the terminal state of language acquisition, may retain multiple grammars, or more precisely, alternate parameter values”. Bybee (2010:112–114) goes further than that: in her view, adult grammars are not “discrete and unchanging” (Bybee 2010:114), but rather statistical and flexible. Yang (2004) combines statistical learning with Principles and Parameters: his theory of acquisition uses different parameter settings to create different (competing) grammars in a learner's mind. Those grammars that do not allow for an item of input are penalized, which over time leaves only a few 'surviving' grammars. The adult speaker then has access to these grammars only. Bybee (2010:64–66)'s theory of acquisition is also statistical (or exemplar-based), with abstract patterns emerging from the repetition and modification of stored exemplars. These patterns are not above the variations of them, however: “the same factors operate to produce both regular patterns and the deviations” (Bybee 2010:6). In this theory, variation is constrained by how it spreads: since exemplars are associated with each other in a rich network, language change tends to affect all forms that are similar in a relevant way. That means that the variation has to make sense, as it were, for all those forms. In first language acquisition, variation appears to be limited to “consistently” replicating variation in the input (Bybee 2010:117). However, because she does not assume fixed adult grammars, she can explain variability in adult speech with just one (flexible) grammar as the end result of acquisition. Of course, a flexible grammar does not have to change dramatically to be flexible, and dramatic changes are intuitively unlikely in adult grammar. This important point is also realized by the more fleshed-out theories of Batali (2002) and Bod (2006), which will be discussed below.

It is worth pointing out that Batali (2002) technically describes an exemplar-based modeling and simulation system for the negotiation of a shared language, not a theory of the evolution of language or its subsequent acquisition by following generations. This system is nevertheless deserving of discussion because it is quite detailed and because, through that level of detail, it reveals a fact about exemplar-based approaches that is central to the discussion here.

In Batali (2002)'s modeling system, two simulated agents are picked out of a larger population

at random. One is given a message to transmit to the other and has to use its stored message-to-signal formulae to find the signal that will best express the message to the other agent. This receiving agent then uses its own formulae to find the most likely meaning for that signal. Batali demonstrates in some detail how this gives rise to a stable but flexible language over time even if the agents are given no ‘words’ at all to begin with: through random string generation and imitation, words spread; given enough time, even reflexive and passive markers as well as phrase or clause delimiters and complementizers arise spontaneously in simulations of this system. One might argue that giving the receiving agent both the signal and the intended message in 90% of simulated interactions (Batali 2002:138, 166) invalidates the system as a whole, since hearers of natural language utterances need to infer the message from the signal. However, the 10% of interactions where only the signal was given to the receiving agent were explicitly used as testing interactions; and the intended message may well be inferable from non-linguistic clues in many cases, at least in early language evolution (and acquisition). The relevant point is a different one: this exemplar-theoretic system is nativist.

Although the agents are not given words to start with, they do start out with a full understanding of the messages that they have to express in the simulation. This includes not only the simple words, but also the two-place predicates (verbs, in effect). The simulated agents know how these message segments combine to form tree-like structures. In other words, the agents have an ideal and shared innate grammar of meanings and negotiate a grammar of strings based on that. Furthermore, they all share the same operations for combining the strings of the language they are to create. Though Batali (2002) does not say so, it can be inferred from his use of these innate operations that they are necessary for the simulation to work. This is not an argument against his simulation system, of course, but it supports nativist approaches. The arguments against this system being nativist are not strong enough to remove this support: firstly, while it is arguably true that the innate operations are “similar to those proposed in the domains of perception, analogical reasoning, and planning” (Batali 2002:128), there is no proof that they are identical. To dispel domain-specific nativism, this identity of operations is crucial, however. Secondly, the fact that these operations “might occur as solutions to the problem of encoding complex meanings into linear sequences” (Batali 2002:129) does not mean that they do for each real-world agent (speaker) individually. Finally, the argument that “deficiencies of the model might be fixed by modifying the agents’ system of internal representations, the algorithms and data-structures underlying their communicative behavior, and/or their learning mechanisms” (Batali 2002:169) only if these changes “seem biologically plausible, **or** of general cognitive utility” (Batali 2002:169, emphasis added) is obviously not an argument **against** this exemplar-based system being nativist, but rather **for** that.

The Data-Oriented Parsing model of Bod (2006:307–315) is based on Lexical Functional

Grammar (LFG), but it exhibits the same inconsistency. Briefly, it assumes that language acquisition and processing are cognitively realized by representations being disassembled (partially or completely) and reassembled from these fragments. This is achieved by strict “decomposition” and “composition operations” (both Bod 2006:295). Bod shows that acquisition based on this can be statistical or exemplar-based. However, this model relies on “grammatical functions and semantic roles” (Bod 2006:317) pre-existing in the learner’s mind. These are called “Universal Representation” (UR) by Bod (2006:318), and he argues that this differs from the idea of Universal Grammar in the result: a UR-based grammar can be statistical, whereas a UG-based grammar is restricted to categorical principles. Thus, the difference between nativist and empiricist theories of language acquisition is in the (possibly psychologically real) representation of grammatical knowledge. The mechanisms or operations that give rise to and apply to this knowledge apparently have to be somewhat specific to language, and they have to exist in the learner’s mind before acquisition begins. Thus, Pinker (2004)’s conclusion that most (if not all) linguistic frameworks are ‘innatist’ to some degree is borne out:

“For the behaviourists, the innate constraints reside in the generalization gradients and response classes. For the connectionists, they reside in the features defining the units and the topology of the networks. For Chomskyans, they reside in categories, operations, and principles. For MacWhinney, they reside in the cues, items, alternatives pitted in competition, and categories whose absence constitutes ‘indirect negative evidence.’” (Pinker 2004:949)

To summarize this discussion, there is evidence for statistical or item-based learning as well as a theoretical need for some component of the human language faculty to be innate. ‘Compromise’ theories that incorporate this have been proposed: those discussed above obviously fit the description, but Yang (2004), for example, also arrives at a compromise theory after starting with a Principles and Parameters approach. He does theorize that learners maintain several grammars and use the input to constantly evaluate them, but the system allows for statistical rather than categorical learning. Abbot-Smith and Tomasello (2006) argue that the human language faculty can turn exemplar-based frames or constructions into more abstract categorical rules once they are frequent enough to allow the necessary generalizations. A similar model of item-based learning that leads to generalized rules is described in Bowerman and Croft (2008:293–296).

The rest of this section will briefly explain how different acquisition theories relate to the dative alternation. As shown in Sections 2.3 to 2.6, the current best view of the dative alternation approaches it as a choice between two constructions that is guided by an interaction of many factors of varying strengths. Thus, all that is necessary to embed the acquisition of the dative alternation into an acquisitional framework is to explain how these factors come to influence the

dative alternation and how their strengths can change over time.

Logically, there are two possibilities for how the factors start to have their effect: either they are innately effective, or they are added over the course of acquisition. As shown above, the factors that are relevant for the dative alternation are known to have other effects in the language of quite young children. Therefore, they could in principle also have their effects on the dative alternation from as soon as children produce ditransitive utterances. If the language faculty comes pre-equipped with these factors affecting the dative alternation, this is of course trivial; if the factors are added on the basis of evidence in the input, it appears that there is enough evidence for them before ditransitives become productively used. Thus, the data in principle does not seem to allow a distinction between these two possibilities. Although de Marneffe et al. (2012:42) explain the absence of an animacy effect in their child language model as an effect of certain verbs being overrepresented in their corpus, the results of Bresnan et al. (2007) and de Marneffe et al. (2012) might be used to support the notion of the animacy effect only coming into existence later during acquisition. The strongly innatist view of the effects in the dative alternation could, however, even explain that: it only says that the factor is available immediately, not that it has a non-zero effect strength immediately.

This shifts the burden of explanation to the changing effect strengths: on the basis of existing data, both possible explanations can be made to agree with the idea that “child speech only differs from the speech of their adult interlocutors in degree, not in kind” (de Marneffe et al. 2012:54). For an empiricist, change in effect strengths is unproblematic: it arises naturally from children making statistical generalizations based on the input. After hearing many dative constructions with the shorter object before the longer one, for example, a frame that contains this ordering preference would be more likely to be used in constructing dative utterances. For a nativist, change in effect strengths is less easy to explain. One possibility would of course be to locate one of the dative alternation constructions in the core grammar and remove the other construction to the periphery, and then let the (peripheral) choice between them be susceptible to change. Hinterhölzl (2004) argues that this is enough to account for diachronic changes in constituent order based on information structure. The dative alternation can easily be viewed as an information-structural constituent order alternation, of course, but it is not at all clear that it is an epiphenomenon of some current diachronic change from one construction to the other. Moreover, while minor variation in effect strengths could be reformulated in this approach as changes in thresholds (see Hinterhölzl 2004:145), not all features that appear to affect the dative alternation are discrete enough to allow for a threshold to be formulated: for animacy, it could be a certain level on the animacy hierarchy; but what could a threshold for pronominality or givenness be? Postulating a very large number of mental grammars would be another possibility for a nativist explanation of changing effect strengths, where each possible combination of effect

strengths is realized in one grammar. However, considering the fact that there are many other (unrelated) parameters or cues to be acquired at the same time, this mass of grammars would have to be further multiplied to yield one grammar for each combination of dative alternation effect strengths and all possible values of unrelated parameters. This profusion of grammars is intuitively unlikely.

A more reasonable possibility would be to reduce the myriad combinations of effect strengths to a ranking of effects: the effects that are based on features of the objects can be reduced to simple ordering preferences, which could then be ordered as in Optimality Theory. Bresnan and Hay (2008)'s finding that speakers of New Zealand English are more likely than speakers of American English to use the double object construction with inanimate recipients could thus be rephrased to say that the constraint *INANIMATE-BEFORE-ANIMATE is ranked lower in New Zealand English. Existing OT approaches to acquisition and change (see for example Hendriks and van Rij 2011) could then be used to account for the development of constraint rankings in acquisition and diachrony, and the same constraints could even be useful in accounting for HNPS, the benefactive alternation, and other reordering phenomena (see Section 2.2 and Wasow 2002). Even the variability of the (fully learned) dative alternation could be modeled if Stochastic OT were used, and Lin (2005) shows with computational learning simulations that Stochastic OT is theoretically learnable. However, OT only aims at descriptive power and is specifically not claimed to have any psychological reality (Prince and Smolensky 2004:232–233). Thus, while it is in principle possible to account for the dative alternation and its acquisition in OT, there is no justification for doing so, and no explanatory power is gained. Therefore, the best way for a nativist to account for changing effect strengths is not OT or some other way of implementing a large number of mental grammars, but rather a 'compromise' theory that includes statistical learning and variability as described above.

3 Research questions

Following Chapter 2's review of the extensive literature investigating the dative alternation, this chapter presents the five main research questions of this thesis. The first three concern the effects of specific features on the dative alternation as an ordering choice, the fourth is about the order of acquisition or emergence of these effects, and the fifth arises out of the methodology for investigating the other four questions (described in detail in Section 5.1.2). These research questions were addressed with a categorization experiment (experiment 1, Chapter 4), an interactive visual world presented on a touchscreen (experiment 2, Chapter 5), and an elicited production task (experiment 3, Chapter 6). All three experiments were combined into a single experimental session (each lasting around 45 minutes), in the order as they are presented here: participants completed experiment 1 before moving on to experiment 2 and finished with experiment 3. This means that all three experiments were administered to the same participants, and each participant was presented with all three experiments in a single session.

3.1 Is there any evidence for an animacy ordering preference?

Previous studies have found that adult speakers produce more dative alternation sentences with an animate object before an inanimate one than sentences with an inanimate object before an animate one (Bresnan et al. 2007, Bresnan and Hay 2008). In other words, an animate-before-inanimate preference is apparent in the dative alternation. This preference has not been attested in child language, but this may simply be due to limitations of the data (de Marneffe et al. 2012:42). Experiment 1 of the present study establishes children's and adults' concept of animacy. Experiment 2 is aimed at determining whether children and adults expect animate-before-inanimate order in the dative alternation and whether they conform to this animate-before-inanimate order when filling gaps in dative alternation sentences. Experiment 3 tests for this preference in production by investigating whether speakers can be made to reproduce inanimate-before-animate sentences just as easily as animate-before-inanimate sentences.

3.2 Is there any evidence for a number ordering preference?

The same recent literature indicates that adult speakers appear to prefer placing a plural object before a singular one over the reverse ordering. No study has attempted to show this ordering preference in child language. Therefore, experiment 2 of this study is also aimed at determining whether children and adults expect plural-before-singular order and whether they conform to plural-before-singular order when filling gaps in dative alternation sentences. Experiment 3 tests

for this preference in production by investigating whether speakers can be made to reproduce singular-before-plural sentences just as easily as plural-before-singular ones.

3.3 Is there any evidence for a length ordering preference with a minimal length difference?

The preference for shorter objects before longer one is attested so well that it is taken as a given in most recent studies of the dative alternation: length is either treated as a factor that affects speakers' choices (see for example de Cuypere and Verbeke 2013), or it is kept constant by experimental design to rule out its influence (see for example Stephens 2010:45). The length effect is generally seen as a monotonic effect: smaller length differences between the two objects lead to a weak preference for the shorter-before-longer order, larger length differences cause a stronger shorter-before-longer preference. However, we do not know the smallest amount of linguistic material that has to be added to one object to make it 'longer' in the eyes of this length effect. Experiment 2 of the present study therefore also tests whether the difference between monosyllabic and bisyllabic words is large enough for speakers to expect objects in monosyllabic-before-bisyllabic order and whether they conform to monosyllabic-before-bisyllabic order when filling gaps in dative alternation sentences.

3.4 Do these preferences emerge in a particular order?

Viewed as a construction choice that is governed largely by ordering preferences relating to the features of the objects (Bresnan et al. 2007:81), the dative alternation is a complex system. Its acquisitional trajectory has not been mapped yet—we do not know whether children start by using some of these features and add others into the system gradually, or whether even their earliest expectations and choices are influenced by the same features that influence adults' expectations and choices. All three experiments in this study were completed by participants from three different age groups (four-year-olds, eight-year-olds, and adults—see Section 4.1.1 for details), which means the results can be compared across (apparent) developmental time. The age of four years was chosen because four-year-olds can be expected to be able to use *give* and both dative alternation constructions (see Gropen et al. 1989:213–216), and the age of eight years was chosen because it and the age of four bracket many developmental milestones. Thus, if there is an acquisitional trajectory for dative alternation, these two age groups can be expected to show it.

3.5 Does touchscreen input reflect attention?

It is an axiom of eyetracking methodology that gaze reflects attention—people will generally look at what they are currently attending to (reading, processing, and so forth). This correlation between gaze and attention extends to interaction. When interacting with an ‘intelligent’ home or office, people overwhelmingly look at the specific device they are trying to interact with at the time (Brumitt and Cadiz 2000, Maglio et al. 2000). Pressing a specific button on a smartphone touchscreen also requires looking at the screen, whereas performing touch-input gestures (such as a curly ‘pigtail’ shape to delete an item) does not (Bragdon et al. 2011). For touchscreens specifically, a close correlation between touch location and gaze target point is commonly assumed: the patents of Kawalkar (2011) and Sukumar (2012) unobtrusively filter out unintentional touch inputs by accepting input only when both touch and gaze are on (or near) the same touchscreen button. In other words, they are based on the assumption that the (intentional) touch of a target point will be accompanied by a gaze at the same target. Similarly, Hagiya and Kato (2014) show that asking people to touch one target while gazing at another, distant on-screen target leads to larger variance in the gap between touch target and touch input—in other words, when people cannot look at what they are touching, they are more likely to miss it (and miss it by more). This reflects the everyday intuition that finer motor control (which most touchscreen input will be) requires closer visual attention for hand-eye coordination. The success of these patents and inventions shows that touch and gaze are closely correlated, without any need for training users to look where they touch.

However, these findings and implementations are based on general, common-sense assumptions. I am not aware of any quantitative studies of the distance between simultaneous gaze and touch during touchscreen use. As the interactive task in the present study records eye gaze during touchscreen interaction (see Sections 5.1.2, 5.2.1, and 5.2.2 for details), it will supply such data in addition to addressing the linguistic research questions. This data will show whether the above assumptions are justified. If they are, future technical designs can rely more robustly on touch and gaze being close, and researchers in the cognitive sciences can use touchscreen input as an alternative attention measure. If they are not, further research into gaze during touchscreen use will be necessary in order to replace these common-sense assumptions.

4 Experiment 1: Categorization

The aim of this first experiment is not to answer any of the five research questions (Chapter 3) directly. Rather, it tests how participants categorize the animals and inanimate objects used in experiment 2 (Chapter 5) by asking whether these animals and objects are capable of moving and playing. One of the features of interest in experiment 2 is animacy, but it would be presumptuous to use the simple a priori categories of animals and inanimate objects without checking whether participants (especially children) have those same categories.

4.1 Methodology

4.1.1 Participants

Adult participants (N = 22; 18 female and 4 male;⁴⁷ age range 18 to 41 years, median age 21 years) were recruited through notices posted on campus and on the online course platform Learn at the University of Canterbury, Christchurch (New Zealand). Their participation was incentivized with a NZD 10 shopping voucher.⁴⁸ These adult participant were analyzed as one age group. The children formed two age groups: four-year-olds (N = 20, 10 female and 10 male, mean age 4;3) and eight-year-olds (N = 20, 10 female and 10 male, mean age 8;4). They were recruited through Christchurch kindergartens, schools and after-school programs as well as home education networks, the New Zealand Institute of Language, Brain and Behavior's participant pool "Team Tamariki", notices posted on campus and on the online learning platform, and word of mouth. Child participants were incentivized with their choice of one item from a 'box of treasures' (containing things like toy cars, bags of balloons, and sheets of stickers; monetary value less than NZD 5 each); since parents or caregivers had to accompany the child participants to the experiment at the university because of ethical and legal considerations, they received a NZD 10 fuel voucher.

All participants were being or had been raised in New Zealand. All participants named English or New Zealand English as their first and home language, and the impression of the experimenter, herself a native speaker of New Zealand English, was that all participants did indeed speak New Zealand English. Adult participants consented to participate; for child participants, their assent (where possible) and their caregivers' consent was obtained before the experiment. No

⁴⁷The adult participant group was not balanced for gender. Among the studies of the dative alternation that I am aware of (see Chapter 2), none report a gender difference within the realm of the dative alternation. Therefore, none was expected here.

⁴⁸These vouchers were funded by the University of Canterbury's School of Social and Political Sciences (since reorganized and renamed).

information or data apart from that on these forms and participants' categorizations, touchscreen input and eye gaze, and sentence reproductions (see Sections 4.1, 5.1, and 6.1 for details) was recorded.

Participants and parents were given directions to the room where the experiment was conducted when they scheduled an experiment session. On their arrival, they were greeted by the experimenter, a female native speaker of New Zealand English experienced in talking to and working with children. Child participants in particular were given time to get used to the room and equipment. Seats and reading material (magazines and university publications) were available for parents and other accompanying persons. Some parents brought other children beside the participant along; toys as well as picture and coloring books were available for these children, who were encouraged not to disturb the participant. Participants and parents were given information sheets to read⁴⁹ as well as the opportunity to ask questions about the experiment. Once they had read them and their questions had been answered, participants and parents signed consent or assent forms. These sheets and forms are reproduced in Appendix F.

4.1.2 Procedure

This experiment was the first in the experimental session for all participants. It was designed to test their concept of animacy. Participants were asked to take a seat and shown a succession of laminated print-outs of all the images that were used to represent objects and animals in experiment 2 (see Chapter 5 for a description of that experiment, and Appendix C for the pictures used). Only the pictures for singulars were used in experiment 1, since no animacy distinction between one dog and three dogs, or between one key and three keys, was expected. For each image, participants were asked "Could this one move towards you?", and the images were sorted into piles according to participant categorization. The same task was then repeated with another set of laminated printouts of the same pictures, this time with the question "Could this one play with you?". Thus, participants made a second, separate set of piles for this second question. These 'yes' and 'no' piles (four in total) were kept separate, and lists of the images in each one were made after the end of the session. The printouts were then shuffled (in two stacks, one for each question) again to randomize the order of presentation for the next participant. To avoid giving phonetic cues, which might have led to categorization based on similar-sounding names, the pictures had no labels on them and were not named by the experimenter. Color photographs were used instead of the actual household objects and stuffed animal toys in order to rule out the possibility of some participants treating the stuffed animals as inanimate objects and others treating them as (representations of) animals. This categorization task was used instead of a

⁴⁹The information sheets and assent forms for child participants were kept very simple and were printed in a slightly larger size, since that helps with comprehension in children as old as eight (Katzir et al. 2013).

more usual experimental procedure (such as a Likert scale with several options from ‘absolutely able to move towards me’ to ‘absolutely unable to move towards me’) because I expected the youngest participants would have problems with those (unfamiliar) procedures, but would easily be able to stack pictures.

This experiment was approved by the Human Ethics Committee of the University of Canterbury (reference number HEC 2013/166).

4.1.3 Limitations

The methods of participant recruitment could be argued to limit the possible pool of participants in unwanted ways. The adult participants were all recruited on campus. While the campus is open to the public, and data about participants’ enrolment status and educational history were not collected, this method of recruitment together with their shared language background (as speakers of New Zealand English) and ages (young adults on a university campus are most likely university students, and most university students will be from a relatively narrow age range) make it entirely reasonable to assume that all the adult participants were in fact university students of a Western background. This population is not typical for humankind, which creates problems when results based on it are generalized (Henrich et al. 2010). However, the study was by necessity limited to speakers of New Zealand English, and the adult participants are a reasonable sample of that particular population (considering that in 2006, 40% of New Zealanders aged fifteen and over held a post-secondary qualification). Most of the adult participants in the present study were female, but gender was not expected to have an effect, as argued above. Therefore, this gender imbalance in the sample was not expected to negatively affect its representativeness for the phenomena under investigation. The criticisms in Henrich et al. (2010) are also more applicable to behavioral patterns in psychology than to (sub-conscious) linguistic effects of animacy, number, and length.

The recruitment of child participants through schools and research participants pools also had the potential of creating a biased sample: as parents had to take some initiative in responding to the recruitment letter, all participants could conceivably share a background that is more open to science than the population average, which could in turn affect behavior. However, as with the adult participants, it is important to note that the measured behavior was unlikely to be strongly and directly affected by affinity to science. This experiment was designed specifically to measure participants’ concept of animacy without assuming a target or ‘correct’ answer based on the biological definition of life, which of course would be affected by affinity to and knowledge of science.

4.2 Results

Participants' answers to the two questions asked in this experiment ("Could this one move towards you?" and "Could this one play with you?") are shown in Fig. 4.1, where each dot is one of the 28 images (or nouns). The x-axis shows the percentage of responses to the question "Could this one move towards you?", and the y-axis shows the percentage of responses to the question "Could this one play with you?". To use the most striking noun, *ball*, as an example, 12 four-year-olds (of 20 total, or 60%) answered "yes" to "Could this one move towards you?", and 11 (55%) answered "yes" to "Could this one play with you?". It is clear why *ball* in particular (being a toy or sporting implement designed to roll) would get this response. Apart from this special case, this categorization experiment distinguishes animals (orange dots in Fig. 4.1) from inanimate objects (blue dots) very clearly. Table 4.1 gives the answer counts together with Yule's coefficient of association Q (which ranges from -1 for perfect negative association to 1 for perfect positive association) for each age group. The association between the two categorizations is strong and positive.

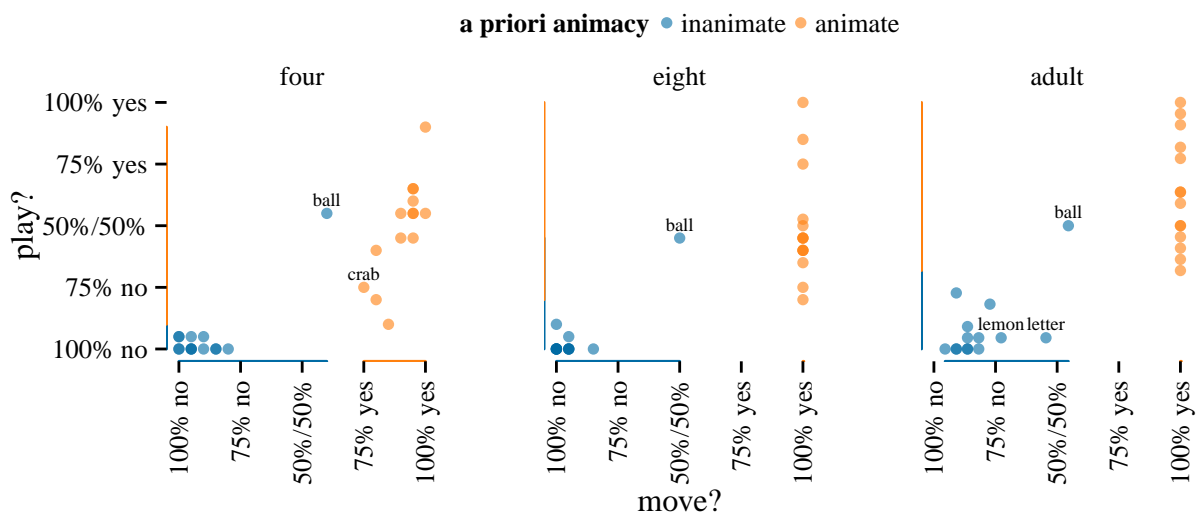


Figure 4.1: Categorizations by age group (each dot represents one noun; more solid dots represent several nouns with the same percentages for both categorization questions)

The question "Could this one move towards you?" by itself divides the nouns very cleanly, and this clean split replicates the a priori distinction between animals and inanimate objects (in fact, no animal got a single "no"-response to this question from eight-year-olds and adults, as is apparent in Fig. 4.1). Table 4.2 shows that there are very few cases of inanimates being labelled as capable of motion, or of animates as incapable, and the Q coefficients confirm that this is a

	cannot move	can move		cannot move	can move		cannot move	can move
cannot play	266	142	cannot play	254	153	cannot play	231	164
can play	10	142	can play	6	144	can play	17	204
(a) Four-year-olds ($Q = 0.93$)			(b) Eight-year-olds ($Q = 0.95$)			(c) Adults ($Q = 0.89$)		

Table 4.1: Categorization answers to both questions, and Yule’s coefficient of association Q

perfect positive association within empirical variation.

	a priori inanimate	a priori animate		a priori inanimate	a priori animate		a priori inanimate	a priori animate
no	250	26	no	261	0	no	248	0
yes	30	254	yes	18	280	yes	60	308
(a) Four-year-olds ($Q = 0.98$)			(b) Eight-year-olds ($Q = 1$)			(c) Adults ($Q = 1$)		

Table 4.2: Categorization answers to “Could this one move towards you?” by a priori animacy, and Yule’s coefficient of association Q

The answers to “Could this one play with you?”, on the other hand, are more varied. Apart from *ball*, the a priori inanimates (inanimate objects like *ball* and *lemon*) also received mostly “no” answers to this question, but the animates are much more divided (see Table 4.3). While there is an association between these responses and the a priori categorization, it is weaker for this question (Q coefficients around 0.9). The association is mostly driven by the a priori inanimates, which overwhelmingly received “no” answers, while the a priori animates received almost exactly as many “no” as “yes” answers from four- and eight-year-olds.

	a priori inanimate	a priori animate		a priori inanimate	a priori animate		a priori inanimate	a priori animate
no	265	143	no	267	140	no	282	113
yes	15	137	yes	12	139	yes	26	195
(a) Four-year-olds ($Q = 0.89$)			(b) Eight-year-olds ($Q = 0.91$)			(c) Adults ($Q = 0.90$)		

Table 4.3: Categorization answers to “Could this one play with you?” by a priori animacy, and Yule’s coefficient of association Q

4.3 Summary

This categorization experiment revealed that children and adults will categorize animates and inanimates fairly predictably with regard to motion and much less predictably with regard to

ability to play. This may be due to the fact that motion is fairly limited in its interpretation while the questions about the ability to play afforded more interpretations. The question about the ability to play is one of Gelman et al. (1983)'s reciprocal action questions. Their four-year-old participants replicated the a priori animate–inanimate distinction about as well with reciprocal action questions as with simple action questions (like “Can this one move?”). The results of experiment 1 paint a different picture: across all age groups, answers to this reciprocal question were more varied than answers to the simple motion question, with no clear binary split between animates and inanimates emerging from just the answers to the reciprocal question.

The answers to these two questions taken together, however, do show that all three participant groups categorize the animals and inanimate objects used in experiment 2 into two distinct groups. This supports the argument that animacy-related differences found in that experiment are really due to a categorical difference between animate and inanimate. For studies of gradual animacy phenomena (involving for example an animacy hierarchy), measuring participants' animacy conception with several questions is recommended, as it will result in more fine-grained animacy values than simple a priori categories would.

5 Experiment 2: Reaction and completion

This chapter presents an experiment that addresses all five of the research questions given in Chapter 3. Participants' choices for filling gaps in the instruction sentences of an act-out task were elicited to measure their expectations and preferences for animacy, grammatical number, and length of the words in these positions depending on the animacy, grammatical number, and length of the other (non-gapped, explicit) object and the order of the two objects. Their eye gaze was also recorded to provide an online measure of expectations and preferences. As this experiment uses a touchscreen and an eyetracker at the same time, it also provides data to assess the usefulness of touchscreens as attention-measure devices.

Section 5.1 describes the methodology used in this experiment, and Section 5.2 introduces and discusses the statistical methods used to analyze the results in Section 5.3.

5.1 Methodology

5.1.1 Participants

The participants of experiment 1 (22 adults, 20 eight-year-olds, and 20 four-year-olds) all participated in this experiment as well.

5.1.2 Procedure

This experiment followed immediately after experiment 1 in the same session. Participants were seated at a desk with a HP EliteBook 2740p 12.1-inch touchscreen computer (displaying 1280 by 800 pixels) and a Tobii X120 head-free eyetracker. The experimenter explained to them that they would be moving images on the touchscreen by simple touching and dragging. The experimenter also explained that the eyetracker would be recording their eye gaze. After eyetracker calibration, the task (run in PsychoPy, version 1.80.00;⁵⁰ see Peirce 2007 and 2009) was this: after a fixation dot (presented in the center of the screen for 500 ms), three images were shown on a black background in a horizontal row, either near the top of the screen (as 'themes') or near the bottom (as 'recipients'). Their relative positions were randomized per participant. To reinforce that they were intended as recipients, the 'recipient' image or images were near

⁵⁰The timing errors that Garaizar et al. (2014) and Garaizar and Vadillo (2014) report with PsychoPy were only present with stimuli presented for a much shorter time than any of the stimuli presented in the present study, and with non-optimal programming (Garaizar et al. 2014 did not use the more exact frame timing; Garaizar and Vadillo 2014, which found fewer problems, and the present study did). Therefore, it is reasonable to expect no timing errors here.

the bottom of the screen, had a white frame around each of them, and were not moveable (see Figs. 5.2 and 5.3). When one of the three images was first touched, a sound recording of the respective noun spoken by a female New Zealand English speaker was played over headphones (Moshi VLH for participants with smaller heads, Panasonic RT-HT 161 for participants with larger heads)—in the three-themes trial shown in Figure 5.1a, a recording of the word *baskets* was played the first time the image of three baskets was touched, a recording of the word *camel* was played the first time the image of the camel was touched, and a recording of the word *lock* was played the first time the image of the lock was touched. In the three-recipients trial shown in Figure 5.1b, *crab* was played when the image of a crab was touched for the first time, *letter* was played when the image of a letter was touched for the first time, and *squirrels* was played when the image of three squirrels was touched for the first time. (This was independent of the order of these first touches.) This was done to ensure that participants registered all three objects and to prime the intended nouns (so that participants thought of the rabbit as a *rabbit* instead of, for example, a *bunny-wabbit*).

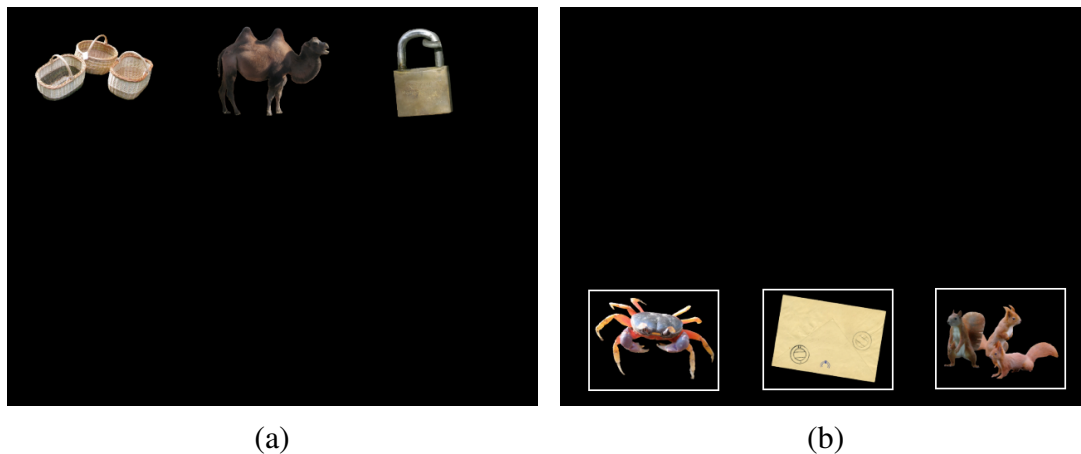


Figure 5.1: Initial state of trials

After all three of these images had been touched and the corresponding recordings had played, the larger image appeared. This image was centered horizontally and positioned near the bottom of the screen (as a ‘recipient’) if the three smaller images were ‘themes’, or near the top (as a ‘theme’) if the smaller images were ‘recipients’. An instruction including the appropriate noun for the larger image and a gap in the place of the smaller images was then played: Once a participant had touched the baskets, the camel, and the lock in the trial shown in Figure 5.1a, an image of three dogs appeared (see Figure 5.2a) and the instruction sentence *Now give the _____ to the dogs.* was played. Similarly, once a participant had touched the crab, the letter, and the squirrels in the trial shown in Figure 5.1b, an image of three keys appeared (see Figure 5.2b) and the instruction *Now give the _____ the keys.* was played. The gap (in place of the theme noun in this example) represents 500 ms of Brownian noise (generated with Audacity,

version 2.0.2). Brownian (or ‘brown’) noise was chosen because Shirakawa (2013) showed that it is not distracting or irritating for participants, particularly children.



Figure 5.2: State of trials during instruction sentence

Once this instruction had finished playing, the theme image(s) could be moved by dragging them on the touchscreen (see Figure 5.3). Once a theme image was moved inside the white frame of a recipient image, orange and white stars were shown over that recipient accompanied by the sound of a trumpet fanfare, chimes, or drums as a reward stimulus. This reward stimulus was shown regardless of the choice that was made, and concluded the trial.

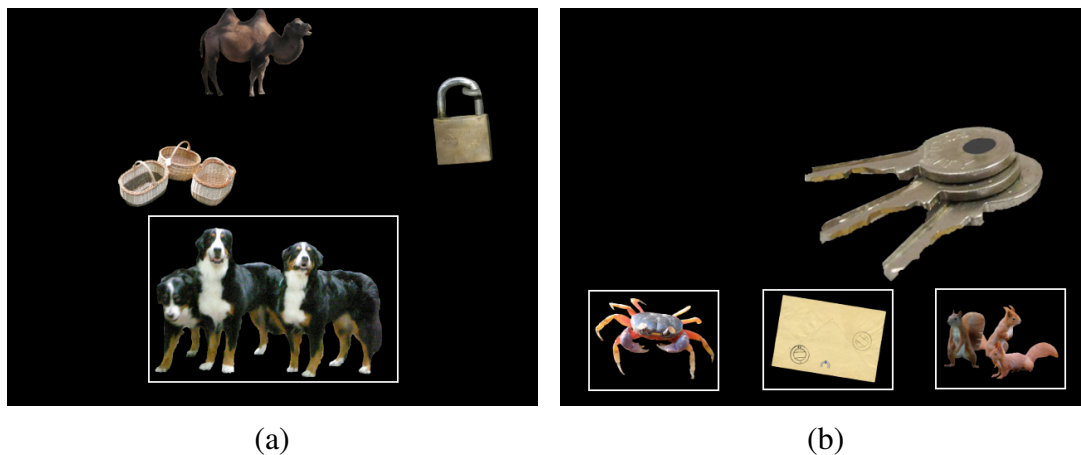


Figure 5.3: Interactive state of trials

Thus, participants made a series of choices to fill the gaps in the instructions in this experiment. Their touchscreen input (touches and dragging paths) and eye gaze during each trial were also recorded, and they were analyzed according to order of touching, the gaze percentages for each of the four images in that trial, the sequences of dragging and gaze, and the correlation between the two.

There were 64 trials, presented in four blocks of 16 (see Appendix C for the images used, and Appendix B for a list of trials). Since the four audio stimuli (three names and one instruction sentence) were always played successively and never allowed to overlap, the shortest possible duration for one trial was about 3 seconds, depending on the lengths of the particular stimuli. Participants took longer than that, of course, with the average time per trial being more than 7 seconds (see Section 5.3 for details)

All trials within one block had the same type of instruction.

Block 1: prepositional construction with gap in place of the theme (as in the *dogs* example above)

Now give the _____ to the dogs.

Block 2: double object construction with gap in place of the recipient (as in the *keys* example above)

Now give the _____ the keys.

Block 3: double object construction with gap in place of the theme

Now give the hammer the _____ .

Block 4: prepositional construction with gap in place of the recipient

Now give the monkey to the _____ .

There was a non-interactive automatic demonstration⁵¹ followed by a practice phase of four trials at the start of the first two blocks. As the illustrated examples and the list of blocks show, there were only two different types of trials, in that either the theme or the recipient was replaced by a gap and participants thus had to choose either between three themes or between three recipients. Since the first two blocks presented both of these types of trials, further training in the third and fourth block was deemed unnecessary. Participants were given the opportunity to rest between the blocks, and the eyetracker was re-calibrated after such a break if necessary.

5.1.3 Stimuli

The same order of blocks was used with all participants: the first block had prepositional dative instructions with a noise gap in place of the theme (*Now give the _____ to the dog.*), the second block had double object instructions with the recipient gapped (*Now give the _____ the bone.*),

⁵¹This demonstration presented one trial with an image of a finger apparently ‘clicking’ and ‘dragging’; all participant input was ignored during this. An image of a finger appeared over the three images sequentially, as if touching them, and the corresponding recordings of the nouns were played as this finger ‘touched’ them. Then, the fourth image appeared and the instruction was played (as described above), after which the finger moved with one ‘theme’ image to the ‘recipient’, or with the ‘theme’ to one of the ‘recipients’, as if dragging it there. On reaching the/a recipient, the reward stimulus was presented.

the third had double object instructions with the theme gapped (*Now give the dog the _____*), and the fourth had prepositional instructions with the recipient gapped (*Now give the bone to the _____*). The order of items within each block was randomized for each participant. The four images of one trial always covered the four possible combinations of animacy and number: there was always one animate singular (the camel in Figure 5.2a), one animate plural (the dogs in Figure 5.2a), one inanimate singular (the lock in Figure 5.2a), and one inanimate plural (the baskets in Figure 5.2a). Table 5.1 gives the different combinations of animacy and length used to design trials for this experiment (see Appendices B and C for full details).

explicit object	choices		
animate monosyllabic	animate bisyllabic	inanimate monosyllabic	inanimate bisyllabic
inanimate monosyllabic	inanimate bisyllabic	animate monosyllabic	animate bisyllabic
animate bisyllabic	animate monosyllabic	inanimate bisyllabic	inanimate monosyllabic
inanimate bisyllabic	inanimate monosyllabic	animate bisyllabic	animate monosyllabic

Table 5.1: Combinations of animacy and length used in trial design

Additionally, grammatical number was manipulated such that the choice that was different from the explicit object in the instruction in both animacy and length always had the same number as that object, whereas the other two choices had the other number. To rule out possible effects of different amounts being more or less prototypically plural, all images representing plurals showed a group of exactly three animals or objects. To take the first row of Table 5.1 as an example, whichever object was the one that was mentioned explicitly in the instruction sentence for that trial (the dogs in Figure 5.2a), there always was one choice that matched this explicit object in animacy (but not length or number; the camel in Figure 5.2a) as well as one length-matching (but animacy- and number-mismatching; the lock in Figure 5.2a; the length difference in this experiment is mono- versus bisyllabic) and one number-matching (but animacy- and length-mismatching; the baskets in Figure 5.2a) choice. In the pool of 16 nouns used as explicit objects (pronounced in the instruction), there were four nouns for each combination of (binary) animacy and length. Each of these 16 nouns appeared only once per block, whether it was a recipient or theme; the nouns for the smaller images that participants chose between to fill the gap in the instruction were taken from a pool of 12 different nouns, each of which was shown four times per block (see Appendix C for all of these images).

5.1.4 Limitations

Using the verb *give* is problematic because it is probably not as syntactically productive in early childhood as corpus data might suggest and because it appears much more often in a double object construction (V—NP—NP) than in the prepositional counterpart. As for the first of

these problems, while *give* utterances are attested before the second birthday for some children (Gropen et al. 1989:214), these may not constitute a truly productive use of that verb: in my study of parts of the CHILDES corpus (Bürkle 2011), 379 of the 815 *give* tokens (46.5%) were double object constructions with pronominal first person singular recipients—in other words, *give me (some item X)*. The transcripts in CHILDES (and elsewhere) are always an orthographic abstraction of the actual speech produced, and it is reasonable to assume that at least some children first use not the ditransitive verb *give*, but rather a monotransitive *gimme*.⁵² Transcription guidelines and practices can then lead to this being transcribed as *give me*, which can give rise to spurious conclusions. Similarly, the general preference for the double object construction with *give* might cause misinterpretations of variables and effects: it is well documented that *give*, on the whole, is found much more often with the double object construction than with the prepositional construction (Bürkle 2011:32 and 38–40). Therefore, a study using only *give* may attribute some double object uses to a feature that happens to be common in the particular dataset, when they merely reflect the verb’s inherent bias.

However, these two problems are not insurmountable, and using *give* also has some benefits. The false identification of *gimme* as adult-like *give* will not be a problem in the present study, as the processing part will not contain *me* as a recipient and the full audio recordings of the production part will allow identification of any verbs or predicates not found in standard adult speech. Furthermore, even if children do start with monotransitive *gimme* early on, they also start to use *give* productively and as clearly ditransitive earlier than other verbs (at least most of them do; Gropen et al. 1989:212–216). Therefore, using *give* in this study means minimizing possibilities for confusion or misunderstanding on the child participants’ part. And even if *give* shows a bias for one dative alternation construction over the other, all other verbs do as well (see Section 2.6). Keeping this particular bias in mind will allow for a clear analysis of the data. Additionally, using *give* means benefiting from its status as “a ‘canonical’ ditransitive verb” (Lambert 2010:13,⁵³ see also Gries 2003:12 and Ibbotson and Tomasello 2009:60–61 and 65), which reflects its high frequency (in English as well as in other languages, see Haspelmath 2004:33, fn. 15). As pointed out by Wasow (2002:68–72), *give* also shows virtually no meaning difference between the two dative alternation constructions. Finally, *give* is simply the most (and perhaps only) natural verb to use for this experiment: unlike *show*, for example, *give* has a recipient party in a very real sense. This means that the action of ‘giving’ one image to another is obviously only completed once it has been ‘physically’ moved there—participants might instinctively do ‘showing’ only by pointing or vague movements, making a reliable interpretation of all responses difficult. Thus, using *give* does introduce some challenges, but the benefits clearly outweigh the

⁵²Interestingly, French seems to have a parallel to this monotransitive verb in young children’s “*donnemoi X*” (Bruyn et al. 1999:361).

⁵³It should be noted here that Lambert even goes as far as defining the dative, which is the central phenomenon in her crosslinguistic study, as the marking on the recipient object of a language’s *give*-equivalent (Lambert 2010:12).

drawbacks.

The verb *give* also imposes certain restrictions on the animacy of its two objects, or at least there is a prototypical or preferred pattern: typically, we give inanimate themes to animate (or even human) recipients. Other patterns, particularly the ‘reversed’ pattern of giving an animate theme to an inanimate recipient, are intuitively less acceptable and less frequent. This may affect participants’ choices in experiment 2: in the trials with three theme options and the noise-gap in place of the theme noun in the instruction sentence (such as Fig. 5.1a), inanimate options may be chosen more frequently not for reasons of ordering preferences, but to fulfil the preferred animacy pattern. The same is true of animate choices in trials with three recipient options and recipient-gap instructions. However, because children as young as 5 years have given animate themes to inanimate recipients in act-out tasks elsewhere (see Cook 1976), this possible preference was expected to be at most a tendency, and statistical analysis by animacy was expected to reveal it if it does arise.

While there is some evidence that the prepositional construction is easier to process, at least for children, no interaction between this processing advantage and the prototypical animacy pattern is expected. This is despite Cook (1976)’s finding of just such an interaction. With a simple act-out task, Cook (1976:437) showed that “the child of five is as ready to give a man to a car as a car to a man” when instructed to do so with a prepositional construction. This suggests that even reversing the prototypical animacy of the two objects of *give* does not affect comprehension. When the instruction sentence used the double object construction, however, participants produced the intended act-out response more often in trials with the prototypical animacy pattern (inanimate theme and animate recipient) than in trials with the reverse pattern (animate theme and inanimate recipient). These reverse-pattern trials in turn elicited the intended response more often than trials with two inanimate objects (both theme and recipient) did. These results suggest that children are sensitive to the prototypical animacy pattern, but this prototypical account does not explain the difference between the reverse and two-inanimate conditions. Moreover, there is some doubt regarding the statistical validity of these results: Cook (1976) reports *p*-values of the rank tests for these comparisons, but does not mention any adjustment for the multiple comparisons problem: if every result that is less than 5% likely under the null hypothesis is judged to be evidence against the null hypothesis, some cases where the null hypothesis is true anyway will be wrongly analyzed as cases where the null hypothesis is false. This is normally acceptable, and perfect certainty is impossible. When one performs many such hypothesis tests, such false positives become more and more likely (just as the chance to win the lottery increases with every lottery ticket one buys). It is possible to ameliorate this problem by performing certain adjustments to the *p*-values (see Section 5.2.6), but Cook does not report having performed any. In other words, some of the supposedly significant differences in Cook

(1976) may be much less significant under standard assumptions than what they are presented as. Also, the items used to represent the animate objects for the act-out task were dolls. As discussed in Chapter 2, participants may understand dolls to be less animate than intended, particularly if the somewhat strict instructions given by Cook (1976:436) put some participants into what Gelman et al. (1983:108) calls “non-play mode” (where dolls and the like are inanimate objects). For these two reasons, the animacy-related results of Cook (1976) cannot be assumed to be strongly supported (though children’s general capability of giving inanimates to animates is of course proven, and the advantage of the prepositional construction appears to be statistically strong).

In experiment 1 (Chapter 4), participants had to access their explicit knowledge about which of the entities shown in the pictures are capable of motion; and in experiment 2, participants were asked to move these same images. Using motion as a criterion for animacy here certainly brought the idea of motion to participants’ attention, and therefore it is possible that their behavior in experiment 2 was not guided by the linguistic features under investigation, but by the idea of motion—for example, after identifying a dog as being able to move and a lemon as being immobile, participants could conceivably be more inclined to move a (picture of a) dog rather than a lemon. This, however, may just as well be turned on its head, with participants being more inclined to move a lemon precisely because it cannot move by itself and thus needs their ‘help’ to do so. The fact that this potential problem is consistent with two opposite outcomes suggests that this is unlikely to be a systematic problem.

Motion had to be used because motion is considered to be central to children’s (and probably adults’) concept of animacy, as discussed in Section 2.3 above. Directly asking participants “Is this one an animal?” would not provide an insight into their concept of animacy either: the finding that “infants [as young as 12 to 14 months] commence the process of acquisition equipped with an initially general expectation linking words to object categories” (Waxman 1999:253) does not mean that all children will link all words to categories in an exact or adult-like way. Children may understand and categorize for a feature without using the word for that category correctly or at all, and conversely their adult-like use and understanding of the word for a category may be accidental and not a result of correct categorization (Taylor 1995, Waxman 1999, Sandhofer 2001). In a study of animacy, merely asking child participants whether one stimulus object is an animal will therefore not get at whatever concept of animacy or ‘animal’ the children actually have at the time. Using just one feature of animacy (like motion) in experiment 1 would not get at the concept of animacy either, but only at the concept of that particular feature. To see which underlying features of animacy are at work for participants, it is necessary to test for features instead of testing for animacy itself, and to test for more than one feature. Gelman et al. (1983) found that animals’ capability for simple as well as reciprocal actions work well as a test of

young children's concept of animacy, and so (goal-directed, following Opfer and Gelman 2011) motion and play were used here to cover both of these categories and thus get a clearer view of participants' concept of animacy.

Furthermore, it should be noted that experiment 2 asks participants to give, rather than move, one image to another, which further reduces the possibility of an unwanted effect here. Also, the images are all static, so the fact that the animals depicted can move and the inanimate objects depicted cannot is not brought to the conscious attention of the participants who are asked to make them move. The entire experiment rather depends on the fact that "young children readily distinguish between what an object looks like and what it truly is" (Booth and Waxman 2002:B20), which could be seen to negate the above point: if children are aware of the difference between picture and depicted, the fact that all pictures used are static cannot affect children's awareness of the depicted animals' ability of motion. However, following Opfer and Gelman (2011), children's concept of animals' motion is best expressed as the capability for **autonomous** goal-directed motion. Experiment 1 accessed that. Moving images on a touchscreen is a different kind of motion, and so should not be confused with autonomous motion: when dragging an image of a dog on a touchscreen, it looks like the dog is moving, but it is not truly moving (as in walking). Moreover, the pilot studies tested all the combinations of images for any strong effects of factors (like the knowledge of animals' ability to move), and found only a general preference for animates to go with animates and inanimates to go with inanimates. Not even the broad type of locomotion that an animal uses had an effect: pilot trials with dolphins showed animacy-matching just like trials with dogs, cows, and other four-legged walking animals.

While the experimental session as a whole was structured in a way that maintained attention and minimized confusion, the random order of trials within each block of experiment 2 might be considered a bad choice in that it could have unwanted and hard-to-spot confounding effects on the results. For example, one participant may randomly see several successive trials that all contain the 'lemon' or 'lemons' images as a choice (as one of the three smaller images). This participant may then be influenced by this reoccurrence to choose the lemon(s) more, or conversely to avoid this choice. Similarly, an imaginative participant could potentially have started out making the choices that the animacy, number, and length of the choice images suggested for the particular instruction, but then built a 'story' from these and subsequently chosen images based on what fitted best into that story. In the obvious alternative to randomization, the trials could have been pre-randomized into one order and presented in this same order to all participants. However, this set order would then have needed to be tested first, to see whether there are any effects of one trial on another. This would have been prohibitively time-consuming and expensive even for just one such order, and presumably several orders would have been evaluated against each other. Per-participant randomization of trials within blocks coupled with data analysis **per group**, on

the other hand, can be expected to cancel out any such effects, should they arise. Therefore, the random order of trials within each block for each participant was the most economical way of ordering trials as well as the one most likely to produce valid, usable data.

Order of blocks has been reported to have a ‘carry-over’ effect on young children’s behavior in experiments (Gelman and Koenig 2001:693–695 and Sloutsky and Fisher 2011), but the order of blocks in experiment 2 varied in two relevant factors (construction and gapped object) and was therefore not expected to cause a similar strong effect on participant behavior. As discussed in Section 2.6, recent constructions are likely to reoccur (in the dative alternation and in other phenomena). Experiment 2 does not suffer from this problem, as production was not its aim.

The above description of this experiment may create the impression that this is too complex a task for young children, particularly since it involves a touchscreen as the only computer input device—although many children may have experience with touchscreens, not all of them can reliably be expected to be ‘proficient’ touchscreen users. Touchscreens have not been widely used in research involving young children, but based on the studies that did use them, it was reasonable to assume that children can generally use touchscreens competently (Sutton 2006). The only common problem seems to be that the touchscreens failed to register children’s clear and intentional touch input (see for example Luciana and Nelson 2002:610). This problem is eliminated by simply using touchscreens based on newer technology (capacitive or optical, for example) rather than the resistive touchscreen Luciana and Nelson (seem to have) used. The problem of initial unfamiliarity still has to be considered, of course, but it is easily dealt with by giving participants “a period of time” to learn how to use the touchscreen (Romeo et al. 2003:334)—the practice phases described above, in other words. Thus, the mere use of touchscreen technology could not be reasonably expected to present a major problem to the research, and indeed it did not (see Section 5.3.3 for details). Participants only interacted with the touchscreen by pointing and dragging, and the pointing gesture is used frequently and with different intentions by children as young as 1 year (see for example Liszkowski et al. 2006). This form of touchscreen use thus builds on a very early gesture and was therefore not expected to be difficult or confusing to participants (see also Colombo et al. 2003). No such problems occurred during the experiment.

Although an eyetracker is most likely unfamiliar to children (and parents or caregivers), I did not expect this fact to present a problem in and of itself. The Tobii X120 eyetracking hardware is a head-free table-mounted device. As used in the present experiment, it appears as a (wired) box sitting on a stand on the desk, behind the touchscreen computer, and it was introduced and explained to participants (and parents or caregivers) in that way. Boot et al. (2012), using a screen with built-in eyetracking hardware, report no problems due to unfamiliarity or other

negative reactions to the technology itself. Therefore, no such effects were expected here, and indeed no participant showed any.

The rationale behind the eyetracking methodology is that people focus their gaze on the current focus of their attention. This requires some control over eye gaze. Children aged 10 to 12 are capable of fixations that cover as little as 0.65 degrees of visual angle (Eden et al. (1994:1350), and typical adults are capable of fixations with around 0.6 degrees of variance (Russo et al. 2003:1840). Oculomotor control and executive control over gaze direction continue to develop well into the second decade of life (Paus et al. 1990, Fischer et al. 1997, Munoz et al. 1998, Fukushima et al. 2000, Bucci and Kapoula 2006). The children in the two younger age groups in the present study are younger than that, however, and so may not be able to control their gaze as well as the adult participants. However, children's gaze is not entirely random: while it may be less exact than in adults, children (as young as three years!) have reasonably good control over the direction of their eye gaze (Scerif et al. 2005). Eyetracking studies with children as participants have to be interpreted with this development in mind, but the underlying rationale that gaze location indicates attention or processing with high temporal resolution is sound.

Finally, if the differences in animacy, length, and number do indeed affect the choices that children make in filling gaps, these differences can be expected to be quite subtle. Any confounding factors must therefore be eliminated where possible, and controlled for elsewhere. The largest concern here is whether any of the combinations of four pictures to be shown together contain two pictures that are strongly associated semantically or through other non-linguistic knowledge—for example, with an instruction of *give the _____ to the dog*, a choice of *bone* would be so strong as to overpower and overshadow any linguistic ordering preferences. This obvious example is easily avoided, of course, but less obvious associations could still cause bias regardless of instructions. However, these associations would then also be evident in a similar task with no instructions at all. Therefore, a pilot study with just such a task was carried out.

5.1.4.1 Pilot study

There were two age groups of participants in this pilot: 19 children (age range 5;11 to 11;0, mean age 8;6) who were recruited through a contributing primary school in the Canterbury region of New Zealand and given their choice of small item from a 'box of treasures' as a reward, and 20 adults (age range 18 to 39 years, median age 20) who were recruited with no incentive on campus at the University of Canterbury.

This pilot was programmed and run in PsychoPy, version 1.76.00, on the same touchscreen laptop as experiments 2 and 3. Participants were presented with a succession of screens consisting of one larger 'recipient' image, centered near the bottom of the screen, and three smaller and

drag-able ‘theme’ images next to each other near the top of the screen. A pre-programmed demonstration showed them that dragging one of the smaller images to the large one caused stars to appear and a fanfare to play (the ‘reward’ stimuli), and a short practice period showed that any of the three theme images could be dragged and would lead to the same ‘reward’. These practice trials went from ones with a very strong association between the recipient and one theme to being more like the trials to be tested later, as shown in Table 5.2. This was intended to encourage participants to look for semantic associations like this in order to increase the likelihood of any such association being revealed in the results.

recipient	associated theme	other themes	
key	lock	camel	dog
pencil	ruler	crab	snake
fox	duck	cherry	cup
pig	apple	camel	lock

Table 5.2: Training trials in the pilot study

This practice phase was followed by 64 individual trials (four for each of the 16 recipient images which were to be used in the later experiment). In other words, the task in this pilot study was very similar to the trials with noise replacing the theme in experiment 2, but without an auditory stimulus sentence.⁵⁴ This pilot study was approved by the Human Ethics Committee of the University of Canterbury (reference numbers HEC 2012/172 and HEC 2013/31/LR-PS).

Two of the combinations of four images were ruled out after this pilot because participants appeared to have strong preferences for one particular ‘theme’ image, presumably based on functional connections between it and the ‘recipient’: with a *key* as the recipient, participants strongly preferred the *basket* choice over the *dolphin* and *fox* options; for a *pencil* as the recipient, participants preferred the *lock* over the *duck* and the *squirrel*. I make no claims regarding the specific functional connections in these combinations. The other combinations of nouns did not show similarly strong preferences and were therefore assumed to be free of any strong individual connections between the nouns. 16 of them were chosen for use in the main experiment.

More interestingly, the results of this pilot study showed a strong preference for animacy-matching: both children and adults were much more likely than chance to pick the one smaller image that expressed the same value for animacy as the recipient image did—in other words, participants preferred animate choices for animate recipients and inanimate choices for inanimate recipients. Some adult participants specifically mentioned using this strategy after the experiment. The pilot

⁵⁴It is true that the omission of pre-recorded speech from the pilot made it different from the later experiment, but the recordings were in that experiment only to elicit ordering-based preferences and to control for length. Neither of these considerations is relevant to the aims of the pilot. Furthermore, it could be argued that the absence of speech stimuli made the pilot less processing-intensive and therefore easier.

study had no per-trial instruction sentences and indeed no instruction beside the pre-programmed demonstration. Animacy-matching was taken to be an explicit strategy adopted in the absence of specific instructions, and therefore less likely in the full experiment, where each trial had an instruction sentence.

5.2 Analysis

Getting from the raw data collected in this experiment to meaningful results is far from straightforward. This section describes the issues in more detail, starting with an overview of different touchscreen technologies and the specific setup of the touchscreen and eyetracker used in experiment 2. Statistical methods for eye gaze data are then introduced and evaluated. The chapter also provides an introduction regression modelling, containing an overview of the modelling strategy adopted for this study, and a brief discussion of the multiple comparisons problem and the steps taken to address it in this thesis.

5.2.1 Touchscreens

Resistive touchscreens have two panes of electrically conductive material behind the screen, and the user's touch input makes a connection between the two. Capacitive touchscreens set up an electrical field covering the screen and measure the deviations caused by a finger entering that field. Finally, there are technologies that emit waves of light or sound (outside of the spectra that are visible or audible to humans, of course) across the screen surface and measure the turbulence that a finger (or any other solid object) touching the screen causes in these waves.

The touchscreen in the HP EliteBook 2740p used in this study is a capacitive model. This strikes a good balance between useability and screen response time: Luciana and Nelson (2002) found that children aged four had trouble using a resistive touchscreen device,⁵⁵ meaning capacitive or wave-based touchscreens are more suited for studies with young children as participants. However, wave-based technologies tend to be slower to record touch input (Holzinger 2003:392), making them less intuitive to use as well as reducing the rate at which touchscreen data can be collected. Therefore, a capacitive screen was judged ideal for the present study.⁵⁶ The screen was calibrated before the first experiment session.

⁵⁵Some of this trouble can be attributed to young children's developing motor skills, as Romeo et al. (2003:333) note. Nevertheless, capacitive touchscreens eliminate at least the problem of having to press the screen quite hard, and are thus relatively better suited for child research than resistive touchscreens.

⁵⁶Holzinger (2003:391) notes that non-conductive objects like gloved hands do not cause deviations in electrical fields, and can therefore not be used as input devices on capacitive touchscreens. As the experimental sessions in the present study were conducted in an enclosed room, participants did not wear gloves, so this was not a problem.

Eyetrackers like the one used here (see Section 5.2.2 for details) work out the direction of a participant's gaze from the position and shape of the pupils. To know what this gaze was directed at, it is necessary to either record the scene in front of the participant with a separate camera, or (as here) to use a computer screen. If the eyetracker is successfully tracking a participant's eyes, it obviously knows the position of the participant's head and the direction of their gaze. To compute where on the screen this gaze was directed, the position and orientation of the screen relative to the eyetracker also needs to be known. I constructed a simple wooden stand for the screen and eyetracker and configured the eyetracker with the distances and angles that this stand enforced. The stand (including touchscreen and eyetracker) could be tilted up and down by adding or removing blocks underneath the back. This allowed the whole setup to be adjusted for participants of different heights and for mitigating the "gorilla arm" problem (see for example Way and Paradiso 2014:138): holding out one arm to interact with a vertical screen for a longer time period is fatiguing and painful. A tilted and lowered screen is more comfortable to use. As the eyetracker and screen were fixed on the stand, the relative distances and angles between them were not affected by these adjustments.

5.2.2 Eyetracker setups

The eyetracker used in this study is a Tobii X120, a head-free model that relies on near-infrared light. Head-free eyetrackers are the obvious choice for studies with young children as participants because do not constrain the head or place some unfamiliar apparatus on the head or face, which would be distracting at best and scary at worst for children. Near-infrared eyetracking makes use of the fact that the inside of the human eye reflects near-infrared light very well: the pupil, being the opening that lets light into and back out of the inside of the eye, shows up very well in near-infrared light (Morimoto et al. 2000).⁵⁷ The direction of the eye gaze can then be extrapolated from the pupil's relative position within the eye, and the point where the gaze 'line' intersects with the display area is the point where the participant's gaze is currently centered. This requires only that the location of the display area is known, which is trivial when, as in this study, the display area is a computer screen in a fixed position and angle relative to the eyetracker.

It is normal procedure with the Tobii X120 and similar eyetrackers to position the eyetracker below the display area (see Morgante et al. 2012). The eyetracker is angled upwards towards the participant's eyes and the cameras inside the eyetracker are also angled upwards relative to the case itself. With an interactive display like the task in experiment 2, this position is not practical:

⁵⁷The user manual of the Tobii X120 certifies the infrared light to be safe for exposure to the eyes within industry standards (Tobii Technology AB 2008:2). The dangers of near-infrared diodes in immediate skin contact (Bozkurt and Onaral 2004) are obviously of no concern for an eyetracker.

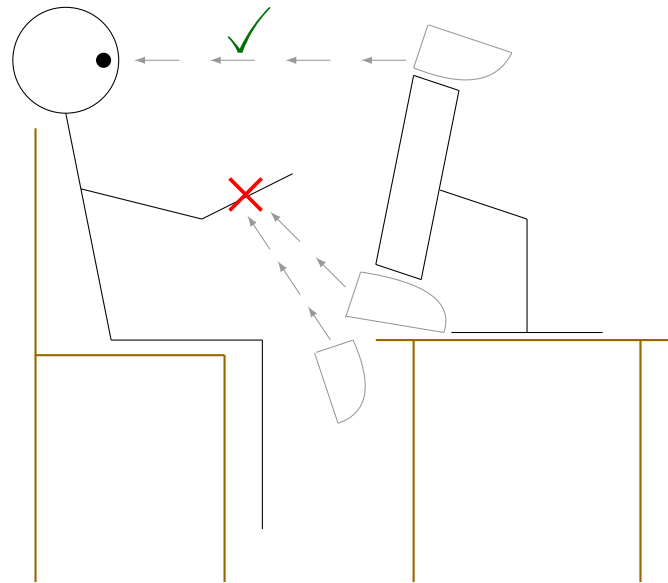


Figure 5.4: Schematic view of different positions of a head-free eyetracker (grey boxes) in a touchscreen experiment

as Fig. 5.4 shows, the participant’s hand and arm would be between the eyetracker and the eyes during any interaction, which would make it impossible for the eyetracker to collect gaze data. Therefore, the manufacturer’s recommendation as well as apparent general practice in this kind of study (see for example Biedert et al. 2012) is to place the eyetracker above the display area. With the Tobii X120 in particular, it is also necessary to turn the eyetracker upside-down: as mentioned above, the cameras inside the eyetracker are angled upwards relative to the eyetracker casing, which is helpful in the normal below-the-screen setup. When the eyetracker is above the screen, however, this would mean that the cameras view an area high above the participant’s head (see Fig. 5.5a). Inverting the eyetracker solves this problem, as Fig. 5.5b shows.

It is true that “the incoming gaze data needs then to be post-processed to match the flipped order of axes” (Biedert et al. 2012:385), but this is in fact trivial: defining the center of the display area as the coordinates $(0, 0)$, while making initial stimulus design somewhat less intuitive,⁵⁸ means that all the ‘post-processing’ that is necessary is multiplying both coordinates of all gaze points by -1 (to change their signs). This changes the incorrect coordinates in Fig. 5.6b to the correct ones as in Fig. 5.6a.

The major drawback of head-free eyetrackers (particularly in the inverted above-the-screen setup) is that it does not yield as much data as other kinds of eyetrackers. Firstly, since a

⁵⁸Choosing a corner of the screen to be $(0, 0)$ is arguably more intuitive because it means a coordinate pair is can be very easily understood: with $(0, 0)$ as the top left corner (the standard in E-Prime, for example; see Schneider et al. 2002:20 and Hudson 2011:17), a coordinate pair (x, y) is simply “ x across, y down”. Also, $(0, 0)$ as the top left (or really any) corner means that there can only be positive screen coordinates, while $(0, 0)$ as the center makes the left half of the screen have negative x coordinates and the bottom half have negative y .

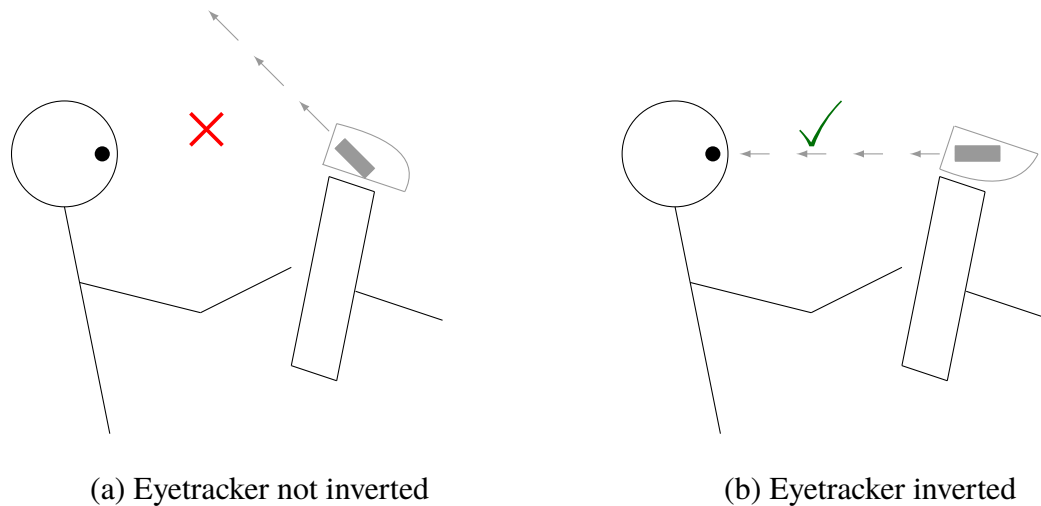


Figure 5.5: Schematic view of the orientation of the Tobii X120 eyetracker camera (solid grey rectangle) within the eyetracker casing (larger grey box)

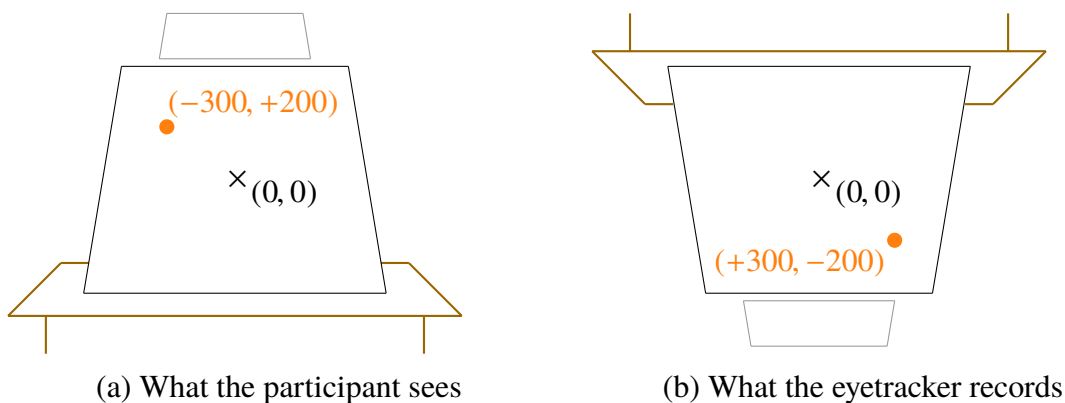


Figure 5.6: Schematic views of a possible time point, with hypothetical gaze (round orange dot). Numbers in parentheses are (x, y) coordinates.

head-free eyetracker allows participants to move their heads freely, many of them do (particularly young children and particularly in a longer experimental session). Although the eyetrackers are programmed to continually search for eye-like near-infrared reflections in a ‘tracking box’ in the space before them ($36 \times 22 \times 30$ cm at 70 cm distance from the eyetracker and a data rate of 60 Hz in the case of the Tobii X120, see Tobii Technology AB 2008:21), it is easy for the participant’s head to leave this relatively small area, especially by slow drift over the course of a longer session. Stopping the experiment every time this happens would mean losing the advantage of unobtrusive eyetracking; letting the experiment continue means losing some data. Secondly, the near-infrared light can be partially obstructed by eyelashes (Tobii Technology AB 2011:5) or glasses (Morimoto et al. 2000:334), and the eye/pupil detection algorithms cannot always recover from this. Thirdly, near-infrared eyetracking seems to just not work with some people’s eyes, for idiosyncratic reasons. On this point, I can only agree with Morgante et al.

(2012:29): “We have no concrete explanation for these effects—in our experience, we simply cannot get much data from some participants”. These three problems together make for a ‘lossy’ data collection method, with data for only part of the session for almost all participants. Mulak et al. (2013:2072) achieved an average yield of 72.5% with 15- and 19-month-old participants. Even with cooperative and motivated adults in studies as short as two minutes, Morgante et al. (2012) got data for more than 80% of those two minutes from only about half of their participants. In the present study, there was no (0 data points in the experimental period) or effectively no (less than 60 data points, which at 60 Hz would be 1 second) gaze data from eight of the 62 participants (four four-year-olds, three eight-year-olds, and one adult).

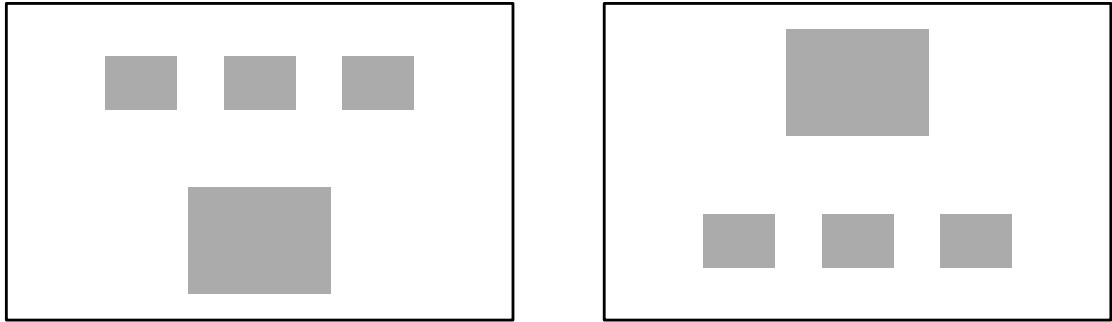
5.2.3 Correcting systematic errors in eye gaze data

While setting up the eyetracker above the screen and inverted is recommended practice for studies using touchscreen devices with the Tobii X120, it means that some participants’ eyes will not be recognized at all, and that eye detection will be very spotty for some others (Tobii Technology AB 2011:5). This can be reduced somewhat by calibrating the eyetracker carefully at the start of an experimental session and asking the participant to try and maintain the calibrated position, as well as having the option to re-calibrate at several points during the session. Even so, the raw eyetracking data gathered in this study often appeared to have a consistent and obvious per-participant error that was easily corrected. This section explains why these errors were obvious and how they were corrected.⁵⁹

Although their gaze was not constrained, it is reasonable to expect that participants would look at pictures more than at the black background behind them; and although the pictures could (and indeed had to) be moved, it is reasonable to expect that the time before any pictures were moved was a significant portion of the total time taken in each trial. It follows from these two assumptions that a large portion of gazes in a particular trial will fall into the regions where the images first appeared in that trial. Since there were only two distinct layouts of pictures in the whole experiment (three smaller pictures near the top and one larger picture near the bottom, or the reverse; see Fig. 5.2), this means that a large portion of gaze data collected throughout the experiment should fall into these clearly-defined regions. The grey rectangles in Fig. 5.7 schematically show these regions of the screen.

Some participants’ data bears this out fully—Fig. 5.8, for example, is a heatmap of the raw gaze

⁵⁹I do not know what caused these errors, and I will not speculate here. Curiously, Morgante et al. (2012:28, emphasis added) found that “most deviations were **downward** from the calibration points” in their evaluation of a similar Tobii eyetracker, while most of the errors corrected for here were effectively **upward** from the expected pattern (while this section discusses adjusting eye gaze data from apparent downward errors, remember that this was applied after all eye gaze data was ‘flipped’ to correct for the inverted eyetracker setup described in Section 5.2.2 above).

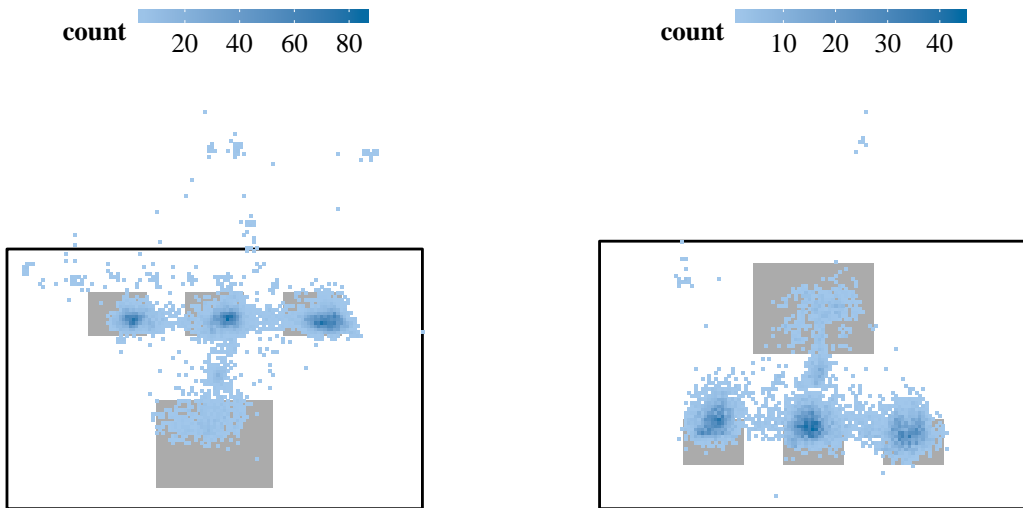


(a) Theme-gap blocks

(b) Goal-gap blocks

Figure 5.7: Locations of pictures at beginning of trial in experiment 2

data from one participant's session (areas shown in darker shades had more gaze data) overlaid onto Fig. 5.7. The pattern of the picture starting positions, as based on the above assumptions, is reflected strikingly well in this participant's eye gaze data.



(a) Theme-gap blocks

(b) Goal-gap blocks

Figure 5.8: Density of gaze data from one participant, by block type

Other participants' data had the same very apparent pattern, but offset by a constant amount: most of the data points in Fig. 5.9a and Fig. 5.9b fall into four rectangles with the same relative positions to each other as those in Fig. 5.7, but in a different absolute position (most of them

even off the screen in Fig. 5.9b). It is apparent that a simple correction of subtracting a certain value from the x-coordinate and adding another value to the y-coordinate of each gaze data point would bring that pattern in line with the expected pattern. Furthermore, it is reasonable to apply this correction: since the same constant value is added to **all** data points from this participant, there is no overzealous and possibly dishonest maximizing of “useful” data points. The correction values (−887.833 pixels in the x-coordinate and +202.1 pixels in the y-coordinate in this example) were determined algorithmically by finding the peaks in the density plot for the two coordinates separately (three peaks for x, corresponding to the left, middle, and right pictures’ starting position; two peaks for y, corresponding to the upper and lower rows/positions) and checking their distances. The mean value of the three or two peaks was then applied as the correction value, and the output was checked manually.⁶⁰ Note that noise in the data as well as apparent rotation, expansion or contraction (relative to the expected pattern) is retained, as the corrected heat map in Fig. 5.9d shows—the only thing that is removed is the fairly obvious bias.

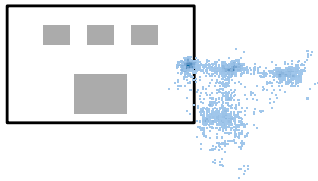
This correction was not necessary for all participants (see Fig. 5.8 above), and was not effective for some others: in cases where the expected pattern was not obvious, such as Fig. 5.10, no correction value was found or applied.

In summary, the simple per-participant correction vector was applied only to the data of participants where the expected pattern was obvious, but offset. Furthermore, the same correction vector was applied to all of one participant’s data, which still left many data points outside of the expected pattern and even outside the screen area. For these two reasons, this method of correcting for errors did not zealously over-correct data into a preconceived pattern, but only removed some obvious errors. Correction as described in this section was applied to eye gaze data from 38 of the 54 participants who provided more than one second of eye gaze data each (12 four-year-olds, 12 eight-year-olds, and 14 adults).

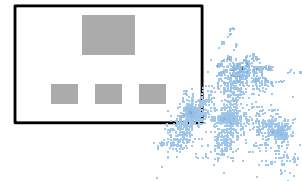
5.2.4 Analyzing eye gaze data

Raw eye gaze data is little more than (x, y) coordinate pairs with time codes: the coordinate pair defines the participant’s point of gaze at the time defined by the time code. Studies of reading see this data as a path of fixations and forward or backward rapid directed eye movements (saccades) over time. Visual world studies like the present one usually define areas of interest on screen (typically the locations of text, images, or video) and check for each data point whether it fell into one of those areas (and into which, if there is more than one). Instead of analysing

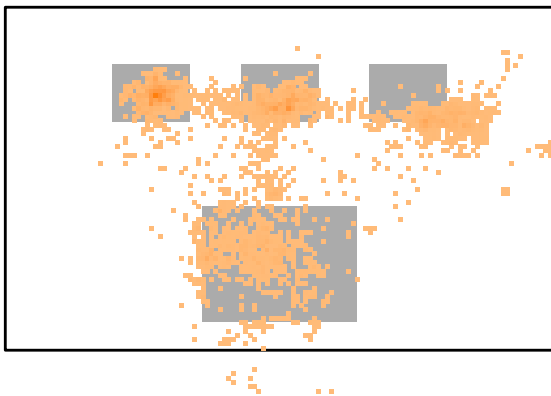
⁶⁰Manually validating the output of this automatic process was necessary because the algorithm was quite simple and used any set of three or two peaks it could find—even moving the data points **away** from the expected patterns in some cases with very scattered data.



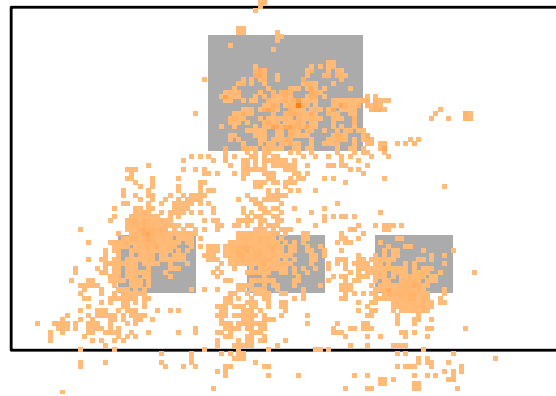
(a) Raw data, theme-gap blocks



(b) Raw data, goal-gap blocks



(c) Corrected data, theme-gap blocks



(d) Corrected data, goal-gap blocks

Figure 5.9: Density of gaze data from another participant, by block type, before and after correcting

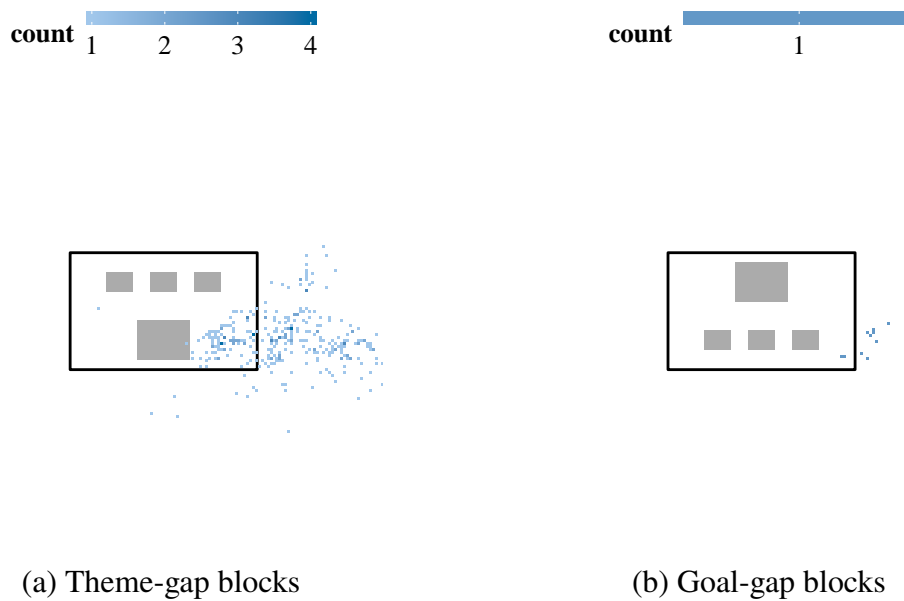


Figure 5.10: Density of gaze data from yet another participant, by block type, deemed uncorrectable

it millisecond-by-millisecond, the data is often binned by time: for example, 50 ms-long bins mean that all data points with time codes between 0 and 50 ms (from the start of the trial or any other fixed time point of interest) are assigned the binned time code 1, all data points between 50 and 100 ms get the binned time code 2, and so forth. Often, all the data from all participants in all trials of an experimental condition is combined. This is particularly beneficial if, as in the present study, the eyetracking method used is not guaranteed to provide a continuous stream of data for each participant: if, for example, there are 20 participants, but only 19 data points at a given time points (due to missing data), the average of those 19 points will not be greatly different from the average of the 20 points at the previous and following time points, meaning the overall smooth pattern of the data is preserved.⁶¹

These steps allow for a more stable and interpretable metric, namely the percentage of gaze points or fixations on each area of interest for a time bin. If this percentage for one area of interest rises during a certain period of time relative to the percentages for the other areas, this suggests that participants looked at that area more during that time, meaning that it drew more attention and interest. Averaging works particularly well for this type of data, since the percentage for a given area is unlikely to change immensely from one time point to the next. Other aspects of eyetracking data are not suited for averaging: the average gaze point at a given time point is not an interesting measure in visual world studies, where the areas of interest are often not individual

⁶¹Averaging over data from different participants or trials to mitigate the problem of missing data assumes the time points where data is missing are random, meaning there is no systematic problem such as the eyetracker not recording between 500 and 700 ms into a trial, for example.

images, but rather types of images as defined by the experimental design (for example, the ‘target’ image that is named in an instruction sentence played over headphones versus a ‘distractor’ image of a different object with the same color as the ‘target’). Percentages or proportions of gazes on an area, distances between gaze and type of area, and other data defined with reference to the experimental design can reveal interesting patterns of gaze behavior when averaged.

As mentioned above, it is very common in these studies to find that the percentages change only by a small amount from one bin to the immediately following bin, particularly if the bins are small. This is not surprising: a sharp increase or drop in one area’s percentage would mean that all participants in all trials suddenly changed their gaze behavior at the exact same time with sub-second precision, which is highly unlikely. However, this smoothness of the percentages over time gives rise to a major problem for statistical analysis: the percentages for one time bin depend to a large extent on the percentages in the preceding bin(s). Therefore, it is statistically unsound to compare the percentages within each bin in isolation. This section presents three ways of dealing with this problem (averaging across larger time windows, growth curve analysis, and smoothing spline analysis of variance) and compares their performance on simulated datasets. This discussion will show that smoothing splines are both well-suited for the analysis of eye gaze data in this study and conservative with regards to rejecting the null hypothesis.

5.2.4.1 Time windows

The authors of many eyetracking studies choose to pool all the gaze data for a certain time window (hundreds of milliseconds to several seconds, generally) and then perform common statistical tests (analysis of variance, *t*-tests) on that averaged data. For example, Huettig and McQueen (2007) perform *t*-tests for each of eight successive time windows of 100 ms each to find significant differences between the gaze percentages for four different areas of interest, Ito and Speer (2008) use the *F*-statistic (and *p*-values derived from it) of data in 300 ms windows, and Mulak et al. (2013) investigate the mean fixation time averaged across a 4-second window. At first glance, this seems like little more than binning with very long bins, with the added benefit of using well-known and easily interpreted statistics. However, as the three examples just given show, there is no standard value for time windows, which is troubling in light of the fact that “different size windows can produce very different results” (Mirman et al. 2008:477)—spurious differences can emerge as significant, or true effects be lost in the noise, simply due to a bad arbitrary choice of window length (be that choice accidental or deliberate). The examples of Barr (2008:458–460; see there for details and graphs) make these possibilities more clear: if, in a study with just two areas of interest, participants looked at one of them more than the other (for unrelated reasons) **before** the beginning of the time window under analysis, this preference

would ‘spill over’ into the time window because the gaze percentages are ‘smooth’ and thus highly unlikely to drop to the level of the other area immediately. When pooling all the gaze data from the time window, it would appear that participants were more likely to look at that first area, which would lead researchers to assume that this area (or its content) drew attention in the context under investigation—when, in fact, the difference is due entirely to an unrelated preference before the time window and the nature of eye gaze percentages. Similarly, if participants started out preferring one area over the other but switched this preference to the other area during the time window, the average percentages would end up at the same level, leading researchers to conclude there is no interesting difference here when, in fact, there is.

Of course, if there are well-founded a priori reasons for a particular study to investigate only a particular time window, it can hardly be called arbitrary. This is the case in Kurumada et al. (2014), for example: they investigate whether the prosody of *looks* in sentences like *It looks like a monkey* had an effect on the interpretation of those sentences (‘It merely appears to be a monkey.’ versus ‘All signs point to it being a monkey.’). Since they manipulated the prosody of the verb, it is reasonable to assume (as they do) that differences in gaze percentages will manifest once the verb has been processed. Under the standard assumption that it takes about 200 ms between stimulus and eye movement response, their choice of a time window starting 200 ms after the onset of the verb in the stimulus and ending 200 ms after the onset of the target noun is reasonable. However, this particular window is reasonable only for this particular study—different designs mean different critical time windows. Moreover, some designs (like this study) do not have just one critical time window of interest where a significant difference in average gaze percentages would be interesting, but rather are interested in several changes in gaze percentages over time. Averaging within time windows cannot possibly handle time with enough detail for such studies, and it does not account for the continuous nature of time when comparing several successive time windows (Barr 2008:460).

5.2.4.2 Growth curve analysis (GCA)

Mirman et al. (2008)’s growth curve analysis (GCA)⁶² does manage time with a high level of detail and with controls for its continuous correlated nature. The method is based on mixed-effects regression models (see for example Barr et al. 2013), which, due to the popularity of mixed-effect models in linguistics, makes it relatively easy for an expert linguistic audience to understand the results of GCA. GCA models the percentage of gaze using time and the area of interest as fixed effects (predictors) and participants as random effects (intercepts and slopes). This allows for fine detail with regard to time and does not limit the analysis to averages in arbitrary

⁶²This section only gives a brief overview of GCA; for a more detailed introduction as well as worked examples in R, see Mirman et al. (2008) and Mirman (2014).

time windows. (It would also be trivial to include other predictors—different conditions of the same experiment, for example.) Regression modelling finds the best estimate for a predictor’s parameter (β_n), which is the amount of change in the response variable for every unit change in that predictor. This obviously allows only for linear changes that are constant over the whole range of predictor variables. Eye gaze percentages are often more complicated, with curved patterns of increases followed by decreases. These patterns can be approximated by using not bare time codes, but polynomials of the time codes (up to a given order). For example, the first-, second-, and third-order polynomials in (5.1) by themselves describe the lines and curves in Fig. 5.11.

- (5.1) a. $y_1 = 20 \times x$
 b. $y_2 = -200 + x + (10 \times x^2)$
 c. $y_3 = (-50 \times x) + x^2 + x^3$

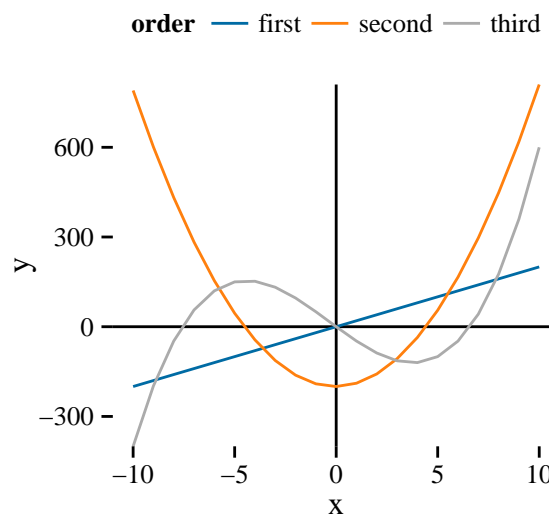


Figure 5.11: First-, second-, and third-order polynomials

The y -values of these polynomials can be multiplied by constants and added together. This operation is ‘blind’, so to speak, to how the y -values were calculated. If there are no exponents in this operation (no y_1^3 , for example), the result is a linear combination of these y -values, even though the y -values themselves were calculated with exponentiation. A linear combination like this can approximate a curve with at most one peak and one curve: the dashed curve in Fig. 5.12 is described by the formula (5.2) in terms of the y -values from (5.1).

(5.2) $y_{curve} = (-0.92 \times y_1) + (0.47 \times y_2) + (0.71 \times y_3)$

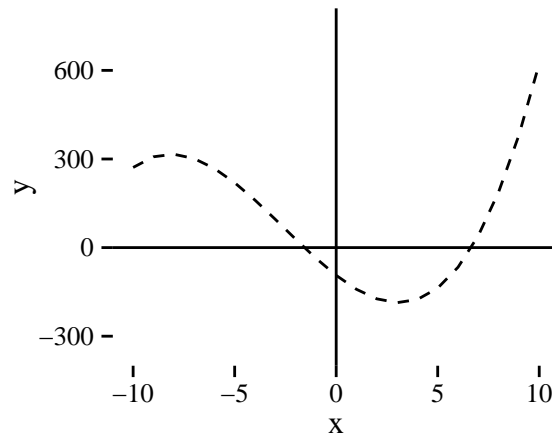


Figure 5.12: Linear combination of first- to third-order polynomials

This is fine for small values of x (10 to 10 here), where the y -values of the different polynomials are not correlated and the shapes of the curves are very different. At large values of x , however, the y -values are correlated and the curves look very similar: Fig. 5.13 shows the y -values for the same polynomials as in (5.1) and Fig. 5.11 above, but for x from 50 to 100.

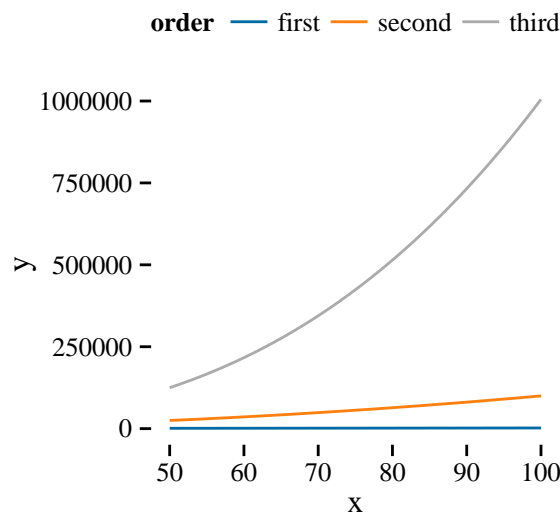


Figure 5.13: The same first-, second-, and third-order polynomials for larger x

Because the curves have reasonably similar shapes in this range, a linear combination of them cannot approximate a curve with peaks or troughs in this range. Furthermore, correlated predictors (which would be the ultimate use for the y -values of these polynomials) are highly problematic in regression modelling.

Orthogonal polynomials are designed to have the ‘interesting’ and useful shapes like in Fig. 5.11 over a specific range. The ‘orthogonal’ aspect of orthogonal polynomials means, briefly, that the

polynomials of different orders are not correlated for the given range. The lines toward the left side in Fig. 5.14 are the orthogonal polynomials up to third order for the range of -10 to 10 , and the lines toward the right are the orthogonal polynomials for the range of 50 to 100 . Of course, the polynomials for different ranges have different formulae: the ones used here were calculated by the `poly()` function in R and are given in (5.3) and (5.4) (with numbers rounded to three decimal places⁶³, and in normalized form where appropriate). For details on how orthogonal polynomials are calculated, see Kennedy and Gentle (1980:343–344).

(5.3) Orthogonal polynomial formulae for the range -10 to 10 (left-hand lines in Fig. 5.14)

a. $y = 0.036 \times x$

b. $y = -0.245 + (-5.920 \times x) + (0.007 \times x^2)$

c. $y = (-6.830 \times 10^{-17}) + (-0.083 \times x) + (7.391 \times 10^{-19} \times x^2) + (0.001 \times x^3)$

(5.4) Orthogonal polynomial formulae for the range 50 to 100 (right-hand lines of Fig. 5.14)

a. $y = -0.713 + (0.010 \times x)$

b. $y = 3.910 + (-0.108 \times x) + (0.001 \times x^2)$

c. $y = -21.991 + (0.923 \times x) + (-0.017 \times x^2) + (5.601 \times 10^{-5} \times x^3)$

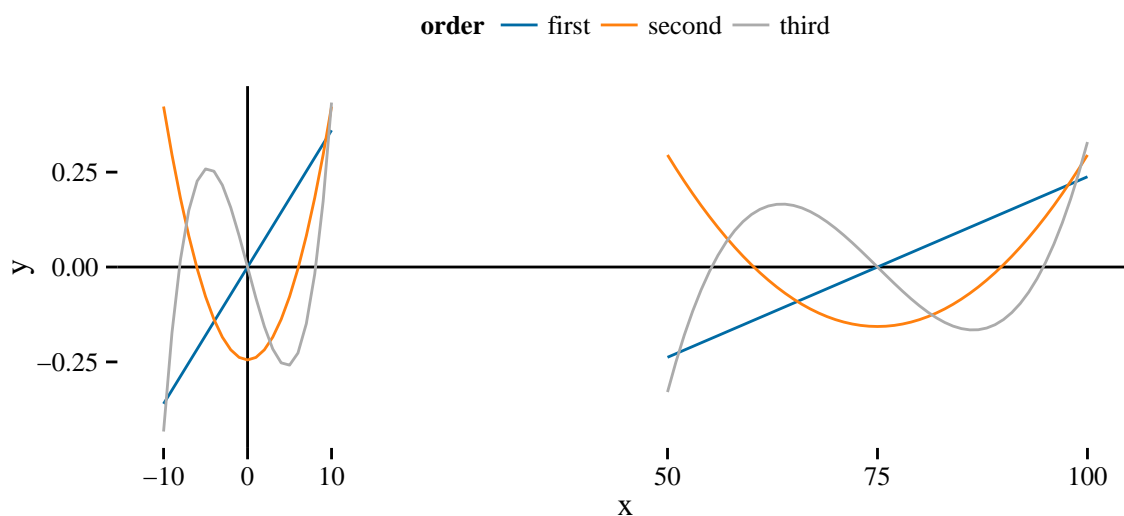


Figure 5.14: Orthogonal polynomials for two different ranges

In GCA, the y -values of orthogonal polynomials up to a given order are calculated from time codes, and these orthogonalized time values, but not the bare time codes themselves, are then entered as predictor variables into a linear model. I will refer to orthogonalized time values of

⁶³The R commands `print(poly.orth(-10:+10,3), digits=16)` and `print(poly.orth(50:100,3), digits=16)`, using the `poly.orth()` function from the `PolynomF` package (Venables 2010), return the exact values for these formulae.

order n as t_n in the interest of brevity: it allows the rather long function in (5.5) (where β_{2o1} is the coefficient for the first-order term in the formula for the second-order orthogonalized polynomial, so for example -5.920 in (5.3b) above) to be expressed as the less confusing (5.6). In both formulae, a is used as a stand-in for some predictor not related to time (area of interest/image, for example), which is of course constant over time, and ϵ is the error term.

$$(5.5) \quad \text{percentage} = \beta_0 + (\beta_1 \times (\beta_{1o0} + \beta_{1o1} \times t)) + (\beta_2 \times (\beta_{2o0} + \beta_{2o1} \times t + \beta_{2o2} \times t^2)) + (\beta_3 \times (\beta_{3o0} + \beta_{3o1} \times t + \beta_{3o2} \times t^2 + \beta_{3o3} \times t^3)) + (\beta_4 \times a) + (\beta_5 \times (a \times (\beta_{1o0} + \beta_{1o1} \times t))) + (\beta_6 \times (a \times (\beta_{2o0} + \beta_{2o1} \times t + \beta_{2o2} \times t^2))) + (\beta_7 \times (a \times (\beta_{3o0} + \beta_{3o1} \times t + \beta_{3o2} \times t^2 + \beta_{3o3} \times t^3))) + \epsilon$$

$$(5.6) \quad \text{percentage} = \beta_0 + (\beta_1 \times t_1) + (\beta_2 \times t_2) + (\beta_3 \times t_3) + (\beta_4 \times a) + (\beta_5 \times (a \times t_1)) + (\beta_6 \times (a \times t_2)) + (\beta_7 \times (a \times t_3)) + \epsilon$$

With the best estimates for the coefficients β (which is a solved problem in mixed-effects modelling) and a high enough order of polynomial, this polynomial decomposition can approximate curves with very complex shapes. This makes it ideal for GCA, which thus can control for time very closely and make use of the continuous nature of both time and eye gaze percentages without suffering from the problems that collinear predictors cause. Of course, outside of the range that a particular set of orthogonal polynomials is designed to be ‘interesting’ in, the values are correlated and the curves are fairly similar. This means that a GCA analysis must specify that range when computing the polynomials, and that GCA models cannot be generalized outside of that range. The former is easy (in R), and the latter is not a problem since GCA is not intended to make generalizations outside of the range. With the well-known tools of mixed-effects modelling, GCA can include random effects for individual participants, and the resulting model of the underlying curves can be analyzed and compared.

The following examples show how well a regression model using orthogonalized polynomial predictor variables can model any appropriate curve shape (given a high enough order) without overfitting. Fig. 5.15 shows two examples of simple (non-orthogonalized!) third-order polynomials (grey lines), data points generated from them by adding random noise (black dots), and the linear models (orange lines) of those data points based on orthogonal polynomials of up to third order. It is immediately apparent how close the orange model lines get to the ‘true’ grey lines.

The first random dataset (Fig. 5.15a) consists of 100 data points based on $y = -7 + (0.01 \times t)$, with uniform random noise added. The coefficients of model of that data, shown in Table 5.3a, show clearly that it did not find the second- and third-order orthogonal polynomial values (t_2 and t_3) to be significant (vanishingly small coefficient estimates and $p > 0.1$), and indeed a model without those terms (not shown here) is virtually identical to this one. The model fit on the



(a) Data generated by first-order orthogonal polynomial (b) Data generated by first- and third-order orthogonal polynomials

Figure 5.15: Third-order orthogonal polynomial-based linear models (orange lines) for data (black dots) generated by polynomials (grey lines)

second random dataset (201 points based on $-7 + (-2 \times t) + (0.1 \times t^3)$, with uniform random noise) performs equally well: Table 5.3b shows that it correctly identifies the contributions of the first- and third-order orthogonal polynomials t_1 and t_3 and correctly finds the second-order polynomial t_2 to be insignificant (small coefficient estimate compared to the standard error, and thus $p > 0.1$).

	estimate	standard error	t	p
intercept	-6.49	0.01	-1186.10	< 0.0001
t_1	2.87	0.05	52.41	< 0.0001
t_2	0.03	0.05	0.60	0.55
t_3	-0.00	0.05	-0.08	0.94

(a) Non-zero parameter for first-order orthogonal polynomial

	estimate	standard error	t	p
intercept	3902.04	3820.40	1.02	0.31
t_1	505089.39	54163.56	9.33	< 0.0001
t_2	60129.12	54163.56	1.11	0.27
t_3	232423.17	54163.56	4.29	< 0.0001

(b) Non-zero parameters for first- and third-order orthogonal polynomials

Table 5.3: Coefficients of linear models based on orthogonal polynomials

These examples show how one can model any curve shape that approximates a third-order polynomial using the three exponential components from orthogonal third-order polynomials instead of the simple time variable as predictors in a linear regression model. Furthermore, the regression model correctly finds insignificant exponential components to be insignificant, even in the presence of considerable noise. With the right random effects structure (see Barr et al. 2013) and the right order of polynomial (see Mirman 2014), GCA, which makes use of orthogonal polynomials, is an appropriate method for modelling and comparing smooth but complex curves, such as eye gaze percentages.

5.2.4.3 Smoothing spline analysis of variance (SSANOVA)

This section uses the `eyetrack` dataset provided as part of the R package `gss` (Gu 2014) to briefly introduce smoothing splines and SSANOVA.⁶⁴ To make this introduction more clear, this first paragraph describes that dataset. The dataset contains the number of gazes that fell on each of three images for several trials and six participants, over a time window of 2.3 seconds. The three images are a target image, an object-matching distractor (of a different color), and a color-matching distractor (different object). As this dataset does not appear to have gaps like those discussed in Section 5.2.2, the percentages of gazes on each image for a given time point are easily calculated. Fig. 5.16 shows the averaged proportions for each of the three images and the implicit ‘elsewhere’ percentage (for gazes on empty regions of the screen), averaged across all trials and participants.

Smoothing splines are curves, not regression lines. Thus, they work best when there is only one data point for each value of x . Time series fit this definition very naturally: for each time point, there is only one value of the variable of interest. This variable of interest can in principle be anything that can be operationalized and that changes over time. Some linguistic examples are frequency values of formants (Derrick and Schultz 2013), distance between tongue and alveolar ridge (Ardestani 2013), and pitch (Zhang et al. 2014), but proportions of eye gazes for different images on screen can also be analyzed using smoothing splines. For example, Fig. 5.17 shows the proportion of the color-matching distractor image in the `eyetrack` data (in other words, Fig. 5.17 is just the blue line from Fig. 5.16, with a different scale on the y-axis for clarity).

It is obvious that this curve is ‘jittery’. If one assumes there is an underlying smooth pattern or function here (caused by the gaze being attracted to this distractor image before being more strongly attracted by another image), this means that there is some noise in the data. Nevertheless, the overall pattern over time is immediately apparent even from this noisy data. One might be

⁶⁴For a much more comprehensive and mathematically rigorous introduction to smoothing splines, see de Boor (2001). For a comprehensive and rigorous introduction to SSANOVA, see Gu (2013). Finally, Davidson (2006) was the paper that introduced SSANOVA to linguistics.

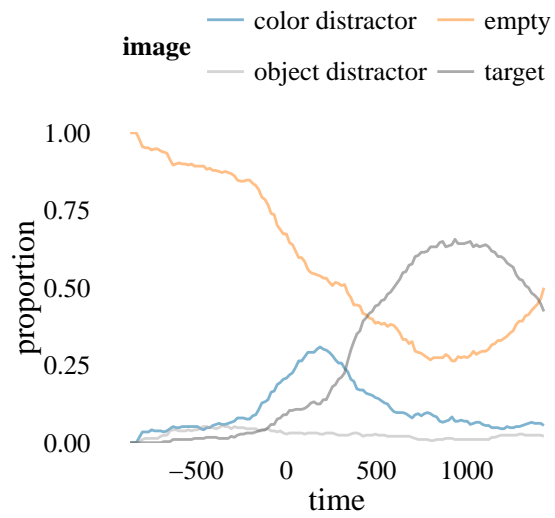


Figure 5.16: Average proportions of gazes on four different regions of the screen over time (in milliseconds)

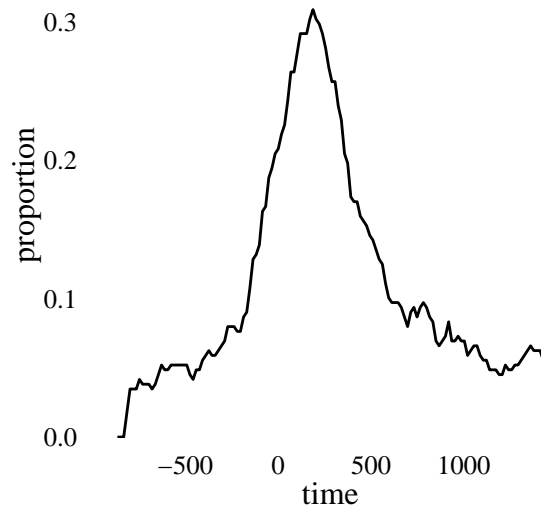


Figure 5.17: Average proportion of gazes on the color-matching distractor image over time (in ms)

interested in representing the pattern more abstractly, without the need to list every single data point and with some reduction in noise. Smoothing splines do this by splitting the range of x (time, here) at a number of points called ‘knots’ and finding the best-fitting polynomial curve of no higher than a set order, usually third, for each interval between two successive knots. These curves are constrained to result in a curve that is ‘smooth’ by requiring that a certain derivative (often the second-order derivative) of the whole curve be continuous.

‘Continuous’ here means that no gaps are allowed: there must be a y -value for each x -value. If a curve has gaps in it, it is not continuous. The derivative function of a curve calculates the slope of that curve at all of its points. If a curve has gaps in it, its derivative also has gaps in it, as the area where the gap is cannot have a slope. If a curve has kinks (abrupt transitions from one straight line to another straight line) in it, there is no slope at the precise point of the kink: at that point, the curve does ‘no longer’ have whatever slope it has just before that point, but it does ‘not yet’ have the slope it has just after that point. Any other choice of value for the slope at the kink would be arbitrary. Therefore, there is no slope here. This means the derivative function does not have a y -value for that x -value. Thinking of the derivative function as a curve itself, this means it has a gap at that precise x -value—it is not continuous. If the derivative function can be thought of as a curve, it too has a slope, of course. The function that calculates this slope of the derivative is the second-order derivative (and the derivative of the ‘original’ curve is the first-order derivative). A curve with a transition between a straight line and a parabolic curve (described by x^2) is continuous itself, and its first-order derivative is also continuous. However, the first-order derivative has a kink in it, and so the second-order derivative (the slope of the first-order derivative curve) is not continuous. In this way, requiring derivatives up to a certain order to be continuous imposes requirements of gaplessness and smoothness on the ‘original’ curve, which is why this requirement is enforced on smoothing splines.

The number of knots is crucial, of course, but quite complex curve shapes can be approximated even with a relatively low number. Fig. 5.18 shows three smoothing splines fit to the data shown in Fig. 5.17 (and repeated here as a dotted line). The blue spline, fit with only five knots, matches the general shape of the data curve, but is not very accurate at most time points and makes curious fitting errors around 1000 ms. The orange spline, fit with 10 knots, is much more accurate. The grey spline, fit with 20 knots, captures some peaks and troughs that the orange one missed, but it does not appear to be a big improvement in accuracy.

The inherent smoothness of smoothing splines may seem like a drawback, considering how hard this makes it for the spline to capture extreme peaks or troughs in the data. In practical use, however, it turns out to be an advantage, as it makes spline-based methods conservative, or less likely to overfit the data. Smoothing splines are unsuitable for data where sudden and extreme spikes can be expected, because such spikes are obviously not smooth and thus hard to

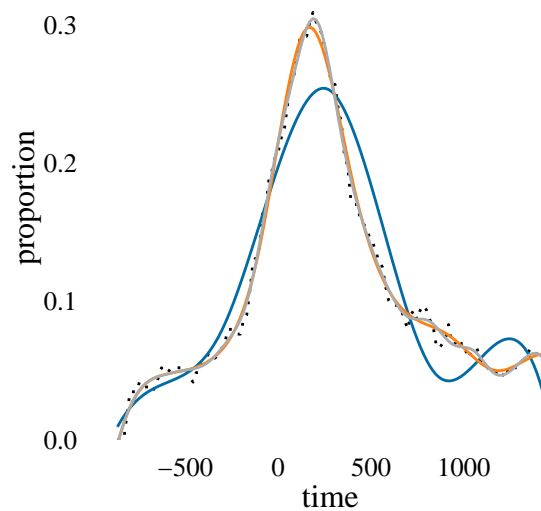


Figure 5.18: Average proportion of gazes on the color-matching distractor (dotted line), and smoothing splines with five (blue line), ten (orange line) and twenty (grey line) knots

approximate with a smoothing spline. Data without such spikes, however, is easily simplified into a smoothing spline—and, as I have argued in Section 5.2.4, eyetracking percentage data fits that description. Their inherent smoothness makes smoothing splines less likely to model a spike even when there does appear to be one due to missing data, which is possible in the present study due to the lossy eyetracking method (see Section 5.2.2).

Several smoothing splines can be fit to different datasets and compared easily, and SSANOVA is a method of making these comparisons in a rigorous way. A smoothing spline is fit for each of the different groups in the data: one spline of gaze percentages for each image in the eyetrack data (and in the present study), one spline of formant frequencies for each phonological environment in Derrick and Schultz (2013), and so on. Then, the 95% confidence interval around each spline is computed (Gu 2013:75–79). Unlike more usual ANOVAs, SSANOVA does not lend itself to *F*-tests (Davidson 2006:411). Rather, the current best practice for SSANOVA appears to be visual analysis: “A more precise technique for determining where the differences . . . lie is to construct 95% Bayesian confidence intervals around the smoothing splines. The two curves are significantly different where the confidence intervals are not overlapping” (Lee-Kim et al. 2013:486–487), and “white space between the . . . lines . . . represents a statistically significant difference” (Derrick and Schultz 2013:5).⁶⁵ Returning to the eyetrack example data, Fig. 5.19 shows the SSANOVA fits for that dataset as dashed lines, and the faded band around each dashed line represents the 95% confidence interval (1.96 standard errors either way, meaning that the errors are assumed to be normally distributed).

⁶⁵For other examples of this visual analysis of SSANOVA in practical use in a variety of fields, see Ratkovic

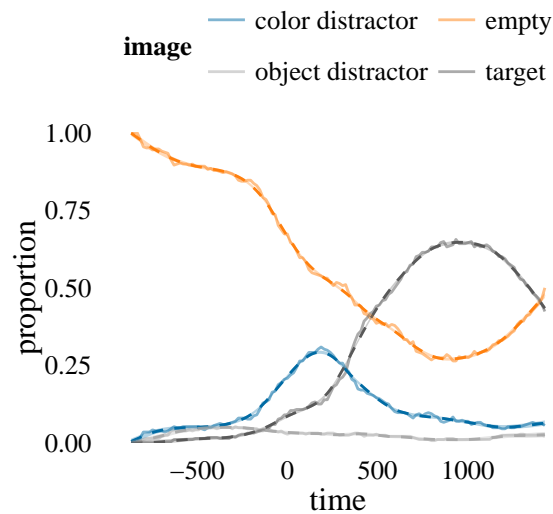


Figure 5.19: Average proportion of gazes (solid lines) and SSANOVA fits (dashed lines) with 95% confidence intervals (faded bands)

The SSANOVA splines fit the data very well, even showing how the percentage of gazes on empty space drops initially but then levels off slightly at around -500 ms, for example. At the same time, the splines are smooth and do not overfit the ‘jittery’ noise. The confidence intervals are very narrow here because the `eyetrack` dataset is quite large and does not have missing data. Nevertheless, the intervals for the two distractor images (blue and light grey bands) overlap around -500 ms, showing how the percentages of gazes for these two images are not significantly different at that time.

Thus, smoothing splines can model even complex curves with good accuracy but without overfitting, even when there are missing values and noise in the data. Comparing different splines is easy. Random grouping effects can be included in an SSANOVA model (Gu 2013:216). Finally, since SSANOVA achieves complex curve shapes not through the trick of orthogonalized polynomials, but rather by combining a succession of standard polynomial curves, it is simpler and more elegant. In conclusion, therefore, SSANOVA is also an appropriate method for modelling and comparing eye gaze percentages.

5.2.4.4 Comparing these methods

I have argued here that comparing mean percentages across time windows is conceptually a misguided method of analyzing eye gaze data, and that GCA and SSANOVA seem more suited for that task. Which of the latter two is better? There is no immediately apparent answer to

(2009), Chanethom (2011), Ardestani (2013), Stevenson et al. (2014), and Zhang et al. (2014).

that question. It is true that GCA (like mixed-effects models generally) seems most appropriate when there are many independent variables and one wants to test which of them are significant predictors of the data while controlling for all others. SSANOVA, on the other hand, is designed to model (and investigate the differences between) just a few curves—based on only one or two grouping variables of interest. In eyetracking studies that fit broadly into the visual world paradigm, there are generally only a few curves that are compared. While this appears to make SSANOVA a better fit, GCA is not incapable of dealing with just a few variables just because it is designed to handle many. To answer the question of which method is better, this section compares the performance of GCA and SSANOVA using the `eyetrack` dataset and two sets of randomly generated, simulated data.⁶⁶

The `eyetrack` dataset has data for all timepoints from all participants in all trials, as mentioned in Section 5.2.4.3. To see how GCA and SSANOVA perform on less ideal data, this section will use only a subset of the dataset, consisting of six (out of 48) sessions. Fig. 5.20 shows that this subset exhibits the same broad pattern of proportion over time, but with certain differences: for example, gazes on the ‘empty’ region appear to increase and gazes on the color-matching distractor image appear to decrease at around 500 ms in the subset, whereas the whole dataset (shown in Fig. 5.16 above) does not have this peak and trough.

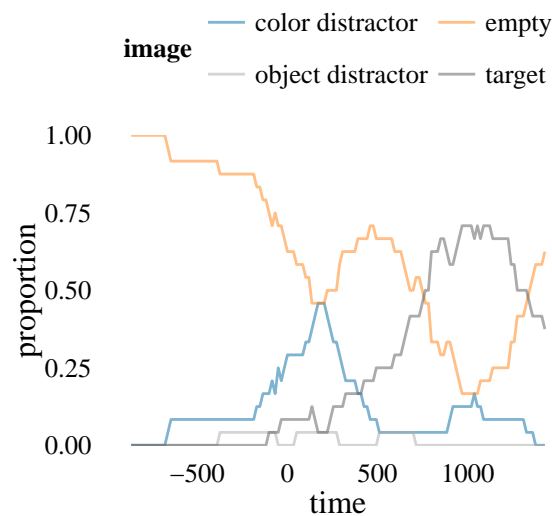


Figure 5.20: Average proportions of gazes on four different regions of the screen over time (in milliseconds) in the subset

As demonstrated in Section 5.2.4.2 above, a linear model with orthogonalized polynomials of time as predictors (as in GCA) can approximate data with non-linear changes over time, while

⁶⁶I am grateful to Dave Kleinschmidt for suggesting simulated data as a test of performance. Any errors in the implementation of this idea are of course my own.

still being linear and therefore allowing the use of tools developed for linear models. Fig. 5.21 shows the best fit of a GCA model with third-order polynomial terms to the subset used here. While the parameters of a GCA model are not easy to interpret, this method allows the relatively simple statistical method of regression modelling to be extended to non-linear data. This model is still a linear regression model: for example, the formula in (5.7) describes the model curve of gazes on ‘empty’ regions (orange dashed line in Fig. 5.21).

$$(5.7) \quad Y = 0.626 + (-2.598 \times t_1) + (0.407 \times t_2) + (0.480 \times t_3) + \epsilon$$

This formula multiplies predictors by parameters and adds constants (the intercept 0.626 and the error term ϵ). It is entirely ignorant of the fact that the predictors t_2 and t_3 were calculated using quadratic and cubic polynomials.

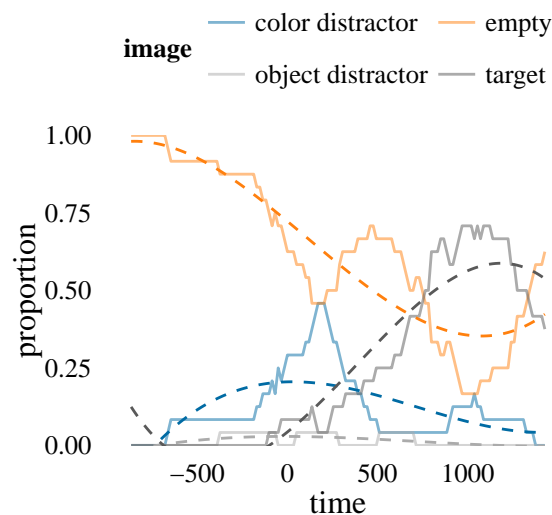


Figure 5.21: Average proportions of gazes in the subset (solid lines) and third-order GCA model fits (dashed lines)

Third-order polynomials are essentially limited to one peak and one trough. Therefore, this model does fairly well on those curves that can be approximated with one peak and trough (target and object-matching distractor), but only captures a very rough overall shape of more complex curves (color-matching distractor and empty region). To allow for more peaks and troughs and thus enable GCA to fit the curve shape more accurately, orthogonalized polynomial terms of higher order are necessary.

The best-fitting GCA model using seventh-order polynomial terms is shown in Fig. 5.22, and (5.8) is its formula for the ‘empty’ (orange) curve. This model is obviously a closer fit than the third-order GCA above, but that comes at a cost. Firstly, this is very computationally expensive. The somewhat arbitrary order of 7 was chosen here simply because models with

higher orders took very long to fit on a typical desktop computer. Moreover, the model fitting function (`lmer()` from the `lme4` package; Bates et al. 2014) warned that it failed to converge on the data with seventh-order polynomials, which suggests (very briefly) that the model as specified is not appropriate for the data as it is. Secondly, the model is all but impossible to interpret: most of its 32 parameters achieve significance (under the assumption that the t -values of estimated parameters approximate a standard normal distribution and that parameters whose t has an absolute value of more than 1.96 are therefore significant at $\alpha = 0.05$). While this at first glance seems to reflect the fact that the seventh-order curves are obviously better fits than the third-order ones, there is no objective way of finding the best order for a GCA model. A model of lower order (fifth, say) would be less likely to overfit, but that would mean dropping the ‘significant’ higher-order terms. Thirdly, the seventh-order model appears to be overfitting the data. Consider for example the color-matching distractor and empty region gazes before -500 ms, where this model suggests that there is an initial rise in the proportion of gazes toward the color-matching distractor (and concurrent fall in gazes on empty regions), followed by a drop before the major rise in proportion. This pattern is not apparent in the raw data, meaning the model is inaccurate at best and misleading at worst here. This is a known problem: terms of higher order “are just capturing differences in the tails” (Mirman 2014; “tails” here means the beginning and end of the time window under analysis).

$$(5.8) \quad Y = 0.626 + (-2.598 \times t_1) + (0.407 \times t_2) + (0.480 \times t_3) + (0.472 \times t_4) + (0.738 \times t_5) + (0.280 \times t_6) + (0.240 \times t_7) + \epsilon$$

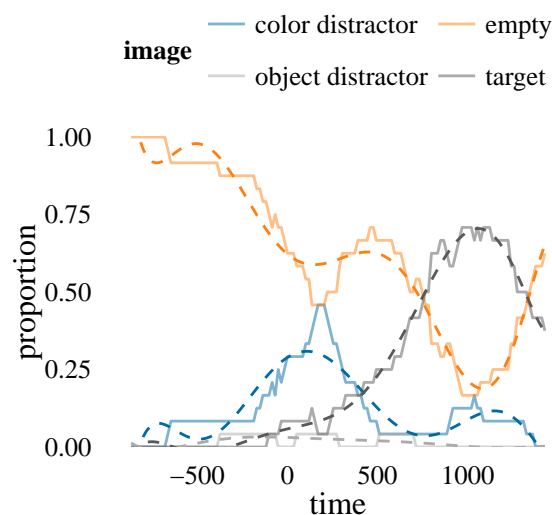


Figure 5.22: Average proportions of gazes in the subset (solid lines) and seventh-order GCA model fits (dashed lines)

The best SSANOVA fit to the same subset manages a fairly good fit, but without overfitting:

Fig. 5.23 shows that the SSANOVA model captures the initial minor increase in gazes on the object-matching distractor image (around -500 ms), but without modelling a fictitious drop shortly thereafter. At the same time, it provides smooth curves: SSANOVA correctly does not model the color distractor and empty curves to be on the same level at around 200 ms, even though the data suggest this. The smoothness of splines makes them unlikely to fit this extreme peak, and comparing the SSANOVA fits in Fig. 5.23 to the full data in Fig. 5.16 shows that the larger dataset is very well approximated by SSANOVA even based on the small subset.

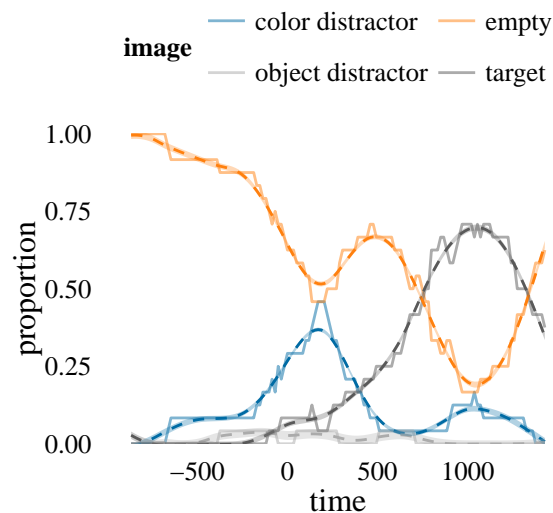


Figure 5.23: Average proportions of gazes in the subset (solid lines) and SSANOVA model fits (dashed lines) with their confidence intervals (faded bands)

Thus, on this one realistic dataset, SSANOVA appears to perform better than GCA: it is less computationally expensive, it is less likely to overfit, and makes finding areas of difference between curves easier thanks to the 95% confidence intervals.

However, this may be coincidence: it is conceivable that the `eyetrack` dataset is singularly well suited to SSANOVA (the dataset is provided as part of an SSANOVA package, after all). Therefore, I tested the performance of GCA and SSANOVA on simulated data in two sets of simulations. This allowed more robust testing, since it is easy to simulate thousands of simple random datasets. It also allowed me to make these simulated datasets much less perfect than the `eyetrack` data by introducing a large random error into the data, and thus make these simulated datasets more like the real eyetracking data collected in the present study. (These simulations were run in R (R Development Core Team 2011); see Appendix G for details and code.)

The starting point for each of these two sets of simulations were data-generating functions. A basic polynomial of the form given in (5.9) was defined for each simulation by specifying the twelve β -values (resulting in effectively a GCA-like model, to counter the possible bias

for SSANOVA in the eyetrack dataset). The variables x_1 , x_2 , and x_3 are the values of orthogonalized polynomials of first, second, and third order for the range of a continuous integer variable x (simulating time bins); d_1 and d_2 are two of the three levels of the grouping variable (thus simulating a study with three areas of interest; note that there is no variable for the third level because that was used as the reference level); and the variables $x_n \times d_n$ are the interaction terms between the orthogonal polynomials and the two grouping variable levels (which are necessary in order to allow for different curve shapes over time for the different levels).

$$(5.9) \quad y = \beta_0 + (\beta_1 \times x_1) + (\beta_2 \times x_2) + (\beta_3 \times x_3) + (\beta_4 \times d_1) + (\beta_5 \times d_2) + (\beta_6 \times (x_1 \times d_1)) + (\beta_7 \times (x_1 \times d_2)) + (\beta_8 \times (x_2 \times d_1)) + (\beta_9 \times (x_2 \times d_2)) + (\beta_{10} \times (x_3 \times d_1)) + (\beta_{11} \times (x_3 \times d_2))$$

Each of the two sets of simulations had two of these data-generating functions, a null function with $\beta_1 \dots \beta_{11}$ set to 0 and an effect function with some of those β -values different from 0. To introduce randomness,⁶⁷ grouped errors were added to simulate 20 participants: firstly, the constant β -values were changed by adding or subtracting random values for each simulated participant, and a random error ϵ was added to each y -value. Both of these randomizations were based on normal distributions with mean 0. Thus, each simulated participant had their individual generation function, and the output of these was subject to further noise. As a result, each dataset in these simulations had 20 distinct y -values for each combination of x (time) and d (area)—one for each simulated participant.

Each function was run 1000 times (with the randomization being (randomly) different for each run). GCA and SSANOVA models were then fit onto each of these datasets, and their performance was compared. The following paragraphs describe the results of these simulations in two sets: the first set consists of 1000 datasets based on a function with just one β being not 0 and 1000 datasets based on the corresponding function with all β -values being 0. The second set consists of 1000 datasets based on a more realistic and complex function with several β -values being not 0 and 1000 datasets based on its counterpart with all β -values being 0.

In the first set of simulations, only one β in the effect model (apart from the intercept β_0) was not 0: the parameter for the interaction between the ‘response’ level of the image factor and the second-order orthogonalized polynomial value was set to 1, thus giving a parabola-like trajectory to the curves for the percentages of gazes at the ‘response’ and ‘other’⁶⁸ images (see Fig. 5.24b). The null model had this and all other $\beta_1 \dots \beta_{11}$ set to 0 (and the same intercept β_0 as the effect model), thus simulating the three absolutely identical flat lines in Fig. 5.24a (where

⁶⁷Since all simulation data was computer-generated, strictly speaking the errors were only **pseudorandom**. I do not consider the distinction to be relevant for these simulations, and will continue to use the term ‘random’ for simplicity’s sake.

⁶⁸The ‘other’ level of the image variable received the inverse of this effect due to the way contrasts were handled in setting up the dummy variables for that categorical variable in this simple simulation.

the three lines have been given different line types to make it more apparent that they overlap completely).

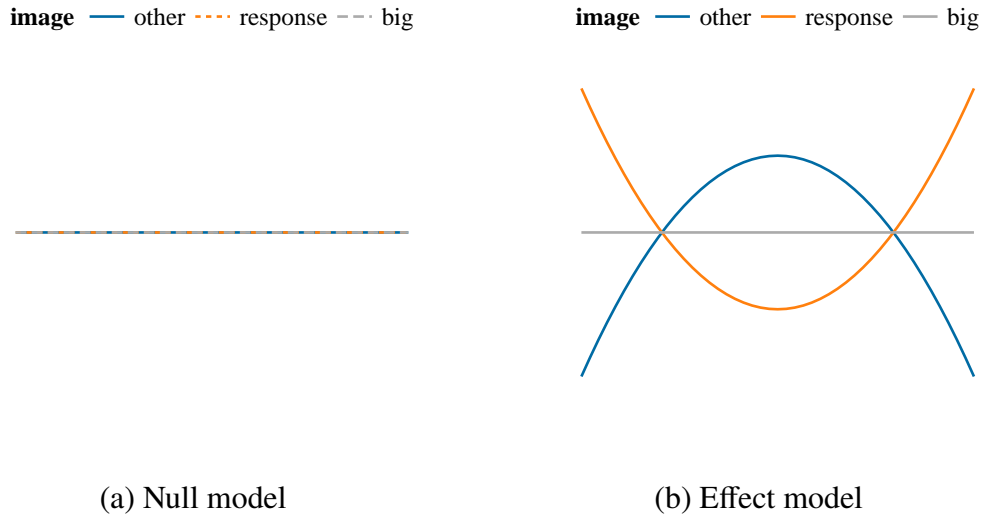


Figure 5.24: Data generation functions for the first set of simulations (before being randomized for simulated participants)

Note that this was the data **before randomization**. With the random errors added to parameters and y-values, the data was much more messy, as is apparent from comparing the two simulated datasets in Fig. 5.25 to the underlying functions in Fig. 5.24.

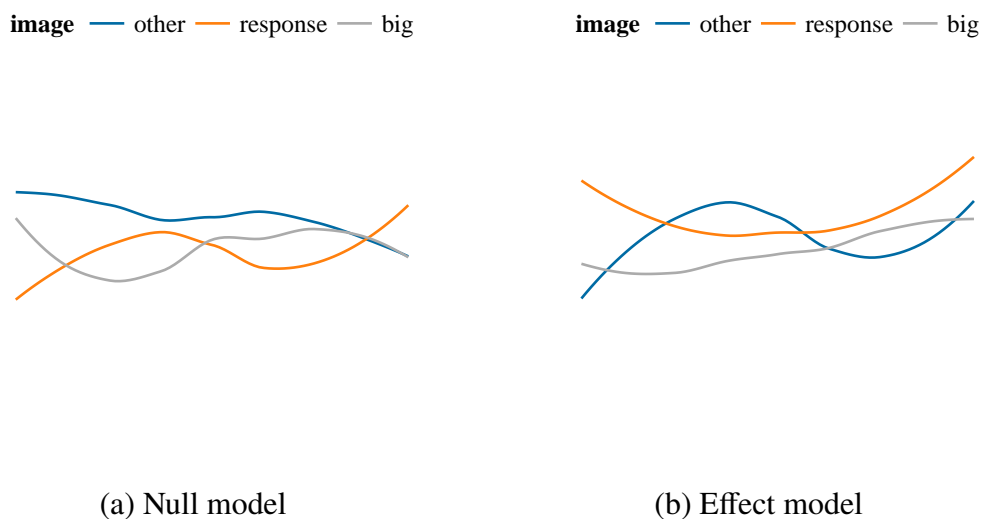


Figure 5.25: Examples of datasets in the first set of simulations (averaged across simulated participants)

1000 datasets like this were generated, a GCA model was fitted to each simulated dataset, and each model’s parameter estimates, their standard errors, and the t -values resulting from these values were saved for analysis. On the assumption that the t -values approximate a standard normal distribution (which is the assumption that underlies how the p -values for parameter estimates are often calculated), the t -values for the parameter of interest should ideally fall between -1.96 and 1.96 in 95% of the simulations without an effect—in the simulation, they fall in this interval 96.4% of the time (see Fig. 5.26). The t -values for the other predictors (which were all zero in the data-generating function) should also be in this interval 95% of the time, and they are 95.1–97.4% of the time.⁶⁹ Similarly, the t -values for this parameter should ideally be outside of that interval in 97.5% of the simulations with an effect—and they are in 91.4% of those simulated models.⁷⁰ Thus, the GCA approach to this kind of data does not have textbook-level/nominal performance (apparently it slightly underestimates the effect and is thus conservative), but it does get very close.

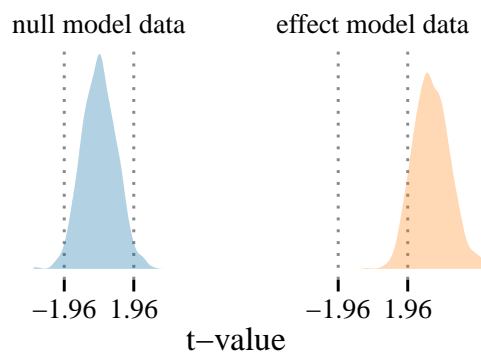


Figure 5.26: Density plot of t -values of GCA models in the first set of simulations

SSANOVA models were also fitted to these 1000 simulated datasets. Since SSANOVA models are analyzed by whether the 95% confidence intervals of different splines overlap (see Section 5.2.4.3 above), this was how these models were evaluated. Fig. 5.27 shows the SSANOVA splines with 95% confidence intervals for nine null-model and nine effect-model datasets as examples.

The number of models that had non-overlapping confidence intervals for at least one time bin was 130 for the simulations without an effect and 331 for the models with an effect—in other words, the false positive rate was 13% and the false negative rate an astonishing 66.9% in this simulation. The size of the differences did not differ between the true and the false positives: some models (like #134 in Fig. 5.27a and #139 in Fig. 5.27b) had only small differences in the time interval (sequence of bins) where they did have a difference; others (#135 in both Fig. 5.27a

⁶⁹The intercept (grand mean) was far from zero in the simulated data, and consequently its t -value is outside the same interval in all 1000 simulation models.

⁷⁰The other predictors are again zero, and their t -values should again be between -1.96 and 1.96 in 95% of models—and they are in 94.3–97.4% of the models.

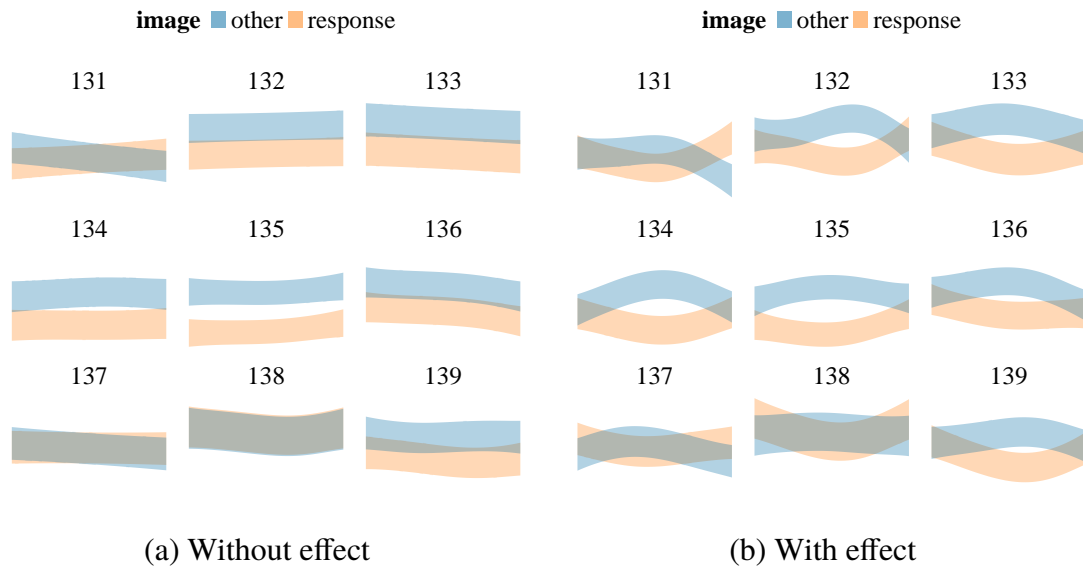
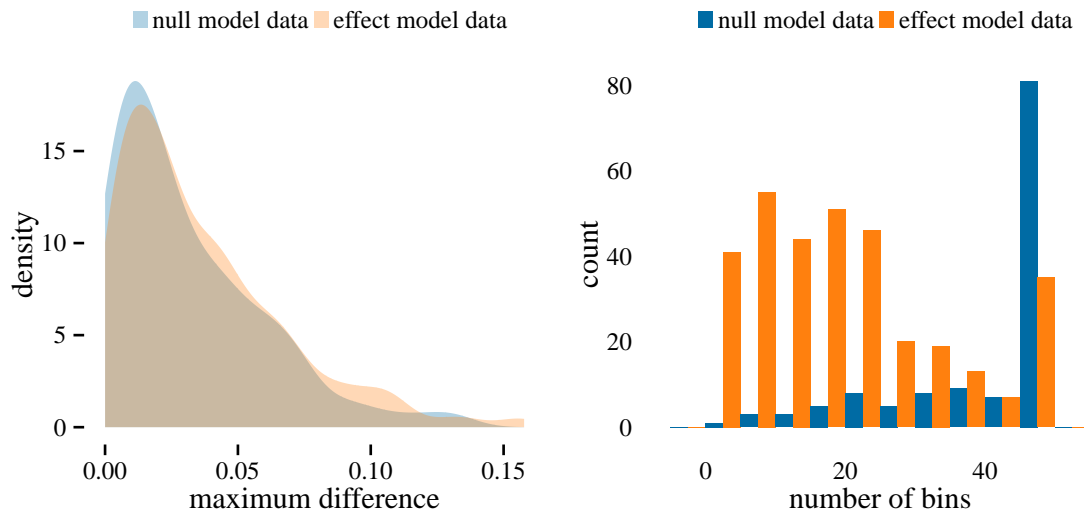


Figure 5.27: Examples of SSANOVA models in the first set of simulations

and Fig. 5.27b) showed a much larger difference. Fig. 5.28a is a density plot of the largest vertical differences (one from each model that did have a difference), with the false positives (shown in blue, based on null-model data) and true positives (shown in orange, based on effect-model data) being virtually indistinguishable. This is puzzling, but the difference between the two models #135 in Fig. 5.27 is enlightening: the false positive #135 has a difference all the way through the simulated range of time (it looks like a constant offset rather than an interaction between image and time), while the true positive #135 (like all the models with a difference in Fig. 5.27b) only shows a difference for part of that range. In fact, 79 of the 130 false positives (60.8%) did show a difference all the way through, whereas only 32 of the 331 true positives (9.6%) did. Fig. 5.28b, a histogram of the length of these differences, makes very obvious that false positives mostly showed their difference for all bins and that true positives mostly had a difference for only a part of the time range.

Redefining a ‘positive’ SSANOVA finding as showing a difference between confidence intervals **and** this difference not lasting for the entire time range seems warranted. This, of course, changes the numbers: there are 51 false and 299 true positives under that definition, so $\frac{299}{51+299} = 85.4\%$ of the positives are now true ones (up from $\frac{331}{130+331} = 71.8\%$ under the previous definition). The SSANOVA approach to this type of data as a whole still appears to be overly conservative, especially when compared to the GCA’s performance above, but it must be remembered here that the data was generated by what is essentially a GCA model—and an unrealistically simple one, at that. As the examples in Mirman et al. (2008), Barr (2008), and Mirman (2014) show (and as attempts at modelling subsets of the data gathered in the present study showed), GCA models of real data generally have either no significant parameter β_n or several ones, but not



(a) Density plot of models by largest difference between the two confidence intervals (b) Histogram of models by number of bins showing significant difference

Figure 5.28: Comparison of false and true positives in SSANOVA models in the first set of simulations

just one. Simulations where the effect model had several significant parameters would therefore be more realistic.

The second set of simulations was designed to meet that goal. These simulations were like the ones in the first set, except that the underlying data-generation function for the effect model had non-zero values for the intercept, the main effect for the third-order time polynomial values, and three interactions between time polynomials and image levels. These were chosen from the effects with $p < .05$ in a GCA model of the adult participants' data from experiment 2 of this thesis (the data that is presented in Section 5.3), split by whether the area contained their ultimate response choice, one of the other two choices, or the larger image representing the explicit object in the instruction. All parameters corresponding to effects which had $p \geq .05$ in that empirical model were set to 0 in the effect model here. Fig. 5.29b shows this underlying model. The null-model again had all parameters except the intercept set to 0, see Fig. 5.29a. Just as in the first set of simulations, random errors were introduced into these models to generate 2000 datasets (1000 null-effect datasets and 1000 effect datasets). Fig. 5.29c shows the resulting gaze percentages over time from one simulated dataset generated by the null model, and Fig. 5.29d shows the gaze percentages for one dataset generated by the effect model. GCA and SSANOVA models were fitted to these 2000 simulated datasets.

For each of these 2000 GCA models, the parameter estimates, their standard errors, and the t -values resulting from these values were calculated. On the assumption that the t -values

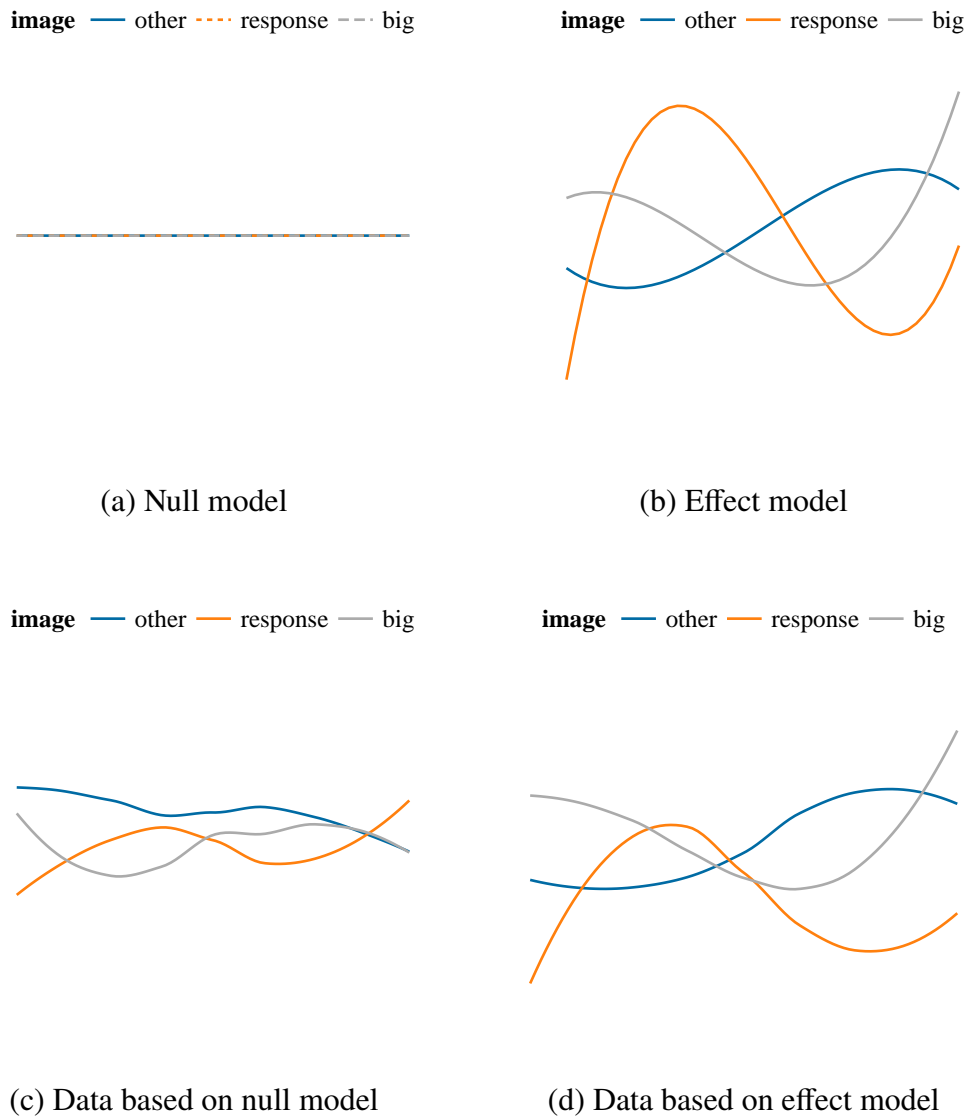


Figure 5.29: Data generation functions for the second set of simulations (top panels) and examples of simulated datasets (bottom panels)

approximate a standard normal distribution (Mirman 2014), the t -values for all parameters⁷¹ should ideally fall between -1.96 and 1.96 in 95% of the null-model simulations. As the panels on the left in Fig. 5.30 show, they do so 93.4% to 96.3% of the time (slightly different percentages for the different parameters). In the effect-model simulations, the t -values for the parameters that were not zero should ideally be outside of that interval in 97.5% of the simulations with an effect under the same assumption—and they are in 98.4% to 100% of those simulated models (see the panels on the right in Fig. 5.30). The other predictors in the effect model were 0,

⁷¹Since the intercept was not set to zero in the null model, its t -value should be outside the -1.96 – 1.96 interval in at least 97.5% of cases—in fact, it is outside this interval in all 1000 null-model simulation models.

and their t -values should again be between -1.96 and 1.96 in 95% of models—as they are in 94.5–96.3%. Thus, the GCA approach to this kind of data is again demonstrated to be good enough for practical use.

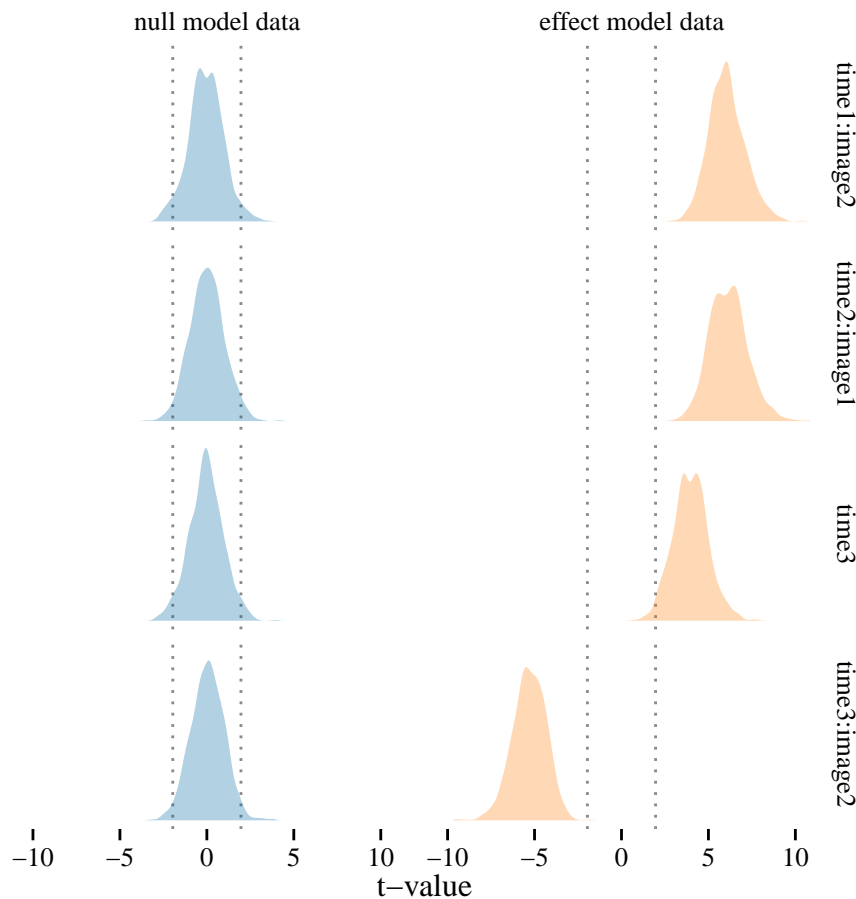


Figure 5.30: Density plot of t -values of GCA models in the second set of simulations (dotted lines at -1.96 and 1.96)

For each SSANOVA model in this set of simulations, the confidence intervals around the smoothing splines for all three images were calculated, and the number of bins where the confidence interval of the ‘response’ image did not overlap with the other two (and was above them)⁷² was recorded. Fig. 5.31 is a histogram of the number of models by number of bins with such a difference.

109 of the 1000 simulations based on the null model data did have a difference like that in at least one time bin, and might therefore be considered false positives. 57 of those showed this difference for the entire simulated range of time, while **no** model out of the other 1000 simulations

⁷²This may seem arbitrary, but the early peak for ‘response’ is certainly the most striking aspect of Fig. 5.24b and therefore the one that would most likely attract the attention of researchers.

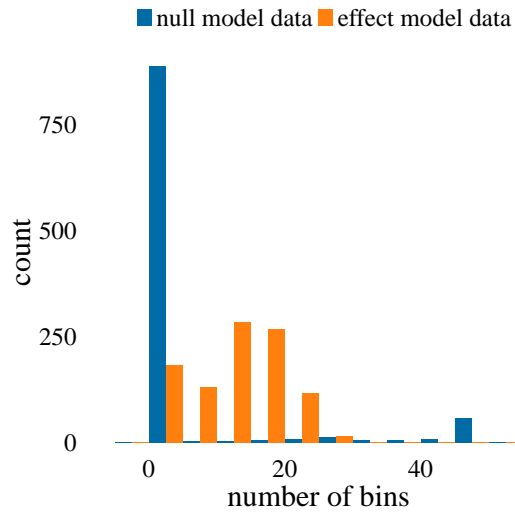


Figure 5.31: Histogram of SSANOVA models in the second set of simulations by number of bins with difference deemed interesting (see text for details)

(based on effect model data) showed the difference all the way through (the maximum number of time bins with a difference was 30, out of a total of 47 (63.8%), and only one model out of 1000 exhibited this). The difference between the positives in the two different simulation runs is made clearer in Fig. 5.32 (which is just Fig. 5.31 without the values 0 and 47): the models in the simulations based on data with an actual underlying effect cluster around 20 bins with difference, whereas the false positives (based on null model data) are more evenly spread.

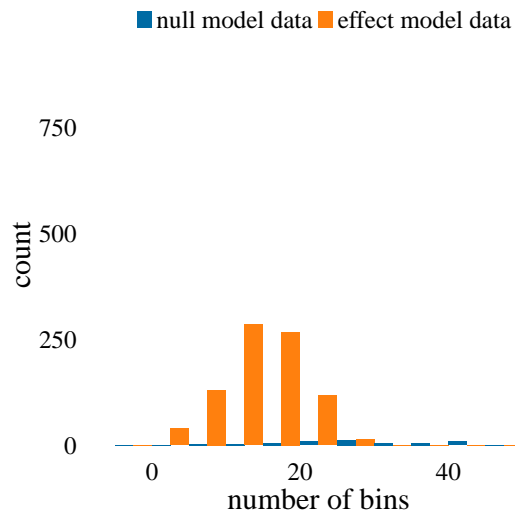


Figure 5.32: Histogram of SSANOVA models in the second set of simulations by number of bins with 'interesting' difference, without extreme values

This neatly replicates the pattern that allowed the elimination of many 'false positives' in the first set of simulations, so it seems appropriate to not consider SSANOVA models with a difference

throughout the entire time range to be positive findings. That leaves 52 (5.2%) false positives for SSANOVA in this second set of simulations.

143 of the 1000 SSANOVA models on effect-model data failed to show such a difference, so the false negative rate is 14.3% and the true positive rate 85.7%.

The cosine diagnostics (following Gu 2013:98–102) were also calculated for each SSANOVA model in this set of simulations. Briefly, if one thinks of predictors, the response, and the error term as vectors in some n -dimensional space, a good predictor would be almost orthogonal to the error term but collinear to the response. Taking the cosine of the angles between these vectors operationalizes these diagnostics into simple values between 0 and 1: the cosine of the angle between predictor and error should be close to 0, and the cosine of the angle between predictor and response should be large (approaching 1 ideally, but Gu 2013 accepts predictors for which this cosine is as small as 0.38). If a variable in an SSANOVA model has a large cosine to the error term or a small cosine to the response term, it is not a good predictor of the response and should be removed from the model.

Fig. 5.33 shows density plots of these cosine values for the second set of simulations. No major difference between the models based on the two different data-generation functions is apparent—the error-cosines for the ‘bin’ predictor and its interaction with the ‘image’ predictor tend to be larger when there is an effect in the data, but even the largest of these cosine values are still small and would not be removed in Gu (2013)’s worked examples. The cosines to the response are far from the ideal value of 1, but they are mostly in the range of values that Gu (2013) accepts. This suggests that the predictors are helpful even when there is no underlying effect to predict, which is puzzling. Recall, however, that time bin and area level really were the only variables used in generating the data as well, and that their corresponding parameters were randomized for each simulated participant. It is therefore possible that these two predictors were actually made significant by chance for some of the participants, and that the cosine diagnostic values reflect this. Furthermore, consider that Gu (2013:99-102) introduces these cosine diagnostics as a method of identifying the best few predictors in a model with many predictors and all their interactions. Since SSANOVA does not use several orthogonalized polynomial values of time, the SSANOVA models here have just two variables (time and area) and their interaction. It is hard to conceptualize what would be meant by ‘the best few predictors’ out of just three. Therefore, the cosine diagnostics are conceptually inappropriate for the analysis of eyetracking data with just one grouping variable.

In conclusion, the SSANOVA approach to this type of data still appears to be somewhat conservative, especially when compared to the near-textbook performance of GCA. However, conservativity is not a bad thing in statistical tests. At 5.2%, the false positive rate of SSANOVA

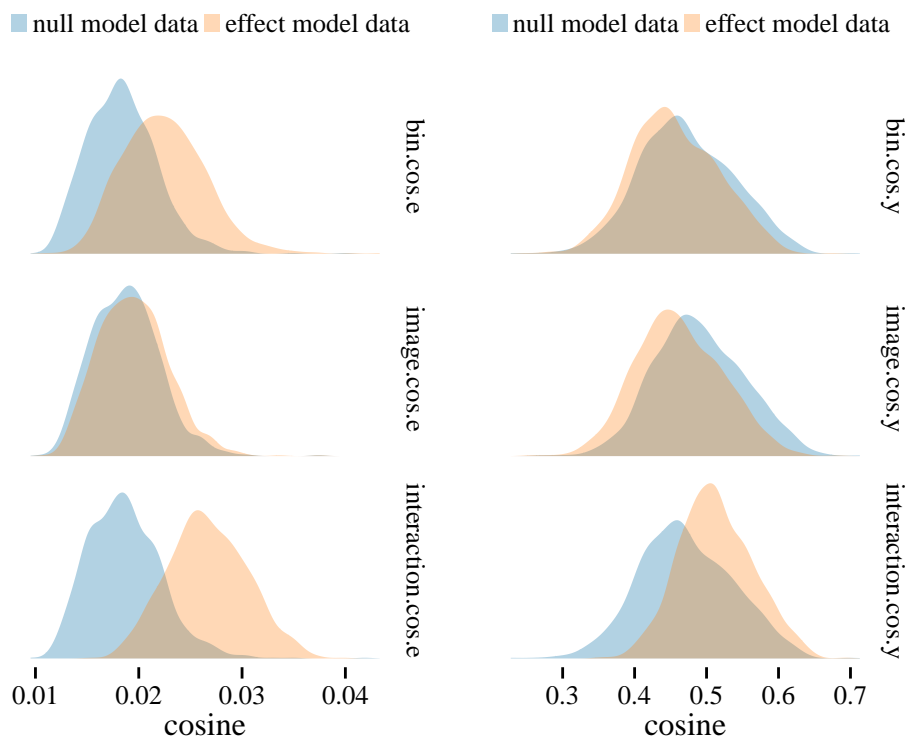


Figure 5.33: Cosine diagnostics of SSANOVA models the in second set of simulations (the y-axes of the left and right panels are different, but this is only to show the overlapping data in both columns)

is also near textbook levels. Therefore, a negative SSANOVA finding may not be a true negative, but a positive SSANOVA finding (keeping in mind the definition of ‘positive’ as above) can be accepted with confidence.

5.2.4.5 Summary

The discussion and evaluation in this section can be summarized in three points.

Firstly, analyzing the percentages of gazes or fixations for different areas of interest by averaging the percentages for each area across a time window longer than, say, 100 ms is appropriate only when there are clear a priori reasons for the length of this time window. Many eyetracking studies that analyze their data this way do not seem to have any such reasons, which casts doubt over their results. In the present study, no single well-defined time window appears to be interesting a priori: participants’ eye gaze could be attracted to one of the three ‘choice’ images more than to the other two at the start of the trial, or as soon as they hear the explicit object in the instruction, or after syntactic processing of the entire instruction sentence has finished, for example. Therefore, to be able to find any significant difference in gaze percentages without the risks of seeing spurious effects or not seeing true effects, the gaze percentages in the present study will not be analyzed by averaging across longer time windows.

Secondly, GCA and SSANOVA, while being based on quite different statistical methods, are both appropriate tools to model percentages of gazes or fixations for different areas of interest. It is important to choose the size of the time bins that the data are combined into well, but at bin sizes smaller than 100 ms (half of what is generally assumed to be necessary to plan a single eye movement saccade), different sizes are unlikely to make a difference in the results and thus would not hide any interesting pattern in the data or make spurious patterns appear. The choice of 50 ms in the present study might be smaller than strictly necessary, since it combines at most 3 consecutive samples of data (collected at 60 Hz) and thus might not do as much averaging as is justified. However, too small bins will not obscure patterns in the data and are therefore unproblematic.

Thirdly, SSANOVA appears to be more conservative than GCA, not rejecting the null hypothesis of no difference between percentage curves when in some cases where there is one.

For these reasons, the eyetracking data collected in experiment 2 is analyzed using SSANOVA.

5.2.5 Regression modelling

Much of the data in this thesis will be analyzed using linear and generalized linear regression models. This section provides a basic practical introduction to this method⁷³ and discusses the issues of predictor significance and overfitting, and how they were addressed in the models presented in this thesis. For readers who are familiar with regression modelling, the following sentences summarize the approach taken in this thesis: regression modelling is used as hypothesis testing, meaning all variables that are expected to have an effect are included and no model reduction or selection takes place. The standard normal Wald assumption is made, and variables with $p < 0.05$ are deemed to be significant. Overfitting is evaluated using ten-fold cross-validation to yield M_{10} , the mean of the estimates of the mean of the squared prediction errors. Random effects were not included for practical reasons.

5.2.5.1 Numeric independent variable

Regression analysis is a method of determining whether one variable in a dataset has an overall effect on another variable. For example, a linguist might be interested in whether words that are used more frequently also share other features. One possible feature might be the number of other words that share the same morphological base: *table*, for example, has many morphological ‘family members’ (*tablespoon*, *coffee table*, the verb *table*, and so forth), while *desk* has far fewer. The etymology dataset provided as part of the R package `languageR` (Baayen 2013) gives this family size and the frequency (in writing) for 285 Dutch verbs. To find out whether words with larger families are used more frequently, we can fit a linear regression model with frequency as the dependent variable and family size as the independent variable. This model is represented by the dashed line in Fig. 5.34 and in more detail in Table 5.4.

variable	parameter estimate	standard error	t	p
(Intercept)	5.45	0.28	19.28	< 0.01
family size	0.78	0.08	9.43	< 0.01

Table 5.4: Coefficients of regression model for writing corpus frequency on morphological family size

The ‘intercept’ in Table 5.4 represents this model’s fitted or predicted value for a verb with a family of size 0 (no other words derived from the same base). The dashed line in Fig. 5.34 intercepts the vertical axis (where the variable on the horizontal axis, family size in this case, is 0) at 5.45, which is the the estimate for the intercept in Table 5.4. The next value along the

⁷³I will not discuss the mathematical model-fitting methods or their computational implementation here—see Freedman 2009 or Yan and Su 2009 for the former and Baayen 2008 or Gries 2009 for the latter.

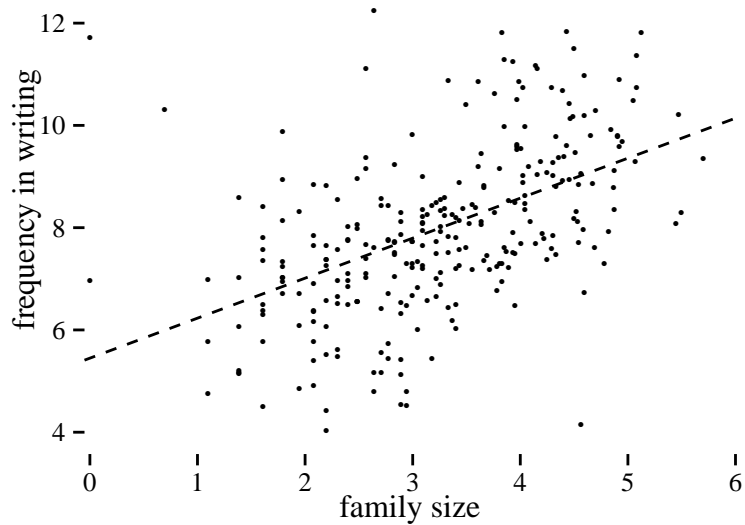


Figure 5.34: Corpus frequency of Dutch verbs by their morphological family size (dots) and linear regression model thereof (dashed line)

same row, 0.28, is the standard error of this estimate, which is a measure of the certainty of the estimate. This measure can be used to calculate the likelihood of a data sample yielding the given estimate under certain assumptions: the Wald test divides the difference between the estimate and the expected value by the standard error of the estimate. In a typical hypothesis-testing approach, the expected value would be the value according to the null hypothesis, often 0. This means the Wald test statistic for this intercept estimate is simply $\frac{5.45-0}{0.28} = 19.5$, and this value will be shown in regression model tables throughout this thesis as t or z (depending on whether the standard error of the estimate was itself estimated, as is the case in this example, or was based on an a priori assumption). Note that the figures in Table 5.4 were rounded to two decimal places **after** all of them had been calculated. The un-rounded values for the estimate and the standard error have more decimal places (5.4481 . . . and 0.2826 . . .), and thus give the slightly different Wald test statistic of 19.28 shown in the table. It is commonly assumed (and I will assume the same here) that the values of this statistic for regression model parameter estimates are approximately normally distributed, with mean 0 and standard deviation 1. Under this assumption, the likelihood of a Wald test statistic is given by the cumulative distribution function of this assumed distribution: in a standard normal distribution (mean 0 and standard deviation 1), a value of 19.28 is extremely rare. The fourth and final value in the Intercept row of Table 5.4 quantifies this rarity: under the null hypothesis and ancillary assumptions described above, we would expect the estimate of a model based on a randomly drawn dataset to be close to 0. Since randomly drawn datasets can be expected to show some random variation, and since we have to take into account the standard error of the estimates, we can allow for this variation and error by

quantifying “close to 0” to mean the Wald test statistic should fall within a certain range of 0 if the null hypothesis were true. Assuming a standard normal distribution for the Wald test statistic is useful for this, since it allows us to use a significance level. I will go with common practice here and use $\alpha = 0.05$. This means that any estimate with a Wald test statistic that is either in the smallest 2.5% or in the largest 2.5% of a standard normal distribution will be deemed to be so unlikely to arise by chance if the null hypothesis were true that the null hypothesis (of the corresponding estimate being 0) can be rejected. Some results in this thesis are reported as significant with a p -value of 0.05. In those cases, the calculated p -value was between 0.045 and 0.05 and was rounded to the latter for clarity and consistency (as statistics are rounded to two decimal places wherever possible in this thesis). Under the assumption of normally-distributed Wald statistics, significant results at $\alpha = 0.05$ will have z or t either less than -1.96 or more than $+1.96$, and these Wald statistics will be given with all regression models. The fourth column in Table 5.4 gives the probability of the Wald test statistic given these terms. For 19.28, this probability is vanishingly small (less than 0.00001, though I will show such values as “ < 0.01 ” here in order to have consistent rounding practices). In other words, if the frequency of verbs with family size 0 is really 0, and the corresponding Wald statistic is normally distributed, the present dataset would be extremely unlikely to emerge.

The second row in Table 5.4 shows the same values for the effect of the family size. The model estimates that for every increase of 1 in the family size variable, the frequency goes up by 0.78. As discussed above, the model estimates that verbs with family size 0 have a frequency value of 5.45⁷⁴; this estimate for family size means that the model estimates verbs with family size 1 to have a frequency value of $5.45 + (1 \times 0.78) = 6.23$, verbs with family size 3.8 to have a frequency value of $5.45 + (3.8 \times 0.78) = 8.41$, and so on. In other words, this estimate is the slope of the dashed line in Fig. 5.34. This estimate has a standard error of 0.08, which means the Wald test statistic for this estimate (with the null hypothesis being that family size has no effect on frequency and that the parameter is thus 0) is $\frac{0.78-0}{0.08} = 9.75$ (again slightly different from the table due to rounding, but the difference is immaterial here). Assuming that this statistic is normally distributed, this value is very unlikely: $p < 0.01$. Therefore, given the etymology dataset, I would reject the null hypothesis that family size and frequency are independent of each other and argue that family size has an effect on frequency. As the estimate (0.78) is a positive number, this effect is mathematically positive in the mathematical formulation used here: higher family size values go together with higher frequency values (or lower family size goes together with lower frequency).

⁷⁴There are two verbs with a family size of 0 in the etymology data, namely *vergen* and *moeten*. It is apparent from Fig. 5.34 that both of them are rather more frequent in the data than the model estimates, since the two dots on the far left of that figure (which represent those two verbs) are far higher than the regression line. Regression modelling as used here is a tool of hypothesis testing is not perfectly predictive, but it does show whether there is an effect overall.

Linear regression in effect is a way of finding the closest line to all the points in a scatterplot like Fig. 5.34. It does not respect direction: the discussion here assumes that morphological family size is a more basic feature of verbs and has some causal effect on how often speakers (or rather writers) choose to use these verbs. This is an assumption outside of the regression model: a reverse model, of family size being ‘predicted’ by frequency, would find a similar and significant effect. The decision of how to conceptualize and interpret a model is crucial. In many applications, however, an experimental outcome is used as the dependent variable and various conditions of the experimental stimuli are used as independent variables. It is reasonable to assume that the causality of effects, if the model finds any, does indeed go that way.

The etymology data also provides the ratio of the frequency in writing to the frequency in speech for each verb. This could be useful in testing, for example, if the frequency extracted from a written corpus is affected by the fact that it was a written and not a spoken corpus: intuitively, some words may well be more frequent in writing than in speech, particularly in the genres of writing that are commonly used for corpora (*fracas*, for example, may be more frequent in newspaper writing than in other genres and modes). If there were such an effect, we might reasonably expect that frequency in a written corpus and the ratio of written to spoken frequency are not independent: if words with a higher ratio of written to spoken frequency tend to have a higher written frequency overall, written frequency would be driven in part by this ratio and thus by the mode or genre. A linear regression model shows that this is not the case for the Dutch verbs in the etymology data. Although the parameter estimate for the effect of the written/spoken frequency ratio is not 0 (but rather 0.06, which is the slope of the dashed line in Fig. 5.35), this estimate is small relative to the (estimated) standard error of this estimate. The Wald test statistic (with the null hypothesis being that there is no effect here) is $\frac{(0.06-0)}{0.07} = 0.857$ (as above, the *t*-value given in the table is different due to rounding; as above, this mathematical difference makes no difference for the interpretation of the model). Assuming a standard normal distribution centered on 0, this value is not very unlikely: its probability *p* is 0.41. As this is above the significance level established above, there is no evidence here that the written/spoken frequency ratio has an effect on the written frequency. In other words, the written corpus frequency does not overvalue verbs that are more frequent in writing than they are in speech.

variable	parameter estimate	standard error	<i>t</i>	<i>p</i>
(Intercept)	7.96	0.10	77.97	< 0.01
written/spoken ratio	0.06	0.07	0.82	0.41

Table 5.5: Coefficients of regression model for writing corpus frequency on ratio of writing to speech corpus frequencies

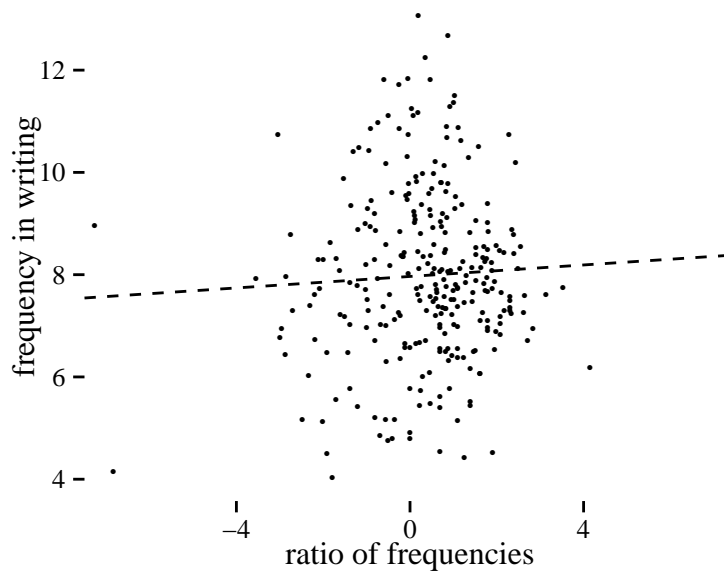


Figure 5.35: Frequency in writing of Dutch verbs by the ratio of written to spoken frequencies (dots) and linear regression model thereof (dashed line)

5.2.5.2 Categorical independent variable

Regression modelling can also be used when the independent variable is not a number, but a variable that can take one of a specific set of values (a categorical variable or ‘factor’). The etymology data contains information regarding the inflectional paradigm for each verb. The inflectional paradigm of a verb can be either regular or irregular, so those two options are the two values for this factor (its ‘levels’). The frequencies of verbs may differ between the two: verbs with irregular inflectional paradigms, for example, may be more frequent. Fig. 5.36 suggests this, at least: the two boxes represent the distributions frequencies for verbs with regular and irregular inflection. The thick black line inside each box shows the median frequency for the respective group of verbs. The lower and upper ends of the boxes represent the first and third quartiles: one quarter of the verbs with regular inflections have frequency values of less than 7.1 and thus fall below the box; one quarter of the verbs with regular inflections have frequency values of more than 8.5 and thus fall above box; and it follows that the remaining two quarters (or one half) have frequency values that fall inside the box. The vertical lines beyond the ends of the boxes extend to the most extreme value within 1.5 times the difference between the first and third quartile (the ‘interquartile range’, or the height of the box)—the end-point of the vertical line extending from the top of each box is the largest value in the data that is less than or equal to the third quartile value plus 1.5 times the interquartile range, and the end-point of the line extending from the bottom is the smallest value that is equal to or larger than the first quartile value minus 1.5 times the interquartile range. (This explains why these lines are not of equal length in plots of this type: for example, the largest quarter of data points (above the box) may

be very clustered, so that the largest value is much less than 1.5 interquartile ranges above the third quartile and the upper line ending at it is therefore very short.) Verbs with values more extreme than that (lower than first quartile minus 1.5 times interquartile range, or higher than third quartile plus 1.5 times interquartile range) are shown individually, as grey dots. The grey X shows the mean value for each group; here, these are almost exactly the same as the medians. It is thus apparent that the irregular verbs, on average, have a higher frequency than the regular verbs. However, the difference is not enormous, and there is much variation even within each group of verbs.

The relationship between this regularity factor and the written corpus frequency can be tested statistically with a linear regression model. Table 5.6 shows the coefficients for this model, and the orange⁷⁵ line in Fig. 5.36 is a graphical representation of the model. It is obvious that the orange line ends exactly in the grey X's that show the mean values. With just one categorical independent variable with two levels, this is precisely what linear regression does: the fitted or predicted values of the dependent variable for either level are the mean values of the dependent variable within that level. The parameter estimates for categorical independent variables have a different meaning than the estimates for numerical independent variables: for example, with the numerical independent variable of the ratio of writing/speech corpus frequencies in the model in Table 5.5 discussed above, the intercept parameter is the fitted value for when this independent variable is 0. A categorical variable cannot be 0 by definition—it can only take one of a set of values (for example 'regular' or 'irregular'). With binary variables (like regularity here), a common practice is to choose one of these values as the 'reference level' and construct a numerical variable based on that, which takes the value 0 for all data points that have the reference level value and 1 for the data points that have the other level. The intercept thus becomes the fitted value for the reference level. In the present model, the reference level is 'regular', and the intercept value of 7.79 is the mean frequency value for regular verbs. For these verbs, the constructed numerical variable is 0, and the coefficient for this variable is thus multiplied by 0 and thus does not change the predicted value. For the non-reference level, however, the numerical variable is 1. This means the fitted values for irregular verbs are calculated by taking the intercept and adding the coefficient for this variable. With the present model, that means the fitted value for all irregular verbs is $7.79 + (1 \times 0.4) = 8.19$, which is the mean of the frequencies of all the irregular verbs.

Happily, the standard error, Wald test, and probability values work much the same way with categorical independent variables as they did with numerical ones (discussed above): the difference between the estimate and the expected value (if the expected value is 0, this is just the

⁷⁵There is no meaning attached to the color of the line here. It was chosen merely to make the line more visible in a figure with several other lines.

estimate) is divided by the standard error of the estimate, and the resulting statistic is located in a standard normal distribution to find a probability value. If this p is less than 0.05, I assume that the difference expressed by the variable estimate is significant, meaning the variable in question has a significant effect. In the model of frequency by regularity, that is the case ($p = 0.04$): irregular verbs are significantly more frequent than regular ones overall.

variable	parameter estimate	standard error	t	p
(Intercept)	7.79	0.14	55.90	< 0.01
irregular	0.40	0.20	2.03	0.04

Table 5.6: Coefficients of regression model for writing corpus frequency on regularity of inflectional paradigm

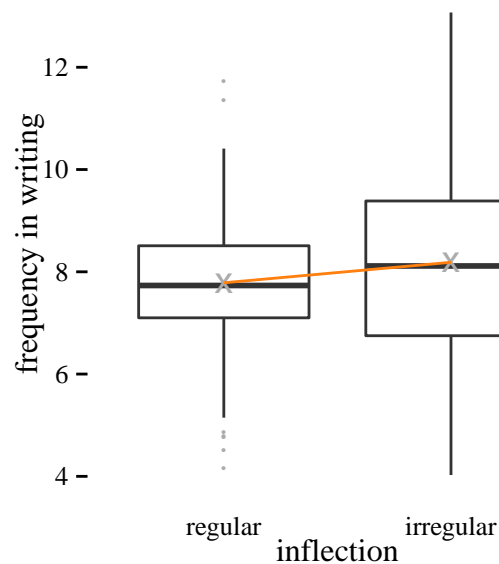


Figure 5.36: Frequency in writing of Dutch verbs by regularity of inflectional paradigm (boxes) and linear regression model thereof (orange line)

5.2.5.3 Multiple variables and their interactions

Crucially, regression models can take more than one independent variable. These multiple regression models estimate a parameter for each independent variable (much in the same way as the simple regression models discussed above) in the presence of the other independent variable. The meaning of one independent variable's parameter is the change in the dependent variable that is correlated with an increase of 1 in this independent variable, keeping all others constant. Multiple regression also allows for interaction effects, which model a change that goes together when two independent variables change at the same time. Interactions typically mean deviations

from the overall tendencies of either variable. For example, a model of general health might show an interaction between quality of diet and amount of exercise: a better diet leads to better health, more exercise leads to better health, but a better diet and more exercise at the same time mean an even greater increase in health than the simple summation of the two individual effects. Interactions can also reverse the trend of the individual effects: a model of election data may show an effect of candidate age (with younger, less experienced, less known candidates winning fewer votes) and an effect of demographics of the district (with districts with a younger, less engaged population voting less overall), but also an interaction effect of the two: young candidates running in 'younger' districts may win **more** votes (because their platform or public profile engage the voters more).

Continuing with examples from the etymology data, it is conceivable that the effect of family size interacts with the effect of regularity: for example, the irregular English verb *see* is probably more frequent than the regular *believe*, even though they have similar family sizes. A regression model that uses these two effects side-by-side, without an interaction, would not reveal this effect. Interaction effects allow the model to account for changes in two independent variables **together** having an effect on the dependent variable. Fig. 5.37 shows the frequency of the 285 Dutch verbs in a corpus of writing by their family size (just as in Fig. 5.34), colored by whether they have a regular inflectional paradigm (blue dots) or an irregular one (orange dots). The regression model using family size, regularity of inflection, and their interaction as predictors for frequency is summarized in Table 5.7. The first three rows of that table can be interpreted much as above: the intercept now is the fitted frequency value for when all variables are 0 or at their reference level. This means that this model predicts a frequency value of 6.1 for regular verbs with no other words derived from the same base. The effects for irregular inflection and family size by themselves (middle rows of Table 5.7) can be added to this to calculate the predicted values for the cases where only one of those variables changes: adding the effect of irregular inflection (−1.61) to the intercept value gives 4.49 as the predicted value for an irregular verb with family size 0; adding the effect of family size (multiplied by the family size) gives the predicted value for a regular verb with that family size, like $6.1 + (3 \times 0.53) = 7.69$ for a verb with family size 3. The standard error, Wald test statistic, and *p* values in Table 5.7 are calculated as above; with 0.05 as the significance level for *p*, all effects are significant.

When both variables change from the intercept/baseline, however, the interaction effect (“irregular : family size”, the bottom row of Table 5.7) comes into play as well: we add both effects **and the interaction effect** to the intercept value. The model thus predicts that an irregular verb with family size 3 has a frequency value of $6.1 + (-1.61) + (3 \times 0.53) + (3 \times 0.58) = 7.82$. As both the main effect parameter for family size and the interaction parameter are multiplied by the same value (a verb’s family size), they can be summed for simplicity into 1.11. This

is how much the frequency value is changed by a change of 1 in the family size if the verb is irregular—in other words, the slope of the regression line for irregular verbs. The interaction effect does not affect the prediction for regular verbs, so the change and slope there are simply the parameter for family size, 0.53. The two lines in Fig. 5.37 show the model fits for regular (blue line) and irregular (orange line) verbs. It is apparent from these values and slopes that the model suggests the effect of family size on frequency is stronger for irregular verbs than it is for regular verbs.

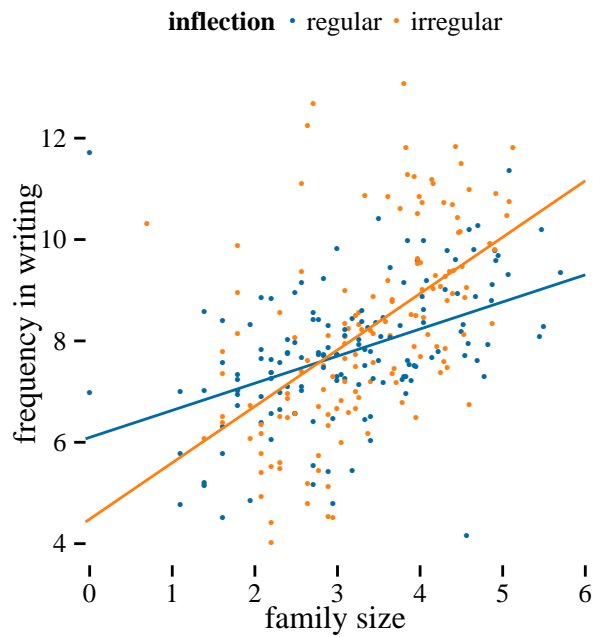
variable	parameter estimate	standard error	<i>t</i>	<i>p</i>
(Intercept)	6.10	0.36	17.14	< 0.01
irregular inflection	-1.61	0.57	-2.85	< 0.01
family size	0.53	0.11	5.05	< 0.01
irregular : family size	0.58	0.17	3.49	< 0.01

Table 5.7: Coefficients of regression model for writing corpus frequency on regularity of inflectional paradigm and family size

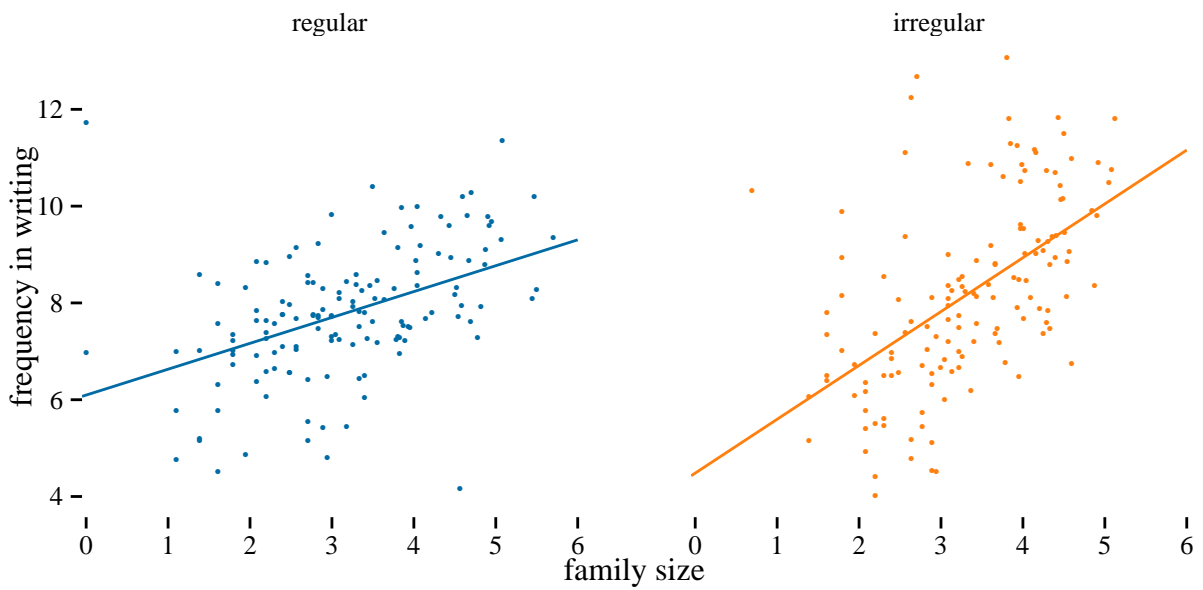
This concludes the illustration of regression modelling in this section. In practice, models with many more independent variables can be fit to test which (if any) of them have a significant effect on the dependent variable in the presence of all the others. This is the approach I will take in this thesis: all (independent) variables that are present in the data and that I have a reason to expect to have an effect on the dependent variable in question will be included in a given model to test whether they do have an effect. In other words, no model reduction or selection will occur. The *p*-values of their parameter estimates under the normality assumption for the Wald statistic will be used as the indicator for whether an independent variable has a significant effect on that dependent variable.

5.2.5.4 Cross-validation

Regression models with more independent variables can often approximate the dependent variable in the data more closely than models with fewer independent variables. This, however, is a blessing and a curse: the closest fit to a given set of data is the one that describes each data point exactly. In practical applications, this phenomenon (called ‘overfitting’) is generally undesirable because the data is assumed to be noisy—to contain some variation that is not due to the independent variables in question, but rather to chance. For example, consider school or university test grades: intuitively, grades in a test can be modelled as being largely determined by time spent studying for that test, previous test performance, and general intelligence. However, an intelligent A student who spent weeks in the library may on occasion do poorly on a given test because they did not sleep well the night before or the on-campus café was out of their favorite



(a) All verbs



(b) Regular and irregular verbs separately (for clarity)

Figure 5.37: Frequency in writing of Dutch verbs by regularity of inflectional paradigm and family size (dots) and regression models thereof (lines)

energy drink. It would be next to impossible to collect all this data, and its effects may not be of interest. Fitting this data point exactly would mean spuriously assuming this noise is also caused by the independent variables, which means their effects may be wildly mis-estimated: an overfit model of this test data could suggest that general intelligence has no effect, or even a negative effect, on test performance.

The method I will use here to check whether a model does fit the noise inherent in the data too closely is ten-fold cross-validation. The dataset used to fit a model is randomly split into ten distinct, non-overlapping sub-sets ('folds') of as close to equal size as possible. For each fold, the model parameters are re-estimated using just the nine other folds, and the model with those re-estimated parameters then makes predictions for all of the data points in the one remaining fold. As the parameters were estimated without those data points, it cannot be overfitted to them. For each data point, the difference between the prediction and the actual data point value is calculated and squared. The mean of all these squared errors is calculated for each fold, and the mean of the ten resulting means of squared errors is used as a measure of overfitting for the model as a whole. If the model was severely overfitted, the parameter estimates will vary greatly between the original model and at least some of the re-estimations. Therefore, at least some of the re-estimated parameters will fit the data in the one fold that the respective re-estimation process did not use less closely than the original model does. In other words, the mean of the ten cross-validation mean squared errors will be large relative to the original model's mean squared error (the mean of the squared differences between that model's predicted values and the data values). If the original model was not overfitted much, on the other hand, the parameters will change little between each re-estimations. The predictions on the un-used fold will not be much worse than the fitted values of the model were as a whole, and the mean of the cross-validation mean squared errors will not be much larger than the mean squared error of the original model.

As the data is split into ten folds **randomly**, it is good practice to repeat this cross-validation process with different random splits.⁷⁶ Each repetition's mean of mean squared errors will be slightly different, of course, but the mean of all these means of mean squared errors remains a good estimate of overfitting. As 'mean of ten-fold cross-validation means of mean squared errors' is an unwieldy phrase, I will report this value as ${}_xM_{10}$, with the number of repetitions of the ten-fold error processes in the place of x . MSE will be used to mean a (full) model's mean squared error.⁷⁷ For example, the model of frequency in writing by morphological family

⁷⁶I use the number 2332 as the initial random seed for each cross-validation process here.

⁷⁷For models predicting a numeric dependent variable (reaction time, for example), the mean squared error and ${}_xM_{10}$ are calculated from the actual and fitted values of that numeric variable. For generalized models, which predict categorical variables by transforming them into a binary (0 or 1) variable, the fitted values are probabilities of the outcome coded as 1.

size, regularity of inflection, and their interaction (Table 5.7 and Fig. 5.37) has $MSE = 2.02$ and ${}_{50}M_{10} = 2.10$. The mean of squared errors in cross-validation is not much larger than the mean of squared errors in the model, which confirms the impression in Fig. 5.37 that this model does not suffer from overfitting.

5.2.5.5 A word on mixed-effects modelling

This section so far has described fixed-effects regression modelling. However, mixed-effects models (Bates et al. 2014) are increasing in popularity. They ‘mix’ fixed effects with random effects: variables that are not of interest can also affect the dependent variable as well as how the variables of interest (the fixed effects) affect the dependent variable. In linguistics, it is reasonable to assume that a dependent variable (word frequency, for example) depends not only on certain fixed effects (regularity of the inflectional paradigm, for example), but also in part on the particular speaker or word. Some speakers may use many more words (types), which would lower the frequency of any one word. A corpus based in part on cooking recipes would probably slightly overrepresent the frequency of *roux* and underrepresent *dog*. Finally, the fixed effects may differ for different speakers or words (some speakers may be more sensitive to the regularity of the inflectional paradigm, or their data may suggest this by random chance). These effects are assumed to follow a random distribution—some speakers will show lower word frequencies, but others will show higher ones. Mixed-effects models allow researchers to control for these random effects and not be misled by random patterns in the data.

Using mixed-effects modelling properly, however, can be computationally expensive (Barr et al. 2013, Bates et al. to appear), and there are practical problems. I fit mixed-effects models to some of the data in this thesis, and in many cases these models did not converge, meaning the fitting algorithm did not arrive at final estimates for the parameters. The parameter estimates that were reported were very similar to those in the corresponding fixed-effects models. Convergence was attained by sequentially excluding effects (both fixed and random), but the converging model formulae were very impoverished (often modelling just two or three of the effects of interest) and thus unable to answer the research questions of this thesis. I am not confident in drawing conclusions from non-converged mixed-effects models. Therefore, this thesis reports the fixed-effects models only to make it as obvious as possible to the reader that random effects were not (successfully) modelled.

5.2.6 The multiple comparisons problem

This section briefly discusses the statistical problem of multiple comparisons and introduces the Holm-Bonferroni correction that is used in this thesis to address that problem.

Many statistical tests are tests of data against a null hypothesis: if there was no actual effect, how likely is it that we would have collected this dataset? Two entirely uncorrelated variables may appear to be correlated in a particular sample dataset by random chance. The strengths of these spurious correlations (or other effect strengths) are assumed to form a certain distribution, often the standard normal distribution around 0—in other words, datasets that show weak spurious effects are more likely than datasets with strong spurious effects. Thus, a strong effect is more likely to be due to an actual underlying effect than to random chance. The effect size is a useful metric in determining the likelihood that a particular finding is true. Some weaker effects are also true, however, and statistical tests can be used to operationalize the likelihood that a particular effect is spurious (or, conversely, that it is a true effect). This is one interpretation of p -values: if the null hypothesis (no effect) was true, how likely would a set of data showing this particular effect be? A result that is sufficiently unlikely given the null hypothesis is deemed to be significant evidence for the alternative hypothesis. 5% ($= 0.05$) is the commonly used value for the significance level (α): if a particular test results in a p -value less than 0.05, the data being tested are deemed to show significant evidence against the null hypothesis. Linear regression modelling as discussed in Section 5.2.5.1 above follows these same principles: larger parameter estimates directly lead to smaller ('more significant') p -values.

This understanding of statistical testing rests on the assumption that there is one randomly collected data sample and one hypothesis (effect) to be tested against it: given one effect strength, how common is this effect strength in the random distribution of effect strengths centered around 0? If it is rather uncommon (unlikely) there, we can assume with some certainty that we did not get it from that distribution, and that the null hypothesis is therefore false. However, many (if not most) researchers test more than one hypothesis at a time. This means that they 'draw' several effect strengths at the same time, which increases the likelihood of finding one that is less than 5% likely to appear by chance. If I think of a number between 1 and 100 (inclusive) and give you one chance to guess it, you are unlikely to guess right by chance, and a correct guess will be surprising ($p = 0.01$). If I give you ten guesses, your chances at a correct guess are much larger ($p = 0.1$, assuming you never guess the same number twice). While there are some relevant differences between this example and hypothesis testing, the principle is the same: more guesses/hypotheses increases the chance of a correct/significant result.

Statistical tests therefore need to account for the number of hypotheses tested simultaneously. Austin et al. (2006) illustrate this problem with a large dataset of hospital diagnoses in Canada: they compared the birth dates of patients with the 223 most common diagnoses using Fisher's exact test. 72 of these diagnoses were 'significantly' more frequent for one astrological sign than for the other eleven under a naive significance threshold of $\alpha = 0.05$ for each test—for example, Canadians born between 23 August and 22 September (under the sign of Virgo) were more likely

than other zodiac signs to experience excessive vomiting in pregnancy ($p = 0.04$), and those born between 19 February and 20 March (Pisces) were more likely to suffer heart failure ($p < 0.01$). Of course, these ‘significantly’ higher risks are entirely “spurious”, in the words of the title of Austin et al. (2006)’s article. They highlight that appropriate corrections show these associations between astrological signs and certain diagnoses to be statistically insignificant.

Having shown why it is necessary to correct for multiple hypotheses (or comparisons) above,⁷⁸ the rest of this section will describe how it can be done. One simple way is to divide the desired significance level α (often 0.05) by the number of comparisons m and take the result as the new significance level that p -values need to be lower than in order to be deemed significant. Austin et al. (2006) tested 223 diagnoses for differences between astrological signs, meaning they performed $m = 223$ comparisons. The Bonferroni-corrected significance level for their study is $\frac{0.05}{223} \approx 0.0002$. All p -values reported in Austin et al. (2006:966) are larger than that, which (rightly) suggests all the associations between diagnoses and astrological signs are insignificant and only appear in the data due to chance.

The Bonferroni correction ($\alpha_{corrected} = \frac{\alpha}{m}$) has been criticized for being too restrictive: if all hypotheses are compared against the same corrected significance level, then only very strong effects will be deemed significant and weaker true effects will falsely be deemed to be insignificant. Moreover, it is intuitively inconsistent, in that effects that appear to be the same can be deemed significant or insignificant depending on apparently unrelated factors. Imagine two researchers shared a dataset, but tested it independently. The first researcher investigates ten hypotheses. Nine of them are obviously not confirmed (all $p > 0.1$), but one of them looks promising with $p = 0.0046137$. The Bonferroni correction provides $\alpha_{corrected} = \frac{0.05}{10} = 0.005$, meaning this one strong effect is deemed to be significant. The other researcher investigates the same ten hypotheses, but is also interested in confirming a well-established effect with this data and therefore adds an eleventh hypothesis. This eleventh hypothesis is confirmed with $p = 0.000001$, and the other ten obviously receive the same p -values as they did with the first researcher. However, because this second researcher tested $m = 11$ hypothesis, their $\alpha_{corrected}$ is $\frac{0.05}{11} = 0.0045$. The additional eleventh hypothesis is still deemed significant, of course. The other strong hypothesis (with $p = 0.0046137$) is now insignificant, even though it was significant for the first researcher using the same data and the same tests.

Holm (1979) offers a simple procedure that is based on the Bonferroni correction and is just as powerful, but less restrictive and less inconsistent. All p -values are calculated and ordered

⁷⁸Gelman et al. (2012) argue that the multiple comparisons problem can be ignored in social science, where the null hypothesis of no true underlying correlation is supposedly “less likely” (Gelman et al. 2012:209). However, this thesis does not test previously established effects, but is aimed at determining whether or not the effects of interest really exist. I do not see how this assumption that the null hypothesis is irrelevant can be maintained when pursuing this more exploratory aim and will therefore correct for multiple comparisons where necessary.

from smallest to largest. For the smallest p -value, the normal Bonferroni correction is applied: the null hypothesis is rejected for this effect if and only if $p < \frac{\alpha}{m}$. If this is the case, the procedure moves on to the next smallest p -value and crucially reduces m by 1—the ‘strongest’ hypothesis has been confirmed already, so it should not affect this and future calculations. Thus, the second-smallest p -value is compared against $\alpha_2 = \frac{\alpha}{m-1}$. If it is smaller than this threshold, the procedure moves on to the third-smallest p -value and checks whether it is smaller than $\alpha_2 = \frac{\alpha}{m-2}$, and so forth. As soon as one p -value is larger than the threshold (and thus deemed to be insignificant), the procedure stops and deems this and all larger p -values to be insignificant. Holm (1979) provides more details. Applying this Holm-Bonferroni procedure to the above example with two researchers sharing the same dataset, the second researcher would test their smallest p -value (0.000001) against $\alpha_1 = \frac{0.05}{11} = 0.0045$, find it significant, move on to the second-smallest p -value (0.0046137), test it against $\alpha_2 = \frac{0.05}{11-1} = 0.005$, find it significant, and so forth. Crucially, the first researcher would test $p = 0.0046137$ against $\alpha_1 = \frac{0.05}{10} = 0.005$ —in other words, both researchers would arrive at the same result.

This Holm-Bonferroni method is not without its shortcomings either (Chen 2014:160–166 discusses them and also describes other methods), but its computational simplicity and statistical usefulness make it a reasonable choice.

For ease of presentation, a set of p -values can be adjusted by simply multiplying each p -value with the denominator of the corresponding α_i (where i is the number of the p -value in the set under consideration, counting from smallest to largest, and m is the number of comparisons being made, meaning the number of p -values in that set):

$$(5.10) \quad \alpha_i = \frac{\alpha}{m-(i+1)}$$

$$(5.11) \quad p_{adjusted} = p \times (m - (i + 1))$$

In the above example, this yields $0.000001 \times 11 = 0.000011$ and $0.0046137 \times (11 - 1) = 0.046137$. These adjusted p -values can then be judged against **any** desired significance level α without further correction. In this example, both of these adjusted p -values are significant at $\alpha = 0.05$, but only the first one is significant at a stricter $\alpha = 0.01$. In this thesis, I will report p -values adjusted by this Holm-Bonferroni method where appropriate and provide the number of comparisons m .

5.3 Results

In this second experiment, participants were presented with sets of images and stimulus sentences with one object masked by noise. They implicitly made a choice out of three of the images

for that noise-masked object. This section analyzes the time taken to make these choices as well as the choices themselves. This is followed by analyses of participants' touchscreen input, their gaze behavior, and finally the correlation between simultaneous touchscreen input and gazes.

5.3.1 Reaction times

For the results of this experiment, I define reaction time as the time between the end of the instruction sentence (from which point images could be moved by dragging) and the point at which a theme image was dragged into the rectangle surrounding a recipient image. This reaction time differed significantly between age groups, as the boxplots in Fig. 5.38 show. A linear regression model⁷⁹ was fit to this data to determine the effects of age group, the trial number in sequence,⁸⁰ and the interaction between these two terms on this reaction time measure. The coefficients of this model are given in Table 5.8. Four-year-olds served as the reference age group, and 0 as the intercept value for trial number in sequence. The significant effect with a negative parameter for the adult age group shows that, compared to four-year-olds, adults reacted significantly quicker (had a lower reaction time value). The main effect of the trial number in sequence is not significant, meaning that the reference group (four-year-olds) did not get significantly quicker (or slower) over the course of the experiment. The interaction between the adult age group and the trial number in sequence has a negative parameter (-0.01), but with a p -value of 0.1, it does not reach significance as defined in Section 5.2.5—the adult participant group got slightly quicker over the course of the experiment, but not significantly so. The main effect associated with the eight-year-old age group and the interaction between this group and the trial number in sequence have to be considered together: the positive and significant effect for the eight-year-olds suggests that, at the intercept level for trial number (which is 0), eight-year-olds are significantly slower than four-year-olds. The data for the first trial in sequence do not support this: four-year-olds' mean reaction time on trial 1 is 7.26 seconds, and eight-year-olds' mean reaction time is 7.06 seconds. This main effect makes more sense when the interaction with trial number is also taken into account: this interaction is negative (-0.03) and significant ($p < 0.01$). This means that the model shows that eight-year-olds did get significantly quicker over the course of the experiment. Regression is vulnerable to overestimating such effects near the ends of the range of one variable (“in the tails”): in capturing the overall downward trend, the values at the start of this experiment (at the intercept level, which is where the main effects for age group are

⁷⁹Mixed-effects regression with random effects of participant was deemed unsuitable for the present study in general because of the small number of participants per age group.

⁸⁰As the sequence of trials within each block was randomized for each participant, different participants saw different trials as their first, second, . . . trials. By “trial number in sequence”, I mean this ordinal number rather than any unique trial (meaning ‘picture combination’) identifier.

estimated) are estimated slightly too high. In the present case, this leads to a significant main effect for the eight-year-old age group. The boxplots of these reaction times in Fig. 5.38 make it obvious that four- and eight-year-olds' reaction times do not differ significantly overall, and that the four-year-olds' reaction times show the most variation.

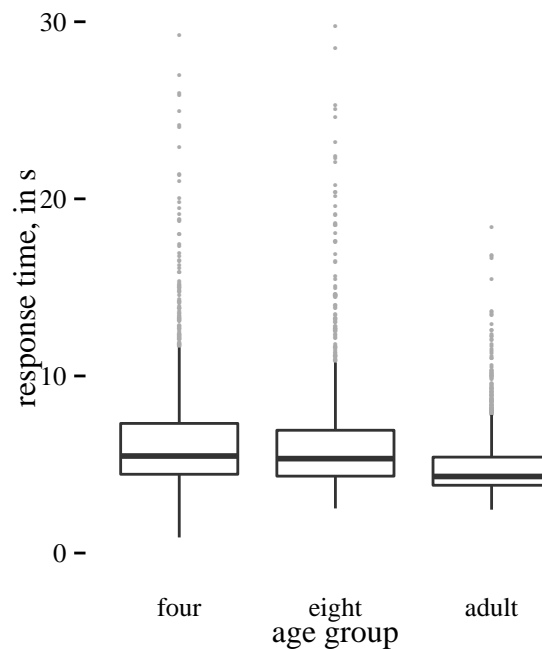


Figure 5.38: Reaction times by age group (lower and upper ends of boxes indicate first and third quartiles, vertical lines extend to the most extreme value within 1.5 interquartile ranges of these quartiles, grey dots are more extreme values; note that 14 reaction times greater than 30 s are not shown)

variable	parameter estimate	standard error	t	p
(Intercept)	6.57	0.24	27.23	< 0.01
trial number in sequence	< 0.01	0.01	0.21	0.83
eight-year-old group	1.02	0.34	2.99	< 0.01
adult group	-1.21	0.33	-3.63	< 0.01
eight-year-olds : trial number	-0.03	0.01	-3.48	< 0.01
adults : trial number	-0.01	0.01	-1.65	0.10

Table 5.8: Coefficients of regression model for reaction time (variables in bold are deemed to have significant effects; $MSE = 18.1$, $_{100}M_{10} = 18.2$)

5.3.2 Choices

In each trial in this experiment, participants made a 'choice': when the noise in the recorded instruction sentence was in the place of the theme object (as in *Now give the _____ to the dogs* or

Now give the dogs the _____), they picked one out of three options for theme objects and dragged it to the recipient (which was explicit in the instruction). When the noise was in the place of the recipient (Now give the _____ the keys or Now give the keys to the _____), participants dragged the explicit theme to one of the three options for recipients. Each of these options matched the explicit object in exactly one of the three features of interest (length in syllable number, binary a priori animacy, and grammatical number), but did not match it in the other two features—in other words, there was one matching choice and two mis-matching choices for each of the three features. This section presents participants’ choices by these features.

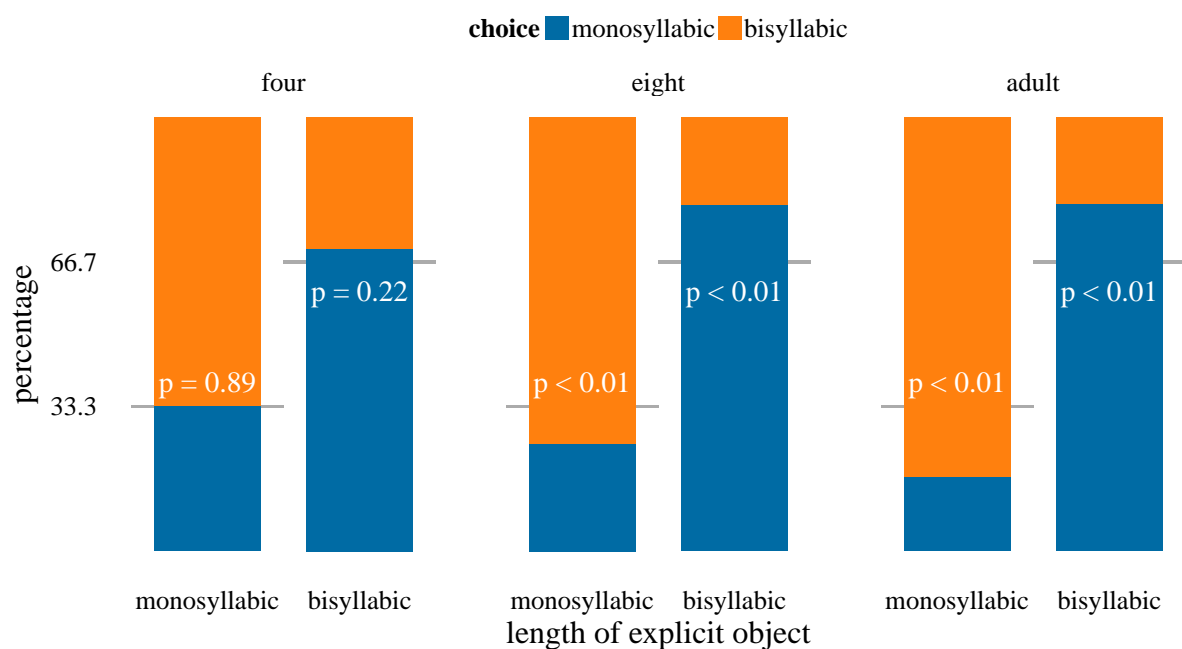


Figure 5.39: Percentage of choices by their length (split by length of the explicit object and by age group), with p -values of χ^2 tests (Holm-Bonferroni-corrected)

Fig. 5.39 compares the percentages of monosyllabic and bisyllabic choices from trials with monosyllabic explicit objects to the same percentages from trials with bisyllabic explicit objects, for each of the three age groups. If the choice was not affected by length—if, in other words, it is random with respect to length—we would expect the proportion of monosyllabic to bisyllabic choices being made to be roughly the same as the proportion of monosyllabic to bisyllabic options being available to choose from. Since there were two bisyllabic options when the explicit object was monosyllabic, and two monosyllabic options when the explicit was bisyllabic, these random percentages would be 66.7% length-mismatching choices and 33.3% length-matching choices. The grey lines with each bar show these expected levels (or, in graphical terms, where the dividing line between the orange and blue segments should be if the random expectation was true). The p -values given inside the bars in Fig. 5.39 result from testing the respective bars against this expectation (using Pearson’s χ^2 goodness-of-fit test on the response counts,

and applying Holm’s sequential Bonferroni procedure for $m = 6$ comparisons to the resulting p -values in order to avoid falsely rejecting the null assumption in multiple comparisons, as described in Section 5.2.6; the same test and correction were also used for the p -values in Figs. 5.40 and 5.41). The four-year-olds’ choices do not differ significantly from these expected values ($p = 0.89$ and 0.22). The eight-year-olds’ and adults’ choices, on the other hand, do: eight-year-olds and adults prefer bisyllabic options when presented with a monosyllabic explicit object and monosyllabic options when presented with a bisyllabic explicit object significantly more than expected by chance (all $p < 0.01$). There were no further differences in this regard between trials with the gap in place of the theme and trials with the gap in place of the recipient in the instruction sentence.

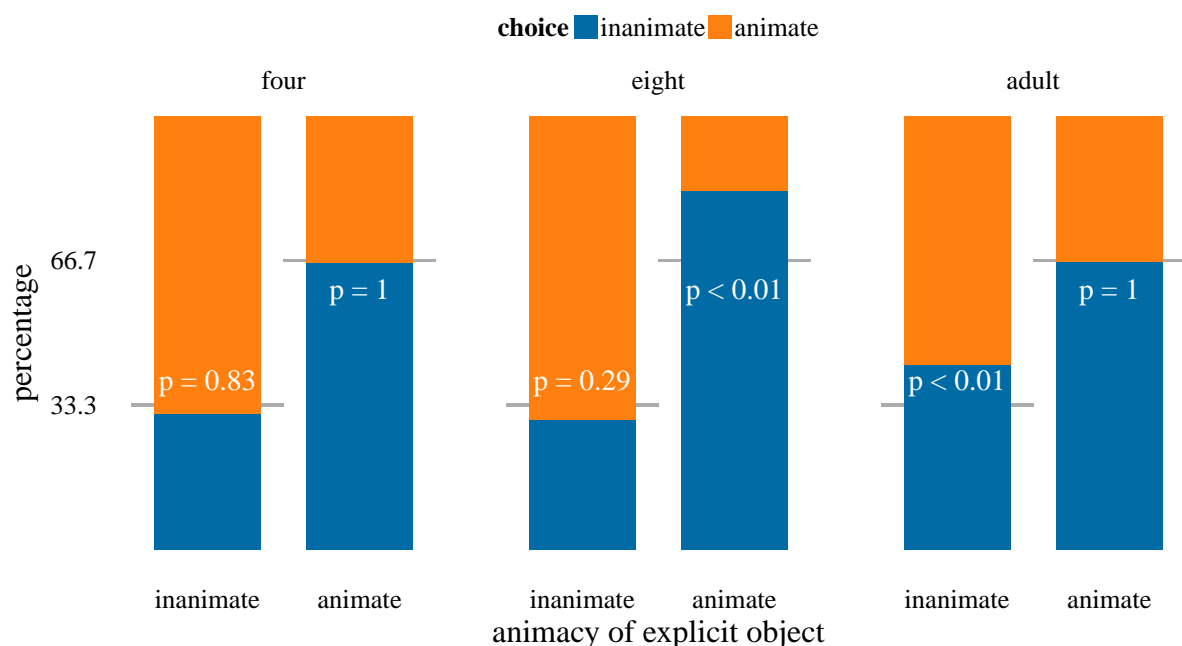


Figure 5.40: Percentage of choices by their animacy (split by animacy of the explicit object and by age group), with p -values of χ^2 tests (Holm-Bonferroni-corrected)

Four-year-olds also were apparently not guided by animacy in their choices, as the two leftmost bars in Fig. 5.40 show.⁸¹ Eight-year-olds were more likely than chance to pick one of the two inanimate options when the explicit object was animate ($p < 0.01$), but apparently did not have a preference for animates or inanimates when the explicit object was inanimate ($p = 0.29$). For adults, the reverse is true: when the explicit object was inanimate, adults chose significantly more inanimate options than expected by chance ($p < 0.01$), but with animate explicit objects,

⁸¹Analyzing the data in this manner requires a categorical split in the animacy variable. A numerical measure derived from the results of experiment 1 (Section 4.2) would therefore be impracticable here. Most categorical splits that could be derived from the same results would simply replicate the a priori split between animals and inanimate objects, as is apparent from Fig. 4.1 (page 69). Therefore, that a priori categorical split is used here.

there was no significant preference. This apparent preference for inanimates is different from the other significant deviations from random chance discussed so far: the latter can all be subsumed under feature-mismatching (**b**isyllabic choices for **m**onosyllabic explicit, and so on), whereas this preference for inanimates is apparent only when the explicit object is also inanimate. In other words, adults apparently tend towards feature-matching choices in that case.

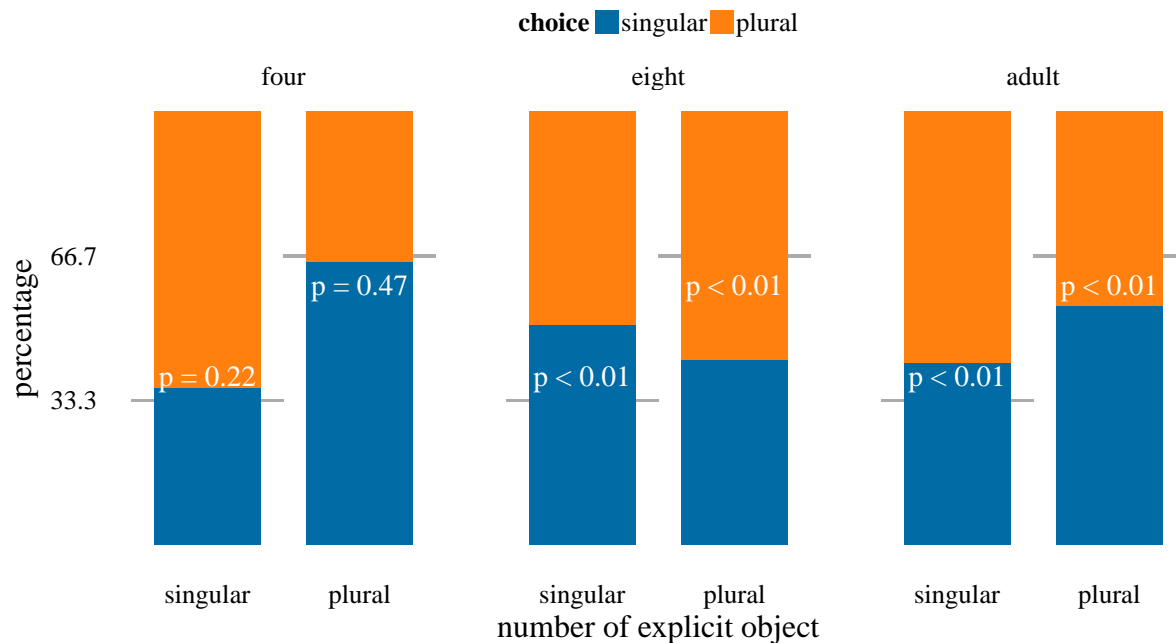


Figure 5.41: Percentage of choices by their grammatical number (split by number of the explicit object and by age group), with p -values of χ^2 tests (Holm-Bonferroni-corrected)

Feature-matching is also apparent in the percentages of singular and plural choices made by eight-year-olds and adults (all significant at $p < 0.01$), shown in Fig. 5.41. The effect is particularly striking in the choices made by the eight-year-old age group: even though each trial offered two mismatching options and only one matching one, eight-year-olds chose this one option that matched the number of the explicit object in 54% of trials. The choices made by four-year-olds again do not differ significantly from random chance here.

If the choices in this experiment reflect speakers' preferences or expectations regarding the dative alternation, **and** if those preferences are (or can be described as) preferred orderings of features, then the features of participants' choices in this experiment should change depending not only on the feature of the explicit object (which it does, as shown above), but also depending on the position of the gap in the instruction sentence. This gap is in the place of either the first or the second of the two objects in linear order, meaning either before or after the other, explicit object. Since a choice implicitly fills this gap, differences in the features of this choice across different positions reveal possible ordering preferences.

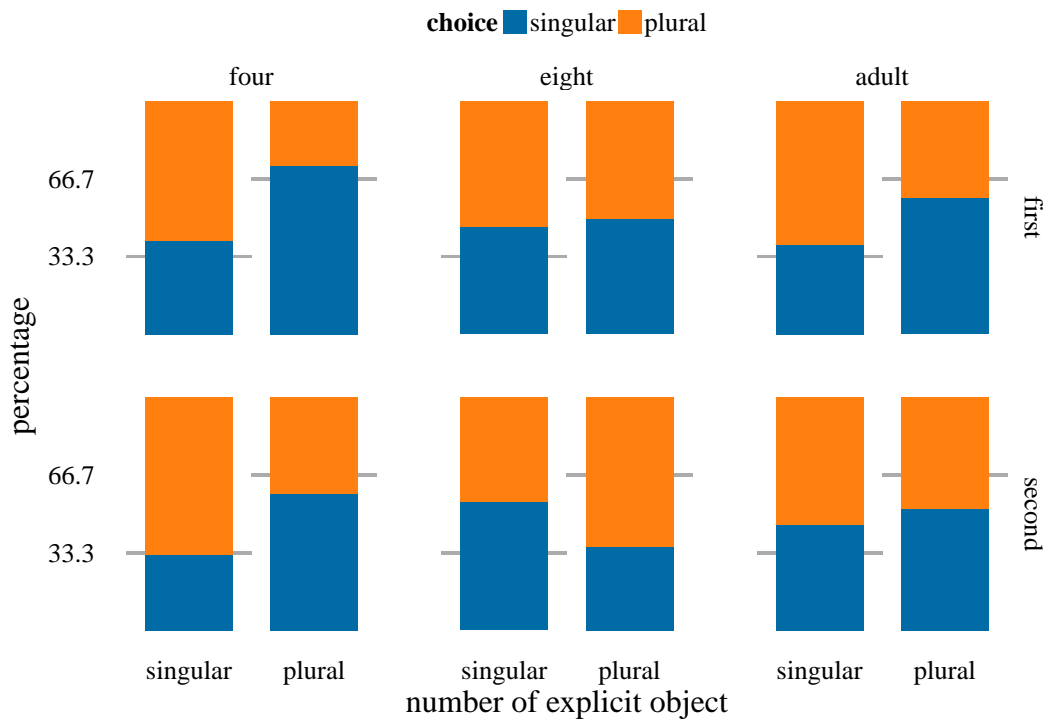


Figure 5.42: Percentage of choices by their grammatical number, split by number of the explicit object, by age group, and by position of gap in instruction sentence

Fig. 5.42 shows the proportions of choices by their number, divided three different ways. This is the same data as shown in Fig. 5.41, but with an additional split between the trials where the gap replaced the first object in the instruction sentence (*Now give the _____ to the dogs* or *Now give the _____ the keys*, shown in the top row of Fig. 5.42) and the trials where the gap replaced the second object in the instruction sentence (*Now give the dogs the _____* or *Now give the keys to the _____*, shown in the bottom row of Fig. 5.42). To make the figure easier to read, the grey lines again show the level that would be expected by chance, even though the above has shown significant deviations from this. If the ordering of objects does have an effect on the grammatical number, there should be a significant difference between a vertical pair of bars: for example, it appears that four-year-old participants were more likely to choose the one plural option in trials with plural explicit when the explicit object preceded the gap in the instruction sentence (second bar in the bottom row of Fig. 5.42) than when the explicit object followed the gap (second bar in the top row). A two-sample χ^2 test for equality of proportions shows this difference to be significant ($p < 0.01$ after Holm-Bonferroni correction for $m = 6$ comparisons). Eight-year-olds chose more plurals overall, as discussed above, but this preference was significantly stronger in plural-explicit trials with the explicit object preceding the gap in the instruction (fourth bar from the left, bottom row) than in plural-explicit trials with the explicit object following the gap (fourth bar, top row; $p < 0.01$). The other four vertical pairs of bars do not differ significantly,

and there are no significant differences in the length or animacy of the choices between trials with the explicit object preceding the gap and trials with the explicit object following the gap (all corrected $p > 0.05$).

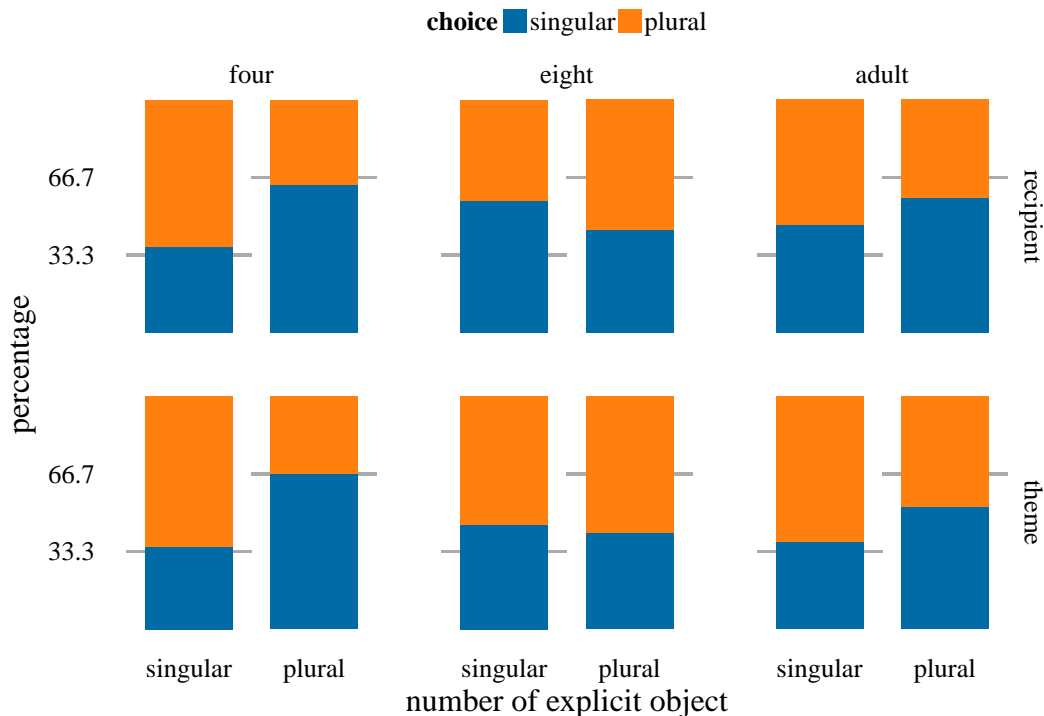


Figure 5.43: Percentage of choices by their grammatical number, split by number of the explicit object, by age group, and by function of gap/choice

Independent of the position of the gap was its function: in half of the trials, the gap replaced the recipient of the *give* in the instruction sentence (*Now give the _____ the keys* or *Now give the keys to the _____*); in the other half, the gap replaced the theme (*Now give the _____ to the dogs* or *Now give the dogs the _____*). In Fig. 5.43, the choices by number have been divided according to this function: the top row of bars shows the choices in the recipient-gap trials, and the bottom row shows the choices in the theme-gap trials. As above, two-sample χ^2 tests for equality of proportions were computed for each vertical pair of bars; here, this test investigates whether any age group differed in its choices depending on the function that that choice ostensibly performed in the instruction sentence. Only for the eight-year-old age group and singular explicit objects does that test show a significant difference ($p = 0.02$ after Holm-Bonferroni correction for $m = 6$ comparisons)—in other words, eight-year-olds were more likely to choose a plural when the gap in the instruction sentence replaced the theme than when it replaced the recipient. The difference between the two types of trials is not significant when the explicit object is plural, and neither is either difference for four-year-olds and adults (all corrected $p > 0.05$).

When analyzing these choices with respect to animacy, the only significant differences are with

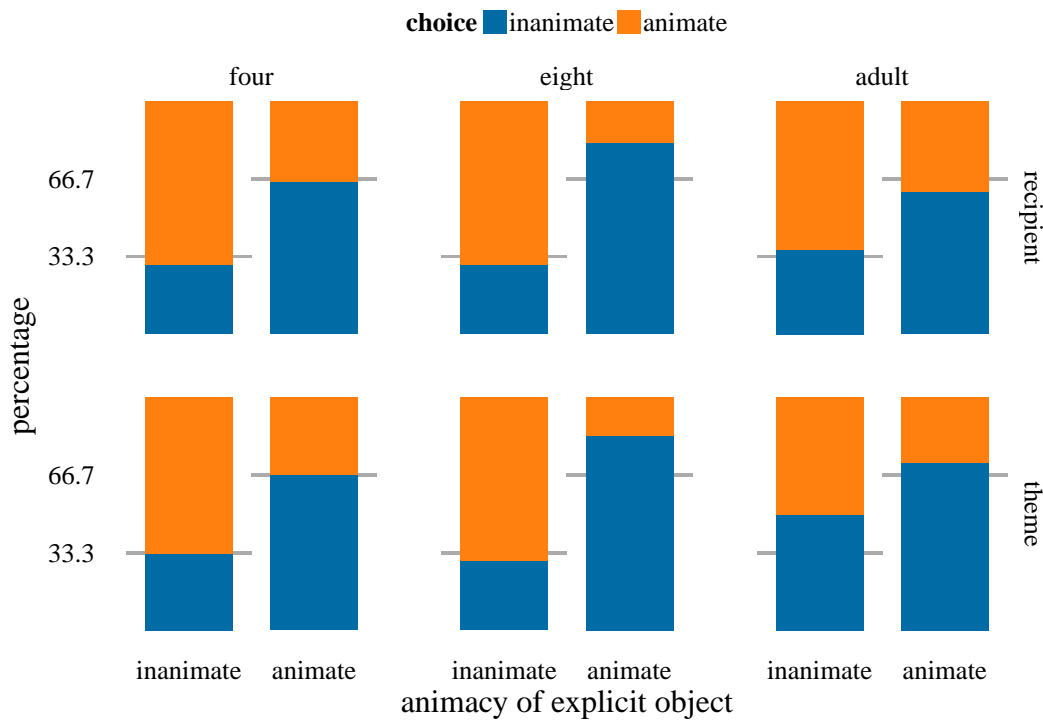


Figure 5.44: Percentage of choices by their animacy, split by animacy of the explicit object, by age group, and by function of gap/choice

adults: Fig. 5.44 shows that adults chose more inanimates as themes (the two rightmost bars in the bottom row) than as recipients (two rightmost bars, top row) regardless of whether the explicit was inanimate or animate, and two-sample χ^2 tests show both of these differences to be significant (corrected $p < 0.01$ and $p = 0.02$, respectively). The differences for the children's age groups appear small in Fig. 5.44 and insignificant in two-sample χ^2 tests (all corrected $p > 0.1$).

By design, the features of each **option** were not independent: the one option that matched the grammatical number of the explicit object always mis-matched its animacy and length, for example (see Section 5.1.2 for details). There were three different options, which (with this controlled manipulation of features) should allow feature-based preferences to surface. However, the explicit object manifested its values of the three features of interest all at the same time, and an analysis of the choices by each feature of the explicit separately (as done above) will therefore be limited in what it can reveal. For a less limited analysis, three generalized linear models were fit to the data, one for each feature of the choice. Each model approximated had one of the features of participants' choices (length, animacy, or number) as its dependent variable. This is reasonable if the dependent variables are not strongly associated with each other, which these three are not: Yule's coefficient of association is between -0.1 and 0.1 for each pairwise association, meaning the three features of participants' actual **choices** are almost

perfectly independent of each other. The dative construction used in the instruction sentence and the function of the gap in the instruction sentence (which the choice ostensibly filled) were included as independent variables. Bresnan et al. (2007) argue that the dative alternation can be viewed as an ordering phenomenon depending on certain features of the two objects, and that there are feature-specific ordering preferences which are independent from each other (short before long independent from animate before inanimate, and so on). Therefore, the interaction between age group, relative position of the gap (first or second of the two objects), and the explicit object's value for the feature of the response that a model was estimating (length of the explicit object in the length model, animacy of the explicit object in the animacy model, number of the explicit object in the number model) were also included. The other two features of the explicit object (animacy⁸² and number in the length model, and so on) were also included as main effects.⁸³

Table 5.9 gives the coefficients of the length model. Here, monosyllabic choices are the reference or 0-level, and bisyllabic choices the alternative level (coded as 1). Variables with positive parameter estimates thus make the (numeric) result of the model formula calculation larger, pushing the model prediction towards bisyllabic choices. Likewise, variables with negative parameter estimates favor monosyllabic choices. Of the significant effects (at $p < 0.05$), that of the length of the explicit object and its interactions with age group is the most striking: monosyllabic choices were much more likely when the explicit object was bisyllabic, and even more so among eight-year-olds⁸⁴ and adults than among four-year-olds. As there were two monosyllabic options and one bisyllabic one in trials with bisyllabic explicit objects, this is not

⁸²For the sake of consistency, a priori animacy was used as the categorical independent variable in all three models. Equivalent models with a numeric independent variable derived from the results of experiment 1 (see Section 4.2) instead of this categorical variable were also fit (with all other variables being the same). This numeric independent animacy variable was derived from the group-wise percentages of “yes” answers in experiment 1, with the percentage of answers to “Could this one move towards you?” weighted double because the answers to that questions were split more cleanly. (For example, 75% of four-year-old participants said that the crab could move towards them, and 25% said that it could play with them (see Fig. 4.1 on page 69). The group-wise numeric animacy value for *crab* for the four-year-old group was thus $\frac{(0.75 \times 2) + (0.25 \times 1)}{3} \approx 0.58$. The adult group gave markedly different answers for *crab*: 100% said it could move, and 36% said it could play, giving $\frac{(1.00 \times 2) + (0.36 \times 1)}{3} \approx 0.79$ as the adults' group-wise numeric animacy value for *crab*.) The models with this numeric animacy value as an independent variable did not differ in interesting ways from the models with categorical a priori animacy as an independent variable: all effects that were significant in one model were also significant in the corresponding other one (and worked in the same direction), and no effect that was insignificant in one was significant in the other. In the interest of clarity, only the models with a priori animacy are reported here.

⁸³For readers familiar with regression modelling in R, the linear model formula for the length model was `choice.length ~ (explicit.length * gap.position * group) + explicit.animacy + explicit.number + construction + gap.function`.

⁸⁴Note that the p -value of the bisyllabic: eight-year-old interaction is between 0.045 and 0.05 and was therefore rounded to 0.05. The z -value of -2.00 is outside the non-significant range of -1.96 to 1.96 , which shows the significance of this effect more clearly. The z of the main effect associated with eight-year-olds, on the other hand, is just inside this range, meaning that that main effect is not deemed to be significant here even though its rounded p -value is also 0.05 (the more specific value being somewhere between 0.05 and 0.055 and thus larger than 0.05).

surprising. The significant and positive main effect associated with the adult age group means that adults prefer bisyllabic options in trials with monosyllabic explicit objects (the reference level for the length effect in this model) and, conversely, monosyllabic options in trials with bisyllabic explicit objects. This was apparent in Fig. 5.39 and is confirmed by an entirely different method here. The effect of number in this model is unexpected: everything else being equal, bisyllabic choices were more likely when the explicit object was plural than when it was singular. When the gap (and thus the choice, implicitly) was the theme of the instruction sentence, monosyllabic choices were preferred. The most interesting interaction, that between the length of the explicit and the relative position of the choice, is significant only in interaction with the age group: above the other effects, eight-year-olds preferred monosyllabic choices when the instruction sentence had the gap after a bisyllabic explicit object. Four-year-olds and adults do not show this ordering effect, however.⁸⁵

variable	parameter estimate	standard error	<i>z</i>	<i>p</i>
(Intercept)	0.65	0.14	4.70	< 0.01
bisyllabic explicit object	-1.63	0.17	-9.44	< 0.01
animate explicit object	-0.13	0.08	-1.71	0.09
plural explicit object	0.22	0.08	2.88	< 0.01
eight-year-olds	0.33	0.17	1.95	0.05
adults	0.86	0.18	4.79	< 0.01
gap after explicit	0.18	0.17	1.08	0.28
gap was theme	-0.17	0.08	-2.28	0.02
prepositional construction	-0.03	0.08	-0.45	0.65
bisyllabic explicit : gap after explicit	0.21	0.24	0.88	0.38
bisyllabic explicit : eight-year-olds	-0.50	0.25	-2.00	0.05
bisyllabic explicit : adults	-1.15	0.26	-4.50	< 0.01
gap after explicit : eight-year-olds	0.22	0.25	0.89	0.37
gap after explicit : adults	0.08	0.26	0.29	0.77
bisyllabic explicit : gap after explicit : eight-year-olds	-0.99	0.36	-2.72	0.01
bisyllabic explicit : gap after explicit : adults	-0.61	0.37	-1.67	0.10

Table 5.9: Coefficients for the model of the length of choices (positive parameters indicate effects favoring bisyllabic choices, variables in bold are deemed to have significant effects; $MSE = 0.178$, $_{100}M_{10} = 0.180$)

The coefficients of the model for the animacy⁸⁶ of the choice are shown in Table 5.10 (note that

⁸⁵It might be argued that the preposition *to* introduces a syllable to the length of the recipient object in the prepositional sentences and thus could confound the true length effect. However, the interaction between length and ordering is not significant in a model (not shown here) based on only the data from trials with a double-object instruction sentence either.

⁸⁶While it would be possible to use a numeric measure of animacy based on the results from experiment 1 as the dependent variable (fitting a linear regression model instead of a generalized one), I see no way of deriving a useful measure: averaged or weighted percentages of “yes” or “no” responses across all participants would go against the fundamental assumptions that the age groups differ from each other. Using by-group percentages would avoid this problem, but it would make the model rather more difficult to interpret: if, for example, a particular option had a lower animacy score for eight-year-olds and adults, a lower parameter estimate for the interaction of one feature with eight-year-olds may still mean an equally strong effect, or it may not. Finally, a by-participant measure would mean introducing individual participant effects via the back door, under the guise of this dependent variable that could be different for each participant. As mentioned above, most categorical splits that could be derived from the

this model was not pruned, for reasons laid out in Section 5.2.5.3). The effect of the animacy of the explicit object here has the same cause as the effect of explicit length in the length model above: since there were two inanimate options in all trials with an animate explicit object, it is not surprising that animate explicit objects go together with inanimate choices. Likewise, it is apparent from Fig. 5.40 that adults on the whole chose more inanimates, and that eight-year-olds chose significantly more inanimates when the explicit was animate, so the significant main effect of adults and the significant interaction for eight-year-olds and animate explicit objects here again serves as independent confirmation of that finding. The two remaining significant effects are more interesting: when the gap in the instruction sentence was the theme, participants tended to choose inanimates to fill the gap. Independent of this, when the gap in the instruction sentence was after an animate explicit object, inanimates were also more likely to be chosen. The theme effect is readily explained as a preference for (or prototypicality of) inanimate themes and animate recipients with *give* (and other ditransitive transfer verbs). The interaction between explicit animacy and relative position, finally, is an order effect predicted by Bresnan et al. (2007): inanimate objects are preferred following animate ones.

Experiment 3 shows that participants of all ages prefer animate recipients and inanimate themes with *give* (see Section 6.2 for details). As the prepositional construction has often been argued to be less restrictive with regard to what verbs it can be used with and what meanings it can encode (Oehrle 1976, Gropen et al. 1989, Rappaport Hovav and Levin 2008), it is conceivable that these animacy preferences may be stronger in trials with double-object instruction sentences. As the double object construction has the recipient as the first object, this would mean that animate choices are more preferred in double-object trials with the gap in place of the first object than in other sorts of trials. If this preference was strong enough, this would result in a dataset showing an apparent preference for animates before inanimates, even though there was none. As this preference is exactly what the model for the animacy of the choice shows, this possibility was investigated by splitting the dataset into two sub-sets, one with all the double-object trials and one with all the prepositional trials, and fitting a model with the same formula (except the main effect of construction, of course) to each of them. If the interaction between animacy and order was significant in the double-object model only, the above account for the ordering effect in the model for all data would be strengthened. However, the interaction is insignificant in both of the smaller models (not shown here). Thus, there is no evidence for a stronger animacy preference with the double object construction. The fact that the interaction that shows the ordering effect is insignificant in both smaller models does not necessarily cast doubt on the significance of that effect in the larger model: the smaller models were based on half the dataset each, and thus had larger standard errors for all parameter estimates. Since the standard error is used in assessing

results of experiment 1 would simply replicate the a priori split between animate and inanimate objects. Therefore, that a priori categorical split is used here.

effect significance (see Section 5.2.5.1 for details), smaller datasets make z - or t -values smaller and p -values larger. The non-significance of the ordering interaction in the smaller models does not invalidate the significance of the same interaction in the larger model.

variable	parameter estimate	standard error	z	p
(Intercept)	0.93	0.14	6.69	< 0.01
bisyllabic explicit object	-0.08	0.07	-1.09	0.28
animate explicit object	-1.23	0.17	-7.42	< 0.01
plural explicit object	-0.05	0.07	-0.72	0.47
eight-year-olds	-0.03	0.17	-0.17	0.87
adults	-0.55	0.16	-3.41	< 0.01
gap after explicit	0.17	0.17	0.98	0.33
gap was theme	-0.24	0.07	-3.52	< 0.01
prepositional construction	-0.07	0.07	-1.03	0.30
animate explicit : gap after explicit	-0.47	0.24	-1.98	0.05
animate explicit : eight-year-olds	-0.96	0.25	-3.83	< 0.01
animate explicit : adults	0.40	0.23	1.75	0.08
gap after explicit : eight-year-olds	0.19	0.24	0.77	0.44
gap after explicit : adults	0.10	0.23	0.43	0.67
animate explicit : gap after explicit : eight-year-olds	-0.01	0.36	-0.02	0.98
animate explicit : gap after explicit : adults	0.18	0.33	0.57	0.57

Table 5.10: Coefficients for the model of the animacy of choices (positive parameters indicate effects favoring animate choices, variables in bold are deemed to have significant effects; $MSE = 0.209$, $_{100}M_{10} = 0.211$)

As with the two other models, the effect of the number of the explicit object in the model for the number of the choices (see Table 5.11 for coefficients) is not surprising: every plural explicit object was accompanied by two singular options and only one plural one, so the strong effect in favor of singular choices is caused by the experimental design. The effect of length (long explicit objects favoring singular choices) is unexpected here, but it is parallel to the number effect in the model for the length of the choice (where plural explicit favor short choices). Plural options were apparently also more likely to be chosen when the choice was to be the theme (or, conversely, singulars were more likely as recipients). Four age-related interactions were significant: both eight-year-olds and adults chose relatively more plural options when the explicit was also plural. This effect works against the expected design effect discussed above to produce the tendency (seen in Fig. 5.41) of eight-year-olds and adults to match the number of the explicit in their choice. Finally, both eight-year-olds and adults chose more singulars when the gap followed the explicit object in the instruction sentence, regardless of the number of that explicit object. The number-sensitive ordering effect (expressed in the model as three interactions: the two-way interaction between number of explicit and relative position of gap; the three-way interaction between number of explicit, relative position of gap and eight-year-old age group; and the three-way interaction between number of explicit, relative position of gap

and adult age group) is not significant, meaning that the number of the chosen object was not affected by the number of the explicit and its position together—all else being equal, singulars were not more preferred after plurals, as the harmonic alignment findings in Bresnan et al. (2007) would predict.

variable	parameter estimate	standard error	<i>z</i>	<i>p</i>
(Intercept)	0.32	0.13	2.42	0.02
bisyllabic explicit object	-0.14	0.07	-2.13	0.03
animate explicit object	0.07	0.07	1.09	0.28
plural explicit object	-1.37	0.17	-8.06	< 0.01
eight-year-olds	-0.26	0.16	-1.60	0.11
adults	0.06	0.16	0.36	0.72
gap after explicit	0.32	0.17	1.93	0.05
gap was theme	0.19	0.07	2.98	< 0.01
prepositional construction	0.05	0.07	0.82	0.41
plural explicit : gap after explicit	0.30	0.24	1.26	0.21
plural explicit : eight-year-olds	1.23	0.23	5.29	< 0.01
plural explicit : adults	0.55	0.23	2.42	0.02
gap after explicit : eight-year-olds	-0.68	0.23	-2.98	< 0.01
gap after explicit : adults	-0.59	0.23	-2.62	0.01
plural explicit : gap after explicit : eight-year-olds	0.64	0.33	1.94	0.05
plural explicit : gap after explicit : adults	0.24	0.32	0.75	0.45

Table 5.11: Coefficients for model of the grammatical number of choices (positive parameters indicate effects favoring plural choices, variables in bold are deemed to have significant effects; $MSE = 0.237$, $_{100}M_{10} = 0.239$)

It seems a truism that speakers prefer using frequently used words over rarer words—that is, after all, how word frequency is defined. Word frequency is known to have many other effects: for example, frequent words are read more quickly than rarer words (Whitford and Titone 2012), speakers are quicker to recognize frequent words as words than they recognize rarer words as words (Keuleers et al. 2011), and frequent words are less likely than rarer words to show sound changes (Hay et al. 2015). It is conceivable that participants in this experiment were affected by the frequencies of the choices: since frequent words are quicker to read and recognize, speakers may have a subconscious preference for more frequent words. To test for this, the base-10 logarithm⁸⁷ of the CELEX lemma frequency of the intended nouns was computed for all images used as options (or small images). For each trial, the mean frequency of the two options that were not chosen was then subtracted from the frequency of the choice. A positive value for this relative frequency measure thus means that the choice noun is more frequent than the other two options, and a negative value means that at least one of the options not chosen is more frequent than the choice.

⁸⁷The distribution of word frequencies covers a large range, and the raw frequencies are commonly converted into their logarithms in order to arrive at more useful, less wide-ranging measure of frequency.

The relative frequency measure is summarized in Fig. 5.45. It is apparent that the relative frequency values fall around to zero, which suggests that more frequent words were not chosen more overall. Because the distribution of these values is not normal, the Kruskal-Wallis rank sum test was used to test whether there are significant differences between the three age groups here, and there are ($p < 0.01$). Specifically, pairwise Mann-Whitney tests show that the relative frequencies of four-year-olds' choices differ significantly from those of eight-year-olds and adults (both $p < 0.01$ after Holm-Bonferroni correction for $m = 3$ comparisons). Eight-year-olds and adults do not differ significantly ($p > 0.1$). This frequency measure therefore reveals that eight-year-olds and adults showed a slight preference for more frequent words here, and that four-year-olds did not. This is intriguing in light of previous findings that also suggest word frequency effects grow stronger with age: in word-recognition tasks, older children show a larger difference between high- and low-frequency words than younger children do (Pearson and Studt 1975); in reading, older adults show a larger difference between high- and low-frequency words than adolescents and younger adults do (Rayner et al. 2006). The present finding is in line with this literature.

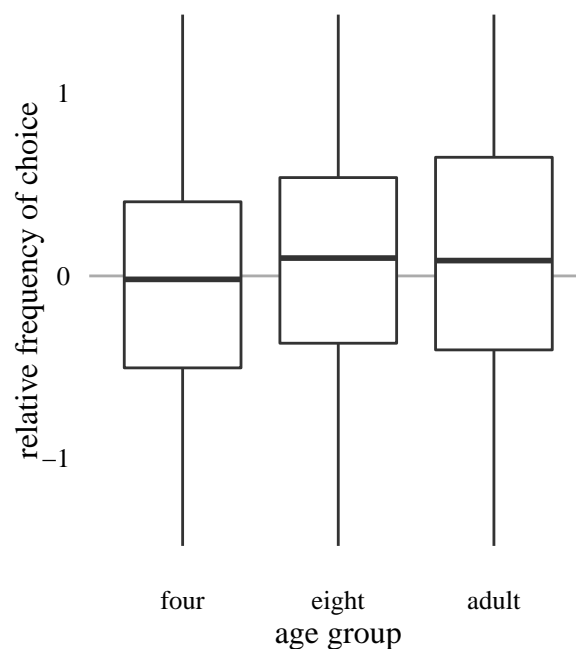


Figure 5.45: CELEX lemma frequency of choice relative to CELEX lemma frequency of other options, by age group (lower and upper ends of boxes indicate first and third quartiles, vertical lines extend to the most extreme value within 1.5 interquartile ranges of these quartiles)

5.3.3 Touch input

Participants dragged images on the touchscreen in this experiment. These input patterns of sequences of touch points (sampled at 60 Hz), or ‘paths’, were recorded. For each of these paths, the length of the path (as the sum of the Euclidean distances between each pair of consecutive touchscreen input points, recorded at 60 Hz) and the Euclidean straight-line distance between its start and end points were measured. These were then used to calculate the sinuosity of each path, which is the path length divided by the Euclidean distance between start and end (Stølum 1996). Sinuosity can be understood as a proportion: a path with sinuosity 1.5 is 1.5 times as long as the Euclidean distance between its start and end points. As a path between two points cannot be shorter than the distance between the two points, sinuosity cannot be less than 1. Paths with sinuosity close to 1 are almost perfectly straight, and larger sinuosity values mean more curved or ‘roundabout’ paths. Fig. 5.46 presents these differences in bins of 0.5 each. It is apparent that the vast majority of dragging paths (90.7%) have a sinuosity value between 1 and 1.5, regardless of age group. This means that participants overwhelmingly produced very straight dragging paths, although participants in the four-year-old age group seem to have shown slightly more variation.

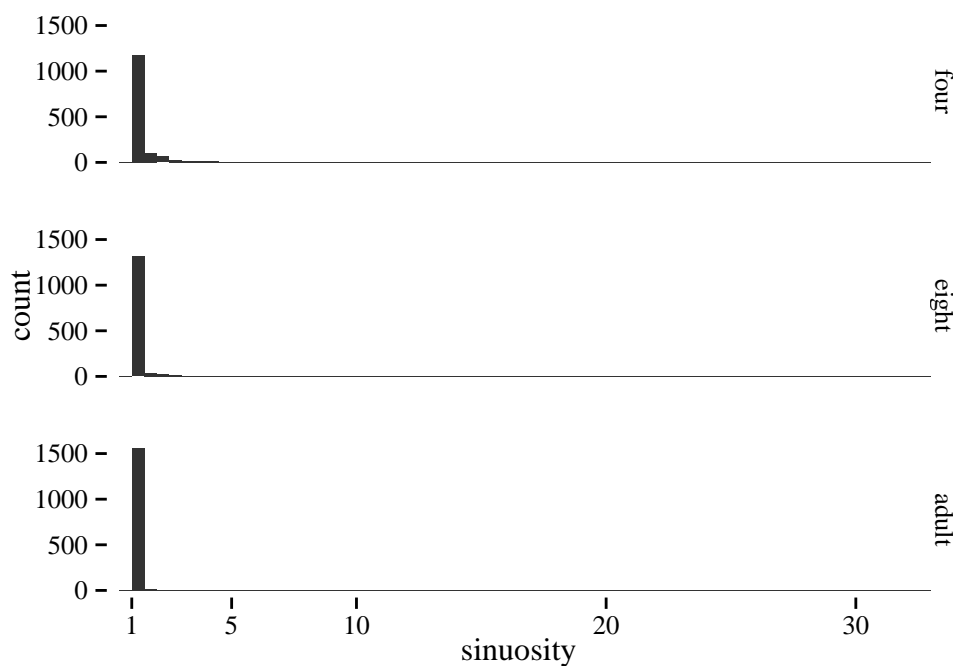


Figure 5.46: Sinuosity of drag paths by age groups

The same age group difference is also apparent in the duration of each drag path, shown in Fig. 5.47 (141 paths (3.2%) lasted longer than 5 seconds and are not shown in this plot for reasons of clarity). Most dragging paths were completed very quickly: 3407 paths (76.3%) were completed in less than 1 second each.

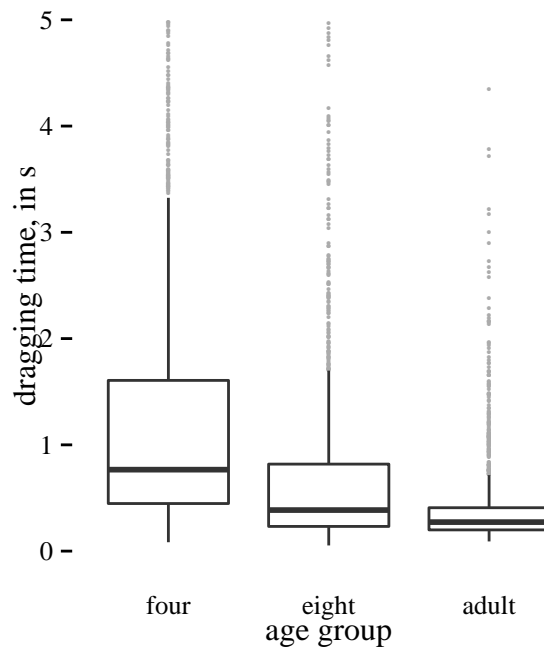


Figure 5.47: Durations of drag paths by age groups (cut off at 5 s; lower and upper ends of boxes indicate first and third quartiles, vertical lines extend to the most extreme value within 1.5 interquartile ranges of these quartiles, grey dots are more extreme values)

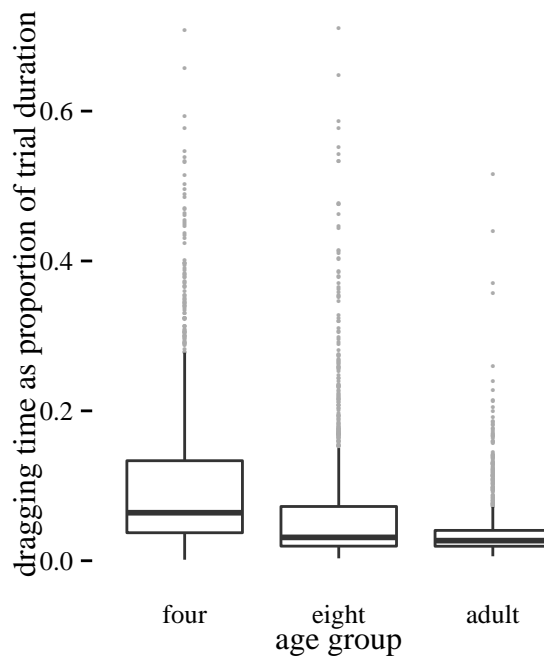


Figure 5.48: Durations of drag paths as a proportion of total trial time, by age groups (cut off at 5 s)

Four-year-olds were slower to complete trials and took longer to drag images overall. The proportion of the total trial time spent dragging may therefore be a better measure of dragging speed than the raw drag path duration. While drag path duration and trial time are correlated across age groups ($\rho = 0.47$, $p < 0.001$), there are differences between the age groups: the proportion of each trial spent dragging was higher for the four-year-old group, as Fig. 5.48 shows. On average, four-year-olds spent 10.1% of each trial dragging images, eight-year-olds took 6.2%, and adults 3.8%.

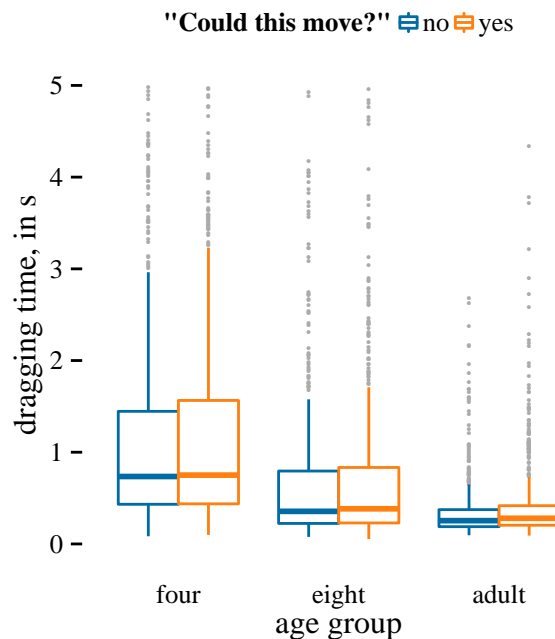


Figure 5.49: Durations of drag paths, by age groups and participant answers to “Could this one move towards you?” (cut off at 5 s)

The durations of drag paths were minutely different depending on whether the participant doing the dragging had answered “yes” or “no” to the question “Could this one move towards you?” in experiment 1 (see Section 4.2): if a participant answered “yes” for an image, they dragged that image and its plural counterpart slightly slower. This is represented by the orange boxes in Fig. 5.49, which are slightly higher than their blue counterparts (which show the durations of dragging paths for images that could not move according to the individual participant). A linear regression model (see Table 5.12) shows that this difference is not significant for any age group, however. More sinuous paths took longer, for obvious reasons, and eight-year-olds and adults were significantly quicker than four-year-olds.

As described in Section 5.1.2, participants had to touch each of the three smaller (‘choice’) images before the trial progressed to the interactive stage. 21 participants (12 adults, 4 eight-year-olds, and 5 four-year-olds) touched first the left, then the middle, and finally the right image

variable	parameter estimate	standard error	t	p
(Intercept)	-0.16	0.06	-2.70	0.01
eight-year-old age group	-0.58	0.09	-6.37	< 0.01
adult age group	-1.40	0.28	-4.97	< 0.01
sinuosity	1.30	0.02	54.64	< 0.01
“yes” answer to move question	-0.06	0.07	-0.85	0.39
eight-year-olds : sinuosity	0.14	0.05	2.92	< 0.01
adults : sinuosity	0.55	0.26	2.09	0.04
eight-year-olds : “yes” answer	-0.13	0.10	-1.34	0.18
adults : “yes” answer	0.12	0.09	1.25	0.21

Table 5.12: Coefficients for regression model of dragging path duration (variables in bold are deemed to have significant effects; $MSE = 1.58$, ${}_{100}M_{10} = 1.55$)

on 60 or more trials. One adult and one eight-year-old had the reverse pattern (right-to-left) in 60 or more trials. Among the 19 remaining participants (who had no consistent pattern throughout the entire experiment), there was still a strong preference for the left-to-right order (used in 57.1% of their trials). These patterns indicate that participants completed this part of the task very routinely and without problems. It is reasonable to assume that the left-to-right order is the most common one due to it also being the reading direction in English, particularly since it was preferred more strongly by adults (who can be assumed to be expert readers) than by young children. Possible effects of handedness cannot be investigated here since handedness was not recorded.

5.3.4 Eye gaze

The eye gaze data was analyzed using smoothing spline analysis of variance (SSANOVA). This statistical method is described in detail in Section 5.2.4; briefly, SSANOVA fits smooth curves with confidence intervals to the curves described by data, allowing for random effects (of participant, here). Where the confidence intervals of two SSANOVA curves do not overlap, a significant difference between the two curves can be assumed. As there is a random component to SSANOVA model fits,⁸⁸ false positives can occur. Using simulated datasets, I have argued in Section 5.2.4.4 that SSANOVA curves whose confidence intervals never overlap throughout the entire time period under analysis are to be treated with caution, and that curves with non-overlapping confidence intervals for only part of the time period are more likely to be true positive findings of significant differences.

In this section, I present plots of several SSANOVA models and discuss them. All models

⁸⁸All SSANOVA models presented here were fit with the same random seed, namely 2332.

are based on gaze data that was processed as described in Section 5.2.3. As the three smaller images differed in the features of interest, the smaller images were divided by these features (into animates and inanimates, for example), and gaze percentages were calculated for each. Since the most important linguistic stimulus in this touchscreen/eyetracking experiment was the instruction sentence (*Now give the _____ to the dog.*), the eye gaze data time stamps were normalized to the start of the instruction sentence for these SSANOVA models. Due to limitations of the Python script used in the experiment, eye gaze data was not reliably recorded for about half a second from the start of the instruction sentence. To rule out possible effects of unreliable or missing data in this time period, only data from after 0.7 seconds after the start of the instruction sentence was included in the analysis. It seems likely that gaze behavior during passive looking and gaze behavior during interaction (touchscreen dragging, in this case) would be different: Mrotek and Soechting (2007) found that participants who were asked to follow a moving stimulus with their eyes exhibit different saccade behavior than participants who were asked to intercept the same moving stimulus with their hand, and Bieg et al. (2010) found that the distance between the gaze location and a mouse cursor differs in different tasks. Even the mere presence of their hand in their visual field changes participants' gaze patterns (Thura et al. 2008). To rule out this possible confound, the start of participant touchscreen dragging input is the end of the time window for the SSANOVA models presented here. The SSANOVA model curves for this dataset are relatively level more than 3 seconds after the start of the instruction sentence, mostly due to sparseness of data. Therefore, data from that less interesting period is not shown here.

The first SSANOVA model was fit to data divided by whether the gazes were on the small image that the participant would eventually choose as their response this trial, one of the other two smaller images, or the larger image (explicitly named in the instruction sentence). The faded and more jagged lines in Fig. 5.50 represent the data for this split, averaged across participants. The smooth lines represent the respective SSANOVA models, and the ribbons surrounding them show the 95% confidence intervals of the models. This shows very clearly the conservativeness advantage of SSANOVA: the peak in four-year-olds' gazes on the response image (faded orange line) just after 1 second from the start of the instruction, for example, is visually impressive, but the corresponding model does not fit it very closely. The peak in adults' gazes on the response image just before 1 second, on the other hand, is better supported across trials and participants, and thus the SSANOVA curve (orange line with ribbon) approximates it. As the orange ribbon does not overlap either of the other two ribbons in this time period, this peak can be assumed to be significant. In other words, around 1 second after the start of the instruction sentence (which is during the instruction sentence still), adults are already much more likely to gaze at the image they will choose as their response than at either of the other two options (which are combined into the blue line). Four- and eight-year-olds do not exhibit the same pattern in their gaze behavior to a significant degree—in fact, the percentage of gazes on the response does

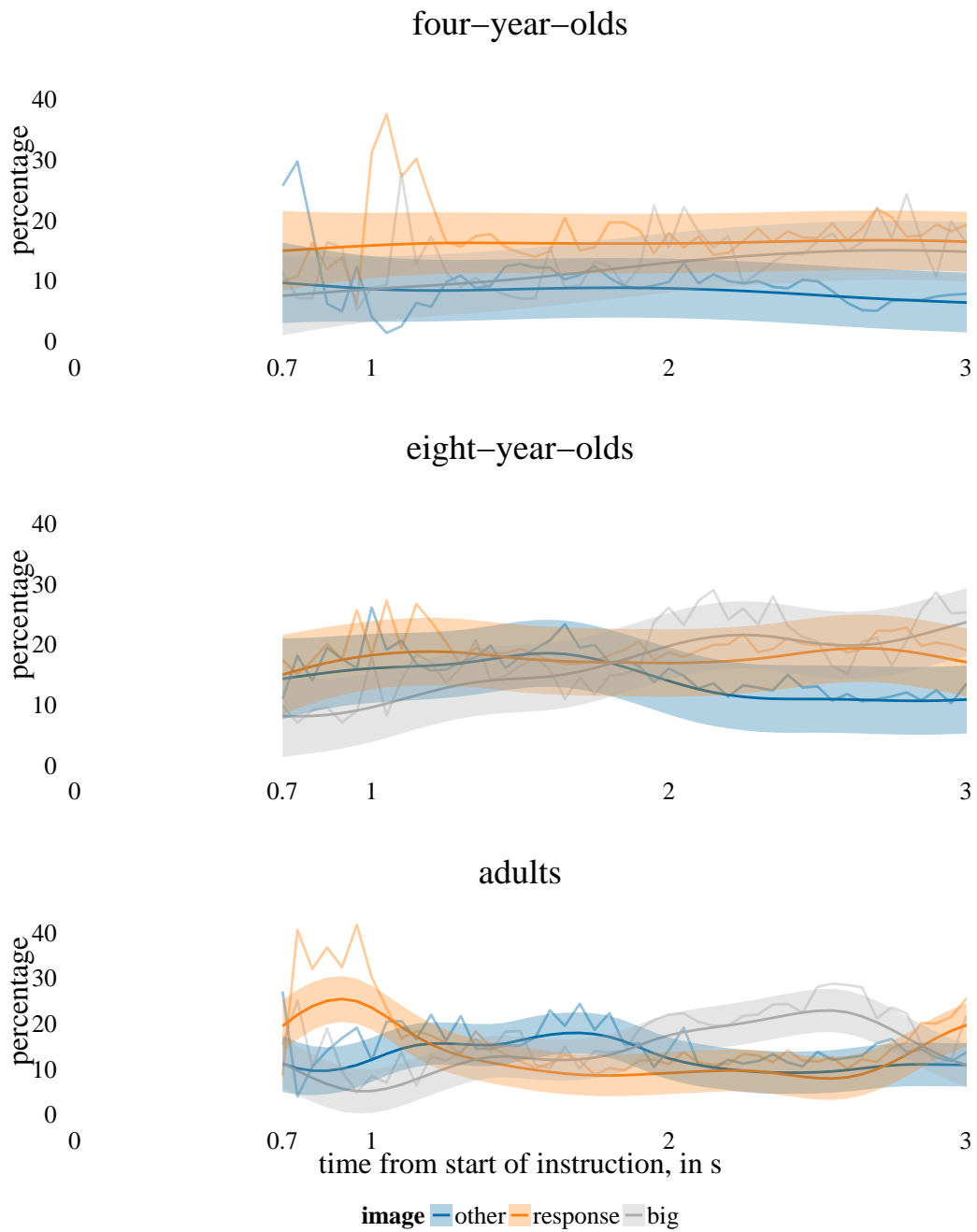


Figure 5.50: Percentages of gaze on images split by response (faded jagged lines) and SSANOVA models with confidence intervals (smooth lines with ribbons)

not appear to be significantly different from the percentage of gazes on the other two options throughout the entire time window. If there was no preference for any of the three smaller images, we would expect that the percentage of gazes on the response would be significantly lower than the percentage of gazes on the other two, since there are two of them. The fact that there is no significant difference either way in the children's gaze data is thus significant in itself: contrary to what we would expect if there was no preference, the response was not less likely to be gazed at than the other two options combined. Therefore, it is reasonable to assume that there is a preference for gazing at the response image; it is merely less pronounced and immediate with the children than it is with the adults.

Each one of the three smaller images matched the larger image in exactly one of the three features of interest—for example, a large image of three dogs would be accompanied by images of one camel (also animate, but singular and bisyllabic), three baskets (also plural, but inanimate and bisyllabic), and a lock (also monosyllabic, but inanimate and singular). Splitting the three smaller images up by which feature of the larger image they matched and fitting SSANOVA models to this data (see Fig. 5.51) reveals that adults' gaze was significantly more likely to be directed at the animacy-matching⁸⁹ smaller image (orange curve and ribbon) than at the number-matching one (blue curve and ribbon) in the time between 0.7 and 1 seconds after the start of the instruction. The children's gaze data does not exhibit a significant difference here.

Fig. 5.52 shows the data and models for the trials where the explicit object was animate, with the smaller images divided into inanimates (blue) and animates (orange). By design, there were two inanimate options and one animate option in these trials. Therefore, the fact that both four- and eight-year-olds were significantly more likely to gaze at inanimates than at animates here (the blue ribbon being higher than the orange one, and not overlapping with it) is not surprising. Adults, however, did not show this expected significant difference (the blue and orange ribbons overlap). This means that, in trials with an animate explicit, adults looked at the one animate option at least as much as at the other two options combined up to about 2 seconds after the start of the instruction sentence (which is where the ribbons diverge, at least for a fraction of a second and by a very short distance). In the trials with inanimate explicit objects (not shown here), all age groups looked at the two animate options significantly more than at the one inanimate one.

When the explicit object was monosyllabic, four- and eight-year-olds looked at one of the two bisyllabic options more than at the one monosyllabic one at around the 1-second mark, but this (expected) preference disappears between 1 and 2 seconds after the start of the instruction

⁸⁹This SSANOVA approach to analyzing eye gaze data by features of the gaze targets requires a categorical split in the animacy variable. A numerical measure derived from the results of experiment 1 would therefore be impracticable here, as a categorical split derived from these results would simply replicate the a priori split between animals and inanimate objects. Therefore, that a priori categorical split is used for all SSANOVA analyses here.

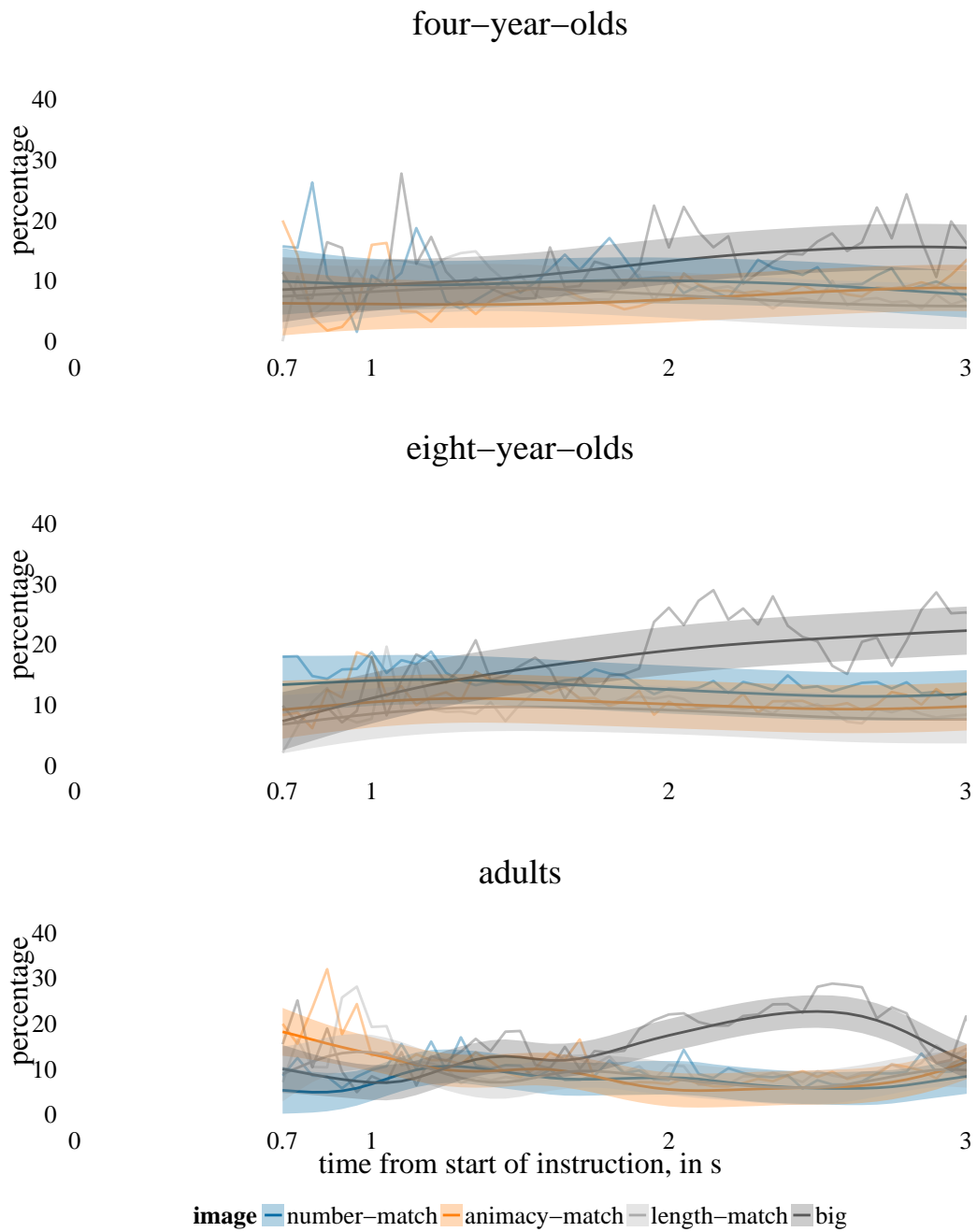


Figure 5.51: Percentages of gaze on images split by which feature of the ‘big’ image they matched (faded lines) and SSANOVA models with confidence intervals (lines with ribbons)

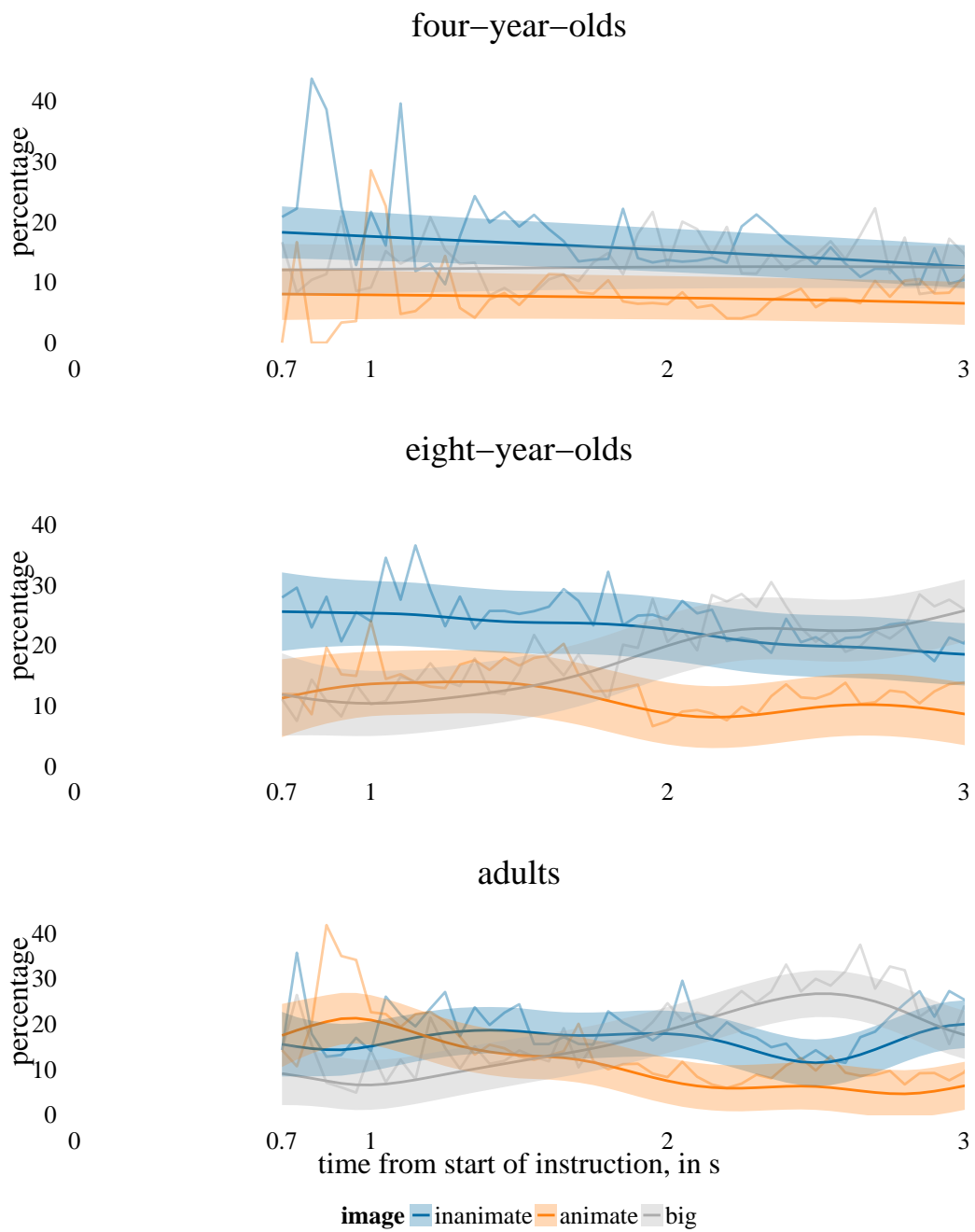


Figure 5.52: Percentages of gaze on images split by animacy in trials with animate explicit object (faded lines) and SSANOVA models with confidence intervals (lines with ribbons)

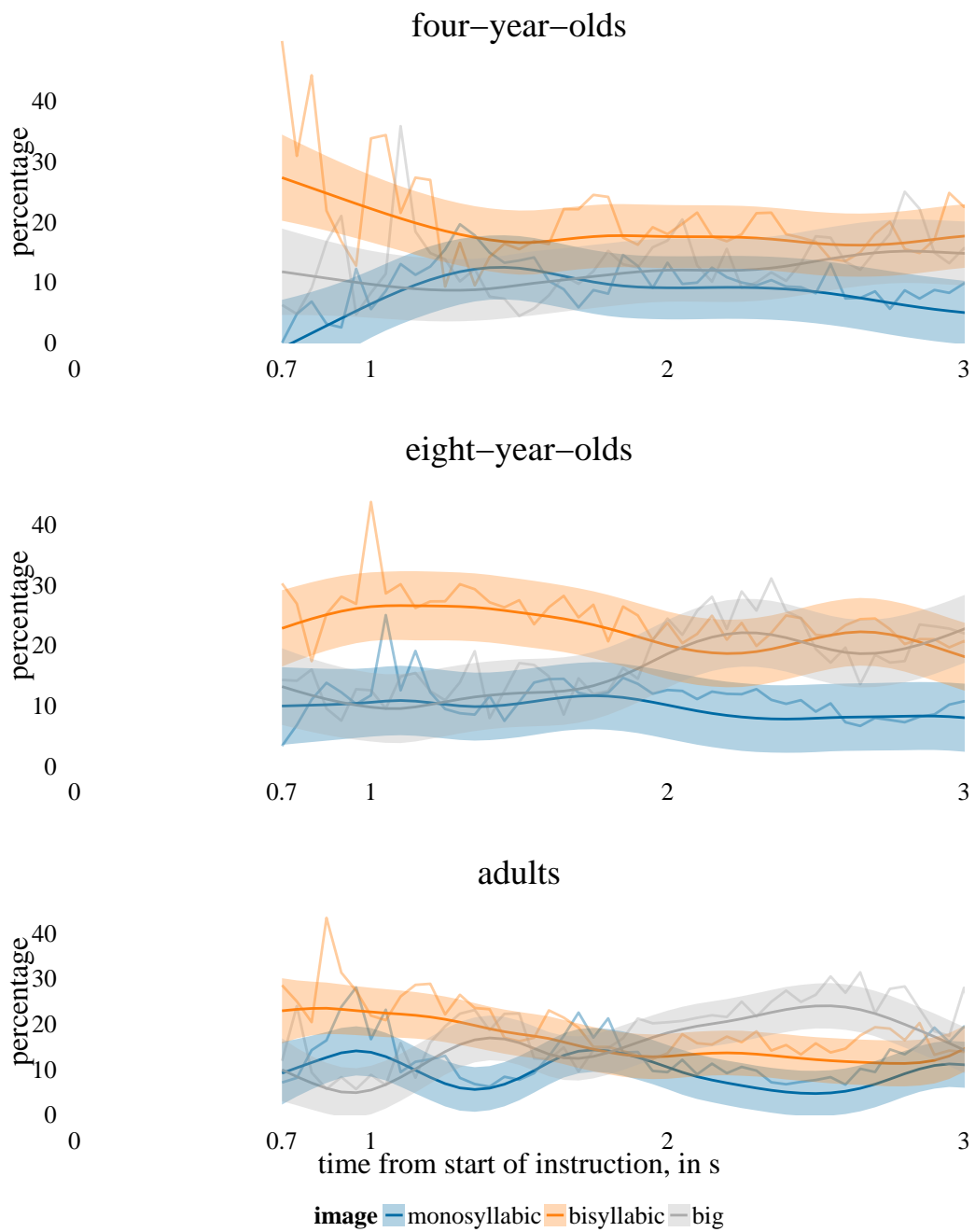


Figure 5.53: Percentages of gaze on images split by length in trials with monosyllabic explicit object (faded lines) and SSANOVA models with confidence intervals (lines with ribbons)

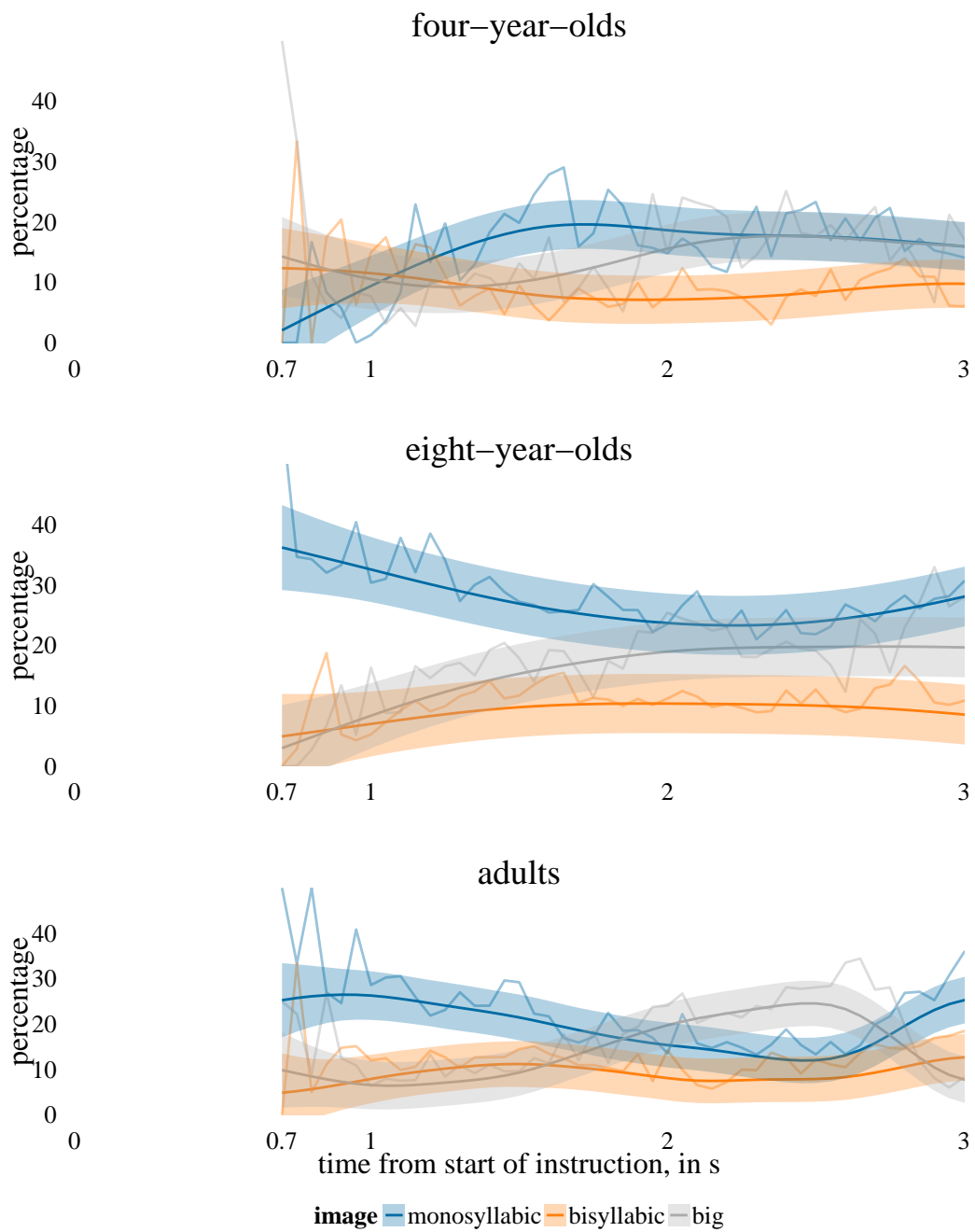


Figure 5.54: Percentages of gaze on images split by length in trials with bisyllabic explicit object (faded lines) and SSANOVA models with confidence intervals (lines with ribbons)

sentence (see Fig. 5.53). Adults' gazes only show this expected difference from the 1-second mark to about the 1.4-second mark, meaning they looked at the monosyllabic option more than expected before and after this time period. When the explicit object was bisyllabic, however, adults and eight-year-olds exhibit the expected pattern: they gazed significantly more at the two monosyllabic options than at the one bisyllabic one (see Fig. 5.54). The difference becomes insignificant for adults at around 1.5 seconds after the start of the instruction, however. More interestingly, four-year-olds initially do not show a significant difference, suggesting they look more at the one bisyllabic option than at either of the two monosyllabic options (though the expected pattern emerges quickly). Taken together with the data from the trials with monosyllabic explicit objects, this means that four-year-olds' gazes were more likely to be on bisyllabic options than on monosyllabic ones overall.

This apparent preference for images with bisyllabic names may be caused by novelty or processing difficulty: if the four-year-old participants (on average) learned some of the nouns much more recently than others, the images representing the newer nouns would likely be more interesting to these four-year-olds. If unknown or recently acquired nouns are harder to process than familiar nouns, participants may gaze at images of unknown or more recently acquired nouns more. Eight-year-olds and adults, on the other hand, would not be expected to be affected by the relative novelty or difficulty of these common words, which would explain why they do not gaze at bisyllabic options more overall. The age of acquisition norms of Kuperman et al. (2012) were used to investigate these possibilities.⁹⁰ The mean Kuperman et al. (2012) ages of acquisition tend to be higher for the bisyllabic words used as options in this experiment than for the monosyllabic ones (see Fig. 5.55), but the difference is not significant (Welch's $t = 1.816$, $p = 0.08$).

The gaze data was also analyzed with the option nouns split by several age-of-acquisition thresholds (3, 4, 5, 6, and 7 as convenient steps and 4.38, the median age of acquisition of the nouns used in this experiment). Of course, these splits do not make for two evenly-sized groups of nouns: while length was controlled such that exactly half of the options across the whole experiment were monosyllabic and the other half bisyllabic, there were many more options with an age of acquisition greater than 3 than there were options with an age of acquisition of 3 and lower. Therefore, the null assumption that the confidence intervals of the SSANOVA model curves for the two groups should overlap if there is no significant preference for gazing at one group more than at the other does not hold for these splits.⁹¹ The resulting graphs (not shown)

⁹⁰Kuperman et al. (2012)'s norms are based on Amazon Mechanical Turk users stating what age they were (or think they were) when they understood a word. However, Kuperman et al. carefully removed unreliable responses, and their norms correlate very well with those developed in other studies. Moreover, their list of words is much more comprehensive than those of previous age-of-acquisition norming databases, and thus (unlike these previous studies) has norms for all of the words used in the present study.

⁹¹The assumption of equality does not even hold for the median split, as 4.38 is the median of the nouns'

can thus only be interpreted as showing tendencies. The curves for both groups are fairly flat in all of them, and the gap between them shrinks as the age-of-acquisition threshold grows. Crucially, no early peak like the ones on bisyllabic options in Figs. 5.53 and 5.54 is apparent in any of these curves, which suggests that those peaks are not caused by novelty or processing difficulty.

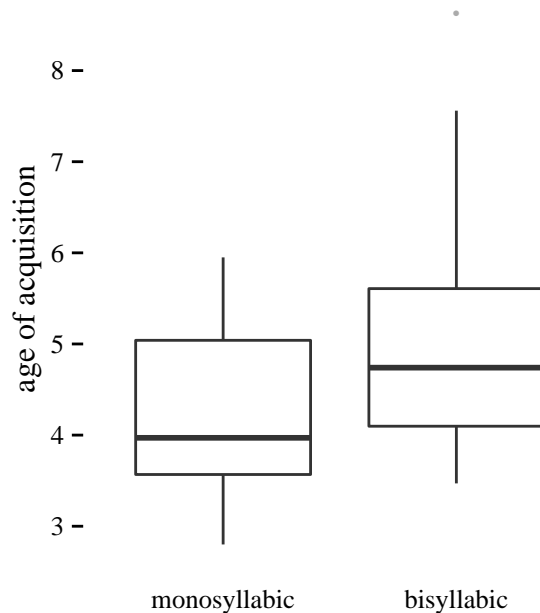


Figure 5.55: Kuperman et al. (2012)’s age-of-acquisition norms for the nouns used in experiment 2, by noun length (lower and upper ends of boxes indicate first and third quartiles, vertical lines extend to the most extreme value within 1.5 interquartile ranges of these quartiles, grey dots are more extreme values)

To test whether the early peak for bisyllabic nouns in four-year-olds’ gazes was carried by only a few of the bisyllabic nouns, further SSANOVA fits were computed, each comparing the gazes on the image for one specific bisyllabic noun (*basket*, for example) to the gazes on all other bisyllabic nouns (*baskets*, *hegehog*, . . .). Of course, the SSANOVA curves for the individual noun will be very low, as only a small percentage of all gazes in the entire experiment are expected to fall on the *basket* image specifically. If the peak was due to just one or a few specific nouns, their individual curves would be higher than those of the other bisyllabic nouns. No such difference is apparent in the graphs of these noun-wise SSANOVA models (not shown here), meaning there is no evidence that a preference for gazing at one or a few specific bisyllabic options caused the apparent bisyllabic preference in the four-year-old age group.

We see a similar initial preference for plurals in trials with singular explicit objects (data and age-of-acquisition values, but trials were not balanced for age of acquisition—in some trials, two or even all three options had ages of acquisition above this median value.

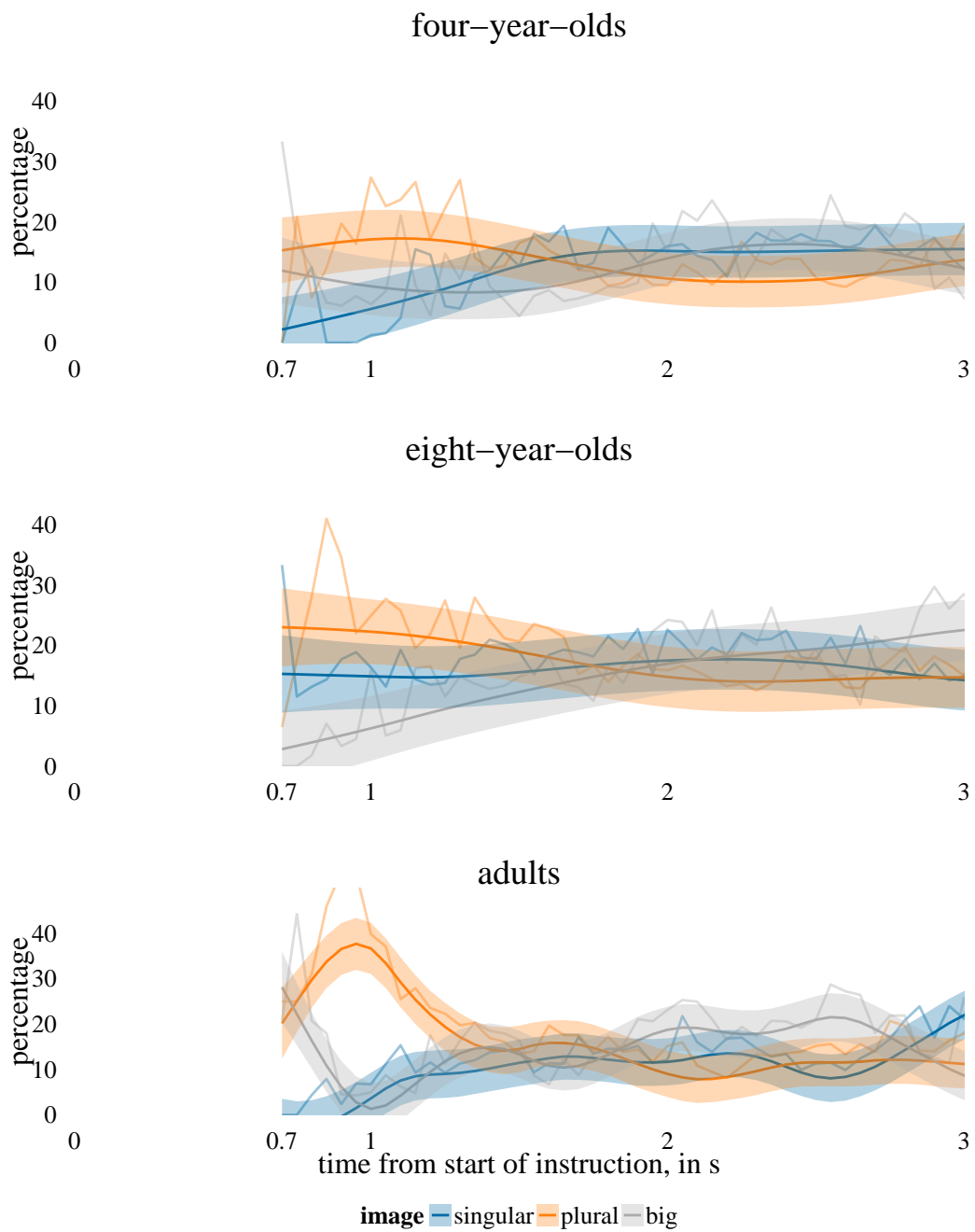


Figure 5.56: Percentages of gaze on images split by grammatical number in trials with singular explicit object (faded lines) and SSANOVA models with confidence intervals (lines with ribbons)

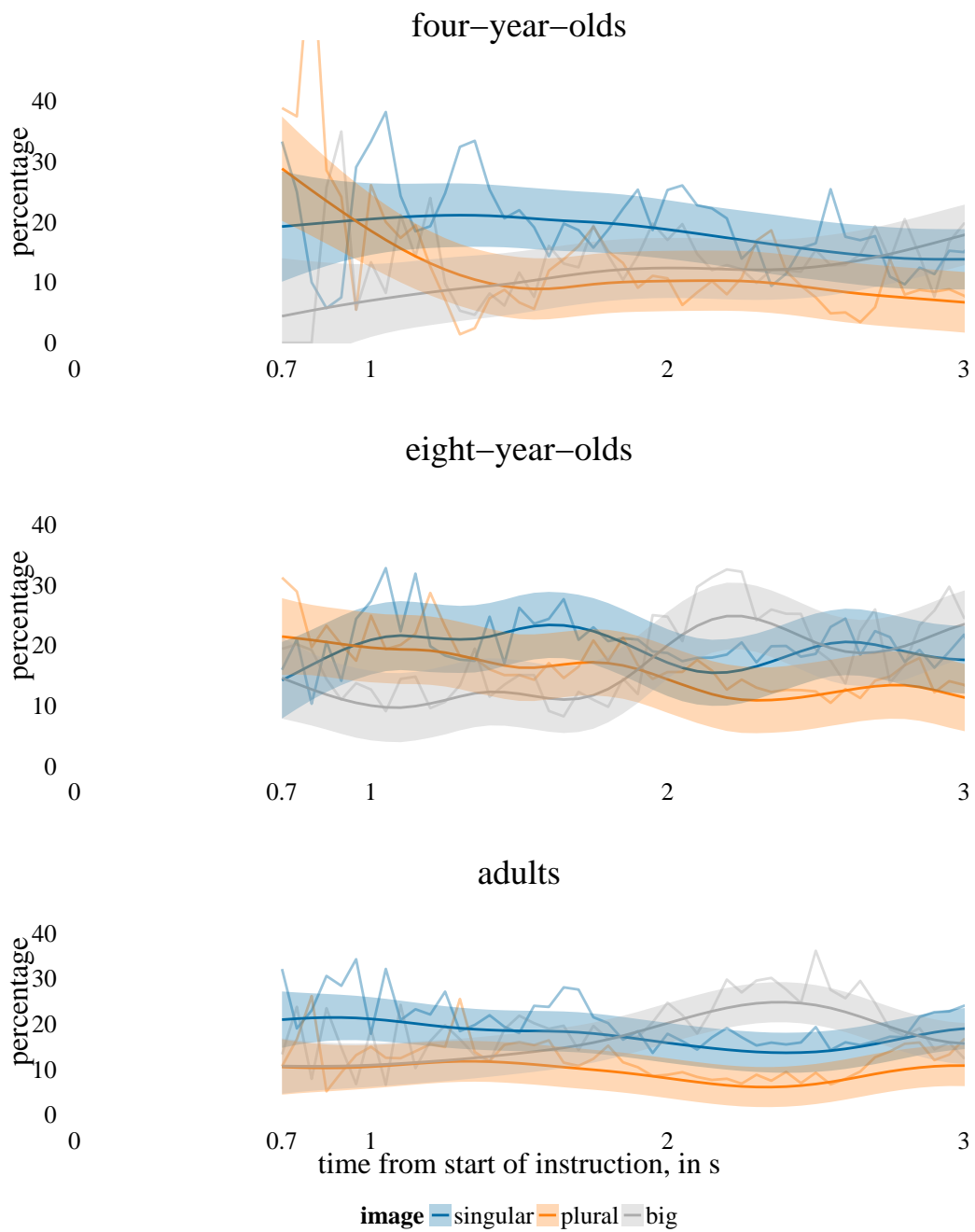


Figure 5.57: Percentages of gaze on images split by grammatical number in trials with plural explicit object (faded lines) and SSANOVA models with confidence intervals (lines with ribbons)

SSANOVA models shown in Fig. 5.56) as well as in trials with plural explicit objects (Fig. 5.57), where four-year-olds as well as eight-year-olds show either no difference or a preference for plural images around the 1-second mark. Strikingly, adults look at the plural options much more often than at the singular option in the singular-explicit trials around the 1-second mark, as the bottom panel of Fig. 5.56 shows. While this fits the expected pattern (since there are two plural options and only one singular one in each of these trials), this peak for plurals is the largest one in all the SSANOVA models shown here. It is difficult to test whether the percentage of gazes at the plural options conforms to the null hypothesis of being about twice as high as the percentage of gazes at the singular option. The magnitude of the difference here strongly suggests that it is higher, meaning that adults' gazes were strongly attracted to plural options in the presence of a singular explicit object.

The two plural options can be disentangled by dividing the three images by which feature of the explicit object ('big' image) they matched. Since this concerns only the trials with a singular explicit, the number-matching choice will be the singular one, and the animacy- and length-matching choices will both be plural. Fig. 5.58 shows the data divided this way (the same split as in Fig. 5.51, except that Fig. 5.51 included data from all trials; the lines and ribbon for the 'big' image are not shown in Fig. 5.58 for reasons of clarity). The blue and light grey lines and ribbons represent the two plural options, and there is a gap between these two ribbons and the orange one representing the singular up to about the 1-second mark. The two ribbons for the plural options, however, do overlap. In other words, adults looked at each of the two plural options significantly more than at the singular option in that time, but did not prefer one of the two plural options over the other.

Fig. 5.59 shows the SSANOVA model for gaze data from trials where the gap in the instruction sentence replaced the theme (see Section 5.1.2 for details), with the smaller images split by their animacy: the blue lines represent inanimate options, and the orange lines animate ones. While each individual trial had two options of one kind and one of the other, this was balanced across trials, meaning there were the same number of inanimate and animate options overall. The datasets contains roughly the same number of gaze samples from trials with animate explicit objects and from trials with inanimate explicit objects.⁹² Therefore, there should be no significant difference between the numbers of gazes at animates and inanimates if there is no preference for gazing at one of them. However, it appears that four-year-olds' gazes are attracted significantly more by the inanimate option or options than by the animate one(s) just before the 1-second mark, and that adults' gazes are attracted significantly more by the animate option(s) just after the 1-second mark. This pattern is only apparent when the options represent possible themes;

⁹²Specifically, gaze samples from animate-explicit trials make up 50.8% of the four-year-olds' gaze data, 48.2% of the eight-year-olds' gaze data, and 48.6% of the adults' gaze data.

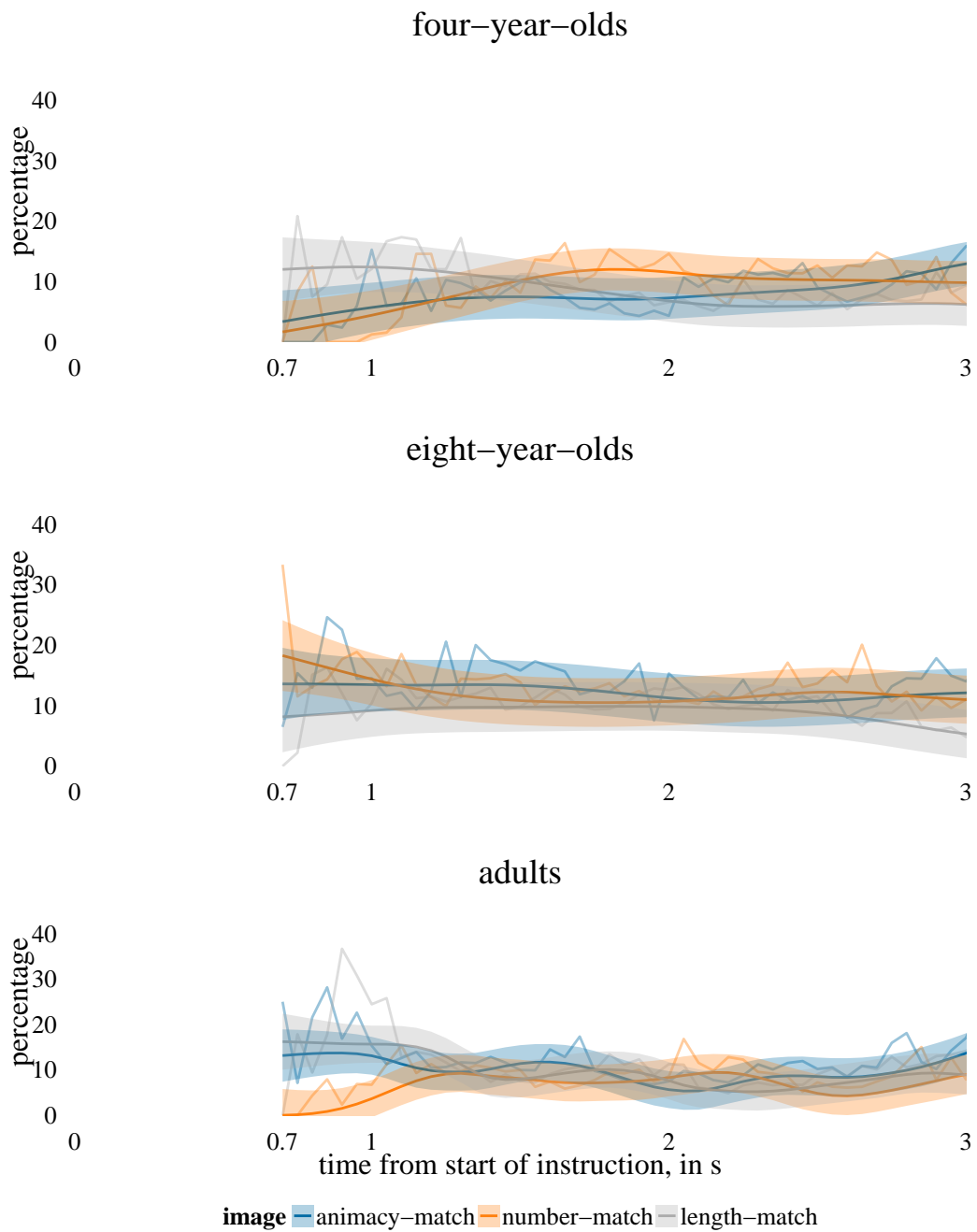


Figure 5.58: Percentages of gaze on images split by which feature of the ‘big’ image they matched in trials with singular explicit object (faded lines) and SSANOVA models with confidence intervals (lines with ribbons)

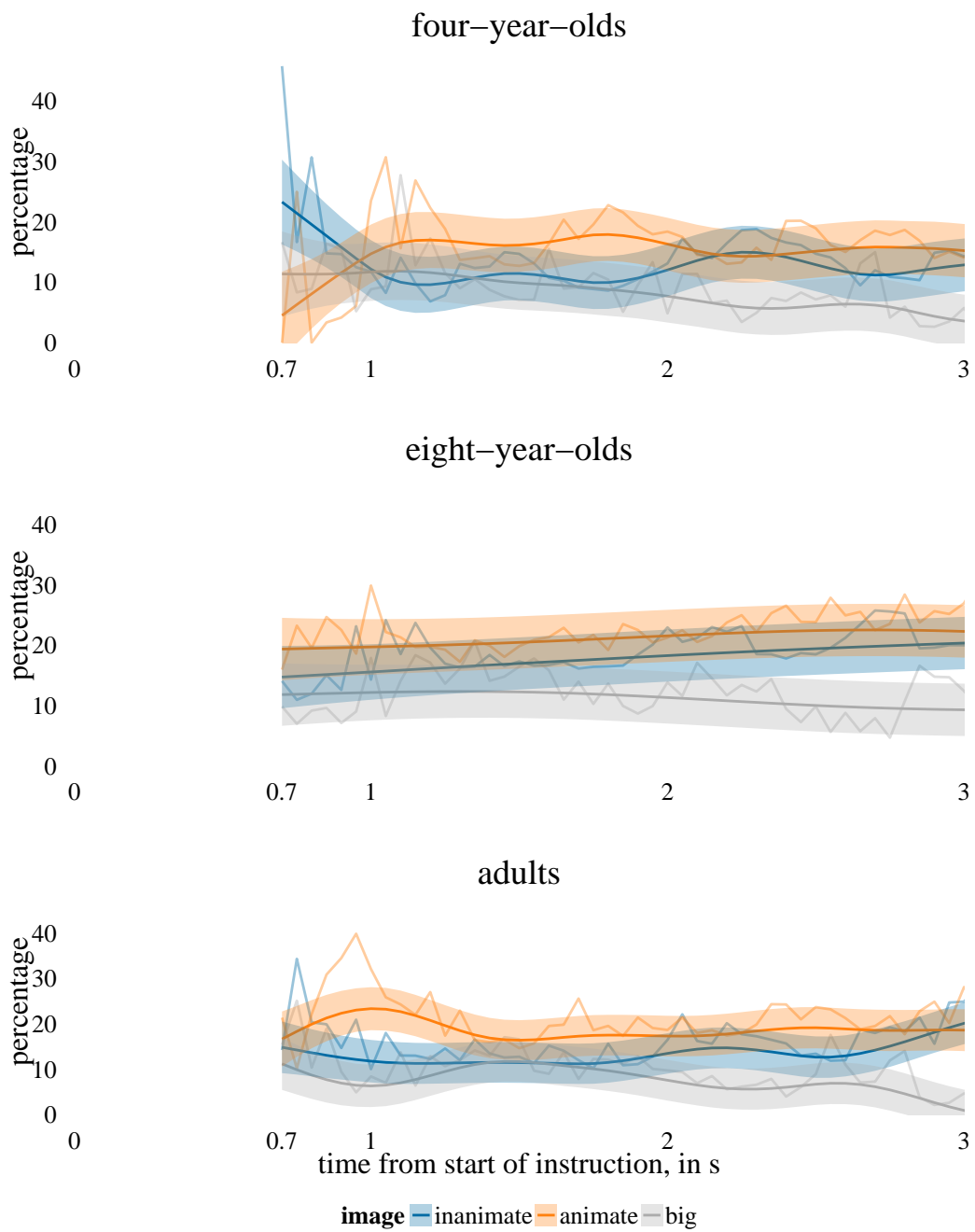


Figure 5.59: Percentages of gaze on images split by animacy in theme-gap blocks (faded lines) and SSANOVA models with confidence intervals (lines with ribbons)

the SSANOVA models for the goal-gap blocks (not shown here) do not suggest any differences between gazes on inanimate and animate images for any of the age groups.

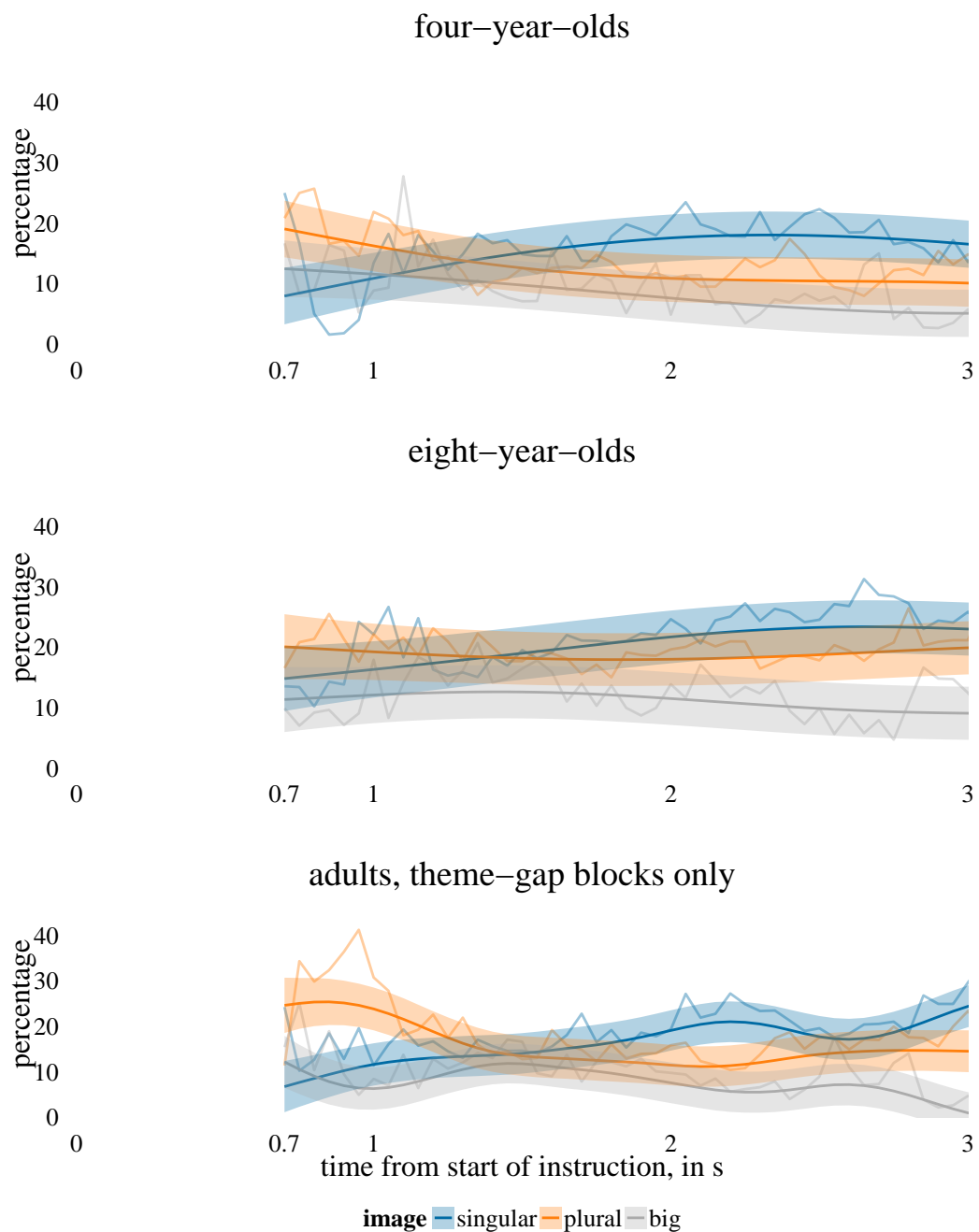


Figure 5.60: Percentages of gaze on images split by grammatical number in theme-gap blocks (faded lines) and SSANOVA models with confidence intervals (lines with ribbons)

Similarly, both four-year-olds and adults were more likely to look at images representing plurals than at ones representing singulars around 1 second after the start of the instruction sentence during theme-gap trials, as Fig. 5.60 shows. This preference was not found in goal-gap trials

(plots not shown here). Eight-year-olds' gazes, on the other hand, do not seem to be affected by number at any point in the time window studied here.

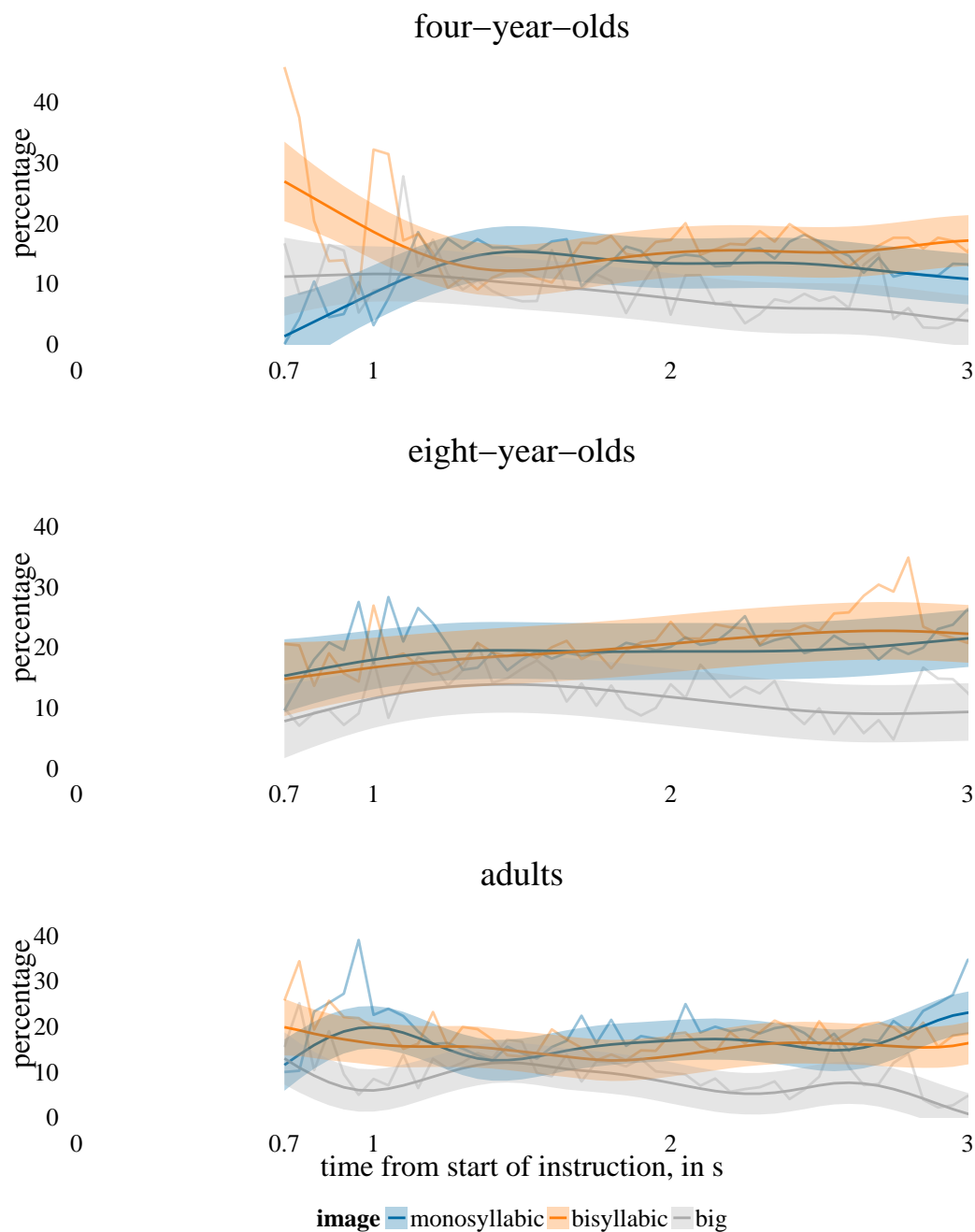


Figure 5.61: Percentages of gaze on images split by length in theme-gap blocks (faded lines) and SSANOVA models with confidence intervals (lines with ribbons)

In the same theme-gap trials and at around the same time, four-year-olds' gaze was also more likely to fall on images representing bisyllabic nouns than on monosyllabic ones (see Fig. 5.61). (Recall that the participants had heard audio stimuli of each of the three nouns before the

instruction sentence.) Eight-year-olds and adults did not gaze at bisyllabic images more than at monosyllabic ones, or vice versa.

None of these preferences reduce to a position effect: the three smaller images were randomly assigned to the three positions dynamically, and post-hoc analyses show no significant association between one feature and one position. Moreover, SSANOVA models (not shown here) do not indicate significantly more gazes on one of the three positions for any age group.

5.3.5 Gaze and touch

Gaze data samples with concurrent touchscreen input samples were separated from the data described above in order to test how closely related gaze and touch are. A pair of one gaze and one touch sample was judged to be concurrent if they were recorded immediately after one another. 98.2% of these pairs had the same timestamp (within one microsecond), and the largest time difference within a pair was 14.7 milliseconds. Since even this largest difference is very small, each of these pairs was accepted as one gaze-and-touch data point. This yielded 14174 data points from 16 four-year-olds, 14348 data points from 17 eight-year-olds, and 8376 data points from 21 adults.

The Euclidean distance between the gaze and touch positions was calculated for each of these data points with a gaze position on the screen (4811 data points, or 13%, had a gaze position outside the screen boundaries). Fig. 5.62 shows density plots of these distances for each of the three age groups (in orange, top three panels). It is immediately apparent that small distances are very common (and measures of the distribution of distance values support this: skewness is 1.23 and excess kurtosis is 1.73, indicating that the values are concentrated on the left). For comparison, a dataset was simulated with two points placed (uniformly) randomly in a window the same size as the one used in the experiment (1280 by 800 pixels). The distribution of these random points, shown in grey in the bottom panel of Fig. 5.62, is much less biased (skewness 0.38, kurtosis -0.49). This illustrates that concurrent gaze and touch positions are very close to each other.

To further establish this connection between gaze and touch positions, a linear regression model of gaze position (in x and y pixel coordinates) on touch position (also in x and y pixel coordinates) was fit to this data. It is conceivable that different age groups show differently strong connections between gaze and touch. To control for this, age group and interactions between age group and the two coordinates were also included in the regression model. This model (coefficients shown in Table 5.13) indicates that each of the two coordinates of the gaze position is correlated with the corresponding coordinate of the concurrent touch position. The interactions between age group and coordinates suggest that the strength of this correlation differs with different age

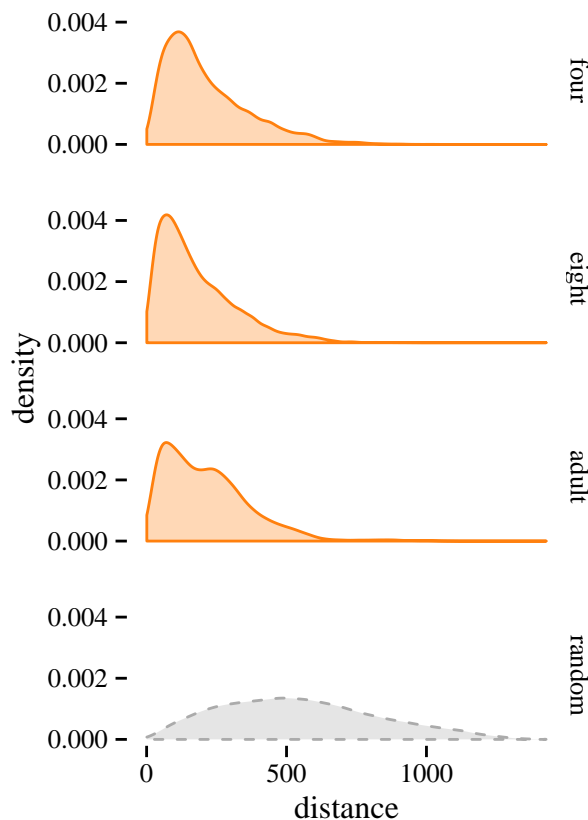


Figure 5.62: Density plots of the distance between concurrent gaze and touch position by age group (orange, top three panels), and the distance between two random points in the same window (grey, bottom panel)

groups, and since the main effect parameter is positive (0.81 in the x-coordinate sub-model) but the interaction effect parameters are negative (-0.12 and -0.28), the suggestion is that the correlation is weaker with eight-year-olds and weaker still with adults than it is with the reference group, the four-year-olds. Correlation statistics support this: Spearman's ρ for this correlation is smaller with the adult group (0.44) than with the eight-year-olds (0.60) and the four-year-olds (0.63).⁹³ The significant main effects of age group means that the age groups' gaze positions overall were different in both coordinates. The two-dimensional density plots of gaze samples by position in Fig. 5.63 show this very well: four-year-olds' gazes are spread widely, while adults' gazes are more concentrated. Surprisingly, the x-coordinate of the touch position is a significant predictor for the y-coordinate of the gaze position. However, the estimated effect strength of the touch x coordinate (0.13) is much smaller than that of the touch y coordinate (0.64), and that sub-model (for the y coordinate of the gaze position) is a worse fit to the data than the other sub-model (y-coordinate sub-model $R^2 = 0.06$; x-coordinate sub-model $R^2 = 0.23$).

⁹³I am not reporting p -values for these correlations because I consider them to be irrelevant: the regression model in Table 5.13a shows convincingly that the x-coordinates of concurrent touch and gaze are correlated; the point of the ρ statistics here is to compare the strength of this established correlation across age groups.

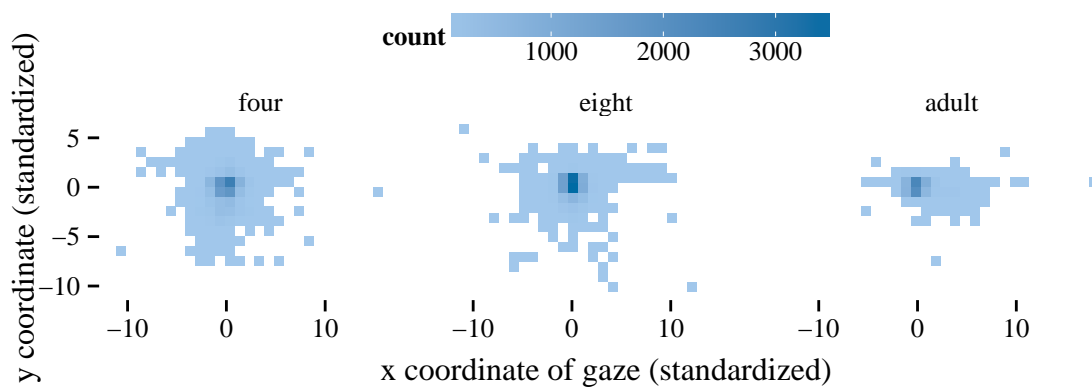


Figure 5.63: Position of gaze data samples by age group (standardized to an overall mean of 0 and an overall standard deviation of 1 for ease of comparison; darker cells contain more data samples)

variable	parameter estimate	standard error	<i>t</i>	<i>p</i>
(Intercept)	-36.83	2.93	-12.58	< 0.01
x coordinate of touch position	0.81	0.01	78.06	< 0.01
y coordinate of touch position	-0.01	0.02	-0.55	0.58
eight-year-olds	29.64	4.11	7.21	< 0.01
adults	47.92	4.71	10.18	< 0.01
touch x coord. : eight-year-olds	-0.12	0.02	-7.89	< 0.01
touch x coord. : adults	-0.28	0.02	-15.16	< 0.01
touch y coord. : eight-year-olds	0.05	0.03	1.88	0.06
touch y coord. : adults	0.23	0.03	7.28	< 0.01

(a) Coefficients for x coordinate of gaze position

variable	parameter estimate	standard error	<i>t</i>	<i>p</i>
(Intercept)	-133.89	3.50	-38.27	< 0.01
x coordinate of touch position	0.13	0.01	10.49	< 0.01
y coordinate of touch position	0.64	0.02	29.01	< 0.01
eight-year-olds	80.80	4.91	16.44	< 0.01
adults	79.82	5.63	14.19	< 0.01
touch x coord. : eight-year-olds	-0.22	0.02	-12.00	< 0.01
touch x coord. : adults	-0.20	0.02	-9.38	< 0.01
touch y coord. : eight-year-olds	-0.06	0.03	-1.66	0.10
touch y coord. : adults	< 0.01	0.04	0.13	0.90

(b) Coefficients for y coordinate of gaze position

Table 5.13: Coefficients for the regression model of gaze position (variables in bold are deemed to have significant effects; $MSE = 73575.2$, $_{100}M_{10} = 73612.8$)

The distribution of the distance between concurrent gaze and touch positions, and the regression of one on the other, show that gaze and touch were closely related or, in other words, that gaze and touch were often focused on the same region or image. While this finding is very clear, it is important to remember that the data supporting it are the time points for which both eye gaze and touch screen input were recorded. Eye gaze recording was far from continuous, as detailed above, and participants actively used the touchscreen only for a fraction of each trial, as described in Section 5.3.3. As the experiment required participants to use the touchscreen mostly for dragging, gaze and touch can thus only be said to be strongly correlated during that particular action.

Look-ahead gazes, meaning gaze behavior that anticipates later interaction, have been identified in very different tasks, such as walking (Patla and Vickers 2003), washing one's hands (Pelz and Canosa 2001), or replicating models using toy construction sets (Mennie et al. 2007). While the connection between touchscreen input and eye gaze has not been studied closely, it is reasonable to expect such look-ahead gazes in this experiment. The data show none, however: within each age group, the cross-correlations between the x coordinates of gaze and touch is strongest at a lag value of 0, meaning the concurrent gaze sample is a better predictor of a given touch sample than any preceding or following touch sample overall. The same is true for the y coordinates of gaze and touch. Thus, no 'look-ahead' behavior is apparent in the data for any age group. It must be stressed, however, that this task likely included several sub-routines, each of which may well have their own associated gaze behavior pattern. A look-ahead tendency in one of these patterns may thus have been drowned out by the absence of a look-ahead tendency in the other patterns. Therefore, this negative finding (that there were no look-ahead gazes in this experiment) must be treated with great caution.

5.4 Summary

Across all age groups, the touchscreen appears to have been easy to use. There were no complaints or other indications of problems. Dragging paths are very straight (Fig. 5.46) and were completed quickly (Fig. 5.47 and Table 5.12), although speed increases with age. When interacting with the touchscreen, participants tended to focus their gaze on the point of touch input (Section 5.3.5). This tendency is stronger for younger participants. Participants' choices out of the three options in each trial reveal several coexisting preferences: eight-year-olds and adults prefer to choose the option that matches the grammatical number of the explicit object (Fig. 5.41). Adults prefer inanimate options if the explicit object is also inanimate or if the gap in the instruction sentence was in place of the theme object (Fig. 5.44). Eight-year-olds, on the other hand, prefer inanimate choices when the explicit object is animate (Table 5.10). There is some

evidence that eight-year-olds also prefer monosyllabic options when the gap in the instruction sentence followed a bisyllabic explicit object (Table 5.9). Similarly, participants in all age groups show a preference for animate-before-inanimate ordering (Table 5.10): inanimate choices were favored more when the gap followed an animate explicit object than when it preceded that object, or conversely animate choices were favored when the gap preceded an inanimate explicit object than when it followed it. The adult participants also showed strong patterns in what images they gazed at: several seconds before making the choice, adults gazed at their eventual choice image more than at the other two options (Fig. 5.50). They also looked more at the option that matched the explicit object (and big image) in animacy than at the other two options, particularly during trials with an animate explicit object (Figs. 5.51 and 5.52). When the explicit object was singular, adults looked at the plural options more than at the singular one (Fig. 5.56). In trials where the gap in the instruction sentence had a gap in place of the theme object, adults were more likely to look at animate options, and four-year-olds were more likely to look at inanimate options (Fig. 5.59). Both four-year-olds and adults also gazed at plural options more than at singular ones in these theme-gap trials (Fig. 5.60), and four-year-olds showed a similar preference for bisyllabic options over monosyllabic ones (Fig. 5.61).

Per-participant analyses replicate the a priori animacy categorization and do not reveal any differences in this study, except in one case—adults dragged images of animals or inanimate objects slightly slower if they had answered “yes” to the question “Could this one move towards you?” for that animal or inanimate object. This could be explained by a subconscious understanding that entities capable of movement need less ‘help’ in moving, but also by a feature of the task design: in all trials, the image being moved implicitly filled the position of the theme in the instruction sentence. The prototypical animacy pattern for *give* is an animate recipient being given an inanimate theme. Thus, motion-capable entities (being somewhat like the prototypical animate entity) make somewhat worse themes. This misfit may be argued to cause a slight hesitation when dragging these motion-capable entities. However, it would be reasonable to expect a similar hesitation effect on the reaction time, and a linear model (not shown here) shows that this effect is only significant among the four-year-old age group there. As even this significant effect is quite small, and the hesitation apparent in adults’ dragging speed is also minimal, there is no strong evidence for this idea that the lack of fit to the prototypical animacy for themes made participants drag motion-capable themes slightly slower. These minor differences in speed that were discovered incidentally without being a specific aim of this study may thus well be a spurious finding.

Four-year-old participants dragged images more slowly than the older participants did, but this small difference could reflect motor development: four-year-olds are undergoing not only outwardly visible physical growth, but also neurological development (Berk 2013). This means

that they are less able and less skilled at performing physical tasks than older children and adults are, which explains the dragging speed difference. Note also that the speed difference was fairly small in absolute terms: the median length of a dragging path was 264 ms among the adult age group and 835 ms among the four-year-olds. Children are certainly slower, but they still finish the simple dragging path in less than a second. Moreover, nothing suggests that the touchscreen itself was difficult to use for even the youngest participants. Therefore, this speed difference probably indicates a general developmental difference rather than a linguistic one.

6 Experiment 3: Elicitation

The aim of the experiment presented in this chapter is to address the production aspect of first two research questions of Chapter 3: Do speakers produce animate-before-inanimate and plural-before-singular order? Participants are asked to reproduce sentences which systematically vary animacy and number of the two objects as well as their order. Various aspects of their productions are used to assess which configurations are produced, which provides insight into participants' knowledge of ordering preferences and their interactions with selectional preferences.

6.1 Methodology

6.1.1 Participants

The participants of experiments 1 and 2 (22 adults, 20 eight-year-olds, and 20 four-year-olds) all participated in this experiment as well.

6.1.2 Procedure

This experiment followed immediately after experiment 2 in the same session. It used drawings with priming prompts to elicit *give* sentences from participants: they were shown a sequence of 24 drawings (not randomized) and played pre-recorded sentences describing the action depicted in each drawing (see Appendix E for these drawings and sentences). These drawings were shown on-screen, and the participants heard the accompanying sentences over headphones.⁹⁴ Participants were asked to repeat each sentence to a stuffed toy (see Figure 6.1), and the experimenter then advanced the presentation program (programmed and run in PsychoPy, version 1.80.00, like experiment 2) to the next drawing and sentence.

Participants were told that this toy was an alien. This was done because they otherwise might not have repeated some of the sentences at all, as they were too odd: the third sentence in this experiment, for example, was (6.1).

(6.1) Mom gave the cushions Anne.

As the accompanying picture (Fig. 6.2) shows, *Anne* was the intended theme and *the cushions* the intended recipient here. The 'oddness' obviously stems from the fact that animates (like

⁹⁴The screen and headphones used were the same as in experiment 2. The speaker who was recorded saying these sentences was the same female New Zealand English speaker that had been recorded for the nouns and instructions in experiment 2.



Figure 6.1: Stuffed toy used as ‘addressee’ in experiment 3



Figure 6.2: Drawing used to illustrate *Mom gave the cushions Anne*.

Anne) are prototypical recipients and inanimates (like *the cushions*) prototypical themes of *give*. These odd sentences had to be used, however, since they are one of the possible combinations of dative alternation construction and object animacy and the aim of this experiment was finding the effect of order, regardless of the construction used or the role of either object.⁹⁵ Models

⁹⁵Another possible source of oddness in the design of the sentences in this experiment (see Appendix E) was the repeated use of the phrase *the parents*. It could be argued that *parents* is typically used with a possessive, but

of the dative alternation choices made by adults (see Section 2.1) appear to show a preference for the realization that places an animate object before an inanimate one and a plural object before a singular one. These two⁹⁶ ordering principles are in conflict when one object is an animate singular and the other an inanimate plural, as in (6.1). This sentence and others that violate one or both of these principles were used to see which principle was easier to violate. (6.1) has an inanimate plural object before an animate singular one, meaning it violates the animate-before-inanimate principle, but not the plural-before-singular one. The twenty-third sentence, *Ben gave the cat the crackers*, reverses these features: it has an animate singular object before an inanimate plural one, meaning it violates the plural-before-singular principle, but not the animate-before-inanimate one. Some sentences (like the seventh one, *The cat gave* [recipient *the basket*] [theme *the kittens*]) violate both principles. The fourth and final option, sentences that violate neither of these principles (for example the sixteenth one, *Mom gave the children to the sofa*), was also included.

Participants' speech during this experiment was recorded, and I later transcribed these recordings. These transcripts were then used to statistically analyze the choice of construction and the likelihood of correct repetition based on the animacy and grammatical number of objects as well as the construction modelled in the pre-recorded sentence. Because violation of the principles may also make sentences more difficult to process, three measures of processing difficulty were also considered in the analysis: response time, disfluencies, and comments or other reactions indicating that a participant found a sentence odd or unacceptable.

This third experiment was the last one in an experiment session. Participants and parents were given a further information or 'debriefing' sheet and the chance to ask questions as well as withdraw their consent. They received their incentive (vouchers or toys, see Section 4.1.1) and remuneration for parking fees (where applicable) together with a small card containing the researcher's contact details in case of further questions. Parents of child participants were also offered small cards with a brief invitation to the experiment, to give to other parents of suitably aged children who they knew. This concluded the experiment session.

This experiment was approved by the Human Ethics Committee of the University of Canterbury (reference number HEC 2013/166).

not with *the*. Possessives may be more difficult to process than *the* because they require a referent, the possessor. Other phrases used here (like *the cat* or *the blocks*) are more common with *the*. To get consistent object phrases without any differences in processing requirements, *the* was chosen for all phrases (except proper names) because it is unnatural with fewer of the nouns.

⁹⁶The length-based ordering principle ('short before long') is well supported in the literature on the dative alternation and other optional ordering phenomena, which is why it was **not** investigated in this third experiment (by systematically varying object length).

6.1.3 Limitations

As discussed in Section 4.1.3, it is conceivable that recruiting participants through schools could bias the sample of parents and children. The one obvious possible factor here would be ‘openness’ in general—perhaps parents who are more likely to agree for their children to participate in research also influence their children to be more open and talkative in general. This, however, would only matter in a relatively unrestricted production task. The picture-based elicitation technique used in experiment 3 is very restrictive, as it is designed to elicit a few specific sentences; individual differences in talkativeness should matter relatively little. Only two of the youngest participants did not cooperate during this experiment, and the rest did not exhibit major differences in openness and talkativeness during it. Therefore, though the sample may be biased, this bias is not expected to affect the data collected in experiment 3 either.

In sentence imitation, a strong priming effect from the prompt or target sentence cannot be avoided. Experiment 2 minimizes these unwanted effects of priming: using the same verb throughout the entire experiment means that lexical priming (*give* being preferred with one or the other construction) was the same for all items and will thus not distort the results (Ivanova et al. 2012). Structural priming is still a possibility, of course: if the first trial was a sentence using the double object construction, for example, this would still be in very recent memory during the second trial and could thus make double object responses more likely during the second trial. However, in this elicitation experiment, the items were presented in an order that alternated between one double object and one prepositional construction (see order of presentation in Appendix E). Therefore, any structural priming would always work in favor of the construction **not** used in the current trial’s target sentence. While the target sentences alternated between the prepositional and the double object construction, participants’ production may not always follow this and may thus be seen to be affected by participants’ own recent constructions. However, priming has been shown to be effective regardless of who used the priming construction (Bock 1986, Gries 2005, Bresnan et al. 2007, Shimpi et al. 2007, de Marneffe et al. 2012). I therefore assume that the most recent prime, the elicitation prompt, will have the strongest priming effect precisely because it was the most recent one. In this way, the problem of uncontrolled ‘self-priming’ was avoided. Thus, the effect strength and direction of structural priming are the same for each trial, and no confounding effect on the results is to be expected (compare Jaeger and Snider 2013:68–73).

As in experiment 2, the verb *give* is arguably the best choice for the target sentences in experiment 3: actions of giving are easy to represent in drawings and are easily understood from these pictures as giving actions. Furthermore, the very high frequency of *give* in child speech as well as child-directed speech (Gropen et al. 1989, de Marneffe et al. 2012) means that other

transfer verbs might be replaced by *give* in participants' productions anyway, just as in Stephens (2010:134, 154). This means that *give* makes this elicitation experiment as easy and naturalistic as possible while still producing the desired data.

Some of the target sentences used in this experiment are rather odd, as they contain animate theme objects being 'given' to inanimate recipients (for example *Mom gave the cushions Anne*). This is by design, as it allows for statistical analysis to reveal whether animacy patterns affected how (or whether) participants reproduced these target sentences.

Children are quicker than adults to lose attention to experimental tasks. The 'dry runs' I did myself when designing and programming the session consisting of experiments 1, 2, and 3 took less than 40 minutes each, but of course child participants cannot possibly be expected to be as quick or concentrated on these experiments. In fact, some of the youngest participants grew bored, restless, and even uncooperative as the session went on. Experiment 1 ("Could this one move/play?") was expected to not bore children, and it did turn out to maintain their interest. Experiment 2 was less simple, of course; however, it was designed with very few repetitions (only two per possible combination of factors) and had ample pauses between the four blocks. Therefore, it was as accommodating to young children's attention span as was possible within the confines of stringent research design. The pre-experiment information sheets as well as the experimenter's instructions also presented the experimental task as a game, and the fact that it was done on a computer (with the perhaps unknown but certainly intuitive touchscreen as an input device, no less) only helped this framing of the experiment. Experiment 3, finally, was certainly not the most exciting, but as even very young children frequently engage in free and guided story-telling (Foster-Cohen 1990:127–128), it was expected that they see it as a story and thus not get bored. The drawings provided enough structure to elicit useful reproductions from 18 of the 20 four-year-old participants, all 20 eight-year-olds, and all 22 adults.

6.2 Results

Two participants in the four-year-old age group did not cooperate during this experiment (likely due to fatigue, as it was the last experiment in sessions that lasted up to an hour for some participants). Other than that, participants appeared to have no problems with the task as such. This section presents the results of this experiment as analyzed by five measures: construction used, accurate reproduction of the target sentence, time to begin reproduction, disfluencies, and indications that participants found a sentence odd. The effects of age group and the objects' animacy and number on these measures are investigated to determine whether they reveal more specific patterns of difficulties with processing or (re-)production.

6.2.1 Construction

One of the most basic measures of participant behavior in this experiment is the type of construction used. Half of the target sentences used *give* in the double-object construction, the other half used the prepositional construction. The participants' productions for each trial were manually transcribed and subsequently tagged for the construction used and various features of the two objects. In 58 trials, the participant did not produce a sentence at all, and with a further four, it was not possible to tell with certainty whether the participant used the prepositional or the double-object construction (due to technical problems with the recording equipment). These 62 trials were excluded from the analysis in this section. Fig. 6.3 shows the percentages of constructions used, split by age groups and further by the construction in the target sentence. (The raw counts and totals differ between age groups: 18 four-year-olds, all 20 eight-year-olds, and all 22 adults cooperated during this experiment, although some trials across age groups were removed as described.) The difference between the double object and prepositional target sentences is obvious, as are the differences between age groups: prepositional targets almost always elicited prepositional productions. Adults use the double object construction for almost all double object targets, but the children use the prepositional construction for those too, with four-year-olds preferring it more than eight-year-olds do.

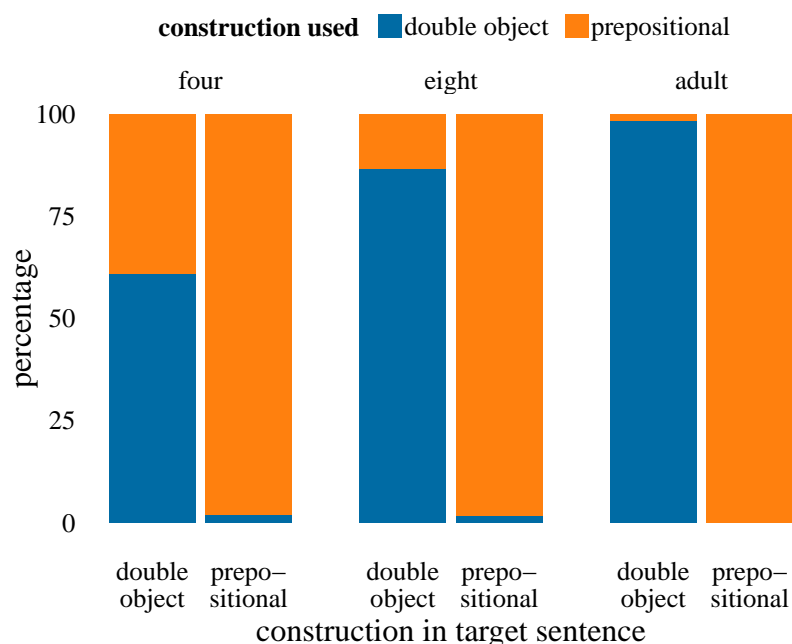


Figure 6.3: Number of sentences by dative constructions

In 110 trials, the participant used a different construction than the target sentence, but still produced a useable response. 102 of them had the double object construction in the target sentence (and thus the prepositional one in the reproduction; these make up the orange segments

	singular recipient	plural recipient
singular theme	16	31
plural theme	34	23

(a) Grammatical number in target sentence

	animate recipient	inanimate recipient
animate theme	15	61
inanimate theme	11	15

(b) Animacy in target sentence

Table 6.1: Features of the objects in the target sentence in trials with double-object target sentences and prepositional reproductions

of the first, third, and fifth bar in Fig. 6.3). They are fairly balanced for the number of the target objects (see Table 6.1a), but show an interesting imbalance for animacy (Table 6.1b): more than half of these trials had an animate theme and an inanimate recipient in the target sentence (Fisher’s exact test confirms this to be significantly different from an evenly-distributed table: $p = 0.03$). In all but two of these 61 cases, the participant effectively retained the order of objects, but inserted the preposition *to* and thus changed the functions of the two objects: (6.2a–6.5a) are the four target sentences with an inanimate recipient and an animate theme in the double object construction, and (6.2b–6.5b) show the pattern of participants’ production in 59 of these 61 trials. These four are all of the sentences with this specific combination of construction and animacies; no other target sentence shared this combination. The 61 trials where participants changed the sentence to the prepositional construction are distributed among the four sentences almost perfectly evenly, and no other sentence elicited a similar amount of it. Therefore, it appears that this change is related to this specific combination of the double object construction, an inanimate recipient, and an animate theme—this combination is evidently harder to reproduce accurately. 24 of the other 41 trials with a double-object target and a prepositional reproduction (Table 6.1) also had this *to*-insertion, 13 saw meaning-preserving construction changes (*Ben gave the kitten to the parents* for the target *Ben gave the parents the kitten*), and the remaining 4 had miscellaneous errors and interruptions.

- (6.2) a. Mom gave the cushions Anne.
b. Mom gave the cushions to Anne.
- (6.3) a. Mom gave the basket the kittens.
b. Mom gave the basket to the kittens.
- (6.4) a. The parents gave the chairs the children.

- b. The parents gave the chairs to the children.
- (6.5)
- a. The cat gave the milk the kitten.
 - b. The cat gave the milk to the kitten.

A generalized linear regression model was fit to the data from the remaining data to model how the construction a participant used was affected by age group, the construction used in the target sentence, the animacy and number of both objects the participant used in their production, and the interaction between age group and all other variables named. Table 6.2 shows the coefficients of this model, rounded to two digits. Variables with positive parameter estimate values favor the prepositional construction, and conversely variables with negative parameter estimate values favor the double object construction. (Note that this model was not pruned, for reasons laid out in Section 5.2.5.3.)

variable	parameter estimate	standard error	<i>z</i>	<i>p</i>
(Intercept)	0.10	0.39	0.26	0.79
eight-year-old age group	-3.12	0.74	-4.22	< 0.01
adult age group	-80.64	3196.15	-0.03	0.98
prepositional target sentence	4.50	0.55	8.24	< 0.01
inanimate theme produced	-0.27	0.33	-0.83	0.41
plural theme produced	-0.25	0.32	-0.77	0.44
inanimate recipient produced	-0.80	0.34	-2.36	0.02
plural recipient produced	> -0.01	0.32	> -0.01	1.00
eight-year-olds : prepositional target	2.41	0.88	2.73	0.01
eight-year-olds : inanimate theme	1.88	0.65	2.91	< 0.01
eight-year-olds : plural theme	0.40	0.50	0.79	0.43
eight-year-olds : inanimate recipient	0.03	0.57	0.06	0.95
eight-year-olds : plural recipient	0.16	0.50	0.31	0.75
adults : prepositional target	79.96	2902.41	0.03	0.98
adults : inanimate theme	31.73	1566.18	0.02	0.98
adults : plural theme	31.64	1557.93	0.02	0.98
adults : inanimate recipient	16.17	873.03	0.02	0.99
adults : plural recipient	15.38	873.03	0.02	0.99

Table 6.2: Coefficients for regression model of the construction used in participants' productions (positive parameters indicate effects favoring the prepositional construction, variables in bold are deemed to have significant effects; $MSE = 0.056$, $_{100}M_{10} = 0.058$)

Unsurprisingly, the construction in the target sentence had a strong effect: prepositional target sentences elicited more prepositional responses, and conversely double-object targets elicited mostly double-object responses. This model takes the four-year-old age group as the reference level for age. This means that the negative effect associated with eight-year-olds shows that eight-year-olds were significantly less likely to choose a prepositional construction than four-year-olds were—in other words, the apparent preference for the prepositional construction is strongest for the four-year-olds. The strikingly large negative effect associated with adults does not reach

significance under the standard assumptions (as detailed in Section 5.2.5) because its standard error is even larger. I suspect that this is because the produced construction is almost perfectly determined by the target construction within the adult age group: ignoring missing data, the target and produced construction do not match in only 4 out of 522 trials (or 0.8%). Within both child age groups, however, there is much more variation. Against that varied baseline, the complex regression model presented here is unable to quantify effects in the neater data from adult participants with certainty, leading to large standard errors for the main effect of adults (third row from the top in Table 6.2) and all interactions with it (bottom five rows). A model with the same formula was also fit to only the four- and eight-year-olds' data. The differences between its coefficients (not shown here) and those in Table 6.2 are small, and there are no differences in significance.

Inanimate recipients are significantly associated with the double object construction. As this construction puts the recipient after the theme, this effect reflects previous corpus-based findings of an animate-before-inanimate preference. Finally, the interactions between eight-year-olds on the one hand and target construction or theme animacy on the other hand are significant, both with a positive parameter estimate. The first of these interactions is apparent in Fig. 6.3: while the eight-year-olds produced fewer prepositional constructions than the four-year-olds (relatively speaking), they did still prefer the prepositional construction very strongly when the target sentence was also prepositional. In other words, this positive interaction effect of eight-year-olds and prepositional targets adjusts for the negative main effect of eight-year-olds, which by itself models the drop in prepositional constructions when moving from the reference baseline of four-year-olds and double-object targets to the eight-year-olds and double-object targets. The other interaction (eight-year-olds : inanimate theme produced) is more straightforward to interpret: for eight-year-olds, the animacy of the theme and the construction used are more strongly correlated than for the four-year-olds. The parameter estimate is positive, meaning the eight-year-olds apparently preferred the prepositional construction with an inanimate theme and the double-object construction with an animate one. In part, this reflects the *to*-insertion pattern discussed above, which was strongest with inanimate recipients for animate themes: the productions with an inserted *to* effectively turned the inanimate recipients into inanimate themes. 23 of the 61 responses with this *to*-insertion were produced by eight-year-olds, which makes them stand out against the otherwise mostly accurate reproductions by eight-year-olds. 36 of the *to*-insertion responses came from four-year-olds, but they had much more varied productions, and I assume this is why this sub-pattern does not make the corresponding effect significant for four-year-olds in the present regression model.

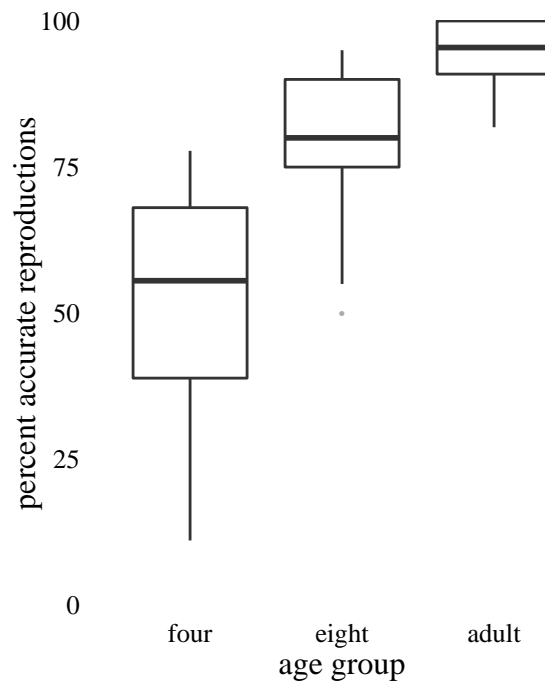


Figure 6.4: Percentages of accurate reproductions of target sentences by age group (lower and upper ends of boxes indicate first and third quartiles, vertical lines extend to the most extreme value within 1.5 interquartile ranges of these quartiles, grey dots are more extreme values)

6.2.2 Accurate reproduction

Based on the audio recordings of each trial, I noted whether the participant had accurately reproduced the target sentence. If a participant repeated one or more words in an otherwise accurate reproduction or if they revised their production to an accurate reproduction, that trial was not counted as an accurate reproduction, because the accuracy of reproduction was intended to measure difficulty in reproduction here and these types of disfluencies have been connected to processing difficulty in children (Buhr and Zebrowski 2009). Therefore, counting disfluent reproductions as accurate would mean counting trials where participants evidently had difficulties as free of difficulty. Otherwise accurate reproductions with pauses were counted as accurate reproductions, as were reproductions that were accurate considering systematic participant idiosyncrasies: for example, it was obvious from all of one participant's recorded trials that they always produce [d] in the place of /g/ and [t] in the place of /k/. *Dad dave Anne the toat* was counted as an accurate reproductions of the target *Dad gave Anne the coat* for this participant. There were 277 trials with inaccurate reproductions in total. The construction changes discussed above are among them, of course, as are many trials with repeated words in an otherwise accurate reproduction and trials where (four-year-old) participants did not cooperate. In some trials, participants changed words in the target sentence to phonologically close ones (like *Dan* for target *Dad*) or morphosyntactic alternatives (like *a* for target *the*). A few trials saw

participants producing just one object of the target sentence. In only 3 trials, participants used a verb form different from the ubiquitous target *gave* (two *carried*, one *brang*, all three usages ditransitive).

The percentage of correct reproductions was then calculated for each sentence and age group. Fig. 6.4 shows a summary of these percentages. The difference between the age groups is striking: adults performed very well on all sentences; eight-year-olds were less accurate on average and show much more difference between sentences; and four-year-olds are less accurate again and showed even more variation. I take this to mean that, at least for the four- and eight-year-olds, the stimulus sentences were too long for simple ‘parroting-back’ or verbatim repetition (Lust et al. 1998:73) and thus were re-produced in the strict meaning of the term. The adult participants may have been able to hold an entire target sentence in working memory and repeat it from working memory without processing (Vinther 2002).

Fig. 6.5 shows the same data, with two lines for the mean percentages of accurate reproductions depending on which dative construction was used in the target sentence. There is an obvious upwards slope with increasing age for both lines, but the blue line (showing mean percentages of correctly reproduced double object sentences) is lower than the orange line (mean percentages of correctly reproduced prepositional sentences) for all age groups. All three age groups were more likely to accurately reproduce prepositional sentences than double-object sentences, and the difference is particularly striking for four-year-olds.

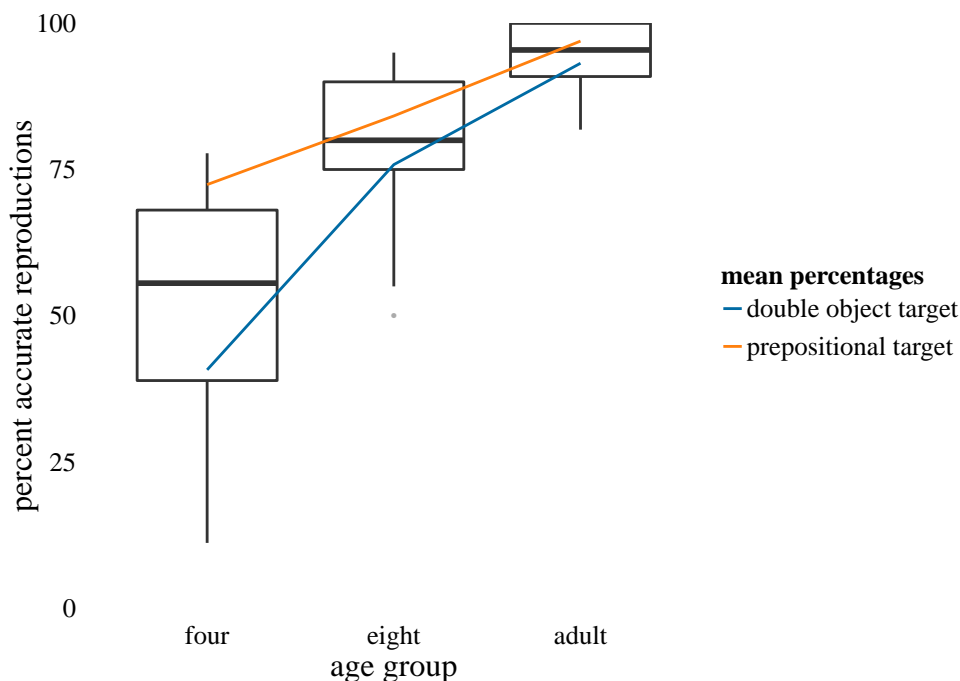


Figure 6.5: Percentages of accurate reproductions by age group (boxes), with mean percentages for construction in the target sentence (lines)

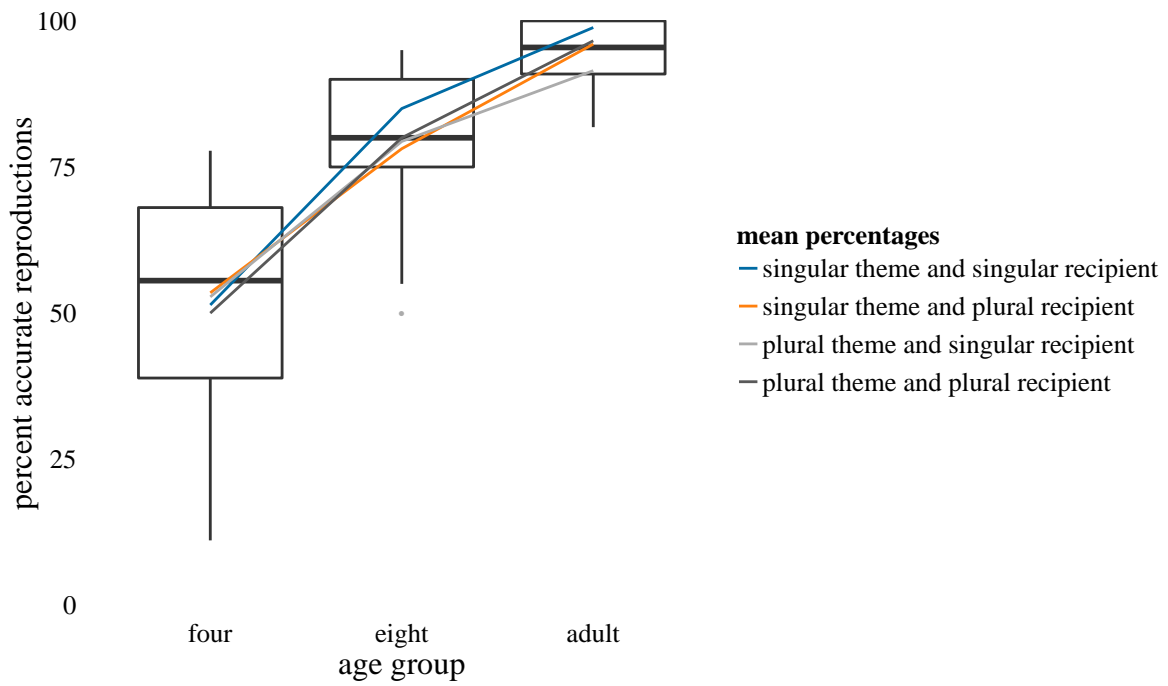


Figure 6.6: Percentages of accurate reproductions by age group (boxes), with mean percentages for combinations of number (lines)

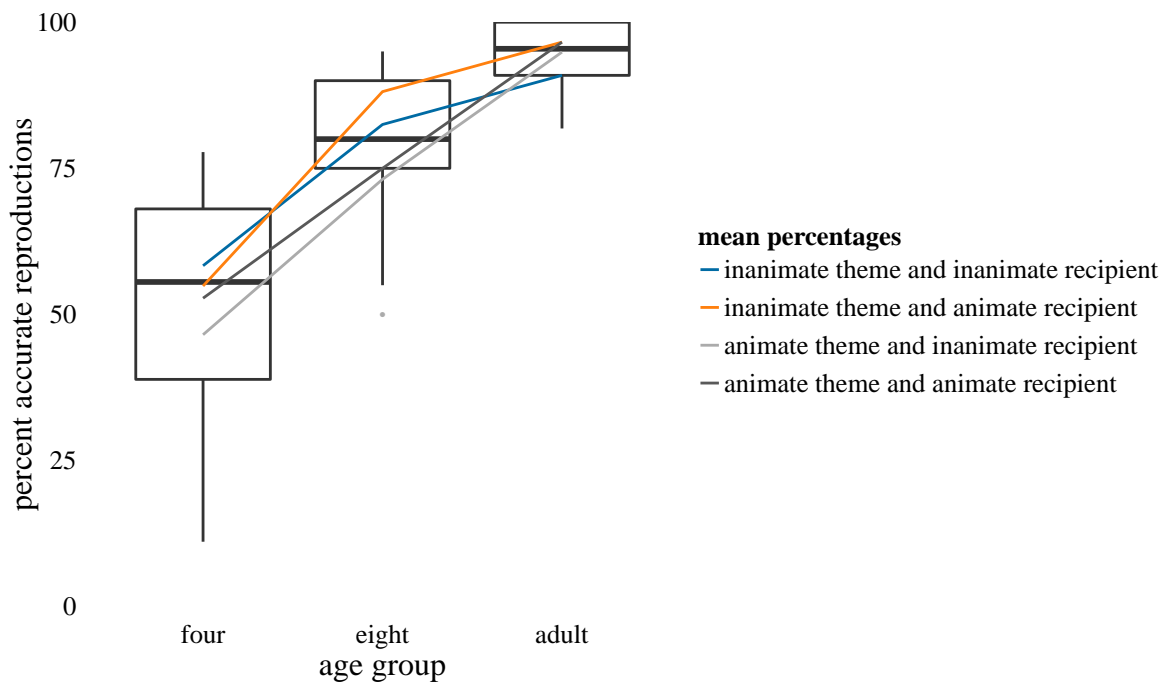


Figure 6.7: Percentages of accurate reproductions by age group (boxes), with mean percentages for combinations of animacy (lines)

In Figs. 6.6 and 6.7, the mean percentages for different combinations of object number animacy have been added as colored lines (the boxes showing the overall percentages are the same as in Fig. 6.4). The lines in Fig. 6.6 form a very tight bundle, meaning there is very little difference between the mean percentages of accurate reproductions for the four combinations of number. In Fig. 6.7, on the other hand, the lines are more spread out, suggesting that animacy may be a source of systematic variation here. Fig. 6.8 shows the same data with colored lines for the different orderings of animacy. It suggests that children reproduced sentences with an animate object as the second object (orange and dark grey lines) less accurately than sentences with an inanimate second object.

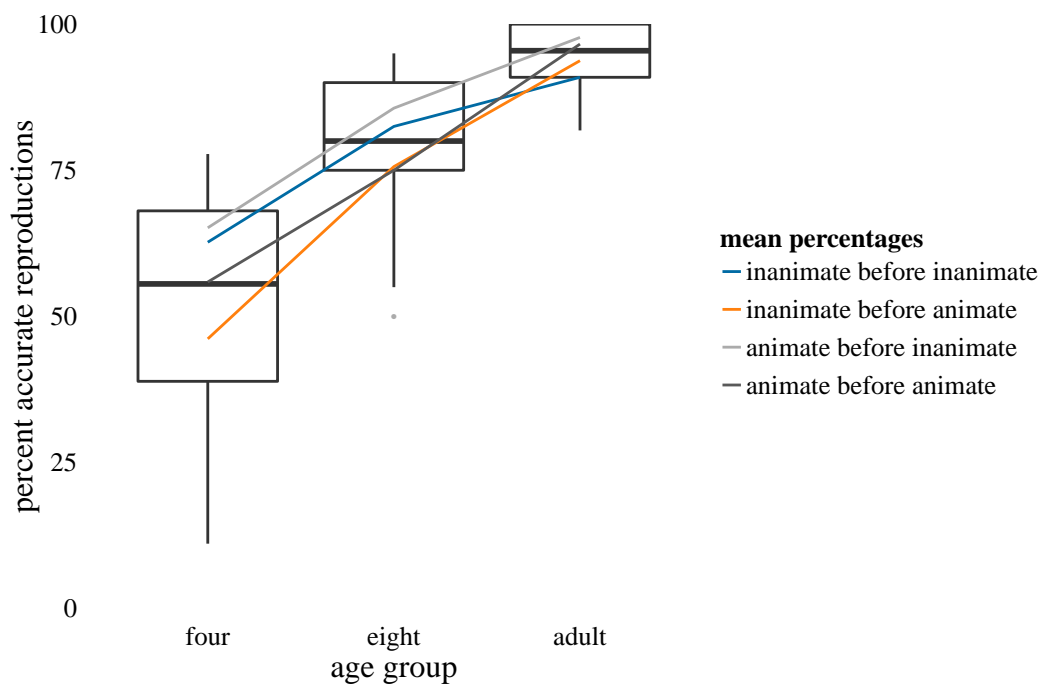


Figure 6.8: Percentages of accurate reproductions by age group (boxes), with mean percentages for combinations of animacy and order (lines)

A linear regression model was fit to this data to test whether the percentage of accurate reproductions was affected by age group, target sentence construction, animacy of the two objects in the target sentence, and all two- and three-way interactions between age group, target construction, and object animacy. This model (see Table 6.3 for all coefficients) provides evidence that older participants and target sentences using the prepositional construction all favor higher percentages of accurate reproductions. Inanimate recipients, on the other hand, have a negative parameter, meaning sentences with an inanimate recipient saw a lower percentage of accurate reproductions. The significant interactions between age groups and construction type also have negative parameters. Since the main effect of construction type is a positive one, these negative interaction effects work to (mostly) cancel it out: the calculation in (6.6a) shows this model's fitted or predicted percentage of correct repetitions by four-year-olds and the target sentence

variable	parameter estimate	standard error	<i>t</i>	<i>p</i>
(Intercept)	0.36	0.06	6.11	< 0.01
eight-year-old group	0.36	0.08	4.35	< 0.01
adult group	0.57	0.08	6.83	< 0.01
prepositional target	0.33	0.08	3.99	< 0.01
inanimate recipient	-0.15	0.07	-2.11	0.04
inanimate theme	0.11	0.07	1.53	0.13
eight-year-olds : prepositional	-0.28	0.12	-2.40	0.02
eight-year-olds : inanimate recipient	0.05	0.10	0.52	0.61
eight-year-olds : inanimate theme	0.04	0.10	0.38	0.71
adults : prepositional	-0.27	0.12	-2.24	0.03
adults : inanimate recipient	0.14	0.10	1.38	0.17
adults : inanimate theme	-0.07	0.10	-0.64	0.52
inanimate recipient : inanimate theme	0.15	0.10	1.49	0.14
prepositional : inanimate recipient	0.18	0.10	1.76	0.08
prepositional : inanimate theme	-0.18	0.10	-1.76	0.08
eight-year-olds : inanim. recipient : inanim. theme	-0.10	0.14	-0.71	0.48
adults : inanim. recipient : inanim. theme	-0.26	0.14	-1.76	0.08
prepositional : inanim. recipient : inanim. theme	-0.11	0.14	-0.77	0.45
eight-year-olds : prepositional : inanim. recipient	-0.02	0.14	-0.12	0.90
adults : prepositional : inanim. recipient	-0.19	0.14	-1.33	0.19
eight-year-olds : prepositional : inanim. theme	0.14	0.14	0.99	0.33
adults : prepositional : inanim. theme	0.09	0.14	0.62	0.54
eight-year-olds : prepositional : inanim. recipient	-0.06	0.20	-0.31	0.76
adults : prepositional : inanim. recipient	0.24	0.20	1.15	0.25

Table 6.3: Coefficients for regression model of percentages of accurate reproductions (variables in bold are deemed to have significant effects; $MSE = 0.005$, ${}_{10}M_{10} = 0.01$)

Anne gave the drawing to the parents.

(6.6) a. $0.36 + 0.33 + 0.11 - 0.18 = \mathbf{0.62}$

b. $0.36 + 0.33 + 0.11 - 0.18 + 0.57 - 0.27 - 0.07 + 0.09 = \mathbf{0.94}$

As this is a prepositional sentence with an inanimate theme, the parameters for these two variables (0.33 and 0.11, respectively) are added to the intercept value (0.36). The parameter for the interaction (−0.18) is also added, resulting in a fitted percentage of 62% (the actual percentage in the data is 50% for this combination). In (6.6b), the predicted value for the same sentence with adult participants is calculated by adding the adult main effect (0.57) and the relevant interactions (−0.27 for adults and prepositional, −0.07 for adults and inanimate theme, and 0.09 for adults and prepositional and inanimate theme) to all these, with 94% as the result (the actual value observed in the data is 91%). The negative parameter for the interaction of adults and prepositional targets (−0.27) cancels out most of the parameter for the main effect of prepositional targets (0.33) here, and the similar negative parameter for the interaction of eight-year-olds and prepositional targets (−0.28) has a similar effect for the eight-year-olds—meaning this main effect of prepositional targets is strong only among the four-year-olds according to this model. This is not surprising: as seen in Fig. 6.4, adults have high percentages of accurate repetitions for all sentences, meaning they are unlikely to show any (strong) effect of construction type or object features. Four-year-olds, on the other hand, are less likely to accurately reproduce a double object construction sentence than a prepositional one, as Fig. 6.4 suggest and the significant and strong effect of construction type in the regression model summarized in Table 6.3 shows.

The interactions between construction type and the animacy of the two objects were included to test for effects of animacy order effects (to test, for example, whether inanimate-before-animate targets are less likely to be reproduced accurately). If the interaction between prepositional construction and inanimate recipient reached significance, it would serve as evidence for an ordering effect: in the double-object animate-theme baseline, an inanimate recipient is significantly associated with fewer accurate reproductions. As the double object construction has the recipient before the theme, this is consistent with an ordering effect: in the baseline, sentences with an inanimate object first elicited fewer accurate reproductions than other sentences. However, if this were truly an effect of animacy ordering, the interaction of the prepositional construction and inanimate recipients would have to be significant as well (with a positive parameter estimate): since the prepositional construction has the recipient after the theme, an inanimate recipient should not be associated with fewer accurate reproductions there. As this interaction is not significant (for any age group), there is no strong evidence here for any animacy ordering being preferred or more easy to process.

A linear regression model testing for effects of number was also fit to the data. No effect or

interaction involving number reached significance, which is why that model is not shown here. This does mean, however, that there is no evidence for **any** systematic difference in accurate reproductions that relates to number, and thus no evidence for any combination of number, order, and role of objects being systematically preferred or easier than any other.

6.2.3 Reproduction initiation time

Participants may take longer to start reproducing sentences that are harder to process. The time from the end of the target sentence stimulus to the start of the participant’s reproduction serves as a measure of this. For each recorded trial, the starting point of the participant’s reproduction of the target sentence was determined by the “To Text Grid (silences)” method of Praat (Boersma 2001), and these starting points were then manually checked and adjusted where necessary (for example, where the first detectable sound in the recording was a background noise rather than the participant speaking). The resulting response initiation times are shown in Fig. 6.9. A linear regression model was fit to this data to test whether response initiation time was affected by accuracy of reproduction, age group, target sentence construction, animacy and number of the two objects in the target sentence, and all two- to four-way interactions between age group, target construction, object animacy, and object number.

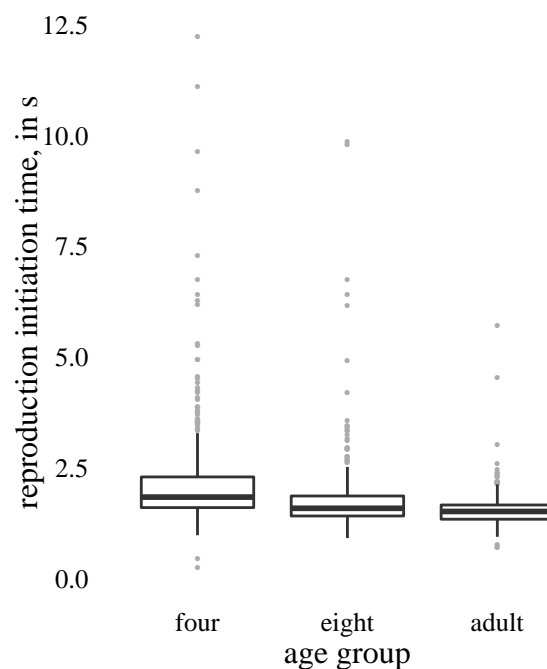


Figure 6.9: Reproduction initiation times by age group (lower and upper ends of boxes indicate first and third quartiles, vertical lines extend to the most extreme value within 1.5 interquartile ranges of these quartiles, grey dots are more extreme values)

variable	parameter estimate	standard error	<i>t</i>	<i>p</i>
(Intercept)	2.20	0.22	10.18	< 0.01
accurate reproduction	-0.21	0.06	-3.49	< 0.01
prepositional target	-0.15	0.31	-0.50	0.61
eight-year-old age group	-0.85	0.29	-2.94	< 0.01
adult age group	-0.60	0.29	-2.08	0.04
inanimate theme	0.04	0.18	0.20	0.84
inanimate recipient	0.14	0.18	0.74	0.46
singular theme	0.22	0.18	1.22	0.22
singular recipient	0.02	0.18	0.09	0.93
eight-year-olds : prepositional target	0.64	0.41	1.58	0.11
adults : prepositional target	0.25	0.40	0.63	0.53
inanimate theme : inanimate recipient	-0.13	0.27	-0.49	0.63
singular theme: singular recipient	0.38	0.28	1.36	0.17
prepositional : inanimate theme	0.13	0.25	0.53	0.60
prepositional : inanimate recipient	-0.16	0.25	-0.62	0.53
prepositional : singular theme	-0.36	0.26	-1.41	0.16
prepositional : singular recipient	0.04	0.26	0.14	0.89
eight-year-olds : inanimate theme	0.37	0.24	1.53	0.13
adults : inanimate theme	0.18	0.24	0.76	0.45
eight-year-olds : inanimate recipient	0.51	0.24	2.09	0.04
adults : inanimate recipient	-0.06	0.24	-0.24	0.81
eight-year-olds : singular theme	0.24	0.24	1.01	0.31
adults : singular theme	-0.11	0.24	-0.47	0.64
eight-year-olds : singular recipient	0.28	0.24	1.14	0.25
adults : singular recipient	0.04	0.24	0.18	0.85
prepositional : inanimate theme : inanimate recipient	0.56	0.38	1.45	0.15
prepositional : singular theme : singular recipient	-0.17	0.39	-0.44	0.66
eight-year-olds : inanimate theme : inanimate recipient	-0.58	0.37	-1.59	0.11
adults : inanimate theme : inanimate recipient	-0.06	0.36	-0.17	0.86
eight-year-olds : singular theme : singular goal	-0.58	0.37	-1.56	0.12
adults : singular theme : singular goal	-0.22	0.36	-0.61	0.54
eight-year-olds : prepositional : inanimate recipient	-0.32	0.34	-0.93	0.35
adults : prepositional : inanimate recipient	0.19	0.34	0.57	0.57
eight-year-olds : prepositional : inanimate theme	-0.48	0.34	-1.42	0.16
adults : prepositional : inanimate theme	-0.30	0.34	-0.89	0.37
eight-year-olds : prepositional : singular recipient	-0.44	0.34	-1.28	0.20
adults : prepositional : singular recipient	-0.08	0.34	-0.24	0.81
eight-year-olds : prepositional : singular theme	-0.33	0.34	-0.97	0.33
adults : prepositional : singular theme	0.17	0.34	0.51	0.61
eight-year-olds : prepositional : inanimate theme : inanimate recipient	0.11	0.52	0.21	0.83
adults : prepositional : inanimate theme : inanimate recipient	-0.59	0.51	-1.16	0.25
eight-year-olds : prepositional : singular theme : singular recipient	0.53	0.52	1.03	0.30
adults : prepositional : singular theme : singular recipient	0.04	0.51	0.07	0.94

Table 6.4: Coefficients for regression model of reproduction initiation times (variables in bold are deemed to have significant effects; $MSE = 0.57$, ${}_{100}M_{10} = 0.62$)

This model (see Table 6.4 for all coefficients) provides evidence that older participants are quicker to initiate their reproductions, which is apparent in Fig. 6.9. The model shows that accurate reproductions are initiated significantly faster than inaccurate ones. This is made more clear in Fig. 6.10, where the orange line for the average initiation time of inaccurate reproductions is above (and approximately parallel to) the blue line for the average initiation time of accurate reproductions. The model also suggests that eight-year-olds are faster to initiate sentence reproduction when the target sentence contains an animate recipient than when the target sentence recipient is inanimate. In Fig. 6.11, the blue line (for mean reproduction initiation time of target sentences with animate recipients) is lower than the orange line (mean reproduction initiation time for inanimate recipient target sentences) only for the eight-year-old age group. In other words, four-year-olds and adults were not quicker to reproduce sentences with animate recipients than sentences with inanimate recipients, but eight-year-olds were. No other effects or interactions in this model are significant.

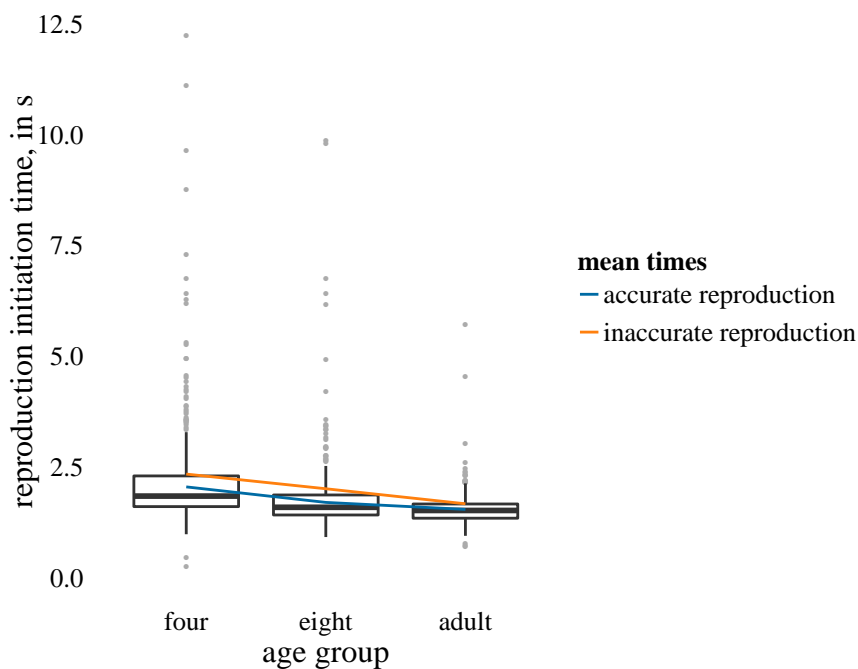


Figure 6.10: Reproduction initiation time by age group (boxes), with means for accurately and inaccurately reproduced sentences (lines)

6.2.4 Disfluencies

Once reproduction has been initiated, reproduction of more difficult sentences may be more likely to be interrupted. Therefore, all trials were manually annotated for the presence of speech disfluencies. I used the following disfluency categories (based on the disfluency types of Conture 1990:15):

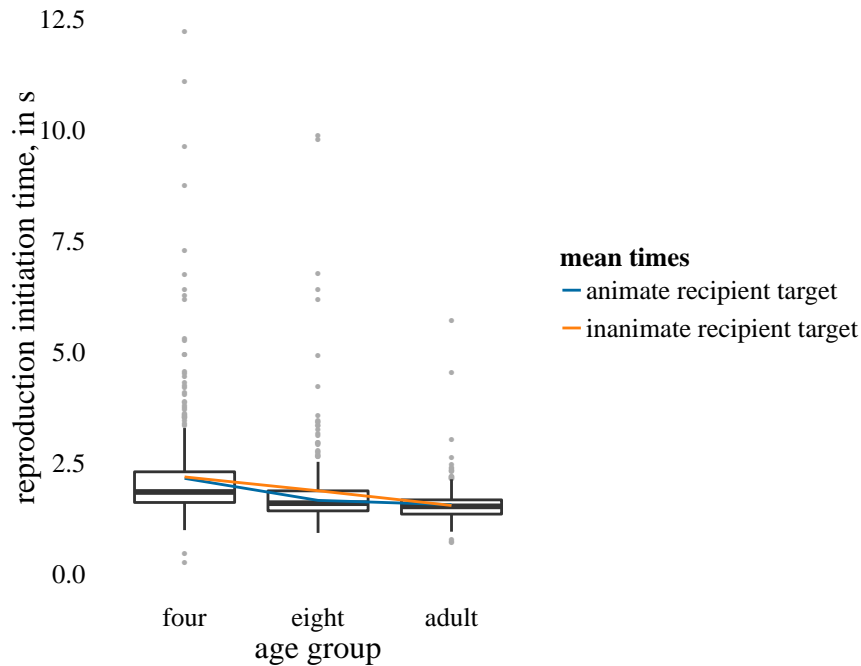


Figure 6.11: Reproduction initiation time by age group (boxes), with means for target sentences with animate and inanimate recipients (lines)

- one-word repetitions (*Anne gave– gave . . .*)
- multi-word repetitions (*Anne gave– Anne gave . . .*)
- interjections (*uh*, but also *no*, and the like)
- pauses not in line with the impression of the participant’s speech rate during this trial (within or between words)
- revision (*Anne made– gave . . .*)
- sound or syllable repetitions (*Anne ga– gave . . .*)

Because the incidence of some disfluencies was low (for example, only 34 trials contained revisions and only 7 had sound or syllable repetitions), all disfluency categories were collapsed into a single variable for the presence of any disfluency. The percentage of trials with any disfluency was calculated for each target sentence and age group to measure difficulties in reproduction. Fig. 6.12 compares these percentages across the three age groups. Adults were hardly ever disfluent (7 out of $22 \times 24 = 528$ trials, or 1.3%), while eight- and particularly four-year-olds produced many disfluencies. Disfluencies have been used in many studies of child language as a measure of processing or production difficulty (for example Gordon et al. 1986, Bernstein Ratner and Sih 1987, Buhr and Zebrowski 2009, McDaniel et al. 2010, and Williams 2014), and they were frequent enough among four- and eight-year-olds in the present data to be a

useful measure of difficulty here. The fact that adults were almost never disfluent is not surprising, considering their well-developed memory capacity and the length of the sentences: McDaniel et al. (2010) found that children aged between three and eight years produce disfluencies in various places across the clause, whereas adult disfluencies tend to be nearer the beginning of the clause. They argue that children plan their speech less far ahead than adults do, and this is most likely due to children being limited by memory or processing capacity. If there is a capacity limitation that becomes less restrictive with development in the first decade of life, it is only to be expected that four-year-olds produce more disfluencies than eight-year-olds and eight-year-olds in turn produce more disfluencies than adults.

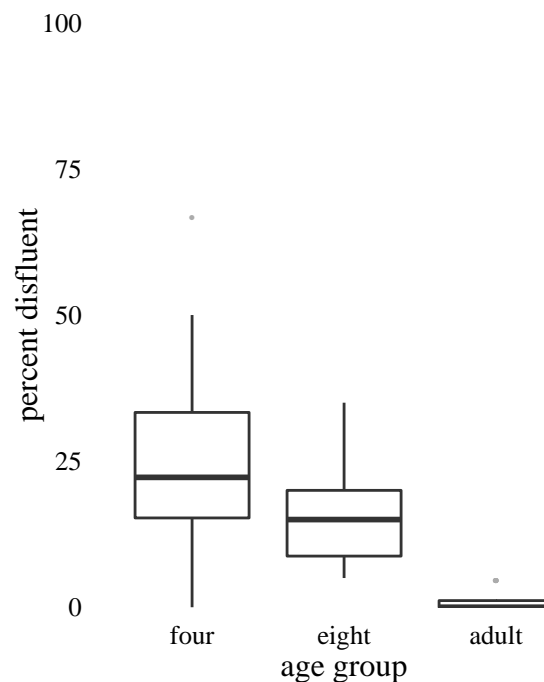


Figure 6.12: Percentages of disfluencies in reproduction, by age group (lower and upper ends of boxes indicate first and third quartiles, vertical lines extend to the most extreme value within 1.5 interquartile ranges of these quartiles, grey dots are more extreme values)

A linear regression model was fit to this data to test whether the percentage of disfluent reproductions was affected by age group, target sentence construction, animacy and number of the two objects in the target sentence, the two-way interactions between object numbers and between object animacies, and all three-way interactions between age group and those two-way interactions. This model (see Table 6.5 for all coefficients) shows that eight-year-olds and adults were significantly less likely than four-year-olds to produce disfluent reproductions, which is readily apparent from Fig. 6.12. There are also significant effects of animacy: inanimate theme and recipient objects each have a marginally significant (both $p = 0.06$) negative or lowering effect on the percentage of disfluent reproductions (against the baseline of two animate objects), and the combination of an inanimate theme with an inanimate recipient has a positive effect that

variable	parameter estimate	standard error	<i>t</i>	<i>p</i>
(Intercept)	0.41	0.06	6.46	< 0.01
prepositional target	-0.08	0.02	-3.37	< 0.01
eight-year-olds	-0.16	0.07	-2.18	0.03
adults	-0.31	0.07	-4.29	< 0.01
inanimate theme in target sentence	-0.12	0.06	-1.93	0.06
inanimate recipient in target sentence	-0.12	0.06	-1.93	0.06
singular theme in target sentence	-0.06	0.04	-1.53	0.13
singular recipient in target sentence	-0.04	0.04	-1.03	0.31
inanimate theme : inanimate recipient	0.23	0.09	2.51	0.01
eight-year-olds : inanimate theme	0.05	0.09	0.62	0.54
adults : inanimate theme	0.09	0.09	1.05	0.30
eight-year-olds : inanimate recipient	0.11	0.09	1.25	0.21
adults : inanimate recipient	0.10	0.09	1.18	0.24
singular theme : singular recipient	0.08	0.06	1.32	0.19
eight-year-olds : inanim. theme : inanim. recipient	-0.15	0.12	-1.21	0.23
adults : inanim. theme : inanim. recipient	-0.16	0.12	-1.29	0.20

Table 6.5: Coefficients for regression model of percentages of disfluent reproductions (variables in bold are deemed to have significant effects; $MSE = 0.01$, ${}_{10}M_{10} = 0.01$)

almost cancels out each of these negative effects (0.23 against -0.12 and -0.12). In other words, one inanimate object means a lower percentage of disfluencies than in the baseline, but this is cancelled out in the case of two inanimate objects. Thus, sentences with two inanimate objects as well as sentences with two animate objects were more likely to be reproduced with disfluencies than sentences with one inanimate and one animate object, regardless of the function of those objects. Note that the corresponding interaction effects with eight-year-olds and adults are in the reverse direction of these effects (positive for one inanimate object: 0.05, 0.09, 0.11, and 0.10; negative for the interaction effects with both objects being inanimate: -0.15 and -0.16). Although they are not significant according to the Wald test, these effects mostly cancel out the main effects and interactions discussed above. In other words, sentences with one animate and one inanimate object were less likely to cause disfluencies **among the four-year-olds**. This is illustrated in Fig. 6.13. Although none of the corresponding effects are significant, the fact that the parameter for the interaction of eight-year-olds and inanimate themes, 0.05, is smaller than those of the other age-animacy interactions (0.09, 0.11, and 0.10) would appear to replicate the intriguing pattern in the middle column of Fig. 6.13: eight-year-olds are slightly less likely to be disfluent when reproducing sentences with inanimate themes and animate recipients than with any other animacy pattern, though the difference is not significant. (This is more apparent in Fig. 6.14, where the disfluency distributions have been further split between target sentences with this prototypical pattern (orange boxes) and all other patterns (blue boxes), and the orange box is noticeably lower than the blue one for the eight-year-old age group). In the regression model, inanimate themes increase the likelihood of disfluency less than inanimate recipients

do for eight-year-olds (the estimate for the eight-year-olds:inanimate theme interaction, 0.05, is smaller than that of the eight-year-olds:inanimate recipient interaction, 0.11). This, however, is nothing more than an interesting trend, as the parameter does not achieve significance in the model.

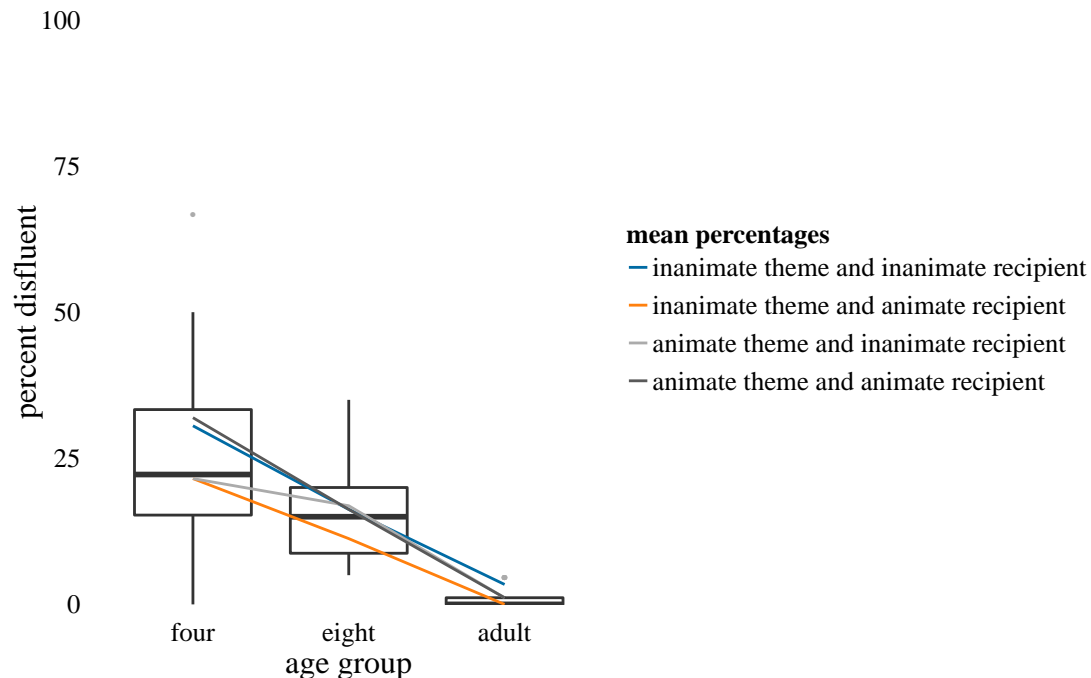


Figure 6.13: Percentages of disfluencies by age group (boxes), with mean percentages for combinations of animacy (lines)

6.2.5 Oddness reactions

After the experiment had been concluded, I noticed reactions indicating that a participant found the target sentence funny or odd in some recordings. These include comments like “That’s silly!” or “That doesn’t sound right!”, laughter or chuckling, and intonations that indicate amusement or questioning/disagreement. I will call them ‘oddness reactions’ here. All trials were manually annotated for the presence or absence of oddness reactions. They were rare (24 from four-year-olds, 15 from eight-year-olds, and 9 from adults for a total of 48), and most of them (43 of the total 48) occurred in trials without any noticeable disfluency. Percentages of oddness reactions were calculated for each sentence and age group (see Fig. 6.15), and these percentages were analyzed using linear regression (percentage of oddness reactions as affected by target sentence construction type, age group, target sentence objects’ animacy and number, and all two-and three-way interactions between age group, target sentence objects’ animacy, and target sentence objects’ number).

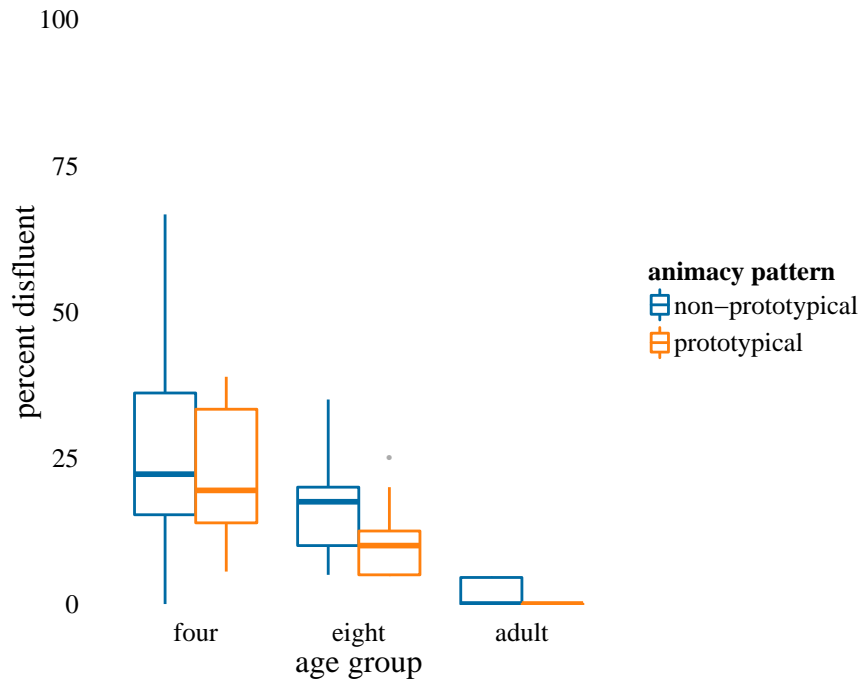


Figure 6.14: Percentages of disfluencies by age group in trials with inanimate themes and animate recipients (orange boxes) and other animacy combinations (blue boxes)

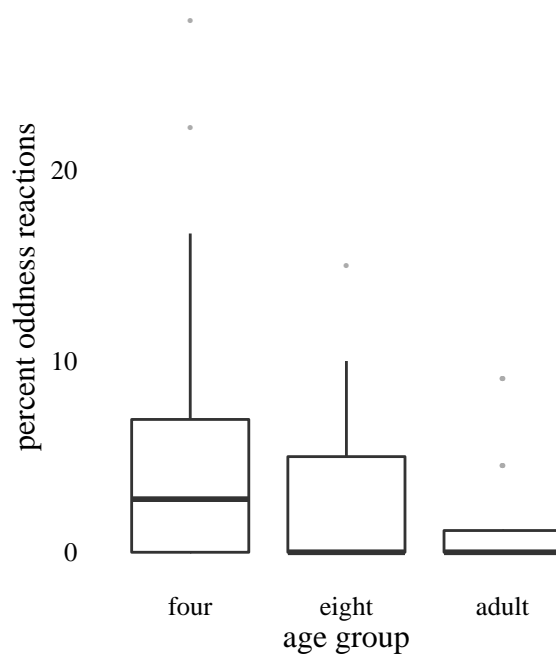


Figure 6.15: Percentages of oddness reactions to target sentences by age group (note the labels on the vertical axis; lower and upper ends of boxes indicate first and third quartiles, vertical lines extend to the most extreme value within 1.5 interquartile ranges of these quartiles, grey dots are more extreme values)

variable	parameter estimate	standard error	<i>t</i>	<i>p</i>
(Intercept)	-0.02	0.03	-0.76	0.45
prepositional target	0.01	0.01	1.07	0.29
eight-year-olds	> -0.01	0.03	> -0.01	1.00
adults	> -0.01	0.03	> -0.01	1.00
inanimate theme	0.04	0.03	1.53	0.13
inanimate recipient	0.06	0.03	2.04	< 0.05
singular theme	0.03	0.02	1.52	0.14
singular recipient	< 0.01	0.02	0.30	0.77
inanimate theme : inanimate recipient	0.07	0.04	1.72	0.09
eight-year-olds : inanimate theme	-0.03	0.04	-0.75	0.45
adults : inanimate theme	-0.02	0.04	-0.62	0.54
eight-year-olds : inanimate recipient	0.01	0.04	0.37	0.71
adults : inanimate recipient	-0.01	0.04	-0.38	0.70
singular theme : singular goal	-0.03	0.03	-1.16	0.25
eight-year-olds : inanim. theme : inanim. recipient	-0.10	0.05	-1.91	0.06
adults : inanim. theme : inanim. recipient	-0.12	0.05	-2.20	0.03

Table 6.6: Coefficients for regression model of percentages of oddness reactions (variables in bold are deemed to have significant effects; $MSE = 0.001$, ${}_{10}M_{10} = 0.003$)

The resulting model (see Table 6.6 for all coefficients) shows that inanimate recipient objects lead to a higher percentage of oddness reactions. Marginally ($p = 0.09$), the combination of an inanimate recipient with an inanimate theme further increases the percentage of oddness reactions for four-year-old participants. This two-way interaction effect is cancelled out for eight-year-olds (marginally: $p = 0.06$) and adults (significantly: $p = 0.03$) by the respective three-way interaction effects. This is made more apparent in Fig. 6.17, where the blue and light grey lines showing the mean percentage of oddness reactions for sentences with inanimate recipients are above the other two lines across age groups, and the blue line (inanimate theme and recipient) rises above all others for the four-year-old age group. There is no significant effect of age group by itself, but this may be caused by the relatively low percentages of oddness reactions overall (note the scale in Fig. 6.15). Since the values are so small, the differences between age groups are quite small as well. The grammatical number of objects has no significant effect,⁹⁷ and neither does the type of construction used in the target sentence.

6.3 Summary

Crucially, the effects of ordering that were expected following the corpus studies of the dative alternation were not found in this sentence imitation experiment. The other interesting or significant results can be condensed into three points. Firstly, the prepositional construction appears to be easier overall for children than the double object construction. Both the four-

⁹⁷Note that the figures in this section have been scaled to adequately show the rare oddness reactions—the orange line for mean percentage of oddness reaction in sentences with singular themes and plural recipients is above the other lines in Fig. 6.16, for example, but the difference is minute: 6.9% with the next-highest line at 6.2% for the four-year-olds, 5% versus 2.5% for eight-year-olds, and 2.8% versus 2.3% for the adults.

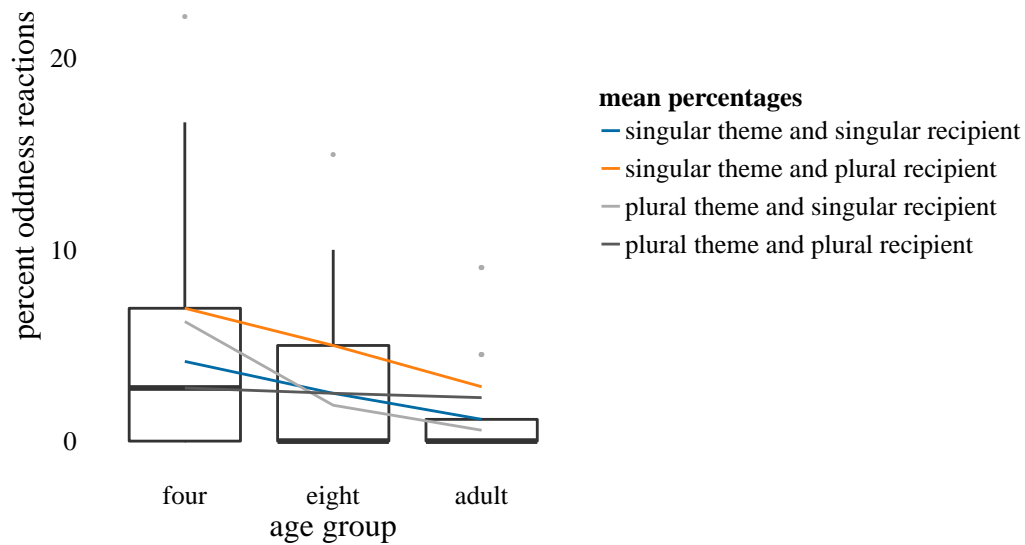


Figure 6.16: Percentages of oddness reactions by age group (boxes), with mean percentages for combinations of number (lines)

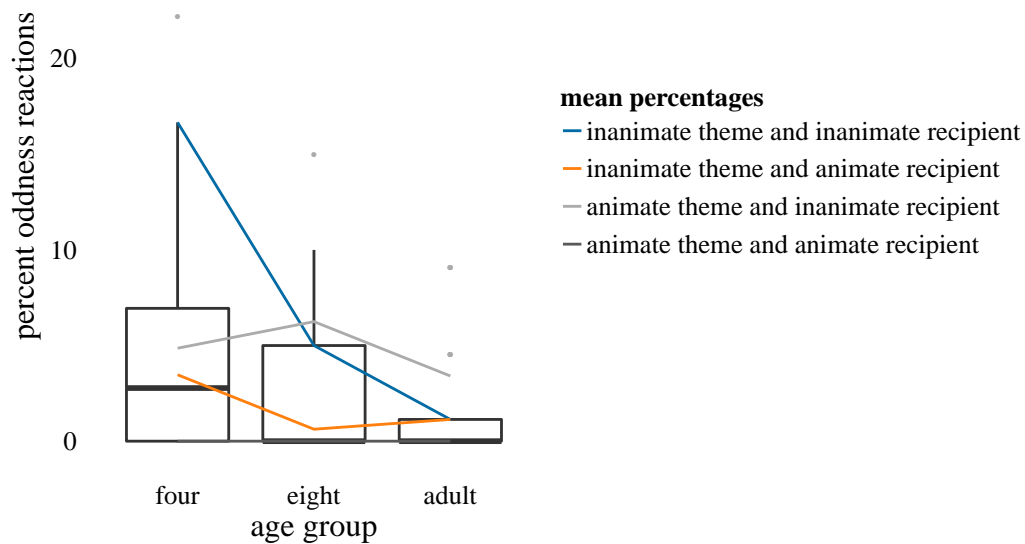


Figure 6.17: Percentages of oddness reactions by age group (boxes), with mean percentages for combinations of animacy (lines—note that the dark grey line (animate/animate combination) is at 0 for all age groups)

and eight-year-old age groups used the prepositional construction more than the double object construction in their reproductions of sentences (Fig. 6.3). Four-year-olds were also more likely to reproduce a prepositional target sentence accurately (Fig. 6.5). Secondly, animate recipients are easier to reproduce, and recipients are arguably even expected to be animate: eight-year-old participants reproduced sentences with animate recipients quicker than sentences with inanimate recipients (Table 6.4). Sentences with inanimate recipients also caused more “oddness reactions” (like laughter or comments about the sentence being “silly” accompanying the reproduction) than sentences with animate recipients did (Fig. 6.17 and Table 6.6). Thirdly, and in addition to the second point, the prototypical animacy pattern for the objects of *give* (an inanimate theme and an animate recipient) seems to be easier to reproduce as well, and participants will even change sentences to get this prototypical pattern. Four-year-olds are quicker to reproduce sentences with one inanimate and one animate object, regardless of the order or roles of these objects (Fig. 6.11). This curious disregard for order and roles is explained in part by a common pattern of changes: participants often changed double object sentences with an inanimate recipient and an animate theme (like *Mom gave the cushions Anne*) to a more prototypical prepositional sentence by simply inserting *to* between the two objects (yielding *Mom gave the cushions to Anne*). There were also more oddness reactions to sentences with two inanimate objects (Fig. 6.17), which cannot be amended towards the prototypical pattern in this way. Finally, eight-year-old participants showed slightly fewer disfluencies in sentences with inanimate themes and animate recipients than in other types of sentences (Fig. 6.13), though the difference is not significant.

7 Discussion

This thesis took a psycholinguistic approach to a syntactic question in the domain of first language acquisition, and its results and contributions are as multifaceted as that description. I found evidence for an animate-before-inanimate preference in experiment 2, but not in the elicited productions of experiment 3. There was no evidence for a plural-before-singular preference, and a length difference of just one syllable does not appear to be enough to trigger the shorter-before-longer preference. Touch input is closely related to eye gaze, making touchscreens a potential alternative to eyetracking.

7.1 Is there any evidence for an animacy ordering preference?

Experiment 2 (Chapter 5) asked participants to choose between three options for filling a gap in a dative sentence where their choice would fill either the theme or the recipient slot. When the gap came after an animate explicit object in the instruction sentence, participants were more likely to select inanimate choices than when the gap came before an animate object. In other words, an animate-before-inanimate preference appears to have guided participants' choices, at least in part. The interaction effects in the regression model that reveals this effect (Table 5.10) suggest that it does not differ significantly across age groups, meaning that **all** participants, from the four-year-olds to the adults, show this same order effect. This is independent experimental evidence for the the animacy-related ordering effect reported by Bresnan et al. (2007) for adult speakers. Previous studies of the dative alternation in child language (de Marneffe et al. 2012 and an unpublished analysis of the data in Bürkle 2011) did not find a significant animacy ordering effect in child speech corpora, probably due to the uncontrolled and unbalanced nature of corpus data. The fact that experiment 2 did find this effect with child participants after all demonstrates the usefulness of the experimental approach as a supplement to corpus studies, and adds to an emerging literature that seems to converge on the idea that the dative alternation choice is just as complex in the language of children (from four years of age) as in the language of adults (Stephens 2010, de Marneffe et al. 2012, van den Bosch and Bresnan 2013)—in the words of de Marneffe et al. (2012:54), “child speech only differs from the speech of their adult interlocutors in degree, not in kind”.

The verb *give* prototypically selects an inanimate theme and an animate recipient. For example, we are much more likely to give a coffee to Kate (and to talk about this event) than we are to give a coffee to Kate's desk or give Kate to the coffee. The latter pattern may even be hard to parse—what does it mean for a coffee to be given a person? Sandberg et al. (2012) show that speakers with and without aphasia have more trouble understanding sentences with that pattern

than sentences with other patterns. The results of the present study provide some insight on how this animacy pattern is learned. There is a wealth of evidence (discussed in Section 2.3) that children have a basic understanding of the concept of animacy, and that it can have effects in their speech. This is reflected in the finding that sentences with animate recipients were overall more likely to be reproduced accurately in experiment 3 than sentences with inanimate recipients, which violate the prototypical pattern. Other measures also show that four-year-olds are affected by such animacy violations: among four-year-olds, sentences with two inanimate objects met with more oddness reactions such as laughter or comments like “That’s silly!” than other types of sentences (Fig. 6.17). In experiment 2, trials with the gap replacing the theme in the instruction sentence (and participants choosing between three theme options) saw more of four-year-olds’ gazes falling on pictures of inanimate objects than on pictures of animates (Fig. 5.59). This difference was not found in the goal-gap trials. In other words, the four-year-old participants were particularly drawn to inanimates in the condition where the prototypical animacy pattern demands inanimates. Together with the comparatively large number of oddness reactions to sentences with two inanimates, this suggests that four-year-olds are aware that *give* prototypically takes only one inanimate object (its theme) and are trying to fulfil this early version of the prototypical pattern.

This tentative developmental milestone immediately raises one question, however: if four-year-olds’ behavior is evidence of their having learned the prototypical animacy pattern, and if this pattern persists into adulthood (a reasonable assumption), why do the eight-year-olds not show the same behavioral differences? The answer that can be inferred from the present data is that while the prototypical animacy pattern itself persists, its behavioral correlates change. Firstly, some measures do reflect an understanding of the prototypical animacy pattern in eight-year-olds: they were more likely to accurately reproduce sentences with an inanimate theme than those with an animate theme (Fig. 6.7). Together with the overall facilitating effect of animate recipients on reproduction accuracy (see Table 6.3), this effect of inanimate themes completes the prototypical animacy pattern. The fact that eight-year-olds’ sentences show both of these effects is evidence for eight-year-olds (still) possessing this pattern. Moreover, they also produced fewer disfluencies when reproducing sentences with the prototypical pattern than when reproducing any other type of sentence (Fig. 6.13). The absence of any correlate of this animacy pattern in eight-year-olds’ gaze data can be explained by charting the development of this animacy pattern and its effects further, into adulthood. The gaze data collected from the adult participants shows the reverse of the effect found in the four-year-olds’ gazes. In trials asking them to choose one of three theme options, adult participants looked at the animate options significantly more than at the inanimate options. Maintaining the basic assumption that adults know the prototypical animacy pattern, this can be explained as a different type of attention: whereas four-year-olds attended more to the options that fit the pattern, adults attended more to the unexpected options that

violate the pattern. With this tentative explanation, the absence of a difference between animate and inanimate theme options in the eight-year-olds' gaze data is no longer surprising—they are undergoing development of their attention and expectations, and so are sometimes attending to the expected options and sometimes to the unexpected ones. Averaging over many trials and participants then would then lead to a mean result of no difference, which is exactly what was found here.

The data from adult participants in experiment 3 is uninformative not because they are unaware of the prototypical animacy pattern, but because of floor or ceiling effects: with fully-developed meta-linguistic skills, adults are capable of reproducing any sequence of words, so a violation of animacy patterns in an otherwise grammatically correct sentence in their native language poses no challenge for reproduction. In other words, because adults reproduced virtually all sentences accurately and without disfluencies or oddness reactions, there is no way for any significant differences relating to the animacy pattern to emerge in their data. Because of children's developing cognitive skills, inaccurate reproductions, disfluencies, and oddness reactions were more common in the two children's age groups. This makes it possible to gain insights from their data.

7.2 Is there any evidence for a number ordering preference?

Participants' choices between the three smaller images in experiment 2 did not show an ordering preference related to grammatical number: they did not choose plural options more in trials where the gap preceded a singular object than in trials where the gap followed a singular object, and they did not choose singular options more when the gap followed a plural object than when it preceded a plural object. In experiment 3, none of the measures used revealed systematic differences that could be explained by an ordering preference for number. In light of the plural-before-singular preference in the literature following Bresnan et al. (2007), such preferences would be expected (assuming the experiments in the present study allow such effects to surface). There are two possible explanations for this negative finding: either the plural-before-singular ordering preference do not hold for the participants in this study, or the experiments in this thesis were not suitable for testing for this effect. The first of these possibilities is not as far-fetched as it may first appear, since most of the studies that reported a plural-before-singular ordering preference are based on corpus data and thus have to be interpreted with the limitations of corpora in mind. However, the second possibility is more likely. The task in experiment 2 was designed specifically for the present study, meaning it is simply not known whether this task would replicate the ordering effects found in corpus studies. The task was also somewhat long, which means at least some (particularly younger) participants likely got fatigued and changed

their behavior (probably to random choice, in order to ‘just finish’) part of the way through the task.⁹⁸ Moreover, the features of one option image systematically depended on the features of the other two: each of the three options matched the explicit object in exactly one of the three features of interest. This was a deliberate design choice to keep the number of choices small, but it does mean the features are not perfectly independent, which is a challenge in statistical analysis. Thus, there are several possible confounding factors in the experiment, and any of them could be expected to mask a subtle ordering preference. The fact that this study did not find the plural-before-singular preference attested in Bresnan et al. (2007) therefore does not constitute evidence against this ordering preference.

The choices in experiment 2 show some other effects of animacy and grammatical number in children and adults. Similarly to the results of the pilot study (see Section 5.1.4.1), these effects can be seen as tendencies to choose options that match or mismatch one of the features of the explicit object. In experiment 2, the eight-year-olds overall tended to choose animacy-mismatching options, meaning they were more likely than chance to select an inanimate option in trials with an animate explicit object and more likely than chance to select an animate option in trials with an inanimate explicit object. As mentioned above, the features of the three options were not independent of each other. There was always one animacy-matching option and two animacy-mismatching ones, one of which matched the grammatical number (but not the length) and one of which matched the length (but not the number) of the explicit object. This means that a very strong mismatching tendency for one feature will manifest as a matching tendency for the other features: in the hypothetical extreme case where all participants always chose one of the two animacy-mismatching options and had no preference regarding number or length, they would choose the number-match and the length-match around 50% of the time each. This would appear as significant deviations from the 33% random chance of choosing the number-match and the 33% random chance of choosing the length-match (compare the analysis in Section 5.3.2), even though participants were absolutely insensitive to either of these features. Of course, this hypothetical extreme case is rather unlikely, and a less extreme mismatching tendency for one feature would not cause these misleading significant deviations. Nevertheless, multiple matching/mismatching tendencies within the same participant groups have to be interpreted cautiously, as it is not clear which of them are true effects and which are merely caused by the interaction of these effects with the details of the experimental design (and, of course, it is conceivable that there is more than one concurrent true effect).

Happily, there is only one such pair of tendencies: eight-year-old participants made more

⁹⁸Pupillometry data, which some eyetrackers can collect during gaze-tracking tasks, may be able to reveal whether there was such a change in behavior. Human pupils dilate under cognitive load (Porter et al. 2007), so if a participant changed their behavior to a less processing-intensive random choice, their pupils may not dilate as much after this change as they did before.

animacy-mismatching choices than expected by chance, as mentioned above (see Fig. 5.40), and they also made more number-matching choices than chance (Fig. 5.41). It is not clear which of these reflects the preferences of the eight-year-old participants, but it is intriguing that the adult participants also show the number-matching preference, but not the animacy-mismatching one. Moreover, an animacy-matching tendency is apparent in adults' gazes (Fig. 5.52). This at least suggests that the number-matching preference is the true one, as the development from a strong number-matching preference in eight-year-olds (manifesting as an epiphenomenal animacy-mismatching tendency) to a more moderate number-matching preference in adults (without epiphenomena) is a more parsimonious account of this data than the development from an animacy-mismatching preference in eight-year-olds to a number-matching preference in adults. Moreover, a tentative explanation for number-matching is more forthcoming than one for animacy-mismatching: singular recipients can plausibly be given singular or plural themes (*give Kate a coffee* and *give Kate three bagels* are both very imagine-able), but it would perhaps be more plausible to give a plural theme rather than a singular theme to a plural recipient (*give three bagels to the boys* is fine, but *give a coffee to the guys* makes one wonder which of the guys really received the one coffee). In other words, a plural-for-plural strategy is semantically more plausible, and it is easy to imagine this strategy being generalized to singular-explicit trials and thus becoming a number-matching strategy—particularly since all trials after the initial training featured the very obvious contrast between two plural images (depicting three animals or objects each) and two singular images (depicting just one animal or object each). Eight-year-old and adult participants looking for a strategy or 'solution' to the task would have been likely to base it on this obvious difference, which would explain the number-matching tendencies observed with these two age groups.

7.3 Is there any evidence for a length ordering preference with a minimal length difference?

There is no evidence for a monosyllabic-before-bisyllabic preference in the results of experiment 2. The effect of length on the dative alternation and similar two-item ordering effects is well established: speakers of all ages prefer shorter items to be ordered before longer items (Wasow 2002, Bresnan et al. 2007, de Marneffe et al. 2012). The question in this study is whether a length difference of one syllable (and zero words) is enough to cause this effect. As with the animacy and number effects discussed above, there is some uncertainty inherent in interpreting the results, as the methodology used here may not be able to uncover such ordering effects at all. In light of this, the fact that eight-year-olds' choices appear to manifest a bisyllabic-before-monosyllabic preference (Table 5.9) has to be treated with great caution. In light of the well-established

strength of the short-before-long effect on postverbal constituent ordering, it seems unlikely that this one finding is evidence of a long-before-short preference—especially given that there are other, more parsimonious explanations for the present data: the task may be unable to reveal length-related ordering effects, and a length difference of one syllable may not be enough to trigger the short-before-long effect. In either case, the apparent bisyllabic-before-monosyllabic preference would then be a spurious finding, as would the four-year-olds preference for gazing at bisyllabic options. The length-mismatching tendency in adults' choices is most likely an epiphenomenon of another matching tendency, as discussed above.

7.4 Do these preferences emerge in a particular order?

As the three experiments did not find evidence for number- or length-based ordering effects, the question of an order of emergence of these effects cannot be answered here. One interesting developmental trajectory is apparent, however: in experiment 3, children (in both age groups) were more likely to reproduce target sentences using the prepositional construction than sentences using the double object construction. Reproductions of prepositional sentences also contained fewer disfluencies. The adults' reproductions do not show these differences, as they performed essentially without inaccuracies or disfluencies. It is clear that the prepositional construction carries some processing advantage for children. I do not mean that it is acquired earlier, which Gropen et al. (1989) already showed to be an unfounded assumption—rather, it appears that children who have already acquired both constructions have an easier time (re-)producing prepositional sentences or, in other words, a harder time (re-)producing double object sentences. It is interesting in light of this latter formulation that many of the inaccurately reproduced double object sentences were reproduced using the prepositional construction. (Note that this does not logically follow from the fact that more double object sentences were reproduced inaccurately: there are many other changes or errors that would render a reproduction inaccurate, for example exchanging the two objects without inserting a preposition, omitting one or both objects, using different objects or verbs, and so forth.) This suggests that both factors play a role: the double object construction can be difficult to reproduce, **and** there is some processing advantage to the prepositional construction. The obvious candidate for this processing advantage is the preposition *to* itself. It clearly marks the object following it as the recipient and thereby makes the sentence less ambiguous. This would be particularly helpful in sentences where other hints to the roles of the two objects are less clear, for example with sentences that deviate from the prototypical pattern of inanimate theme and animate recipient. The double-object sentences with animate themes and inanimate recipients (the strongest possible deviation from the prototypical animacy pattern) were more likely to be reproduced inaccurately in experiment 3 than prepositional sentences with the same pattern. It thus seems plausible that the prepositional

construction easier for children to process and produce because the recipient is marked more clearly in the prepositional construction than in the double object construction. This is in line with previous findings regarding the use and usefulness of *to* as a cue: the presence of *to* “is by far the strongest cue” that hearers use in identifying the recipient (McDonald 1987:114), and persons with and without aphasia are least likely to arrive at the intended meaning of a dative alternation sentence when the objects violate the prototypical animacy pattern (animate theme, inanimate recipient) **and** the sentence uses the preposition-less double object construction (Sandberg et al. 2012).

7.5 Does touchscreen input reflect attention?

When actively interacting with the touchscreen, participants looked at regions close to the point of touch—in other words, their touch input is correlated to their gaze, and gaze is an established attention measure. Therefore, the assumption that touchscreen input and gaze target are closely correlated (made commonly in touchscreen implementations, as discussed briefly in Section 3.5) is justified. In light of this finding, researchers can use touchscreens as attention-measure devices more confidently, and designers can be confident in assuming that touchscreen input requires and reflects attention. The distance between the gaze and touch locations appears to grow with age, but it is possible that this is an artifact of hardware problems: the fact that gaze data collection failed completely for four four-year-olds, three eight-year-olds, but only one adult could be argued to suggest that the particular eyetracker used here is less likely to capture gaze data from younger participants with more varied gaze behavior, while adults with similar variation can still be recorded. This would in effect make the adults’ gaze data more messy. As the accuracy of the touchscreen data does not differ between age groups, messier eye gaze data would make the correlation between touch and gaze weaker. Although the pattern of gazes does differ between age group (as the regression model in Section 5.3.5 shows), this difference is not necessarily due to one group’s data being messier overall. The fact that algorithmic correction was successfully applied to the gaze data from 12 four-year-olds, 12 eight-year-olds, and 14 adults (see Section 5.2.3 for details) further suggests that the eye gaze data collected from most adult participants was bunched just as tightly as the data collected from children. The difference in correlation strengths could therefore be due to gaze data from a few ‘messier’ adult participants, while no gaze data was recorded from their child counterparts (which would make the child gaze data appear ‘cleaner’ and thus more closely correlated with the corresponding touch locations).

8 Conclusion

Previous studies of the English dative alternation (almost exclusively corpus work, exemplified by Bresnan et al. 2007; see Section 2.1) found that speakers tend to use the construction that places shorter objects before longer ones, animate objects before inanimate ones, and plural objects before singular ones. Apart from the animate-before-inanimate preference revealed by experiment 2, the present study did not find evidence for these ordering effects. This overall negative finding does not invalidate those previous results, of course—the highly artificial tasks used here and the various data collection methods used to construct these corpora are different in many ways. The present finding does show, however, that the ordering tendencies are no more than tendencies: large corpora of (mostly) naturalistic speech allow them to be detected, but in an individual processing or production situation, they can easily be overpowered by myriad other influences. Syntactic priming is known to be one such influence, task strategies and task demands appear to be others.

By ‘task strategies’, I mean consistent and conscious strategies that participants can adopt for an experimental task. Some of the adult participants in the pilot study described in Section 5.1.4.1 commented that they chose the animate option in trials with an animate as the big picture and the inanimate option in trials with an inanimate big picture. They were evidently trying to find the ‘solution’ or pattern, and animacy presented itself. The task strategy here was to choose the option that matched the animacy of the big picture, at least for some participants. Number was not varied in that pilot—all pictures showed one animal or object. In experiment 2, number was varied. It is possible that number then presented itself as the most obvious pattern, which would explain why the eight-year-old and adult participants appear to have used a number-matching strategy: when presented with a singular big image representing the explicit object in the instruction sentence, eight-year-olds and adults chose the one singular option more often than chance would suggest; when the big image representing the explicit object was plural, they chose the plural option more than chance would suggest.

No major strategy or pattern is apparent in the four-year-olds’ choices. The demands of the experimental tasks, however, explain this: all three experiments presented in this thesis were administered in one session. Participants were free to take breaks between the experiments as well as between the blocks in experiment 2, and even a session without breaks could take more than 45 minutes. Loss of interest and fatigue due to this long duration, combined with the discovery that there did not seem to be a wrong choice in experiment 2 (any choice led to the ‘reward’ stimulus), may very well have led (some) four-year-old participants to simply get through the task without trying to ‘solve’ it like the older participants did. Moreover, four-year-olds’ general cognitive skills are less well developed than those of the older children and

adults—even if they did try to adopt a strategy, they may have been less successful at formulating or implementing it consistently.

Ordering preferences, task strategies and task demands cannot explain every single choice made in experiment 2: participants did not try to match the animate-before-inanimate ordering in absolutely every trial, eight-year-olds and adults did not always choose the option that matched the explicit object in grammatical number, and four-year-olds did not seem to make systematic choices to the same extent as eight-year-olds and adults did. However, these influences are the strongest ones apparent in the choice data, and any other potential pattern is simply lost under their strength and the noise that is common to experiments of this nature.

The eye gaze data collected during experiment 2 is also fairly noisy, which is due in part to the limitations of head-free eyetracking technology. The gaps in the eye gaze recording (explained in Section 5.2.2) as well as the fact that I analyzed the raw gaze samples (as opposed to post-processed fixations defined by some algorithm) make for a less even dataset, further adding to this noise. Moreover, the methodological evaluation in Section 5.2.4 indicates that SSANOVA is conservative when applied to eye gaze data, which means that SSANOVA models are less likely to deem a given difference between images or conditions to be a significant difference than other statistical methods would be. However, with these limitations in mind, the significant differences that were found in the eye gaze data can be accepted much more confidently.

More participants in the adult participant group provided useful gaze data than participants in either of the two child groups. It is therefore not surprising that their data also provides more cases of significant differences. As they heard the instruction sentence, adults already looked at the option that they would later choose more than at the other two options combined (see Fig. 5.50). It is impossible to establish causation from the data—adults may immediately have decided on their choice and then looked at it, or they may have been more likely to choose the option that they happened to be looking at during the instruction sentence (resembling a mere exposure effect).⁹⁹ No matter which of these two possibilities is true, this finding shows that the act-out choices and the eye gaze data captured some of the same features of adult participants' behavior in this experiment. If the four-year-old participants did choose randomly, as I have argued above, it is not surprising that their gaze behavior does not correspond to these random choices: whether their gazes were random or patterned, they are unlikely to show interesting correspondences to a random set of data.

⁹⁹Mere exposure effects are preferences for previously seen items over novel ones, regardless of attitude towards or conscious awareness of these previously seen items—the mere exposure to them is apparently enough to cause this preference. In the same vein, the results and gaze data from experiment 2 may reflect a preference for the image that happened to be being seen during the instruction sentence, simply because it was being seen as the sentence was heard. A mere exposure effect would be particularly likely here because the stimuli used were photographs, which have been found to elicit stronger mere exposure effects than line drawings (Bornstein 1989:279).

The gazes of eight-year-old participants, however, are interesting: their choices do not appear to be random, so a correspondence between choices and gazes (like the one found with adult participants) is possible. However, there is none. Two explanations suggest themselves: either the eight-year-olds really did not look at their subsequent choices more than at the other options; or they did look at their choices more, but this difference was smaller than it was among the adults and therefore got lost in the noise or the conservative analysis. Hardly any interesting difference between gazes at one image and gazes at another were found at all in the eight-year-olds' gaze data, while the four-year-old and adult age groups each showed several such differences. This is suggestive of considerable variation in the eight-year-olds' gaze behavior. The significant patterns of gazes found with younger children and adults differ from each other, which could be explained by a cognitive development between four years of age and adulthood. It is reasonable to assume that some participants in the eight-year-old age group are further along that developmental path than others. This would mean some of them display more adult-like gaze behavior while others display more four-year-old-like gaze behavior—in other words, that there is considerable variation in the eight-year-olds' gaze data. As the groups are fairly small already, trying to separate the eight-year-olds into more and less adult-like sub-groups would not be statistically sound: smaller datasets (from smaller groups) are more susceptible to false positive findings.

The fact that adults' choices were mirrored in their gazes shows neatly how an attention measure (like eye gaze) and a simpler measure (like the choice between the three options made on the touchscreen) can provide similar results, and that both are useful. As eyetracking has a very fine temporal resolution, it can also reveal more information within trials. As eye gaze is not under constant conscious control, it is a more useful measure when participants may not engage with the task fully, like the four-year-old participants in the present study. The touchscreen, however, requires very little setup and no explanation (with most participants), and careful experiment design can turn it into an attention measure. Participants were allowed to take their hands off the screen in experiment 2, and they did so for most of the time. When participants are instructed to keep their finger on the screen, or when the task naturally requires that they do, touch-tracking can provide data that is similar to eye gaze data. The advantage that touch-tracking has over mouse-tracking is that the mouse is often used as a reading or viewing aid, for example following gaze “roughly” in a visual search (Bieg et al. 2010:90) or following gaze in the vertical dimension while remaining quite distant in the horizontal (Rodden and Fu 2007). **Touch** and gaze are much closer even in an entirely unconstrained task like the one in this study. In a task that is designed to keep participants' fingers on or near the point of processing, such as Hatfield (to appear)'s self-guided reading task, touch-tracking can approximate eye gaze data very closely. This is particularly useful to know because touchscreen-based experiments can be delivered as smartphone apps to thousands of users with relative ease (Dufau et al. 2011), which makes it

possible to carry out attention-measure studies with unprecedented numbers of participants and the corresponding gain in statistical power.

8.1 Further research

This study used several measures of difficulty in the data from experiment 3, chief among them the binary distinction between accurate and inaccurate reproductions of the target sentence. This analysis assumes that participants are generally capable of reproducing sentences of the length and complexity used here and that an additional difficulty caused by the factors that were varied as part of the experiment design negatively affects this capability. However, the most striking difference in reproduction accuracy was that between the three age groups. Further research might explore in more detail how the various factors used here as well as others influence this reproduction difficulty, and how much of the difference in reproduction accuracy is really a more general effect of children's memory limitations.

In the realm of the dative alternation, it would be interesting to assess the length effect in more detail. It is established beyond doubt that shorter objects are preferred before longer ones, but the results of this thesis suggest that this does not apply in ditransitive sentences where the two object NPs differ in length by just one syllable. The question that this result leaves open is this: What is the smallest length difference that will give rise to the length effect? Future studies with incrementally larger length differences (several syllables, one word, and so on) would answer that question, giving further insight into the more peripheral features of the human language faculty that govern these gradient phenomena.

As the dative alternation varies across different varieties of English, a replication of the present study in varieties other than New Zealand English would be a worthwhile addition to the literature on the dative alternation as well as a possible source of data for general theories of language variation.

Finally, while I am confident in the animate-before-inanimate preference found in experiment 2, this particular ordering effect has not been demonstrated with children before. It would thus be desirable for another study, preferably with a different methodology, to test for this animacy effect in child language. If the same effect is found again, it would become more firmly established, and the complementary nature of experiments and corpora would be given much-needed prominence and attention.

9 References

The blue numbers following each reference item are the pages in this thesis where that item is cited. Clicking on one will make many PDF readers turn to the corresponding page.

- Abbot-Smith, Kirsten, and Michael Tomasello. 2006. Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *Linguistic Review* 23:275–290.
- Abeillé, Anne. 1990. Lexical and syntactic rules in a Tree Adjoining Grammar. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, 292–298. Morristown: Association for Computational Linguistics. URL <http://acl.ldc.upenn.edu/P/P90>, retrieved 13 January 2014.
- Aissen, Judith. 1999. Markedness and subject choice in Optimality Theory. *Natural Language and Linguistic Theory* 17.4:673–711.
- Aissen, Judith. 2003. Differential object marking. *Natural Language and Linguistic Theory* 21.3: 435–483.
- Akasaka, Yukiko, and Koichi Tateishi. 2001. Heaviness in interfaces. In *Issues in Japanese phonology and morphology*, eds. Jeroen van de Weijer and Tetsuo Nishihara, 3–46. Berlin: Mouton de Gruyter.
- Alfons, Andreas. 2012. *cvTools: Cross-validation tools for regression models*. URL <http://CRAN.R-project.org/package=cvTools>.
- Anderson, John. 1977. *On case grammar*. London: Croom Helm.
- Anderssen, Merete, Paula Fikkert, Roksolana Mykhaylyk, and Yulia Rodina. 2012. The dative alternation in Norwegian child language. *Nordlyd (Tromsø University Working Papers on Language and Linguistics)* 39.1:24–43. URL <http://septentrio.uit.no/index.php/nordlyd/issue/view/205>, retrieved 28 February 2014.
- Anttila, Arto, Matthew Adams, and Michael Speriosu. 2010. The role of prosody in the English dative alternation. *Language and Cognitive Processes* 25.7–9:946–981.
- Ardestani, Reza. 2013. Gradient and categorical consonant cluster simplification in Persian. Doctoral Dissertation, University of Ottawa.
- Arnold, Jeffrey B. 2014. *ggthemes: extra themes, scales and geoms for ggplot*. URL <http://CRAN.R-project.org/package=ggthemes>.
- Arnold, Jennifer, Anthony Losongco, Thomas Wasow, and Ryan Ginstrom. 2000. Heaviness vs. newness: the effects of structural complexity and discourse status on constituent ordering. *Language* 76.1: 28–55.
- Au, Terry Kit-fong, and Laura Romo. 1999. Mechanical causality in children’s “folkbiology”. In *Medin and Atran (1999)*, 355–401.
- Auguie, Baptiste. 2012. *gridExtra: functions in Grid graphics*. URL <http://CRAN.R-project.org/package=gridExtra>.
- Austin, Peter, Muhammad Mamdani, David Juurlink, and Janet Hux. 2006. Testing multiple statistical hypotheses resulted in spurious associations. *Journal of Clinical Epidemiology* 59.9:964–969. URL <http://dx.doi.org/10.1016/j.jclinepi.2006.01.012>.

- Baayen, R. Harald. 2008. *Analyzing linguistic data*. Cambridge: Cambridge University Press.
- Baayen, R. Harald. 2013. *languageR: Data sets and functions with “Analyzing Linguistic Data”*. URL <http://CRAN.R-project.org/package=languageR>.
- Barner, David, Toni Lui, and Jennifer Zapf. 2012. Is *two* a plural marker in early child language? *Developmental Psychology* 48.1:10–17.
- Barr, Dale. 2008. Analyzing ‘visual world’ eyetracking data using multilevel logistic regression. *Journal of Memory and Language* 59.4:457–474.
- Barr, Dale, Roger Levy, Christoph Scheepers, and Harry Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68.3:255–278.
- Barrios, Edison. 2012. Knowledge of grammar and concept possession. *British Journal for the Philosophy of Science* 63.3:577–606.
- Bassène, Alain-Christian. 2010. Ditransitive constructions in Jóola Banjal. In Malchukov et al. (2010b), 190–203.
- Batali, John. 2002. The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In *Linguistic evolution through language acquisition*, ed. Ted Briscoe, 111–172. Cambridge: Cambridge University Press.
- Bates, Douglas, Reinhold Kliegl, Shravan Vasishth, and Harald Baayen. to appear. Parsimonious mixed models. *Journal of Memory and Language* URL <http://arxiv.org/abs/1506.04967>, retrieved 19 June 2015.
- Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2014. Fitting linear mixed-effects models using lme4. (*Submitted to Journal of Statistical Software*) URL <http://arxiv.org/abs/1406.5823>.
- Becker, Misha. 2007. Animacy, expletives, and the learning of the raising-control distinction. In *Proceedings of the 2nd Conference on Generative Approaches to Language Acquisition North America*, eds. Alyona Belikova, Luisa Meroni, and Mari Umeda, 12–20. Somerville: Cascadilla Proceedings Project.
- Becker, Misha. 2009. The role of NP animacy and expletives in verb learning. *Language Acquisition* 16: 283–296.
- van der Beek, Leonoor. 2004. Argument order alternations in Dutch. In *Proceedings of the LFG ‘04 Conference*, eds. Miriam Butt and Tracy Holloway King, 39–59. Stanford: CSLI. URL <http://csli-publications.stanford.edu/LFG/9/lfg04.html>, retrieved 7 June 2013.
- Behaghel, Otto. 1928. *Deutsche Syntax*, volume 3: *Die Satzgebilde*. Heidelberg: Carl Winters.
- Berk, Laura. 2013. *Child development*. Boston: Pearson, ninth edition.
- Bernaisch, Tobias, Stefan Gries, and Joybrato Mukherjee. 2014. The dative alternation in South Asian English(es). *English World-Wide* 35.1:7–31.
- Bernstein Ratner, Nan, and Catherine Sih. 1987. Effects of gradual increases in sentence length and complexity on children’s dysfluency. *Journal of Speech and Hearing Disorders* 52:278–287.
- Berwick, Robert. 1985. *The acquisition of syntactic knowledge*. Cambridge: MIT Press.

- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan, eds. 1999. *Longman grammar of spoken and written English*. Harlow, Essex: Pearson Education.
- Biedert, Ralf, Andreas Dengel, Georg Buscher, and Arman Vartan. 2012. Reading and estimating gaze on smart phones. In *Proceedings of the Seventh ACM Symposium on Eye Tracking Research and Applications*, 385–388. Safety Harbor, Florida.
- Bieg, Hans-Joachim, Lewis Chuang, Roland Fleming, Harald Reiterer, and Heinrich Bülthoff. 2010. Eye and pointer coordination in search and selection tasks. In *Proceedings of the 2010 Symposium on Eye Tracking Research & Applications*, 89–92.
- Bloom, Paul. 1990. Subjectless sentences in child language. *Linguistic Inquiry* 21.4:491–504.
- Bloom, Paul. 1993. Grammatical continuity in language development. *Linguistic Inquiry* 24.4:721–734.
- Bock, J. Kathryn. 1986. Syntactic persistence in language production. *Cognitive Psychology* 18.3: 355–387.
- Bock, J. Kathryn, and Willem Levelt. 1994. Language production: grammatical encoding. In *Handbook of psycholinguistics*, ed. Morton Gernsbacher, 945–984. San Diego: Academic Press.
- Bod, Rens. 2006. Exemplar-based syntax. *Linguistic Review* 23:291–320.
- Boersma, Paul. 2001. Praat, a system for doing phonetics by computer. *Glott International* 5.9/10: 341–345.
- de Boer, Carl. 2001. *A practical guide to splines*. New York: Springer, second edition.
- Boot, Fleur, Johan Pel, Heleen Evenhuis, and Johannes van der Steen. 2012. Quantification of visual orienting responses to coherent form and motion in typically developing children aged 0–12 years. *Investigative Ophthalmology and Visual Science* 53.6:2708–2714.
- Booth, Amy, and Sandra Waxman. 2002. Word learning is ‘smart’. *Cognition* 84:B11–B22.
- Bornstein, Robert. 1989. Exposure and affect: Overview and meta-analysis of research, 1968–1987. *Psychological Bulletin* 106.2:265–289.
- van den Bosch, Antal, and Joan Bresnan. 2013. Modeling dative alternations of individual children. Draft for comments.
- Bowerman, Melissa. 1994. From universal to language-specific in early grammatical development. In Trott et al. (2004), 131–146. Reprint.
- Bowerman, Melissa. 2011. Linguistic typology and first language acquisition. In *The Oxford handbook of linguistic typology*, ed. Jae Jung Song, 591–617. Oxford: Oxford University Press.
- Bowerman, Melissa, and William Croft. 2008. The acquisition of the English causative alternation. In *Crosslinguistic perspectives on argument structure: implications for learnability*, eds. Melissa Bowerman and Penelope Brown, 279–306. Mahwah: Erlbaum.
- Bozkurt, Alper, and Banu Onaral. 2004. Safety assessment of near infrared light emitting diodes for diffuse optical measurements. *BioMedical Engineering OnLine* 3.1. URL [dx.doi.org/10.1186/1475-925X-3-9](https://doi.org/10.1186/1475-925X-3-9).
- Bragdon, Andrew, Eugene Nelson, Yang Li, and Ken Hinckley. 2011. Experimental analysis of touch-screen gesture designs in mobile environments. In *Proceedings of the SIGCHI Conference on*

- Human Factors in Computing Systems*, 403–412. URL <http://dx.doi.org/10.1145/1978942.1979000>.
- Brainerd, Charles. 1973. Neo-Piagetian training experiments revisited. *Cognition* 2.3:349–370.
- Branigan, Holly, Martin Pickering, and Mikihiro Tanaka. 2008. Contributions of animacy to grammatical function assignment and word order during production. *Lingua* 118:172–189.
- Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2007. Predicting the dative alternation. In *Cognitive foundations of interpretation*, eds. Gerlof Bouma, Irene Krämer, and Joost Zwarts, 69–94. Amsterdam: Royal Netherlands Academy of Science.
- Bresnan, Joan, and Marilyn Ford. 2010. Predicting syntax. *Language* 86.1:168–213.
- Bresnan, Joan, and Jennifer Hay. 2008. Gradient grammar: An effect of animacy on the syntax of *give* in New Zealand and American English. *Lingua* 118.2:245–259.
- Bresnan, Joan, and Tatiana Nikitina. 2003. On the gradience of the dative alternation. URL <http://www.stanford.edu/~bresnan/new-dative.pdf>, retrieved 4 July 2011. Manuscript, Stanford University.
- Brown, Geoffrey, and Charles DesForges. 1979. *Piaget's theory: a psychological critique*. Abingdon: Routledge.
- Brown, Meredith, Virginia Savova, and Edward Gibson. 2012. Syntax encodes information structure. *Journal of Memory and Language* 66.1:194–209.
- Bruening, Benjamin. 2010. Double object constructions disguised as prepositional datives. *Linguistic Inquiry* 41.2:287–305.
- Brumitt, Barry, and JJ Cadiz. 2000. “Let there be light!” Comparing interfaces for homes of the future. Technical report, Microsoft. Technical report MSR-TR-2000-92.
- Bruyn, Adrienne, Pieter Muysken, and Maaïke Verrips. 1999. Double-object constructions in the creole languages. In *Language creation and language change*, ed. Michel DeGraff, 329–373. Cambridge: MIT Press.
- Bucci, Maria, and Zoï Kapoula. 2006. Binocular coordination of saccades in 7 years old children in single word reading and target fixation. *Vision Research* 46:457–466.
- Buhr, Anthony, and Patricia Zebrowski. 2009. Sentence position and syntactic complexity of stuttering in early childhood. *Journal of Fluency Disorders* 34:155–172.
- Bürkle, Daniel. 2011. Weight effects in the acquisition of English: evidence from first and second language acquisition. Master's thesis, University of Konstanz. URL <http://nbn-resolving.de/urn:nbn:de:bsz:352-186631>, retrieved 28 February 2014.
- Bybee, Joan. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.
- Byrne, Brian, and Elizabeth Davidson. 1985. On putting the horse before the cart. *Journal of Memory and Language* 24:377–389.
- Canty, Angelo, and Brian Ripley. 2015. *boot: Bootstrap R (S-Plus) functions*. URL <http://CRAN.R-project.org/package=boot>.
- de Cat, Cécile. 2011. Information tracking and encoding in early L1. *Journal of Child Language* 38: 828–860.

- Chanethom, Vincent. 2011. Dynamic differences in child bilinguals' production of diphthongs. In *Proceedings of the Fourth ICSA Workshop ExLing*. Paris.
- Charney, Rosalind. 1980. Speech roles and the development of personal pronouns. *Journal of Child Language* 7:509–528.
- Chen, John. 2014. *Multivariate Bonferroni-type inequalities*. Boca Raton: Taylor & Francis.
- Chomsky, Noam. 1980. *Rules and representations*. Oxford: Blackwell.
- Chomsky, Noam, and Morris Halle. 1968. *The sound pattern of English*. New York: Harper & Row.
- Chomsky, Noam, and Howard Lasnik. 1993. Principles and Parameters theory. In *Syntax*, ed. Joachim Jacobs, volume 1, 506–569. Berlin: Mouton de Gruyter.
- Clark, Eve, and Tatiana Nikitina. 2009. One vs. more than one: antecedents to plural marking in early language acquisition. *Linguistics* 47.1:103–139.
- Colleman, Timothy, and Bernard de Clerck. 2009. 'Caused motion'? The semantics of the English to-dative and the Dutch aan-dative. *Cognitive Linguistics* 20.1:5–42.
- Colombo, Carlo, Alberto Del Bimbo, and Alessandro Valli. 2003. Visual capture and understanding of hand pointing actions in a 3-D environment. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 33.4:677–686. URL <http://dx.doi.org/10.1109/TSMCB.2003.814281>.
- Colunga, Eliana. 2006. The effect of priming on preschooler's extensions of novel words. In *Proceedings of the 30th annual Boston University Conference on Language Development*, 96–106.
- Connolly, Andrew, Jerry Fodor, Lila Gleitman, and Henry Gleitman. 2007. Why stereotypes don't even make good defaults. *Cognition* 103:1–22.
- Conture, Edward. 1990. *Stuttering*. Englewood Cliffs, NJ: Prentice Hall, second edition.
- Cook, VJ. 1976. A note on indirect objects. *Journal of Child Language* 3.3:435–437.
- Cooper, William, and John Ross. 1975. World order. In *Papers from the parasession on Functionalism*, eds. Robin Grossman, L. James San, and Timothy Vance, 63–111. Chicago: CLS.
- Corrêa, Letícia Sicuro, Marina Augusto, and José Ferrari-Neto. 2005. The early processing of number agreement in the DP. URL <http://www.bu.edu/buclid/proceedings/supplement/vol30/>, retrieved 8 April 2013. Poster presented at the 30th Boston University Conference on Language Development.
- Cuervo, María Cristina. 2003a. Datives at large. Doctoral Dissertation, Massachusetts Institute of Technology.
- Cuervo, María Cristina. 2003b. Structural asymmetries but same word order. In *Asymmetry in grammar*, ed. Anna Di Sciullo, volume 1: *Syntax and semantics*, 117–144. Amsterdam: Benjamins.
- de Cuypere, Ludovic, and Saartje Verbeke. 2013. Dative alternation in Indian English. *World Englishes* 32.2:169–184. URL <http://dx.doi.org/10.1111/weng.12017>.
- Dahl, David. 2014. *xtable: Export tables to LaTeX or HTML*. URL <http://CRAN.R-project.org/package=xtable>.
- Davidson, Lisa. 2006. Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. *Journal of the Acoustical Society of America* 120.1:407–415.

- Davidson, Thomas. 1874. *The Grammar of Dionysios Thrax*. St. Louis, MO: Studley. URL <http://archive.org/stream/grammarofdionysios00dionuoft#page/n3/mode/2up>, retrieved 28 February 2014.
- Demuth, Katherine, 'Malillo Machobane, Francina Moloi, and Christopher Odat. 2005. Learning animacy hierarchy effects in Sesotho double object applicatives. *Language* 81.2:421–447.
- Derrick, Donald, and Benjamin Schultz. 2013. Acoustic correlates of flaps in North American English. *Proceedings of Meetings of Acoustics* 19.1:no page numbers. URL <http://dx.doi.org/10.1121/1.4798779>.
- Desmond, William. 2008. *Cynics*. Berkeley: University of California Press.
- Dewart, M. Hazel. 1979. The role of animate and inanimate nouns in determining sentence voice. *British Journal of Psychology* 70:135–141.
- den Dikken, Marcel. 2001. “Plurilinguals”, pronouns and quirky agreement. *Linguistic Review* 18.1: 19–41.
- Dixon, Robert. 1979. Ergativity. *Language* 55.1:59–138.
- Dąbrowska, Ewa, and Elena Lieven. 2005. Towards a lexically specific grammar of children’s question constructions. *Cognitive Linguistics* 16.3:437–474.
- Downing, Angela, and Philip Locke. 2006. *English grammar: a university course*. London: Routledge, second edition.
- Dufau, Stephane, Jon Andoni Duñabeitia, Carmen Moret-Tatay, Aileen McGonigal, David Peeters, F.-Xavier Alario, David Balota, Marc Brysbaert, Manuel Carreiras, Ludovic Ferrand, Maria Ktori, Manuel Perea, Kathy Rastle, Olivier Sasburg, Melvin Yap, Johannes Ziegler, and Jonathan Grainger. 2011. Smart phone, smart science. *PLoS ONE* 6.9:e24974. URL dx.plos.org/10.1371/journal.pone.0024974.
- Duguine, Maia, Susana Huidobro, and Nerea Madariaga, eds. 2010. *Argument structure and syntactic relations*. Amsterdam: Benjamins.
- Duranti, Alessandro. 1979. Object clitic pronouns in Bantu and the topicality hierarchy. *Studies in African Linguistics* 10.1:31–45.
- Eden, Guinevere, John Stein, HM Wood, and Frank Wood. 1994. Differences in eye movements and reading problems in dyslexic and normal children. *Vision Research* 34.10:1345–1358.
- Emonds, Joseph. 1976. *A transformational approach to English syntax*. New York: Academic Press.
- Erteschik-Shir, Nomi, and Lisa Rochman, eds. 2010. *The sound patterns of syntax*. Oxford: Oxford University Press.
- Fischer, Burkhardt, Monica Biscaldi, and Stefan Gezeck. 1997. On the development of voluntary and reflexive components in human saccade generation. *Brain Research* 754.1–2:285–297.
- Fisher, Anna. 2009. Does conceptual information take precedence over perceptual information early in development? Evidence from perseveration errors. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, eds. Niels Taatgen and Hedderik van Rijn, 1330–1335. URL <http://csjarchive.cogsci.rpi.edu/proceedings/2009/papers/297/index.html>, retrieved 14 January 2014.
- Fodor, Jerry. 1998. *Concepts*. Oxford: Oxford University Press.

- Ford, Marilyn, and Joan Bresnan. 2013. "They whispered me the answer" in Australia and the US. In *From quirky case to representing space: papers in honor of Annie Zaenen*, eds. Tracy Holloway King and Valeria de Paiva, 95–107. Stanford: CSLI. URL <http://csli-publications.stanford.edu/site/9781575866628.shtml>, retrieved 28 February 2014.
- Foster-Cohen, Susan. 1990. *The communicative competence of young children*. London: Longman.
- Freedman, David. 2009. *Statistical models*. Cambridge: Cambridge University Press, second edition.
- Freudenthal, Daniel, Julian Pine, and Fernand Gobet. 2007. Understanding the developmental dynamics of subject omission. *Journal of Child Language* 34:83–110.
- Friedmann, Na'ama. 2001. Agrammatism and the psychological reality of the syntactic tree. *Journal of Psycholinguistic Research* 30.1:71–90.
- Fukushima, Junko, Tatsuo Hatta, and Kikuro Fukushima. 2000. Development of voluntary control of saccadic eye movements: I. age-related changes in normal children. *Brain and Development* 22.3: 173–180.
- Gabora, Liane, Eleanor Rosch, and Diederik Aerts. 2008. Toward an ecological theory of concepts. *Ecological Psychology* 20:84–116.
- Garaizar, Pablo, and Miguel Vadillo. 2014. Accuracy and precision of visual stimulus timing in PsychoPy. *PLoS ONE* 9.11:e112033. URL dx.doi.org/10.1371/journal.pone.0112033.
- Garaizar, Pablo, Miguel Vadillo, Diego López de Ipiña, and Helena Matute. 2014. Measuring software timing errors in the presentation of visual stimuli in cognitive neuroscience experiments. *PLoS ONE* 9.1:e58108. URL dx.doi.org/10.1371/journal.pone.0085108.
- Gelman, Andrew, Jennifer Hill, and Masanao Yajima. 2012. Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness* 5.2:189–211. URL <http://dx.doi.org/10.1080/19345747.2011.618213>.
- Gelman, Rochel, Elizabeth Spelke, and Elizabeth Meck. 1983. What preschoolers know about animate and inanimate objects. In *The acquisition of symbolic skills*, eds. Don Rogers and John Sloboda, 297–326. New York: Plenum.
- Gelman, Susan, and Melissa Koenig. 2001. The role of animacy in children's understanding of 'move'. *Journal of Child Language* 28:683–701.
- Gentner, Dedre, and Lera Boroditsky. 2001. Individuation, relativity, and early word learning. In *Language acquisition and conceptual development*, eds. Melissa Bowerman and Stephen Levinson, 215–256. Cambridge: Cambridge University Press.
- Gentner, Dedre, and Laura Namy. 2006. Analogical processes in language learning. *Current Directions in Psychological Science* 15:297–301.
- Genz, Alan, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl, and Torsten Hothorn. 2014. *mvtnorm: Multivariate normal and t distributions*. URL <http://CRAN.R-project.org/package=mvtnorm>.
- Gilquin, Gaëtanelle. 2008. The place of prototypicality in corpus linguistics. In *Corpora in cognitive linguistics*, eds. Stefan Gries and Anatol Stefanowitsch, 159–191. Berlin: Mouton de Gruyter.
- Girouard, Pascale, Marcelle Ricard, and Thérèse Gouin Décarie. 1997. The acquisition of personal pronouns in French-speaking and English-speaking children. *Journal of Child Language* 24:311–326.

- Gordon, Pearl, Harold Luper, and Harold Peterson. 1986. The effects of syntactic complexity on the occurrence of disfluencies in 5 year old nonstutterers. *Journal of Fluency Disorders* 11:151–164.
- Gries, Stefan. 2003. Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics* 1:1–27.
- Gries, Stefan. 2005. Syntactic priming: a corpus-based approach. *Journal of Psycholinguistic Research* 34.4:365–399.
- Gries, Stefan. 2009. *Statistics for linguistics with R*. Berlin: Mouton de Gruyter.
- Gries, Stefan, and Anatol Stefanowitsch. 2004. Extending collocation analysis. *International Journal of Corpus Linguistics* 9.1:97–129.
- Gropen, Jess, Steven Pinker, Michelle Hollander, Richard Goldberg, and Ronald Wilson. 1989. The learnability and acquisition of the dative alternation in English. *Language* 65.2:203–257.
- Gu, Chong. 2013. *Smoothing spline ANOVA models*. New York: Springer, second edition.
- Gu, Chong. 2014. Smoothing spline ANOVA models: R package gss. *Journal of Statistical Software* 58.5:1–25. URL <http://www.jstatsoft.org/v58/i05/>.
- Hagiya, Toshiyuki, and Tsuneo Kato. 2014. Probabilistic touchscreen keyboard incorporating gaze point information. In *Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices & Services*, 329–333. URL <http://dx.doi.org/10.1145/2628363.2628370>.
- Hammer, Rubi, Gil Diesendruck, Daphna Weinshall, and Shaul Hochstein. 2009. The development of category learning strategies. *Cognition* 112:105–119.
- Haspelmath, Martin. 2004. Explaining the ditransitive person-role constraint. *Constructions* 2004.2:1–71. URL <http://elanguage.net/journals/constructions/article/view/3073>, retrieved 7 May 2013.
- Haspelmath, Martin. 2007. Ditransitive alignment splits and inverse alignment. *Functions of Language* 14.1:79–102.
- Hatfield, Hunter. to appear. Self-guided reading: touch-based measures of syntactic processing. URL http://www.otago.ac.nz/englishlinguistics/linguistics/pub/Hatfield_selfguidedreading.pdf, retrieved 14 April 2014.
- Hawkins, John. 1994. *A performance theory of order and constituency*. Cambridge, Cambs.: Cambridge University Press.
- Hay, Jennifer, Janet Pierrehumbert, Abby Walker, and Patrick LaShell. 2015. Tracking word frequency effects through 130 years of sound change. *Cognition* 139:83–91. URL <http://dx.doi.org/10.1016/j.cognition.2015.02.012>.
- Hendriks, Petra, and Jacolien van Rij. 2011. Language acquisition and language change in bidirectional Optimality Theory. In *Bidirectional Optimality Theory*, eds. Anton Benz and Jason Mattausch, 97–123. Amsterdam: Benjamins.
- Henrich, Joseph, Steven Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences* 33.2–3:61–83.
- Hinterhölzl, Roland. 2004. Language change versus grammar change. In *Diachronic clues to synchronic grammar*, eds. Eric Fuß and Carola Trips, 131–160. Amsterdam: Benjamins.

- Holm, Sture. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6.2:65–70. URL <http://www.jstor.org/stable/4615733>, retrieved 5 August 2015.
- Holzinger, Andreas. 2003. Finger instead of mouse: touch screens as a means of enhancing universal access. In *Universal access (papers from the 7th ERCIM International Workshop on User Interfaces for All, Paris, 24–25 October 2002)*, eds. Noelle Carbonell and Constantine Stephanidis, 387–397. New York: Springer.
- Hudson, Kerry. 2011. *Getting a Tobii eyetracker to work (version 1.2)*. University of Reading, Reading. URL <https://sites.google.com/site/drkerryhudson/eye-tracking-how-to/>, retrieved 22 January 2015.
- Huettig, Falk, and James McQueen. 2007. The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language* 57.4:460–482.
- Hundt, Marianne, and Benedikt Szmrecsányi. 2012. Animacy in early New Zealand English. *English World-Wide* 33.3:241–263.
- Hyams, Nina. 1998. Underspecification and modularity in early syntax. In Trott et al. (2004), 241–258. Reprint.
- Hyams, Nina, and Kenneth Wexler. 1993. On the grammatical basis of null subjects in child language. *Linguistic Inquiry* 24.3:421–459.
- Ibbotson, Paul, and Michael Tomasello. 2009. Prototype constructions in early language acquisition. *Language and Cognition* 1.1:59–85.
- Inkelas, Sharon. 1990. *Prosodic constituency in the lexicon*. New York: Garland.
- Ito, Kiwako, and Shari Speer. 2008. Anticipatory effects of intonation. *Journal of Memory and Language* 58.2:541–573.
- Ivanova, Iva, Martin Pickering, Janet McLean, Albert Costa, and Holly Branigan. 2012. How do people produce ungrammatical utterances? *Journal of Memory and Language* 67.3:355–370.
- Jaeger, T. Florian, and Neal Snider. 2013. Alignment as a consequence of expectation adaptation. *Cognition* 127:57–83.
- Jenkins, Lyle, ed. 2004. *Variation and universals in biolinguistics*. Amsterdam: Elsevier.
- Johnston, Judith. 1985. Cognitive prerequisites: the evidence from children learning English. In *The crosslinguistic study of language acquisition*, ed. Dan Slobin, volume 2: *Theoretical issues*, 961–1004. Hillsdale: Erlbaum.
- Jones, Horace. 1950. *The Geography of Strabo: with an English translation*, volume 6. London and Cambridge, MA: William Heinemann and Harvard University Press.
- Junge, Bianca, Anna Theakston, and Elena Lieven. 2015. Given–new/new–given? children’s sensitivity to the ordering of information in complex sentences. *Applied Psycholinguistics* 36.3:589–612.
- Katzir, Tami, Shirley Hershko, and Vered Halamish. 2013. The effect of font size on reading comprehension on second and fifth grade children. *PLOS ONE* 8.9:e74061. URL dx.doi.org/10.1371/journal.pone.0074061.
- Kawalkar, Amit. 2011. Touch screen and method for providing stable touches. URL https://www.lens.org/lens/patent/US_8766936_B2, patent US 8,766,936 B2.

- Kayne, Richard. 2003. Some remarks on agreement and on heavy-NP shift. In *Movement and silence*, ed. Richard Kayne, 261–276. Oxford: Oxford University Press. Reprint.
- Kayne, Richard. 2012. Comparative syntax. *Lingua* 130:132–151.
- Keil, Heinrich, ed. 1961. *Grammatici latini*, volume 7: *Scriptores de orthographia*. Hildesheim: Georg Olms.
- Kendall, Tyler, Joan Bresnan, and Gerard van Herk. 2011. The dative alternation in African American English. *Corpus Linguistics and Linguistic Theory* 7.2:229–244.
- Kennedy, William, and James Gentle. 1980. *Statistical computing*. New York: Dekker.
- Kent, Roland. 1951. *Varro on the Latin language: with an English translation*, volume 2. London and Cambridge, MA: William Heinemann and Harvard University Press.
- Keuleers, Emmanuel, Paula Lacey, Kathleen Rastle, and Marc Brysbaert. 2011. The British Lexicon Project. *Behavior Research Methods* 44:118. URL <http://dx.doi.org/10.3758/S13428-011-0118-4>.
- Kilgarriff, Adam. 2007. Googleology is bad science. *Computational Linguistics* 33.1:147–151.
- Kizach, Johannes, and Laura Winther Balling. 2013. Givenness, complexity, and the Danish dative alternation. *Memory and Cognition* xx:yy–zz.
- Koster, Jan. 1994. Toward a new theory of anaphoric binding. In *Syntactic theory and first language acquisition*, eds. Barbara Lust, Margarita Sener, and John Whitman, volume 2: *Binding, dependencies, and learnability*, 41–69. Hillsdale: Erlbaum.
- Kouider, Sid, Justin Halberda, Justin Wood, and Susan Carey. 2006. Acquisition of English number marking. *Language Learning and Development* 2.1:1–25.
- Kovačević, Melita, Marijan Palmović, and Gordana Hržica. 2009. The acquisition of case, number and gender in Croatian. In *Development of nominal inflection in first language acquisition*, eds. Ursula Stephany and Maria Voeikova, 153–177. Berlin: Mouton de Gruyter.
- Kuperman, Victor, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods* 44.4:978–990. [Http://crr.ugent.be/archives/806](http://crr.ugent.be/archives/806).
- Kurniasih, Nia. 2009. Benefactive verbs in double object construction (DOC) in English sentences. *Jurnal Sositologi* 8.16:575–586.
- Kurumada, Chigusa, Meredith Brown, Sarah Bibyk, Daniel Pontillo, and Michael Tanenhaus. 2014. Is it or isn't it: Listeners make rapid use of prosody to infer speaker meanings. *Cognition* 133:335–342.
- Laka, Itziar, and Kepa Erdocia. 2012. Linearization preferences given “free word order”; subject preferences given ergativity. In *Of grammar, words, and verses*, ed. Esther Torrego, 115–142. Amsterdam: Benjamins.
- Lambert, Silke. 2010. Beyond recipients: towards a typology of dative uses. Doctoral Dissertation, University of Buffalo, State University of New York.
- Langendoen, D. Terence, Nancy Kalish-Landon, and John Dore. 1973. Dative questions: a study in the relation of acceptability to grammaticality of an English sentence type. *Cognition* 2.4:451–478.
- Laurendeau, Monique, and Adrien Pinard. 1962. *Causal thinking in the child*. New York: International Universities Press.

- Law, Vivien. 1987. An unnoticed late Latin grammar. *Rheinisches Museum für Philologie* 130.1:67–89.
- Leddon, Erin, Sandra Waxman, and Douglas Medin. 2009. Unmasking “alive”: children’s appreciation of a concept linking all living things. *Journal of Cognition and Development* 9.4:461–473.
- Lee-Kim, Sang-Im, Lisa Davidson, and Sagjin Hwang. 2013. Morphological effects on the darkness of English intervocalic /l/. *Laboratory Phonology* 4.2:475–511.
- Legerstee, Maria. 2001. Domain specificity and the epistemic triangle: the development of the concept of animacy in infancy. In *Emerging cognitive abilities in early infancy*, eds. Francisco Lacerda, Claes von Hofsten, and Mikael Heimann, 193–212. Hillsdale: Erlbaum.
- Legerstee, Maria, Andree Pomerleau, Gérard Malcuit, and Helga Feider. 1987. The development of infants’ responses to people and a doll: implications for research in communication. *Infant Behavior and Development* 10.1:81–95.
- Lempert, Henrietta. 1989. Animacy constraints on preschool children’s acquisition of syntax. *Child Development* 60:237–245.
- Lieven, Elena, Dorothé Salomo, and Michael Tomasello. 2009. Two-year-old children’s production of multiword utterances. *Cognitive Linguistics* 20.3:481–507.
- Lightfoot, David. 2003. The development of grammars. In *The second Glot International state-of-the-article book*, eds. Lisa Cheng and Rint Sybesma, 1–24. Berlin: Mouton de Gruyter.
- Lightfoot, David. 2010. Language acquisition and language change. *Wiley Interdisciplinary Reviews: Cognitive Science* 1.5:677–684.
- Lin, Ying. 2005. Learning stochastic OT grammars. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 346–353. Ann Arbor: Association for Computational Linguistics. URL <http://acl.ldc.upenn.edu/P/P05>, retrieved 26 June 2013.
- Liszkowski, Ulf, Malinda Carpenter, Tricia Striano, and Michael Tomasello. 2006. 12- and 18-month-olds point to provide information for others. *Journal of Cognition and Development* 7.2:173–187.
- Liu, Feng-hsi. 2006. Dative constructions in Chinese. *Language and Linguistics* 7.4:863–904.
- Lorimor, Heidi. 2007. Conjunctions and grammatical agreement. Doctoral Dissertation, University of Illinois at Urbana-Champaign.
- Loveland, Katherine. 1984. Learning about points of view. *Journal of Child Language* 11:525–556.
- Luciana, Monica, and Charles Nelson. 2002. Assessment of neuropsychological function through use of the Cambridge Neuropsychological Testing Automated Battery. *Developmental Neuropsychology* 22.3:595–624.
- Lupyan, Gary, and David Rakison. 2006. What moves in a mysterious way? In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*. Mahwah: Erlbaum.
- Lust, Barbara, Suzanne Flynn, and Claire Foley. 1998. What children know about what they say. In *Methods for assessing children’s syntax*, eds. Dana McDaniel, Cecile McKee, and Helen Smith Cairns, 55–76. Cambridge, MA: MIT Press.
- Maglio, Paul, Teenie Matlock, Christopher Campbell, Shumin Zhai, and Barton Smith. 2000. Gaze and speech in attentive user interfaces. In *Proceedings of the Third International Conference on Advances in Multimodal Interfaces*, 1–7. URL dx.doi.org/10.1007/3-540-40063-X_1.

- Malchukov, Andrej, Martin Haspelmath, and Bernard Comrie. 2010a. Ditransitive constructions: a typological overview. In Malchukov et al. (2010b), 1–64.
- Malchukov, Andrej, Martin Haspelmath, and Bernard Comrie, eds. 2010b. *Studies in ditransitive constructions*. Berlin: Mouton de Gruyter.
- Maratsos, Michael. 1974. Preschool children's use of definite and indefinite articles. *Child Development* 45:446–455.
- Maratsos, Michael. 1983. Some current issues in the study of the acquisition of grammar. In *Handbook of child psychology*, eds. John Flavell and Ellen Markman, volume 3: *Cognitive development*, 707–786. New York: Wiley, fourth edition.
- de Marneffe, Marie-Catherine, Scott Grimm, Inbal Arnon, Susannah Kirby, and Joan Bresnan. 2012. A statistical model of the grammatical choices in child production of dative sentences. *Language and Cognitive Processes* 27.1:25–61.
- McDaniel, Dana, Cecile McKee, and Merrill Garrett. 2010. Children's sentence planning. *Journal of Child Language* 37:59–94.
- McDonald, Janet. 1987. Assigning linguistic roles. *Journal of Memory and Language* 26.1:100–117. URL [http://dx.doi.org/10.1016/0749-596X\(87\)90065-9](http://dx.doi.org/10.1016/0749-596X(87)90065-9).
- McDonald, Janet, J. Kathryn Bock, and Michael Kelly. 1993. Word and world order. *Cognitive Psychology* 25.2:188–230.
- McElree, Brian, and Thomas Bever. 1989. The psychological reality of linguistically defined gaps. *Journal of Psycholinguistic Research* 18.1:21–35.
- Medin, Douglas, and Scott Atran, eds. 1999. *Folkbiology*. Cambridge, MA: MIT Press.
- Mennie, Neil, Mary Hayhoe, and Brian Sullivan. 2007. Look-ahead fixations. *Experimental Brain Research* 179:427–442.
- Meyer, David, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. 2014. *e1071: Misc functions of the Department of Statistics (e1071), TU Wien*. URL <http://CRAN.R-project.org/package=e1071>.
- Miller, Karen, and Cristina Schmitt. 2012. Variable input and the acquisition of plural morphology. *Language Acquisition* 19.3:223–261.
- Mirman, Daniel. 2014. Growth curve analysis. URL <http://www.danmirman.org/gca/>, retrieved 14 December 2014.
- Mirman, Daniel, James Dixon, and James Magnuson. 2008. Statistical and computational models of the visual world paradigm. *Journal of Memory and Language* 59.4:475–494.
- Morgante, James, Rahman Zolfaghari, and Scott Johnson. 2012. A critical test of temporal and spatial accuracy of the Tobii T60XL eye tracker. *Infancy* 17.1:9–32.
- Morimoto, Carlos, Dave Koons, Arnon Amir, and Myron Flickner. 2000. Pupil detection and tracking using multiple light sources. *Image and Vision Computing* 18.4:331–335.
- Morra, Sergio, Camilla Gobbo, Zopito Marini, and Ronald Sheese. 2008. *Cognitive development*. New York: Erlbaum.

- Mrotek, Leigh, and John Soechting. 2007. Target interception: hand-eye coordination and strategies. *Journal of Neuroscience* 27:7292–7309.
- Mulak, Karen, Catherine Best, Michael Tyler, Christine Kitamura, and Julia Irwin. 2013. Development of phonological constancy. *Child Development* 84.6:2064–2078.
- Munoz, Douglas, JR Broughton, JE Goldring, and Irene Armstrong. 1998. Age-related performance of human subjects on saccadic eye movement tasks. *Experimental Brain Research* 121.4:391–400.
- Nisbet, Tim. 2005. Benefactives in English: evidence against argumenthood. *Reading Working Papers in Linguistics* 8:51–67. URL <http://www.reading.ac.uk/internal/appling/wp8/index.htm>, retrieved 29 April 2013.
- Niyogi, Partha. 2004. Phase transitions in language evolution. In Jenkins (2004), 57–74.
- Oehrle, Richard. 1976. The grammatical status of the English dative alternation. Doctoral Dissertation, Massachusetts Institute of Technology.
- Oehrle, Richard. 1977. Review of Semantics and syntactic regularity, by Georgia M. Green. *Language* 53.1:198–208.
- Ogawa, Yoshiki. 2008. The dative alternation as A-movement out of a small clause CP. *English Linguistics* 25.1:93–126.
- Okita, Sandra, and Daniel Schwartz. 2006. Young children’s understanding of animacy and entertainment robots. *International Journal of Humanoid Robotics* 3:393–412.
- Oliphant, Margaret. 1893. *The makers of Venice*. London: Macmillan.
- Opfer, John, and Susan Gelman. 2011. Development of the animate-inanimate distinction. In *The Wiley-Blackwell handbook of childhood cognitive development*, ed. Usha Goswami, 213–238. Chichester: Wiley-Blackwell, second edition.
- Ormazabal, Javier, and Juan Romero. 2010. The derivation of dative alternation. In Duguine et al. (2010), 203–232.
- Ormazabal, Javier, and Juan Romero. 2012. PPs without disguises: reply to Bruening. *Linguistic Inquiry* 43.3:455–474.
- Otsuka, Tatsuo. 2006. On the thematic roles of beneficiary and recipient in the benefactive alternation in English. *Bulletin of the Faculty of Education, Chiba University* 54:257–261.
- Oyharçabal, Beñat. 2010. Basque ditransitives. In Duguine et al. (2010), 233–260.
- Paczynski, Martin, and Gina Kuperberg. 2011. Electrophysiological evidence for use of the animacy hierarchy, but not thematic role assignment, during verb-argument processing. *Language and Cognitive Processes* 26.9:1402–1456.
- Patla, Aftab, and Joan Vickers. 2003. How far ahead do we look when required to step on specific locations in the travel path during locomotion? *Experimental Brain Research* 148:133–138.
- Paul, Hermann. 1880. *Principien der sprachgeschichte*. Halle: Niemeyer. Facsimile reprint, 2009. Cambridge: Cambridge University Press.
- Paus, Tomáš, Vitalij Babenko, and Tomáš Radil. 1990. Development of an ability to maintain verbally instructed central gaze fixation studied in 8- to 10-year-old children. *International Journal of Psychophysiology* 10:53–61.

- Pearson, P. David, and Alice Studt. 1975. Effects of word frequency and contextual richness on children's word identification abilities. *Journal of Educational Psychology* 67.1:89–95.
- Peirce, Jonathan. 2007. PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods* 162:8–13. URL <http://www.psychopy.org>.
- Peirce, Jonathan. 2009. Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics* 2.10.
- Pelz, Jeff, and Roxanne Canosa. 2001. Oculomotor behavior and perceptual strategies in complex tasks. *Vision Research* 41:3587–3596.
- Piaget, Jean. 1978. *Success and understanding* (transl. Arnold Pomerans). Abingdon: Routledge.
- Pinker, Steven. 2004. Clarifying the logical problem of language acquisition. *Journal of Child Language* 31:949–953.
- Porter, Gillian, Tom Troscianko, and Iain Gilchrist. 2007. Effort during visual search and counting. *The Quarterly Journal of Experimental Psychology* 60.2:211–229. URL <http://dx.doi.org/10.1080/17470210600673818>.
- Prince, Alan, and Paul Smolensky. 2004. *Optimality Theory*. Malden: Blackwell.
- Qi, Feng, A-Xing Zhu, Mark Harrower, and James Burt. 2006. Fuzzy soil mapping based on prototype category theory. *Geoderma* 136:774–787.
- R Development Core Team. 2011. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. URL <http://www.R-project.org>, ISBN 3-900051-07-0.
- Rakison, David, and Diane Poulin-Dubois. 2001. Developmental origin of the animate-inanimate distinction. *Psychological Bulletin* 127.2:209–228.
- Ransom, Evelyn. 1977. Definiteness, animacy, and NP ordering. In *Proceedings of the Third Annual Meeting of the Berkeley Linguistics Society*, 418–429.
- Rappaport Hovav, Malka, and Beth Levin. 2008. The English dative alternation: the case for verb sensitivity. *Journal of Linguistics* 44:129–167.
- Ratkovic, Mark. 2009. Are there red states and blue states? URL http://users.polisci.wisc.edu/gehlbach/documents/StateLines_RD4.pdf, manuscript.
- Rayner, Keith, Erik Reichle, Michael Stroud, Carrick Williams, and Alexander Pollatsek. 2006. The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers. *Psychology and Aging* 21.3:448–465.
- Revelle, William. 2015. *psych: Procedures for psychological, psychometric, and personality research*. Northwestern University, Evanston, Illinois. URL <http://CRAN.R-project.org/package=psych>.
- Roberts, Ian. 2007. *Diachronic syntax*. Oxford: Oxford University Press.
- Rodden, Kerry, and Xin Fu. 2007. Exploring how mouse movements relate to eye movements on web search result pages. In *Proceedings of the 2007 SIGIR workshop “Web information seeking and interaction”*, eds. Kerry Rodden, Ian Ruthven, and Ryen White, 29–32.
- Rolfe, John. 1927. *The Attic Nights of Aulus Gellius: with an English translation*, volume 2. London and New York: William Heinemann and Putnam's.

- Romeo, Geoff, Suzy Edwards, Sue McNamara, Ian Walker, and Christopher Ziguras. 2003. Touching the screen: issues related to the use of touchscreen technology in early childhood education. *British Journal of Educational Technology* 34.3:329–339.
- Rosch, Eleanor, and Carolyn Mervis. 1975. Family resemblances: studies in the internal structure of categories. *Cognitive Psychology* 7.4:573–605.
- Rosch, Eleanor, Carolyn Mervis, Wayne Gray, David Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive Psychology* 8:382–439.
- Rouder, Jeffrey, and Roger Ratcliff. 2006. Comparing exemplar- and rule-based theories of categorization. *Current Directions in Psychological Science* 15:9–13.
- Rozendaal, Margot, and Anne Baker. 2008. A cross-linguistic investigation of the acquisition of the pragmatics of indefinite and definite reference in two-year-olds. *Journal of Child Language* 35: 773–807.
- Russo, Francesco Di, Sabrina Pitzalis, and Donatella Spinelli. 2003. Fixation stability and saccadic latency in elite shooters. *Vision Research* 43:1837–1845.
- Sadeh-Leicht, Oren. 2007. The psychological reality of grammar. Doctoral Dissertation, University of Utrecht.
- Sandberg, Chaleece, Swathi Kiran, Edward Gibson, and Evelina Fedorenko. 2012. The effect of plausibility in sentence processing. In *Proceedings of the 42nd Clinical Aphasiology Conference*. URL <http://aphasiology.pitt.edu/archive/00002396/>, retrieved 6 August 2015.
- Sandhofer, Catherine. 2001. Structure in parents' input. In *Proceedings of the 25th annual Boston University Conference on Language Development*, 657–667.
- Santesteban, Mikel, Martin Pickering, and Holly Branigan. 2013. The effects of word order on subject-verb and object-verb agreement. *Journal of Memory and Language* 68:160–179.
- Saylor, Megan, Mark Somanader, Daniel Levin, and Kazuhiko Kawamura. 2010. How do young children deal with hybrids of living and non-living things. *British Journal of Developmental Psychology* 28: 835–885.
- Scerif, Gaia, Annette Karmiloff-Smith, Ruth Campos, Mayada Elsabbagh, Jon Driver, and Kim Cornish. 2005. To look or not to look? Typical and atypical development of oculomotor control. *Journal of Cognitive Neuroscience* 17.4:591–604.
- Schneider, Walter, Amy Eschman, and Anthony Zuccolotto. 2002. *E-Prime reference guide*. Psychology Software Tools, Inc., Pittsburgh. URL <http://step.psy.cmu.edu/materials/manuals/reference.pdf>, retrieved 22 January 2015.
- Scholes, Robert. 1981. Developmental comprehension of third person personal pronouns in English. *Language and Speech* 24.1:91–98.
- Schwartz, Richard. 1980. Presuppositions and children's metalinguistic judgments. *Child Development* 51.2:364–371.
- Selkirk, Elisabeth. 1980. *The phrase phonology of English and French*. New York: Garland.
- Selkirk, Elisabeth. 1995. The prosodic structure of function words. In *Papers in Optimality Theory*, eds. Jill Beckman, Laura Walsh Dickey, and Suzanne Urbanczyk, 439–469. Amherst: University of Massachusetts Graduate Linguistic Student Association.

- Shimpi, Priya, Perla Gámez, Janellen Huttenlocher, and Marina Vasilyeva. 2007. Syntactic priming in 3- and 4-year-old children: evidence for abstract representations of transitive and dative forms. *Developmental Psychology* 43.6:1334–1346.
- Shirakawa, Mineko. 2013. Experimental study of morphological case marking knowledge in Japanese-English bilingual children in Christchurch New Zealand. Master's thesis, University of Canterbury. URL <http://hdl.handle.net/10092/8715>, retrieved 11 April 2014.
- Siewierska, Anna, and Willem Hollmann. 2007. Ditransitive clauses in English with special reference to Lancashire dialect. In *Structural-functional studies in English Grammar*, eds. Mike Hannay and Gerard Steen, 83–102. Amsterdam: Benjamins.
- Sinclair, John, Gwyneth Fox, Stephen Bullon, Ramesh Krishnamurthy, Elizabeth Manning, John Todd, Mona Baker, Jane Bradbury, Richard Fay, and Deborah Yuill, eds. 1990. *Collins Cobuild English grammar*. London: HarperCollins.
- Skemp, Joseph. 1952. *Plato's Statesman*. London: Routledge.
- Sloutsky, Vladimir, and Anna Fisher. 2011. The development of categorization. In *The psychology of learning and motivation*, ed. Brian Ross, 141–166. Amsterdam: Elsevier.
- Snider, Neal. 2011. Investigating syntactic persistence in corpora. In *Language from a cognitive perspective*, eds. Emily Bender and Jennifer Arnold, 247–268. Stanford: CSLI.
- Stallings, Lynne, and Maryellen MacDonald. 2011. It's not just the "heavy NP": relative phrase length modulates the production of heavy-NP shift. *Journal of Psycholinguistic Research* 40:177–187.
- Stallings, Lynne, Maryellen MacDonald, and Pádraig O'Seaghdha. 1998. Phrasal ordering constraints in sentence production. *Journal of Memory and Language* 39:392–417.
- Stephens, Nola. 2010. Given-before-new: the effects of discourse on argument structure in early child language. Doctoral Dissertation, Stanford University. URL <http://www.personal.psu.edu/nms20/stephens2010.pdf>, retrieved 12 April 2013.
- Stevenson, Andrew, Chenhao Chiu, Dana Maslovat, Romeo Chua, Bryan Gick, Jean-Sébastien Blouin, and Ian Franks. 2014. Cortical involvement in the StartReact effect. *Neuroscience* 269:21–34.
- Stølum, Hans-Henrik. 1996. River meandering as a self-organizing process. *Science* 271.5256: 1710–1713. URL <http://dx.doi.org/10.1126/science.271.5256.1710>.
- Subrahmanyam, Kaveri, Rochel Gelman, and Alyssa Lafosse. 2003. Animates and other separably moveable objects. In *Category-specificity in brain and mind*, eds. Emer Forde and Glyn Humphreys, 341–371. Hove: Psychology Press.
- Sukumar, Sharath. 2012. System and method for reducing the effects of inadvertent touch on a touch screen controller. URL https://www.lens.org/lens/patent/US_2014_0043241_A1, patent application US 2014/0043241 A1.
- Sutton, Jennifer. 2006. The development of landmark and beacon use in young children. *Developmental Science* 9.1:108–123.
- Szabóné Papp, Judit. 2003. Reconsidering the dative shift. Doctoral Dissertation, University of Debrecen. URL <http://hdl.handle.net/2437/79661>, retrieved 24 May 2013.
- Szmrecsányi, Benedikt. 2004. On operationalizing syntactic complexity. In *Le poids des mots. Proceedings of the 7th International Conference on Textual Data Statistical Analysis*, eds. Gérard

- Purnelle, Cédric Fairon, and Anne Dister, volume Second, 1032–1039. Louvain-la-Neuve: Presses universitaires de Louvain. URL <http://www.benszm.net/omnibuslit/Szmrecsanyi2004.pdf>, retrieved 21 May 2013.
- Taylor, John. 1995. *Linguistic categorization*. Oxford: Oxford University Press, second edition.
- Thal, Donna, and Melanie Flores. 2001. Development of sentence interpretation strategies by typically developing and late-talking toddlers. *Journal of Child Language* 28:173–193.
- Theijssen, Daphne. 2009. Variable selection in logistic regression: The British English dative alternation. In *Proceedings of the 14th Student Session at the 21st European Summer School in Logic, Language and Information*, 85–95.
- Theijssen, Daphne, Hans van Halteren, Karin Fikkers, Frederike Groothoff, Lian van Hoof, Eva van de Sande, Jorieke Tiems, Véronique Verhagen, and Patrick van der Zande. 2009. A regression model for the english benefactive alternation. In *Proceedings of the 19th Meeting of Computational Linguistics in the Netherlands*, eds. Barbara Plank, Erik Tjong Kim Sang, and Tim van de Cruys, 115–130.
- Thura, David, Driss Boussaoud, and Martine Meunier. 2008. Hand position affects saccadic reaction times in monkeys and humans. *Journal of Neurophysiology* 99:2194–2202.
- Tobii Technology AB. 2008. *Tobii X60 & X120 eye trackers user manual (manual revision 3.0)*.
- Tobii Technology AB. 2011. *Tobii X60/X120 mobile device testing solution user manual (manual revision 1)*. URL <http://www.tobii.com/eye-tracking-research/global/products/hardware-accessories/tobii-mobile-device-stand/>, retrieved 22 January 2015.
- Tomasello, Michael. 2000. Do young children have adult syntactic competence? *Cognition* 74:209–253.
- Tomasello, Michael. 2003. *Constructing a language*. Cambridge, MA: Harvard University Press.
- Trott, Kate, Sushie Dobbinson, and Patrick Griffiths, eds. 2004. *The child language reader*. London: Routledge.
- Valian, Virginia. 1986. Syntactic categories in the speech of young children. *Developmental Psychology* 22.4:562–579.
- Venables, Bill. 2010. *PolynomF: Polynomials in R*. URL <http://CRAN.R-project.org/package=PolynomF>.
- Vinther, Thora. 2002. Elicited imitation: a brief overview. *International Journal of Applied Linguistics* 12.1:54–73. URL <http://dx.doi.org/10.1111/1473-4192.00024>.
- Warden, David. 1976. The influence of context on children's use of identifying expressions and references. *British Journal of Psychology* 67.1:101–112.
- Wasow, Thomas. 1997. Remarks on grammatical weight. *Language Variation and Change* 9.1:81–105.
- Wasow, Thomas. 2002. *Postverbal behavior*. Stanford: CSLI.
- Waxman, Sandra. 1999. The dubbing ceremony revisited: object naming and categorization in infancy and early childhood. In *Medin and Atran (1999)*, 233–284.
- Way, David, and Joseph Paradiso. 2014. A usability user study concerning free-hand microgesture and wrist-worn sensors. In *Proceedings of the 11th International Conference on Wearable and Implantable Body Sensor Networks*. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6855631, retrieved 27 January 2015.

- Whitford, Veronica, and Debra Titone. 2012. Second-language experience modulates first- and second-language word frequency effects. *Psychonomic Bulletin & Review* 19.1:73–80.
- Whong-Barr, Melinda, and Bonnie Schwartz. 2002. Morphological and syntactic transfer in child L2 acquisition of the English dative alternation. *Studies in Second Language Acquisition* 24:579–616.
- Wickham, Hadley. 2009. *ggplot2: elegant graphics for data analysis*. New York: Springer. URL <http://had.co.nz/ggplot2/book>.
- Wickham, Hadley. 2011. The split-apply-combine strategy for data analysis. *Journal of Statistical Software* 40.1:1–29. URL <http://www.jstatsoft.org/v40/i01/>.
- Williams, Molly. 2014. Normally fluent preschoolers' response to linguistic complexity. Undergraduate honors thesis, University of Redlands.
- Wolk, Christoph, Joan Bresnan, Anette Rosenbach, and Benedikt Szmrecsányi. 2013. Dative and genitive variability in Late Modern English: Exploring cross-constructural variation and change. *Diachronica* 30.3:382–419.
- Wood, Justin, Sid Kouider, and Susan Carey. 2009. Acquisition of singular-plural morphology. *Developmental Psychology* 45.1:202–206.
- Yan, Xin, and Xiaogang Su. 2009. *Linear regression analysis*. Singapore: World Scientific.
- Yang, Charles. 2004. Toward a theory of language growth. In Jenkins (2004), 37–56.
- Yonge, Charles. 1853. *Diogenes Laertius' "Lives and opinions of eminent philosophers"*. London: Bell.
- Zapf, Jennifer, and Linda Smith. 2009. Knowing more than one can say: the early regular plural. *Journal of Child Language* 36:1145–1155.
- Zec, Draga, and Sharon Inkelas. 1990. Prosodically constrained syntax. In *The phonology-syntax connection*, eds. Sharon Inkelas and Draga Zec, 365–378. Chicago: University of Chicago Press.
- Zehler, Annette, and William Brewer. 1980. Acquisition of the article system in English. Technical report 171, Center for the Study of Reading (University of Illinois at Urbana-Champaign and Bolt Beranek and Newman Inc.).
- Zhang, Shuo, Rafael Caro Repetto, and Xavier Serra. 2014. Study of the similarity between linguistic tones and melodic pitch contours in Beijing Opera singing. In *Proceedings of the Fifteenth International Society for Music Information Retrieval Conference*.
- van der Ziel, Maria Elisabeth. 2012. The acquisition of scope interpretation in dative constructions. Doctoral Dissertation, University of Utrecht.
- Zovko Dinković, Irena. 2007. Dative alternation in Croatian. *Suvremena lingvistika* 63:65–83.

Appendices

A Human ethics

Approval for the research involving human participants described in this thesis was granted by the Human Ethics Committee of the University of Canterbury under the following reference numbers:

- HEC 2012/172
- HEC 2013/31/LR-PS
- HEC 2013/166

As per the regulations for PhD & Staff Low-Risk Applications current at the time of submission of the application for HEC 2013/31/LR-PS, that application is reproduced on the following pages.

UNIVERSITY OF CANTERBURY
HUMAN ETHICS COMMITTEE



PhD & STAFF LOW RISK APPLICATION

(For research proposals which are not considered in full by the University Human Ethics Committee)

ETHICAL APPROVAL OF LOW RISK RESEARCH INVOLVING
HUMAN PARTICIPANTS REVIEWED BY DEPARTMENTS

Please read the important notes appended to this form before completing the sections below

- 1 RESEARCHER'S NAME: Daniel Bürkle
- 2 NAME OF DEPARTMENT OR SCHOOL: Linguistics
- 3 EMAIL ADDRESS: daniel.buerkle@pg.canterbury.ac.nz
- 4 TITLE OF PROJECT: The acquisition of sentence alternations (pilot study)
- 5 PROJECTED START DATE OF PROJECT: September 2013
- 6 STAFF MEMBER/SUPERVISOR RESPONSIBLE FOR PROJECT: Dr Heidi Quinn
- 7 NAMES OF OTHER PARTICIPATING STAFF AND STUDENTS: Professor Jen Hay, Dr Susan Foster-Cohen (associate supervisors)
- 8 STATUS OF RESEARCH: pilot study in PhD thesis
- 9 BRIEF DESCRIPTION OF THE PROJECT:
Please give a brief summary (approx. 500 words) of the nature of the proposal in lay language, including the aims/objectives/hypotheses of the project, rationale, participant description, and procedures/methods of the project:

The aim of this pilot study is to explore the validity of a research protocol designed to investigate the effects of animacy, grammatical number, and word length on the English dative alternation (which is the choice between sentences like "give the coffee to Kate" and "give Kate the coffee"). Specifically, this pilot study is designed to find out which groups of pictures are well-suited for the main experiment, which will show pictures in sets of four. In that main experiment, participants will be asked to "give" (by dragging) one given picture to their choice out of the three others, or to choose one out of three and "give" it to the given goal picture. A world-knowledge connection between two pictures (for example, "dog" and "bone") would skew the results there, because this connection would override the subconscious/non-intentional grammatical preferences that are the focus of the main study. The connection between "dog" and "bone" is obvious, of course; but there may well be similarly strong, but less obvious connections between other pictures in this task that would affect people's choices in the main experiment. Therefore, any such connections must be ruled out beforehand. This pilot study will present groups of pictures in the same way as they would be presented in the main experiment, but without the instructions. A previous pilot study (HEC 2012/172; same methodology, but with children as participants) suggests that children show a significant preference for animacy matching: animate choices were preferred with animate goals, and inanimate choices were preferred with inanimate goals. It is important to explore whether adults exhibit a similar preference, because such a preference could skew the results of the later experiment. The hypotheses (and motivation for the pilot) are (1) that there are some world knowledge connections between pictures in the task that may not be obvious to the researcher, but could affect participant responses to the main experiment that this study is a pilot to; and (2) that adults will show a preference for animacy matching in the pilot task.

Participants for this pilot study will be 20 adult speakers of New Zealand English (10 female, 10 male) without language disorders. They will be recruited through university classes and notices on campus. Participants will be asked to choose one of three pictures shown on a touchscreen computer and drag their choice to a target picture. There will be 64 trials; this is unlikely to take more than 20 minutes. The data collected will be the chosen picture for each trial, the time taken to do so, and the “path” of the dragging finger on the touchscreen.

10 WHY IS THIS A LOW RISK APPLICATION?

There is no danger to or deception of participants.

This pilot study extends an already approved pilot study (HEC 2012/172, where children were the participants) to adult participants without language disorders who are able to give informed consent (see attached forms).

Data will be stored for no more than ten years in locked drawers in lockable offices in the Department of Linguistics and in password-protected and encrypted computer files on the researcher’s “P: drive” (data storage system provided by the University), and any publications resulting from this research or future research based on this research will not allow identification of individual participants.

No Treaty of Waitangi issues present themselves.

No conflicts of interest arise.

Participants will not be accessed through other individuals or bodies, so no permission is needed.

No inducements will be offered.

11 PROVIDE COPIES OF INFORMATION & CONSENT FORMS FOR PARTICIPANTS

These forms should be on University of Canterbury departmental letterhead. The name of the project, name(s) of researcher(s), contact details of researchers (and for PhD students, the supervisor), names of who has access to the data, the length of time the data is to be stored, that participants have the right to withdraw participation and data provided, and what the data will be used for should all be clearly stated. A statement that the project has been reviewed approved by the appropriate department and the UCHEC Low Risk Approval process should also be included.

Please ensure that Section A (where appropriate), B and C below are all completed

Applicant's Signature: [Signature] Date 13/8/2013

A SUPERVISOR'S DECLARATION FOR PhD RESEARCH:

- 1 I have made the applicant fully aware of the need for and the requirement of seeking HEC approval for research involving human participants.
- 2 I have ensured the applicant is conversant with the procedures involved in making such an application.
- 3 In addition to this form the applicant has individually filled in the full application form which has been reviewed by me.

Signed (Supervisor): [Signature: Heidi A.] Date 13-8-2013

B SUPPORTED BY THE DEPARTMENTAL/SCHOOL RESEARCH COMMITTEE:

Name LYNN CLARK

Signature: [Signature] Date 14/8/2013

C APPROVED BY HEAD OF DEPARTMENT/SCHOOL:

Name Kevin Watson

Signature: [Signature] Date 13-8-2013

SUBMISSION OF APPLICATION:

- Please attach copies of any Information Sheet and Consent Form
- Forward two hard copies to: The Secretary, Human Ethics Committee, Okeover House
- Forward an electronic copy to: human-ethics@canterbury.ac.nz

NOTES ON PROCEDURE:

The Chair of the University of Canterbury Human Ethics Committee and two other Human Ethics Committee members will review this application.

In normal circumstances queries will be forwarded via email to the applicant within 7 days

If you are a PhD student, please include a copy of this form as an appendix in your thesis

ACTION TAKEN BY HUMAN ETHICS COMMITTEE:

- Added to PhD & Staff Low Risk Reporting Database
- Referred to University of Canterbury HEC
- Referred to another Ethics Committee – please specify:

.....

REVIEWED BY: (HEC Chair)

..... (HEC Member)

..... (HEC Member)

Date

NOTES CONCERNING LOW RISK REPORTING SHEETS

1. This form should only be used for proposals which are Low Risk as defined in the University of Canterbury Human Ethics Committee Principles and Guidelines policy document and which may therefore be properly considered and approved at departmental level and by the Chair and two members of the University of Canterbury Human Ethics Committee under Section 5 of that document.
2. Low Risk applications are:

PhD thesis, pilot studies and staff research where the projects do not raise any issue of deception, threat, invasion of privacy, mental, physical or cultural risk or stress, and do not involve gathering personal information of a sensitive nature about or from individuals.
3. No research can be counted as low risk if it involves:
 - (i) invasive physical procedures or potential for physical harm
 - (ii) procedures which might cause mental/emotional stress or distress, moral or cultural offence
 - (iii) personal or sensitive issues
 - (iv) vulnerable groups
 - (v) Tangata Whenua
 - (vi) cross cultural research
 - (vii) investigation of illegal behaviour(s)
 - (viii) invasion of privacy
 - (ix) collection of information that might be disadvantageous to the participant
 - (x) use of information already collected that is not in the public arena which might be disadvantageous to the participant
 - (xi) use of information already collected which was collected under agreement of confidentiality
 - (xii) participants who are unable to give informed consent
 - (xiii) conflict of interest e.g. the researcher is also the lecturer, teacher, treatment-provider, colleague or employer of the research participants, or there is any other power relationship between the researcher and the research participants.
 - (xiv) deception
 - (xv) audio or visual recording without consent
 - (xvi) withholding benefits from "control" groups
 - (xvii) inducements
 - (xviii) risks to the researcher

This list is not definitive but is intended to sensitise the researcher to the types of issues to be considered. Low risk research would involve the same risk as might be encountered in normal daily life.

4. Responsibility

Supervisors are responsible for:

Theses where the projects do not raise any issues listed below.

Heads of Department are responsible for:

- (i) Giving final approval for the low risk application.
- (ii) Ensuring a copy of all applications are kept on file in the Department/School.

NOTE: If the HOD is the applicant, then a senior member of staff and preferably also the department and/or school research committee should give final approval. The HOD is still responsible for (ii) above.

4. A separate low risk form should be completed for each research proposal involving human participants and for which ethical approval has been considered or given at Departmental level.
5. Two completed and signed Application forms, together with a copies of Information Sheets and/ or Consent Forms, should be submitted to the Secretary, Human Ethics Committee, Okeover House, as soon as the proposal has been considered at departmental level. Please also submit an electronic version to human-ethics@canterbury.ac.nz.
6. The Information Sheet and Consent Form include the statement "This proposal has been reviewed and approved by the University of Canterbury Human Ethics Committee low risk process".

PhD & Staff Low Risk Application Form

7. Please ensure the Consent Form and the Information Sheet are on University of Canterbury letterhead and have been carefully proof-read; the institution as a whole is likely to be judged by them.
9. The research must be consistent with the University of Canterbury Human Ethics Committee Principles and Guidelines. Refer to the appendices of the University of Canterbury Human Ethics Committee Principles and Guidelines for guidance on information sheets and consent forms.
10. Please note that if the nature, procedures, location or personnel of the research project changes after departmental approval has been given in such a way that the research no longer meets the conditions laid out in Section 5 of the Principles and Guidelines, a full application to the Human Ethics Committee must be submitted.
11. This form is available electronically at: <http://www.canterbury.ac.nz/humanethics>

CHECKLIST

Please check that your application/summary has discussed:

- Procedures for voluntary, informed consent
- Privacy & confidentiality
- Risk to participants
- Obligations under the Treaty of Waitangi
- Needs of dependent persons
- Conflict of interest
- Permission for access to participants from other individuals or bodies
- Inducements

In some circumstances research which appears to meet low risk criteria may need to be reviewed by the University of Canterbury Human Ethics Committee. This might be because of requirements of:

- The publisher of the research
- An organisation which is providing funding resources, existing data, access to participants etc.
- Research which meets the criteria for review by a Health and Disability Ethics Committee – see HRC web site.

If you require advice on the appropriateness of research for low risk review, please contact the Chair of the University of Canterbury Human Ethics Committee

B Blocks and trials in experiment 2

recipient	animacy-matching option	theme	number-matching theme option	length-matching theme option
cows	penguin		baskets	hat
dogs	camel		baskets	lock
frog	squirrels		bottle	hats
bear	penguins		letter	balls
hedgehogs	bee		hats	basket
monkeys	crab		balls	basket
kiwi	pigs		hat	bottles
rabbit	pigs		ball	bottles
keys	letter		squirrels	crab
pears	bottle		penguins	crab
pot	letters		camel	bees
shirt	letters		camel	crabs
hammers	ball		pigs	squirrel
lemons	lock		pigs	squirrel
pillow	locks		bee	penguins
pencil	locks		bee	camels

Table B.1: Experiment 2, block 1: prepositional instruction sentence with gap in place of the theme (*Now give the _____ to the cows.*)

theme	animacy-matching recipient option	number-matching recipient option	length-matching recipient option
cows	penguin	baskets	hat
dog	camels	basket	locks
frogs	squirrel	bottles	hat
bear	penguins	letter	balls
hedgehogs	bee	hats	basket
monkey	crabs	ball	baskets
kiwis	pig	hats	bottle
rabbit	pigs	ball	bottles
keys	letter	squirrels	crab
pear	bottles	penguin	crabs
pots	letter	camels	bee
shirt	letters	camel	crabs
hammers	ball	pigs	squirrel
lemon	locks	pig	squirrels
pillows	lock	bees	penguin
pencil	locks	bee	camels

Table B.2: Experiment 2, block 2: double object instruction sentence with gap in place of the recipient (*Now give the _____ the cows.*)

recipient	animacy-matching option	theme	number-matching theme option	length-matching theme option
cow	penguins		basket	hats
dogs	camel		baskets	lock
frog	squirrels		bottle	hats
bears	penguin		letters	ball
hedgehog	bees		hat	baskets
monkeys	crab		balls	basket
kiwi	pigs		hat	bottles
rabbits	pig		balls	bottle
key	letters		squirrel	crabs
pears	bottle		penguins	crab
pot	letters		camel	bees
shirts	letter		camels	crab
hammer	balls		pig	squirrels
lemons	lock		pigs	squirrel
pillow	locks		bee	penguins
pencils	lock		bees	camel

Table B.3: Experiment 2, block 3: double object instruction sentence with gap in place of the theme (*Now give the cow the _____.*)

theme	animacy-matching recipient option	number-matching recipient option	length-matching recipient option
cow	penguins	basket	hats
dog	camels	basket	locks
frogs	squirrel	bottles	hat
bears	penguin	letters	ball
hedgehog	bees	hat	baskets
monkey	crabs	ball	baskets
kiwis	pig	hats	bottle
rabbits	pig	balls	bottle
key	letters	squirrel	crabs
pear	bottles	penguin	crabs
pots	letter	camels	bee
shirts	letter	camels	crab
hammer	balls	pig	squirrels
lemon	locks	pig	squirrels
pillows	lock	bees	penguin
pencils	lock	bees	camel

Table B.4: Experiment 2, block 4: prepositional instruction sentence with gap in place of the recipient (*Now give the cow to the _____.*)

C Images used in experiment 2

In addition to own works, the following photographs were modified and used to represent the respective noun in the experiments described here.

apple: User “Pupuli”. 10 February 2008. Apples. <http://www.flickr.com/photos/10943651@N00/3044525754>. Licensed under [Creative Commons Attribution-NonCommercial-ShareAlike 2.0](#).



ball: User “bpariedl”. 23 August 2009. Einsamer Ball. <http://www.flickr.com/photos/bpariedl/3861962002>. Licensed under [Creative Commons Attribution-NonCommercial-ShareAlike 2.0](#).

balls: User “Joe Shlabotnik”. 1 November 2012. Soccer Balls. <http://www.flickr.com/photos/joeshlabotnik/8168644501>. Licensed under [Creative Commons Attribution-NonCommercial-ShareAlike 2.0](#).



basket: User “Susann_Schweden”. 18 July 2011. A basket. http://commons.wikimedia.org/wiki/File:A_basket.jpg. Licensed under [Creative Commons Attribution-ShareAlike 3.0 Unported](#).

baskets: User “Oop”. 20 July 2012. Korviväljapanek Tartu hansapäevade Jõelinnas. http://commons.wikimedia.org/wiki/File:Korviv%C3%A4ljapanek_Tartu_hansap%C3%A4evade_J%C3%B5elinnas,_20._juuli_2012.jpg. Licensed under [Creative Commons Attribution-ShareAlike 3.0 Unported](#).



bear: User “Bastet78”. 15 November 2006. Niedźwiedź brunatny (Ursus arctos L.)Ursus arctos. http://commons.wikimedia.org/wiki/File:6_maja_06_r._Z00_161.jpg. Licensed under [Creative Commons Attribution-ShareAlike 3.0 Unported](#).

bears:

- Annette Elisabeth Rudolph. 17 February 2008. mmmm...smelling good today...new perfume ?. http://www.flickr.com/photos/a_rud_beth/2273717093. Licensed under [Creative Commons Attribution-NonCommercial-ShareAlike 2.0](#).
- User “Hswaton”. 9 May 2013. Baer im Alpenzoo. <http://commons.wikimedia.org/wiki/File:Baer1.jpg>. Licensed under [Creative Commons Attribution 3.0 Unported](#).



bee: Marc Andrighetti. 5 May 2012. Abeille butineuse et son pollen. http://commons.wikimedia.org/wiki/File:Abeille_butineuse_et_son_pollen.JPG. Licensed under Creative Commons Attribution-ShareAlike 3.0 Unported.

bees: User "Wiredwitch". 28 February 2008. Wild bee hive. <http://www.flickr.com/photos/60509459@N00/2302523896>. Licensed under Creative Commons Attribution-ShareAlike 2.0.



bottle: User "keepps". 18 August 2007. green glass. <http://www.flickr.com/photos/isg-online/1167028571>. Licensed under Creative Commons Attribution-NonCommercial-ShareAlike 2.0.

bottles: Matthew Solle. 1 September 2012. (Bottles) Still Life. http://www.flickr.com/photos/matthew_solle/7995697812. Licensed under Creative Commons Attribution 2.0.



camel: User “Bouette”. 11 October 2006. Chameau de Bactriane. http://commons.wikimedia.org/wiki/File:Chameau_de_bactriane.JPG. Licensed under Creative Commons Attribution-ShareAlike 3.0 Unported.

camels: User “bobosh_t”. 10 June 2011. Three Wisemen’s camels. <http://www.flickr.com/photos/frted/5823449584>. Licensed under Creative Commons Attribution-ShareAlike 2.0.



cherry: Lisa Connolly. 20 July 2005. Cherry. <http://www.flickr.com/photos/lisaconnolly/27339281>. Licensed under Creative Commons Attribution-NonCommercial-ShareAlike 2.0.



cow: User “Roland zh”. 1 October 2010. Forch towards Pfannenstiel Guldenen (Switzerland). http://commons.wikimedia.org/wiki/File:Forch_-_Pfannenstiel_2010-10-01_14-22-10.JPG. Licensed under Creative Commons Attribution-ShareAlike 3.0 Unported.

cows:

- User “Jsome1”. 21 July 2008. Azores. <http://www.flickr.com/photos/jsome1/3240489010>. Licensed under Creative Commons Attribution 2.0.
- USDA. Pre-November 2003. Holstein cows large. http://en.wikipedia.org/wiki/File:Holstein_cows_large.jpg. Public domain work.
- Anna Briggs. 5 May 2007. Cow on Kegworth canal bridge. <http://www.flickr.com/photos/anna-b/3160390295>. Licensed under Creative Commons Attribution-NonCommercial 2.0.



crab: User “Bhny”. 27 April 2006. *Gecarcinus quadratus* (Halloween crab) at Nosara, Costa Rica. http://commons.wikimedia.org/wiki/File:Gecarcinus_quadratus_%28Nosara%29.jpg. Public domain work.

crabs: User “eileendelhi”. 6 August 2006. Fighters. <http://www.flickr.com/photos/eileendelhi/212775704>. Licensed under Creative Commons Attribution-NonCommercial 2.0.



cup: Louis Kreusel. 22 November 2005. Teacup. <http://www.flickr.com/photos/louiskreusel/82682900>. Licensed under [Creative Commons Attribution-NonCommercial 2.0](#).



dog: Małgorzata Miłaszewska-Duda. 2 October 2011. Bernese Mountain Dog. http://commons.wikimedia.org/wiki/File:Bernese_Mountain_Dog_Miedzynarodowa_wystawa_psow_rasowych_rybnik_kamien_pazdziernik_2011_1.jpg. Licensed under [Creative Commons Attribution-ShareAlike 3.0 Unported](#).

dogs: User “barrio dude”. 2 July 2007. i want one!. <http://www.flickr.com/photos/97111149@N00/751565592>. Licensed under [Creative Commons Attribution 2.0](#).



duck: User “Paula”. 17 February 2008. Four ducks. http://commons.wikimedia.org/wiki/File:Four_ducks.jpg. Licensed under [Creative Commons Attribution-ShareAlike 2.0](#).



fox: Edd Deane. 14 August 2011. Renard roux marchant lentement dans l'eau. <http://commons.wikimedia.org/wiki/File:Renardriviere.jpg>. Licensed under [Creative Commons Attribution 2.0](#).



frog: User “Contrabaroness”. 21 May 2010. The Northern Green Frog. http://commons.wikimedia.org/wiki/File:Northern_Green_Frog_-_Tewksbury,_NJ.jpg. Licensed under Creative Commons Attribution-ShareAlike 3.0 Unported.

frogs: User “4028mdk09”. 15 July 2011. Frösche, fotografiert in einem Teich mit Seerosen in Heidelberg (Baden-Württemberg, Deutschland). http://commons.wikimedia.org/wiki/File:Fr%C3%B6sche_in_einem_Seerosenteich.JPG. Licensed under Creative Commons Attribution-ShareAlike 3.0 Unported.



hammer: User “Tjcase2”. 22 October 2008. Standard Hammer. <http://commons.wikimedia.org/wiki/File:Hammer.JPG>. Licensed under Creative Commons Attribution-ShareAlike 3.0 Unported.

hammers: Mauro Cateb. 19 October 2010. Some hammers used in jewellery. http://commons.wikimedia.org/wiki/File:Jewellery_hammers_%281%29.jpg. Licensed under Creative Commons Attribution-ShareAlike 3.0 Unported.



hat: Clément Bucco-Lechat. 6 November 2012. Indiana Jones fedora hat. http://commons.wikimedia.org/wiki/File:Indiana_Jones_fedora_hat.jpg. Licensed under Creative Commons Attribution-ShareAlike 3.0 Unported.

hats: Jorge Royan. 26 April 2008. Hat stall owner in a sunday fair. Amsterdam, The Netherlands. http://commons.wikimedia.org/wiki/File:Amsterdam_-_Hats_-_0932.jpg. Licensed under Creative Commons Attribution-ShareAlike 3.0 Unported.



hedgehog: Jan-Erik Finnberg. 17 July 2006. A hedgehog. <http://www.flickr.com/photos/17424735@N00/2252528405>. Licensed under Creative Commons Attribution 2.0.

hedgehogs:

- User “Nopple”. 3 July 2013. Hedgehogs in Garden. http://commons.wikimedia.org/wiki/File:Hedgehogs_Siesta.JPG. Licensed under Creative Commons Zero 1.0.
- Mikael Miettinen. 20 June 2008. #119 Hedgehog. <http://www.flickr.com/photos/30261851@N06/3324010908>. Licensed under Creative Commons Attribution 2.0.



key: Honza Groh. 23 August 2008. Keys from padlocks. http://commons.wikimedia.org/wiki/File:Klice_od_visacu.jpg. Licensed under Creative Commons Attribution 3.0 Unported.

keys: Tomasz Sienicki. 4 November 2004. padlock. http://commons.wikimedia.org/wiki/File:Padlock_kl%C3%B3dka_ubt.JPG.



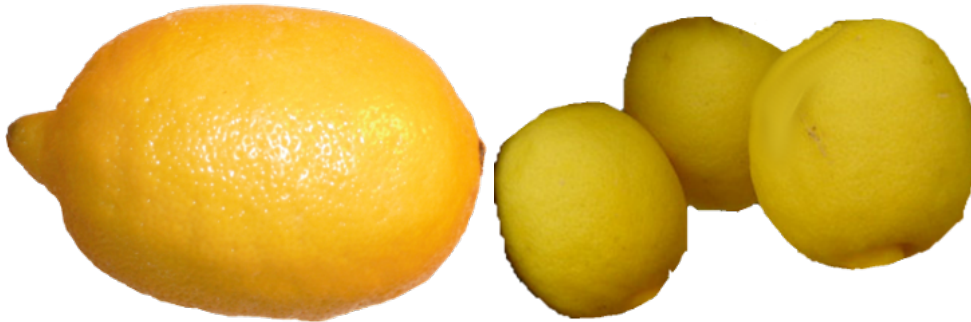
kiwi: User “The.Rohit”. 13 May 2010. Apteryx mantelli. http://en.wikipedia.org/wiki/File:Apteryx_mantelli_-_Rotorua,_North_Island,_New_Zealand-8a.jpg. Licensed under Creative Commons Attribution 2.0.

kiwis: User “barb”. 11 July 2010. Kiwi: right strange looking birds. <http://www.flickr.com/photos/barb/4783570229>. Licensed under Creative Commons Attribution-NonCommercial-ShareAlike 2.0.



lemon: Gábor Hanák. 24 December 2009. A big yellow lemon. <http://commons.wikimedia.org/wiki/File:Citrom.jpg>. Licensed under [Creative Commons Attribution-ShareAlike 3.0 Unported](#).

lemons: User “Dodo-Bird”. 16 April 2006. Lemons. <http://www.flickr.com/photos/dodo-bird/477499717>. Licensed under [Creative Commons Attribution 2.0](#).



letter: User “sludgegulper”. 17 January 2011. Briefumschlag VEB Schwerarmaturenwerk ERICH WEINERT Magdeburg. <http://www.flickr.com/photos/sludgegulper/5362224401>. Licensed under [Creative Commons Attribution-ShareAlike 2.0](#).

letters: User “Bahar101”. 22 July 2007. SayilmisSecimZarflari. <http://commons.wikimedia.org/wiki/File:SayilmisSecimZarflari.JPG>. Licensed under [Creative Commons Attribution-ShareAlike 3.0 Unported](#).



lock: Nino Barberi. March 2007. Padlock. http://commons.wikimedia.org/wiki/File:-_Padlock_- .jpg. Licensed under [Creative Commons Attribution-ShareAlike 2.5](#).

locks: Mike Baird. 22 March 2008. Multiple Padlock Farm Gate Mechanism. <http://www.flickr.com/photos/mikebaird/2354116406>. Licensed under [Creative Commons Attribution 2.0](#).



monkey: User “whologwhy”. 29 July 2011. MONKEY BUSINESS. <http://www.flickr.com/photos/hulagway/6010643359>. Licensed under [Creative Commons Attribution 2.0](#).

monkeys:

- Daisy Seneviratne. 3 November 2011. Monkey Grooming. <http://www.flickr.com/photos/daisysen/6369979195>. Licensed under [Creative Commons Attribution-Non-Commercial-ShareAlike 2.0](#).
- User “BotheredByBees”. 30 November 2007. monkey. <http://www.flickr.com/photos/botheredbybees/2081791806>. Licensed under [Creative Commons Attribution 2.0](#).



pear: User “Genet”. 4 October 2008. Birnen der Sorte ‘Prinzessin Marianne’. http://commons.wikimedia.org/wiki/File:Pyrus_-_Prinzessin_Marianne_-_Kaiserkrone.JPG. Licensed under Creative Commons Attribution-ShareAlike 3.0 Unported.

pears: USDA. No date. Image number K5302-1. http://commons.wikimedia.org/wiki/File:More_pears.jpg. Public domain work.



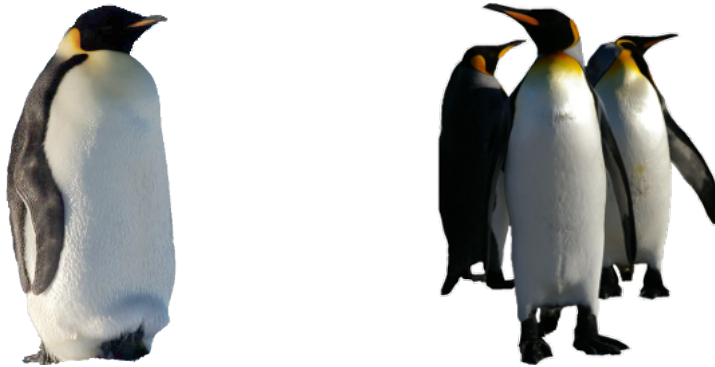
pencil: User “Démsthène”. 24 June 2012. Matériel de dessinateur: gomme ‘Mie de pain’ et crayon HB. http://commons.wikimedia.org/wiki/File:Crayon%2Bmie_de_pain.JPG. Licensed under Creative Commons Attribution-ShareAlike 3.0 Unported.

pencils: User “Baselmans”. 15 June 2007. Conte pencil. http://commons.wikimedia.org/wiki/File:Conte_pencil.jpg. Public domain work.



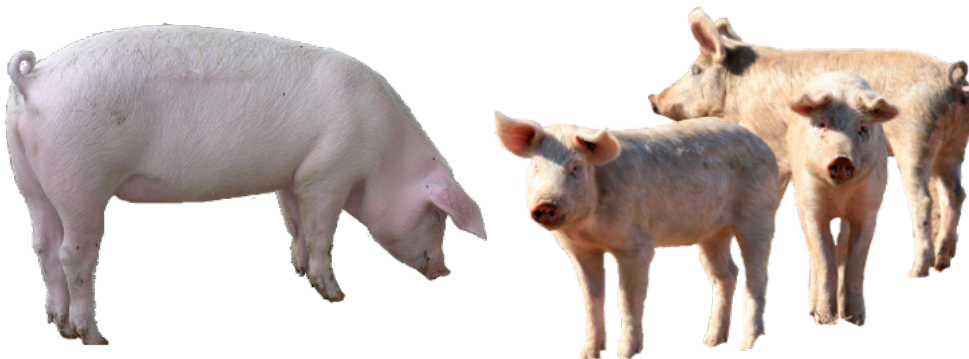
penguin: Hannes Grobe. 2004. Emperor Penguin, Atka Bay, Weddell Sea, Antarctica. http://commons.wikimedia.org/wiki/File:Emperor-cold_hg.jpg. Licensed under Creative Commons Attribution 3.0 Unported.

penguins: Ben Tubby. 20 March 2007. Kings on the beach. <http://www.flickr.com/photos/tubby/435342058>. Licensed under Creative Commons Attribution-NonCommercial 2.0.



pig: Joshua Lutz. 2 October 2004. *Sus scrofa scrofa*. http://commons.wikimedia.org/wiki/File:Sus_scrofa_scrofa.jpg. Public domain work.

pigs: Alex Proimos. 22 August 2011. Three Little Pigs. <http://www.flickr.com/photos/proimos/6073731653>. Licensed under Creative Commons Attribution-NonCommercial 2.0.



pillow: Pavel Ševela. 15 April 2011. Living room, Building Fairs Brno 2011. http://commons.wikimedia.org/wiki/File:Building_Fairs_Brno_2011_%28210%29.jpg. Licensed under Creative Commons Attribution-ShareAlike 3.0 Unported.

pillows: User “Andrevruas”. 20 February 2010. Almofadas verde, amarela e azul. <http://commons.wikimedia.org/wiki/File:Almofadas.jpg>. Licensed under Creative Commons Attribution 3.0 Unported.



pot: User “FiveRings”. 7 November 2006. Copper saucepot. <http://commons.wikimedia.org/wiki/File:Copper-saucepot.jpg>. Public domain work.

pots: User “Nemo bis”. 14 July 2011. Pentolini e tegami assortiti. http://commons.wikimedia.org/wiki/File:Pentolini_e_tegami_assortiti,_lungo.jpg. Licensed under Creative Commons Attribution-ShareAlike 3.0 Unported.



rabbit: Ann Harrison. 4 July 2009. A wild rabbit at Lossiemouth. <http://www.geograph.org.uk/photo/1441920>. Licensed under [Creative Commons Attribution-ShareAlike 2.0](#).

rabbits: Chris Downer. 16 June 2012. bunnies at the M3 services. <http://www.geograph.org.uk/photo/3007908>. Licensed under [Creative Commons Attribution-ShareAlike 2.0](#).



ruler: User "Grikalmis". 31 July 2008. Stahlmaßstab. <http://commons.wikimedia.org/wiki/File:Stahlma%C3%9F.jpg>. Public domain work.



shirt: McArthurGlen Designer Outlets. 1 February 2013. DesignerOutletSalzburg_Strenesse_-gestreiftes Hemd. <http://www.flickr.com/photos/93153549@N06/8467163511>. Licensed under [Creative Commons Attribution-ShareAlike 2.0](#).



snake: User “Chiswick Chap”. 2 September 2008. Sinuously meandering snake on road, Jianxia, China. http://commons.wikimedia.org/wiki/File:Jiangxia-snake-9704_%28cropped%29.jpg. Licensed under Creative Commons Attribution-ShareAlike 3.0 Unported.



squirrel: Hernan De Angelis Campephilus. 4 June 2008. Red squirrel (*Sciurus vulgaris*) on a tree. Bromma, Stockholm, Sweden. http://commons.wikimedia.org/wiki/File:Sciurus-vulgaris_hernandeangelis_stockholm_2008-06-04.jpg. Licensed under Creative Commons Attribution-ShareAlike 3.0 Unported.

squirrels:

- Kamil Gwóźdź. 23 August 2011. Wiewiórka pospolita. http://commons.wikimedia.org/wiki/File:Wiewi%C3%B3rka_pospolita2.jpg. Licensed under Creative Commons Attribution-ShareAlike 3.0 Unported.
- User “Rave”. 21 March 2011. Squirrel in Lazienki 2. http://commons.wikimedia.org/wiki/File:Squirrel_in_Lazienki_2.JPG. Licensed under Creative Commons Attribution-ShareAlike 3.0 Unported.
- User “Rave”. 21 March 2011. Squirrel in Lazienki 1. http://commons.wikimedia.org/wiki/File:Squirrel_in_Lazienki_1.JPG. Licensed under Creative Commons Attribution-ShareAlike 3.0 Unported.



D Sound files used in experiment 2

In addition to own works, the following sound files were modified and used in the experiments described here.

User “BristolStories”. 13 April 2008. u_chimes4. <http://www.freesound.org/people/BristolStories/sounds/51711>. Licensed under Creative Commons Attribution-NonCommercial 3.0 Unported.

User “tobyk”. 23 November 2006. Medieval_Fanfare. <http://www.freesound.org/people/tobyk/sounds/26198>. Licensed under Creative Commons Attribution 3.0 Unported.

User “dobroide”. 17 July 2005. violin.open.strings.chords. <http://www.freesound.org/people/dobroide/sounds/4503>. Licensed under Creative Commons Attribution 3.0 Unported.

User “dangerbabe”. 22 April 2007. Drum. <http://www.freesound.org/people/dangerbabe/sounds/34091>. Licensed under Creative Commons Attribution 3.0 Unported.

E Sentences and drawings used in experiment 3

The target sentences in Task 3 were presented in the order given below, each accompanied by the respective drawing. The drawings are reproduced here by kind permission of the artist, Laura McKinley.

1. Dad gave Anne the coat.



2. Anne gave the drawing to the parents.



3. Mom gave the cushions Anne.



4. Dad gave the parents to the chairs.



5. Mom gave the shelves the drawing.



6. Dad gave the toys to Anne.



7. The cat gave the basket the kittens.



8. Mom gave the baby to the toys.



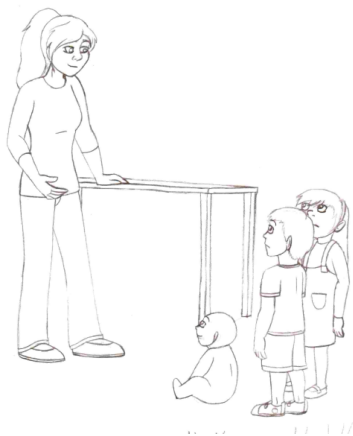
9. The parents gave the chairs the children.



10. Dad gave the kittens to the baby.



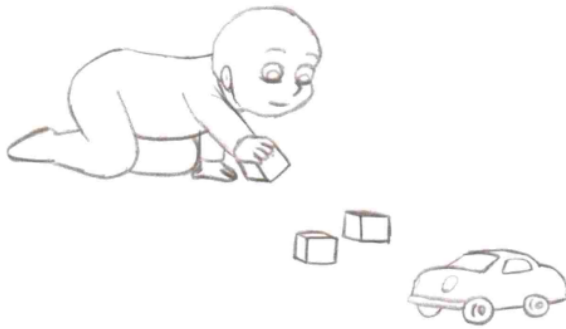
11. Mom gave the children the table.



12. Anne gave the cat to the parents.



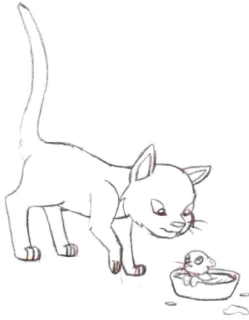
13. The baby gave the car the blocks.



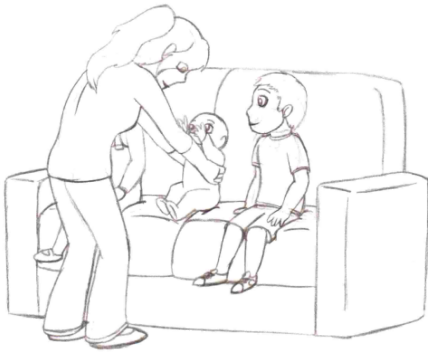
14. Dad gave the glasses to the children.



15. The cat gave the milk the kitten.



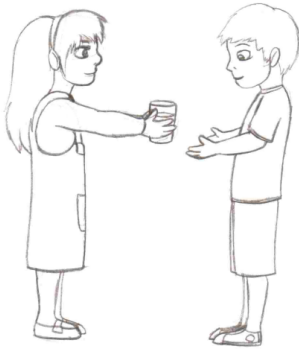
16. Mom gave the children to the sofa.



17. Ben gave the parents the kitten.



18. Anne gave the glass to Ben.



19. Dad gave the cat the kittens.



20. Anne gave the cat to the basket.



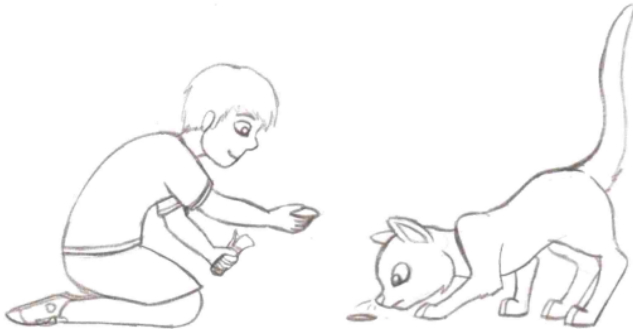
21. Mom gave the kittens the crackers.



22. The baby gave the cracker to the blocks.



23. Ben gave the cat the crackers.



24. The cat gave the crackers to the basket.



F Information sheets and consent forms

HEC2013/166

The acquisition of sentence alternations

Daniel Bürkle

☎ +64 3 364 2987 extension 8131

Email: daniel.buerkle@pg.canterbury.ac.nz



Information for adult participants

You are about to participate in a research project on how people use certain types of English sentences. Your involvement in this project will be to sort pictures according to two criteria, follow voice instructions while using a touch-screen, and repeat a short story told to you. This won't take more than an hour.

The data I will collect is this:

1. how you sorted the pictures
2. your touch-screen computer input
3. where you were looking during the touch-screen task
4. what you were saying in the story task
5. your age, gender, and what language(s) you speak.

You will get a \$10 voucher at the end of the experiment.

Please feel free to ask any questions you may have about the project now, or via email, post, or on the phone later.

You have the right to withdraw from the project at any time with no disadvantage, up until all individual data is entrenched in the general findings of the study and/or up until submission of the thesis or any articles stemming from it. If you withdraw, I will delete any information and data you have provided.

I expect the results of this project will help us understand how these types of sentences are formed and understood. This may give us other insights into how language works in the brain, and how children learn language. Also, touch-screens and eye-tracking are relatively new research methods, and have only been used together very rarely. The results of this project will show how well they can be used together. If you would like a summary of the results of this project once they are ready, please be sure to leave your email or postal address on the consent form.

This project is being carried out as a requirement for a PhD degree by me, Daniel Bürkle, under the supervision of Dr Heidi Quinn, who can be contacted at ☎ 03 364 2008. She will be available to discuss any concerns you may have about participation in the project. I will also be happy to address any concerns and answer any questions you may have about the project.

University of Canterbury Private Bag 4800, Christchurch 8140. www.canterbury.ac.nz

The results of the project will be published as a PhD thesis and possibly in other forms, but you can be sure of the complete confidentiality of all gathered data: the identity of participants will not be made public. To ensure anonymity and confidentiality, data will not be stored together with participant names—only anonymous participant numbers will be used in the actual data. All data will be stored in locked filing cabinets or on password-protected and encrypted computers, so only my supervisors and I will have access to this data. Research assistants like Ailsa will have access to the data where necessary, but are legally bound to strictest confidence. I will destroy all data after 10 years.

The project has been reviewed and approved by the University of Canterbury Human Ethics Committee. If you have any concerns about the ethical conduct of the project, you can contact the Committee at ☎ 03 364 2987, human-ethics@canterbury.ac.nz, or through its Secretary, Okeover House, Private Bag 4800, Christchurch 8140. Any issues you raise will be treated confidentially, and you will be informed of the outcome.

Thank you.

HEC2013/166
The acquisition of sentence alternations

Daniel Bürkle
Telephone: +64 3 364 2987 extension 8131
Email: daniel.buerkle@pg.canterbury.ac.nz



Adult participant consent

The information on the information sheet has been explained to me, and I understand it.

I voluntarily agree to take part in the project.

I understand that I may withdraw at any time (up until individual data is entrenched in general findings and/or submission of the thesis or articles) without disadvantage, including withdrawal of information provided.

I consent to the recording of my picture sorting, computer input, eye gaze, and storytelling. I consent to the secure storage of this data for 10 years. I consent to confidential access to the data by research assistants if that access is needed for this research project.

I consent to any publication of anonymous results from this project. I understand that the published results will be anonymous and that the data will be confidential.

I note that the project has been reviewed and approved by the University of Canterbury Human Ethics Committee.

Name: _____

Age: ____ years

Born and raised in New Zealand _____

Language(s) spoken at home during childhood: English _____

Date: _____ Signature: _____

Yes, I would like to be sent a summary of the results of this research project.

Email or postal address for summary of results: _____

University of Canterbury Private Bag 4800, Christchurch 8140. www.canterbury.ac.nz

HEC2013/166
The acquisition of sentence alternations

Daniel Bürkle
☎ +64 3 364 2987 extension 8131
Email: daniel.buerkle@pg.canterbury.ac.nz



Information for parents/caregivers

Your child is invited to participate as a subject in a research project on how children use certain types of English sentences. Their involvement in this project will be to sort pictures according to two criteria, follow voice instructions while using a touch-screen, and repeat a short story told to them. I expect that this will not take more than an hour. The session will take place at UC (room 201c in the Locke building), and at least one parent or caregiver has to be with the child throughout the session.

The data I will collect is this:

1. how your child sorted the pictures
2. your child's touch-screen computer input, to see what they were doing
3. a recording of your child's eyes during the touch-screen task, to see what they were looking at
4. an audio recording of your child's story-telling, to see how they say certain sentences
5. your child's age, gender, and language(s).

I will of course answer any questions you may have about the project via email, post, or on the phone.

If you are happy for your child to participate, please talk to them about the project. Then, simply go to www.tinyurl.com/UCstudy to set up a session, or call me at (03) 364 2987 ext. 8131. When you come in, you will get a \$10 fuel voucher, and your child will get a small gift of their choice from a "box of treasures".

You and the child have the right to withdraw from the project at any time with no disadvantage, up until all individual data is entrenched in the general findings of the study and/or up until submission of the thesis or any articles stemming from it. If you withdraw, I will delete any information and data you and your child have provided.

I expect the results of this project will help us understand how these types of sentences are formed and understood. This may give us other insights into how language works in the brain, and how children learn language. Also, touch-screens and eye-tracking are relatively new research methods, and have only been used together very rarely. The results of this project will show how well they can be used together. A summary of the results of this project will be available for you, once they are ready.

This project is being carried out as a requirement for a PhD degree by me, Daniel Bürkle, under the supervision of Dr Heidi Quinn, who can be contacted at ☎ 03 364 2008 or

University of Canterbury Private Bag 4800, Christchurch 8140. www.canterbury.ac.nz

heidi.quinn@canterbury.ac.nz. She will be available to discuss any concerns you may have about participation in the project. I am also happy to address any concerns and answer any questions you may have about the project.

The results of the project will be published as a PhD thesis and possibly in other forms, but you can be sure of the complete confidentiality of all gathered data: the identity of participants will not be made public. To ensure anonymity and confidentiality, data will not be stored together with participant names—only anonymous participant numbers will be used in the actual data. All data will be stored in locked filing cabinets or on password-protected and encrypted computers, so only my supervisors and I will have access to this data. Research assistants will have access to the data where necessary, but are legally bound to strictest confidence. I will destroy all data after 10 years.

The project has been reviewed and approved by the University of Canterbury Human Ethics Committee. If you have any concerns about the ethical conduct of the project, you can contact the Committee at ☎ 03 364 2987, human-ethics@canterbury.ac.nz, or through its Secretary, Okeover House, Private Bag 4800, Christchurch 8140. Any issues you raise will be treated confidentially, and you will be informed of the outcome.

Thank you.

University of Canterbury Private Bag 4800, Christchurch 8140. www.canterbury.ac.nz

HEC2013/166
The acquisition of sentence alternations

Daniel Bürkle
☎ +64 3 364 2987 extension 8131
Email: daniel.buerkle@pg.canterbury.ac.nz



Information for participants

My name is Daniel and I'm doing a project at the university. I want to find out some things about how children understand and speak English.

If you want to help with the project, I will ask you to sit at a computer and move pictures on a touchscreen according to what the computer voice asks you to do. After that, I will ask you to listen to a story about some pictures and repeat the story. When you're done with that, you will get to choose a small gift from a "box of treasures" as a thank-you.

The whole thing will take less than one hour, and your mum/dad/other caregiver will be in the room with you all the time.

The computer will record how you move the pictures on the screen, and a little box on the table will record where you are looking. When you tell the story, we will also record your voice. All of these recordings will be stored on safe computers or in locked drawers. In 10 years, I will delete all that. Your name will not be on the recordings, there will only be a code number. Some people will maybe need to see the recordings, but they can't tell anything about that to anybody.

When the project is finished, you and your parents/caregivers can get a report on what I found out.

If you have any questions, you can talk to your parents/caregivers or to me. If you change your mind about being in the project, that's fine too. All you need to do is tell me.

Thank you for helping!

HEC2013/166
The acquisition of sentence alternations

Daniel Bürkle
☎ +64 3 364 2987 extension 8131
Email: daniel.buerkle@pg.canterbury.ac.nz



Caregiver consent

The information on the information sheet has been explained to me, and I understand it.

I voluntarily agree for the participant named below to take part in the project.

I understand that I or the participant may withdraw at any time (up until individual data is entrenched in general findings and/or submission of the thesis or articles) without disadvantage, including withdrawal of information provided.

I consent to the recording of the participant's picture sorting, computer input, eye gaze, and story-telling. I consent to the secure storage of this data for 10 years. I consent to confidential access to the data by research assistants if that access is needed for this research project.

I consent to any publication of anonymous results from this project. I understand that the published results will be anonymous and that the data will be confidential.

I note that the project has been reviewed and approved by the University of Canterbury Human Ethics Committee.

Name(s) of parent(s)/caregiver(s): _____

Child's name: _____ and age: ____ years, ____ months

Child was born and raised in New Zealand _____

Language(s) spoken at home: English _____

Date: _____ Signature(s) _____

 Yes, we would like to be sent a summary of the results of this research project.

Email or postal address for summary of results:

University of Canterbury Private Bag 4800, Christchurch 8140. www.canterbury.ac.nz

HEC2013/166
The acquisition of sentence alternations

Daniel Bürkle
☎ +64 3 364 2987 extension 8131
Email: daniel.buerkle@pg.canterbury.ac.nz



Participant assent

I have been told about Daniel's project, and I am happy to help with the project.

I know that any information collected about me will not be told to anyone else and will be stored away in a locked cabinet or a safe computer for 10 years. Daniel will not use my name or my parents' names in the project. I know that only very few people will see my information, and they can't tell anything about me to anybody. I know that my parents can ask for a report on the project.

I understand that I can change my mind about being in this project and no-one will mind.

I know that if I have any questions I can ask my parents or Daniel.

Name: _____

Date: _____ Signature: _____

HEC2013/166
The acquisition of sentence alternations

Daniel Bürkle
☎ +64 3 364 2987 extension 8131
Email: daniel.buerkle@pg.canterbury.ac.nz



Information for adult participants

Thank you for your cooperation in the project. The project is aimed at how children's brains expect certain sentences to work out and how children say these sentences. Your results will be incorporated into a 'baseline', so that I know the end-point of children's development. As you may have guessed from the experiment, I am interested in sentences like these:

- "Charlie gave the dog a bone."
- "Charlie gave a bone to the dog."

These sentences are part of what linguists call the 'dative alternation'. What I'm particularly interested in is the way children choose between the two alternatives, based on the animacy, length, and grammatical number of the two objects. In the example above, "*the dog*" is alive, "*a bone*" isn't, and it seems that adults subconsciously prefer the first sentence because it has the living object before the inanimate one. We don't know if children also show these kinds of preferences, which is why I'm doing this study.

If I had given you this detailed information beforehand, your reactions and language may not have been perfectly natural. Natural data is always the best, so I kept the initial information to a minimum in order to get more reliable results.

This was reviewed and approved by the University of Canterbury Human Ethics Committee, who can be contacted at ☎ 03 364 2987, human-ethics@canterbury.ac.nz, or through its Secretary, Okeover House, Private Bag 4800, Christchurch 8140.

Note that all the information from the first information sheet is true. You have the right to withdraw at any time up until submission of the thesis or any articles stemming from it, and the information and data you provided will be deleted if you withdraw. Only touch-screen usage, eye gaze, and story-telling were recorded. Participant data will be stored securely. Participant information will be stored securely and separately from data. Access to data is restricted. Anyone with access is bound to confidence. The results will only ever be published without participant names. You may of course exercise your right to withdraw now or at a later date; should you withdraw, information and data you provided will be deleted. I will be available to answer your questions or concerns, and I'll send you a summary of results (once they are ready) if you wish.

Thank you again.

University of Canterbury Private Bag 4800, Christchurch 8140. www.canterbury.ac.nz

HEC2013/166
The acquisition of sentence alternations

Daniel Bürkle
☎ +64 3 364 2987 extension 8131
Email: daniel.buerkle@pg.canterbury.ac.nz



Information for caregivers

Thank you for your cooperation in the project. The project is aimed at how children's brains expect certain sentences to work out and how children say these sentences. As you may have guessed from observing the experiment, I am interested in sentences like these:

- "Charlie gave the dog a bone."
- "Charlie gave a bone to the dog."

These sentences are part of what linguists call the 'dative alternation'. What I'm particularly interested in is the way children choose between the two alternatives, based on the animacy, length, and grammatical number of the two objects. In the example above, "*the dog*" is alive, "*a bone*" isn't, and it seems that adults subconsciously prefer the first sentence because it has the living object before the inanimate one. We don't know if children also show these kinds of preferences, which is why I'm doing this study.

If I had given this detailed information beforehand, the child's reactions and language may not have been perfectly natural. Natural data is always the best, so I kept the initial information to a minimum in order to get more reliable results.

This was reviewed and approved by the University of Canterbury Human Ethics Committee, who can be contacted at ☎ 03 364 2987, human-ethics@canterbury.ac.nz, or through its Secretary, Okeover House, Private Bag 4800, Christchurch 8140.

Note that all the information from the first information sheet is true: At no time was the child in any risk. You have the right to withdraw at any time up until submission of the thesis or any articles stemming from it, and the information and data you provided will be deleted if you withdraw. Only touch-screen usage, the child's eyes, and their story-telling were recorded. Participant data will be stored securely. Participant information will be stored securely and separately from data. Access to data is restricted. Anyone with access is bound to confidence. The results will only ever be published without participant names. You and your child may of course exercise your right to withdraw now or at a later date; should you withdraw, information and data you provided will be deleted.

I will be available to answer your questions or concerns, and I can send you a summary of results (once they are ready) if you wish.

Thank you again.

University of Canterbury Private Bag 4800, Christchurch 8140. www.canterbury.ac.nz

HEC2013/166
The acquisition of sentence alternations

Daniel Bürkle
☎ +64 3 364 2987 extension 8131
Email: daniel.buerkle@pg.canterbury.ac.nz



Information for participants

Thank you for helping with the project!

I can now tell you what exactly I'm trying to find out: I want to know what children like you do with sentences like

- “Charlie gave the dog a bone.” and
- “Charlie gave a bone to the dog.”

Other projects have found out that the first sentence in this example would be said more often. This happens probably because it has “*the dog*”, which is alive, before “*a bone*”, which isn't. However, nobody has done a detailed project on what children do with these sentences, and that's why I'm looking at it.

If I had told you that earlier, you might have said the sentences differently than you normally would.

All the things I said before are still true. The computer did only save what you pressed on the screen, the little box recorded where you were looking, and we recorded your story. All of these recordings will be stored on safe computers or in locked drawers. Your name will not be on the recordings, there will only be a code number. Some people will maybe need to see the recordings, but they are forbidden to tell anything about that to anybody.

If you have any questions, you can talk to your mum/dad/caregiver or to me. If you change your mind about being in the project, that's fine too. All you have to do is to tell me.

Thanks again!

G Statistical computation in R

Data analysis and simulations presented in this thesis were run in version 3.1.1 of the R language (R Development Core Team 2011), using the base packages plus

- `boot` (Canty and Ripley 2015), version 1.3-16,
- `cvTools` (Alfons 2012), version 0.3.2,
- `e1071` (Meyer et al. 2014), version 1.6-4,
- `ggplot2` (Wickham 2009), version 1.0.0,
- `ggthemes` (Arnold 2014), version 1.7.0,
- `gridExtra` (Auguie 2012), version 0.9.1,
- `gss` (Gu 2014), version 2.1-3,
- `lme4` (Bates et al. 2014), version 1.1-7,
- `mvtnorm` (Genz et al. 2014), version 1.0-1,
- `plyr` (Wickham 2011), version 1.8.1,
- `PolynomF` (Venables 2010), version 0.93,
- `psych` (Revelle 2015), version 1.5.1,
- `xtable` (Dahl 2014), version 1.7-4,

and their dependencies.

The code used for the examples and simulations presented in Section 5.2 is available online at [dbuerkle.github.io/ssanova.R](https://github.com/dbuerkle/ssanova.R) and [.../simulations.R](https://github.com/dbuerkle/simulations.R) or on request.