Language diversity in the psycholinguistic study of sentence form variation

Dutton, E.M.

Author: Dutton, E.M.
Title: Language diversity in the psycholinguistic study of sentence form variation
Issue Date: 2019-12-12

Language diversity in the psycholinguistic study of sentence form variation

Eleanor Dutton

Stellingen

behorende bij het proefschrift

Language diversity in the psycholinguistic study of sentence form variation

Eleanor Dutton

1. A sample skewed towards West European and particularly Germanic languages has led to an overstatement of the role of *subject*, and an underappreciation of the role of *topic*, in the psycholinguistic account of sentence form. *(this thesis)*
2. Listeners do not combine discrete cues to thematic interpretation in a purely additive fashion, but instead integrate linguistic and non-linguistic evidence in a probabilistic fashion to predict thematic structure during comprehension. *(this thesis)*
3. A wide range of phenomena in sentence production have been attributed to a single underlying dimension of Accessibility (i.e. speed of information retrieval); however, this single latent variable is insufficient to fully account for variation in sentence form, particularly across diverse languages. *(this thesis)*
4. The aforementioned skewed sample should not be attributed to lack of awareness, but rather to the challenges inherent in attempting to integrate unfamiliar linguistic phenomena with existing theory. *(this thesis)*
5. Over time, sampling bias becomes entrenched as methodological and theoretical bias; we must therefore not only expand the typological reach of psycholinguistics, but also critically evaluate our existing approaches and assumptions.
6. Attempts to resolve the linguistic sample bias in psycholinguistics will ultimately fail as long as typological and descriptive linguistic training is not considered a requirement for psycholinguists.
7. Language is the sound of brains sharing content; all linguistic phenomena therefore demand a psychological or neuroscientific explanation.
8. There is so much inter-individual variability in the PhD process, which is so vastly impacted by the differences in (often unwritten) protocol between fields, the support received, and personal circumstances, that the only characteristic that one can assume with certainty about a person who has been awarded a PhD is that they have been awarded a PhD.
9. Not everything that is valuable is measurable.
10. It’s all about the niche.

Language diversity in the psycholinguistic study of sentence form variation

Published by

Ridderprint BV
Ohmweg 17
2952 BD Alblasserdam
The Netherlands

phone: +31 (0)180 463 962
www.ridderprint.nl

Cover illustration: Eleanor M. Dutton (a collection of stimulus materials from this research)

Language diversity in the psycholinguistic study of sentence form variation

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof. mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op donderdag 12 december 2019
klokke 10:00 uur

door

Eleanor M. Dutton

geboren te Groot-Brittanië

Promotores: prof. dr. N.O. Schiller
prof. dr. M.G. Kossmann

Promotiecommissie: dr. F.K. Ameka
prof. dr. L.L. Cheng
prof. dr. A.S. Meyer (Radboud University and MPI, Nijmegen)
prof. dr. M.J. Pickering (University of Edinburgh)

Welcome to my paper bag!

Contents

Acknowledgements

Before I began this PhD project, I discussed the possibility of carrying it out at Leiden with a number of people at LUCL. I would like to thank them for their belief in my capabilities as I embarked on this phase of my academic life. Among those are my supervisors, Niels Schiller and Maarten Kossmann. I would also like to thank Florian Jaeger for his encouragement, later in my PhD, to continue pursuing psycholinguistic research on understudied languages, as well as for comments on an earlier draft of Chapter 5. Towards the end of the PhD, when I was also active as Lab Manager at LUCL, I benefited from the support of Lisa Cheng.

The studies that I present in this thesis would not have been possible without the help and involvement of many people. In particular, none of this would have happened without the kind cooperation of study participants in the Netherlands, Morocco and Indonesia. In addition, my trip to Morocco in 2014 was partially supported by the Leids Universiteits Fonds.

I am truly indebted to the Boudihi family for their great generosity in hosting me on my trips to Morocco. I was welcomed into family life with a hospitality that I had never before experienced and will never forget. Moreover, without the diligent assistance of Hanae in my data collection, this thesis would never have come to fruition. Khalid Mourigh additionally provided help with preparation of Tarifiyt stimuli and transcription of data. I also owe a debt of gratitude to Prof. Mostafa Ben-Abbas, who not only supported and facilitated my research at Faculté Pluridisciplinaire de Nador but also made sure that I felt welcome on the campus. Additionally I would like to thank Dr. Fouad Saa, who kindly received me at Université Sidi Mohamed Ben Abdellah in Fes. My sincere gratitude also goes to everyone who participated in the research in Selouane, in particular Farida and Nafi3a, whose friendly curiosity made the experience so positive.

Thanks to Ernanda, whose unrelenting positive-thinking attitude was the driving force behind the experiment reported in Chapter 5. I would like to extend special thanks to everyone involved in the experiment in Pondok Tinggi, particularly those who assisted with recruitment.

I would like to thank all my colleagues and friends at LUCL for valuable companionship: a problem shared is a problem halved. Having been around a while, there are simply too many names for me to mention here. However, if you ever laughed at one of my jokes (or attempts at humour), your name is on the list.

I was carried through this experience day by day on the support of Olga Kepinska, Viktorija Kostadinova and Bobby Ruijgrok. The PhD journey without you guys is simply unimaginable. You were there at the beginning, at the end, at the high points and the low points, and everything in between.

Aside from colleagues and friends in the Netherlands, I have also been able to count on all the people back home, including those whose presence in my life goes so far back that they make this PhD seem like a brief moment in time. There's a little village in the English countryside where I know that there are people who care what's happening in my life and want me to succeed. Thanks go in particular to those who participated in an especially boring norming survey (and won't let me forget it).

Special thanks also go to Abdo, who is always ready to remind me that it's all just words.

The final word of thanks is of course reserved for my family, in particular my dear mum and dad, who are unfailingly supportive (even though I drink all the water in the tap, etc.)

Abbreviations

1	first person
3	third person
ABS	absolute
ACC	accusative
ACT	active
ADVPASS	adversative passive
AS	annexed state
AUX	auxiliary
DO	direct object
EPTH	epenthetical subject
F	feminine
FS	free state
HITHER	ventive particle
IMPF	imperfective
IO	indirect object
M	masculine
NOM	nominative
OBL	oblique
PASS	passive
PERF	perfective
PRES	present
PRESREL	present relevance
PST	past
PRT	participle
RECP	reciprocal
SG	singular

CHAPTER 1

Introduction: the psycholinguistic study of sentence form variation

1.1 Overview

This thesis is concerned with the processes by which speakers organise their ideas into felicitous, well-formed sentences, with particular focus on how this proceeds given that languages vary in myriad ways. The questions that I aim to address in the current thesis – which are both theoretical and methodological in nature – are questions that are raised as we aim to conduct more cross-linguistic work on typologically distinct languages. I am interested in the question of how speakers and listeners navigate the language-specific tools in order to communicate conceptual content; for this it is clearly vital to grasp how different languages meet the communicative needs of speakers. However, insights in this area have been limited by the fact that psycholinguistics has only covered a specific fraction of the world's languages in any depth. Moreover, when researching new linguistic types, challenges arise at the point of applying existing methods and paradigms. It is for this reason that the research questions I formulate in this thesis relate to both theory and method.

The studies reported in this thesis aim to contribute to the *theoretical* understanding of the psycholinguistic processes underlying sentence organisation, specifically grammatical encoding. Primarily, I investigate this in production (Chapters 3, 5 and 6), but I also conduct an investigation of the same structures and variables from the point of view of the hearer (i.e. in comprehension; Chapter 4). Sentence production research seeks to shed light on the mechanisms by which speakers structure elements and ideas, and the different forms that sentences can take. An essential question here is how speakers decide on the roles and positions for arguments to take in the sentence. From the cross-linguistic point of view, we can also ask: how do speakers successfully navigate these choices in light of the restrictions posed by specific languages? Furthermore, how can cross-linguistic differences and similarities inform us about the underlying psycholinguistic mechanisms?

For this, it is clearly important to consider relevant variables across languages with different typological profiles. However, as we seek to work with a wider variety of linguistic types and speaker communities, we typically encounter new *methodological* challenges. When conducting experimental research with linguistic types and speaker communities that are not yet represented in the empirical base of the discipline, we have to think flexibly not only in terms of theory, but also in terms of how we plan and implement our experimental studies. Therefore, a supplementary goal of this thesis is to bring to discussion the various practical and methodological challenges that arise when conducting psycholinguistic research with typologically diverse languages. With this thesis, I hope to contribute some suggestions of ways in which we could adapt our approaches to design, data processing and analysis, so as to be able to engage more fully with the variation found across the world's languages, while still maximising continuity with the existing body of experimental work on sentence formulation.

1.1.1 Outline of the thesis

This introductory chapter aims to further delineate the theoretical and methodological scope of this thesis. To do so, I provide an outline of relevant aspects of sentence production research. Firstly, I review the theory of grammatical encoding in sentence production, also drawing links to the related field of information structure. Secondly, I describe the methods that are typically used to investigate the phenomena in question. In this chapter I raise some key areas of inquiry that are explored in the studies in the later chapters. These questions chiefly concern the consensus account of accessibility (viz. ease of information retrieval) as a driving force in sentence production, in particular the nature and scope of accessibility as an explanatory variable for sentence form phenomena.

In Chapter 2, I move on to discuss the issue of typologically diverse, cross-linguistic research in psycholinguistics. While the importance of linguistic diversity in psycholinguistic research is well recognised, it seems that working with a diverse range of languages raises new conceptual and practical challenges which need to be overcome. The aim of this chapter is therefore to review and discuss these challenges. They include the practicalities of working outside the lab, the applicability of paradigms across languages and participants, the problem of resource scarcity in understudied languages, and lastly, issues that arise at the point of analysis and interpretation.

The four studies subsequently presented in Chapters 3 – 6 then report empirical investigations to address certain theoretical and methodological questions raised. These studies feature three languages: Dutch (Chapters 3 and 4), Tarifiyt Berber (Chapters 3, 4 and 6) and Pondok Tinggi (Chapter 5).[1]

Chapter 3 reports a picture description experiment in Tarifiyt Berber and Dutch. The chapter forms a new presentation, analysis and discussion of an existing dataset which was first reported in Dutton (2012). A key question concerns the manner in which the patterns of sentence form in the languages both resemble each other and diverge from each other, and how we can capture these commonalities and differences under the standard theory of sentence production as introduced in Chapter 1. Chapter 4 follows up on this study by investigating the comprehension of the forms produced by participants in the study in Chapter 3. Following this, Chapter 5 investigates how conceptual accessibility impacts structural choice when there are more than two felicitous structures to choose from. Chapter 6 then investigates the effect of contextual accessibility on how individual arguments are realised within the sentence, rather than the effect on the choice of overall sentence structure.

[1] A short overview of each language is provided at the end of Chapter 2 (Section 2.3.1).

In order to shed further light on some methodological points raised in Chapter 2, the method in Chapter 3 is presented in detail. The issues of interest concern the choices that we must make in adapting experimental paradigms to new languages and settings, and in preprocessing, coding and analysing the data. This presentation then forms the basis for discussion and further exploration of these issues in Chapters 5 and 6. Chapter 5 considers the problem of analysing sentence production data when there are many different possible structures in the output. Chapter 6 reflects further on the issues of analysing high variability sentence production data, considering in addition the issues of bias and data loss that are paramount when working with unfamiliar and/or understudied languages. In each of these chapters a possible solution is proposed and implemented; limitations of the approach are also discussed, with a view to developing better solutions in future work.

1.2 The study of grammatical encoding

1.2.1 The phenomena under study

Sentence production research investigates the production of clause level units. The questions of interest in this field concern how we organise ideas into sentence forms that are grammatical in the language. This is particularly challenging given that for any given proposition there may be multiple ways of expressing it in linguistic form.

Although the study of sentence production can be understood to encompass any processes that are required for producing a well-formed felicitous utterance, typically the focus of this field is on phenomena which overlap with the traditional field of information structure. In simple terms, we are interested in what provokes speakers to choose one form rather than another, particularly when both forms express the same semantic or propositional meaning. A well-studied example of this is the active/passive alternation in English, illustrated in example (1).

(1) a. The boy chased the dog
b. The dog was chased by the boy

This example is provided by Palmer (1994), introducing the cross-linguistic existence of such pairs of sentences which "differ grammatically in the marking of the arguments, but with very little change of meaning" (1994:4).[2]

The study of information structure could be characterised as an investigation of different forms with comparable propositional content. We can borrow this characterisation from the seminal study of information structure by Lambrecht: information structure concerns "the comparison of semantically equivalent but formally and pragmatically divergent sentence pairs" (Lambrecht, 1996:6). Such descriptive or typological studies of information structure are primarily concerned with how different sentence forms relate to different discourse 'meanings', while psycholinguistics is more concerned with the processes underlying their production. Nonetheless, the psycholinguistic study of sentence production rests on a similar conceptual basis: we study the choice between "linguistically different but semantically equivalent utterances" (Gleitman, January, Nappa, & Trueswell, 2007:548).

At this point we can ask: in what relevant way are forms 'linguistically different' from each other? For psycholinguists, what are the variables of interest, as regards linguistic or formal differences

[2] Although here we are primarily concerned with the realisation of arguments, it must be noted that a key feature of the passive cross-linguistically is concurrent marking on the verb (Siewierska, 1984).

between sentences? Much of sentence production research has focused on two key dimensions: differences in grammatical function assignment and variations in the relative order of elements in a sentence.

Grammatical function assignment refers to how referents in a proposition are associated with grammatical functions in the language, such as subject and object. A key point here is that there is a distinction between the semantic roles that are played by referents – for example, the agent, or the patient – and the grammatical functions that they are assigned – for example, subject or object. A semantic role may have a grammatical function that it is typically associated with (such as agent typically being associated with subject function in English). However, crucially, a semantic role is not necessarily always mapped to the same grammatical function. The fact that semantic roles can be mapped to different grammatical functions is what gives us different forms of diathesis, or voice. In (1), the two sentences express the same interaction: a chasing event where the boy is agent and the dog is patient. However, they differ in terms of the grammatical functions assigned to the boy and the dog. When patient, rather than agent, is mapped to subject function, we term it a passive structure. Typically this is accompanied by special morphological marking on the verb (Siewierska, 1984), as can be seen in the English examples in (1).

While psycholinguistic research on sentence production tends to revolve around the distinctions agent/patient and subject/object, it should be noted that this is only part of the picture. Firstly, a wider range of semantic roles have been argued for in theoretical linguistics (instrument, theme, undergoer, etc.). Secondly, there is diversity across languages in the nature of grammatical functions: it is not the case that all the grammatical systems of the world's languages can be appropriately characterised with the notions of "subject" and "object". A useful typological survey of grammatical roles (viz. semantic roles) and grammatical functions is the book by Palmer (1994).[3] When viewing the situation from a cross-linguistic perspective, there is an important distinction to be drawn between semantic roles and grammatical functions: on the one hand, grammatical functions are linguistic: they are defined by specific marking in linguistic form (for example, morphology or word order); on the other hand, semantic roles are properties of the event that can be defined independently of linguistic form (Palmer, 1994:5).

Note that the sentences in (1) also differ in terms of the relative order of the two elements: in example (1a) the boy precedes the dog and in (1b) the dog precedes the boy. In English, grammatical function assignment and relative ordering of arguments (viz. linear order) are strongly intertwined, or "correlated" (Comrie, 1989:75). However, in other languages, it is possible for grammatical function assignment and linear order to vary independently of each other. Consider the Japanese examples below, which are stimuli from the sentence recall study by Tanaka, Branigan, McLean, and Pickering (2011). In all examples, the roles of the boat (*booto*) and the fisherman (*ryoshi*) remain the same. Example (2a) shows an active transitive with SOV order. In example (2b), the grammatical functions of the boat and the fisherman remain the same (subject and object, respectively) but the relative order changes – the order is now OSV. By contrast, in example (2c) the grammatical functions of the fisherman and the boat are switched – now the fisherman is the subject and the boat is the object – but the relative order is SOV once again. Since the roles of the two entities remain the same while the functions are remapped, we characterise (2c) as a passive. Lastly, in example (2d) we again see a passive construction, but the relative order is the same as in (2b): the boat precedes the fisherman.

[3] Note that Palmer uses the term *grammatical relations* for what I have termed *grammatical functions*.

(2) a. booto-ga ryoshi-o hakonda.
boat-NOM fisherman-ACC carried.ACT [4]
'The boat carried the fisherman.'
b. ryoshi-o booto-ga hakonda.
fisherman-ACC boat-NOM carried.ACT
'The fisherman, the boat carried.'
c. ryoshi-ga booto-niyotte hakobareta.
fisherman-NOM boat-OBL carried.PASS
'The fisherman was carried by the boat.'
d. booto-niyotte ryoshi-ga hakobareta.
boat-OBL fisherman-NOM carried.PASS
'By the boat, the fisherman was carried.'

The examples given so far concern transitive sentences where the key differences are in (i) the mapping of agent and patient to grammatical functions and (ii) their relative linear ordering. Of course, the study of sentence production is broader than just this. For example, we may study variations in how relative clauses are formed, how conjuncts are ordered, and the relative ordering of other elements besides subject and object, among other aspects. In this thesis, I choose to focus on variation in the realisation of transitive sentences, and I leave to one side other types of form variation mentioned; nonetheless, many topics touched upon may be relevant beyond this subsection of the field.

1.2.2 Theory of grammatical encoding

The standard model of sentence production (Bock & Levelt, 1994)

The theoretical basis for sentence production in psycholinguistics owes much to the model of sentence production proposed in Bock and Levelt (1994), reproduced here in Figure 1.1. The model consists of four levels or stages. Firstly, there is the message level; in other words, the prelinguistic message. It is assumed that conceptual information is accessed at this stage, including the roles played by different event participants (i.e. agent, patient and so forth). Then there is the process of *grammatical encoding*, which in this model is subdivided into two stages: *functional processing* and *positional processing*. Functional processing involves lexical selection (including the retrieval of grammatical information associated with sentence elements, such as gender) and the assignment of grammatical functions. Positional processing involves organising these words and morphemes into a linearly ordered structure. As marked in the figure, the term 'grammatical encoding' thus encompasses the phase of retrieving and integrating linguistic material into a grammatical structure. Following this, we have the stage at which articulation is planned and executed, that is, *phonological encoding*.

The processes of interest to this thesis fall within the scope of grammatical encoding, including the transition from message level to grammatical encoding. I therefore leave to one side the processes involved in phonological encoding. We should note at the outset that this model is not uncontroversial: in particular, there has been some debate about whether grammatical encoding should indeed be seen

[4] The glossing in this example reflects that found in the original article, Tanaka et al. (2011). Note that the intention here is not to claim discourse equivalence of these forms, but just to demonstrate the possibilities for formal variation in transitive sentences along the dimensions of both linear order and grammatical function assignment.

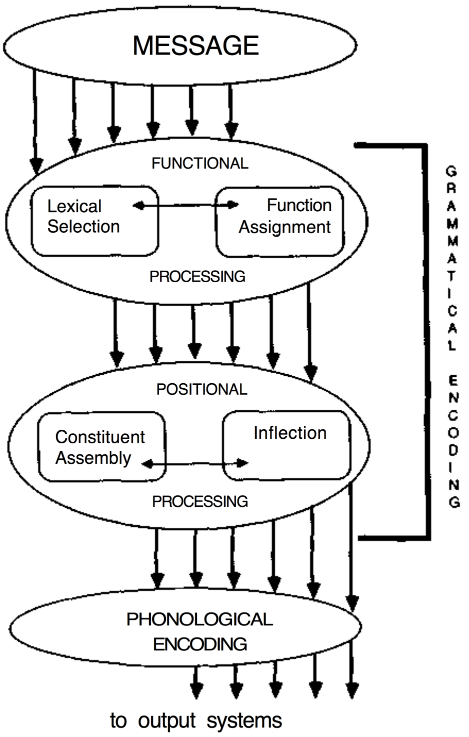


Figure 1.1: Bock and Levelt's "overview of language production processes" (reproduced from Bock & Levelt, 1994:946)

as two consecutive stages as depicted in Figure 1.1, or whether structure is formulated at one stage (cf. Pickering, Branigan, & McLean, 2002).

It is important to keep in mind that any model which purports to describe some aspect of human language processing carries with it a claim of universality, even if that is not made explicit. Accompanying a model such as in Figure 1.1 is the tacit assumption that the same set of processes underpin sentence production in all languages. After all, it is a basic tenet of psycholinguistic or cognitive linguistic research that at some level, human linguistic communication is underpinned by a set of species-wide cognitive mechanisms, regardless of which language is actually being used. However, as we have seen above, languages differ in terms of their specific ordering and function assignment constraints. Therefore, the question arises of how these language-specific features constrain the process described by the model, and at what stage of the production process they are brought into play. Indeed, important work has been done using eye-tracking to probe the time-course of sentence planning also in typologically diverse languages. These studies indicate that there are indeed likely to be differences in the process of formulation, stemming from the typological profile of the language (e.g. Norcliffe, Konopka, Brown, & Levinson, 2015; Sauppe, Norcliffe, Konopka, & Levinson, 2013). The model as depicted in Figure 1.1 is not explicit about how language-specific constraints are to be integrated. Nonetheless, one thing is clear: a cross-linguistically valid theory must be able to account for how general cognitive principles interplay with language-specific constraints.

What underlies differences in sentence form?

In the foregoing section, I identified the processes of grammatical encoding as the focus of this thesis. In the model above, grammatical encoding is seen as a largely autonomous set of processes. If they are indeed autonomous, it follows that when a speaker produces different grammatical forms, this cannot be due to a different set of encoding procedures. This means that the differences that we see in sentence forms must relate to differences either (a) in terms of the message that is passed to the grammatical encoding stage, or (b) in terms of the *way* that the message is passed on. In other words, differences in sentence forms should be relatable to the properties of the prelinguistic message, or alternatively, the way the message is passed to the grammatical encoding process. We can flesh out this idea with a more concrete example. Recall example (1), repeated here as (3).

(3) a. The boy chased the dog
b. The dog was chased by the boy

The difference between these two constructions is not due to differences in the proposition conveyed. The question is then, what *does* this difference derive from? One intuition is that the difference between the forms relates to something about the way we perceive (or wish to represent) the relationship between elements involved in the sentence. This viewpoint is familiar from traditional approaches to the study of information structure (cf. Lambrecht, 1996), reviewed in a psycholinguistic processing context by Arnold, Kaiser, Kahn, and Kim (2013). For example, the difference in forms here may be due to a difference in terms of which referent the speaker wants to say something *about*. So, the speaker presents the state of affairs differently depending on whether she wants to say something *about* the boy or say something *about* the dog – even though the event described is the same.

We could understand the difference in forms, then, as relating to a difference in *topic*, or that the two referents differ along the dimension of topicality. Another such dimension familiar from traditional studies of information structure is *givenness* versus *newness*. Generally speaking, this refers to the contrast between information that is already established in the discourse, versus information that is new. There is a widely observed tendency across languages for given information to precede new information within a sentence structure (V. S. Ferreira & Yoshita, 2003; Halliday, 1967; MacWhinney & Bates, 1978). The consequence for the study of sentence form variation is that speakers are likely to represent states of affairs differently depending on whether one or other referent has already been introduced in the discourse.

Although the notions of topic and givenness (among others) are mainstays of the study of information structure, they are notoriously difficult notions to pin down. Definitions consequently vary widely between researchers (Arnold et al., 2013; M. Wagner, 2016). Practically speaking, it is often difficult to say objectively and reliably what the 'topic' of a sentence is. An exception, of course, are languages which display dedicated topic morphology (such as Japanese, with the morphological topic marker *wa*; Kuno, 1973). However, this of course does not provide a cross-linguistically applicable criterion for defining topicality. The upshot of this situation is that notions such as topic are very difficult to investigate experimentally (although see Cowles & Ferreira, 2011).

Another intuition about the sentences in (3) is that the two referents differ inherently, in terms of their *animacy*. In this case, we are not concerned with the speaker's perspective or the previous discourse, but on inherent properties of the referents involved. Animacy has been found to relate to grammatical form choices cross-linguistically, and its influence on linguistic form is attested well beyond the realm of psycholinguistic experiments (Comrie, 1989; Lockwood & Macaulay, 2012). Animacy can be seen on a hierarchy, with humans at the top, followed by animals, and lastly inanimate objects or forces. The proposal of such a hierarchy for describing linguistic variation goes back at least to Silverstein (1976). Overall, the finding from a range of languages is that the relative positioning of referents on this prominence hierarchy has consequences for linguistic form (Comrie, 1989; Lockwood & Macaulay, 2012). As regards the production of transitive utterances, the general finding based on experimental studies is that more animate entities tend to precede less animate entities, and animates are more likely than inanimates to be assigned the grammatical function of subject (Jaeger & Norcliffe, 2009; Pickering & Ferreira, 2008). Based on these findings, animacy has been argued to have an impact on both the ordering and the grammatical function of referents.

The psycholinguistic account of sentence form variation

Although the fields of information structure and sentence production have conceptual overlap, the psycholinguistic approach diverges in that it aims to be able to conceptualise the variables that affect sentence production in terms of cognitive processing. Perhaps the most prominent cognitive account of information structural phenomena is the **Accessibility Hypothesis**, proposed by Bock (starting with Bock, 1982). The Accessibility Hypothesis sees speed of information retrieval as a fundamental dimension along which referents may differ from one another. The more quickly the necessary information can be retrieved for a referent, the more likely that referent is to be assigned the more prominent grammatical function (e.g. subject in English) or an earlier linear position in the sentence. The accompanying notion of a hierarchy of prominence in grammatical functions, of which subject is the highest, is elaborated by Bock and Warren (1985), drawing upon the noun phrase accessibility hierarchy proposed by Keenan and Comrie (1977). To give a practical example: perhaps for the speaker of (1b), the concept of the dog – the patient referent – was simply accessed more quickly than the concept of the boy, with the result that the patient-prominent passive structure was produced. The more quickly information is retrieved, the faster it can be integrated into the structure being built. This view integrates well with experimental findings indicating that speakers can initiate production before a sentence is fully planned – i.e. that sentence production can proceed incrementally (see for example Allum & Wheeldon, 2007; Brown-Schmidt & Konopka, 2008; Konopka, 2012). In summary, the notions of accessibility and incrementality combine logically under a view that sentence form reflects efficiency in production, whereby speakers produce material as soon as it is available: the *principle of immediate mention* (V. S. Ferreira & Dell, 2000).

Although accessibility is thought to impact both linear ordering and grammatical function assignment, there has been debate regarding the manner in which these effects take place. Early evidence from English (e.g. Bock & Warren, 1985) showed that more accessible referents were more likely to take the initial position in the sentence; however, initial sentence position in English is strongly associated with subject function, meaning that it was difficult to distinguish ordering and function effects. Cross-linguistic data from languages such as Greek (Branigan & Feleki, 1999), Spanish (Prat-Sala & Branigan, 2000) and Japanese (Tanaka et al., 2011) pointed towards independent effects of accessibility on both grammatical function assignment *and* linear order. Alternations such as that in examples (2a) and (2b) are taken as evidence of accessibility effects on "purely positional word order changes" (M. Wagner, 2016:559). In other words, the fact that these two sentences differ in referent order but not in grammatical function assignment shows that accessibility effects on linear order are not always ‘mediated’ by grammatical role assignment. In terms of the sentence production model in Figure 1.1, it seems that we must then propose independent effects of accessibility at both functional and positional stages of encoding. Some authors have proposed that, in terms of constituent structure building, it would therefore be more parsimonious to see functional and positional processing as occurring in a single stage (Pickering & Ferreira, 2008).

The example of functional and positional effects above is a key example of the importance of cross-linguistic evidence in the progress of theory, in particular in informing debates about the impacts of accessibility (or processing speed, more generally) on sentence form. M. Wagner (2016:559) notes a range of typological phenomena that seem to form counter-examples to the predictions made by processing accounts. In the face of counter-examples, it is possible that a better appreciation of the specific linguistic features in the languages under study is what is needed in order to find a way forward. An example of this is the study of sentence production in English and Korean by Hwang and Kaiser (2015). English and Korean might at first seem to show contradictory results in terms of accessibility effects; however, when we factor in that Korean uses nominal case marking and English does not, it seems that a unified processing account is still viable (Hwang & Kaiser, 2015:202).

Are there different types or sources of accessibility?

The dominance of the Accessibility Hypothesis can be attributed to its power to account for a range of variables that appear to have similar effects on sentence form choice. In the case of givenness, it is straightforward to assume that the previously mentioned referents in a discourse are more active in memory, and therefore more easily accessed in the course of production. With the Accessibility Hypothesis, Bock proposed that more animate entities, too, gain linguistic priority by virtue of being more easily accessed on a conceptual level. Bock and Warren (1985) argued that easier conceptual access for more animate entities may be because being more animate means a referent is more ‘predicable’, i.e. it can enter into a greater number of predicative relations. On this view, animacy is seen as a variable that ‘indexes’ the underlying level of accessibility (Branigan & Feleki, 1999; McDonald, Bock, & Kelly, 1993). Following this, it seems equally plausible to account for similar effects of variables such as prototypicality or imageability by assuming that these are all variables that index the underlying accessibility of a referent in the course of producing a sentence (for review see Jaeger & Norcliffe, 2009:869).

We can summarise this view, then, as proposing that accessibility is a latent variable that can account for much, if not all, of the observed variability in sentence form choice. In the strongest sense, accessibility may be viewed as a single latent variable – that is, it is a single dimension of information readiness that is indexed by givenness, animacy, prototypicality and other such variables. In this case, the different variables may be seen as contributing to a referent’s overall accessibility; in other words, that all these sources combine in a quantitative (additive) fashion to influence sentence form. However, various authors view discourse effects (givenness) and inherent properties (animacy) as representing different ‘sources’ of accessibility (Christianson & Ferreira, 2005; Prat-Sala & Branigan, 2000). On the one hand, Prat-Sala and Branigan’s Spanish results indicated that inherent and derived (discourse) accessibility combined in an additive manner, amplifying the overall likelihood of structural effects (Prat-Sala & Branigan, 2000:178). On the other hand, Christianson and Ferreira’s results from Odawa suggested that different sources of accessibility may have similar but qualitatively distinct effects (Christianson & Ferreira, 2005:128).

How much of sentence form variation can be explained by accessibility?

It is important to note that, despite its dominance in the theory of sentence production, the Accessibility Hypothesis was originally conceived of as complementary to traditional information structural notions (Bock, 1982). In his seminal work *Speaking* (1989), Levelt suggests that in some cases topicality could fall out from the order of conceptual access; however, he also describes an alternative or complementary account of topicality as being a property assigned to a referent at the message level, which then affects the way it is processed in the grammatical encoding stage (Levelt, 1989:260-261). Under the second account, then, sentence form phenomena are not reducible to general cognitive dimensions such as attention or information retrieval speed.

It seems the attempt to grapple with the psychological instantiation of nebulous notions such as 'topic' has been shelved, in favour of accounts that reduce such notions to the effects of information retrieval. Indeed, from a processing point of view, the way forward for experimental approaches is much clearer if we believe that "information status emerges naturally out of human memory and attentional systems" (Arnold et al., 2013:409; see also M. Wagner, 2016:558 for a similar observation of this trend). To illustrate this with a concrete example: instead of the speaker in (1b) intending to say something *about* the dog, it may simply be that the speaker's attention was for some reason directed primarily towards the dog, leading to prominence or precedence in the resulting sentence form. By drawing a direct causal link from well-established cognitive processes (e.g. attention, memory) to sentence structure phenomena (e.g. passivisation, linear ordering), we bypass the need for difficult-to-pin-down information structural categories and terms. A demonstrative example of this trend can be found in the study by Tomlin (1995), where traditional notions of 'topic' are eschewed explicitly because of their vagueness, in favour of an account based on attention.

Therefore, it is perhaps testament to the preference for cognitively grounded variables that the notion of accessibility seems to have won more and more explanatory ground over the years: for some authors it has completely usurped the traditional pragmatic notion of topicality (e.g. MacDonald, 2013). Given the widespread adoption of the Accessibility Hypothesis account, it is striking that this hypothesised latent variable is very rarely subject to objective investigation. Often, it is simply invoked as the *de facto* cognitive underpinning for variables that influence sentence form choice. A notable exception to this is Slevc (2011), who investigated whether putative accessibility effects can indeed be objectively related to linguistic information retrieval processes. In that study, Slevc investigated the effect on sentence production when verbal working memory load was manipulated. He indeed found a relationship, suggesting that information retrieval processes do have an effect on the forms of sentences.

Even if we accept a role for general cognitive principles such as memory and attention in determining sentence form variation, there is still an unanswered question: how do these general principles interface with language-specific features? Moreover, given the variation among languages, what would a unified account of this interface look like? In order to answer these questions, we need to first ascertain the *degree* to which sentence form variation can legitimately be attributed to universal cognitive factors (attention, memory) and the degree to which linguistic systems themselves constrain the form of the sentence. Clearly, in order to ascertain the common denominators of grammatical encoding, it is vital to pursue experimental studies in a range of languages that have different typological profiles.

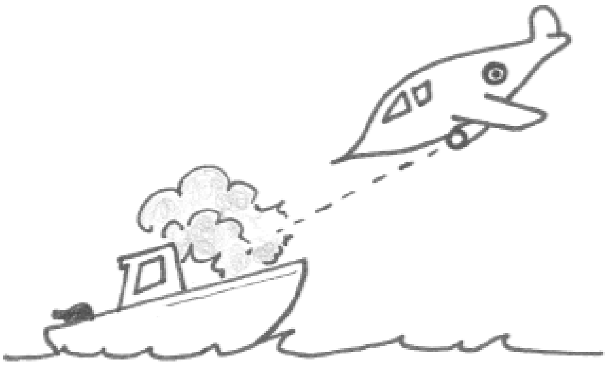
Sentence elements vs. sentence structures in grammatical encoding

Cognitive accounts that are concerned primarily with attention and information retrieval speed tend to focus on relative characteristics of the elements within the sentence, rather than characteristics of the abstract structure as a communicative unit. A strictly information-retrieval based view may see holistic structures (such as passive) as a by-product of the underlying information retrieval processes (Kempen & Hoenkamp, 1987). For example, the passive in example (3b) might be seen as an epiphenomenon arising from the fact that the dog was selected first for output. On the other hand, it is possible to conceive of sentence formulation as beginning with structure building. Under a structure-driven view, speakers consider the holistic structural form and its felicity for encoding the intended message.

Bock and Ferreira (2014) review the evidence in favour of each of these views. They note that the debate about whether sentence formulation is **word-driven** or **structure-driven** dates back at least a century. This dichotomy – whether the parts dictate the whole, or the whole dictates the parts – is also echoed by Arnold et al. (2013), in their distinction between categorical and gradient conceptions of information structure. For descriptive-theoretical accounts, information structure is often understood as the pragmatic ‘meaning’ or ‘function’ of a structure, which can be seen as another unit of linguistic representation on a par with semantics and syntax (e.g. Lambrecht, 1996). Under this view, sentence formulation involves a *categorical* choice among possible abstract structures. By contrast, cognitive viewpoints tend to account for structural variation in terms of the properties of the elements that make up the structure. These properties, such as animacy or salience, are *gradient*; the differing levels of these variables at the time of formulation (such as patient being more animate than agent) is thought to result in variation in sentence form.

Both views have their limitations. On the one hand, structure-driven approaches must account for evidence that we do not always fully pre-plan our utterances. On the other hand, word-driven approaches must account for the fact that elements cannot be arranged just anyhow – structures must be well-formed and discourse-appropriate if they are to be properly comprehended. Bock and Ferreira note that despite the evidence for structure-driven sentence formulation, the word-driven approach has gained more traction in the field of sentence production research. They remark that “the typical question is not about whether the word-first approach is right, but about the factors that determine what the first word, the starting point, will be.” (Bock & Ferreira, 2014:27). Ultimately, the two accounts need not be mutually exclusive, but can represent two approaches to formulation, with one or other approach being preferred depending on the conditions under which the sentence is produced, or the ease of formulating the plan (Bock & Ferreira, 2014; Konopka & Brown-Schmidt, 2014).

Regardless of how we ultimately reconcile these two views, it is vital to bear in mind the basic fact that different languages constrain sentence form in different ways, and that part of a native speaker’s knowledge must be the awareness of what is possible (Lambrecht, 1996). A basic demonstration of this fact is that a linear order (and/or grammatical function assignment) that is well-formed in one language may be ungrammatical in another. For example, the SOV ordering that is unmarked for Japanese is not feasible in English; similarly, not all languages have a passive (in the sense of Siewierska, 2013). Furthermore, from any given starting point, a speaker must end up with a well-formed utterance. Again, the possibilities for integrating a particular starting point into an appropriate sentence may also differ between languages. For a concrete example, consider the image below. When describing the scene below, speakers of different languages may in fact select the same starting point – for example, the boat. The description of this scene may then begin with ‘(the) boat...’ in both languages. But while the speaker of one language, such as Dutch, would typically complete this with a passive structure as in (4), a speaker of a language where passive is marginal or absent, such as Tarifiyt Berber, may go on to produce an object-topicalised form as in (5), as confirmed in a *simply describing* experiment with these two languages (Dutton, 2012; presented in Chapter 3 of this thesis).[5]



(4) een boot wordt beschoten door een vliegtuig
a boat AUX.3SG shoot.PRT by a plane
'a boat is shot by a plane'

(5) ijj uɣarrabu tewt-it ijj n ṭṭiyara
one AS.boat 3SG.F.hit.PERF-3SG.M.DO one of plane
Lit. 'a boat, shot it a plane'

Not only are there different possibilities for integrating starting points into well-formed sentences, there may also be different possibilities for starting points themselves, depending on the language's typological profile. For example, languages with canonical verb-initial word order pose problems for any account which rests on the assumption that sentences are formulated by picking one (nominal) argument as the starting point (for an analogous critique of the 'topic-first' principle, see Lambrecht, 1996:200).

Summary

In this section, I reviewed some key elements of psycholinguistic theory as regards the way that speakers organise elements into felicitous grammatical sentences, focusing particularly on the linguistic dimensions of grammatical function assignment and linear order. I then considered what may underlie the differences between the variety of sentence forms within and across languages, with particular reference to topicality, givenness and animacy. I noted that psycholinguistic study of sentence production and the descriptive realm of information structure have a degree of overlap; however, psycholinguistic accounts of sentence form variation have been strongly influenced by information retrieval approaches, with the Accessibility Hypothesis emerging as the dominant explanatory account. In particular, notion of topic, which plays a fundamental role in descriptive theories of inform-

[5] Throughout this thesis I adopt the transcription style for Tarifiyt Berber that is used in Oulad Saddik (2013).

ation structure, is minimal or absent within the current accessibility-driven account of grammatical encoding.

Accounts of sentence form variation that view linguistic form as reflecting the process of information retrieval lead us to a view of sentence formulation as being driven by the relative properties or status of the elements to be produced. With regard to the Accessibility Hypothesis, this refers to the properties of referents in the sentence which make them more easy or costly to retrieve from memory. These properties could be related to discourse variables (e.g. givenness) or inherent properties (e.g. animacy). The idea here is that accessibility leads to priority in grammatical function and linear ordering. Overall, the Accessibility Hypothesis account is powerful in that effects of different variables can be unified under a single latent dimension (ease of information retrieval). However, this elegant account is threatened by evidence which points towards different qualitative effects of these variables. I also noted that despite this strong focus on the properties of referents, we can also view sentence formulation as structure-driven, meaning that the production process also considers the overall structural form, with the planning process proceeding hierarchically rather than linearly (cf. Konopka & Brown-Schmidt, 2014).

Where there are different theoretical views possible, we usually find debates. Here, I noted debates around whether accessibility affects sentence form through an effect on linear order or grammatical function assignment (or both), around whether or not functional and positional processing form two distinct stages of encoding, and around whether sentence structures are holistically chosen or better viewed as a by-product of information retrieval processes. Data from typologically diverse languages both aids and challenges us as we approach these debates. On the one hand, evidence from languages such as Japanese and Greek has helped us to tease apart the effects of accessibility on grammatical function assignment and linear order. On the other hand, the standard (or consensus) model of sentence production is not yet clear about exactly how language-specific differences should factor into the process – for example, which sentence elements are viable as starting points (only nominal arguments, or also the verb), and how a sentence form that is felicitous in one language is ruled out in another.

Clearly more experimental work on typologically diverse languages is needed to expand the psycholinguistic understanding of sentence production. However, working with a more typologically varied range of languages poses challenges in itself. On the one hand, language types that are unfamiliar to the psycholinguistic field may present phenomena that are difficult to reconcile with current theory. On the other hand, there are various practical issues involved in collecting data from a wider range of speaker communities. Before embarking on a more in-depth discussion of these conceptual and practical issues (in Chapter 2), it is first necessary to touch upon the methodological basis of sentence production research. Therefore, in the next section I consider the typical methods used in the study of sentence production.

1.2.3 Experimental methods for studying grammatical encoding

We can identify several classes of experimental methods for studying how speakers build and produce felicitous sentence units. According the methodological review of Bock (1996), the most direct method to tap into sentence production processes is the *simply describing* technique (as coined by Osgood, 1971). This involves 'simply describing' a visual scene or a video clip. Examples of visual scenes used in *simply describing* experiments are provided in Figure 1.2. This straightforward task has featured in a range of studies over the past few decades including Bock (1986b), Sridhar (1988), Tomlin (1995), Prat-Sala and Branigan (2000), Christianson and Ferreira (2005), Myachykov, Garrod, and Scheepers (2012) and Konopka and Meyer (2014). Related to this are experiments using visual arrays, where participants are asked to describe the relative arrangement of a few objects or shapes. Much work using this paradigm is in the area of sentence planning, such as Allum and Wheeldon (2007); Smith and Wheeldon (1999); V. Wagner, Jescheniak, and Schriefers (2010). This method differs from the more traditional *simply describing* approach in that participants describe movement or location of visual elements relative to each other (such as a star and a circle), rather than a transitive interaction between an agent and a patient.

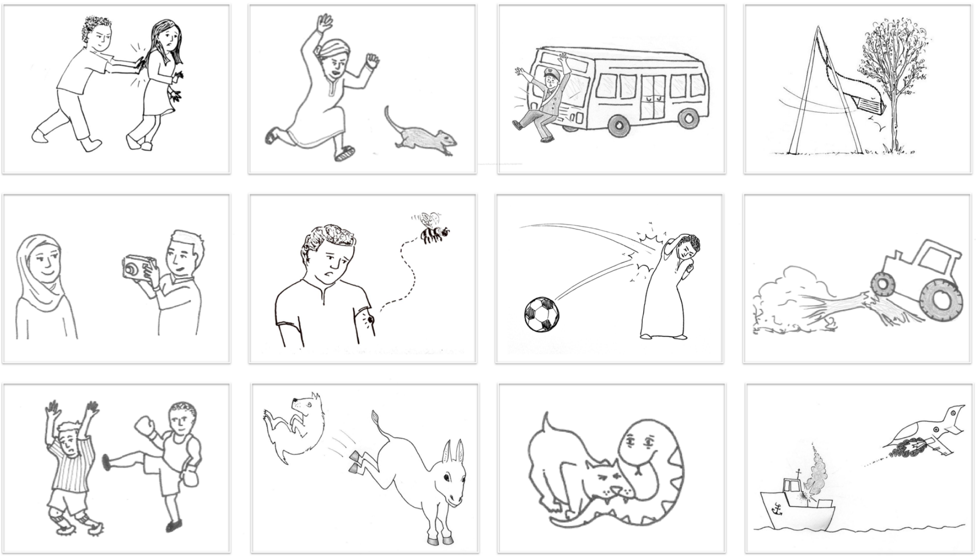


Figure 1.2: A selection of images of the type used in *simply describing* experiments. All images here were used as stimuli in the studies reported in this thesis.

Two more techniques for tapping into speakers' preferences in production are *sentence recall* and *constrained production* (the term adopted by Stallings, MacDonald, & O'Seaghdha, 1998). In *sentence recall*, rather than spontaneously constructing sentence forms, participants are asked to memorise a number of sentences and recall them when prompted. The success of this technique relies on the fact that speakers can normally remember the gist of the sentence (or the propositional content) but struggle to recall the exact wording. The result is that, with an appropriate interval between memorisation and recall, speakers are not able to remember the exact form of the sentences, and frequently reproduce them in a slightly different configuration. This configuration is then taken to reflect a more preferred form. For example, participants may recall a passive stimulus sentence in an active form (e.g. Bock & Irwin, 1980; Kelly, Bock, & Keil, 1986; Konopka & Bock, 2009; McDonald et al., 1993; Onishi, Murphy, & Bock, 2008; Tanaka et al., 2011). Meanwhile, in *constrained production* participants are provided with some linguistic material, which they must make use of to produce fully formed utterances. For example, participants see two nouns and a verb displayed on-screen and must construct a sentence that involves these three lexical elements. Studies using this paradigm include F. Ferreira (1994), V. S. Ferreira (1996), Stallings et al. (1998) and Verhoeven (2014).

Variations on these methods are possible, such as by including priming manipulations and/or by recording eye movements while participants are viewing stimuli. Such approaches have been instrumental in the study of sentence planning; for example, in trying to establish how planning processes vary depending on the structures that are ultimately produced (Brown-Schmidt & Konopka, 2008; Ganushchak, Konopka, & Chen, 2014; Griffin & Bock, 2000; Sauppe et al., 2013). This type of study can be seen as an augmentation of the basic paradigms, inasfar as these additions are supplementary to the core sentence production task. For example, eye movements allow us to ask questions about the planning of specific sentence structures, but this is additional to the question of what provokes a certain structure to be planned in the first place.

A key observation regarding these three different methods (*simply describing*, *sentence recall* and *constrained production*) is that they vary in terms of the amount of constraint or freedom that participants have when producing their utterances. For example, *simply describing* allows participants to respond in any way they see fit, but *constrained production* severely limits their choices. At the most extreme, *sentence recall* pre-specifies the ‘correct’ form of each utterance and relies on the fact that it is difficult to recall sentence forms precisely. The extent of control that we see in a given paradigm has an impact on the naturalness of the task and responses, and therefore the ecological validity of the data.

Overall, it seems that the *simply describing* paradigm should give us the most naturalistic, spontaneous production and therefore arguably the most ecologically valid results. However, this paradigm also tends to yield data that have a high degree of variability. Participants may respond with utterance forms that are unexpected, difficult to categorise or simply “exuberant” (cf. Bock, 1996:407). By contrast, the *constrained production* and *sentence recall* paradigms avoid this problem by narrowing the range of possible responses. It is true that constraining participants in this way, to use certain forms and certain lexical items, may give us a clear dataset from which we can derive a clear set of findings. However, the possibility of relating these findings to normal production processes in everyday speech is thereby compromised. As Bock puts it, “if the desired response is stipulated to speakers in advance, the investigator risks distorting or circumventing the underlying processes” (1996:407). Such a trade-off between control and validity is a perennial problem of experimental research; however, it seems to be particularly problematic for sentence production research. The phenomenon under study is quintessentially variable, spontaneous and creative, meaning that “hard problems arise in crafting viable experimental methods” (Bock, 1996:406) and extracting theoretically meaningful measures.

In terms of data preprocessing, it is customary to transcribe freely spoken responses and then code them according to the dependent variable. As part of this process, erroneous or ‘non-target’ responses are typically discarded from the dataset, according to a set of (predefined) acceptability criteria. Response screening of this type is standard for experimental data processing in general; however, sentence production experiments stand out in that the data exclusion rate can be extremely high. The discarded data can sometimes represent almost half of the raw data; an extreme example of this is an extemporaneous picture description task conducted by Griffin and Bock where over 50% of responses were considered “deviant” (2000:276). Given the potential for fluent, exuberant responding, it is unsurprising that non-target responses should be high in number. Moreover, given that *simply describing* experiments exert the least control over participant responding behaviour, it is no surprise that *simply describing* experiments have some of the highest rates of data loss among sentence production studies.

1.3 Summary and directions for the current thesis

The study of grammatical encoding is concerned with how the pre-linguistic message is encoded into grammatical structures. We can conceive of the linguistic communication as a process by which a speaker transmits a message to a listener, through the medium of language. The role of grammatical encoding is to realise not only pieces of information, but their relationships, in linguistic form. In terms of linguistic form, we are concerned here primarily with the notions of grammatical function assignment and the relative ordering of elements. These dimensions (or variables) not only carry great importance for the psycholinguistic study of sentence organisation, but can also be related more generally to the study of information structure. Much of sentence production research has been concerned with the assignment of grammatical functions to participants in a described event (*functional processing*). A prominent role in the literature has been played by the passive structure, where the patient role is mapped onto the subject function (Bock, 1982). Equal attention must be paid to the relative ordering of elements: ordering is another variable which relates to the mapping between message and form, distinct from the assignment of grammatical functions ('positional processing').

Looking cross-linguistically, it is clear that the exact way that grammatical functions are assigned and the way elements are ordered in the sentence is realised differently across languages. That is to say, the exact way that message elements and their relationships are realised in form – and the felicity of various possible forms – is different between languages. Nonetheless, as discussed in Section 1.2.2, we operate under the assumption that the non-linguistic aspect of sentence production is common across all speakers. This means that our theory of sentence production must ultimately be able to explain how cognitive (non-linguistic) factors, thought to be universal, interface with linguistic features and patterns, which can vary widely between different types and families of languages. In this area of research – perhaps more than other areas of psycholinguistics – it is therefore absolutely vital to engage with typological variation. When the fundamental research questions concern how universal cognitive principles interface with language-specific forms, we cannot afford to be restricted or limited in our understanding of what 'language-specific' can really involve.

However, psycholinguistic research in general has been criticised of late for its bias towards a handful of closely related languages and the tendency to assume that features of these languages are typical of human language, rather than typological quirks of the languages under study (Evans & Levinson, 2009). The field of sentence production is no exception: it has long been limited by an unrepresentative sample of linguistic diversity (Jaeger & Norcliffe, 2009). Although this has begun to change (cf. Norcliffe, Harris, & Jaeger, 2015), we are still far from having a cross-linguistically representative sample in the field. Given the importance of linguistic diversity for the field, we may wonder: how come it is so underrepresented? One likely cause of this situation is that there are additional challenges involved in doing this kind of research. In the next chapter, I turn towards this issue. I first consider why linguistic diversity is important for the field, before reflecting on the fact that the empirical base is still typologically narrow. I relate this latter observation to a number of challenges that arise when doing typologically diverse psycholinguistic research, both conceptual and practical. In doing so, I set the scene for the studies in subsequent chapters that feature typologically divergent languages.

CHAPTER 2

Language diversity in psycholinguistics

In the previous chapter I pointed towards the fundamental need for data from a wide range of languages, with different typological profiles, both in the study of sentence production and in the field of psycholinguistics more generally. However, as mentioned, the empirical base of the field remains typologically narrow. My aim in this chapter is to consider why this situation may have come to be, and, moreover, to reflect on some issues that may be serving to perpetuate the problem.

In the first part of this chapter, I outline the movement towards a more diverse empirical base for psycholinguistics in general, indicating previous publications and events that have played an important role in drawing attention to this issue in the field. Following this, I review a number of challenges that are faced by researchers wishing to conduct linguistically diverse psycholinguistic research, with a view to highlighting issues that still need to be resolved. To conclude, I outline the way in which the studies in Chapters 3–7 of this thesis relate to the discussion in this chapter. This is followed by a section providing a linguistic overview of the languages that feature in the studies of following chapters.

2.1 The need for a diverse linguistic sample

Even though a diverse linguistic sample is clearly beneficial to gaining a richer, more comprehensive theory, linguistic diversity is not well-represented in psycholinguistics. This lack of linguistic diversity in psycholinguistic research has been identified by several authors in recent years, notably Evans and Levinson (2009), Norcliffe, Jaeger and colleagues (Jaeger & Norcliffe, 2009; Norcliffe, Harris, & Jaeger, 2015) and, from an interdisciplinary perspective, Chung (S. Chung, 2008, 2012). Estimates of the number of languages and language families in the world vary widely; however, Evans and Levinson suggest a rough estimate of 7,000 languages falling into 300 or 400 families or "groups" (although note that Jaeger and Norcliffe take 200 as their estimated number of language families;

2009:877). In order to assess the coverage of psycholinguistic research, Anand, Chung, and Wagers (2011) surveyed more than 4,000 psycholinguistics abstracts, taken from top journals and conferences. This survey found that just 57 languages were represented. Moreover, their survey indicated that that at least 85% of the research was accounted for by just ten languages (Anand et al., 2011:2). Turning to sentence production specifically, a review by Jaeger and Norcliffe found in 2009 that less than 30 of the world's languages were represented in the literature; of these 30, there were only seven languages for which the body of research consisted of more than five papers (2009:877). Although these numbers may have increased slightly in the intervening years, Norcliffe, Harris, and Jaeger still asserted in 2015 that psycholinguistic work in general is "based on a very small sample of the world's languages, primarily Germanic and Romance, to a lesser extent Finnish, Hebrew, Chinese, Korean, and Japanese" (Norcliffe, Harris, & Jaeger, 2015:1009).

Beyond simply a small sample, there is an additional issue compounding the problem: the handful of languages that have seen substantial research is dominated by a number of closely related Indo-European languages spoken in geographically close proximity. In the case of sentence production specifically, the seven better-researched languages identified by Jaeger and Norcliffe were English, Dutch, German, French, Spanish, Italian, and Japanese; this small sample itself demonstrates an over-representation of two language families (Germanic and Romance) which are both Indo-European, thus further shrinking the coverage that this sample offers (Jaeger & Norcliffe, 2009).

If we understand psycholinguistics to be part of the field of linguistics, it is self-evident that the full range of linguistic diversity should be taken into consideration. The phenomenon under study is the human capacity for linguistic communication; it seems clear that in order to understand how the system *works*, we need to understand what the system *is*: what characterises it, and what the extent of its variability is. An empirical approach to understanding the essential nature of human language processing demands "careful analysis of data from a suitably large and diverse number of languages" (Bowerman, 2010:598).

Even for researchers whose focus leans more towards mapping psychological processes than charting linguistic variation, the advantages of taking a cross-linguistic approach are clear. The most compelling reason is the need to avoid sampling bias. Generally speaking, whenever we have a restricted and unrepresentative sample, we run a high risk of sampling-based confounds. In psycholinguistics, this takes the form of assuming that certain aspects of linguistic form are part of the architecture of the human language capacity, when in fact they are typological features that happen to be over-represented in the sample (Jaeger & Norcliffe, 2009:878). This sampling bias in empirical research is far from confined to psycholinguistics: it is a recognised issue in the field of psychology more generally, where putatively universal theories of human cognition are based largely on work with a narrow demographic – typically participants who are Western, educated, industrialised, rich and democratic, or "WEIRD" (Henrich, Heine, & Norenzayan, 2010).

Apart from the issue of sampling bias, there are several ways in which diverse cross-linguistic work is vital for the progress of psycholinguistic theory. In recent years these have been explored in a number of publications, particularly in the work of Norcliffe, Jaeger and colleagues (Jaeger & Norcliffe, 2009; Norcliffe, Harris, & Jaeger, 2015). In the first place, working with more languages expands the opportunities to test – or simply validate – existing theories and models (cf. the example of agreement processing described by Norcliffe, Harris, & Jaeger, 2015:1014). Furthermore, the exploration of different language types may be the only way to distinguish between different theoretical accounts. A number of cases are described in the work of Norcliffe and colleagues (see also Costa, Alario, & Sebastián-Gallés, 2007); of particular relevance here is of course the cross-linguistic work on sentence production that revealed independent effects on grammatical function assignment and word ordering, something which could not be resolved through work that focused only on English (cf. Norcliffe, Harris, & Jaeger, 2015:871). Ultimately, though, linguistic diversity should not be considered as purely supplemental to existing work on well-studied languages. When we encounter new linguistic features and categories, these lead us to pose questions and hypotheses that would simply not have arisen otherwise: in other words, “we are bound to discover new phenomena themselves in need of explanation, data points that existing theories do not make predictions about” (Norcliffe, Harris, & Jaeger, 2015:1014).

Aside from individual journal articles, evidence of growing momentum towards research on a more diverse range of languages comes from other activities within the research community. Recent years have seen workshops and conferences dedicated specifically to the issue of linguistic diversity in psycholinguistics, such as *EXAL+ (Experimental Approaches to Arabic and Other Understudied Languages)* held in 2016 at NYU Abu Dhabi, and *Linguistic Diversity Meets The Brain: Future Directions in the Language Sciences* held in 2017 at the University of Zürich. There is also sufficient interest in the topic to warrant special issues, such as Language Cognition & Neuroscience volume 30 issue 9: *Laboratory in the Field: Advances in cross-linguistic psycholinguistics* (edited by Norcliffe, Harris, and Jaeger), published in 2015, and a forthcoming special issue of the Journal of Cultural Cognitive Science on structural priming in less-studied languages and dialects (edited by Pickering and Branigan). It is notable that the discussions of linguistic diversity in language (processing) research, and the sampling bias of psychological science more generally, have even appeared in popular news outlets such as the New Yorker and Slate Magazine, respectively (Brookshire, 2013; Burdick, 2018).

At first glance, it might seem that the importance of linguistic diversity in the field has only been recognised recently, and that the movement towards a more diverse sample is a recent phenomenon. However, a historical perspective on the field reveals that this is not accurate. More than three decades ago, Cutler’s review of developments in psycholinguistics heralded the reemergence of a diversity-oriented approach (Cutler, 1985). Cutler’s review charts three ‘ages’ in the history of psycholinguistics, each characterised by the relative dominance of linguistics-based versus psychology-based approaches, i.e. the two parent disciplines. Following a period in which psycholinguists “virtually ignored” the benefits of cross-linguistic work, Cutler points towards a “revived interest in cross-language psycholinguistics”, arising thanks to the reconnect of psycholinguistics with its linguistics roots (Cutler, 1985:659). In the 2015 special issue of Language, Cognition & Neuroscience, the opening article by Norcliffe, Harris, and Jaeger provides an up-to-date review of the situation, which is supplemented by a historical perspective. In line with Cutler, Norcliffe, Harris, and Jaeger (2015) identify the 1980s as a point from which cross-linguistic approaches began to re-emerge, following a period of disconnect between the psychological and linguistic threads of the field.

In summary, what is striking from this historical perspective is that over 30 years after Cutler heralded a renewed momentum towards the incorporation of diverse languages, “the cross-linguistic scope of language processing research ... still falls far short of what is required” (Norcliffe, Harris, & Jaeger, 2015:1009). With specific reference to the area of sentence level processing, Evans and Levinson noted in 2009 that psycholinguists had “hardly begun” to exploit “the full variety of syntax types” (Evans & Levinson, 2009:19). In other words, despite its importance and potential benefits being long acknowledged and demonstrated, typologically diverse research is still a niche, rather than a norm, in the field of psycholinguistics.

The question arises: why should this be? Having established the importance of cross-language research for psycholinguistic theory, why would it be that the field has still not addressed the issue – what is holding us back? The answer to this question seems to have its roots in the additional challenges involved in doing this work. To some extent these challenges are self-evident (e.g. the need for portable equipment), while others are tacitly understood (e.g. the personal effort required to do research in unfamiliar contexts). There are however many challenges that are not fully recognised by field at large (e.g. the limitations of using written stimuli, or the difficulties encountered in data handling and analysis). In order to shed more light on this situation, in the next section I review four broad areas of difficulty encountered by researchers doing non-lab-based work on less well-studied languages.

2.2 Challenges in doing typologically diverse psycholinguistics

In the foregoing discussion, I argued that the benefits and importance of diverse cross-linguistic research in psycholinguistics have been acknowledged for some time, but that it is far from being the norm in the field. Therefore, it seems that more is needed to achieve language diversity in psycholinguistic research beyond drawing attention to its potential benefits. It turns out that, besides simply being aware of the issue, there are a number of challenges that need to be overcome in order to successfully conduct such research. In this section, I explore these challenges in more detail. However, before proceeding with the discussion of what is challenging in conducting typologically diverse psycholinguistics, it is first crucial to consider what is actually implied by this phrase.

2.2.1 What does typologically diverse psycholinguistics involve?

From a practical point of view, the most fundamental implication is that the researcher or experimenter may need to take the experiment to the participant, rather than inviting the participant to a university or laboratory environment. This situation has given rise to the notion of ‘field psycholinguistics’, as a shorthand to describe the act of collecting experimental psycholinguistic data in non-lab settings. This sometimes has the effect of calling up associations with ‘linguistic fieldwork’, which may give rise to the idea that researchers must be prepared to travel to geographically remote locations to track down minority speaker communities (Tsegaye, 2017:214). This idea may be compounded by authors who leverage the emotive topic of language endangerment to propel interest (and urgency) in conducting research with a wider range of languages (Burdick, 2018; Evans & Levinson, 2009).

We should acknowledge the need to be geographically and logistically flexible; however, it is important to note that there is still much psycholinguistic ground to be covered through work on majority languages spoken by easily accessible, urban speaker communities (cf. Speed, Wnuk, & Majid, 2017:192, in reference to Sauppe’s work on Tagalog). In addition, not every language that is understudied in psycholinguistics is understudied more generally in linguistics. An interesting case to consider at this point is Arabic, varieties of which are mother tongue for millions of people across the Middle East and North Africa, and which has a long tradition of scholarship, also in Europe (Loop, Hamilton, & Burnett, 2017). However, psycholinguistic research on this language family is scant (Boudelaa, 2013). This is regrettable, since not only is Arabic typologically distinct from the over-represented (Western) Indo-European languages, but also the linguistic variation among the Arabic dialects could provide unique opportunities for the controlled study of processing and production differences – in other words, a kind of ‘natural laboratory’ (Brustad, 2000; Evans & Levinson, 2009).

For the rest of this discussion, I take as the point of departure that, as we try to consider a wider range of languages, we need to be prepared to go beyond the formal, laboratory setting, and to work with participants who are not familiar with the conventions of research (or the typical aims of researchers) (S. Chung, 2012; Speed et al., 2017). The actual range of contexts of this research may, however, vary – non-lab research could even include web-based experiments, but this form of research does not feature in the current thesis (however, a useful primer on this method is provided by Speed et al., 2017). Challenges in conducting typologically diverse psycholinguistic research are of course not only practical, but also conceptual, as we try to build on this research to advance our theory. In the discussion below, I consider first the logistical and practical issues, before going on to consider less salient problems that concern experimental design and implementation, and finally the theoretical embedding of such work.

2.2.2 Logistical and practical issues

The most salient challenges associated with working on understudied languages relate to the fact that they are typically less accessible in some way to the researcher. The need to work outside the lab setting can give rise to issues regarding portability of equipment, and challenges in finding an appropriate location for experimental sessions away from interference of ambient noise or background events. Consequently, it is sometimes necessary to compromise, allowing for some environmental influences that would normally be absent in a lab setting (practically-oriented reviews of such issues are provided in Speed et al., 2017; Whalen & McDonough, 2015).

At first glance, the logistical challenges of transporting equipment and finding testing locations seem to be the most salient. However, there are additional, underlying, practical issues involved in making the transition out of the lab. A first practical point is that undertaking this kind of research may require additional time and funding, beyond what is typically needed for lab-based work on well-studied languages. This point is raised by Norcliffe, Harris and Jaeger, who also remark that the extra effort required to conduct this work is not always recognised by readers, reviewers, editors, and funding agencies (Norcliffe, Harris, & Jaeger, 2015:1024). These authors also suggest that sentence production research suffers even more from these kinds of costs, due to the labour-intensive nature of processing production data (Norcliffe, Harris, & Jaeger, 2015:1014).

Norcliffe, Jaeger and colleagues also indicate that conducting psycholinguistic research outside the lab, or in understudied languages more generally, may call for a broader training profile compared to lab-based work (Jaeger & Norcliffe, 2009; Norcliffe, Harris, & Jaeger, 2015). The history of disconnect between psychology-oriented psycholinguistics and descriptive, field-based linguistics may pose obstacles here. As noted above, the field of psycholinguistics has both a psychological and a linguistic (or typological) component, but these two poles have sometimes become rather separate (Cutler, 1985); this may lead to research groups being oriented towards one pole, at the expense of the other. For the promotion of typologically diverse psycholinguistics, it is therefore important to ensure comprehensive support for the full range of skills and knowledge entailed (cf. Norcliffe, Harris, & Jaeger, 2015:1024).

Chung and colleagues also note that the necessary skills to successfully conduct research in the field situation may be broader than those required to conduct research in the lab situation (S. Chung, 2012). In particular, flexibility with regard to cultural or contextual norms is an important point for consideration. For example, in an experiment conducted in a university lab, it is quite normal for a participant to show up at a specific prescribed time, promptly sit down and undertake the experimental task, receive a small amount of money as compensation, and leave immediately. Such a brief, functional interaction is efficient for both researcher and participant. However, in the non-lab context (such as visiting a participant at home, or while staying as a guest in a speaker community), the procedure adopted for an experimental session will obviously need to be less ‘clinical’. Such considerations imply a certain amount of work on the part of the researcher to adapt procedures without sacrificing too much experimental control or continuity with previous studies. Another example of a practical issue when planning an experiment for the non-lab setting is the question of how to compensate participants. For example, in the lab context described above, it is customary to offer financial compensation for participation. However, in other settings, the offer of money could be perceived in quite different ways, or may simply offend participants whose involvement is meant as a gesture of interest or friendship.

Overall, conducting experimental research outside the lab may incur extra costs and effort in order to approach certain practical challenges which arise due to the new setting. At the least, conducting field-based research demands a greater level of flexibility, as researchers must adapt their testing procedures for success in new (cultural) settings. An important point to consider is that the greater the practical issues that have to be overcome, the more likely there is to be an overall deterrent effect against embarking on this kind of research. In addressing logistical and practical challenges such as these, it seems crucial to engage in open discussion of the issues at hand. With regard to the issue of training and support, much can be gained from maintaining and stimulating dialogue between the psychological and linguistic poles of the field. In this vein, a cross-disciplinary approach can also be helpful, encouraging knowledge and skills exchange with (native-speaker) linguists from other subfields who are interested in collaborative work.

2.2.3 Applicability of paradigms

One of the first issues faced when embarking on work in a diverse range of languages is the reliance on orthography in psycholinguistic paradigms. Much of psycholinguistic research findings have been obtained using experimental paradigms where participants respond to written stimuli. Written stimuli are widely used in psycholinguistic experiments, even when the research question of the study does not concern orthography or reading processes – the visual form is simply used as a proxy for the oral form. This presumes a reliably high level of literacy among participants. One example of this is the use of written stimuli to investigate sentence processing: particularly demanding in terms of literacy are techniques that use Rapid Serial Visual Presentation, where a sentence is presented word-by-word, with individual lexical items appearing one by one on-screen in quick succession (an interesting reflection is provided by S. Chung, in considering the limitations of “off-the-shelf” experimental designs for work in Chamorro; 2012:10).

The fundamental issue here is that many of the world’s languages are either not written at all, or are predominantly spoken (Nettle & Romaine, 2000:32). It may be the case that, when the language is written, it is written in a form that is unstandardised, or in ways that are not equally familiar for all speakers. An example of this is Tarifiyt Berber, one of the languages investigated in this thesis (see Section 2.3.1). It is a predominantly oral language, with education and official business being conducted in prestige languages such as standard Arabic or French. On the one hand, there does exist a standardised writing system for Tarifiyt Berber: the Tifinagh script, refined and supported in recent years by IRCAM (*Institut Royal de la Culture Amazighe*). However, this script is not typically used by speakers of the language for personal communication. On the other hand, speakers of Tarifiyt have widely adopted an unstandardised latin-based writing system in their communications online via messaging apps and social media. This ersatz orthography is part of a wider use of latin characters for writing colloquial Arabic, known informally as ‘Franco-Arabic’ (also termed “ASCII-ized Arabic” by Palfreyman and al Khalil, 2003).

Even in languages that have established literary traditions, literacy may not be at such a reliably high level in the speaker community. This can of course also apply when we test populations who speak the familiar languages of psycholinguistic research (such as English), but who are not recruited from the highly-educated or university student population. Another common situation is that although speakers are highly literate in a prestige variety, their first language is a colloquial, non-standard variety. Arabic again serves as an example here, of a situation where the literary or standard form of the language is used for formal, public contexts (such as media, education and public signage), but differs significantly from the colloquial forms (e.g. Moroccan *Darija*) that children acquire as a first language. This diglossia has significant implications for processing research (Ibrahim & Aharon-Peretz, 2005).

Ultimately, it is clear that the assumption of many psycholinguistic paradigms, namely that the written language is a reliably good proxy for the spoken language, is no trivial assumption for the world at large. In addition, if we opt to use written language in an experiment, this could mean we are testing participants on linguistic processing in a linguistic variety that is quite different from their native variety, because of a situation of diglossia between the standard written variety and the colloquial oral variety. For true cross-linguistic applicability, we therefore have to move away from a dependence on written stimuli in our experimental paradigms, unless we are specifically researching orthographic processing. In terms of sentence production research, the most straightforward way to avoid the use of written material is to focus on using the *simply describing* paradigm, potentially in conjunction with audio stimuli. In such cases, instructions should also be presented in audio form. With recent advances in technology, it is also very easy to support the instruction phase of experiment with video clips or animations. In cases where paradigms rely on orthographic stimuli, an alternative possibility is to adapt these paradigms to be purely audio and image based (for example, using audio rather than written distractors in a picture-word interference task, as in Tsegaye, 2017). Importantly, to the extent that these are novel adaptations of existing formats, they should be validated also for languages where the written modality has been used until now, such as English.

In *simply describing*, we ask participants to interpret non-linguistic, image stimuli, meaning that issues with orthography can easily be sidestepped. However, issues still arise here. Firstly, it is difficult to control the way that pictures are interpreted. This leads to potential variability and noise in terms of the input to grammatical encoding. Bock terms this the “comprehension contamination problem” (1996:407). The comprehension contamination problem can be amplified when working with understudied languages. We may find that pictures are interpreted quite differently depending on different cultural backgrounds. Furthermore, objects or scenes that are highly familiar for one speaker community may be unusual or unknown for another speaker community (Tsegaye, 2017:218).

Secondly, results can be affected not only by how people interpret pictures, but also how they interpret the experimental task, or at a more general level how they perceive the role of participating in an experiment. It is possible that participants across different speaker communities have wildly varying interpretations of what is intended (and expected) while performing the task. According to S. Chung, a major neglected issue in this area is the fact that “experimental methodology is highly culturally circumscribed ... [and] presupposes that participants are familiar with tests, accept the norms of test-taking, and are willing to maintain exclusive focus on tasks that are often solitary and unnatural" (2012:2). This issue is echoed by Speed et al., who note that when conducting research outside the lab, we face the practical issue of working with "participants who are not used to being tested ... [or] who are not socialized into being compliant responders" (2017:192). Overall, it seems that experimental paradigms with low ecological validity are likely to cause more problems in this regard. An example of this again comes from the work of Tsegaye (2017), who conducted experiments in Konso, a Cushitic language spoken in Ethiopia. In the picture-word-interference paradigm, participants are asked to name an image while ignoring a distractor. While this is a commonly used task in the university lab setting, Tsegaye found that the Konso participants struggled to ignore the distractor words, with some participants even offering commentary on the relationships between the picture and distractor while performing the task (Tsegaye, 2017:217).

With regard to sentence production in particular, the aforementioned 'exuberance' of responses has typically led researchers to give up some ecological validity in favour of increased experimental control. This is done by guiding participants in terms of the kinds of responses they should produce. This may even be explicit, such as when participants are instructed to name both actors in the scene, or are instructed to avoid using pronominal forms (e.g. Prat-Sala & Branigan, 2000). Evidently, such explicit directions rely on participants having some degree of metalinguistic knowledge. Such approaches are unlikely to be very effective for the majority of participants outside the university lab setting, whose metalinguistic intutitions may be weak or non-existent. A more implicit, ecologically valid approach to this problem is to 'familiarise' participants with the kinds of responses that are appropriate (e.g. Norcliffe, Konopka, et al., 2015).

2.2.4 Scarcity of resources

One fundamental issue hindering research on understudied languages is exactly that: the fact that they are understudied. What this means in practice is that there is a lack of resources for these languages. For example, for English we have access to large, annotated corpora such as CELEX (Baayen, Piepenbrock, & Gulikers, 1995). These corpora contain information such as the frequency of individual lexical items across a sample of millions of words. Frequency of words has been established as a variable affecting response times in a range of word recognition and production experiments (Jescheniak & Levelt, 1994), meaning that it is an important control variable in experiments that employ these kinds of tasks. The lack of such a resource means it may be difficult to rule out confounding effects of this variable.

In general, such resources are heavily reliant on the existence of pre-existing text resources in the language. For example, CELEX is based entirely on written sources (Baayen et al., 1995). More recently, the SUBTLEX corpora developed by Brysbaert and colleagues have capitalised on the rise of same-language television subtitling (or 'captioning') in order to provide more representative corpora for spoken language (including Brysbaert & New, 2009; Cai & Brysbaert, 2010; Keuleers, Brysbaert, & New, 2010; van Heuven, Mandera, Keuleers, & Brysbaert, 2014). Nonetheless, corpus-building is reliant on the existence of text, whether written material or transcription. Furthermore, television subtitles still skew us towards favouring majority languages with a standardised written form and high literacy (without which it is not meaningful to provide subtitles). Additionally, these languages are likely to be those associated with higher-profile speech communities who have sufficient resources to produce television programming and broadcast subtitles – presuming that television programmes are even being produced in the language.

Overall, the scarcity of such resources is closely related to the issue that many languages are predominantly oral, or only written in unstandardised forms. Christianson and Ferreira explicitly highlight this issue in their 2005 study of sentence production in the Algonquian language Odawa. They note that "frequency data in Odawa are virtually non-existent since the language is not normally written" and that the transcribed texts that do exist tend not to be representative of daily language use (Christianson & Ferreira, 2005:117). This is likely to be true of many languages where documentation efforts have focused on the collection of oral histories and legends, which may have specific linguistic conventions different from everyday speech. To overcome this issue the authors propose to view the data elicited in the control condition of the experiment as a baseline (Christianson & Ferreira, 2005:117). Tsegaye suggests that such issues could also be tackled by using corpus data from related languages; an alternative approach is of course to conduct additional linguistic fieldwork to bridge the gap (2017:221). This point again underscores the issues discussed in the previous section: working with such languages brings extra demands in terms of workload (and potentially field-linguistic training).

Another type of resource that plays a role in experimental design are stimuli resources. When designing an experiment we can draw on existing stimuli sets made available by previous researchers. By doing so we not only avoid duplicating the workload, but we also bring our work more in line with previous research. An example of such a stimulus database is the set of line drawings provided by Snodgrass and Vanderwart (1980). The norming of these datasets for a new language is also vital work which can aid other researchers. For example, the Snodgrass and Vanderwart set has formed the foundation for the collection of norms (and improved picture sets) for a number of different languages (Bonin, Peereman, Malardier, Méot, & Chalard, 2003). However, the selection of languages for which this work is undertaken again reflects the familiar skew, thus perpetuating the disparity in resource availability across languages. An additional concern here is the aforementioned point that stimuli designed for an experiment in a certain language (or set of languages) may also be culturally biased (Speed et al., 2017). To some extent this can be resolved with a good understanding of the relevant cultural norms, to make an appropriate sub-selection from such a stimulus set. Even so, the fact remains that for understudied languages, we do not have the convenience of pre-existing norming data, and must collect it ourselves, and/or work with native speaker consultants to derive and evaluate appropriate materials (Tsegaye, 2017:221). This again results in a higher threshold for working with such languages rather than more familiar ones.

Lastly, it is important to appreciate that previous research is also a type of resource, especially as far as experimental research is concerned. When we are able to refer to previous findings for a given language, we have a better framework within which to formulate hypotheses, and analyse and interpret our data. Without such a framework, it may be challenging to pose clear hypotheses and we are more likely to face inconclusive results. In many cases, we will need to consider drawing on published research from other (sub)fields, such as descriptive or theoretical treatments. Overall, when working with less familiar languages, we need to be more resourceful in looking beyond our specific discipline for support (cf. Norcliffe, Harris, & Jaeger, 2015). A pivotal resource in this context is the native speaker linguist, who can provide an invaluable source of linguistic intuition and insight for the purposes of hypothesis, experimental design and data processing. Beyond this, native speaker linguists and assistants also act as a personal link between researchers and the language community, thereby facilitating the entire process of data collection. Chung remarks that the role for native speaker assistants in experimental studies actually represents a positive means for collaboration with the language community (S. Chung, 2012:23).

All in all, the existence of richer resources for certain languages will draw researchers towards selecting those languages. This results in a sort of feedback loop, which ends up perpetuating the bias: without a concerted effort and sustained investment, we will not break out of the cycle. It is interesting to note that the issues surrounding resource-poor languages are not confined to psycholinguistics. It is also an issue for computational linguistics, where again most existing research – and therefore, established theoretical and practical approaches – has grown up around English and easily accessible languages. The result of this is that also in the field of Natural Language Processing, "the vast majority of the world's languages ... including some widely spoken ones, are low resource languages" (Bender, 2016:652).

2.2.5 From data to theory, and theory to data

Once we have designed, implemented and run our experiment, we must then process the raw data, in order to perform analyses and interpret the results. The process of going from raw data to results involves many decisions on the part of authors, regarding exactly how the data should be prepared and analysed. For example, which responses constitute errors, and should be discarded? Are there other types of responses which should be discarded, such as outliers? Do the data need to be grouped or transformed in any way?

In sentence production studies in particular, the amount of work that is required to go from raw data to analysis is very extensive (cf. Norcliffe, Konopka, et al., 2015). However, the more processing steps involved in preparing the data for analysis, the more scope there is for researcher bias to have an influence on the outcomes. As described in Section 1.2.3, the raw recordings must be transcribed, and usually annotated or glossed; in a *simply describing* experiment we then typically categorise the responses into groupings that relate to the dependent variable. However, our understanding of what that dependent variable *is* will largely colour the way that we prepare data for analysis. The bias can also be manifested in decisions about the form of the analysis itself. There may be a tendency to simply re-apply familiar procedures for data analysis, which were originally selected for their appropriateness in analysing familiar languages. However, new languages may exhibit fundamental differences in how experimental variables are realised in linguistic form. Since the linguistic form itself is our dependent variable, new forms of data may call for reevaluation of our statistical approaches. For example, a linguistic feature that seems to be binary in one language may be better analysed as ternary in another; variables that seem to operate independently in one language may seem to interact in another.

This issue of bias applies more generally as we extend our theory to incorporate new language types. There is a risk that new languages are always viewed and evaluated against familiar languages as the 'standard'. On the one hand, existing categories may be overgeneralised to new languages. Based on our linguistic experience, we tend to see certain categories and features as fundamental (or even universal), and are therefore inclined to look for them in every language we encounter (Gil, 2001). This relates clearly to the issue of sampling confounds, discussed in Section 2.1: the features we believe to be universal – and which therefore take centre stage in our theories – may simply be ones that are over-represented in our skewed sample. On the other hand, unfamiliar patterns and categories may be ignored, simply because their relation to the existing theory is unclear. Note that even when we find similar categories and structures across languages, we must be open to the possibility that these categories and structures function differently within their particular linguistic systems. For example, it may be tempting to assume *a priori* that any structure that exhibits the classic features of 'passive' (i.e. in the sense of Siewierska, 2013) is the result of the same psycholinguistic processes across all languages. A prediction of this view would be that if a language has a passive structure, it will therefore be elicited in an experiment like that of Prat-Sala and Branigan (2000), namely, when describing scenes less animate agents act on more animate patients. However, assuming this kind of parity of structures across languages is not trivial. In fact, a comparative experimental study in English and Korean by Hwang and Kaiser (2015) has lately cast doubt on this assumption: in that study, although passives were produced by both English and Korean participants, only in English was the rate of passive related to the accessibility of referents (as manipulated through the visual cueing technique, cf. Gleitman et al., 2007).

Overall, to combat this kind of bias, we need to have a good understanding of the language we are dealing with, and how the features that we are interested in fit into that linguistic system as a whole (Norcliffe, Harris, & Jaeger, 2015). The role of native speaker linguists, and collaborative work with linguists from other sub-fields in general, is again very important in ensuring this. Added to this is the need for flexibility in the way we process and analyse data: we must pay careful attention not only to the steps taken to prepare data for analysis, but also to the appropriateness of the statistical method.

In Section 1.2.3, I raised an additional issue regarding the preprocessing of sentence production data more generally, namely is that the exclusion rate (or number of 'non-target responses' can be particularly high, especially for *simply describing* experiments. This issue is compounded when we contemplate working outside the lab, particularly when the language of study requires a significant extra investment of time and cost in terms of recruitment and testing. On the one hand, a high rate of exclusions may be bearable when working in the lab with easily accessible participants. However, it is less tolerable when we are making special trips to conduct experiments in the field.

In summary, we may find that the need to incorporate new language types raises questions not only for psycholinguistic theory, but also for the way we collect, process and analyse data. Working with a wider range of languages is not simply a case of learning to collect data under different circumstances, but also requires flexibility and resourcefulness on the part of the researcher, in order to be able to elucidate the theoretical relevance of linguistic features that are not yet familiar within the domain psycholinguistic research. This is a vital area of attention: the choices we make in data processing and analysis ultimately shape the theoretical value that we take away from a study, and thereby impact on the overall development of the field.

2.3 Summary and directions for the current thesis

In 1985, Cutler observed that "the fact that cross-linguistic research is routine in linguistics and in other areas closely related to psycholinguistics makes it remarkable that core psycholinguistics succeeded in ignoring it for so long" (1985:665). The fact that now, almost 35 years later, we are still seeing calls for more linguistic diversity in psycholinguistics, may indicate the existence of systemic issues in the field that remain unresolved. In simple terms, simply raising awareness of the advantages of cross-linguistic work is not sufficient; we also need to resolve a number of challenges in order to facilitate research of this type. The foregoing discussion reviewed a number of such challenges, which can be seen as obstacles in the way of increasing linguistic diversity in this area of research. These challenges include logistical or practical difficulties, but they may also come in the form of difficulties in trying to apply existing methodology or integrate with current theory. Ultimately, the last of these issues relate to the bias that can easily arise when making the transition to working with unfamiliar language types.

What becomes clear as we discuss these issues, though, is that the problem lies not only in the challenges themselves, but also in the deterrent effect that challenges have. The extra effort required to investigate an unfamiliar language experimentally does not seem worth it if we are likely to be left with results that we cannot relate meaningfully to theory.

Resolving this situation entirely will clearly take time and effort far beyond just the studies presented in a single thesis. In particular, the discussion above has noted the importance of cross-disciplinary training and dialogue (as well as appropriate investment and support) in facilitating researchers as they adapt their familiar modes of research to unfamiliar languages and contexts. Another important way in which to facilitate this process of adaptation is through open discussion and critical evaluation of methodological choices when working with typologically divergent languages. It is in this vein that the studies presented in the following chapters aim to contribute to this theme, by highlighting and examining the choices that must be made in preprocessing, coding and analysing data.

2.3.1 The languages featured in this thesis

The languages featured in Chapters 4–6 are Dutch, Tarifiyt Berber and Pondok Tinggi. Dutch and Tarifiyt Berber provide a wide range of contrasts across relevant dimensions. From a practical point of view: Dutch is a language spoken in the vicinity of university labs, the language is widely studied and has a strong literary tradition, with high rates of literacy among participants; meanwhile, Tarifiyt Berber is a far less widely studied language, spoken predominantly in provincial cities and rural areas of Northern Morocco, with speakers who either do not write the language, or write it using an ersatz script. From a theoretical point of view: Dutch is an Indo-European (Germanic) language, with a typological profile that will be familiar to speakers of other Germanic languages such as English; meanwhile, Tarifiyt is an Afro-Asiatic language which displays typological features that diverge sharply from the profile of Dutch. Pondok Tinggi adds another dimension to the work on this pair of languages. From a practical view, it is a little-studied, endangered linguistic variety spoken by a small community in Indonesia, used almost exclusively in the oral modality. In this sense, the study here on Pondok Tinggi can be seen as a step further into the realm of field psycholinguistics than the studies on Tarifiyt, exemplifying also the value of collaborative work with native-speaker linguists, even from different disciplines of linguistics. The primary importance of this language for the thesis, however, is the theoretical and methodological questions posed by the variety of structures encountered in a straightforward sentence production experiment.

As noted above, it is important for psycholinguists to have overall familiarity with the languages they are studying (Norcliffe, Harris, & Jaeger, 2015). In light of this, this section will provide a brief overview of each of the three languages that feature in this thesis (Dutch, Tarifiyt Berber and Pondok Tinggi) including observations about the participant cohorts.

Linguistic features in the study of grammatical encoding

As discussed in Chapter 1, the dependent measures of interest in previous sentence production research (with relation to grammatical encoding in particular) have been grammatical function assignment and linear order. As described above, we are primarily interested here in transitive forms with an agent and patient argument. As previously mentioned, agent and patient represent language-independent role concepts, i.e. the semantic roles that different entities can play in events. These roles are mapped to grammatical functions in a given language. Grammatical functions serve to express the roles and relationships of arguments in a predication, in linguistic form. Exactly how this is implemented may change across languages. For example, although the term ‘subject’ is used cross-linguistically, it should not be assumed *a priori* that subjects in all languages exhibit the same distributional and grammatical properties (Comrie, 1989).

The same is true of sentence structures cross-linguistically. While the notion of “passive” may be useful for generalising between structures across a variety of different languages, it cannot be assumed *a priori* that a passive construction in one language exhibits the same set of properties as a passive construction in another language. In particular, analogous constructions in two different languages may have different pragmatic properties – in other words, they differ in their distributional properties as regards discourse situations. For example, English has the possibility of object-topicalisation as in sentences such as “John I saw yesterday”; however, this structure would be ruled out when describing isolated scenes such as in Figure 1.2. Meanwhile, object-topicalisation is frequently used to describe such scenes in other languages such as Tarifiyt, Spanish and Japanese (as discussed in Chapter 1).

As also discussed in Chapter 1, the passive construction has received much attention in the sentence production literature. However, this focus is undoubtedly due to the narrow typological sample of the field. There are a sizeable number of languages which have no passive construction (Siewierska, 2013). In addition, it should be noted that there are other combinations of semantic roles and grammatical functions possible (cf. Palmer, 1994). Although these are not treated in the current thesis, they potentially offer new ground for sentence production research to cover.

Turning to word order, this refers to the arrangement of elements in the sentence, independently of differences in how roles are mapped to functions. In studies of transitive picture descriptions where we are interested in how agent and patient are realised, the variable ‘word order’ often refers to the information structural notion of ‘topicalisation’ – typically of the object argument. However, it is important to recognise that word order may also vary in ways that do not relate clearly to discourse functions, such as the ordering of nouns in a conjunct (M. Wagner, 2016).

Apart from these two main dimensions of interest, there are certainly other aspects of sentence production affected by similar kinds of variables. One such aspect is pronominalisation. Pronominalisation has not played such a major role in *simply describing* studies, which often involve isolated visual scenes. In such experiments, where all referents in a visual scene are new in the discourse context, attenuated forms of reference such as pronouns, clitics or argument dropping are indeed rare and usually constitute a non-target response. However, in experiments where discourse availability is an experimental variable, pronominalisation cannot be reasonably ignored. Pronominalisation may also restrict or license certain structural forms, meaning that structural choice is differentiated on this point (examples can be found in Pondok Tinggi, section (11)).

In summary, languages vary in their possibilities and tendencies for realising agent and patient relative to each other. I provide linguistic examples that have been provided by native speakers (or produced in the experiments in this thesis) or taken from published works on these languages. In what follows, I describe how these dimensions play out in the specific typological profiles of each of the three languages under study. The issue of pronominalisation is picked up again in Chapter 6, which presents a study on Tarifiyt. Therefore, in the presentation of languages below, extra attention is paid to attenuated reference forms in Tarifiyt, alongside the overview of agent-prominent and patient-prominent forms in the language. Note that these descriptions are intended to provide an overview of the possibilities for the realisation of grammatical function assignment and linear order in the languages; it is therefore not the intention of this section to review how various discourse functions might be attributed to the structures mentioned.

A note on multilingualism among participants

A monolingual participant is often considered the ideal (or the standard) for psycholinguistic studies, due to potential influences of bilingualism. However, this is not really a tenable approach going forward: huge swathes of the world's population speak two or more languages on a daily basis (Speed et al., 2017:192), and even supposedly monolingual participants are often exposed to other languages (Tsegaye, 2017:218). Therefore, as we aim to work with a more diverse range of languages – and thus a more diverse range of speaker communities – we may need to start seeing multilingual participants as the norm, and monolinguals as the exception, rather than vice versa.

In all studies in this thesis, the linguistic backgrounds of participants were assessed using a pre-experiment questionnaire. The participants in the Dutch experiments reported in Chapters 3 and 4 had grown up in the Netherlands and Dutch was their first-acquired language as a child. As is common for Dutch speakers and especially typical of university students, all participants in this study also spoke English with a high level of fluency. The vast majority of participants also spoke other languages such as French and German, typically due to these languages being taught in the Dutch school system. For the Tarifiyt participants (Chapters 3, 4 and 6), there was a similar or greater degree of multilingualism. Firstly, being university students, all participants in these experiments were highly fluent in French and (standard or literary) Arabic, the languages of higher education in Morocco. In addition, all were fluent in colloquial Moroccan Arabic (*Darija*). Berber varieties can be considered 'minority' languages – the predominant language of colloquial discourse across Morocco is Moroccan Arabic. However, it should be noted that the current research conducted in Morocco was undertaken in an area where Tarifiyt Berber is the dominant spoken language. The situation for Pondok Tinggi participants (Chapter 6) can be seen as a more extended version of the situation as described for the Tarifiyt participants. The participants in this study all acquired Pondok Tinggi Malay in early childhood and reported that it was their dominant language of daily life, in some cases alongside other regional varieties such as Minangkabau or Jambi Malay.

The form of Dutch used in this experiment is a standardised form familiar to all speakers throughout the Netherlands. However, I note that when working with understudied languages, we are often dealing with language varieties that are unstandardised and intelligible only to neighbouring dialects. This is the case for both Tarifiyt and Pondok Tinggi. Such a situation also tends to relate to proportionally higher levels of multilingualism among speakers, because speakers of such languages require one or more additional languages in order to operate successfully in broader contexts, particularly official environments. For example, business or study are usually conducted in a standardised, national language – in the case of Pondok Tinggi, Indonesian; in the Tarifiyt case, French or Arabic. Additionally, it is notable that multilingualism increases for small local language communities when families intermarry between different regions, whereby there may be a patchwork of language varieties used even within an individual household.

Dutch

Dutch features in Chapters 3 and 4. Chapter 3 features a sentence production study carried out in Dutch (as well as Tarifiyt Berber). Chapter 4 reports a mousetracking study that was carried out in Dutch (as well as Tarifiyt Berber). Both experiments were carried out in non-lab settings in the Netherlands. In both these studies, the Dutch participant cohorts were largely made up of students at Leiden University.

Dutch is spoken by around 16 million people in the Netherlands and Belgium (Simons & Fennig, 2017). Varieties of the language are also spoken by populations in former Dutch colonies. Dutch is a West-Germanic language, related closely to German and English, with which it therefore shares some typological properties. A variety of dialects of the Dutch language are spoken throughout the geographical area of the Netherlands and Belgium. The form of Dutch studied in this thesis is the standardised form of Dutch, as spoken by young adults attending university in the urban area around Amsterdam, Leiden and the Hague (the ‘Randstad’).

Grammatical function assignment and linear order

In transitive sentences, Dutch displays subject and object grammatical functions. Dutch does not display case-marking on nouns (although there is a nominative-accusative distinction in pronouns, analogous to English ‘he/him’). Finite verbs agree with the subject argument in number. There is also person agreement, but not every person is distinguished (i.e. there is syncretism). Verbs do not agree in gender. The grammatical functions of arguments are therefore often known from sentence position and context.

Dutch declarative sentence forms can be characterised by the observation that the finite verb occurs in the second position of the main clause. This position can be considered one of the “poles” of the sentence frame, around which other elements are organised (Haeseryn, Romijn, Geerts, de Rooij, & van den Toorn, 1997).

(1) a. Mijn vader **schreef** een boek
my father write.3SG.PST a book
b. Mijn vader **heeft** een boek geschreven
my father have.3SG.PRES a book write.PRT
‘My father wrote a book’

As can be seen in example (1b), when there is an auxiliary verb, it is this finite auxiliary which occurs in the second position; other verbal forms such as the past participle then occur towards the end of the sentence, following the object (this latter verbal position can be considered another ‘pole’ of the sentence frame).

In principle any other element can appear in the first (i.e. preverbal) position (Haeseryn et al., 1997). This means that objects can also appear in this position, however, the object-initial sentence structure is more restricted than the subject-initial. For example, (2) is typically interpreted with Marie as subject and Jan as object; it is only interpreted with Marie as object with a specific pragmatic context and intonational contour (this interpretation is given below in square brackets).

(2) Marie zag Jan
marie see.PST.3SG jan
‘Marie saw Jan’ [˜‘Jan saw Marie’]

Dutch has a passive construction, where the patient is realised as subject (typically at the start of the sentence). The patient-subject is followed by an auxiliary verb (*wordt* in the example below) marking passivisation of the main verb; this is followed by the verb in participle form and optionally the agent in a ‘by-phrase’, introduced by the preposition *door* ‘through’, ‘by’. The by-phrase can occur before or after the participle:

(3) a. een boot wordt beschoten door een vliegtuig
a boat AUX.3SG shoot.PRT by a plane
b. een boot wordt door een vliegtuig beschoten
a boat AUX.3SG by a plane shoot.PRT
'a boat is shot by a plane'

As mentioned above, the structure of the Dutch declarative main clause requires the verb to be in the second position; however, which element is in the first position can vary (Haeseryn et al., 1997). This flexibility means that there is another possibility in Dutch as regards the relative ordering of agent, patient and verb; namely, that both the agent and patient follow the (first) verb. In this case, another element appears in the first position. This could, for example, be a time phrase, such as in example (4):

(4) vanmiddag heeft Jan Marie gezien
this.afternoon has Jan Marie see.PRT
'this afternoon, Jan saw Marie'

Another possibility, however, is that the initial position is occupied by an epenthetical element, namely *er*. This gives rise to the construction termed 'presentative *er*' by Haeseryn et al. (1997:8.6.3). In this construction, both the agent and patient referents are realised following the verb, but the verb or first auxiliary must still appear in second position. The first position of the clause is then filled by the particle *er*. This is illustrated below in the pair in (5): in comparison with (5a), (5b) places the subject (*een agent*) following the first verb (*stond*) (examples taken from Haeseryn et al., 1997:8.6.3.3.i)

(5) a. een agent stond het verkeer te regelen
an agent stand.PST the traffic direct.INF
b. er stond een agent het verkeer te regelen
EPTH stand.PST an agent the traffic direct.INF
'an agent was directing the traffic'

This structure can be realised with both active and passive verbs. In the passive situation, the verb in second position is the passive auxiliary *werden*). An examples of 'presentative *er*' in the passive is provided in example (6).

(6) er wordt een man aangereden door een auto
EPTH AUX.3SG a man run.into.PRT by a car
'a man was hit by a car'

Tarifiyt Berber

Chapter 3 features a picture description study that was carried out in Tarifiyt Berber (as well as Dutch). The participants in this experiment were native speakers of Tarifiyt residing in the Netherlands, of various ages and occupations. Chapter 4 reports a mousetracking study that was carried out in Tarifiyt Berber (and Dutch). The story descriptions in Chapter 6 were recorded with the same group of Tarifiyt Berber participants as took part in the mousetracking study. The participants in these sessions were undergraduate students of the Faculté Pluridisciplinaire de Nador, Morocco. The data collection in the latter studies was kindly facilitated by Prof. Mostafa Ben-Abbas and assisted by Hanae Boudihi, a native speaker of Tarifiyt Berber.

Tarifiyt Berber is a Berber language spoken in the north of Morocco, specifically in the Rif region (from which the name of the language derives). I use the term ‘Berber’ to refer to the family of languages spoken across North Africa before the spread of Arabic, of which Tarifiyt is one variety that is still spoken today. The Berber languages belong to the Afro-Asiatic phylum, which also includes the Semitic languages (of which Arabic and Hebrew are perhaps the most familiar members to psycholinguistic research). Berber languages tend to carry lower prestige than colloquial varieties of Arabic (which in turn hold lower prestige than standard or literary Arabic). Extended co-existence with these higher-prestige languages, as well as the colonial languages French and Spanish, has resulted in a significant amount of lexical borrowing into Tarifiyt (cf. Kossmann, 2009).

Tarifiyt is typically only used in spoken form; therefore it is a prime example of a language where the reliance of psycholinguistics on literacy causes difficulties if we want to apply well-established experimental techniques. As mentioned previously, a standardised orthography has recently been established in the form of the modern Tifinagh script, supported by IRCAM (*Institut Royal de la Culture Amazighe*). However, much of the small body of written Tarifiyt in recent times has been using orthography based on the latin alphabet. In recent decades, Tarifiyt writers have used this type of transcription to produce novels, memoirs, poetry and to preserve traditional songs. Alongside this, the rise of internet forums and online messaging has spawned an increase in written Tarifiyt communication using a non-standard latin-based orthography; this form of writing is comparable to “ASCII-ized Arabic” (Palfreyman & al Khalil, 2003).

Fuller descriptions of the language can be found in Lafkioui (2017) and Mourigh and Kossmann (to appear). Investigations of topicalisation phenomena in Tarifiyt and other Berber languages are provided by Lafkioui (2010, 2011 and 2014). For a study of the relation of information structure and sentence configuration across several varieties of Berber, the reader is recommended to consult Mettouchi and Fleisch (2010).

Grammatical function assignment and linear order

Tarifiyt is typically classified as a verb-initial language (Cadi, 2005) but this should not be taken to entail that the majority of sentences produced in the language are verb-initial (Mettouchi & Fleisch, 2010). The situation can be better expressed in terms of a sentence core, with peripheral positions for expressing certain discourse roles. The core of a verbal sentence[1] consists minimally of a verb, with obligatory subject marking (indicating the number and gender of whichever referent is assigned the subject function). This is illustrated by the minimal verbal sentence below in (7).

(7) yus-d (ijj umucc)
3.sg.m.come.perf-hither (one as.cat)
‘(a cat) approached’

In the sentence core, the verb may be followed by its arguments expressed as noun phrases. However, arguments are typically pronominalised if known from the context (note that forms of reference will

[1] Non-verbal sentences are also common in Tarifiyt for expressing existence or equivalence, typically in combination with the copula *d*.

be discussed further below). When expressed as full noun phrases, the subject argument is the first one to follow the verb, followed by the direct and/or indirect object(s). In Tarifiyt, lexical arguments may carry case-like morphology marking whether they are in the ‘annexed’ (or ‘free’) state. Annexed state is found on postverbally-occurring subject nouns and nouns following (most) prepositions. Arguments not in the annexed state are in the free state, which is also the citation form. Note, however, that many lexical borrowings do not display the state alternation, nor do kinship terms.

Note that Tarifiyt does not have articles; however, the introduction of a new character in the discourse is typically marked by the use of *ijj n* ‘one (of)’ (as in example (7) above)[2]. When referring anaphorically to previously mentioned arguments, the deictic suffix *-nni* ‘the aforementioned one’ is typically used. Note, however, that noun phrases do not *need* to carry either of these markings, that is, they may be realised as bare noun phrases.

In Tarifiyt, active transitive sentences used to describe scenes such as in Figure 1.2 typically have the agent first in the sentence preceding the verb. The verb itself carries a subject prefix that is coreferential with the agent. Following the verb is the patient, in object function. This kind of sentence is illustrated in example (8a). The preverbal position in Tarifiyt has a special information-structural status (Kossmann, 2016), often being considered as a “topic” position (Lafkioui, 2014). Any nonverbal sentence element can felicitously be placed in this position (cf. Mourigh & Kossmann, to appear). This includes the object, as illustrated in (8b). Here, the object is the farmer, *ijj ufedjaḥ*:

(8) a. ijj n weɣyuř yarceř ijj ufedjaḥ
one of AS.donkey 3SG.M.kick.PERF one AS.farmer
‘a donkey kicks a farmer’

b. ijj ufedjaḥ yarcř-**it** ijj n weɣyuř
one AS.farmer 3SG.M.kick.PERF-3SG.M.DO one of AS.donkey
‘a farmer is kicked by a donkey’ (lit. a farmer kicks him a donkey)

In example (8a), the subject is preverbal and the object is postverbal. The verb only carries the obligatory subject prefix. In example (8b) the object is preposed before the core of the sentence. When the object is in preverbal position, it is obligatorily realised as an object clitic pronoun, cliticised to the verb, as can be seen in (8a) (meaning that the example in (8b) cannot be interpreted with the meaning of (8a)). This preverbal-object structure may be associated with a subtle intonation break before the sentence core (cf. Lafkioui, 2014).

In terms of linear ordering of arguments in a sentence, Tarifiyt allows flexibility in the order of subject and object noun phrases, both relative to each other and relative to the verb. By contrast, there is little flexibility in terms of grammatical function assignment. That is, the use of alternative verb forms to allow flexibility of expressing a proposition, such as passive or inverse, are marginal in Tarifiyt. The Tarifiyt passive is extremely infrequent and obligatorily truncates the agent (Cadi, 2005). An illustrative example of the Tarifiyt active-passive alternation is provided below in example (9).

(9) a. ye-qqen ufedjaḥ tafunast
SG.M-attach.PFV AS.farmer FS.cow
‘(a) farmer tied up (a) cow’

[2] The lack of the *n* in example (7) is attributable to phonological processes.

b. te-ttwa-qqen tfunast
3sg.f-pass-attach.pfv as.cow
'(a) cow was tied up' **[by someone]*

Congruent with the notion of passive in the sense of Siewierska (2013), the verb in (9b) is morphologically marked, with the prefix *ttwa-*. The object argument in the active example (9a) corresponds to the subject of in the passive counterpart (9b), as evidenced by the difference in subject agreement on the verb and state marking on the postverbal argument.

Forms of reference

Tarifiyt has two forms of pronouns: unbound (full) pronouns and bound (clitic) pronouns. It is possible to refer to an argument using a full pronoun, such as *nettat* 'she', in a way that resembles the unbound pronominal forms of English. However, the usage of full pronouns is marginal compared to how such full pronouns are used in English. Instead, the bound forms typically suffice: objects are typically referred to using postverbally attached clitic pronouns rather than full pronouns. In addition, for anaphoric reference to subjects, the agreement marking on the verb often suffices, leading some Berberologists to consider it as having pronominal status (Galand, 1964; Mettouchi & Fleisch, 2010). Note also that forms of reference can double: as we have seen in (8b), a preposed object must co-occur with an object clitic pronoun in the sentence core.

When considering how forms of reference interact with linear ordering, it is important to note that the clitics and affixes (usually used to refer to contextually available arguments) have fully predictable bound positions, and only noun phrases and full pronouns are free. This contrasts with the situation in English, where lexical noun phrases and pronouns are both unbound, with pronouns used to "stand in" for a full noun phrase. In example (10), when 'the girl' is referred to instead with a pronoun, this simply swaps in instead of the lexical noun:

(10) a. **The girl** was chased by a dog
b. **She** was chased by a dog

However, the situation looks different in Tarifiyt. When 'the girl' is the object, a clitic pronoun *-itt* appears on the verb, and this remains whether or not the lexical noun phrase is used. Therefore, the difference between the two sentences below is not in terms of replacement of one lexical form by another, but rather in terms of whether or not the girl *tahenjiat* is included as a full noun phrase, in addition to the clitic pronoun:

(11) a. **tahenjiat** iḍfar-**itt** ijj n weqzin
tahenjiat 3.sg.m.chased-3.sg.f.do one of as.dog
'The girl was chased by a dog'
b. iḍfar-**itt** ijj n weqzin
3.sg.m.chased-3.sg.f.do one of as.dog
'She was chased by a dog'

In summary, English pronouns are free and thus can have the same range of possible sentence positions as full noun phrases. In contrast, Tarifiyt relies on clitics and affixes for pronominal reference, meaning that when using attenuated forms, the ordering is fixed. In general, this issue must be borne in mind for any language which employs clitic pronouns.

Pondok Tinggi

The picture description experiment featured in Chapter 5 was carried out in Pondok Tinggi village. This data was collected in the context of a collaborative project with Ernanda, a linguist and native speaker of the language. The participants in this study were native speakers who were local residents in the village. The participants had a wide range of occupations, educational backgrounds and ages.

Pondok Tinggi is spoken in the Kerinci area of Sumatra, Indonesia. Pondok Tinggi is a variety within the highly diverse Kerinci dialect continuum, itself a sub-variety of the Malay language family which includes standard Indonesian. The number of native speakers is not known, however it is spoken primarily in the Pondok Tinggi village which has around 16,500 inhabitants. As is the case for a large number of languages, Pondok Tinggi faces endangerment through the shift to other languages used in the community, catalysed by the fact that surrounding dialects (e.g. standard Malay, Minangkabau) carry higher prestige (Ernanda, 2015). The Pondok Tinggi language is only used in spoken form, making it another example of a language where we cannot as researchers rely on the use of written experimental materials such as orthographic stimuli or instruction texts.

As well as an in-depth grammatical investigation of the Pondok Tinggi language, an introduction to the sociolinguistic context of Pondok Tinggi can be found in Ernanda (2017). Apart from the examples in (14), which are taken from that study, the examples in this section are taken from the dataset reported in Chapter 5.

Grammatical function assignment and linear order

Pondok Tinggi does not employ case marking to express grammatical functions. In this sense it bears similarities to Dutch and Tarifiyt. However, in addition, Pondok Tinggi does not display subject (or object) agreement morphology. Grammatical function is therefore principally indicated by the position of the argument in the sentence and the verbal morphology (which will be discussed in a moment).

An integral morphological feature of the language is the presence of two dimensions of lexical form alternation. Firstly, there is the “phrasal alternation”, which is the alternation of most lexical units between two forms: the absolute form and the oblique form (Steinhauer & Usman, 1978). Its essential import is to indicate whether a lexical unit should be interpreted as being followed by a restrictive element or not, respectively. However, nuanced and contextually conditioned meanings derive from this essential function. There is also an alternation between so-called “K-words and G-words”, where the choice between two forms of a word is conditioned by phonological context. The two alternations combine to form a complex system of lexical alternation. While alternations in Pondok Tinggi therefore form an integral feature of the language, they do not directly impact the current discussion. An extensive analysis of the alternations can be found in Ernanda (2017).

As noted, in Pondok Tinggi, the mapping of roles (agent, patient) to grammatical functions is realised largely by means of linear ordering. The core grammatical functions are subject and (in)direct objects. Subject is typically associated with sentence-initial position, being followed by the verb and objects.

Active constructions in Pondok Tinggi typically display a prefix on the verb involving a nasal consonant (12). Pondok Tinggi displays a passive construction in the sense of Siewierska (2013). In other words, the patient argument is realised in the subject function, and there is some morphological marking on the verb: a prefix, *di-*. The agent is optionally expressed, taking object or adjunct status. When the agent is expressed it may be introduced using the preposition *wot* 'by', but this is not obligatory. Note that the use of the preposition here may in fact have arisen through contact with Malay, itself influenced by Dutch (cf. Ernanda, 2017). Both active and di-passive structures in Pondok Tinggi typically show SVO order, as visible in (12) and (13).[3] However, in certain contexts SOV ordering can be found in the passive (i.e. patient-agent-verb order; see Ernanda, 2017:187).

(12) umpun kayau nimpok uto
stem.obl wood.abs act.crush.obl car
'a tree trunk crushes a car'

(13) uto di-impok [wot] kayau
car pass-crush.obl [by] wood.abs
'a car is crushed by a tree'

There is a second construction that has been termed 'non-canonical passive' (cf. Ernanda, 2017), illustrated in (14a). Here the patient is realised in the initial position; the agent participant must be either speaker or hearer. This construction does not display verb affixation; in fact, the verb appears in the bare form. Object topicalisation is also possible in Pondok Tinggi, realised as active patient-agent-verb (OSV) order (see (14b)). The verb displays nasal prefix and is in oblique form. However, this construction is again restricted, as the agent argument should be pronominal.

(14) a. umoh itoh akau beloi
house.obl that 1sg buy.abs
'that house was bought by me'

b. buku itoh akau nuleih
book that 1sg act.write.obl
'that book, I wrote [it]'

Note that neither the non-canonical passive nor the object topicalisation construction are attested in the Chapter 5 dataset. This can be attributed to the fact that the experimental design did not elicit pronominal forms, or references to speaker or hearer.

A range of additional verb forms in Pondok Tinggi can be derived by productive verbal affixes, which may affect the argument structure associated with that verb. Of particular interest to this introduction is the perfective prefix *ta-*, which was attested in a small number of responses in the picture description study in Chapter 5. An example of this construction is given in example (15). This form expresses that the patient is an involuntary or unwilling participant in the event (cf. Ernanda, 2017). Additionally, the patient argument tends to appear in initial position; however, in the dataset in Chapter 6, instances of agent-initial perfectives were found, indicating that word order may be somewhat relaxed in this construction.

[3] Examples are taken from the dataset in Chapter 5. Note that the word *uto* 'car' is a loan which does not display the phrasal alternation described above.

(15) uha ta-simbak bola
people.ABS PERF-touch.ABS ball
'a person is touched by a ball'

Another sentence form which is extensively attested in the Chapter 5 dataset is the adversative passive. This construction involves the verbal form *kena/kenao* (the absolute and oblique forms, respectively). The verb originally means 'to touch', however its use as an adversative does not always require that physical contact takes place (Ernanda, p.c.). This verbal form can occur alone in the role of verb, expressing simply that the subject was the patient of some adverse event. However it can also be followed by a bare verb, which provides more specific information about the event. In both cases, the initial argument in the structure is the undergoer of the event but typically understood as the subject of the construction. The two possibilities are illustrated in example (16). In this way we can consider this to be another passive construction, distinct from the di-passive (and the non-canonical passive). The communicative emphasis lies on the affectedness of the patient-subject.

(16) a. uha kenao bola
people.ABS ADVPASS.ABS ball
'a person is adversely affected by a ball'
b. uha kenao timbok bola
people.ABS ADVPASS.ABS hit.OBL ball
'a person gets hit by a ball'

This adversative passive form is attested in other Malay varieties (although there may be interlinguistic differences in the exact distribution of the construction; S.-F. Chung, 2005). Palmer notes the existence of "adversity passives" in unrelated languages such as Japanese and Korean; remarking that a similar adversative meaning is conveyed by the English *I had a book stolen* (1994:235). The issue here is that the grammatical status of these types of constructions is not so clear-cut, or easy to compare cross-linguistically. Nonetheless, a potentially interesting hypothesis that arises from this for sentence production research is that adverse affectedness of the patient is likely to provoke patient-prominent constructions cross-linguistically, which may in some languages be distinct from (and coexistent with) other patient-prominent constructions. The issue of speakers selecting between multiple patient-prominent forms is revisited in Chapter 5.

CHAPTER 3

Simply describing in Tarifiyt Berber and Dutch

3.1 Introduction

In this chapter, I present data from a *simply describing* picture description experiment conducted with Tarifiyt and Dutch speakers. In this experiment, participants were asked to describe line-drawn scenes in one sentence. Each scene depicted either an inanimate agent acting on an inanimate patient (e.g. a car hitting a lamppost) or an inanimate agent acting on an animate patient (e.g. a car hitting a woman). The design of this experiment was based on Prat-Sala and Branigan (2000). The goal of this experiment was to investigate how animacy balance in the scene affected the structure of the spoken response in these two languages. Through a comparison of results between the two languages, this study provides further insight into how divergent language types can inform the theory of grammatical encoding.

This experiment was previously reported in a basic form in Dutton (2012); however, in this chapter I provide a new, fully revised presentation of this study, which has a number of advantages for the current thesis. Firstly, this revision presents an entirely new analysis. In particular, this analysis uses appropriate, up-to-date statistical techniques for the type of data involved. The original data analysis in Dutton (2012) followed procedures that were previously standard in the field, but which are now generally recognised as inappropriate (Jaeger, 2008); moreover, the original analysis permitted only limited insight into the data. Thanks to the updated analysis presented in this chapter, it is possible in turn to engage in an updated discussion of the findings from this study. The identical design and execution of the two experiments permits direct comparison of a newly studied language (Tarifiyt) with a language that falls within the well-researched handful (Dutch). To this end, I also conduct a (new) combined analysis, in order to inform our understanding of between-language differences from a statistical perspective.

Secondly, through a more detailed than usual presentation of the practical steps taken in data collection, preprocessing and analysis, I lay the basis for further discussion of the methodological issues involved in conducting sentence production experiments in typologically diverse languages. As described in Chapters 1 and 2, on the one hand there exists an emerging consensus that we need to work with a wider, more diverse range of languages; meanwhile, on the other hand, there are a number of challenges that we still need to overcome in order to implement this change. On the surface, the most salient challenges seem to stem from the fact that we need to be able to conduct research in less-controlled, non-laboratory settings with unfamiliar cultural groups. However, there are additional questions and challenges that arise at the point of data processing, analysis and interpretation. It is these issues which I aim to highlight and explore further in the ensuing chapters, and the current chapter provides the necessary conceptual basis for doing so.

In what follows, I will first describe the aims, design and procedure, after which I will present and explore the data gathered in this experiment, including new visualisations and a new analysis of the data using mixed effects binomial logistic regression modelling. In a discussion of the results of this analysis, I will explore the insight that can be gained with regard to the theory of sentence production described in Chapter 1. Points made in this chapter form the basis for the studies reported in the subsequent three chapters. Therefore, this chapter concludes with a review of questions raised and how they are to be addressed in Chapters 4, 5 and 6.

3.1.1 Aims of the current study

As indicated in the foregoing section, this chapter has both theoretical and practical goals. Firstly, this chapter seeks to compare and contrast structural choice in two typologically divergent languages in order to shed light on the question of cross-linguistic differences in structural choice. Secondly, the goal of this chapter is to highlight certain practical and methodological issues that will be explored further in later chapters.

As discussed in Chapter 1, there has been significant discussion in the literature about whether animacy affects sentence production through an impact on grammatical function assignment, or through linear order (through functional and positional processing). A number of studies on various different languages have led to the consensus that both grammatical function assignment and linear order are affected by the animacy of arguments. In particular, the consensus is that a highly animate (viz. highly accessible) patient argument is both more likely to be assigned subject function, *and* more likely to appear in an earlier linear position. This explains the finding that when the patient is more animate than the agent, there is an increased incidence of both passivisation and object topicalisation. These two structures have in common that they both encode prominence of the patient in linguistic form. However, they do not have identical distribution across languages. In particular, Tarifiyt and Dutch, the languages of this study, display complementary profiles in this regard.

The overall prediction of the original study Dutton (2012) was that, in line with previous studies, speakers of Dutch and Tarifiyt would make increased use of alternative structures such as passive and object topicalisation to describe pictures where the patient was more animate than the agent. The discussion in the original study was focused on the passive structure in Tarifiyt. The motivation for this focus was the importance of the passive structure in previous *simply describing* studies (as highlighted in Chapter 1). Tarifiyt is described as having a passive structure that is truncated – that is, it does not permit the inclusion of a by-phrase; the same as is described for Odawa (Christianson and Ferreira, 2005; see also Section 2.3.1). The original study in Dutton (2012) sought to establish whether this structure would be used in Tarifiyt picture description responses; the discussion was then centred on why the passive was not attested in the Tarifiyt data, despite being possible in the language. The main question raised was: if a certain structure is possible in the language, what prevents it from being considered for output? The current chapter revisits this question with new perspectives. Rather than focusing narrowly on the absence of the Tarifiyt passive, this chapter takes a broader perspective to discuss the status of patient-prominent structures in both Tarifiyt and Dutch.

The current experiment, being the first of its kind conducted in Tarifiyt Berber (and as far as I am aware, the first sentence production experiment on any Berber language), presented a range of challenges in design, processing and execution, of the type described in Chapter 2. For example, it was not possible to use the written modality; in addition, it was necessary to test participants in a non-lab environment; and furthermore, there was (and still is) also a lack of similar research on related languages to draw on in forming hypotheses and interpreting data. The manner in which these issues were resolved is indicated only obliquely in the original work. In the current chapter, these processes and choices are now presented with more transparency. A number of practical and methodological issues were raised in Section 2.2; through that discussion, however, it became apparent that the more intractable issues facing researchers working on diverse languages arise at the interface of theory with practice, in particular, the way we resolve issues related to study design and control and the difficult decisions that must be made in preprocessing and coding messy or complex data. In providing a more transparent account of practical decisions taken in this study, I lay the basis for considering the ways in which the standard approach to *simply describing* studies might be adapted in order to gain better insights from the study of diverse languages.

3.2 Method

3.2.1 Design and procedure

Generally speaking, this experiment was a replication of the animacy condition in the study of Prat-Sala and Branigan (2000). In this experiment, participants were asked to simply describe line-drawn scenes. These line-drawn images depicted either an inanimate agent acting on an inanimate patient (e.g. a car hitting a lamppost) or an inanimate agent acting on an animate patient (e.g. a car hitting a woman). The critical materials comprised 16 pairs of line-drawn pictures, depicting transitive scenes. In one picture of a pair, an inanimate agent acted on an inanimate patient, and in the other picture, the same inanimate agent was depicted acting in the same way upon an animate agent. Example stimuli are provided in Figure 3.1. Stimuli were counterbalanced so that each participant only saw one picture from each pair. The direction of the action in the picture was also counterbalanced; this was done using horizontally flipped (i.e. mirror-image) versions of the pictures. In all, each participant received 16 critical trials. Each participant also responded to 16 filler pictures. The filler pictures depicted equal animacy interactions. An extensive presentation of the procedure, stimuli and design can also be found in Dutton (2012).

Predictions regarding structural choice

The predictions for this experiment were as follows. As described above, the expectation based on previous research was that speakers would make increased use of alternative structures when the patient argument in the scene was more animate than the agent. Namely, the prediction was that they

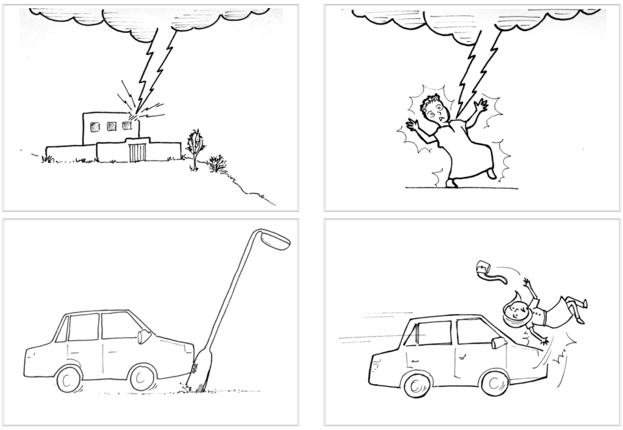


Figure 3.1: Example of stimuli used in the current experiment.

would make use of structures that afford the patient more prominence, in terms of grammatical function assignment and/or linear ordering. However, given the typological profiles of the two languages, there were different expectations for the specific structures used to achieve this in the two languages.

Based on previous studies, including related languages such as English (Bock, 1986b) and German (van Nice & Dietrich, 2003), the expectation was that the Dutch speakers would have an overall preference for SVO active structures, but produce a higher rate of passives in response to the animate-patient condition (also SVO; i.e. where patient argument is mapped to subject). Despite the existence of an object-topicalisation structure in Dutch (cf. Section 2.3.1), previous research did not indicate that this would be a likely structure in this experiment. For the Tarifiyt data, the expectation was less clear-cut. Either the verb-initial or the subject-initial transitive structure was predicted to be the dominant structure. The question was then, if an alternative structure was found to be produced more often in response to the highly animate patient condition, to what extent this would be the object-topicalisation structure, or the passive. Although the passive in Tarifiyt is truncated (i.e. preventing explicit mention of the agent), increased use of truncated passive was nonetheless attested in the study by Christianson and Ferreira (2005).

It is of additional interest to consider the possibility of encountering verb-initial structures in the data. While Tarifiyt is recognised as a verb-initial language, structures where the preverbal position is occupied by an epenthetical subject are productive in Dutch (cf. Section 2.3.1). The incidence of these structures in the data, and the surrounding issue of how to categorise and analyse them, will also form part of the discussion in this chapter.

3.2.2 Participants and procedure

Both experiments were conducted in the Netherlands. Twenty-five native speakers of Tarifiyt Berber took part in the Tarifiyt experiment. The Tarifiyt participants (4 females; age range 26–66 years, M^{age} = 40) were people from Morocco residing in the Netherlands, typically along with other family members. Thirty-three native Dutch-speaking participants took part in the Dutch experiment. Dutch participants (20 females; age range 19–68 years, M^{age} = 28) were students at Leiden University. For all participants, the language of the experiment was acquired in infancy and was the first language of family life. Both groups had a bilingual profile: Dutch participants all used English in daily life at the university, while Tarifiyt participants, in turn, had Dutch as a prevalent language of daily life. It should be noted that the Tarifiyt participant group had a more heterogeneous profile than the Dutch one; more discussion of the participant groups and languages can be found in Section 2.3.1.

The experiment was presented on a 10-inch Dell laptop. Instructions for the experiment were entirely presented in the language of the experiment. Due to the issues surrounding the use of written language, the instructions were presented in spoken rather than written form. These audio instructions were pre-recorded, in order to ensure uniformity across participants. The recordings were made by an early bilingual speaker of Tarifiyt and Dutch – that is to say, the speaker in these recordings was the same in both of the languages. This was done to control for any confounding effects that could arise from using different voices in the different experiments. The audio instructions informed participants that they would see a number of pictures and each time hear a prompt question, 'What has happened?', upon which they should describe the scene they saw. They were asked to describe the picture in one sentence, as if to someone who could not see the picture. There were four pictures provided as a short practice session, following which the instruction to use one sentence only was emphasised if necessary. Participants were not further directed as to the form their responses should take. The experiment was entirely self-paced. Participants' responses were recorded using a digital audio recorder.

3.3 Data processing

3.3.1 Exclusions

As mentioned in Section 1.2.3, in *simply describing* experiments, there is usually a high rate of response exclusion. Typically, the responses that are included in the dataset are those which conform to some target form that the researcher has in mind, meaning that non-target responses are the ones that "uncertain bearing on the questions of interest" (Bock, 1996:407). For example, the agency of the sentence should reflect the agency which is intended in the picture (for example, if a boy is hit by a ball, it is in most experiments 'wrong' to describe this as *a boy is playing with a ball*, even though that would be for many speakers a felicitous description of the scene); in addition, responses must normally mention both entities in the picture (so, a response such as *a boy has been hit in the head* would normally also be rejected). Lastly, many studies require that the sentence that the participant produces should be fluent, with the description of the transitive action occurring within the first clause of the response (although see Sauppe, 2017:39 for an example of a study where a certain level of disfluency is tolerated).

In line with the approach of previous sentence production work, I applied these general principles to the current dataset. This procedure resulted in the exclusion of 39% of the Tarifiyt data and 34% of the Dutch data, which reflects the proportions of exclusions found in earlier studies. The number (and percentages) of exclusions by specific criteria are displayed in Table 3.1. Below, I provide a short description of each of the exclusion criteria in the table.

- **non-target agency** Responses must construe the action as having been initiated by the inanimate agent. If the intended patient is framed as an agent, or the inanimate agent is construed as an animate (e.g. the driver of a car, rather than just a car), then the response is excluded.

Likewise, responses where the action is described as arising from an unseen agent are rejected (e.g. ‘someone is throwing a ball at the man’ where the person doing the throwing is not pictured).

- **multiple clauses** The description of the scene should not be spread across multiple clauses (e.g. ‘there is an old man; he gets hit by a bike’. Note however that phrases of the type ‘an old man who gets hit by a bike’ are permitted, following Prat-Sala and Branigan (2000).
- **only one entity mentioned** This is usually where the speaker has described the scene with an intransitive, involving only one of the two entities (or an agentless passive, which was found in one Dutch response).
- Responses are naturally excluded where there is **no response**, or where the response shows a **general misinterpretation** of the scene (such as describing an image of a kite hitting a chair as ‘the chair is blown away by the wind’).
- **non-verbal construction** Essentially, each response must include a verbal phrase that describes the action of the inanimate agent on the patient.
- **rephrased or fragmented response** A response is rejected if the speaker initially makes a number of failed attempts to describe the picture (even if a satisfactory description is eventually produced).
- **agent expressed as instrument/prepositional** Responses must encode the inanimate agent as an agent (leading to exclusion of responses like ‘a balloon exploded with a cactus’). However, responses were accepted if the patient was expressed in a goal/direction-oriented phrase, such as ‘a tree fell on a car’. This approach follows van Nice and Dietrich (2003).
- **reflexive** The agent and patient should be presented as being in a transitive, not reflexive, relationship (leading to exclusion of responses like ‘the boat and the plane are shooting each other’).

Following exclusions, the Tarifiyt dataset contained a total of 245 responses and the Dutch dataset 350 responses. In both languages, the responses were still evenly distributed across the two experimental conditions (Tarifiyt: 49% of responses were in the IA condition; Dutch: 47% of responses were in the IA condition). Following exclusions, every response was coded for the structure it represented. In the next section, I describe the decisions taken in preparing the raw data for analysis.

3.3.2 Preparing for analysis

In *simply describing* sentence production experiments, it is customary to analyse the data as a binary outcome variable. In English this has typically been done by comparing the proportion of passives to actives (Bock, 1982). Even when participants produce more than two types of structure in their picture descriptions, previous studies have tended to analyse the data along binary lines. For example, in Spanish, both passives and object topicalisations that can be used to prioritise the patient argument; however, in their experiment, Prat-Sala and Branigan (2000) amalgamated both of these structures into a single ‘non-canonical’ category, to be contrasted with ‘canonical’ structures (i.e. SVO actives). Alternatively, there are languages where grammatical function assignment and linear order intersect

	Tarifiyt		**Dutch**	
Total exclusions	155	(of 400 responses)	178	(of 528 responses)
Exclusion Reason		% of exclusions		% of exclusions
non-target agency	41	(26.5)	56	(31.5)
multiple clauses	30	(19.4)	61	(34.3)
only one entity mentioned	21	(13.5)	19	(10.7)
no response	20	(12.9)	3	(1.7)
general misinterpretation	18	(11.6)	7	(3.9)
non-verbal construction	14	(9.0)	3	(1.7)
rephrased or fragmented response	7	(4.5)	6	(3.4)
pronominalisation	3	(1.9)	0	(0)
agent expressed as instrument/prepositional	1	(0.6)	22	(12.4)
reflexive	0	(0)	1	(0.6)

Table 3.1: Distribution of exclusions in Tarifiyt and Dutch datasets

each other. For example, in some languages, it is possible to have an active with or without object topicalisation, and a passive with or without object topicalisation, resulting in four different permutations (such as is described for Japanese in Tanaka et al., 2011). Studies concerning languages with this kind of profile have also opted to code and analyse the two variables (grammatical function assignment and linear order) separately. For example, Tanaka et al. (2011) ran one analysis with “verb form” (i.e. Active vs. Passive) as the binary outcome, and another analysis with “word order” as the binary outcome (see also Norcliffe, Konopka, et al., 2015). An even more complex language production experiment analysed along binary lines is Christianson and Ferreira’s study in Odawa (2005).

Unless all the responses happen to fall into a clean binary split on a variable of interest, then some decisions need to be taken in order to prepare the data for binary analysis. The example of Prat-Sala and Branigan (2000) offers one solution, which is to group all the data under a more general dichotomous variable, namely whether the structures are ‘canonical’ or not. The examples of Christianson and Ferreira (2005), Tanaka et al. (2011) and Norcliffe, Konopka, et al. (2015) offer another solution, namely running separate analyses for two or more dichotomous outcome variables.

In the current data, languages, the responses fell largely into two categories; however, the categories were not the same for the two languages. Table 3.2 displays the different types of structures per condition, for each of the two languages.

We can immediately note that there are no passives found in the Tarifiyt data, and no object-topicalisation structures found in the Dutch data (despite the existence of a passive structure in Tarifiyt, and the existence of an object-topicalisation structure in Dutch, as described in Section 2.3.1). In line with expectations, in the Tarifiyt data, most of the responses were either subject-initial (SVO) active or object-initial (OVS) active. In Dutch, the responses were largely either active or passive (SVO). In 16% of all responses, the verb preceded mention of either of the two referents in the picture (Dutch presentative *er*, Tarifiyt verb-initial). In all of the Tarifiyt verb-initial structures, the agent preceded the patient, and the patient was realised in a prepositional structure rather than as a direct object. Among the Dutch presentative *er* structures, the active/passive distinction was present, as illustrated in the upper panel of Table 3.2.

Tarifiyt

Grammatical **function assignment**	**Linear order** Agent - Verb - Patient	 Patient - Verb - Agent	 Verb - Agent - Patient
Active	76% ṭṭiyyara tenɣa aɛeskar aeroplane 3SG.F.kill FS.soldier	20% aḥenjir tekka xa-s traktur FS.boy 3SG.F.pass on-3SG tractor	4% tewḍa-d essjart x ṭṭumubin 3SG.F.fall-HITHER tree on car
Passive	0%	0%	0%

Dutch

Grammatical **function assignment**	**Linear order** Agent - Verb - Patient	 Patient - Verb - Agent	 Verb - Agent - Patient
Active	63% een auto is tegen een lantaarnpaal opgereden a car AUX.3SG against a lamppost run.into.PST	0%	9% er rijdt een stoomtrein over een bezemsteel EPTH drive.PRES a steam.train over a broomstick
Passive	0%	25% een man wordt geraakt door een voetbal a man AUX.3SG hit.PST by a football	3% er wordt een politieagent aangereden door een bus EPTH AUX.3SG a police.agent run.over.PST by a bus

Table 3.2: Overview of structures produced in Tarifiyt (top) and Dutch (bottom) in the current experiment. The tables are organised by linear order (order of mention of agent, patient and verb) and grammatical function assignment (active or passive, meaning whether agent or patient was assigned to subject function, respectively). For each combination attested in the data, the table shows the percentage of the dataset (by language) and an illustrative example taken from the responses.

The research questions concern the grammatical functions assigned to agent and patient, and their relative ordering in the sentence. On the basis of previous studies, and given the distribution of data seen in Table 3.2, the decision was taken to group all data according to the following binary variables: for Tarifiyt, the binary grouping was *subject-initial* vs. *object-initial*, while in Dutch, the binary grouping was *active* vs. *passive*. In both languages, the decision was taken to group these 'verb-initial' structures according to these same binary categories. That meant that all the Tarifiyt VSO structures were all assigned to the *subject-initial* category, since the agent precedes the patient, and the agent is assigned the subject function. However, each of the Dutch presentative *er* structures were allocated to the relevant active or passive category.

In order to be able to make cross-linguistic generalisations, including a pooled analysis, a further stage of abstraction is required. Namely, we need to identify a common grouping that forms an umbrella for the binary splits in both languages. As discussed in Section 3.1.1, the prediction was that both languages would use alternative structures that prioritise the patient by either assigning it subject function or giving it an earlier linear position. Therefore, a generalisation that spans the two languages is whether speakers use more *patient-prominent* structures when the patient is more animate than the agent. In other words, the cross-linguistic grouping concerns whether the agent is the more 'prominent' argument (Dutch active, Tarifiyt subject-initial active) or the patient is the more 'prominent' argument (Dutch passive, Tarifiyt object-initial active). Therefore, from a cross-linguistic perspective, the binary categorisation of the data is in the categories *agent-prominent* and *patient-prominent*. Figure 3.2 shows the relative distribution of these categories in the two languages.

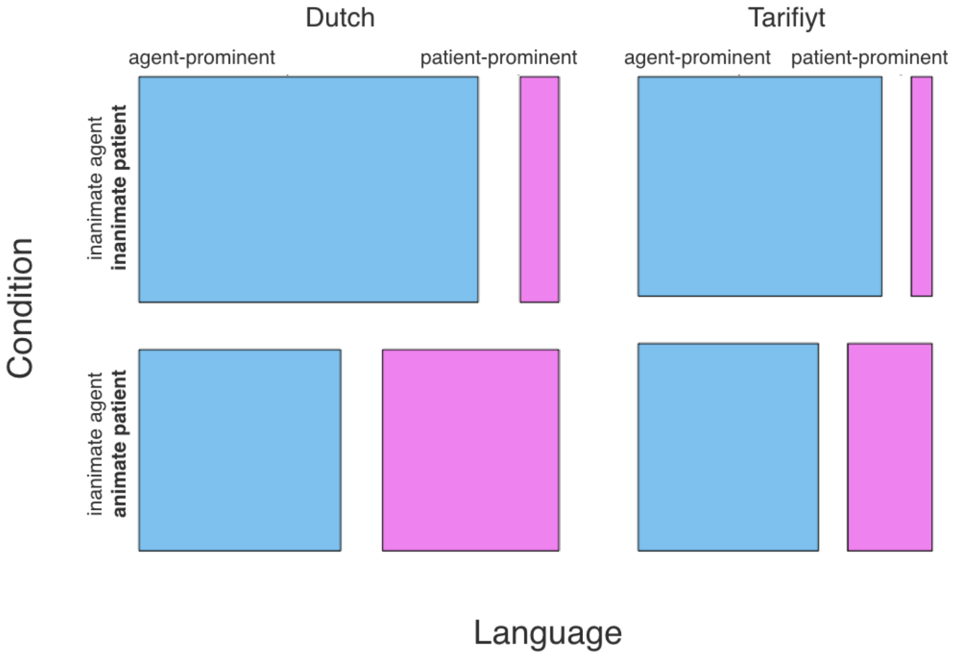


Figure 3.2: Distribution of agent-prominent and patient-prominent structures in the inanimate patient condition (top) and the animate patient condition (bottom), in Dutch (left) and Tarifiyt (right). Blue represents agent-prominent structures, and purple represents patient-prominent structures.

3.4 Analysis

As mentioned above, the data from sentence production experiments are customarily analysed along binary lines. Earlier studies in the psycholinguistic literature opted to analyse the data in terms of proportions – for example, whether the proportion of passives is higher in one condition than the other. This technique relied of the use of ANOVA or T-tests, and analysed the proportions as if they were scores on a continuous scale. This is the approach that was used for the current data in Dutton (2012), following the implementation of Prat-Sala and Branigan (2000) and Christianson and Ferreira (2005) among others. However, analysing categorical data in this way is statistically inappropriate. This issue is discussed in detail by Jaeger (2008). In particular, the ANOVA over proportions approach violates key assumptions of ANOVA and can lead to spurious results. One option to sidestep this problem is to transform the data; however, even when data is transformed, the analysis is still not optimal: as Jaeger (2008) notes, this solution was only ever used as an approximation for more appropriate categorical data analysis techniques. Nowadays, however, the increased affordability of computing power and accessibility of statistical tools such as R (R Core Team, 2017) have enabled us to move towards a new standard of analysing the data with the appropriate techniques, including logistic regression modelling.

Logistic regression allows us to model the probability of a given observation falling into one or another outcome category, given the particular combination of scores on the predictor variables. Logistic regression models can also incorporate random effects structure (i.e. we can have mixed effects models). This means that we can simultaneously control for the correlation among responses from a single participant and correlation among responses to a single item (Baayen, Davidson, & Bates, 2008). This again represents an improvement over older approaches such as that employed in Dutton (2012).

In what follows, I present a new analysis of the Tarifiyt and Dutch data using binomial logistic mixed effects regression. I use the *glmer* function of the *lme4* package in R (D. Bates, Mächler, Bolker, & Walker, 2015; R Core Team, 2017). The aim is to model the effect of the experimental condition on the prominence of agent and patient in the scene descriptions. condition is a factor variable with two levels: inanimate patient and animate patient. prominence is a factor variable with two levels, agent-prominent (Dutch active, Tarifiyt subject/verb-initial active) and patient-prominent (Dutch passive, Tarifiyt object-initial active).

For each of the two languages, the modelling procedure was as follows. Binomial regression models were built with random intercepts for stimuli, and random slopes for the effect of condition across participants. First an intercept-only model was built, with prominence as the outcome variable. Then, prominence was modelled as a function of condition (i.e. whether the picture depicted an inanimate or animate patient). These two models were then compared using ANOVA (note that ANOVA here is not being used for the analysis itself, but for model comparison). Tables 3.3 and 3.4 provides the summary of the models and comparisons for each language.[1]

In both languages, the inclusion of condition in the model resulted in a significantly better fit than the intercept only model. The results indicate a strong effect of condition on prominence. In other words, in both languages, when the patient is more animate than the agent, patient-prominent structures are significantly more likely. We can also interpret the model output in terms of the structures found in the individual languages. The indication is that in Tarifiyt, object topicalisations are significantly more likely in the animate patient condition than the inanimate patient condition; meanwhile, in Dutch, passives are significantly more likely in the animate patient condition than the inanimate patient condition.

An additional possibility afforded by this method of analysis is to include both languages in one model. This is useful in order to assess how the effect of condition on prominence differs between the languages. In other words, we can ask the question: is the probability of producing a patient-prominent response for the animate-patient scenes different between Dutch and Tarifiyt? This comes down to assessing the existence of an interaction between the critical manipulation of the experiment (i.e. the condition variable) and the language of the experiment. Results of this analysis should be interpreted with caution, because the two participant groups are not closely matched, essentially meaning an increased chance of between-group confounds. Nonetheless, bearing this in mind, such a post-hoc analysis can still provide further insight into the data and point towards possible areas for further study.

In this pooled analysis, language is a factor variable, with two levels, Dutch (*NL*) and Tarifiyt (*TB*). For the purposes of this analysis, Dutch is taken as the reference level.[2] First, an intercept-only

[1] All model output tables in this thesis make use of the *TexReg* package in R (Leifeld, 2013).

[2] This means that the Tarifiyt terms in the model reflect how Tarifiyt diverges from Dutch; this is arbitrary and not an

	intercept only	condition
(Intercept)	-3.67^{**}	-5.21^{**}
	(1.14)	(1.71)
ConditionIA		3.56^{*}
		(1.72)
ANOVA (model comparisons) χ^2		5.98^{*}
AIC	191.73	187.75
BIC	209.23	208.76
Log Likelihood	-90.86	-87.87
Num. obs.	245	245
Num. groups: STIM	31	31
Num. groups: ParticipantID	25	25
Var: STIM (Intercept)	8.08	6.93
Var: ParticipantID (Intercept)	1.96	4.02
Var: ParticipantID ConditionIA	5.08	0.77
Cov: ParticipantID (Intercept) ConditionIA	3.15	1.76

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3.3: Binomial logistic regression analysis modelling the effect of animacy condition on the likelihood of patient-prominent structures in Tarifiyt Berber.

model was built, with the same random effects structure as before. Then, successive models were built including fixed effects for condition, then adding language, and finally also the interaction of condition with language. All models were again compared using ANOVA. The results are displayed in Table 3.5. In addition, the plot in Figure 3.3 provides a visual representation of the interaction, by plotting the probabilities of each type of response per condition and per language, as predicted by the full model in 3.5.[3]

Model comparison using ANOVA reveals that including a term for language improves the fit only marginally ($\chi^2 = 3.36$, $p = 0.07$). Including an interaction term does not improve the fit. This suggests that there may be a marginal difference in the prevalence of the patient-prominent structures between Dutch and Tarifiyt, but the model does not provide evidence of an interaction.

The predicted probabilities plot in Figure 3.3 sheds more light on the situation. We indeed see that the direction of the effect is similar between the two languages: in both languages, there is a higher incidence of patient-prominent structures in the animate patient condition. However, we do see a larger discrepancy between the languages in the animate patient condition than the inanimate patient condition. Here, the (mean) probability of a patient-prominent structure in Dutch is around 50%, whereas in Tarifiyt it is around 20%. This reflects the distribution of responses in Figure 3.2; which is presumably what underpins the difference in magnitude of the effect of condition in the Tarifiyt and Dutch models. However, there appears to be much greater variability overall in the animate-

implication that Dutch is viewed as being in any way more ‘default’.

[3] The interaction plot here is made with the R package *sjPlot*, version 2.4.0 (Lüdecke, 2018).

	intercept only	CONDITION
(Intercept)	-2.35^{***}	-5.41^{*}
	(0.61)	(2.19)
ConditionIA		5.17^{*}
		(2.25)
ANOVA (model comparisons) χ^2		17.92^{***}
AIC	321.69	305.77
BIC	340.98	328.92
Log Likelihood	-155.84	-146.88
Num. obs.	350	350
Num. groups: ParticipantID	33	33
Num. groups: STIM	32	32
Var: ParticipantID (Intercept)	2.87	11.44
Var: ParticipantID ConditionIA	11.54	14.84
Cov: ParticipantID (Intercept) ConditionIA	-2.59	-10.80
Var: STIM (Intercept)	3.02	2.17

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3.4: Binomial logistic regression analysis modelling the effect of animacy condition on the likelihood of patient-prominent structures in Dutch.

patient condition (i.e. in both languages). The error bars here are extremely wide, indicating a high degree of uncertainty in these estimates. This reflects the finding that despite apparent differences between the Tarifiyt and Dutch models, there is only a marginal effect of LANGUAGE in the overall bilingual analysis.

3.5 Discussion

Both the Tarifiyt and Dutch data showed overall predominance of the subject-initial active structure. However, in both languages, there was increased probability of patient-prominent structures in the animate patient (IA) condition compared with the inanimate patient condition (II). The current, updated analysis validates the findings of the original study in Dutton (2012). Additional insight was provided in this new analysis, through the possibility to additionally probe the effects of animacy from a cross-linguistic perspective. Based on plots and by-language analyses, it appears that the effect of animacy condition on patient-prominence may be larger in Dutch than in Tarifiyt. However, the bilingual binomial regression analysis does not provide statistical support for this difference.

Overall, we see a higher incidence of patient-prominent structures in the condition where patient is more animate than agent. However, the specific structures used to achieve this are notably different between the two languages. Namely, the Dutch speakers produce a higher rate of passives in response to the animate-patient condition, and the Tarifiyt speakers produce a higher rate of object topicalisa-

	intercept only	condition	cond+lang	cond∗lang
(Intercept)	-2.24^{***}	-3.94^{***}	-3.62^{***}	-3.82^{***}
	(0.52)	(0.80)	(0.83)	(0.89)
ConditionIA		3.33^{***}	3.50^{***}	3.84^{***}
		(0.90)	(0.94)	(1.00)
LanguageTB			−1.16	−0.38
			(0.64)	(0.82)
ConditionIA:LanguageTB				−1.17
				(0.88)
ANOVA (model comparisons) χ^2		15.98^{***}	3.36	1.78
AIC	501.11	487.13	485.77	485.99
BIC	523.05	513.46	516.49	521.09
Log Likelihood	−245.55	−237.56	−235.88	−234.99
Num. obs.	595	595	595	595
Num. groups: ParticipantID	58	58	58	58
Num. groups: STIM	32	32	32	32
Var: ParticipantID (Intercept)	2.10	3.74	4.22	3.81
Var: ParticipantID ConditionIA	3.03	2.49	2.57	2.55
Cov: ParticipantID (Intercept) ConditionIA	0.72	−0.79	−1.41	−1.09
Var: STIM (Intercept)	4.16	2.56	2.67	2.69

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3.5: Binary logistic regression analysis modelling the effect of animacy condition on the likelihood of using patient-prominent structures, including how this likelihood is affected by the language of the speaker.

tions in response to the animate-patient condition. Conversely, the Dutch speakers did not produce any object-topicalisation structures in response to the animate-patient condition; similarly, the Tarifiyt speakers did not produce any passive structures in response to the animate-patient condition.

A simplistic way of interpreting this dataset could be to say that in both languages, linear order is what is ultimately targeted; that is, although the Dutch data displayed a voice alternation, the ultimate effect was that the patient could thereby appear in sentence-initial position. Subject function in Dutch, as in English, is highly correlated with early linear position. Indeed, the fact that subject is so strongly correlated with linear order in English has meant that passivisation is sometimes characterised as a means to achieve linear priority for a patient argument. In other words, under this view, the passive is a means to accommodate an early positioned patient (e.g. Kempen & Hoenkamp, 1987; van Nice & Dietrich, 2003). However, studies from languages where subject assignment need not correlate so closely with sentence-initial position indicate that functional processing and positional processing may be independently affected (as discussed in Section 1.2.2). Ultimately, this means that not all findings can be accounted for as emanating from linear ordering effects alone (Pickering & Ferreira, 2008).

A more nuanced perspective on the data would be to say the two languages studied here respectively demonstrate the two predominant strategies to realise a patient argument in a more promin-

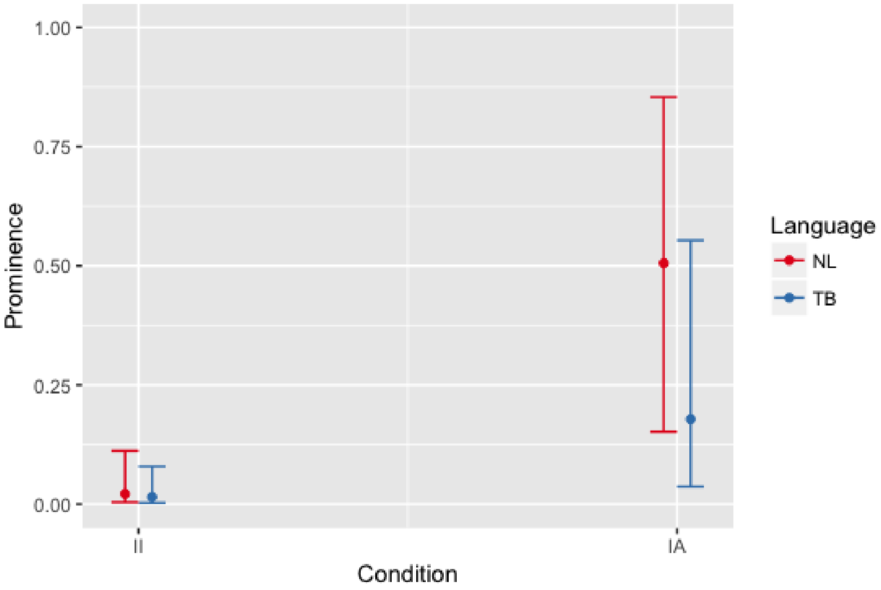


Figure 3.3: Predicted probabilities of the CONDITION*LANGUAGE interaction model, showing predicted values for Dutch (red) and Tarifiyt (blue), with error bars indicating 95% confidence intervals.

ent sentence position, which have been identified in cross-linguistic studies: (1) using grammatical function assignment to map the patient to subject function, which is in turn correlated with early sentence position (Dutch); (2) using linear ordering to give the patient an early sentence position, but still assigning it the canonical object function (Tarifiyt). Moreover, the current data suggests that the situations which provoke the passive in Dutch are actually the same ones that provoke object topicalisation in Tarifiyt.

The two languages are not only complementary in the structures that are attested in the data, but in the structures that are notably *absent* from the data. In Dutch, the passive structure seems to be the primary manner of encoding agent prominence; meanwhile, the passive is not attested in the Tarifiyt data at all. Likewise, in Tarifiyt, the object-topicalisation structure appears to be the primary manner of encoding agent prominence; meanwhile, the Dutch object-topicalisation structure is not attested in the dataset. To summarise this situation, we could say that superficially similar structures in Dutch and in Tarifiyt (passive, object topicalisation) have different distributions in the two languages.

Recalling the question posed in Section 3.1.1 (and with reference to Dutton, 2012), it appears that even though a structure is possible in the language, it is not necessarily considered for output. A question then arises for our universal model of production: given a uniform communicative situation (such as that represented by the current experiment), how can we account for the differences in output across languages? If the model of sentence production applies across all languages, how can it be that functional and positional processing are differently affected under the same communicative circumstances? The answer to this question seems to lie in a better understanding of the manner in which typological features of a language constrain the effect of general cognitive mechanisms in grammatical encoding.

Myachykov, Thompson, Scheepers, and Garrod (2011) provide a possible account for the role of typological differences in the outcomes of sentence production studies. Their account rests on the idea of a hierarchy of functional and positional processing. This account is based on findings in Russian and Finnish sentence production, where a similar pattern to Tarifiyt is attested: passive is marginal, but ordering variations are permitted (Myachykov & Tomlin, 2008), contrasting with languages like English that make use of passivisation to encode patient prominence (similar to Dutch in the current experiment). The authors propose that the typological differences between languages like Tarifiyt and Russian on the one hand, and languages like Dutch and English on the other hand, fall out straightforwardly from a hierarchical two-stage model. They propose that, where possible, variation in subject assignment is used to encode a prominent patient; however, when this is not available, linear ordering is used instead as an alternative strategy (Myachykov et al., 2011:104).

At first sight, this approach ties in well with the model of Bock and Levelt (1994) model, since it relates cross-linguistic variation directly to the two 'stages' of functional and positional processing (as shown in Figure 1.1, Section 1.2.2). However, it is not as parsimonious as it first seems. Specifically, there are two assumptions necessary for this account, which are both open to debate. Firstly, we must first agree that functional and positional processing are two distinct stages of production. However, as pointed out by Kempen and Harbusch (2004) and Pickering and Ferreira (2008), the idea of the two-stage model in itself is not parsimonious. Essentially, to account for animacy effects on sentence production, a two-stage model requires us to posit an effect of animacy at two separate stages of processing. Secondly, a two-stage hierarchical account does not seem parsimonious to account for the findings from languages where the passive is marginal or absent. Even languages such as Tarifiyt, where passive is not attested in the data, do make use of grammatical functions to express relationships between arguments in the sentence (so, it is not the case that functional processing is irrelevant for such languages). However, if we consider functional prominence as the primary strategy, it appears that languages that do not make use of this strategy must first reject the possibility every time a patient-prominent structure is required. Moreover, there is still another component required to support this account: the notion of the grammatical function assignment option in certain languages being 'unavailable' is apparently not due to *absence* of passive in languages such as Russian and Finnish (cf. Myachykov & Tomlin, 2008). The mechanism of structural 'availability' thus needs additional elaboration.

To explore what might make certain strategies more or less *available*, a more fruitful perspective could be to see passivisation (a functional mapping of the patient argument) and object topicalisation (the positional mapping of the patient argument) as strategies that are not hierarchically organised, but rather differently weighted across languages. The question then arises: how can we begin to understand such a weighting cross-linguistically? It seems helpful to come up with some kind of framework which we could use to predict the likely weighting of functional and positional strategies in a given language.

Subject vs. topic prominence account (Li & Thompson, 1976; Butler et al., 2012, 2014)

An alternative cross-linguistic approach to the encoding of prominence, which sees subject assignment as one strategy rather than the predominant strategy, derives from the subject and topic prominence typology of Li and Thompson (1976). This typology informs us about the different strategies afforded by specific linguistic systems for encoding prominence in sentence form. So-called ‘subject-prominent’ languages rely heavily on the grammatical function of subject to encode prominence. The result of this situation is that, in subject-prominent languages, the grammatical function of subject is closely correlated with agenthood, sentence-initial position and animacy (cf. Comrie, 1989). So-called ‘topic-prominent’ languages, on the other hand, have a tendency to encode prominence through marking an argument as sentence topic. Importantly, in these languages, the topic is distinct from subject function, and usually the topic is associated with the initial position in the sentence (Li & Thompson, 1976:465). Another feature of topic-prominent languages is that they tend to make little or no use of passivisation for expressing prominence of the patient (Li & Thompson, 1976:467). The account provided for this is that in topic-prominent languages, the availability of the topic role means that it is not necessary to map the patient to subject function in order to give it prominence.

Let us consider this typological generalisation with regard to the two languages in this experiment. Dutch could be said to be subject prominent: there is a reliance on subject function for the encoding of prominence, leading to an increased use of passives when the patient argument is more prominent. The characterisation of Dutch as subject-prominent is corroborated by Kiss (1995). Meanwhile, Tarifiyt seems to be a topic-prominent language: the passive is a very marginal structure, meanwhile, there is a sentential position associated with topic (i.e. the preverbal position; Lafkioui, 2014). When the patient is more animate than the agent, topicalisation allows a patient-object to be more prominently encoded. This view of Tarifiyt as topic-prominent is also in line with the characterisation of Berber by Kiss (1995:5), based on work by Calabrese (1987) and Ouhalla (1991).

This typological generalisation relates directly to cross-linguistic differences in sentence form, and has received considerable attention in the descriptive (and theoretical) literature. However, it has received much less attention in the psycholinguistic literature on sentence production, and has consequently had little impact on the formulation of sentence production theory. However, an exception to this can be found in the study of Butler, Jaeger, and Bohnemeyer (2012), conducted in Spanish and Yucatec Maya (also Butler, Jaeger, & Bohnemeyer, 2014). In this experiment, Butler and colleagues had participants describe short animations depicting transitive events, where relative animacy was manipulated. In addition, the topicality was manipulated through the use of different prompt questions, which asked about the agent, about the patient, or about the general situation (similar to the approach of Christianson & Ferreira, 2005). Sentence form in Yucatec Maya is noted to be sensitive to topicality, rather than grammatical relations (Bohnemeyer, 2009). However, the authors found that topicality not only played a mediating role in Yucatec Maya, but also was a stronger predictor than animacy in accounting for the Spanish data.The authors indicate a manner in which the topic and subject prominence typology could be integrated with the view of sentence production as described in Section 1.2.2. In particular, they suggest that functional processing may not only concern grammatical function assignment, but also discourse role assignment. In other words, grammatical and discourse roles may be considered functionally equivalent, but with cross-linguistic variability in the degree to which each of these strategies is used to encode argument prominence.

The link between the discourse role of topic and the *functional processing* stage of the model is tricky, given that when object topicalisation is found in the responses, this has typically been interpreted as an effect on *positional processing*. In other words, until now, ‘purely positional choices’ in word order in languages like Tarifiyt have been labelled as linear ordering effects; however, it is possible to view these as ‘grammatical choices’ on a par with subject assignment (M. Wagner, 2016:560). In other words, we may consider the possibility that the formal realisation of topic in many languages could ultimately be best understood as a functional effect, rather than a positional effect. This would allow us to account for the fact that yet other languages have a formal realisation of topic that is morphological rather than positional in nature (such as the *wa* particle in Japanese). Moreover, it would also concur with earlier findings indicating a distinction between linear ordering effects with a functional relevance, vs. pure ordering effects such as conjunct ordering (see M. Wagner 2016 for an overview). We may also bear in mind that the idea of positional effects as being functional in nature is echoed by the observation that some functional effects are positional in nature. For example, subject function in a subject-prominent language such as English is also indicated primarily by word order. Acknowledging this blurring of the functional/positional distinction may potentially lend further support to the idea that constituent structure is decided at a single stage (Pickering & Ferreira, 2008). Resolving these issues entirely is beyond the scope of the discussion here; but in any case, if we agree that there is a specific formal realisation of topic role in certain languages, the mechanism by which arguments are assigned to this formal role should be described in the sentence production model.

Clearly, future cross-linguistic comparative studies that target questions of topic and subject prominence are needed to assess whether this typological dimension can provide an account for the language-specific component of sentence production. One advantage of this account, however, is that it provides an external framework for generating language-specific hypotheses for such studies. That is, it is first possible to assess languages in terms of their characteristics, drawing on typological and theoretical work such as Li and Thompson (1976) and Kiss (1995), in concert with descriptive accounts of the language(s) in question. Based on this characterisation, we can then hypothesise a likely distribution of passivisation (functional mapping of patient to subject) and object topicalisation (patient-to-object mapping with early linear position) in *simply describing* studies. Generally speaking, to the degree that a language is subject-prominent, we would expect passivisation to be the main mode of encoding a prominent patient; to the degree that a language is topic-prominent, we would expect linear ordering to be the primary strategy. More fundamentally, this approach provides the means to gather experimental evidence for whether the typology proposed by Li and Thompson (1976) is a useful account of cross-linguistic differences in sentence production, by assessing whether, over a range of languages, the predictions are borne out.

With such a framework, we can also make more nuanced hypotheses. We can go beyond the broad, language-blind question of whether animacy affects grammatical function assignment, linear ordering or both. Instead, we can ask more targeted questions about how typological tendencies of languages interface with general cognitive principles, such as the accessibility of animate concepts. Over time, this approach could contribute to building an updated model of sentence production processes, in which universality is no longer a tacit assumption, but where the degree of universality and the degree of language-specificity form an explicit part of the picture.

3.5.1 Conclusion

In this *simply describing* study I investigated the effect of patient animacy (relative to the agent) on the form of transitive picture descriptions in Dutch and Tarifiyt Berber. In both languages, there was an overall predominance of the subject-initial active. In both languages, there was also an increased probability of patient-prominent structures in the animate patient (IA) condition compared with the inanimate patient condition (II). However, the two languages exhibited complementary profiles in terms of the linguistic forms used to encode patient prominence. Passivisation was used in Dutch even though object topicalisation is possible in the language; in Tarifiyt, object topicalisation was used, and no passives were attested.

The complementary profiles of these two languages reflect two key strategies for encoding patient prominence in sentences, namely subject assignment and linear order. However, it is unclear why different strategies should be preferred in different languages under the same communicative circumstances. In particular, it is difficult to account for this with a universal sentence production model that does not make explicit mention of how language-specific properties interface with general cognitive principles.

A proposal was considered that these differences could fall out from a hierarchical, two-stage model of grammatical encoding, but ultimately this account is lacking in parsimony, and seems weakly motivated from the point of view of linguistic diversity. Another account, that captures language-specific strategies as under a well-known typological generalisation (subject vs. topic-prominence) seems to provide more fruitful avenues for further research, potentially even shedding new light on the question of whether constituent structure building is best conceptualised as occuring in two distinct stages or in one integrated functional-positional process. Ultimately, more comparative studies across language types, such as the study in this chapter, will aid us in updating and refining the sentence production model to better account for both cross-linguistic unity and diversity.

3.6 Questions for the following chapters

The discussion in the previous section raised the possibility that the preference of Tarifiyt speakers to use object topicalisation, and Dutch speakers to use passivisation, could be understood against the backdrop of typological generalisations about subject and topic. In Kiss's description of the insight of Li and Thompson (1976), "the structural role that the grammatical subject plays in the English sentence may be fulfilled by a constituent not restricted with respect to grammatical function or case in other languages" (Kiss, 1995:3). In terms of the current experiment, Tarifiyt may be considered topic-prominent. However, as noted above, Tarifiyt does make use of the grammatical function of subject; indeed, it is not the case that all topic-prominent languages lack a grammatical category of subject. Therefore, even if we consider grammatical and discourse roles to be functionally equivalent, functional processing in languages like Tarifiyt must also involve the assignment of grammatical relations. To elucidate this issue further, a first avenue of investigation would be to re-examine **the relationship between accessibility and grammatical subject in Tarifiyt**. This idea is taken up again in Chapter 6.

A common finding for both Dutch and Tarifiyt Berber here was that, overall, speakers tend to order their sentences agent-patient; however, when the patient is more animate than the agent, we observe an increased tendency to put the patient earlier in the sentence than the agent – namely, in sentence-initial position. However, it is evident that the structural forms used by speakers in Dutch and Tarifiyt Berber to achieve this are not analogous: in particular, the passive in Dutch and the object-topicalised structure in Tarifiyt differ in terms of how grammatical functions are assigned to agent and patient. In addition to this, we saw that although there was a similar finding between the two languages, the data suggested that the strength of the animacy effect might differ between the languages. In the following chapter (Chapter 4) I explore the interplay of these same variables – thematic role, linear order and animacy, along with linguistic differences in grammatical function assignment strategies – from the perspective of the listener. In particular, I explore **how listeners comprehend the kinds of picture descriptions produced in the current experiment**.

In terms of the analysis, I followed approaches that are customary for this type of experiment based on previous literature. These involved classifying the response into a binary variable, in order to analyse the data using mixed binary logistic regression. However, this raised some challenges. In particular, it was necessary to make a subjective decision about how to deal with structures that did not fit in with the distinction of interest. There were a number of responses that could feasibly be considered to represent a different *structure*, namely verb-initial. However, the binary form of analysis did not provide the option to consider the possibility of three-way choice. The decision was therefore taken to triage or classify these responses in terms of their grammatical function assignment or linear ordering profile (active/passive, or agent-patient/patient-agent, respectively).

The point about generalising into two structural categories touches upon a question that has not been fully addressed in sentence production research: how do speakers choose when there are *more than two* felicitous structures, such as in languages like Spanish or Greek (cf. Chapter 1)? Although previous studies have shed light on what prompts speakers to produce a patient-prominent rather than agent-prominent structure, the binary approach to data analysis means that we are still unsure what governs choice between structures that are equally patient-prominent (or agent-prominent, for that matter). In short: **a highly animate patient leads to a higher rate of patient-prominent structures, but how do speakers choose among *several* patient-prominent structures?**

One way to approach this question is to first build on existing studies looking at animacy. We know that patient-initial structures are triggered by certain animacy conditions. We can then start by broadening this to identify how animacy factors into the choice between more than two structures. For example, considering the animate patient condition in the current experiment: if there were several patient-prominent structures, would they all be equally likely? In order to approach this question, we also need to overcome a practical challenge with regard to data analysis. **How can we analyse sentence production data when there are more than two structural categories?** With binary logistic regression, we can only investigate two response categories; if there are more than two, we have to adapt our statistical approach.

These are the questions that underpin the study in Chapter 5, an experiment using the same experimental design as in this chapter, in the Pondok Tinggi language. This language displays a number of different patient-prominent structures, and I therefore aim to explore animacy effects on choice among possible structures. I consider the limitations of applying a binary analysis to understand *simply describing* data, and opt to analyse the Pondok Tinggi data using multi-categorical logistic modelling. An additional practical value of this study is that it is again conducted in an understudied language, under conditions that may be increasingly encountered by psycholinguists wishing to expand the diversity of languages under study.

In the current chapter, we saw that in order to prepare the data for analysis, it was necessary to exclude various responses. In doing this, I also followed the example of previous studies (with the caveat that not all *simply describing* studies use the same response exclusion criteria). The proportion of excluded responses was 39% of the Tarifiyt data and 34% of the Dutch data. This seems to be very high, even though it is in fact comparable to previous sentence production studies (cf. Section 1.2.3). The breakdown of exclusion reasons was provided in Section 3.3.1. Notably, the kind of responses that were excluded were not limited to the cases where participants misinterpreted the task or the image: it also concerned responses which were natural and felicitous, but which did not fit in with the analysis (cf. Section 3.3.1). This issue relates to the problem of so-called ‘exuberant’ responding, introduced in Section 1.2.3. The hypotheses were focused on the relative realisation of agent and patient, and thus the aim was to home in on exactly how agent and patient are produced in a single sentence form. Yet, generally speaking, participants are not always inclined to respond in a single, self-contained transitive clause. For example, instead of producing the desired form *an old man has been hit by a bike*, participants may produce something more like ‘there is an old man; he has been hit by a bike’. Not only does this response contain several clauses (typically a reason for exclusion), but the clause that contains the transitive description has one of the referents pronominalised (another typical reason for exclusion).

The more exuberantly participants respond, the more likely we are to have responses that are problematic for our data analysis. Added to this is the likelihood that this problem will increase, as we work with a wider range of communities. Typically, exuberant responding is mitigated through increased experimental control, such as by training the participants to use desired response forms, or to avoid certain kinds of linguistic forms such as pronouns (cf. Section 2.2.5). For example, in the current experiment, participants were explicitly instructed to use one single sentence, and were encouraged to restrict their responses if they responded too exuberantly during the practice trials. However, as discussed in Section 2.2.3, this approach depends on participants applying a certain level of (meta)linguistic knowledge and adapting their behaviour to comply with seemingly arbitrary instructions.

The less possibility we have to effectively reduce response exuberance through instruction or training, the more likely we are to encounter higher rates of data loss. In other words, if we need participants to provide a certain form of response, but we are unable (or unwilling) to guide them towards that kind of response, we will need to instead remove the problematic responses after the data is collected. This raises another problem with regard to working on understudied languages in the field setting: losing data at the rate of 35–40% severely reduces the yield of the research. With well-studied languages that are spoken in the vicinity of the lab, participants may be fairly easy to come by: we can test more participants with more ease, so data loss is less problematic. However, with understudied languages in less accessible settings, where more investment is required to collect the same amount of data, obtaining the highest yield from that data is paramount.

In summary, exuberant responding is more likely as we work with a more diverse range of languages. At the same time, the techniques that we normally use to mitigate exuberant responding are less likely to be appropriate when we work with a more diverse range of speaker communities. This conspires to increase the overall rate of data loss as we research a wider range of languages. However, when the languages we wish to study require more investment of time and resources to research, we also have less tolerance for data loss. Therefore, we may ask, **is it possible to find better ways to *engage* with response exuberance, rather than needing to mitigate it through metalinguistic instruction or data exclusion?**

This question is taken forward in Chapter 6, where I reflect on why the standard approach to *simply describing* experiments (laid out in the current chapter, and exemplified further in Chapter 5) makes it difficult to address these issues. In that chapter, I then explore an alternative perspective on processing and analysing data. My primary aim in doing so is to investigate variables of sentence production (such as grammatical function assignment and linear order) in Tarifiyt Berber, while allowing participants to respond exuberantly. On top of this, the alternative approach adopted in that chapter gives rise to some additional insights into Tarifiyt sentence production, beyond what was possible using the approach in the current chapter.

CHAPTER 4

The hearer's perspective on transitive picture descriptions: a mousetracking study in Tarifiyt Berber and Dutch

4.1 Introduction

One interpretation of the findings from the previous chapter was that speakers of different languages may have an increased tendency to put more animate arguments in earlier sentence positions – even if the structural forms required to achieve this are not the same. In particular, Chapter 3 provided evidence that the choice of initial argument in a transitive sentence in Tarifiyt and Dutch is affected by animacy; nonetheless, while in Tarifiyt this was achieved with an object-initial active structure, in Dutch it was achieved through passivisation.

In this chapter, I examine the role of sentence-initial arguments in transitive sentences from the perspective of the hearer, in Tarifiyt and Dutch. In particular, I look at how linear order interplays with animacy, as hearers comprehend transitive sentences; moreover, I study this in both languages in order to be able to draw cross-linguistic comparisons. To investigate this, I build directly on the production experiment in the previous chapter, presenting participants with the same linguistic forms that were observed in the production data. In designing the study, I use the same kind of image stimuli, which likewise helps to create the same contexts regarding animacy constraints. Essentially, this study presents an investigation of hearers' behaviour as they comprehend the types of sentences that were produced by participants in the experiment in Chapter 3. This also provides the possibility of relating findings regarding the perception of these sentences to findings about their production, from the previous chapter.

The comparative design of this study capitalises on key similarities and differences between these two languages. Both in Tarifiyt and in Dutch, sentence-initial arguments are not marked for grammatical or semantic role. This means that after hearing the initial argument, there is momentary ambiguity for the hearer about whether the sentence will turn out to be agent-initial or patient-initial. However, the two languages differ in the way that grammatical function and linear order are deployed to encode prominence. Overall, this allows us to zoom in on how animacy and linear ordering information interplay in prediction of thematic structure, and moreover, the chance to contrast this across two languages which differ in terms of available grammatical forms. In the following section I provide an overview of the predictive nature of sentence comprehension, with attention to the role of animacy and linear order in interpreting transitive sentences. Subsequently I proceed to describe the research questions, design and hypotheses for the current study in Tarifiyt Berber and Dutch.

4.1.1 Comprehension as an incremental, predictive process

Sentences unfold over time. There is a range of evidence for incremental processes in both production and comprehension. Not only are speakers able to begin articulating their sentences before planning them in their entirety, but there is a wealth of evidence that hearers also make use of information as it becomes available to help them converge on the interpretation as soon as possible, rather than waiting until the entire sentence has been uttered. For example, this predictive process of interpretation has been the subject of much eye-tracking research, with findings indicating the ability of hearers to use both linguistic and non-linguistic information in an incremental fashion to predict upcoming referents and resolve ambiguities (Altmann & Kamide, 1999; Knoeferle, Crocker, Scheepers, & Pickering, 2005; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Predicting upcoming structure and content based on partial information or non-deterministic cues, however, means running the risk of getting it wrong. Nonetheless, it should not be surprising that prediction is widespread at various levels of representation, since reasonably accurate prediction is likely to be a more efficient strategy overall than not engaging in any predictive processing (Kuperberg & Jaeger, 2016). Indeed, there is an array of evidence pointing towards prediction as a mainstay of language comprehension processes (Kuperberg & Jaeger, 2016).

One of the most pressing tasks in comprehending a sentence is understanding how the referents in the sentence relate to each other. In a sentence where the referents MAN and DOG and the action BITE are mentioned, how do we as hearers know if the dog biting the man, or the man is biting the dog? In other words, hearers must comprehend the thematic structure of the sentence. The question is, is it possible to comprehend thematic structure incrementally as the sentence unfolds? In some languages, where assignments are fully determined by linguistic factors (such as case marking), it seems straightforward. However, in many languages it is not possible to rely on such linguistic indicators: then, ambiguity remains as the sentence unfolds up to a disambiguation point. English provides such an example: on encountering a sentence beginning ‘The man...’, we do not yet know whether the man is the actor (‘The man is running’, ‘The man is chasing the dog’) or the undergoer (‘The man is being chased by the dog’).

Nonetheless, it seems that even in the case of local ambiguities, hearers do engage in predictive, provisional interpretation, by assigning roles to arguments even when there are alternative interpretations possible. Evidence for this across a range of typologically diverse languages comes from work by Bornkessel-Schlesewsky, Schlesewsky and colleagues, primarily using event-related potential (ERP) technique (for a review see Bornkessel-Schlesewsky & Schlesewsky, 2009). Key findings involve the identification of ERP signatures associated with the mismatch when hearers’ expectations about thematic role assignment are not met, particularly in the form of increased negativity in centro-parietal areas approximately 400 ms after critical word onset (the N400; Bornkessel-Schlesewsky & Schlesewsky, 2014).

In recent times, sentence comprehension theory has moved towards the consensus that rather than integrating sources of information in a staged, modular fashion, the parser can and does draw on a range of linguistic and non-linguistic information, including contextual information, right from the earliest stages of comprehending a sentence (van Gompel & Pickering, 2007). The question is rather how these different sources of information are integrated by the hearer in the process of incremental interpretation. Unsurprisingly, the variables that have been shown to be pivotal in explaining the way that speakers formulate sentences, such as animacy, ordering and subject assignment, are also some of the key sources of information for the hearer in the process of comprehension. In the next section, I look more closely at each of these dimensions from the point of view of cross-linguistic research on thematic interpretation.

Predicting based on linear order and animacy information

In the incremental interpretation of sentences, hearers must use information in the order that it becomes available. Moreover, there is broad evidence that linear order itself can form a cue to thematic interpretation (F. Ferreira, 2003; Weyerts, Penke, Münte, Heinze, & Clahsen, 2002). In particular, it appears that speakers across a wide range of languages are predisposed to assign the agent role to the first argument they encounter in a sentence. In drawing together findings from typologically diverse languages, Bornkessel-Schlesewsky, Schlesewsky and colleagues propose that hearers adopt a strategy of identifying the actor (agent) in a sentence as quickly and as unambiguously as possible. A corollary of this strategy is the finding that in many languages, "the processing system prefers actor-initial orders" (Bornkessel-Schlesewsky and Schlesewsky 2014; see also Bornkessel-Schlesewsky and Schlesewsky 2013).

Another principle that has been shown to guide thematic interpretation across different languages is animacy (Paczynski & Kuperberg, 2011). This of course echoes the robust impact already identified for sentence production (as discussed in Chapter 1 and observed for Dutch and Tarifiyt Berber in Chapter 3). It seems that animate entities are considered more typical agents than inanimates; the result of this for thematic interpretation is that hearers expect the agent role to be assigned to an animate referent, as demonstrated by a 'surprisal' effect when their expectations are not met (Bornkessel-Schlesewsky & Schlesewsky, 2014).

The different sources of information that guide hearer interpretations can be captured under the umbrella of cues (E. Bates, McNew, MacWhinney, Devescovi, & Smith, 1982; MacWhinney, 1987), or prominence scales (Bornkessel-Schlesewsky & Schlesewsky, 2009; Bornkessel-Schlesewsky & Schlesewsky, 2014). In the Competition Model of Bates, MacWhinney and colleagues (which also formed an influence on the development of *extended Argument Dependency Model* of Bornkessel-Schlesewsky and Schlesewsky), features such as animacy and word order are viewed as 'cues' to the interpretation of sentences. For example, sentence initial position is seen as a 'cue' for agent interpretation in English (MacWhinney, 2001:71). The extended Argument Dependency Model (eADM) frames such cues as scalar, under the notion of 'prominence scales' (Bornkessel-Schlesewsky & Schlesewsky, 2015). For example, animacy is taken as a scale from more to less animate, and linear order is viewed as a scale from earlier to later in the sentence. The idea is that referents nearer the top of the scales (more animate, earlier position) are more ideal actors, and referents near the bottom of the scales (less animate, later position) are more ideal undergoers (or patients) (for a comparison of the eADM and Competition Model, see Bornkessel-Schlesewsky and Schlesewsky 2014:124-5).

The question that arises is how different sources of information (relating to different types of cues, or different prominence scales) interact with each other. In terms of animacy and linear ordering, we can identify two key possibilities. On the one hand, we have seen that hearers consider both animacy and initial position as cues to preferentially assign the agent role. That is, when an argument is both sentence initial *and* highly animate, the strength of the cues will be combined in an additive manner. Under this view, if an initial argument is animate, hearers will be *even more* likely to assign it the agent role than if it were inanimate. On the other hand, when we consider the evidence from sentence production, we may come to a different hypothesis regarding cue integration. As both discussed and demonstrated in the previous chapter, sentence production research has found that speakers have a strong tendency to give more animate arguments priority in linear order even when they are *not* agents. It has been shown in a variety of languages that the sentence-initial position is more likely to be occupied by the patient argument, if it is more animate than the agent (Prat-Sala & Branigan, 2000; Tanaka et al., 2011; van Nice & Dietrich, 2003). Under this view, animacy and linear order interact, rather than combining additively.

From a cross-linguistic perspective, things become more complex still. In particular, there is cross-linguistic variation in the way that agentivity, animacy and linear order integrate with the assignment of grammatical functions. On the one hand, animacy and agentivity are closely correlated with subject function cross-linguistically (Comrie, 1989). On the other hand, the relationship of linear order to subjecthood varies across languages. As seen in the previous chapter, both Tarifiyt Berber and Dutch speakers showed a predominant preference for subject-initial orders in their transitive picture descriptions; also in both languages, animacy influenced linear order, by increasing the likelihood that a more animate patient would occupy sentence-initial position. However, the languages differed in the grammatical form used to accommodate a non-agentive argument in sentence-initial position. In Dutch, a sentence-initial patient was always assigned subject function, through passivisation. However, in Tarifiyt, the sentence-initial patient was never assigned subject function; indeed, no passive structure was found among the Tarifiyt responses.

In summary, hearers use the available information to make provisional assignments of thematic structure of the sentence. In terms of linear order and animacy, research has shown that hearers expect initial arguments to be agents, and that they also expect agents to be animate. However, what is less clear is how different cues (both linguistic and non-linguistic) are integrated to arrive at such predictions. Moreover, little is known about how these factors play out cross-linguistically, especially with regard to cross-linguistic differences in grammatical function assignment.

In this study, I investigate how animacy and linear order affect thematic predictions for transitive sentences. Moreover, I investigate this in both Dutch and Tarifiyt Berber, which show complementary profiles with regard to how initial arguments are accommodated in grammatical sentence forms. In order to recruit and test speakers of Tarifiyt Berber, I conduct this study within the speaker community in Northern Morocco. Therefore, in designing and conducting this experiment, it is necessary to consider the issues raised in Chapter 2, with regard to field-based psycholinguistic research. In view of these various issues, I opt to use mousetracking technique, with an experimental design implemented in the Mousetracker software package (Freeman & Ambady, 2010). In the following section, I provide a brief introduction to mousetracking technique, along with consideration of why this technique is advantageous as we aim to study a wider range of languages – particularly given the range of issues raised in Section 2.2.

4.1.2 Mousetracking technique

Mousetracking records the *x,y* coordinates of the mouse cursor as participants process and respond to stimuli using the mouse. A participant is faced with two or more response choices, typically located at the top of the computer screen. To begin each trial, the participant clicks a button at the bottom of the screen. On hearing or seeing the stimulus, they must move the mouse from the bottom of the screen to click on the correct response option. The dependent measure is the journey that the mouse makes from the starting point to the response location. Crucially, the more the alternative response options compete with the correct answer, the more likely it is that the participant's mouse movement will deviate towards other options en route to selecting the correct response. This technique provides a rich variety of measures: the deviation of the trajectory towards the competing response option, as well as the velocity, acceleration and angle of the movements. This means that we can gain insight into stimulus processing and decision making as it unfolds.

The primary concern with the journey that the mouse makes – the 'trajectory' – has given rise to an additional debate about what different types of trajectories might tell us about the nature of the decision making process (Spivey, Dale, Knoblich, & Grosjean, 2010; van der Wel, Eder, Mitchel, Walsh, & Rosenbaum, 2009). Previous literature has drawn attention to the existence of two general trajectory shapes, as illustrated in Figure 4.1. On the one hand, a trajectory may show a smooth shape, increasing in overall curvature towards the competing response depending on the amount of attraction. On the other hand, a trajectory that is attracted to the competitor might instead deviate sharply towards the incorrect response location before redirecting towards the correct response (for further discussion and illustration, see Hehman, Stolier, & Freeman, 2015). In terms of psychological processes, a trajectory that gently curves towards the competitor might reflect parallel consideration of competing alternatives, whereas the sharply deviating type could indicate selection of one alternative followed by abrupt revision.

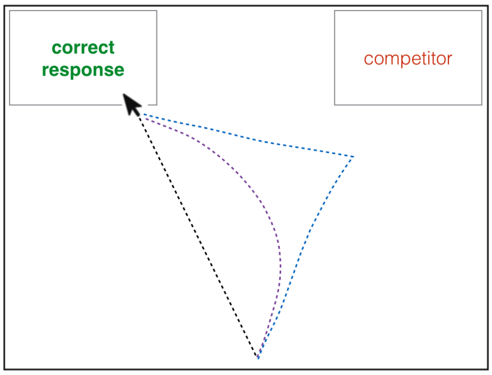


Figure 4.1: Illustration of a Mousetracker trial layout and the three types of trajectory to the response. The black dashed line shows an idealised straight trajectory. When the mouse movement is attracted to the competitor, the resulting trajectory may be gently curving (purple line) or sharply deflecting (blue line).

Although the exact psychological implications of different trajectory shapes is open to debate, it is crucial to examine properties of the underlying trajectories, rather than simply looking at the averaged responses. This is because these important trajectory shapes may be lost in the process of aggregation: differently-shaped trajectories will be blended into a composite shape when the data points are averaged together. The result is that an aggregate trajectory that looks like a gentle curve may actually be a composite of direct trajectories plus sharply deviating trajectories.

Implementing mousetracking studies has become more practical for researchers thanks to the publication of Mousetracker software by Freeman and Ambady (2010). This software facilitates data collection and pre-processing using a trio of software tools. This, along with recent methodological articles (Freeman & Dale, 2013; Hehman et al., 2015), helps to standardise the implementation of the technique across studies, which is an important consideration for emergent data collection techniques. All procedures in this study were implemented using the Mousetracker software.

Moreover, mousetracking lends itself well to data collection outside the laboratory. As discussed in Chapter 2, this kind of approach is crucial if we are to gain data on a wide array of linguistic types. From a logistical point of view, the only apparatus required for this setup is a laptop, a mouse and a pair of headphones. To deal with the practical issues that may arise with finding an appropriate quiet testing environment, I opt to use headphones with active noise-cancelling functionality. Another issue raised in Chapter 2 related to the use of written instructions and stimuli. In this regard, Mousetracker software permits the use of audio and image stimuli, making it appropriate for experiments with predominantly oral languages. Although the software is designed for instructions and on-screen buttons to contain text, it is possible to adapt this to present the instructions to participants through video and audio clip playback within the software. In terms of the paradigm itself, the 'listen-and-click' picture-matching paradigm as used in this study is a viable task for a wide range of populations; note, however, that it is of course required that participants are familiar with computers (I return to this point in Section 4.6). The issue regarding scarcity of cross-cultural stimuli and norming data (also discussed in Chapter 2) is resolved by (a) basing the stimuli for this experiment on the materials and outcomes of the study in the previous chapter, and (b) by developing the stimuli to be appropriate and accessible for both experimental contexts (Dutch and Moroccan), based on experiences in previous research with these groups (and supported by piloting the experiment).

Lastly, another advantage of mouse tracking technique is the high data yield, which is important for field based studies (for example, when travelling specially to conduct experiments in another country, with a certain budget and timeframe, and bearing in mind the potential for changes in circumstances and schedules; cf. Sections 2.2.2 and 3.6). Despite the apparent simplicity of the apparatus and the task, mousetracking technique yields rich quantitative data, including a wide range of dependent measures relating to the trajectory and timing of the response. To some extent, the same could be said of eye-tracking, especially with the advent of slim, portable setups with swift calibration procedures. However, note that the two techniques yield quite different measures and should be viewed as complementary. Eye movements tend to be discrete and ballistic: when items on screen compete for response selection, the eyes move abruptly between the different items. Meanwhile, mouse trajectories are continuous: attraction to competing items is manifested in curvature or deviation of the path taken from the starting point to the response (Freeman & Ambady, 2010). Consequently, eye-tracking studies are typically concerned with the sequence, proportion and duration of fixations on different points on the screen, while mousetracking is concerned with the nature of journey taken *between* points on the screen.

4.1.3 The current study

This study has two broad goals. In the first place, I aim to pinpoint the effects of predictive thematic role assignment in transitive sentences using mousetracking technique. More importantly, however, the key aim of this study is to investigate how the interaction of animacy and ordering expectations affects hearers' predictions of thematic structure. Crucially, I investigate these effects in two typologically distinct languages – which present complementary profiles when it comes to how prominence is realised in grammatical structures.

In each trial, I present participants with a pair of pictures depicting two referents, for example a dog and a cat, engaged in a reversible transitive action. By 'reversible' I mean that either participant in the scene could plausibly be the agent or the patient of the action described by the verb: with a cat and a dog, an example of a reversible action could be that one scratches the other. The only difference between the two pictures in a pair is which referent is playing which role in the action: i.e. in one picture, the dog is scratching the cat, but in the other picture, the cat is scratching the dog. The two pictures appear at the top left and top right corners of the screen, and participants start the trial by clicking a button at the bottom centre of the screen. On hearing a sentence that describes one of the two pictures, participants must move from the bottom of the screen to click on the correct picture. I record the mouse movement between these two points, and compare trajectories between different experimental conditions. Figure 4.2 is provided to give an impression of what participants see during a trial and how a mouse trajectory is achieved; the experimental design is explained in more depth in Section 4.2.

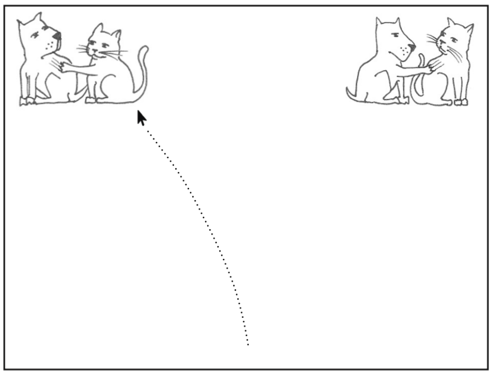


Figure 4.2: General impression of the mouse tracking trial layout in the current study.

In order to investigate the predictions of hearers regarding information structure, I present pictures with both agent-prominent stimulus sentences (*The cat is scratching the dog*, *The dog is scratching the cat*) and patient-prominent stimulus sentences (*The cat is being scratched by the dog*, *The dog is being scratched by the cat*). Figures 4.3 and 4.4 give examples of the agent-prominent and patient-prominent forms in the two languages, which were introduced in sections 2.3.1 and (6). In order to investigate the role of relative animacy in hearer predictions, half of the trials involve scenes where both referents are equally animate ("equal-animacy" trials) as above, but the other half involve scenes where the two referents differ in animacy ("mixed-animacy" trials): for example, in one picture a farmer is kicking a donkey, but in the other picture, the donkey is kicking the farmer.

Predictions

I expect that hearers will use information as it becomes available to make provisional decisions about argument structure, and that this will be evidenced in their mouse movements: in particular, I expect that their mouse movements will be attracted to whichever picture depicts the argument structure they predict. To the extent that the hearers consider the incorrect picture as a possible correct response, their mouse movements should show an increased attraction or deviation to that picture – thus allowing us to probe the online process of thematic role identification.

The primary question then is *how* the hearers' expectations about ordering and animacy interact in the process of assigning provisional argument structure. To explore the possible outcomes here, let us first consider just the effects of **linear order** expectations. This concerns how the hearers will respond to the way semantic role (agent, patient) is assigned to sentence position. I expect that hearers will demonstrate an overall provisional preference for agent-initial structures; in other words, they will expect that the first argument they hear is the agent. If this is the case, the trajectories should show greater deviation towards the competitor when the stimulus sentence turns out to be patient-initial. To illustrate this, recall the example above involving the dog and cat. If the first-heard noun phrase is *a dog*, hearers should be attracted to select the response picture where the dog is the agent, i.e. the picture of the dog scratching the cat. When the sentence is indeed agent-initial, the path to the correct response should thus be fairly direct. However, when this first noun phrase turns out to be the patient (i.e. the dog is *being scratched by* the cat), we should see some deviation towards the incorrect picture even when participants ultimately make the correct response.

Let us now consider the role of **animacy**. Among the mixed-animacy trials, there are two subtypes of trials according to how the animacy balance and the thematic role assignment combine: either the patient of the sentence is more animate than the agent (Patient>Agent, e.g. a donkey kicking a farmer), or the patient of the sentence is less animate than the agent (Patient<Agent, e.g. a farmer kicking a donkey). The goal here is to investigate whether trajectory differences between agent-initial vs. patient-initial stimulus sentences are affected by hearers' knowledge of the relative animacy of patient and agent. In other words, I aim to test whether there is an interaction between sentence type (agent-initial vs. patient-initial) and animacy balance (equal, Patient<Agent, Patient>Agent). So, the question is: what argument structure do hearers predict when the first noun phrase they hear is either the less animate, or the more animate, of two referents?

One possibility is of course that hearers are not sensitive to the animacy differences. In this case, the pattern of trajectories would be the same in all three animacy conditions. However, given the extent of findings indicating that animacy information does influence the predictions that hearers make, there is a strong possibility that we will see different patterns of data between animacy conditions. There are then two possibilities for how they might use this information (as outlined in Section 4.1.1).

On the one hand, we can see animacy as an independent indicator of agenthood, which predisposes hearers to assign the agent role to the more animate referent. With this in mind, we may expect that when the first-heard noun phrase is the more animate of the two referents, hearers would be highly likely to consider this referent as agent. When this is then combined with a preference for agent-initial structures, we should see an even greater attraction to the incorrect response for patient-initial trials when the patient is more animate than the agent (P>A) than when animacy is equal. To illustrate, take the scenario involving a farmer and a donkey: when the the first-heard noun phrase is *a farmer*, a combined preference for agent-initial structures and for human agents should make hearers strongly attracted to the picture where the farmer is agent (i.e. a farmer is kicking a donkey); if the sentence turns out to be patient-initial, as in *a farmer is being kicked by a donkey*, this effect would then manifest itself in greater deviation to the incorrect picture, where the farmer is the one doing the kicking.

On the other hand, production research has demonstrated that mixed-animacy scenarios with an inanimate agent and an animate patient are associated with increased passivisation or object-topicalisation. This was demonstrated in the speakers' responses in Chapter 3. In both Dutch and Tarifiyt Berber, speakers describing mixed-animacy scenes were more likely to use patient-initial structures when the patient was more animate than the agent. From the hearer's perspective, this would translate to an increased probability, when the first argument heard is the more animate of two referents, that a passive or object-topicalised structure is about to unfold. In other words, hearers may be implicitly aware that in mixed-animacy scenarios, there is a significant likelihood that if a human referent is encountered at the beginning of a sentence then it is the one undergoing the action. Under this account of hearer behaviour, we could expect the data to display the opposite pattern to the one described above. That is, on encountering an initial human argument in mixed-animacy trials, instead of trajectories deviating *more* towards the competing picture, trajectories may in fact deviate *less*, when compared to the equal-animacy condition.

On top of this, there is the possibility for **between-language** differences. Passivisation (in Dutch) and object-topicalisation (in Tarifiyt) are both associated with highly animate patients. However, these two structures differ in terms of their linguistic form. The alternation between the active and passive structure in Dutch involves differences in how the agent and the patient are mapped to grammatical functions, but it does not change linear order. By contrast, the alternation between subject-initial and object-initial structures in Tarifiyt involves differences in how agent and patient roles are mapped to linear order, but it does not change grammatical function assignment. These cross-linguistic differences in the relationship between thematic roles, linear order and grammatical function assignment could possibly manifest in cross-linguistic differences in hearer behaviour based on an initial animate argument.

To summarise, it is expected that there will be deviation of trajectories towards the incorrect response whenever hearers' provisional decisions about argument structure do not match the actual argument structure of the stimulus sentence. It is probable that hearers in both languages will be predisposed to interpreting the initial argument as the agent. Based on this, the main goal of this study is to then examine how hearer expectations based on the sentence-initial argument are affected by knowledge about the relative animacy between agent and patient. I have described two possible hypotheses for how these animacy effects may be manifested. On the one hand, it is possible that high animacy and firstness combine as cues for agency, leading to increased expectation of agent-patient order. Alternatively, it is possible that hearers relate high animacy in the sentence initial position is to an increased likelihood of patient-agent order, potentially reducing their expectation of agent-initial order. Another way to think of the difference between these two accounts is whether hearers simply add the probability of an animate referent being agent to the probability of an initial referent being agent, or whether they are able to consider the overall probability of agenthood when the available linguistic and non-linguistic information is combined. In addition to this, between-language differences in how these variables relate to each other, and to grammatical function assignment, may also impact the patterns we see in the results.

4.2 Method

4.2.1 Stimuli and design

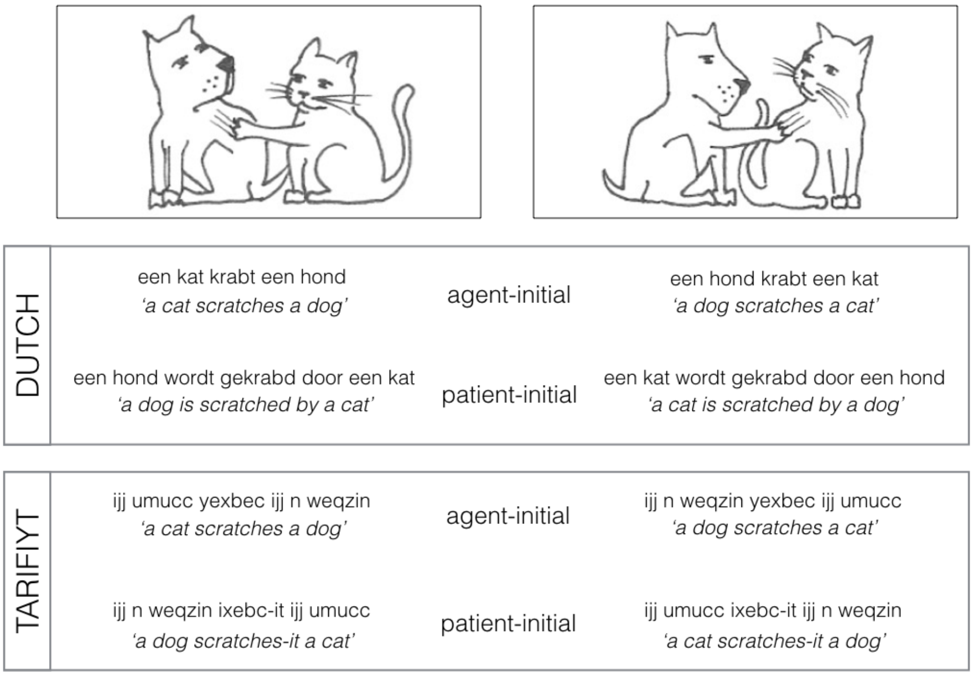
The critical stimuli were 40 picture-pairs, depicting reversible transitive actions between two referents. Pictures were black line-drawn images presented on an off-white background. Each picture was associated with two sentences: one agent-initial description and one patient-initial description (see Figures 4.3 and 4.4). Care was taken to ensure the two pictures within a pair differed only in terms of the roles played by the referents, and not in terms of referents themselves. For example, in the case of a cat scratching a dog and a dog scratching a cat, it is the same dog and cat character in both pictures. The task for the participant was to view the two pictures, hear a sentence and decide which of the two pictures was being described.

As discussed above, one of the goals of this study was to investigate how trajectories would differ under conditions of mixed animacy compared to equal animacy of the two referents in the pictures. Twenty of the picture-pairs depicted equal-animacy interactions (two humans, two animals or two inanimates) and 20 of the picture-pairs depicted mixed-animacy interactions (animal–human or inanimate–human).

Sentence structure was an independent variable with two levels: agent-initial vs. patient-initial. Since each picture was associated with two sentences (one agent-initial description and one patient-initial description), for each picture-pair there were four possible sentences that a participant could hear. The example stimuli in Figure 4.3 and Figure 4.4 provide an example of how sentence structure and thematic structure were interacted in this experiment. Each figure contains a pair of pictures. The two stimulus sentences for each picture are given underneath – an agent-initial and a patient-initial description of each picture. Taken together, the four sentences represent the four possible combinations of argument ordering and thematic role assignment for the two referents featuring in the picture-pair. Each set of sentences is provided in both Tarifiyt and in Dutch.

The experimental design was counterbalanced so that each participant only saw each picture-pair once. Since there were four possible sentences per pair, this meant that there were four counterbalanced lists of 40 trials each. The two sentence structures were also equally balanced across the four lists. As illustrated by Figures 4.3 and 4.4, both pictures in a pair had the same arrangement of the two referents (left/right); only the direction of the transitive action was different. The horizontal flip of the pictures, however, was randomised so that for any picture-pair, a given referent was on the left of both pictures for some participants and the right of both pictures for other participants. The location of the correct picture (top left or top right of the screen) was also randomly determined.

Sentence stimuli in both Dutch and Tarifiyt were checked and recorded by a bilingual speaker. The same bilingual speaker also recorded spoken instructions for both languages (i.e. the video clip described below in Section 4.2.3). This was done to control for any confounding effects that could arise from using different voices in the different experiments. This was the same person who recorded the instructions and prompt questions for the production experiment in Chapter 3. The implementation of the experiment in both languages was piloted with native-speaker participants before testing began.



DUTCH	een kat krabt een hond *'a cat scratches a dog'*	agent-initial	een hond krabt een kat *'a dog scratches a cat'*
	een hond wordt gekrabd door een kat *'a dog is scratched by a cat'*	patient-initial	een kat wordt gekrabd door een hond *'a cat is scratched by a dog'*
TARIFIYT	ijj umucc yexbec ijj n weqzin *'a cat scratches a dog'*	agent-initial	ijj n weqzin yexbec ijj umucc *'a dog scratches a cat'*
	ijj n weqzin ixebc-it ijj umucc *'a dog scratches-it a cat'*	patient-initial	ijj umucc ixebc-it ijj n weqzin *'a cat scratches-it a dog'*

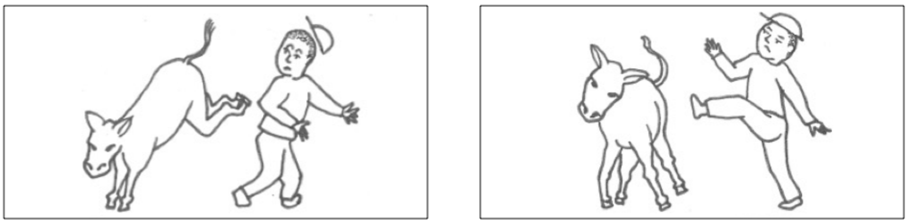
Figure 4.3: Example of an equal-animacy stimulus set

4.2.2 Participants

Fifty-one students participated in the Tarifiyt Berber study (17 males; age range 17-28 years, M^{age} = 20). All were native speakers of Tarifiyt Berber who had grown up in the province of Nador, in north-eastern Morocco; all were right-handed. Seven additional participants who did not fit this description were excluded from the dataset. An additional three who were observed during the session to have misinterpreted the task were also excluded from the data (for example, a participant who vocally repeated the picture descriptions before responding). Participants were recruited and tested at the Faculté Pluridisciplinaire de Nador.

Fifty students participated in the Dutch study (14 males; age range 19-34 years, M^{age} = 22). All were right-handed native speakers of Dutch who had grown up in the Netherlands; four additional participants who did not fit this description are excluded from the dataset. No participants were observed to have misinterpreted the task. Participants were recruited and tested at Leiden University in the Netherlands.

In both Leiden and Nador, testing was conducted in quiet study areas and lasted about ten to fifteen minutes per participant. Participants were rewarded with a small (non-monetary) token of thanks for their time. The research in Nador was kindly facilitated by the invaluable help of Prof. Mostafa Ben-Abbas and native-speaker assistant Hanae Boudihi.



DUTCH	een ezel schopt **een boer** *'a donkey kicks **a farmer**'*	agent-initial	een boer schopt een ezel *'**a farmer** kicks a donkey'*
	een boer wordt geschopt door een ezel *'**a farmer** is kicked by a donkey'*	patient-initial	een ezel wordt geschopt door **een boer** *'a donkey is kicked by **a farmer**'*
TARIFIYT	ijj n weɣyur yarceř **ijj ufedjaḥ** *'a donkey kicks **a farmer**'*	agent-initial	**ijj ufedjaḥ** yarceř ijj n weɣyuř *'**a farmer** kicks a donkey'*
	ijj ufedjaḥ yarcř-it ijj n weɣyuř *'**a farmer** kicks-him a donkey'*	patient-initial	ijj n weɣyur yarcř-it **ijj ufedjaḥ** *'a donkey kicks-it **a farmer**'*

Figure 4.4: Example of a mixed-animacy stimulus set. In each sentence, the more animate referent appears in bold.

4.2.3 Procedure

All data was collected using Mousetracker software (Freeman & Ambady, 2010), running on a Dell Latitude E6330 laptop (Windows 7, screen resolution 1366 by 768 pixels) with a Dell USB optical mouse (model MS111, 1000 dpi). The Windows mouse sensitivity settings were kept as standard, however the mouse speed was set within Mousetracker to 5 (on a scale of 1-20) after piloting. Participants listened to the experimental stimuli through a pair of Sony MDR-ZX110NC on-ear headphones with active noise cancellation.

First, participants were given an information sheet about the study and were asked to indicate their consent, followed by filling in an accompanying background questionnaire. The information sheet and questionnaire for the Berber participants were in Arabic. This was the only part of the recruitment and testing procedure that was not conducted in Tarifiyt.[1]

Participants were seated at the laptop and began by watching a short video clip that provided the experimental instructions. This video clip showed an animation of a trial, and was narrated in the

[1] Despite a recent growth in the written use of Tarifiyt (in both the Latin and Tifinagh scripts), it is still a predominantly oral language with many speakers preferring to use Arabic, French or Spanish when writing, even if they are literate in Tarifiyt; cf. Chapter 2.

target language. The video clip therefore not only instructed participants how to complete the task, it also familiarised them with the visual environment of the experiment. Participants completed a set of eight practice trials in order to further familiarise themselves with the task.

Picture stimuli were displayed on the top left and right corners of the screen, within spaces of 800 by 300 pixels. The procedure of an individual trial is illustrated in Figure 4.5. For each trial, participants were allowed to view the two pictures before clicking on the circle at the bottom of the screen to hear the stimulus sentence, thereby initiating the trial. Their task was simply to click on the correct picture, as soon as possible. If participants did not begin to move the mouse within 1100 ms of clicking the start button, at the end of that trial a warning image would appear on the screen, which encouraged the participant to respond more quickly in future (this warning image and its meaning were explained to participants in the instruction video).[2] Accuracy was also encouraged by negative feedback in the case of an error (a red cross in the centre of the screen).

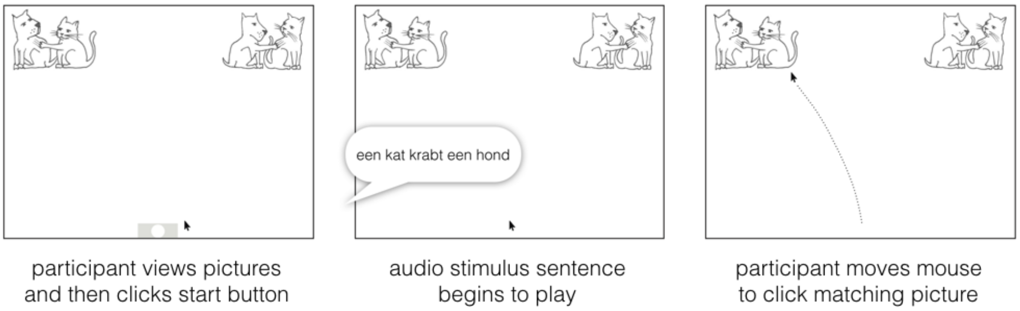

een kat krabt een hond
participant views pictures and then clicks start button
audio stimulus sentence begins to play
participant moves mouse to click matching picture

Figure 4.5: Procedure of an individual trial

4.2.4 Preprocessing and analysis

As described in Section 4.2.2, participants were excluded from the cohort if questionnaire answers indicated that they had not acquired the language of the study in early infancy and if they had not grown up in a community where the target language was the predominant mode of communication, or if they were left-handed (to maximise homogeneity of mouse movements). Participants were also excluded from the cohort if their behaviour during testing indicated that they had misinterpreted the task. Incorrect responses comprised 1% of the data and were excluded from further analysis.

Time taken from initiation click to response click can vary between trials and between participants. Furthermore, different screen sizes and resolutions are used in different studies. In order to be able to compare trajectories within and between studies it is therefore customary in mousetracking research to standardise temporal and spatial parameters (‘normalising’ the trajectories). Therefore, in line with the norms adopted in the mousetracking research literature and implemented in Mousetracker (Freeman & Ambady, 2010), a normalised dataset was exported, with all trajectories resampled to 101 time-points and rescaled to a standard co-ordinate space where the Y axis ranges

[2] 1100 ms was chosen to give participants minimally one second of leeway for beginning their movement without getting a warning, allowing for a margin of error in the timing latency of the software.

from 0 (bottom edge of screen) to 150 (top edge of screen) and the X axis ranges from -100 (left edge of the screen) to 100 (right edge of the screen). Quantitative measures regarding these trajectories were then submitted to linear mixed effects modelling as dependent variables (namely, the area under the curve and the maximum X-axis deviation; see Section 4.3).

The dataset was also exported from Mousetracker with only spatial rescaling, i.e. without resampling of the timepoints. Henceforth I refer to this as ‘raw time’ data following Freeman and Ambady (2010). With the raw time data it is possible to make closer inspection of how trajectories vary in response to a specific moment in the stimulus sentence. Moreover, with raw time data we are able to align the trajectories on a chosen moment in the sentence, even if this is not always the same number of milliseconds after the start of the trial. Here, I align all the trajectories on the offset of the first noun phrase – the point at which hearers have the minimal information that could be integrated into a provisional sentence structure.

Mousetracker raw time data can be exported in time bins of a size specified by the user. I determined that the smallest time bin increment that would still produce a viable dataset (i.e. without missing data points) was 50 ms. Trial data in 50 ms timebins were then exported from the program for each trial. For each audio stimulus sentence, I logged the time-stamp of the offset of the first noun phrase using the CheckFiles tool (Protopapas, 2007). Using this information, I then pinpointed for each trial the time bin that corresponded to this time stamp (for example, a noun phrase offset of 138 ms would correspond to the time bin 100-150 ms). I then added a new variable where the time bin corresponding to the noun phrase offset was labelled zero, subsequent time bins were indexed as 1, 2, 3 and so on, while previous time bins were indexed as -1, -2, -3 and so forth. Having done so, I was then able to align all trajectories on the offset of the first noun phrase.

As explained in section 4.2.1, the design of this experiment manipulates three variables: sentence type (agent or patient initial), thematic role assignment (e.g. cat is agent and dog is patient, or vice-versa) and animacy balance (mixed or equal animacy). In combination, however, the latter two variables are more meaningfully seen as a single factor with three levels (as it is described in Section 4.1.3). On the one hand, mixed-animacy trials present two possible combinations of animacy and thematic role: either the patient is *less* animate than the agent or the patient is *more* animate than the agent (cf. Figure 4.4). On the other hand, in equal-animacy trials there is by definition no such interaction. Nonetheless, for completeness, all possible combinations of each pair of scene referents with sentence types were tested (cf. Figure 4.3). Therefore, the following three sub-conditions were taken forward for analysis: equal animacy (20 trials per participant), patient less animate than agent (P<A; 10 trials per participant) and patient more animate than agent (P>A; 10 trials per participant). These three conditions are henceforth referred to as ‘animacy subconditions’.

AUC, MD and X axis deviation

From the normalised data, it is possible to obtain two quantitative measures indexing the attraction towards the competitor. Firstly, there is the area under the curve (henceforth AUC), which measures the total area between the trajectory curve and an idealised straight path from the start to the target. Secondly, there is the maximum deviation on the X axis (henceforth MD), which measures the point of furthest horizontal deviation of the trajectory from the idealised straight line towards the competitor.

In addition to this, the raw time data can be used to plot the mean X-axis deviation over time. This allows us to see how participants move towards the left or right of the screen as they hear the sentence unfold, and crucially how this differs per condition. I additionally use non-parametric permutation testing (cf. Maris & Oostenveld, 2007) to give further insight into trajectory divergence on the X axis, comparing the trajectories between the two sentence types within each of the three animacy conditions. The aim of this is to shed more light on how the divergence of trajectories between patient-initial and agent-initial sentences varies by animacy subcondition.

Permutation testing is useful for analysis of time-course data. It is used to indicate significant differentiation between conditions over time. For example, it has been used for the analysis of differences in waveform amplitude between conditions in (M)EEG experiments, because it can be used to assess amplitude differences in consecutive time windows and indicate areas of sustained divergence. Furthermore, it does not depend on the assumption that observations are sampled from a normal distribution. A full explanation and discussion can be found in Maris and Oostenveld (2007). In short, the null hypothesis of the test is that observations on both conditions are sampled from the same distribution. Testing this hypothesis works as follows. First, the experimental condition codes are randomly reassigned to the observed values (in effect, randomly swapping the observed values between the conditions). Each time the random swap is done, the difference between the conditions is analysed (in this case using ANOVA, although the choice of test statistic is up to the researcher; see Maris & Oostenveld, 2007). Doing this a large number of times provides us with many test statistics, which together form a frequency distribution. Using this distribution of possible test statistics, it is possible to assess the probability of obtaining the test statistic of the *actual* data under the null hypothesis, and thence a p value. We do this test for each time bin of the data, obtaining a p value for the difference between conditions at every time bin. Taken together, this sequence of p values allows us to assess the time windows relative to the offset of the first NP where trajectories diverge significantly. I repeat this entire procedure for each of the three animacy subconditions, and finally compare the results using a plot, to gain an impression of how the degree of divergence of trajectories relates to animacy of the referents. These plots complement the visualisations of mean trajectories to enable us to reach informed conclusions about the raw time data.[3]

It should be noted that the plots and permutation tests for the raw time data do not include all time bins in the dataset, but rather the maximal part of the dataset for which there is sufficient data to appropriately compare conditions. At the furthest time bins relative to the offset of the first noun phrase, there are fewer observations. This is expected, due to natural variation in response time among participants. In short, the time bins that are not included in the plots and permutation tests are the time bins where the data is too sparse to make inferences about by-condition differences. Since Dutch and Tarifiyt participants differed in the typical time needed to complete the trial, the time-frame of the plots and permutation tests differs per language. Nonetheless, constraining the plots to the time bins with sufficient data permits a direct comparison of mean X deviation profiles with the results of the permutation tests as well as ensuring that visualisations are not unduly influenced by the relative weight of values in areas of data sparsity.

Distributional analysis of MD scores

As mentioned above, recent mousetracking studies have drawn attention to the shape of trajectories and what it can tell us about the participant's decision process. To delve into this point I inspect the distribution of the maximum deviation (MD) scores. This can be done by producing a histogram of the maximum deviation scores. An abrupt revision will typically lead to a horizontal spike in the

[3] Permutation tests and associated plots were executed with the help of R code provided by Cesko Voeten.

trajectory, as the participant abruptly switches direction. Previous studies have used a rule of thumb that MD scores above 0.9 may be taken as symptomatic of abrupt revision (Barca & Pezzulo, 2015; Freeman, 2014). By contrast, trajectories that head straight toward the target have low maximum deviation scores. If the deviating responses in this experiment are characterised by abrupt revision, there should be a peak of trajectories with large maximum deviation scores (where participants changed their minds) accompanied by a peak of trajectories with scores near zero (where there was no revision needed). As a result, such a dataset should show a bimodal distribution of scores. Bimodality can be established by inspecting histograms, and additionally by estimating a coefficient of bi- or multimodality (Freeman & Dale, 2013). If there are effects arising when hearers encounter a mismatch between their expectations and the actual thematic structure (consistent with the work of Bornkessel-Schlesewsky, Schlesewsky and colleagues), then I hypothesise that this will be reflected in abrupt revision of trajectory direction, consistent with the idea that a provisional assignment of thematic structure made and subsequently revised when it becomes apparent that it is incorrect. Accordingly, I expect to see a bimodal distribution of MD scores, as described above (see Dale and Duran (2011) for another mousetracking study hypothesising bimodal distributions as a result of abrupt revision of sentence interpretation). I will assess bimodality using histograms and by calculating Hartigan's Dip (Hartigan & Hartigan, 1985). Hartigan's Dip was chosen following the recommendations of Freeman and Dale (2013) who found this to be the most robust coefficient as regards the assessment of bimodality in mousetracking data. To be precise, Hartigan's Dip indicates whether there is multimodality in the distribution of the data, with the null hypothesis being that data is unimodal.

4.3 Results

In what follows, I first present measures of deviation based on time-normalised data, namely maximum deviation (MD) and the area under the curve (AUC), before inspecting X axis deviation as it unfolds in raw time, relative to the offset of the first noun phrase. Accompanying this is an examination of the distribution of maximum deviation scores (by experimental condition). All visualisations and analyses were conducted in R unless stated otherwise (R Core Team, 2017).

4.3.1 Measures of deviation: normalised data

Tarifiyt Berber: AUC and MD

Figure 4.6 shows the mean trajectories collected in the Tarifiyt study, distinguished by sentence structure (two levels) and animacy subcondition (three levels).[4] Line style indicates sentence type (agent-initial is solid, patient-initial is dashed) and colour indicates the animacy balance of the two entities depicted (equal animacy is red; for mixed animacy, patient less animate than agent is green and patient more animate than agent is blue).

This plot reveals that across all three animacy subconditions, mean trajectories deviate more towards the competing picture on patient-initial than agent-initial trials. The difference between the patient-initial and agent-initial trial trajectories appears fairly similar across the three animacy conditions.

[4] Plots throughout this thesis were created using the *ggplot2* package in R (Wickham, 2009), unless indicated otherwise.

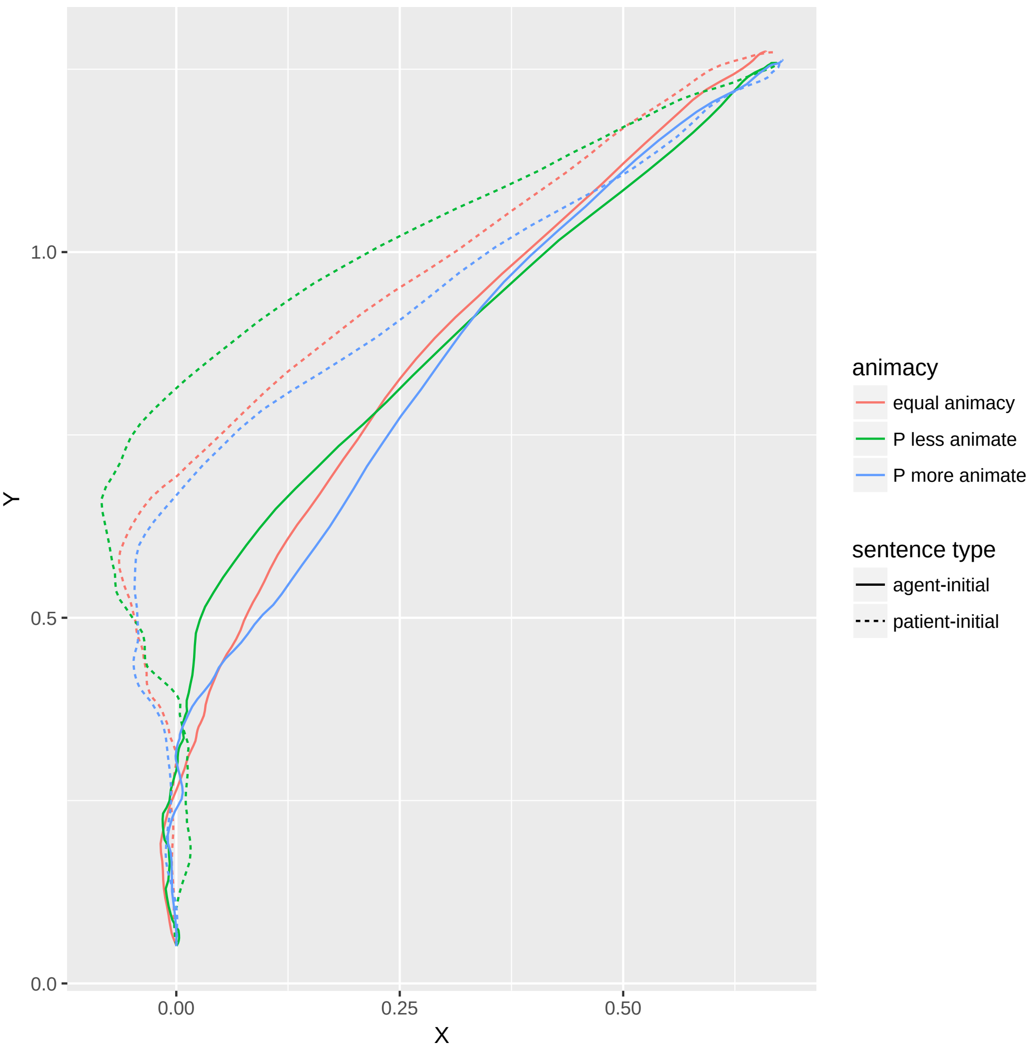


Figure 4.6: Mean trajectories, normalised data, **Tarifiyt Berber**.

As described in section 4.2.4, mousetracking provides us with two quantitative measures of trajectory deviation: area under the curve (AUC) and maximum deviation on the x-axis (MD). I assessed the effects of sentence type and animacy balance on AUC and MD, and possible interactions, with linear mixed effects modelling using the *lme4* package (D. Bates et al., 2015). Models were built and assessed for AUC and MD separately.

	intercept-only	full model
(Intercept)	0.56	0.62
	(0.09)	(0.10)
P-initial sentence		0.54^{***}
		(0.09)
P<A picture		−0.02
		(0.11)
P>A picture		−0.02
		(0.11)
P-initial sentence:P<A picture		0.10
		(0.14)
P-initial sentence:P>A picture		−0.14
		(0.15)
AIC	6805.51	6775.16
BIC	6939.80	6937.43
Log Likelihood	−3378.76	−3358.58
Num. obs.	1989	1989

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4.1: Area under the curve (AUC) model comparison, **Tarifiyt Berber**.

First, I modelled the effect of experimental variables on AUC scores. I established a maximal random effects structure supported by the data that still allowed model convergence; this included random intercepts for picture-pair and participant, and random slopes across participants for the effect of the interaction between sentence type and animacy condition. Only random intercepts were included for picture-pair because random slopes per pair exhibited full correlation with intercepts. I then compared an intercept-only version of this model with a full model that included fixed effects and interactions for sentence type (taking agent-initial as baseline), animacy balance (taking equal animacy as baseline) and trial index. This model improved significantly on the intercept-only model ($\chi^2 = 44.09$, $p < 0.0001$). I then fine-tuned the model by removing trial index, which did not significantly improve model fit ($\chi^2 = 3.75$, $p > 0.05$). The final model is detailed in Figure 4.1. I modelled the effects of the experimental variables on MD following the same procedure as for AUC. The random effects structure is the same as for the AUC model. The full model again improved significantly on the intercept model ($\chi^2 = 40.83$, $p < 0.0001$) and was fine-tuned by the removal of the non-significant trial index variable (model comparison $\chi^2 = 2.59$, $p > 0.05$) to arrive at the model detailed in Figure 4.2.

The models for the two dependent measures show a consistent pattern of results. In each model,

	intercept-only	full model
(Intercept)	0.31	0.28
	(0.03)	(0.04)
P-initial sentence		0.18^{***}
		(0.03)
P<A picture		-0.00
		(0.04)
P>A picture		-0.02
		(0.04)
P-initial sentence:P<A picture		0.07
		(0.05)
P-initial sentence:P>A picture		0.00
		(0.06)
AIC	2870.92	2842.68
BIC	3005.21	3004.94
Log Likelihood	-1411.46	-1392.34
Num. obs.	1989	1989

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4.2: Maximum deviation (MD) model comparison, **Tarifiyt Berber**.

the coefficient for patient-initial condition compared to agent-initial is positive and significant, indicating greater attraction to the competing picture when sentences are patient-initial. The coefficient for animacy condition is not significant, that is, there is no evidence that participants respond differently depending on the animacy balance of the two entities. Finally the interaction terms for the animacy factor are not significant either; in other words there is no evidence that the deviation associated with patient-initial trials is affected by the relative animacy of the patient to the agent.

Dutch: AUC and MD

Figure 4.7 shows the mean of the trajectories collected in the Dutch study, distinguished by sentence structure (two levels) and animacy condition (three levels). Line style indicates sentence type (agent-initial is solid, patient-initial is dashed) and colour indicates the animacy balance of the two entities depicted (equal animacy is red; for mixed animacy, patient less animate than agent is green and patient more animate than agent is blue). Trajectories in the patient-initial condition (dashed lines) deviate more towards the competing response, but only when the patient and agent are equally animate or patient is less animate than agent: when patient is more animate than agent (blue lines) there is little or no divergence between agent-initial and patient-initial trials.

Modelling followed the same procedure as in section 4.3.1 and the maximal supported random effects structure was also the same. Unlike the Tarifiyt data however, in the Dutch data there was a small but significant effect of trial number in both AUC and MD models, so this term was retained in the full model, which was found to improve on the intercept-only model (AUC model comparison:

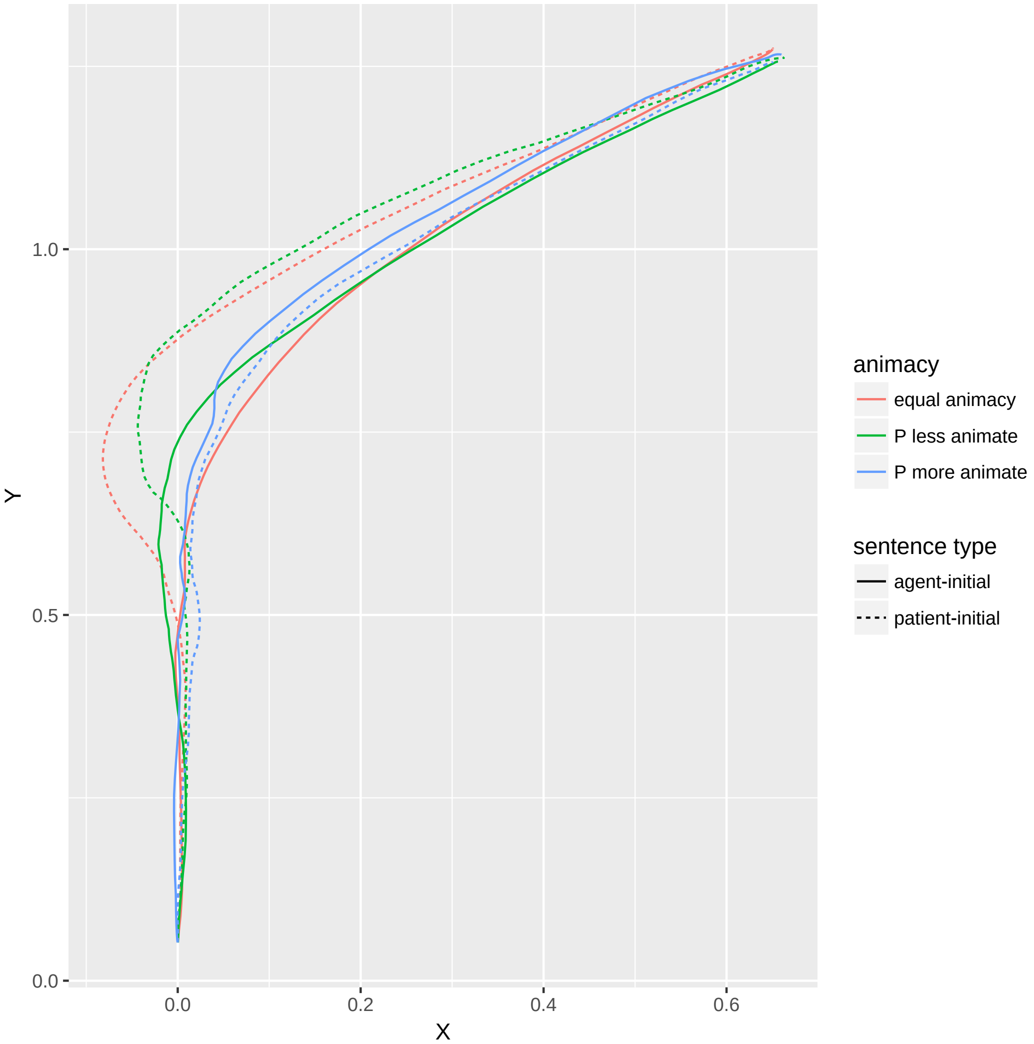


Figure 4.7: Mean trajectories separated by sub-conditions, **Dutch**.

$\chi^2 = 37.19$, $p < 0.0001$; MD model comparison: $\chi^2 = 36.65$, $p < 0.0001$). The effect of trial number was not found to interact with the experimental variables. The final models for MD and AUC are summarised in Tables 4.3 and 4.4.

	intercept-only	full model
(Intercept)	0.57	0.46
	(0.06)	(0.10)
P-initial sentence		0.41^{***}
		(0.09)
P<A picture		0.13
		(0.09)
P>A picture		0.15
		(0.09)
trial		0.01^{**}
		(0.00)
P-initial sentence:P<A picture		-0.00
		(0.11)
P-initial sentence:P>A picture		-0.45^{***}
		(0.12)
AIC	5764.75	5739.56
BIC	5898.89	5907.24
Log Likelihood	-2858.37	-2839.78
Num. obs.	1977	1977

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4.3: Area under curve (AUC) model comparison, **Dutch**.

Differently from the Tarifiyt data, it can be seen that for both MD and AUC there is an interaction of sentence type with animacy condition: patient-initial trials have greater MD and AUC than agent-initial, except when the patient is more animate than the agent. This again corroborates the observations from the mean trajectory plot in 4.7.

4.3.2 Measures of deviation: raw time data

In looking at the deviation of trajectories on the X axis across (raw) time, I am interested in uncovering to what extent participants head towards the side of the correct response or the side of the incorrect response directly after hearing the first noun phrase, and how this may or may not differ between sentence types and animacy conditions. The time bins analysed are the time bins for which there are at least enough data points to compare the conditions. For the Tarifiyt data this was between time bins -18 and +35 relative to the noun phrase offset, while for the Dutch it was between -14 and +26.

	intercept-only	full model
(Intercept)	0.37	0.33
	(0.03)	(0.04)
P-initial sentence		0.15^{***}
		(0.03)
P<A picture		0.05
		(0.03)
P>A picture		0.05
		(0.03)
trial		0.00^{**}
		(0.00)
P-initial sentence:P<A picture		-0.03
		(0.04)
P-initial sentence:P>A picture		-0.16^{***}
		(0.04)
AIC	1810.76	1786.11
BIC	1944.91	1953.79
Log Likelihood	-881.38	-863.06
Num. obs.	1977	1977

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4.4: Maximum deviation (MD) model comparison, **Dutch**.

Tarifiyt Berber: X-axis deviation relative to offset of first noun phrase

The plot in figure 4.8 shows the deviation on the X-axis of mean trajectories per experimental condition in Tarifiyt (colour indicates sentence type and line type indicates animacy balance). As discussed, the trajectories are aligned to the offset of the first noun phrase (time bin 0, marked with a dashed vertical line).

The plots in Figure 4.8 indicate divergence of trajectories between agent-initial and patient-initial trials in all three animacy conditions. Note that in these plots, the Y values in this figure reflect the X coordinates of the mouse trajectory, while the X axis is used to represent time. In other words, a positive-going trajectory (i.e. upwards) indicates movement to the side of the correct response, while a negative-going trajectory (i.e. downwards) indicates movement towards the side of the competing picture. Therefore it can be seen from these plots that trajectories on patient-initial trials (dashed lines) overall deviate more than agent-initial trials towards the side of competing picture. Moreover, in all three animacy conditions, immediately following the offset of the first NP, the trajectories in patient-initial trials deviate directly towards the competing picture – as evidenced by the dip below zero on the Y axis between 0 and 750 ms.

The divergence between agent-initial and patient-initial trials is reflected in the results of the permutation tests. There are large time windows in all three animacy conditions where permutation tests indicate significant divergence on the X axis between conditions. The longest of these time

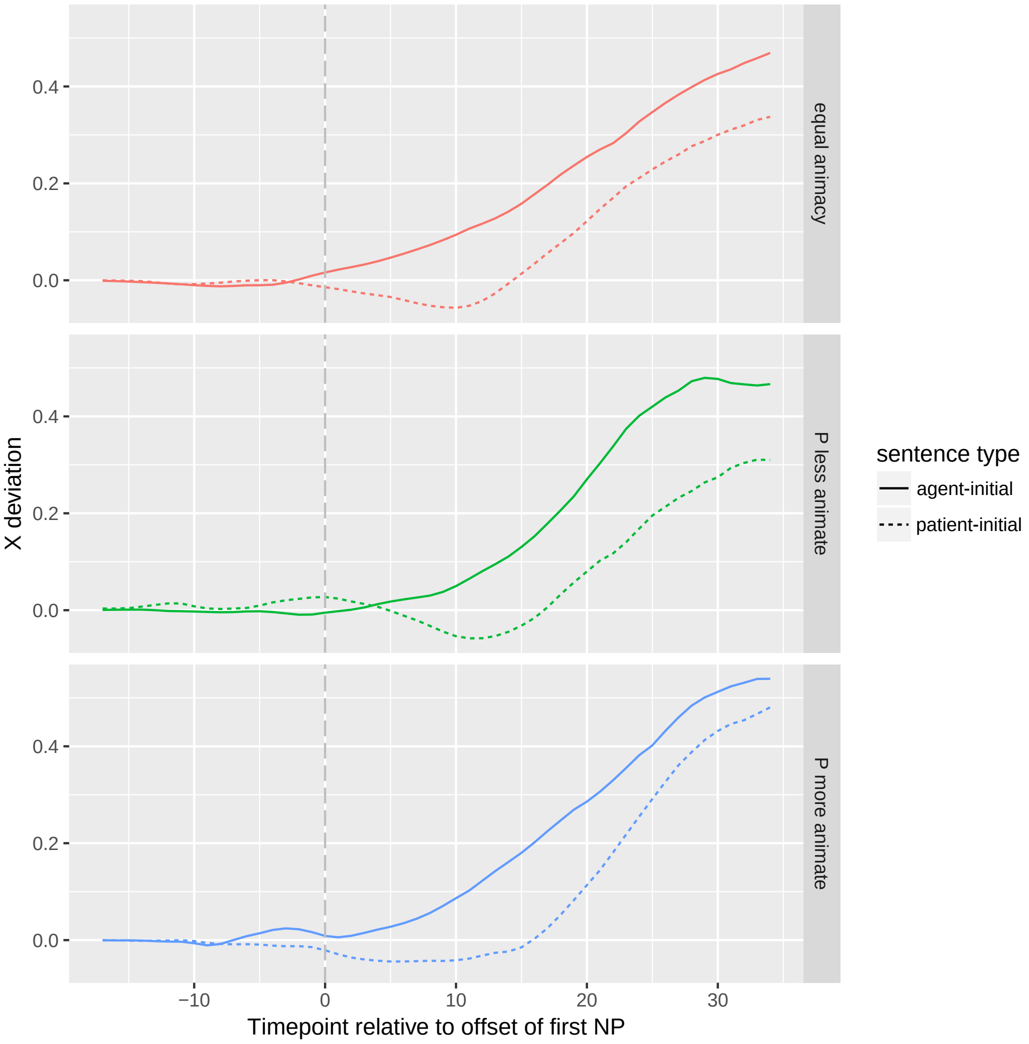


Figure 4.8: Mean deviation on X axis, **Tarifiyt Berber**. Attraction towards the competing response option is represented by lower values on the Y axis: zero on the Y axis can be understood as representing the location at the mid-line of the screen (neither left nor right).

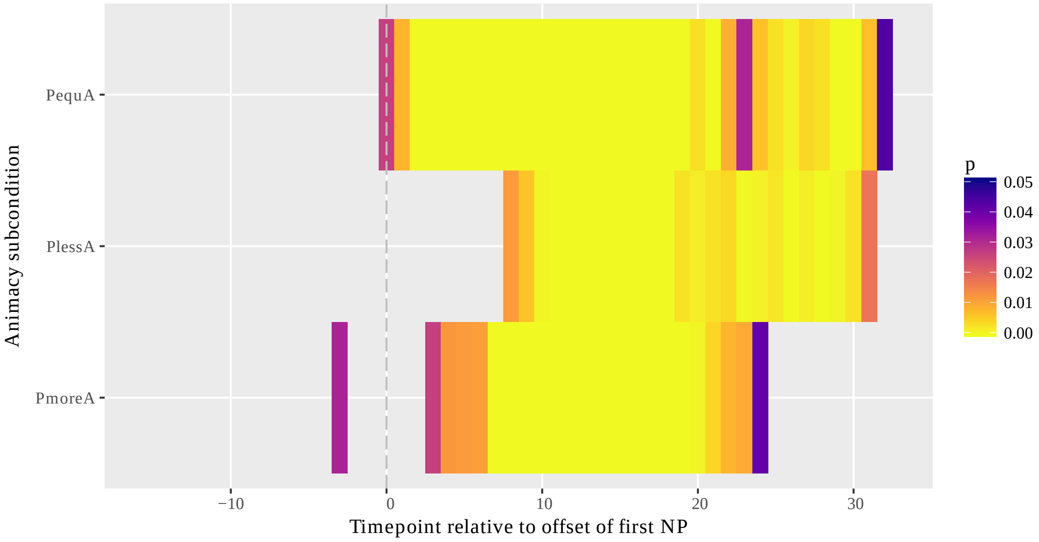


Figure 4.9: Permutation tests of X deviation, **Tarifiyt Berber**. *p* values reflect the significance of difference in X deviation between sentence types at each timepoint relative to the offset of the first NP. Each animacy sub-condition is tested separately.

windows is in the equal-animacy condition, followed by P<A then P>A.

Dutch: X-axis deviation relative to offset of first noun phrase

The graph in Figure 4.10 shows the deviation on the X-axis of mean trajectories per experimental condition for Dutch (again, colour indicates sentence type and line type indicates animacy balance). The trajectories are aligned to the offset of the first noun phrase (timepoint 0, marked with a vertical dashed line). As mentioned above, the timepoints represented in this figure are the timepoints for which there are at least enough data points to compute the permutation test. Permutation test results are displayed in Figure 4.11.

The plots in Figure 4.10 show an overall difference between patient-initial and agent-initial trials except for the P>A condition. This corroborates the observations from the normalised trajectory analysis. There is a clear initial direct deviation towards the competing picture (the line dips below zero) for patient-initial trials in the equal-animacy condition. In the P<A condition the patient-initial dip below zero is very slight, while in the P>A condition both agent-initial and patient-initial only move towards the side of the correct response. In fact, the patient-initial trajectory moves earlier towards the correct response than the agent-initial, that is, the lines display the reverse pattern compared to Equal Animacy and P<A.

The permutation tests clearly reflect the divergence of patient-initial and agent-initial trajectories, that appears to ensue immediately following the offset of the first NP in the equal-animacy condition. For the mixed-animacy conditions, the plot exhibits smaller windows of significant divergence between patient-initial and agent-initial. The window of difference between patient-initial and agent-

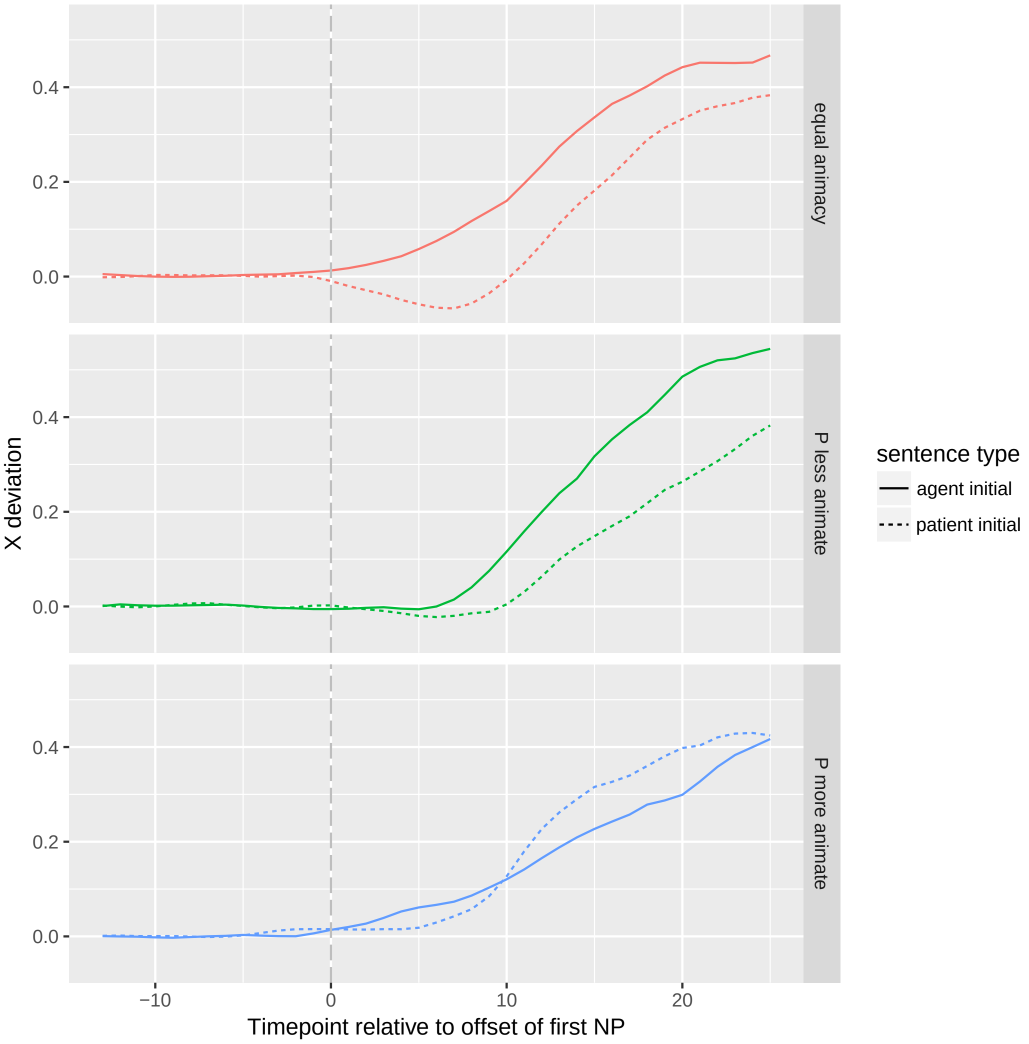


Figure 4.10: Mean deviation on X axis, **Dutch**. Attraction towards the competing response option is represented by lower values on the Y axis: zero on the Y axis can be understood as representing the location at the mid-line of the screen (neither left nor right).

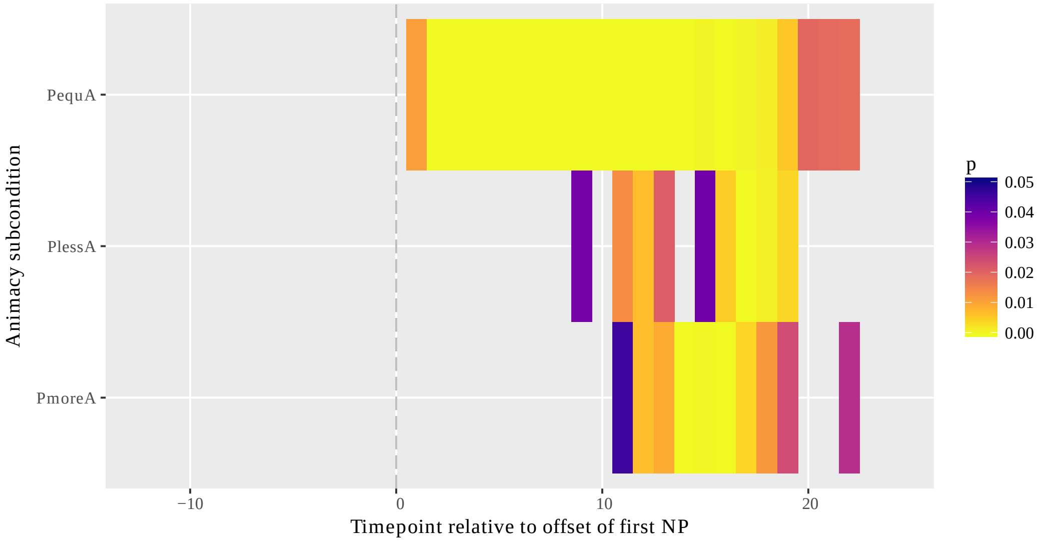


Figure 4.11: Permutation tests of X deviation, **Dutch**. *p* values reflect the significance of difference in X deviation between sentence types at each timepoint relative to the offset of the first NP. Each animacy sub-condition is tested separately.

initial is in fact larger in the P>A condition; however, referring back to the plot in 4.10, note that the difference here is actually in the opposite direction.

4.3.3 Distributional analysis of MD scores

Tarifiyt Berber: histogram and Hartigan's Dip

Figure 4.12 shows the distribution of MD scores, again split by conditions of the experimental variables. On visual inspection, MD shows a bimodal shape in all sub-conditions with two areas of concentration around zero and one. To further assess the bimodality of distributions of MD scores, I estimated the Hartigan's Dip coefficient for each combination of the experimental variables (i.e. the subsets of the data as displayed in the histograms). Hartigan's Dip is explained in more detail in Section 4.2.4; I estimated it using the *diptest* package in R (Maechler, 2016). The coefficients and corresponding *p* values for the test are provided for each of the data subsets.

The test is significant for all sub-conditions of MD; however, the magnitude of the Dip statistic (and significance of the accompanying p value) is much greater when patient is initial, except when the initial patient is more animate than agent.

The distribution plots and tests suggest that overall, the trajectories in this experiment are either quite direct towards the correct response (clusters of MD scores around zero), or deviate sharply towards the competing picture (hence the second cluster of MD scores around 1). The greater magnitude of the Dip statistics for MD in the patient-initial plots correspond with the picture of greater overall deviation for this condition as seen in Figure 4.6. Within the the P>A condition, the relatively smaller Dip statistic for patient-initial trials appears to relate to a more even spread of maximum deviation scores between zero and one, leading to a less stark distinction between the two clusters in the bottom right pane in 4.12.

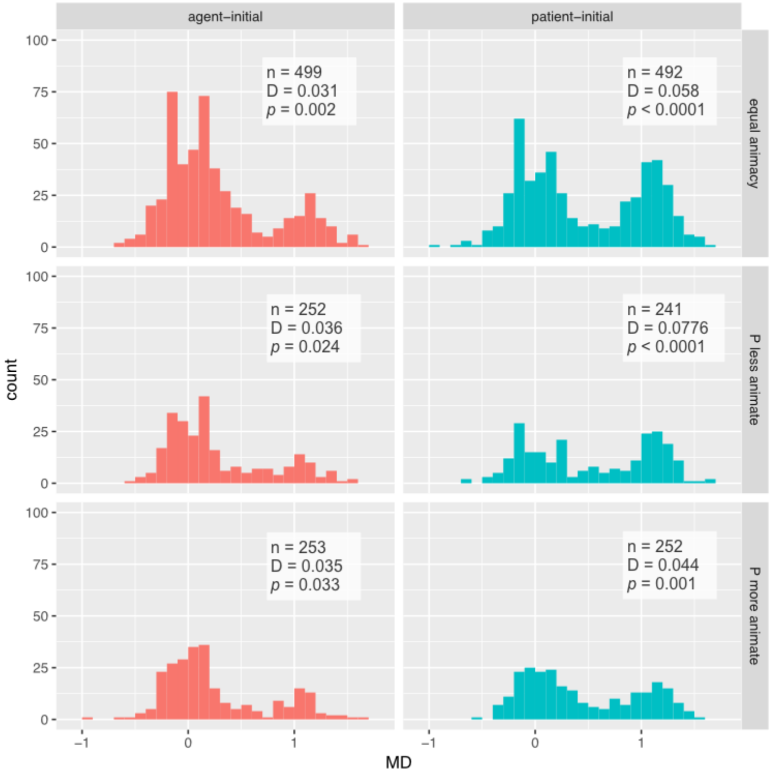


Figure 4.12: Distribution of maximum deviation scores, **Tarifiyt Berber**. Inset in each subplot is the number of observations (n), Hartigan's Dip score (D) and corresponding *p* value (null hypothesis: distribution is unimodal).

Dutch: histogram and Hartigan's Dip

Figure 4.13 shows the distribution of MD values, again split by conditions of the experimental variables. On visual inspection, MD distributions appear to form two peaks at zero and one in most sub-conditions, although this is diminished in the P>A condition.

To assess bimodality statistically, I again computed Hartigan's Dip for all six sub-conditions. The coefficients and corresponding *p* values for the Hartigan's Dip tests are provided for each of the data subsets. The tests indicate that there is a significant departure from unimodality when sentences are patient-initial, except for when the patient is more animate than agent (P>A condition).

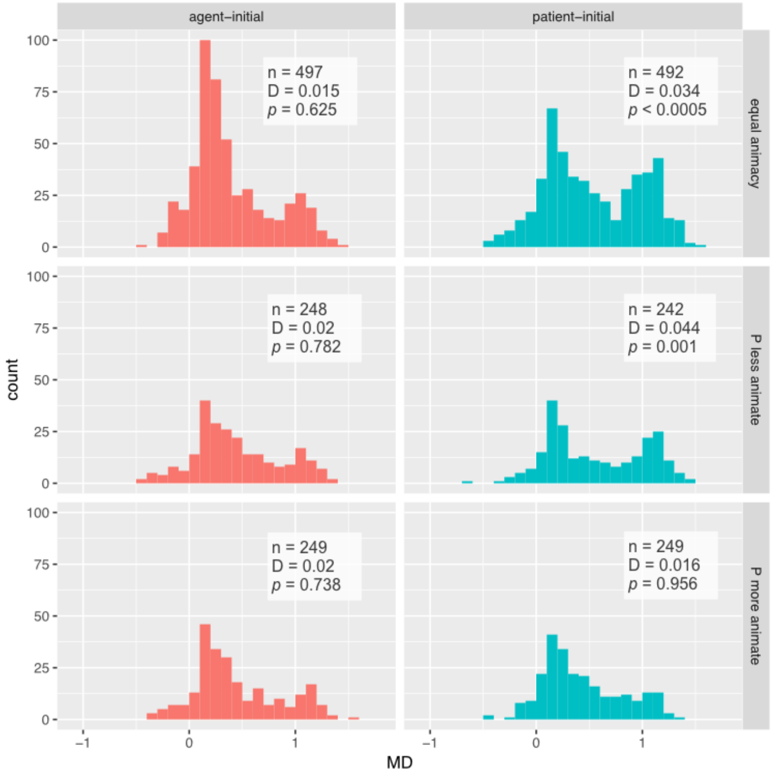


Figure 4.13: Distribution of maximum deviation scores, **Dutch**. Inset in each subplot is the number of observations (n), Hartigan's Dip score (D) and corresponding p value (null hypothesis: distribution is unimodal).

4.4 Preliminary discussion

The primary goal of this study was to use the mouse data to probe the relationship between the relative animacy of agent and patient and their positions in the sentence. The aim was to see whether provisional assignment of thematic structure based on the first-heard noun phrase would be affected by knowledge about whether it was more animate than the other participant in the event. In order to assess the effects of animacy and firstness, it was first important to establish that predictive comprehension based on initial arguments during sentence comprehension is indeed visible in the mouse trajectories. In both the Tarifiyt Berber and Dutch data, we see an overall pattern of larger deviation in the patient-initial condition, with participants heading first towards the response where the first-heard argument is the agent. These results suggest that most of the time, hearers preferentially assign agent role to the sentence-initial argument. This aligns with expectations, and indicates that the design has succeeded in tapping into the targeted mechanisms. In the following sections, I briefly reflect on the findings of the normalised, raw time and distributional analyses before presenting an integrated discussion of these results.

Normalised data

When modelling the Tarifiyt MD and AUC scores using linear mixed effects modelling, patient-initial trials have significantly higher MD and AUC scores overall, but this pattern does not differ significantly between animacy conditions. In other words, there is no evidence of a significant interaction between animacy and sentence position in the Tarifiyt models. The Dutch data show a different picture, however: in both the AUC and MD linear mixed effects models, the interaction between animacy condition and sentence type is significant. In particular, the difference between the two sentence types does not differ when the patient is less animate than the agent compared to when animacy is balanced; however, when the patient is *more* animate than the agent, the difference in both AUC and MD between sentence types is significantly reduced compared to when animacy is balanced (as evidenced by the significant negative parameter estimate for this interaction term in both models). In summary, the picture we get from these results is that animacy and sentence position interact in Dutch but not in Tarifiyt.

Raw time data

The inspection of X deviation relative to the offset of the first noun phrase substantiates the findings based on the MD and AUC models. Firstly, note that the divergence of trajectories in all conditions is indeed associated with the offset of the first noun phrase, indicating again that the design has succeeded in tapping into the predictions made on the basis of the initial argument. In the Tarifiyt data, the mean trajectory for the patient-initial trials ventures into the hemispace of the incorrect response in each of the three animacy conditions. In Dutch this also is the case; *except* for when the initial patient is more animate than the agent. The permutation tests for both languages provide additional evidence beyond the mean trajectory plots by indicating windows of significant divergence between agent-initial and patient-initial trials in all subconditions.

At this point it is appropriate to note that in some of the plots of X deviation (especially in Figure 4.8), there appears to be a small divergence between the conditions already before time-point zero. The permutation tests reveal that there is only one area where this is likely to be significant, namely when patient is more animate in the Tarifiyt experiment (see Figure 4.9). It is difficult to draw a conclusion about why this pattern is observed. Note that timing could play a role here: the datapoints were binned into windows of 50 ms, reducing the timing resolution (which may already have a margin of error as the hardware and software were not optimised for low latencies). In addition, noun phrases had different lengths; this means that in the case of longer noun phrases, participants may have already identified the referent before the offset point, in contrast to shorter noun phrases. Two points for future study can be made here. Firstly, it would be useful to constrain noun phrases to have similar lengths. However, this may lead to difficulties in maintaining a sufficient stimuli set (indeed, this is the reason why it was not possible to constrain it in the current experiment); in any case, different languages will also vary in the relative lengths of noun phrases used to describe the pictured referents, so controlling noun phrase length may not be realistic cross-linguistically. Secondly, future mousetracking research would certainly benefit from the optimisation and validation of software and hardware for timing accuracy. For example, hardware with lower latencies, such as computer gaming peripherals, may provide better means for collecting time-accurate mousetracking data.

Distributional analysis

Turning to the distribution of MD scores, visual inspection of the histograms suggests that responses have a tendency to form two clusters: trajectories with very minimal deviation and trajectories with a considerable deviation towards the competing response. The larger deviation peak is above the cutoff suggested in previous studies for detecting abrupt revision. This pattern of results suggests that trajectories tend to either head directly to the correct response, or directly to the incorrect response and then re-route to the correct response. This dichotomy of trajectory types is consistent with the idea that hearers are using initial argument information to make provisional assignments that must be abruptly revised if found to be wrong.

The Hartigan's Dip statistics indicate bimodality in all subconditions for Tarifiyt; however, note that the magnitude of the Dip statistic (and associated *p* value) is very different between the agent-initial and patient-initial conditions, suggesting that a larger proportion of the patient-initial trials elicited incorrect predictions that were abruptly revised. This pattern, however, is not so pronounced in the case that initial patient is more animate than agent. The bimodality tests for Dutch display the same pattern, though more starkly: the only conditions in which Hartigan's Dip reaches significance is in the patient-initial trials where patient is either equally animate or less animate than the agent. In other words, again we see that responses for the P>A condition follow a different pattern to the other two patient-initial subconditions. Taken together, we see across both languages a tendency for patient-initial trials to elicit patterns of abrupt revision, with this tendency being reduced or absent when the initial patient is more animate than the agent.

In all three of these analyses, we see evidence that the two languages differ with regard to the interaction of animacy with argument ordering (sentence type). In order to fully explore this finding, it is informative to perform a secondary analysis across all data, including language as a variable within the analysis. This will give a stronger idea of the robustness of the difference observed between the two languages.

4.5 The differential role of language

To recap: in Tarifiyt, there was no interaction between sentence type and animacy. Patient-initial structures were associated with more erroneous movements, but the animacy of the two referents did not influence this. In Dutch, patient-initial structures were again associated with more erroneous movements; however, unlike Tarifiyt, this was affected by the relative animacy of the two referents. As a result, the interaction of animacy and sentence type may seem to depend on the language of the experiment. To assess this further, we can investigate whether the interaction between animacy and sentence type interacts with the language of the experiment (i.e. whether there is a three-way interaction).

In what follows, I will pool the data from both languages and explore the relationship between these three variables. I will assess the presence of the three-way interaction that is suggested by the per-language results. First, I will explore the normalised trajectories (compare Sections 4.3.1 and 4.3.1); then I will draw on the raw time data to see how the interaction manifests in real-time following the offset of the first noun phrase (compare Sections 4.3.2 and 4.3.2).

4.5.1 Normalised data: SENTENCE TYPE × ANIMACY × LANGUAGE interaction

First I assess the existence of a language interaction in the normalised data. Here, I report linear mixed models for MD and AUC as before, based on normalised trajectory data; this time, with the addition of language as a factor (two levels), interacting with sentence type (two levels) and animacy (three levels). Before statistically modelling this three-way interaction, it is important to first visualise the relationship between the variables, by plotting the pooled data.

Figure 4.14 displays the mean normalised trajectories for each level of the 3×2×2 interaction. This plot essentially recaps Figures 4.6 and 4.7. In this figure, the upper and lower panels show Tarifiyt and Dutch data, respectively. Animacy subconditions are subdivided into three panels, left to right. Sentence type is distinguished within each panel, with solid trajectories for agent-initial trials and dashed trajectories for patient-initial trials, as in earlier plots.

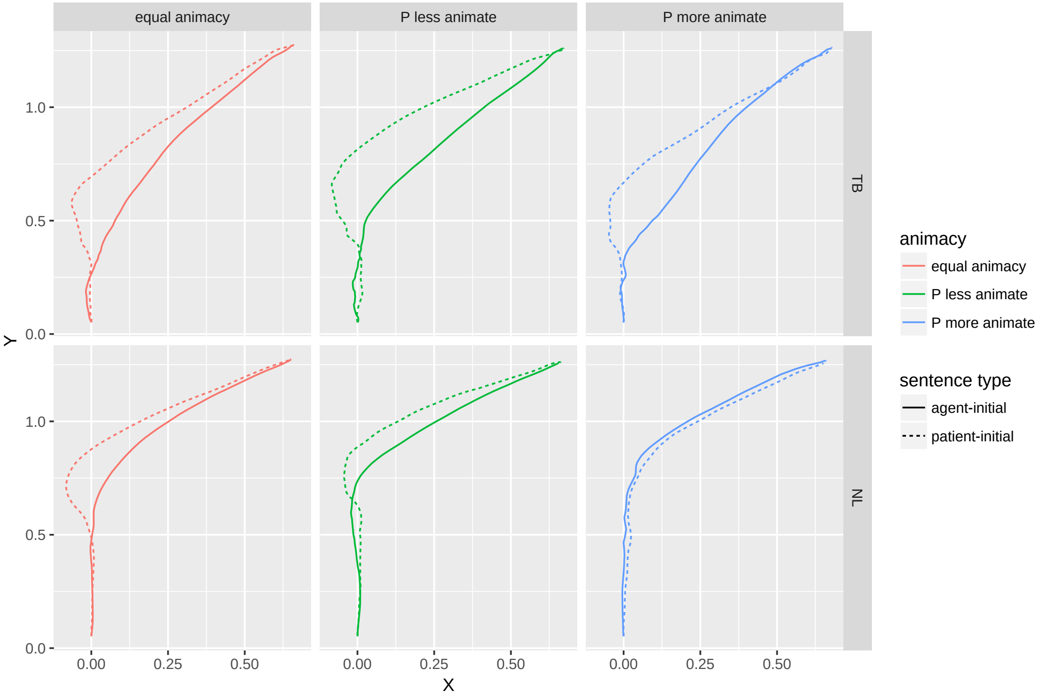


Figure 4.14: Mean trajectories (normalised) compared across animacy, sentence type and language. The upper three panels show the mean trajectories for Tarifiyt Berber, while the lower three panels show the mean trajectories for Dutch.

Here, we see again that the divergence of trajectories on patient-initial versus agent-initial trials (dashed versus solid lines) is similar across all animacy subconditions for Tarifiyt; meanwhile, the separation between patient-initial trials and agent-initial trials in Dutch is different across the three animacy subconditions. Notably, there is hardly any separation between the mean trajectories in Dutch when patient is more animate than agent.

Next, we can proceed to model the area under the curve (AUC) and maximum deviation (MD) for the pooled data. The same model fitting procedure is adopted as previously. The language variable is categorical with two levels, Tarifiyt and Dutch. I select Tarifiyt as the reference category (i.e. the pattern of data in Dutch is evaluated against the pattern in Tarifiyt). Based on the per-language findings, a fixed effect for trial order is also included. Again, in both models, the maximal random effects structure supported by the data was applied: random intercepts for PICTURE-PAIR, and random intercepts and random slopes for the SENTENCE-TYPE X ANIMACY interaction effect. Model comparison using ANOVA confirmed that these models both performed better than intercept-only models, or models without TRIAL ORDER as a covariate. Table 4.5 summarises the fixed effects parameters of these two models respectively.

	Area under the curve (AUC)	Maximum deviation (MD)
(Intercept)	$0.44\ (0.10)^{***}$	$0.22\ (0.04)^{***}$
P-initial	$0.54\ (0.09)^{***}$	$0.18\ (0.03)^{***}$
P<A	$-0.00\ (0.10)$	$0.00\ (0.03)$
P>A	$-0.02\ (0.10)$	$-0.02\ (0.03)$
Dutch	$0.04\ (0.12)$	$0.12\ (0.05)^{*}$
trial order	$0.01\ (0.00)^{***}$	$0.00\ (0.00)^{**}$
P-initial:P<A	$0.09\ (0.13)$	$0.06\ (0.05)$
P-initial:P>A	$-0.14\ (0.14)$	$0.00\ (0.05)$
P-initial:Dutch	$-0.12\ (0.13)$	$-0.04\ (0.04)$
P<A:Dutch	$0.13\ (0.13)$	$0.05\ (0.05)$
P>A:Dutch	$0.17\ (0.13)$	$0.07\ (0.05)$
P-initial:P<A:Dutch	$-0.09\ (0.18)$	$-0.09\ (0.07)$
P-initial:P>A:Dutch	$-0.31\ (0.20)$	$-0.16\ (0.07)^{*}$

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Random effects specification: $(1|picturepair) + (1 + sentencetype * animacy|participant)$

Table 4.5: SENTENCE TYPE X ANIMACY X LANGUAGE linear mixed effects models for AUC and MD.

Our interest here concerns the interaction of sentence type, animacy and language: in other words, the final two parameters in the list. *P-initial:P<A:Dutch* estimates the effect on AUC and MD when (a) the sentence heard is patient-initial (rather than agent-initial), (b) the patient is *less* animate than the agent (rather than equal to it), and (c) the language is Dutch (rather than Tarifiyt). This corresponds to the middle panels of Figure 4.14. Both parameters are slightly negative but not significantly so. *P-initial:P>A:Dutch* estimates the effect on AUC and MD when (a) the sentence heard is patient-initial (rather than agent-initial), (b) the patient is *more* animate than the agent (rather than equal to it), and (c) the language is Dutch (rather than Tarifiyt). This corresponds to the right-hand panels of Figure 4.14. For MD, the effect is negative and significant. This indicates a significantly smaller X-axis deviation for patient-initial P>A trials (e.g. 'the farmer is kicked by the donkey') for Dutch versus Tarifiyt. The parameter estimate for AUC is also negative and quite large, but does not reach statistical significance.

There is a significant main effect for trial in this model, indicating again that AUC and MD increase linearly through the course of the experiment. This parameter is highly significant, however note that the estimate of the linear effect is rather small (0.01 and <0.01 for AUC and MD respectively).

It is interesting to reflect on the significant main effect of Dutch in the MD model. Although the main effect of Dutch is surpassed by the significant interaction effect that includes Dutch, it is interesting to note that the main effect and interaction effect are in different directions. The main effect parameter is positive, which would indicate that Dutch is associated overall with greater X-axis deviation. However, the interaction effect parameter is negative: so, when looking specifically at patient-initial P>A trials, the X-axis deviation is less than in Tarifiyt. This underlines the difference in trajectory patterns between the two languages.

The normalised data is limited in what it can tell us about how the trials unfold in time in response to the initial argument. For a better assessment of this conclusion, we can turn to the raw time data.

4.5.2 Raw time data: trajectory divergence compared between languages

Preparation for between-language analysis of divergence

Figures 4.8 and 4.10 provided a visualisation of the divergence between agent-initial and patient-initial trials as a function of animacy. In those plots, the divergence along the X-axis could be assessed by comparing the lines for agent-initial versus the lines for patient-initial trajectories.

However, simply comparing line plots does not tell us whether the difference between languages is significant. Especially in the case of trajectory data, line plots are the result of averaging over trajectories (as discussed in Section 4.1.2), and may hide the underlying variance. Therefore, in Section 4.3.2 and 4.3.2, I used permutation testing to statistically test for divergence between agent-initial and patient-initial trajectories.

Now, however, I wish to explore *how this divergence itself differs between the two languages*. I wish to do so using the same type of analysis, namely, permutation testing, to directly test for differences in divergence between the two languages. In order to do this, it is necessary to compute a single variable that reflects the divergence. To compute a variable reflecting divergence between sentence-types, I subtract the patient-initial X-deviation from agent-initial X-deviation at each timepoint.[5] For example, if agent-initial deviation is 1, and patient-initial deviation is -0.5, the divergence score will be 1.5, i.e. the distance between the two values. This is analogous to the calculation of difference waves, as is sometimes done in the analysis of EEG data.

Divergence plots and permutation tests

Figure 4.15 visualises the divergence of agent-initial and patient-initial responses in the first 1000 ms following the offset of the first noun phrase. In this figure, the more extreme the values on the Y-axis, the greater the divergence. Positive Y-values indicate that the patient-initial trials involve more movement towards the competitor than agent-initial trials. Negative Y-values indicate the opposite kind of divergence: that the patient-initial trajectories move *less* towards the competitor than agent-initial trajectories. Lastly, Y-values around zero indicate no divergence between the conditions.

We can see from this figure that for equal animacy and P<A, the trajectories diverge increasingly over the first 1000 ms after hearing the first noun phrase. Moreover, the behaviour of the two languages

[5] To do so, it is first necessary to average across participants within each language (i.e. in order to obtain one value per condition per time point).

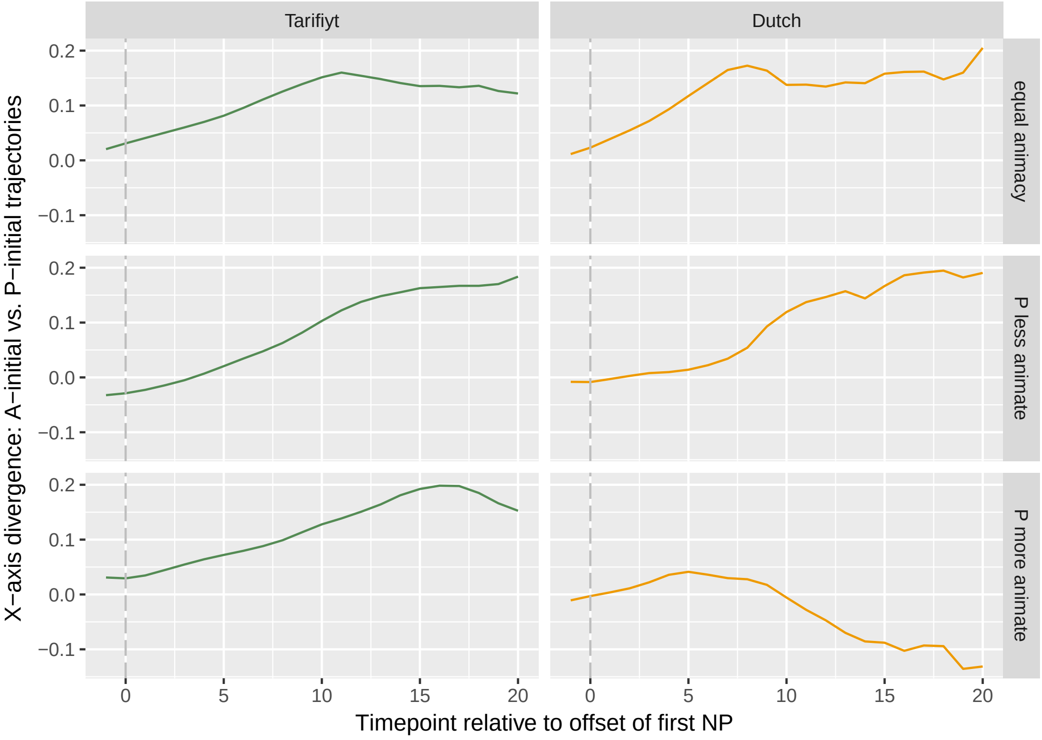


Figure 4.15: Divergence over time of agent-initial versus patient-initial trajectories, in each of the three animacy conditions. The left-hand panels show the divergence in each animacy condition in Tarifiyt, while the right-hand panels show the divergence for each animacy condition in Dutch.

is similar: participants are drawn more to the incorrect response on patient-initial trials than on agent-initial trials. However, we see a stark difference between languages when patient is more animate than agent (P>A, lowest two panels). After hearing the initial argument, Tarifiyt trajectories diverge in the same pattern as before, indicating the same extent of erroneous assumptions that this initial argument is agentive. In Dutch, however, the trajectories do not diverge in the same way over the first 1000 ms after the first noun phrase. In the first 500 ms, there is very little divergence. From 500 – 1000 ms there is negative divergence, indicating that participants are drawn *more* towards the correct response on patient-initial trials.

Figure 4.16 shows the result of permutation testing. As in Sections 4.3.2 and 4.3.2, the three panels from top to bottom represent equal animacy, patient less animate than agent, and patient more animate than agent. The variable being contrasted at each timepoint is language: Tarifiyt versus Dutch. Here, we are testing the null hypothesis that the data were sampled from a distribution where values do not vary systematically with language. At each timepoint, p <0.05 indicates that there is less than 5% chance that the observed datapoints were sampled from a distribution where language has no effect on divergence.

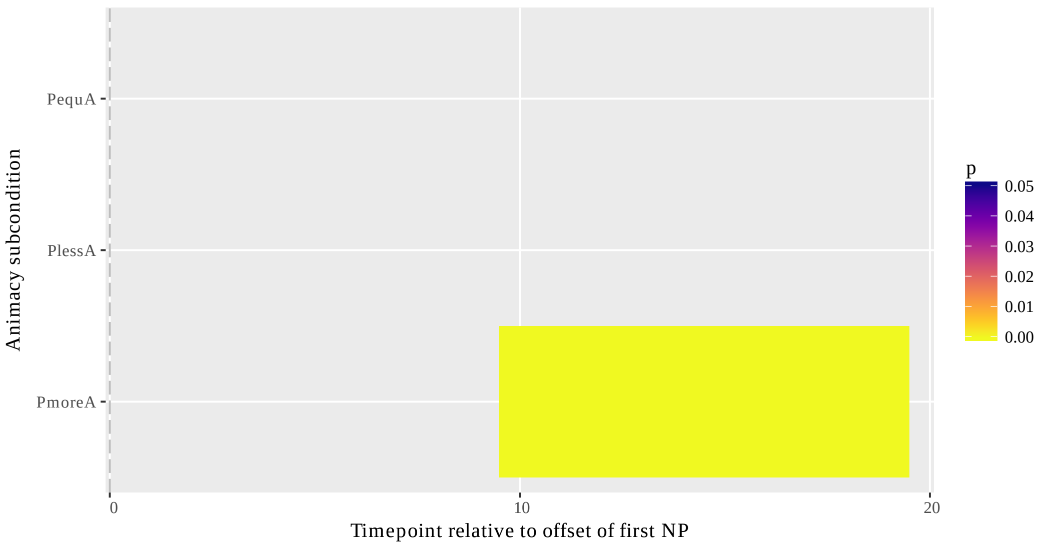


Figure 4.16: Permutation tests comparing divergence over time between Tarifiyt and Dutch, for each animacy subcondition.

The outcome of the permutation tests supports the observed differences in Figure 4.15. Overall, the pattern of trajectory divergence does not differ between the two languages; however, in the case of trials where patient is more animate than agent (P>A), there is a strong difference between the two languages. This corroborates the plot, covering approximately 500–1000 ms after the offset of the first noun phrase (10th–20th timepoints).

4.5.3 Discussion of between-language analyses

Through this combined analysis of the two languages, we are able to draw more informed conclusions about the degree to which languages differ. Turning first to the **normalised data**, the results here provide further evidence that sentence type and animacy interact differently depending on the language of the experiment. The observations based on the plots in Figure 4.14 are borne out in the parameters of the MD model reported in Table 4.5. In other words, when it comes to the P>A condition, MD scores for Dutch are significantly less than those for Tarifiyt. We can interpret this as evidence that Dutch speakers are less likely to assume that an initial argument is an agent when it is also the more animate argument; meanwhile, Tarifiyt speakers do not use animacy information to modulate their predictions of thematic structure.

It is noticeable that the three-way interaction term that is significant for MD is not significant in the AUC model; however, the numerical trends in the AUC model do align with the pattern of results in the MD model. Still we may ask: why would the three-way interaction be stronger for MD than for AUC? This finding is likely to be symptomatic of what is happening to the underlying cognitive dynamics. Recall that the trajectories in this study tend to be either sharply deflected or directly correct (as evidenced by bimodality; see Section 4.3.3). So, attraction to the competitor in this study is characterised by abrupt revision of movement, rather than by gradual curvature. Therefore, the reduction of MD in the patient-initial P>A condition points towards the conclusion that this condition displays fewer abrupt revisions than other conditions. In other words, in this condition, there are fewer trials where the wrong choice is made and then revised.

The **raw time data** analysis provides a more in-depth picture of the divergence between agent-initial and patient-initial trials, and how this unfolds differently in the two languages. Plots and permutation testing underscored the difference already seen between the two languages. Namely, that the Tarifiyt responses do not pattern differently across animacy conditions, but that divergence is affected by animacy in the Dutch data. Specifically, a significant difference was found between the two languages in the P>A condition. This indicates that Tarifiyt speakers in this condition continue to make the same predictions regarding thematic structure, while Dutch speakers significantly reduce their bias towards an agent-initial interpretation. Overall, it seems that Tarifiyt hearers are not sensitive to animacy in predicting thematic structure, but Dutch hearers *are* sensitive to animacy in predicting thematic structure.

4.6 General discussion

With this study, I wished to explore how linear ordering and animacy influence hearer expectations of thematic structure, in sentences and contexts that are based on the production experiment in Chapter 3. On the basis of previous research, I expected that hearers would be more likely to interpret sentence initial arguments as agents. My primary aim was to explore whether and how this pattern would be affected by animacy balance between the two referents. Additionally, I wished to examine these effects in two unrelated languages that, in the foregoing production experiment, showed complementary strategies for the grammatical encoding of prominent arguments.

In the first place, there is strong evidence from these experiments for the preference of hearers to assign agent role to the sentence-initial noun phrase. Notably, this effect is demonstrated in both of the languages, despite differences in the structures that are required in the two languages to accommodate patient sentence-initial position – i.e. passive in Dutch and object-topicalisation in Tarifiyt. This supports previous findings of the cross-linguistic robustness of this tendency (Bornkessel-Schlesewsky & Schlesewsky, 2014:110).

Considering the effects of animacy, we see that trials where the initial argument was human were *not* related to an increased deviation towards the picture where the human was agent. In Tarifiyt, the deviation did not differ significantly across animacy sub-conditions. In Dutch, the human-initial trials were actually associated with significantly *less* initial deviation towards the human-agent response picture.

In Section 4.1.3, two possible views were considered – whether animacy and linear order would combine to increase the likelihood of assigning agent role additively, or whether the information would be integrated in a more probabilistic manner. The findings in this study indicate that high animacy does not purely increase the likelihood of assigning the initial argument the agent role in a purely additive fashion. In Dutch, moreover, the factors of animacy and order seem to interact.

We may ask how exactly hearers weigh up these different sources of information in predicting thematic structure. Generally speaking, this pattern of results fits well with accounts proposing that hearers make use of probabilistic knowledge in predicting thematic structure (Kuperberg & Jaeger, 2016). Under a probabilistic view of sentence comprehension, hearers draw on linguistic experience in interpreting thematic structure. This account sees hearers as predicting thematic structure not purely through the guidance from cues, but by drawing on their knowledge of statistical distributions encountered in previous input. In this way, ambiguity in comprehension is resolved on the basis of statistical probability; in other words, "the processor should prefer the parse with the highest probability" (Chater & Manning, 2006:338).

Turning again to the findings from the current experiment, probabilities in the input can indeed account for why human initial sentences did not trigger higher attraction to human-agent pictures. Since human arguments are prime candidates for topicalisation and/or passivisation (as demonstrated by the sentence production study in Chapter 3), it follows that hearers are in fact *less* likely to commit to the agent interpretation of an initial argument when it is human. The probabilistic approach can also be understood as a strategy to comprehend efficiently and fast in the face of ambiguity. In an equal-animacy scenario, the hearer can make a fairly reasonable assumption that the first argument is likely to be agentive. However, in mixed-animacy scenarios, predicting an agent-initial sentence based on hearing a human argument in initial position is more risky. Given that an inanimate referent is involved in the scene, there is a high likelihood that the human has been placed in initial position due to animacy effects on sentence production. Therefore we are less likely to see agent-initial assumptions being made for human-initial mixed-animacy sentences, because there is in fact a lower probability of being correct than when animacy is balanced.

The cross-linguistic nature of this study also provides evidence of between-language differences. When we compare the two languages with each other, it appears that hearer expectations are different between the two languages. On the one hand, there is a similar effect of sentence position; on the other hand, there are different effects of animacy. How can we account for this finding?

It is possible to extend the probabilistic view to account for this finding, too. As mentioned, probabilistic accounts see hearer behaviour as directly reflecting probabilities in the input (Chater & Manning, 2006; MacDonald, 2013). Therefore, these findings would make sense if we could show that there is a stronger effect of animacy on production in Dutch than in Tarifiyt. In other words, perhaps the Dutch participants are more inclined to use animacy to predict thematic structures because the linguistic input in Dutch shows a higher sensitivity to animacy. Likewise, maybe the Tarifiyt participants are less inclined to use animacy to predict thematic structures, because Tarifiyt input shows a less strong relationship between linear order and animacy. In order to follow up on this possibility, we can reflect on the results of the production experiment Chapter 3. Here we indeed saw that there was some evidence of between-language differences in terms of how sensitive sentence structure was to animacy balance. In particular, the effect of animacy on passivisation in Dutch seemed to be stronger than the effect of animacy on the likelihood of object-initial structures in Tarifiyt. Of course, it is important to recall that these between-language effects were not found to be statistically significant. However, the alignment between the numerical trend in the production data and the clear effects in the comprehension data provides intriguing evidence for the probabilistic view, lending support to the approach of previous authors who have drawn probabilistic links between production and comprehension (MacDonald, 2013). The findings here also corroborate a key insight of the Competition Model: namely that interpretative cues, such as animacy, may differ in strength between languages (MacWhinney, 2001). Future research with directly comparable production and comprehension studies in different languages should elucidate this issue further.

The findings here could also be more directly related to typological differences, in particular with relation to the assignment of the subject function and how this relates to animacy, ordering and agency. As discussed in Section 4.1.1, Dutch and Tarifiyt are complementary in terms of the relationship between subject function, linear order and agentivity. To recap: in Dutch, the grammatical function of subject is strongly tied to linear order, but can be disconnected from the agent role, through use of the passive. In Tarifiyt, by contrast, the grammatical function of subject is strongly tied to agent role, but can readily be disconnected from linear order, through use of object-topicalisation. Nonetheless, it is likely that, in both languages, animacy is closely tied to subject function, through its close association with agentivity. Now consider the situation in mixed-animacy trials. A participant begins to hear a sentence and attempts to assign a thematic structure to it. When the first argument heard is the more animate one of the two, both Dutch and Tarifiyt participants may be predisposed to assign it the subject function. If passivisation is possible (as in Dutch), then it is possible to assign subject function to the animate referent while remaining agnostic about whether this referent is also the agent. However, if passivisation is not available (Tarifiyt) then committing to a subject-first interpretation is the same as committing to an agent-initial interpretation. Note that this account would require us to see the assignment of subject function as a constraining step in thematic interpretation.

Language-specific interpretations offer intriguing avenues for further investigation. However, there are other strong possibilities for explaining the differences between the languages. In particular, there is the chance of confounds with between-group differences. A prime candidate for such a between-group confound is that the Tarifiyt participants on the whole used laptop computers much less frequently than Dutch participants. Among Tarifiyt speakers, smartphones are far more evident in daily usage than laptop or desktop computers, even among university students. This could, for example, mean that the way that mouse movements represent cognitive dynamics is not constant across the two languages. Nonetheless, we may bear in mind that the between-group situation may not actually be as clear-cut as this, since there were also many students in the Dutch cohort who reported that they use a trackpad rather than a mouse. Other possible between-group differences relate to the way that participants perceived the experiment and the task (cf. Section 2.2.3). Note, however, that many efforts were made to counteract (or at least attenuate) this effect. On the one hand, care was taken with matters of ecological validity and cultural norms; on the other hand, the experimental context was held as constant as possible (for example, having the same researcher conduct the studies, testing both groups in non-laboratory settings, and using the same set of equipment).

Another limitation of this study is that, although the task is designed to be cross-culturally accessible, it is nonetheless very constrained. Participants are aware of the two possible thematic interpretations before they hear the sentence, and so the space of possibility is highly restricted. For example, they can be fairly sure that the sentence will not be intransitive, and thus that the first noun phrase is either agent or patient of a transitive verb. Most crucially, at the start of the trial they are already aware of which two referents will be mentioned in the sentence and, therefore, how animacy is balanced between them. Due to this, one may criticise the ecological validity of this design. On the other hand, this aspect of the design is instrumental in demonstrating the role of probabilistic knowledge in thematic structure assignment. Based on this particular form of contextual constraint, we are able to draw the conclusion that hearers are able to make full use of the available information to converge on an interpretation of the thematic structure as soon as possible (cf. van Gompel & Pickering, 2007). Moreover, even if the information provided is not typical of communicative contexts, that does not prevent hearers from making use of it in interpreting what they hear. In this study, hearers seem to show sensitivity to the precise ways in which the complex information available to them constrains the possibilities for sentence completion, and integrate this with other sources of knowledge (such as experience of linguistic input) to inform their expectations for the overall thematic structure of the sentence.

4.7 Concluding remarks

This study demonstrated strong evidence for the preference of hearers to assign agent role to an ambiguous sentence-initial noun phrase, in both Tarifiyt Berber and Dutch. However, in mixed-animacy scenarios, animacy did not increase the likelihood of an initial argument being interpreted as agent. Instead, it either had no effect on the likelihood of agent interpretation (Tarifiyt) or even reduced the likelihood of agent interpretation (Dutch). These findings do not support the view that hearers use ordering and animacy information about an argument as additive cues to assign it a thematic role. Instead, these findings support the view that hearers combine a range of available information in a probabilistic, forward-looking fashion to guide overall thematic interpretation.

The probabilistic view was assessed further on the basis of cross-linguistic, cross-modality comparisons. On the one hand, the difference between Tarifiyt and Dutch hearer expectations in the current experiment could relate to a cross-linguistic difference in production preferences, corroborating a trend seen in the study in Chapter 3. On the other hand, the cross-linguistic differences could possibly be directly derived from typological differences in how grammatical relations are realised. Animacy and linear ordering relate not only to agenthood but also to subjecthood; the pattern of results here is consistent with the idea that hearers in the two languages make combined guesses that pay attention to how linear ordering and animacy interface with grammatical function assignment, as well as thematic role assignment.

Future research is needed to eludicate these issues further, particularly cross-linguistic comparative studies. To maximise the comparability of data from languages spoken in different communities, future studies will also need to pay special attention to the possibility of between-group confounds as alternative explanation for between-language differences. To this end, it would be highly advantageous to explore the possibility for conducting trajectory tracking studies with smartphones or tablets rather than laptops and mice, to provide better validity across a more diverse range of communities.

CHAPTER 5

Multiple choice in sentence form: animacy effects on sentence production in Pondok Tinggi

As discussed in Chapter 1, many languages offer more than two sentence forms for the felicitous realisation of the same message. The focus of this thesis is on transitive sentences, where the key differences are in the mapping of agent and patient to grammatical functions and their relative linear ordering. Arguments may be encoded as more prominent functionally (subject function) or linearly (early sentence position). The study reported in Chapter 3 employed a *simply describing* paradigm to investigate how animacy affects the likelihood of producing patient-prominent vs. agent-prominent structures. However, in a given language, there may be *multiple* structures that afford linear or functional prominence to the patient argument. The question then is how speakers choose between several sentence forms, particularly when there is more than one structure allowing prominent encoding of the patient. In Section 2.3.1, Pondok Tinggi was described as having multiple sentence forms that can be considered patient prominent. In this chapter, I report a sentence production experiment using the same paradigm as in Chapter 3 to look at the effects of animacy on sentence production in Pondok Tinggi. In particular, I wish to investigate whether the animacy manipulation affects the rate of all patient-prominent structures equally, or whether there are differences in how these structures relate to the animacy manipulation.

In the process of asking how speakers choose between more than two structures, a significant practical issue is raised. Namely, it is customary in sentence production studies that the outcome is analysed in a binary fashion, even when more than two structures are possible in the output. In order to analyse the speaker's output as a choice between *more than* two options, it turns out that it is necessary to re-evaluate the choice of statistical analysis. There appears to be a custom of analysing sentence production as binary in the literature. There is a strong possibility that this choice is driven by the difficulties involved in multi-categorical analysis, rather than being motivated on any conceptual or theoretical grounds. It is possible that the custom of treating sentence production as binary impacts or limits the kind of theoretical questions we can address. A secondary goal of this chapter is therefore to consider the advantages that may be gained by analysing sentence production data using a multi-categorical analysis.

In the following sections, I first present the aims of the sentence production experiment in Pondok Tinggi. I then review the approaches that have been taken in the analysis of sentence production data (chiefly from *simply describing* experiments), which treat the data as captured by one or more binary variables. I consider reasons why this approach may have been taken in previous studies, followed by some issues that are faced when using such an approach. Following this, I report the *simply describing* experiment on Pondok Tinggi, where I investigate the effect of animacy on structural choice when there is more than one felicitous patient-prominent structure available. In order to analyse this data, I opt for a statistical approach that permits multiple response categories. Following the presentation of results from this analysis, I then discuss the findings in light of the Accessibility account of grammatical encoding.

5.1 Structural choice in Pondok Tinggi

The consensus from previous research is that speakers more often choose forms that prioritise the patient when it is more animate than the agent. But how do speakers respond when there is more than one structural form that meets these requirements? Pondok Tinggi, an endangered language spoken in Kerinci, Indonesia, is a language that has more than one structure that affords prominence to the patient. The structures were described in Section 2.3.1, and are repeated below in (1) for reference.[1] Example (1a) illustrates the active, agent-initial structure (initial position in Pondok Tinggi being the main cue for grammatical function of subject); (1b) illustrates the di-passive, where patient is realised in initial (subject) position, with passive marking on the verb; (1c) illustrates the perfective, where the patient is realised in initial position and the verb carries a perfective marker; and finally (1d) illustrates the adversity passive using the verb form *kena(o)*, with optional specification with an additional verb (*timbok*).

(1) a. umpun kayau nimpok uto
stem.OBL wood.ABS ACT.crush.OBL car
'a tree trunk crushes a car'

b. uha di-empok [wot] kayau
people.ABS PASS-crush.OBL [by] wood.ABS
'a person is crushed by a tree'

c. uha ta-simbak bola
people.ABS PERF-touch.ABS ball
'a person is touched by a ball'

d. uha kenao [timbok] bola
people.ABS ADVPASS.ABS [hit.OBL] ball
'a person is adversely affected by a ball'

The experiment has the same design as the experiment in Chapter 3. Using the same experimental design as before enables us to more clearly navigate the analysis and interpretation of the multi-

[1] Note that the structures included here only include those where both referents can be non-pronominal.

categorical outcome. The data collection itself was carried out by a native speaker linguist, Ernanda (cf. Ernanda, 2017). A description of the language and the speaker community can be found in Section 2.3.1.

5.1.1 Research questions

Previous research has provided insight into the conditions under which speakers are more likely to produce patient-prominent structures. For example, we have seen that for a range of languages, patient-prominent structures are more likely when patients are more animate than agents. However, what remains unclear is how speakers respond when there is more than one felicitous patient-prominent structure. If more than one structure is available to encode the patient in a prominent functional role and/or sentence position, what is involved in deciding which of these structures is ultimately produced? The key aim of this study is to first assess the impact of animacy on choice between the multiple structures available in Pondok Tinggi. The goal is to then reflect on whether (and how) the outcomes can be explained under an information retrieval account of sentence formulation, such as the Accessibility Hypothesis.

This brings us to the practical challenge raised by taking this approach to the data. How can we analyse sentence production data when there are more than two structural categories? The problem here is that binary logistic regression only permits us to compare two categories; yet, we wish to understand the speaker's choice among multiple categories. The binary approach to analysing sentence production data was introduced in Chapter 3 (Section 3.3.2). Essentially, when we review previous research on structural choice in sentence production, it appears that languages with multiple structures have indeed been studied, but that the outcome variables in such experiments are typically treated as binary. In the following section, I revisit the topic of binary analysis of sentence production data in more detail. I first review the approaches of previous studies, before considering possible motivations for and limitations of the binary approach.

5.2 Review of the binary approach to sentence production data

The standard approach was introduced and exemplified by the *simply describing* experiment in Chapter 3. There, participants were asked to describe scenes involving a transitive action carried out by one referent (the agent) on another referent (the patient). The outcome variable in that experiment was the grammatical construction used by the participant. The participants' freely spoken responses were transcribed and then classified according to structure type. The aim was then to see how choice between structures was affected by the experimental condition. The coding and analysis procedure followed the standard approach in the literature, whereby the outcome of the data is treated as a categorical variable. That is, the dependent measure is a matter of grouping the responses, and the analysis concerns which group the responses fall into, and whether this is systematically affected by the experimental variable. For example, Bock's early studies (conducted in English) investigated effects of the relative accessibility of the referents being described (cf. Bock, 1986a, 1987). The outcome variable was whether the participant's response was active or passive; the analysis then concerned the rate of passive constructions and how this differed between conditions. As discussed in Chapter 3, earlier studies in this area tended to use techniques that were inappropriate for categorical data analysis, such as ANOVA; more recently, however, researchers have been using appropriate categorical data analysis techniques such as logistic regression, as well as incorporating random effects in order to model variation due to participants and items (as described in Chapter 3; cf. Jaeger, 2008). Whether using ANOVA or mixed effects binary logistic regression, it remains the case that the data is treated as binary. In other words, the analysis always compares the likelihood, or rate of choice, between *two* categories.

Crucially, a binary approach to analysis is typically taken even when there are more than two possible structures that are felicitous in a single context, such as Spanish (Prat-Sala & Branigan, 2000), Odawa (Christianson & Ferreira, 2005), Tzeltal (Norcliffe, Konopka, et al., 2015) and Japanese (Tanaka et al., 2011). A closer investigation reveals a number of ways in which researchers have approached multi-structural data with binary analysis techniques.

One approach is to aggregate all responses into two general categories. In Prat-Sala and Branigan (2000)'s *simply describing* experiment in Spanish, participants described the pictures with actives, passives and object-topicalisations. In that experiment, passives and object-topicalisations were aggregated into a 'non-canonical' category, complementing the 'canonical' category (i.e. active structures). The analysis concerned how the proportion of non-canonical structures was affected by the experimental condition (i.e. animacy).

However, such generalisation over the data impacts the theoretical insight we can gain from the data. With regard to Prat-Sala and Branigan (2000)'s Spanish data, the authors found a higher incidence of 'non-canonical' structures, but this still does not inform us about whether this is an effect on the likelihood of passivisation (targeting grammatical function assignment), on object-topicalisation (targeting linear ordering), or both (cf. Prat-Sala & Branigan, 2000:180). Ultimately, we are restricted to drawing conclusions regarding the effect of animacy on the rate of 'non-canonical' structures. In other words, this method provides only limited insight into how speakers navigate the available choices in such a language, and consequently offers little scope for generating more refined hypotheses.

Another approach is to analyse subsidiary variables separately. This kind of approach is found in Christianson and Ferreira, for their study in Odawa, an Algonquian language. Odawa displays a rather more complex array of verb form and word order possibilities than were found in Prat-Sala and Branigan's Spanish data. In order to analyse this complex dataset, the authors conducted an array of T-tests, each time focusing on a different variable within the dataset. Of course, using T-tests again represents the use of inappropriate techniques for categorical data; moreover, the multitude of T-tests leads to issues with Type I error (in other words, increasing the risk of erroneously rejecting the null hypothesis). A similar approach, using appropriate techniques for categorical data, is represented by the approach of Norcliffe, Konopka, et al. (2015), in an eye-tracked picture description study in Tzeltal. In that study, Tzeltal participants produced verb-initial actives, subject-initial actives, verb-initial passives and subject-initial passives (Norcliffe, Konopka, et al., 2015:8). Responses were coded as active or passive, and as verb-initial or subject-initial. Sentence form choice was analysed through two binary logistic regression models: firstly, the effect of the independent variable on the likelihood of passive vs. active (collapsed across verb position), and the effect of the independent variable on the likelihood of verb-initial vs. subject-initial. This approach can be summarised as treating grammatical function assignment and linear order as distinct, cross-cutting dimensions of interest. A similar approach is taken in the Japanese sentence recall study of Tanaka et al. (2011).

However, the approach of analysing subsidiary variables separately does not fully resolve the issues above. This is because the same issues apply when those subsidiary variables are not binary. In the example of Tzeltal, the authors grouped the data along two subsidiary variables, reflecting grammatical function assignment and linear ordering. This approach is warranted when the variables represent binary groupings: in the case of Tzeltal these were active/passive, and subject-initial/verb-initial, respectively. However, it is easy to imagine that in another language, the subsidiary variables themselves may be multi-categorical. For example, with regard to linear ordering, transitive picture descriptions could fall into three groups, namely verb-initial, subject-initial and object-initial. In other words, this approach does not fully circumvent the problems of the binary approach.

Even when a binary split is theoretically justified on the basis of the data, it can still lead us into problems involving spurious effects. This issue is described by Jaeger (2008), as follows. The binary classification of data is normally preceded by the exclusion of 'non-target' responses. These non-target responses are ones that involve errors, misinterpretations, or are simply considered irrelevant for the hypothesis. The problem arises when these 'non-target' or 'error' responses are not randomly distributed. To illustrate: imagine that we exclude 10% of data in condition A and 30% of data in condition B. After exclusions, a structure that originally accounted for 50% of data in condition A and 50% of data in condition B will then account for 55% of data in condition A and 70% in condition B. This then gives the false impression of an effect on the proportion of this structure between conditions. Note that, as discussed previously, in sentence production experiments often quite a sizeable amount of the data is excluded (Bock, 1996). This means that discrepancies in error rates between conditions can be on quite a large scale. As indicated by Jaeger (2008), the solution here would be to have the possibility of treating the error category as a third category on the outcome variable.

In summary, sentence production researchers have found ways to apply binary analysis techniques to gain understanding of how speakers select structural forms in languages where the possible array of structures is complex. However, there are a number of issues that arise when we seek to analyse this data in a binary form. The question then arises: why do researchers typically process the data into binary groupings, if there is more complexity in the data? For example, in a language where broadly four types of structure are produced, why do we not analyse the data as a choice between four types of structure? The choice of binary rather than multi-category analysis is not explicitly discussed in these studies. Despite this, we can nonetheless consider some key reasons as to why this may be the case.

On the one hand, it is possible that researchers adopt a binary analysis in order to ensure comparability with earlier studies. Since the earliest experimental studies of sentence production there has been a focus on active-passive choice (cf. Bock, 1982); this dichotomous perspective may have laid the foundations for subsequent research to pursue comparable groupings for other languages, even when those languages exhibit a range of choice beyond (or even instead of) active-passive alternations. Likewise, it is possible that researchers choose the binary analysis because the data appears to represent a series of binary choices; such as whether the role mapping should be active or passive, and *separately* whether the order of elements should be subject or verb initial (e.g. Norcliffe, Konopka, et al., 2015). This could be seen as a reflection of sentence production as proceeding through functional processing and subsequently positional processing (cf. the model discussed in Chapter 1). However, it is less clear how multiple analyses on different variables can be taken to model speaker choice if structure formulation is thought to occur at a single stage (Pickering & Ferreira, 2008).

On the other hand, it is possible that the binary analysis is not favoured for theoretical reasons, but for methodological reasons. Specifically, mixed effects logistic regression with multiple (i.e. >2) categorical outcomes is more complex than mixed effects binary logistic regression. Not only is the interpretation of the multi-categorical model less straightforward, but at a practical level, the multi-categorical analysis is more difficult to implement. The development of the R statistical programming language (R Core Team, 2017) has given rise to the popularity of packages that can fit mixed effects binary logistic regression models, such as *lme4* (D. Bates et al., 2015); at the time of writing these do not offer the possibility to analyse multiple (i.e. >2) unordered categories on the outcome variable. The issue seems to be down to the significant mathematical challenges that arise in estimating generalised linear mixed models with random effects terms (Agresti, Booth, Hobert, & Caffo, 2000; Breslow & Clayton, 1993; Hadfield, 2010).[2]

In considering these two possibilities, the lack of explicit discussion about the conceptual motivations for the binary approach leads us to consider that practical (methodological) limitations may play a large role. These practical limitations can, however, be resolved straightforwardly by using a Bayesian approach. The particular mathematical challenges mentioned above are moot when using a Bayesian framework, thanks to the powerful technique for estimating models by simulation that lies at the core of Bayesian modelling. In the first instance, the Bayesian approach may seem much more complex than classical frequentist models, but it has the advantage of being able to scale up straightforwardly to complex situations such as the multi-categorical mixed effects model. Moreover, the recent development of the Bayesian statistical modelling language Stan (Carpenter et al., 2017), and the possibility to run Stan-based models in R (R Core Team, 2017) has given rise to packages which permit a straightforward manner of fitting a multi-categorical mixed effects regression, such as the *brms* package (Bürkner, 2017).

In summary, cross-linguistic research has found that patient-prominent structures are more likely when patients are more animate than agents, but it is currently unclear how to account for the choice between several structures that permit functional and/or positional prominence of the patient. In aiming to account for the choice between multiple structures, we come up against the issue that previous studies have tended to code and analyse *simply describing* data in dichotomous form. Although for some languages a dichotomous analysis procedure may well be warranted, the fact that it is applied across the board may stem more from practical limitations than theoretical motivations. In my aim to understand sentence form choice in Pondok Tinggi, I therefore opt to implement an analysis where the outcome variable may be categorical but also polytomous (i.e. multi-categorical logistic regression). For this, I use a Bayesian implementation in R.

5.3 The current study

5.3.1 Experimental design and hypotheses

The experiment employs the same paradigm as reported in Chapter 3, where participants describe transitive scenes depicting two entities. In half of the pictures the two entities are both inanimate (e.g. lightning hitting a house), while the other half of the pictures show an inanimate agent acting on an animate patient (e.g. lightning hitting a man). The design and stimuli were identical, with the exception of one picture which was replaced in this experiment, because a number of participants had found it difficult to interpret in the previous experiment. As in the Chapter 3 experiment, actions were depicted as having just happened, and presented with the prompt question ‘What has happened?’,

[2] In particular, these challenges become more difficult to overcome when models become more complex with more dimensions, such as are introduced by multiple random effects terms. The problem here arises from the need to marginalise out the random effects, arriving at a marginal likelihood function from which model parameters can be estimated. Techniques that are feasible for simple GLMs are no longer robust when models become more complex (Agresti et al., 2000).

which in the case of Pondok Tinggi was *apo nge tajadoi?*. For further description of the design and method, see Section 3.2.

If animacy manipulations have a uniform effect on the likelihood of patients achieving linear and/or functional prominence, then we expect the effect on patient-prominent structures should be more or less consistent. That is, if the realisation of patients in prominent functions and positions is to be attributed to accessibility, it is logical to expect that all patient prominent forms should increase in likelihood given a more accessible patient. However, given that there are different forms available in the language, it seems possible that the pattern of choice will differ across structures. If this is the case, the question will then arise of how and whether it is possible to account for this pattern of results under an information retrieval account of sentence form variation.

5.3.2 Participants and procedure

Fifty native speakers of Pondok Tinggi resident in Pondok Tinggi village, Kerinci (Indonesia), participated in the experiment (25 male, $M^{age} = 41$, range 18–70 years). Following task completion, participants were offered a small financial compensation for their time. The experimental sessions were carried out by Ernanda. Participant recruitment was further assisted by members of the speaker community. Pondok Tinggi is again a language that is not typically written. Therefore, as in Chapter 3, the prompt question and the instructions were audio recorded in the language of the experiment by a native speaker. All participants were tested in a quiet room of a house, with only participant and tester present in the room. The experiment was displayed using Microsoft PowerPoint on a laptop. The responses were recorded on a digital voice recorder and transcribed before coding.

5.3.3 Coding and preprocessing

Once transcribed, each response was coded according to the structure used by the speaker. The classification of structures was based on the description of the language in Ernanda (2017), the relevant parts of which are explored in more detail in Section 2.3.1. In accordance with previous sentence production studies, responses were also screened and coded for whether they met a set of acceptability criteria. In example (2) below, I provide examples of responses from the dataset to illustrate each of the criteria.

To be ‘acceptable’, a response had to be a complete sentence including a verbal phrase describing the action of the inanimate agent on the patient. This meant that reciprocal sentences were excluded (2a). This verb phrase had to occur in the first full clause of the response, so that descriptions involving subordination, or restarted responses of the type in (2b), were not acceptable. Responses where the agency of the action was attributed to the patient or an unseen third party, or where the inanimate agent was realised as an instrument or location were also excluded (2c). Entities did not have to be described with specific target terms, but responses were excluded if animacy was misconstrued, or inferred such as in the case of the agent in (2d).

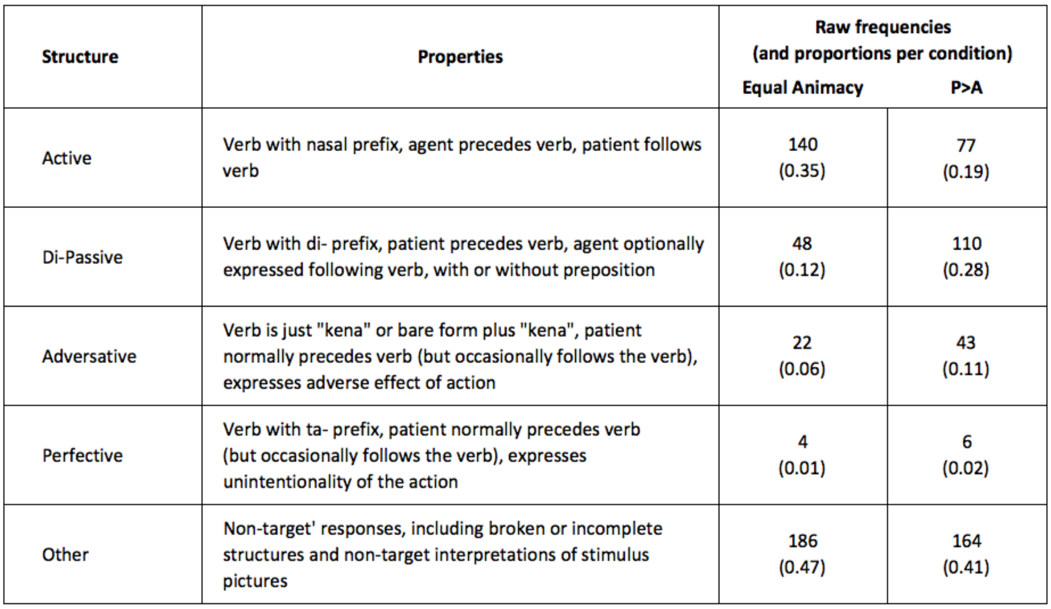
(2) a. kangkao ba-rebeuk ngan kumba
frog.ABS RECP-compete.for.ABS with bumblebee.ABS
‘A frog fights with a bumblebee’ [A picture where frog catches a spider]
b. uha jateuh ... uto numbur uha
people.ABS fall.ABS ... car ACT.hit people.ABS

‘A person falls ... a car hits a person’ [A picture where a car hits a woman]

c. uha jateuh dari ateh uto
people.ABS fall.ABS from up car
‘A person falls from a car’ [A picture where a tractor runs over a person]

d. uha nimbok kepa
people.ABS ACT.shoot.OBL ship.ABS
‘A person shoots a ship’ [A picture where a fighter plane shoots a ship]

Five structure categories accounted for all acceptable responses. Examples of responses of each of these types can be found in Section 2.3.1. The responses that did not meet the criteria above were coded as ‘Other’ in Table 5.1. Table 5.1 gives the raw frequencies of structures (total responses n = 800), and also the proportions these counts represent within each condition (n = 400).



Structure	Properties	Raw frequencies (and proportions per condition) Equal Animacy	P>A
Active	Verb with nasal prefix, agent precedes verb, patient follows verb	140 (0.35)	77 (0.19)
Di-Passive	Verb with di- prefix, patient precedes verb, agent optionally expressed following verb, with or without preposition	48 (0.12)	110 (0.28)
Adversative	Verb is just "kena" or bare form plus "kena", patient normally precedes verb (but occasionally follows the verb), expresses adverse effect of action	22 (0.06)	43 (0.11)
Perfective	Verb with ta- prefix, patient normally precedes verb (but occasionally follows the verb), expresses unintentionality of the action	4 (0.01)	6 (0.02)
Other	Non-target' responses, including broken or incomplete structures and non-target interpretations of stimulus pictures	186 (0.47)	164 (0.41)

Figure 5.1: Descriptions and counts of structural categories in the dataset. Examples of each type can be found in Section 2.3.1.

It is immediately striking that there are many responses in the error category *Other*, that is, the responses that did not meet the criteria for ‘acceptability’. Although seemingly very high, these amounts are rather typical for sentence production experiments, as discussed in Chapters 1 and 3. The *Other* category is not evenly split across conditions: it makes up 47% of the Equal Animacy condition and 41% of the Patient>Agent condition. In light of Jaeger’s suggestions, however, and unlike most previous sentence production studies, this category *is* to be included in the analysis. This allows us to not only assess how the likelihood of error is affected by experimental condition, but moreover means that we avoid the kind of spurious effects as discussed in Section 5.2.

5.3.4 Statistical approach

Multi-categorical analysis

As described above, the dataset in question has one binary independent variable, namely pictures with inanimate agent and inanimate patient (Equal Animacy condition) versus pictures with inanimate agent and animate patient (Patient More Animate condition). The dependent or outcome variable is the sentence form that speakers use to describe the picture. Based on the range of structures produced by the participants, this categorical variable has four levels. Additionally, as described above, we include *all* the data (i.e. all responses to critical trials) in the model, including the error responses in the category 'Other'. This results in five levels on the categorical response variable (Active, Di-Passive, Adversative, Perfective and Other).

The multiple outcome category logistic regression model is sometimes referred to as multinomial, polytomous or softmax regression; however, here I will adopt the term 'multi-category' logistic regression model (cf. Kruschke, 2015). The multi-category logistic regression can be considered an extension of binary logistic regression. As illustrated in Chapter 3, binary logistic regression helps us understand the likelihood of choosing one category versus another, and how this is affected by the independent variable. So, for example, how likely we are to produce passive instead of active, and how this is affected by the animacy balance of the two referents. By extension, the multi-category logistic regression helps us understand choice between *more than two* categories, and *how that choice is affected* by the independent variable.

In terms of interpreting the model, there is an added complexity when we have multiple categories: how exactly do we compare choices? (In simple terms, choice between what and what?) With two categories it is straightforward because the categories simply contrast with each other. But with multiple categories, there are multiple contrasts possible. The standard way to deal with this in categorical logistic regression is to select one of the outcome categories as a so-called 'reference category'. Then we assess each remaining outcome category relative to that reference category. Returning to the example of Spanish, above: alongside "how probable is it that a speaker produces passive rather than active?" we also answer "how probable is it that a speaker produces an object-topicalised structure rather than a canonical active?". Comparing each category against a reference category is termed *baseline category logit* (Agresti, 2007). Of course, what we are really interested in is how these probabilities, these choices, are affected by the independent variable. The question is then, how does the independent variable affect the probability of choosing category x rather than the reference category? So, in our example of the sentence production experiment: how does animacy balance affect the probability that a speaker chooses passive rather than active, and how does it affect the probability of choosing object-topicalised structure rather than active?

Modelling procedure

In order to fit the multi-category response model, I have opted to use the *brms* package in R (Bürkner, 2017; R Core Team, 2017). This package fits Bayesian regression models, including multi-category logistic regression. The *brms* package uses the Bayesian modelling language Stan to fit models (Carpenter et al., 2017).[3]

[3] However, regression models in *brms* are built completely within R and the syntax is based on that of *lme4*, providing maximal comparability with modelling approaches used in Chapters 3 and 4.

The use of a Bayesian rather than frequentist approach means that this study departs from the majority of previous sentence production studies. This means that there are some important differences between Bayesian and frequentist approaches to be taken into consideration when fitting the model and inspecting the output. For the purposes of this study, the key issues concern the use of *MCMC sampling* to estimate the models, the issue of *priors* and lastly how we can interpret the *parameter estimates*. Here, in order to aid the presentation of the analysis, I will now briefly touch upon these three points.

The method at the heart of Bayesian modelling is Markov Chain Monte Carlo sampling, known as *MCMC*. In situations like multi-category mixed logistic regression, where mathematical calculation is not feasible, MCMC provides a way to resolve this by simulating the distribution of interest. In general terms, the MCMC process involves taking random steps around the range of possible parameter values (these steps are known as 'samples'). Each time, the sampler evaluates whether the new values are more or less appropriate than the previous ones. After doing this several thousand times, we end up with a distribution of possible values (posterior distribution) for each parameter. The sampler favours more appropriate parameter values, with the result that the distribution will be at its most dense around the most favoured parameter values. The 'journey' taken by the sampler is known as a chain.

In this analysis, I opt to run four chains. Running a few chains gives us important insight into whether the sampler converges on the same results every time. It is also necessary to specify how many samples we should have per chain (i.e. the number of iterations). It is important to draw enough samples for the model to gain sufficient information about the distribution of the parameters. In this model, 3000 samples are drawn per chain. With four chains, this amounts to a total of 12000 samples. Lastly, it is customary to discard the first part of each chain, considering it a "warm-up" (sometimes termed "burn-in"). The reason for this is that it takes some time for the sampler to settle on the area of the most appropriate values. Here, I will discard the first 1000 samples of each chain (determined in the model by *warmup*). This results in 8000 samples total, when all chains are combined.[4]

The issue of *priors* relates to the fact that the Bayesian approach is concerned with the plausibility, or credibility, of parameter values. We can see the Bayesian modelling procedure as a process of assessing the credibility of different possible parameter values, based on the data that we have. This procedure requires that we state also our 'prior beliefs' about the possible parameter values. This is so that the credibility of parameters can then be (re)assessed on the basis of the data. Note, however, that it is possible to use 'general' priors which express that we do not have any *a priori* belief about which values are the most plausible. In *brms*, general priors as described above are implemented by default. In the current analysis, *brms* will be used with default priors. Information about default priors in *brms* can be found in the overview vignette accompanying the *brms* package (Bürkner, 2017).

When obtaining *parameter estimates*, in a frequentist setting we come up with a singular point estimate – our best estimate. We are not so concerned with the degree to which this is a *plausible* parameter value. Significance of the estimate is determined by how improbable it is under the null hypothesis (i.e. giving us a p value). However, the Bayesian approach contrasts with this, because it *does* care about the plausibility of parameter values. In a Bayesian approach, therefore, the outcome of model estimation is not just one point estimate, but an expression of how credible this estimate is. Furthermore, Bayesian models allow us to assess the credibility of a range of possible values of

[4] Note also that in the current analysis, setting adapt_delta to 0.97 was necessary to achieve a model without divergent transitions. The term *adapt_delta* adjusts the size of the steps taken by the MCMC sampler. Although it is not obligatory to specify this (because it has a default value, namely 0.8), it may be adjusted when the model fitting procedure throws errors regarding divergent transitions (Bürkner, 2017; Stan Development Team, 2017b).

a parameter (i.e. the distribution that forms the output of the MCMC sampling process, explained above). In short, by observing the shape of the posterior distributions, we can assess whether other values are similarly credible, and so assess the amount of (un)certainty there is in the estimate. We can describe this certainty using the 95% Credible Interval (also termed HDI, or Highest Density Interval). This is the area of the posterior distribution that contains most (i.e. 95%) of the probability density, or ‘credibility’. If the posterior distribution is flat, the credible interval will be wide. However, if a small range of values have high credibility, the credible interval will be narrow. The Credible Interval is reminiscent of the frequentist ‘confidence interval’. Even though they represent fundamentally different information, we can use them in a similar way to infer about the estimates (cf. Sorensen, Hohenstein, & Vasishth, 2016): if the 95% Credible Interval does not cross zero, it follows that the parameter estimate is credibly positive or credibly negative. Also, the narrower the Credible Interval, the more confidence we have in the parameter estimate, because credibility is more focused around the estimate.

5.4 Results

The structural choice STRUCTURE is modelled as a function of the type of picture (Equal Animacy or Patient More Animate, i.e. CONDITION). In this model we have random intercepts and random slopes for the effect of condition both within subjects (CONDITION|SUBJECT) and within items (CONDITION|PAIR). This is therefore a maximal model (cf. Barr, Levy, Scheepers, & Tily, 2013). The model summary is provided in Table 5.1. Note that the random effects terms are listed under Group-Level Effects, and the fixed effects terms are listed under Population-Level Effects (regarding this nomenclature, see Bürkner, 2017).

To generate a range of summary and diagnostic plots I have used the *ggmcmc* package (Fernández-i-Marín, 2016). The trace plots (the visualisations of the chains) and the posterior distributions (the distributions that were reached) are displayed in density plots, presented in Figures 5.2 to 5.7.

Diagnostics

In terms of diagnostics, there are several pieces of information we can use to assess how well the model has performed. Firstly, consider the trace plots in Figures 5.2 and 5.3. It can be seen that the chains have settled around the area of certain values, because the chains are horizontal, showing multiple samples around the same area. In other words, the chains appear to have reached equilibrium. Secondly, as noted above, multiple chains were run in order to check that the sampling process converges on similar values even when run several times. Upon consulting the plot, we can see that this is the case, as the chains appear to overlay well (they are “mixing” well); in other words, the chains appear to reach roughly the same distributions. Thirdly, we can also consider the R-hat diagnostic. R-hat values for each model parameter can be seen in the summary output. R-hat values should be as near to 1 as possible (ideally ± 0.1) (Gelman & Rubin, 1992; Sorensen et al., 2016). Here we can see that this is the case.

Model output and interpretation

As described above, from the posterior distributions we obtain a point estimate, such as the mean of the distribution, which is taken as the parameter estimate. These parameter estimates are provided

```
Family: categorical(logit)
Formula: STRUCTURE ∼ CONDITION + (CONDITION — SUBJECT) + (CONDITION — PAIR)
Data: d (Number of observations: 800)
Samples: 4 chains, each with iter = 3000; warmup = 1000; thin = 1;
total post-warmup samples = 8000
ICs: LOO = 1328.19; WAIC = 1282; R2 = NA

Priors:
L ∼ lkj_corr_cholesky(1)
sd ∼ student_t(3, 0, 10)

Group-Level Effects:

∼PAIR (Number of levels: 17)
                                                              Estimate  Est.Error  l-95% CI  u-95% CI  Eff.Sample  Rhat
sd(AdvPassive_Intercept)                                          3.29       1.21      1.59      6.27        3341  1.00
sd(AdvPassive_CONDITIONPMoreAnimate)                              3.40       1.84      0.75      7.89        2592  1.00
sd(DiPassive_Intercept)                                           2.08       0.57      1.23      3.40        2837  1.00
sd(DiPassive_CONDITIONPMoreAnimate)                               0.94       0.61      0.04      2.38        1796  1.00
sd(Perfective_Intercept)                                          9.21       4.80      2.96     21.29        2877  1.00
sd(Perfective_CONDITIONPMoreAnimate)                              9.39       6.22      1.15     24.64        2594  1.00
sd(Other_Intercept)                                               2.41       0.59      1.54      3.85        2732  1.00
sd(Other_CONDITIONPMoreAnimate)                                   1.44       0.50      0.66      2.59        3367  1.00
cor(AdvPassive_Intercept,AdvPassive_CONDITIONPMoreAnimate)        0.61       0.34     -0.27      0.98        4353  1.00
cor(DiPassive_Intercept,DiPassive_CONDITIONPMoreAnimate)         -0.48       0.42     -0.97      0.66        3990  1.00
cor(Perfective_Intercept,Perfective_CONDITIONPMoreAnimate)       -0.26       0.45     -0.92      0.68        5405  1.00
cor(Other_Intercept,Other_CONDITIONPMoreAnimate)                  0.03       0.34     -0.63      0.66        4792  1.00

∼SUBJECT (Number of levels: 50)
                                                              Estimate  Est.Error  l-95% CI  u-95% CI  Eff.Sample  Rhat
sd(AdvPassive_Intercept)                                          2.48       0.83      1.16      4.37        2217  1.00
sd(AdvPassive_CONDITIONPMoreAnimate)                              1.62       0.97      0.10      3.79        1240  1.01
sd(DiPassive_Intercept)                                           1.46       0.36      0.84      2.24        2811  1.00
sd(DiPassive_CONDITIONPMoreAnimate)                               0.71       0.45      0.04      1.68        1123  1.00
sd(Perfective_Intercept)                                          8.32       4.93      2.15     20.65        2873  1.00
sd(Perfective_CONDITIONPMoreAnimate)                              8.25       5.32      0.90     20.81        2102  1.00
sd(Other_Intercept)                                               1.30       0.25      0.84      1.84        3239  1.00
sd(Other_CONDITIONPMoreAnimate)                                   0.56       0.37      0.03      1.36        1597  1.00
cor(AdvPassive_Intercept,AdvPassive_CONDITIONPMoreAnimate)       -0.13       0.51     -0.90      0.89        3039  1.00
cor(DiPassive_Intercept,DiPassive_CONDITIONPMoreAnimate)         -0.07       0.50     -0.89      0.91        3460  1.00
cor(Perfective_Intercept,Perfective_CONDITIONPMoreAnimate)       -0.22       0.47     -0.94      0.78        3594  1.00
cor(Other_Intercept,Other_CONDITIONPMoreAnimate)                 -0.21       0.48     -0.93      0.83        4473  1.00

Population-Level Effects:
                                                              Estimate  Est.Error  l-95% CI  u-95% CI  Eff.Sample  Rhat
AdvPassive_Intercept                                             -4.88       1.48     -8.31     -2.51        2990  1.00
DiPassive_Intercept                                              -2.21       0.72     -3.77     -0.92        2281  1.00
Perfective_Intercept                                            -23.98      12.50    -54.46     -8.06        2431  1.00
Other_Intercept                                                   0.54       0.67     -0.75      1.88        1571  1.00
AdvPassive_CONDITIONPMoreAnimate                                  0.28       2.09     -4.51      3.64        2106  1.00
DiPassive_CONDITIONPMoreAnimate                                   2.47       0.54      1.49      3.64        3008  1.00
Perfective_CONDITIONPMoreAnimate                                 -0.42      12.58    -28.17     23.09        3379  1.00
Other_CONDITIONPMoreAnimate                                       0.87       0.51     -0.09      1.96        4338  1.00

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at convergence, Rhat = 1).
```

Table 5.1: Summary of *brms* multi-category regression analysis, modelling the choice among the five structural categories given in Figure 5.1, and how this choice is affected by the animacy condition (equal animacy, or patient more animate than agent). In this summary, group-level effects correspond to random effects, while the population-level effects correspond to fixed effects.

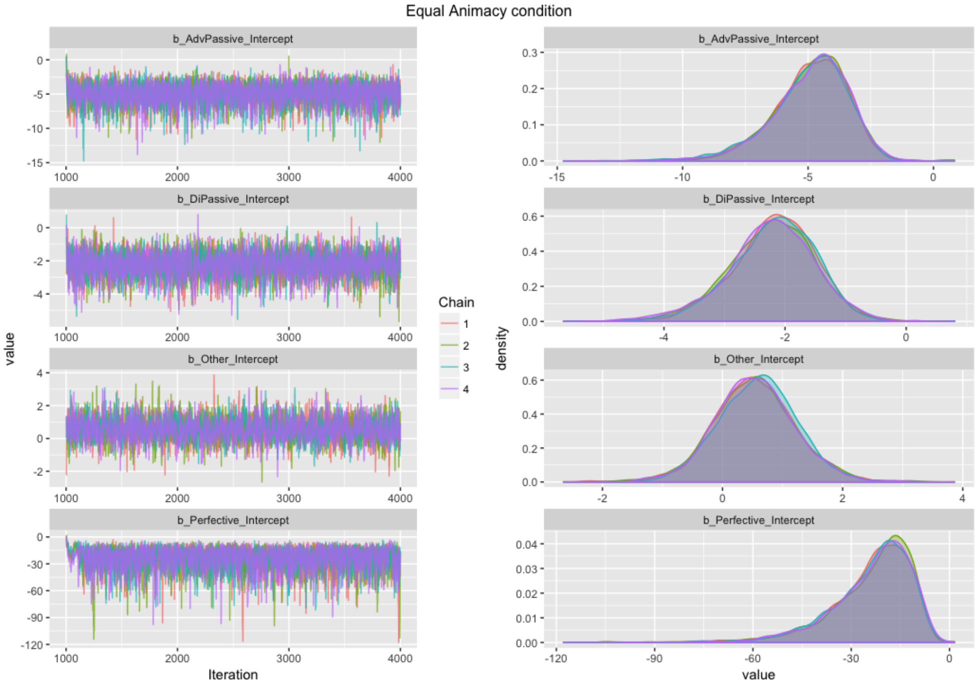


Figure 5.2: Trace and density plots for Equal Animacy parameters.

in the model summary output in Figure 5.1 (along with the bounds of the 95% Credible Intervals). Figure 5.4 visualises the fixed effects parameter estimates with their 95% Credible Intervals.[5]

The fixed effects are listed under ‘Population-Level Effects’ in the model summary. A plot of parameter estimates for the fixed effects, with 95% CIs, is shown in Figure 5.4. Recall that the baseline condition here (i.e. represented by the Intercept term) is the Equal Animacy condition. In addition, one category on the outcome variable is chosen as reference category; namely, Active. The remaining categories are compared with this category one by one.

It is important to take care when interpreting the model coefficients: in particular, the estimates for the Patient More Animate condition do not represent the likelihood of choosing those structures when Patient More Animate, but rather, they inform us about *how the probability of choosing that structure differs between the conditions*. In addition, the coefficients here are expressed in log odds. It is more meaningful to convert this into the odds ratio for each estimate. We can do this by exponentiating the coefficient. The converted coefficient then tells us, for each unit increase in the independent variable (i.e. a change from Equal Animacy to Patient More Animate), how much change there is in the odds of a given structure being chosen. When we interpret odds ratios, we are interested in whether the value is more or less than 1. Values above 1 mean that there is an increase in the odds and values below 1 mean there is a decrease in the odds. A more intuitive way to interpret this is to convert it to

[5] This plot is created using the *Shinystan* package in R (Stan Development Team, 2017a).

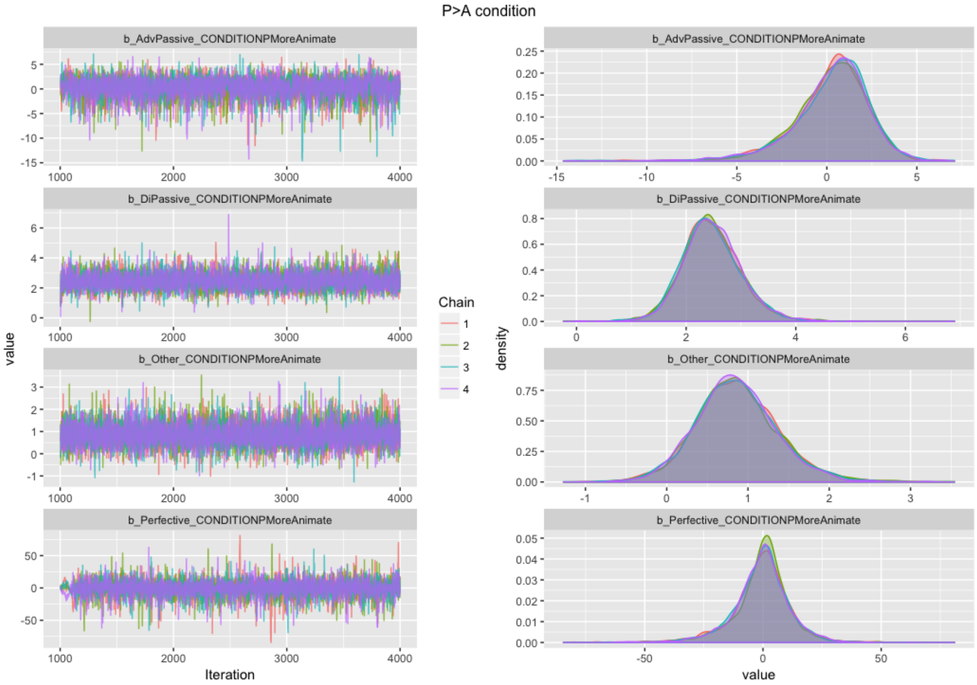


Figure 5.3: Trace and density plots for Patient More Animate parameters.

a percentage that expresses the change in odds (so, using the formula $(x - 1) * 100$, where x is the odds ratio value).

Firstly, we can examine the intercept terms. Often, when we interpret the output of regression models, we are not interested in the intercepts. However in the case of the unordered categorical regression they are informative: these estimates give us information about the odds of choosing a structure apart from Active when animacy is equal (i.e. odds of choosing each non-reference category in the baseline condition).

- *AdvPassive_Intercept* is the parameter estimate for Adversative Passive relative to Active (taken as reference category) for the Equal Animacy stimulus pictures (the baseline condition). The estimate for preferring Adversative Passive to Active is -4.88, i.e. the preference for Adversative is lower than that for Active. The odds ratio is less than 0.01; meaning that the odds of choosing Adversative rather than Active in the Equal Animacy condition are 99% less. The 95% credible interval (CI) of the logit estimate does not cross zero, indicating a significant estimate.

- *DiPassive_intercept* is the logit estimate for producing Di-Passive relative to Active in the Equal Animacy condition. The logit for choosing Di-Passive rather than Active is -2.21. This means odds of 0.1 of choosing Passive than Active when animacy is equal; the odds are 90% lower. The CI does not cross zero.

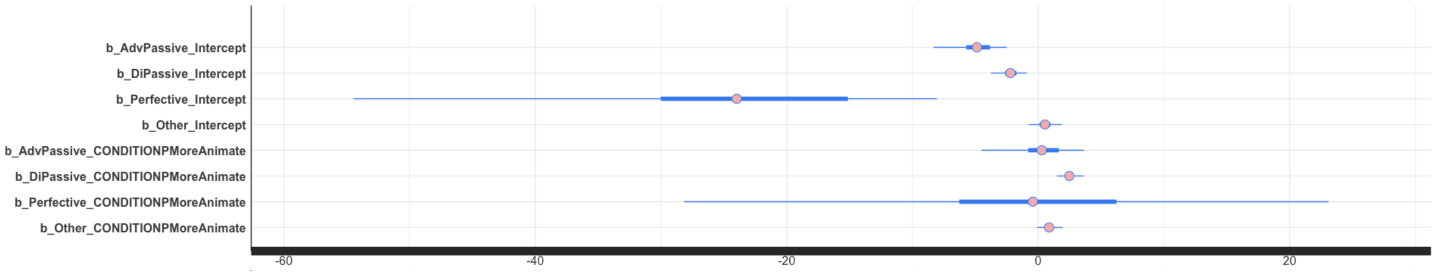


Figure 5.4: Parameter estimates with 95% credible intervals. Bolded sections of the lines indicate 50% credible intervals.

- *Perfective_intercept* is the logit estimate for Perfective relative to active for the Equal Animacy condition. The estimate for choosing Perfective over Active is -23.98, i.e. the odds of choosing Perfective over Active when animacy is equal is 0.0001 (99.99% lower). The CI does not cross zero; however, it is very wide, indicating considerable uncertainty about the estimate.
- *Other_intercept* is the estimate for the *Other* response category relative to Active for Equal Animacy pictures. The logit for producing an 'error' response rather than an acceptable Active structure is 0.54, i.e. the odds of Other rather than an acceptable Active are 1.72. In terms of percentage, this is a 72% increase in odds. However, note that the 95% CI crosses zero.

Now let us consider the estimates for each category when Patient More Animate is compared to Equal Animacy. When interpreting these estimates, it is useful to consider them as addressing the question "how does the likelihood of choosing this structure rather than Active change when patient is *more* animate, rather than equally animate?" Another way to think of this is that it informs us how the change in animacy of the patient affects the odds of choosing the different structures – i.e. the odds we have just seen in the previous paragraph.

- *AdvPassive_CONDITIONPMoreAnimate* is the logit estimate comparing Patient More Animate to Equal Animacy for Adversative Passive relative to Active. The estimate here is 0.28, which translates to odds ratio of 1.32. This means that the odds of choosing Adversative rather than Active increase by 1.32, or 32%, when Patient becomes more animate than Agent. However, note that the CI of the logit estimate falls across zero.
- *DiPassive_CONDITIONPMoreAnimate* is the logit estimate comparing Patient More Animate to Equal Animacy for Di-Passive relative to Active. The estimate is 2.47. In terms of odds, patient being more animate increases the odds of choosing Di-Passive over active by 11.84, i.e. by 1084%. The CI does not cross zero for this estimate. In summary, this indicates that when we compare the Patient More Animate condition with Equal Animacy, there is a stark difference in the probability of producing Di-Passive versus Active.
- *Perfective_CONDITIONPMoreAnimate* is the logit estimate comparing Patient More Animate to Equal Animacy for Perfective vs. Active. The estimate is -0.42, relating to an odds ratio of 0.65. That is, the odds of Perfective vs. Active reduce by 35% when we move to the Patient More Animate condition. However, this time the CI of the logit estimate crosses zero.
- *other_CONDITIONPMoreAnimate* is the logit estimate comparing Patient More Animate to Equal Animacy for the 'Other' response category relative to Active. The estimate here is 0.87; the odds ratio is then 2.38. This translates to a 138% increase in the odds of producing an Other response compared to the Equal Animacy condition. Again, however, the CI of the logit estimate crosses zero.

In summary, the intercept terms indicate that when the animacy of the two characters in the picture is equal, participants are much less likely to choose Adversative, Di-Passive and Perfective than Active structures to describe it. The CIs for these logit estimates do not cross zero, indicating strong evidence that the effect is indeed in the negative direction. The rate of Other structures is a little higher than the rate of Active structures, but not significantly so, as the CI crosses zero.

Concerning how this pattern of preferences changes with animacy condition, we see that when the patient is more animate than the agent, the choice between Di-Passive and Active is affected. The odds of choosing Di-Passive over Active are increased by almost 12 times in the Patient More Animate condition. However, it appears that the choice of Adversative relative to Active and Perfective relative to Active does *not* change notably between the conditions. Nor does the rate of producing a non-target or Other response. In other words, the preferences between Active and Di-Passive are strongly affected by condition, but preferences between Active and each of the other categories (Adversative, Perfective and Other) are not.

Predicted probabilities

A limitation of this analysis is that it is not immediately clear what the relative probabilities of outcome categories are within a condition – all that is known is how the odds differ. Moreover, in the Patient More Animate condition we only gain information about how preferences change *compared to the preferences when animacy is equal*. For example, we know how the odds of choosing Passive rather than Active *changes* when patient is more animate than agent, but we do not know what the basic probabilities of choosing Actives or Passives are per condition. Although we could refer to the proportions shown in Figure 5.1, we can also obtain the probabilities as predicted by the model (Agresti, 2007).

Comparing the predicted probabilities of the model with the proportions in the dataset can also inform us about how the model fits the data. To do so, we can extract the mean predicted probability of each outcome category, and compare these probabilities between conditions, as shown in Table 5.2. This can then be compared to the actual proportions (as shown in Figure 5.1). Overall, it can be seen that the predicted probabilities match up with the actual proportions of the responses as seen in Figure 5.1. This suggests that the model is closely reflecting the proportions found in the data.

	Condition	
Structure	Equal Animacy	Patient More Animate
Active	0.35	0.19
AdvPassive	0.05	0.11
DiPassive	0.12	0.28
Perfective	0.01	0.02
Other	0.46	0.41

Table 5.2: Mean probabilities of structural choice predicted by the model. Probabilities are relative within each condition, i.e. each condition sums to 1 (note that values here are rounded).

It is informative to compare these (predicted) probabilities with the odds ratios derived from the model parameter estimates in the previous section. This way we can see whether numerical differences in proportions of structures are – or are not – corroborated by changes in structural preferences (i.e. odds ratios). As before, we see that in both categories the model predicts a high rate of error or Other responses. However, recall that according to the parameter estimates of the model, the odds of producing an error response is not significantly greater than producing an acceptable Active structure in the Equal Animacy condition; nor does this change when moving to the Patient More Animate condition. Considering Active versus Di-Passive, the predicted probability of Active is higher in Equal

Animacy, but the predicted probability of Di-Passive is higher in Patient More Animate. This pattern is corroborated by the finding from the model that the odds of choosing Di-Passive rather than Active is strongly affected by the animacy condition. There is a higher predicted probability of Adversative Passive in Patient More Animate than Equal Animacy, but we know from the odds ratios (i.e. the model coefficients) that this change does not relate to a meaningful difference in the preference for Adversative rather than Active.

Perfective has a very low probability overall; the model parameter estimates also indicate that it is much less likely than Active to be produced in the Equal Animacy condition and that this situation is not different in the Patient More Animate condition. However, the 95% CIs for the Perfective category are extremely wide (cf. Table 5.1); the spread of the posterior distributions is also clearly visible on the density plots in Figures 5.2 and 5.3. This indicates a great deal of uncertainty in the estimates. This is most likely due to the extremely low number of Perfective responses overall, making it difficult to estimate parameters.

Random effects

Parameter estimates for the random effects are indicated in 'Group-Level Effects' in the model summary in Table 5.1. The first block of random effects terms concerns variability among picture-pairs (CONDITION|PAIR), while the second block concerns variability among participants (CONDITION|SUBJECT). Trace and density plots for these estimates are found in Figures 5.6, 5.5 and 5.7.

The *sd* terms in each block reflect the variability – standard deviation – in the estimates of intercept and slopes. By examining these, we can assess the degree of variability in how participants responded (subject random effects), or important differences among picture-pairs in terms of the proportions of structures they elicited (item random effects). Again in both subject and item random effects, the Perfective estimates stand out. The estimates of intercept and slope terms for both participants and picture-pairs are higher than for other structures. Additionally the estimates have highly spread out posterior distributions as can be seen in Figures 5.5 and 5.6.

The *cor* terms indicate the correlations between intercepts and slopes. Most of the correlation estimates are negative, suggesting that as intercepts increase, the slope decreases. However, the CIs for all correlation estimates cross zero. We can also take a look at the trace and density plots for these correlations in 5.7. The posterior distributions are spread between -1 and 1, indicating a large degree of uncertainty about the estimates.

5.5 Discussion

In this experiment, I aimed to investigate the effects of animacy of agent and patient on the choice among a range of transitive structures in Pondok Tinggi. The results provide evidence that the relative animacy of patient and agent does affect structural choice in Pondok Tinggi. In the Equal Animacy condition, the Active structure is most likely to be chosen over each of the other possible structures (with the caveat that there is a high rate of error responses). When patient is more animate than agent, participants are more likely to choose Di-Passives than Actives. However, the choices between Adversative and Active, and Perfective and Active, do not vary as a function of the animacy manipulation. On the one hand, the result that Di-Passives are more preferred in the Patient More

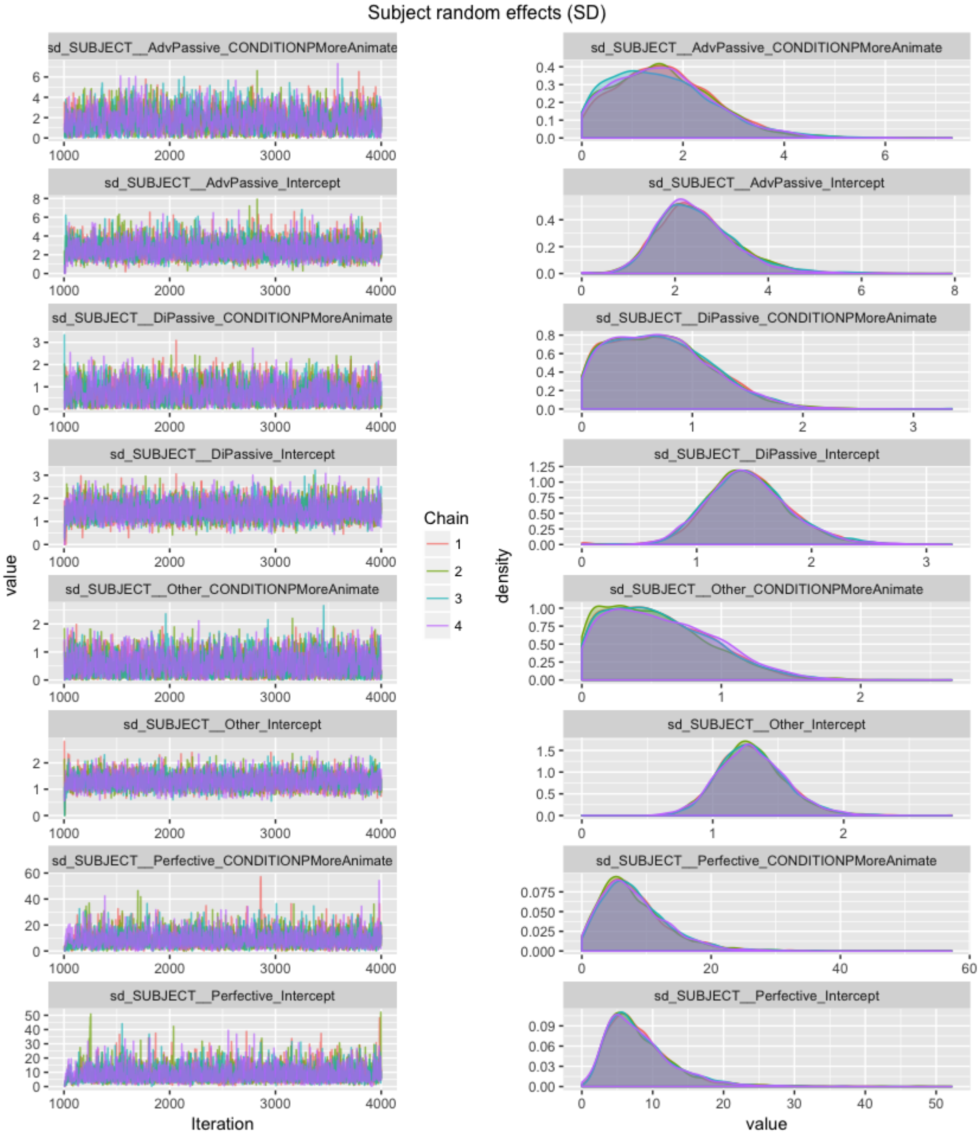


Figure 5.5: Trace and density plots for by-subject random effects.

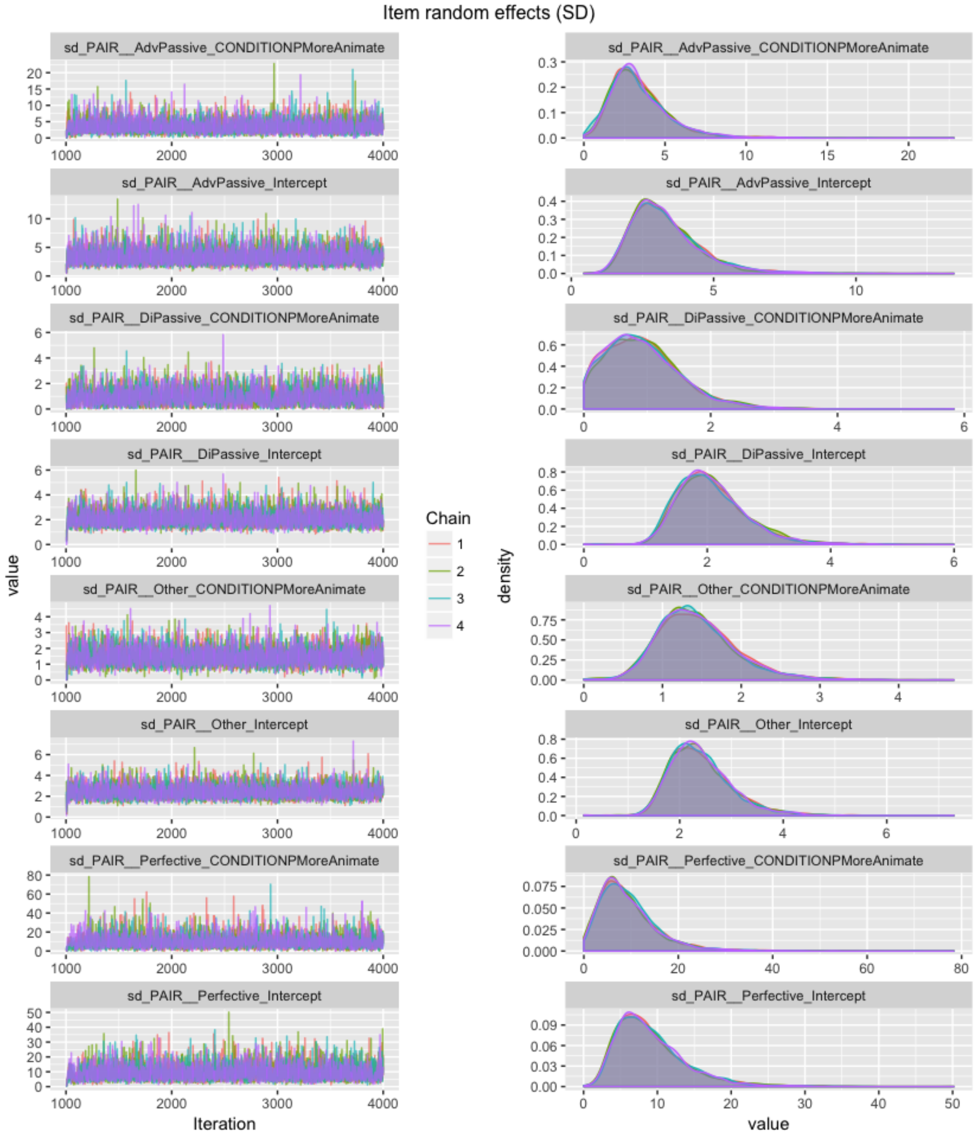


Figure 5.6: Trace and density plots for by-item fixed effects.

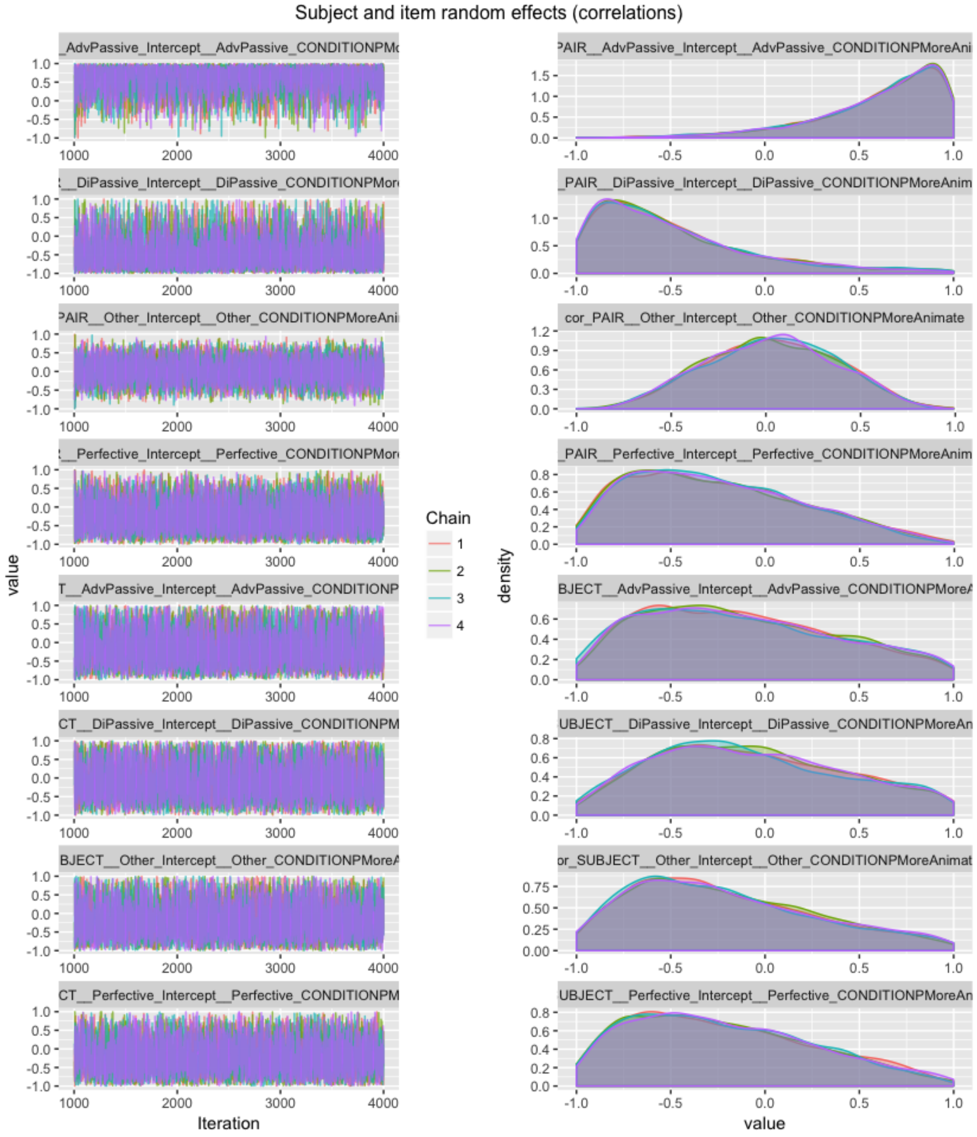


Figure 5.7: Trace and density plots for random effects correlations.

Animate condition echoes cross-linguistic findings that higher animacy of patient relative to agent triggers passivisation, in the sense of Siewierska (2013) (i.e. the assignment of patient to the grammatical role typically occupied by the agent and a concurrent marking of the verb). On the other hand, the analysis does not provide statistical support for the hypothesis that there is a higher probability of choosing *any* patient-prominent structure rather than Active. In other words, although there is a higher probability of patient-prominent structures in the Patient More Animate condition, not every structure that allows for patient prominence is affected.

This finding poses a challenge for accounts based purely on information retrieval. If variation in grammatical function assignment and linear order is considered to be the manifestation of varying levels of accessibility, it is difficult to see how we can account for how speakers choose between these various structures. In simple terms, it is unclear how choices among Adversative, Perfective and Di-Passive could be driven by information availability, since all three achieve the goal of prioritising patient over agent. Since these structures cannot be differentiated purely on the basis of the patient argument being more accessible (or more attention-grabbing), it follows that structural choice in Pondok Tinggi must be sensitive to more than just this.

Note that it is not possible to account for these findings on the basis of the frequency of different structures or verb forms, either. All structures that were encountered in the Patient More Animate condition were encountered in the Equal Animacy condition. Moreover, by modelling the choice between all structures in both conditions, we factor out the overall effects of structure frequency, looking purely at how the animacy manipulation impacted the relative distributions of structures among the responses.

The question therefore arises of how we can account for the findings. For the current data, we can draw on the descriptive study of Ernanda (2017) for an idea about other differences between the kinds of situations that these three structures are typically used to convey. For example, the Adversative may be preferred only when the speaker particularly needs to convey negative affectedness of the patient (Ernanda, 2017). Since the picture sets only differ in the animacy of the patient and not in the nature of the action, the dimension of patient affectedness would indeed not be expected to vary much between the conditions. This is concurrent with the finding that the rate of choosing Adversative was not affected by animacy. Similarly, Perfective may be used to express the unintentionality of the action undergone (Ernanda, 2017). This structure also would not be expected to differ much between conditions either, given that the same agent was present in both pictures of a stimulus pair. Moreover, this could also explain why Perfective was so infrequent in the current study: all pictures depicted inanimate agents; potentially, the notion of 'involuntary' action is only invoked when the agent is human (i.e. capable of volition).

A limitation here is that such variables do not have a clear basis in cognitive mechanisms, unlike accessibility; at present, it is not clear how we should go about integrating such variables into a psycholinguistic model. That is, the dominance of the Accessibility Hypothesis in explaining sentence production phenomena means that the ways in which other factors might be cognitively realised – or how they might feed into the sentence production model – has not received a great deal of attention in the recent sentence production literature (cf. Chapter 1). It would be useful to first undertake further experimental studies that examine how sentence formulation may be influenced by variables relating to "communicative goals" and not just "processing efficiency" (cf. Bock, 1982:41; see also Arnold et al., 2013). On the basis of such experimental findings, it would then be appropriate to evaluate whether and how such variables should be framed in terms of cognitive mechanisms.

While the current study only scratches the surface of this topic, it gives us a departure point and rationale for exploring the issue further. For example, future investigations could assess how the nature of *actions* affects sentence form in a range of languages while agents and patients are kept the same. This design could be instrumental for tapping into the impact of adversity (negative affectedness of the patient). The possible importance of such a variable is not only suggested by the structure of Pondok Tinggi, but is noted to be a factor in the choice of passive constructions in a wide range of languages, including languages that already feature in sentence production literature, such as Japanese and Chinese (Palmer, 1994; Toyota, 2011). Alternatively, a future study could assess how the nature of *agents* affects structural choice when actions and patients remain the same. This could be useful for tapping into the impact of agent-related properties such as the degree to which the action is volitional (or the degree to which it *can* be volitional, for example when considering inanimate agents). This is another factor noted to impact on the realisation of sentence form in a range of languages, far beyond Pondok Tinggi (cf. Fauconnier, 2012).

One limitation of the theoretical interpretation of this study is the small number of 'Perfective' structures in the data. Ultimately, due to this sparsity, we cannot draw very strong conclusions about the effect of animacy manipulations on the use of this particular structure. In this study, we saw that Perfective only accounted for a fraction of all responses in each condition; consequently the posterior distributions for the Perfective parameters showed a very wide range of values. In simple terms, there still was not enough information to really inform us about how the choice of this structure relates to animacy balance in the stimuli pictures. However, the Bayesian approach is helpful in that we can directly pinpoint a possible range of parameter values using the 95% credible interval.

A key goal of this study was to analyse sentence form choice in terms of a choice between more than two structures; that is, to not be restricted to binary choice. This approach has indeed allowed us to explore patient prominence in a more nuanced fashion, comparing not simply patient prominence and agent prominence, but comparing also different types of patient prominence with each other. However, there are still a number of criticisms or limitations of this approach.

Firstly, in Section 5.2 the multi-categorical analysis was proposed as an alternative to the use of multiple binary regressions. However, it may be said that this form of analysis actually resembles a series of binary logistic regressions, on statistical grounds: after all, each category is compared one by one with the baseline category. However, the difference is that when we adopt the multi-category approach, these models are fit all together, giving as an all-in-one analysis. This is more efficient and can lead to more precise estimates because the standard errors are smaller (Agresti, 2002). We also maintain perspective over all the categories the response can fall into.

A second criticism of this approach is that does not totally resolve the issues related to generalising or grouping the data. Although multiple categories are allowed on the outcome variable, it is still necessary to code and classify the structures into these categories. This process still necessarily involves the decision as to what groupings are relevant – a decision about the appropriate way to generalise across the true variability in the data. Nonetheless, being able to take more than two structural categories into account at least presents us with more flexibility in how we approach this. For example, here, I have been able to base the categories on the outcome variable on those presented by a descriptive analysis of the language (i.e. Ernanda, 2017). This categorisation procedure was also aided by the input of a native speaker linguist. This possibility is important for less well-studied languages where resources are scarce and where the range of structures encountered in the data do not immediately relate to the binary classifications seen in previous psycholingusitic studies of sentence form choice. It is also a demonstration of the interdisciplinary approach advocated by Norcliffe, Harris, and Jaeger (2015), among others.

A third limitation of the current analysis concerns the selection of a reference category. This is required in order to set up the 'baseline logit' model. Here, the Active category was chosen, being the largest category (apart from Other). The choice of reference category is to some degree arbitrary, in that the categories still relate to one another in the same way. However, the choice of reference category can affect what we learn from this type of model. Here, Active was chosen as the baseline category. Although Active is the largest category, it is not the intention to imply that this category is somehow more 'basic' than the others. Nonetheless, what we learn from the analysis is the odds of choosing something *other than* the Active, and how odds of different choices are affected by animacy. Therefore, we do not in fact gain a clear picture about how speakers choose between e.g. Adversative and Passive: we can just compare the likelihood of choosing Adversative over Active with the likelihood of choosing Passive over Active. Furthermore, we cannot assume that in every language there will be a numerically dominant structure category to select as reference category. A possible approach to this issue would be to explore the use of sum coding instead of treatment coding; that is, instead of comparing every category to the reference category as I have done here (treatment coding), we would obtain the contrast of each category (including Active) against a combination of all the other categories combined (sum coding). Exploring this option is beyond the scope of the current chapter, but doing so might provide additional enhancement and flexibility for future studies on non-binary structural choices.

A fourth criticism is that this approach does not circumvent problems posed by small subcategories in the dataset (i.e. low cell counts) – for example, as seen in the Perfective category in the current study. On the other hand, it should be noted that the Bayesian approach does have an advantage for datasets that are small: namely, when fitting random effects models to small datasets, the Bayesian implementation can be much more reliable (Sorensen et al., 2016). Given that small datasets are more likely to be an issue when dealing with endangered or minority languages, this is a pertinent consideration for 'field psycholinguistics' more generally (cf. Anand et al., 2011:4).

An additional remark must be made concerning the Other category. The responses in the Other category are all the responses that did not meet the criteria for inclusion in the analysis. The inclusion of this category in the analysis was motivated by Jaeger (2008). Although it is preferable to include these responses in the analysis rather than discarding them, we must be tentative in how we interpret any effects on this category. Specifically, this category does not represent a 'structure' in the way that the other categories do. As described in Section 5.3.3, this category is fairly heterogeneous. Consequently, it is debatable what it means if we see an effect on the likelihood of producing an Other response. Nonetheless, the flexibility of the analysis means that we even have the possibility to categorise into sub-types of error category if appropriate.

5.6 Concluding remarks

In summary, the aim of this chapter was to investigate the effects of animacy on sentence form choice in the case of more than two possible categories of structure. In this study, there was more than one structural category that provided the option to make the patient argument prominent; the fact that these structural categories were not uniformly impacted by the animacy manipulation conflicts with the idea that all such phenomena are relatable to a single latent variable, namely accessibility. Meanwhile, based on the findings from this study, it is possible to formulate hypotheses for effects on structural choice when experimental manipulations target other parameters than relative animacy of agent and patient, namely adversity and volition. The (potential) cross-linguistic impact of these parameters on sentence production should be evaluated through investigations in languages already familiar to psycholinguistic research (such as English, Dutch, German and Spanish), as well as more languages with typological profiles similar to Pondok Tinggi. Through further research, it will also be possible to better characterise such variables as cognitive factors, with the aim of establishing the mechanisms behind their effects on grammatical encoding.

Although the multi-categorical regression modelling technique in this chapter is more complex to implement and interpret than the forms of analysis more commonly used in the study of sentence production, the flexibility and insight afforded by the analysis makes it a useful approach. Ultimately, the choice of analysis of course depends on the data at hand and the research questions. Nonetheless, it is counterproductive if we are restricted to using a mode of analysis that does not allow us to satisfactorily reflect the variability in the data at hand, or to ask increasingly fine-grained research questions. As discussed in Section 2.2.5, flexibility of data analysis techniques also helps us to reduce the impact of subjective bias as we prepare our data for analysis. I have discussed several possible criticisms or limitations of this approach, in order that these can be taken into account in future decisions of whether this analysis is appropriate for a given dataset.

Last but not least, this chapter also forms a small contribution towards addressing the problem of scarcity of resources for understudied languages (Section 2.2.4): this study has provided insight into the effects of animacy on sentence production in a new language type. By investigating an endangered linguistic variety, this study also contributes to the effort to grasp the full range of linguistic diversity in the face of rapid language loss (cf. Evans & Levinson, 2009).

CHAPTER 6

Fluent scene description in Tarifiyt Berber: an element-level approach

In previous chapters, I reported picture description experiments in Tarifiyt Berber, Dutch and Pondok Tinggi. In Chapter 3, I presented a sentence production experiment in Tarifiyt and Dutch, which showed a relationship between referent animacy and priority in grammatical function assignment and/or linear order in Tarifiyt and Dutch. In Chapter 5, I used the same experimental paradigm as in Chapter 3 to investigate a language where there was more than one way of putting the patient in a prominent position. In order to analyse the choice among various outcomes, I opted to use a multi-categorical analysis rather than a binary analysis. This indicated again an effect of animacy on patient prominence. However, it also demonstrated that when our analysis more faithfully captures language-specific features, we are able to raise new challenges for information retrieval accounts; this enables us to go beyond simply validating the general finding that a contextually or inherently prominent referent is realised as a linguistically prominent argument.

In this chapter, I present a study in production in Tarifiyt Berber. This is again a *simply describing* study, but it departs in key ways from the *simply describing* study in Tarifiyt Berber in Chapter 3. The research questions of this study still relate to grammatical function assignment and linear ordering; however, rather than approaching this at the level of the entire structure, I instead investigate it at the level of the realisation of individual referents. Additionally, in this study I manipulate the message by foregrounding one or other referent, rather than through animacy manipulations.

As in the rest of this thesis, there is again a twin theoretical and methodological focus in this chapter. The primary theoretical aim is to better understand the manner in which cognitive variables impact on sentence form, aiming to distil the common features of sentence formulation in the face of the great complexities presented by linguistic diversity (Evans & Levinson, 2009). Meanwhile, the methodological focus concerns the choices made in data coding and analysis, in particular how these choices constrain the manner in which we interact with linguistic variation. Notably, the decision in this chapter to investigate the realisation of individual referents rather than holistic structures is prompted by a number of issues raised as I conducted the foregoing studies. These are limitations which still remain even if we use the multi-categorical approach rather than a binary approach. Specifically, they relate to issues of grouping and generalisation, experimental control and data loss. I argue that we can find new ways through these issues when we take a different approach to coding and analysing picture description responses. Essentially, I suggest that we may obtain additional insights into sentence form cross-linguistically if we code and analyse the outcome variables at the level of individual elements in the response, rather than through the classification of responses into holistic categories such as active and passive. I reflect on the potential insights that can be gained by taking such a perspective towards sentence production phenomena, particularly as we work with languages and speaker communities that are less familiar to psycholinguistic research.

In the following section, I review the challenges that persist after having conducted the studies reported in Chapters 3 and 4. I then present the motivations and setup for the study in Tarifiyt Berber, where I investigate the effect of referent foregrounding on grammatical function assignment and linear order. I then describe how I approached the coding and analysis of the data, by extracting measures that relate to sentence elements rather than holistic structures. Finally, I discuss the findings of the analysis, followed by a reflection on limitations of the approach, and possible future directions.

6.1 Background

6.1.1 Review of the approach in previous chapters

The studies reported in Chapters 3 and 5 followed the previous approach in the literature for *simply describing* experiments. Each stimulus picture was intended to elicit a single sentence. The spoken responses were transcribed, and exclusions made, before analysis. Responses were excluded if they did not form a single sentence that fulfilled certain target criteria with regard to form. Responses were coded, with every response being coded as representing a certain structure – for example, active or passive. The categories for coding the structures were chosen both on the basis of previous research (Chapter 3) and with reference to the descriptive work of a native speaker linguist (Chapter 5).

In order to analyse the data, it was necessary to group the responses in a way that made them amenable to analysis. In Chapter 3, I grouped the responses along a binary distinction, in line with the approach taken in previous *simply describing* studies. To achieve this, I needed to generalise over some of the structural variation in the data. For example, responses which could be seen as another structural type (such as verb-initial structures in Tarifiyt, or even presentative *er* constructions in Dutch) needed to either be rejected, or assigned to one of the two response categories (active/passive, subject/object-initial, respectively).

At the end of Chapter 3 and the beginning of Chapter 5, I argued that a binary analysis may not always be appropriate. In particular, I argued that relying on a binary analysis approach may in fact be limiting the insight that we can gain from *simply describing* experiments, especially as we aim to incorporate a wider diversity of linguistic types in sentence production research. Therefore, by means of the study reported in Chapter 5, I investigated how structural choice is affected by animacy in a language where there are more than two felicitous response structures. To do so, I thus needed to analyse the data in a way that could represent the choice between more than two categories. By using this analysis, it was possible to gain insight into choices among different patient-prominent forms, that would not be possible if the variability along the response variable was reduced to a dichotomy.

However, even if we use a multi-category analysis, there are still issues remaining unresolved. These issues fall into two main categories: firstly, problems concerning the need to group sentences into discrete categories, and secondly, problems concerning the amount of experimental control or data processing that is required to achieve this. I will now outline each of these two issues.

Firstly, at the end of Chapter 5, I pointed out that even when we can have more structures on the response variable, we still have to group the structures. In other words, even though the variability in the data can be more closely represented, we still need to generalise over the data. The researcher must still decide on what defines each of the relevant 'structural categories' in the language, and then classify the data into these groupings by hand – that is, by going through the responses one by one and deciding for each response which structure is represented.

Furthermore, although the use of multi-category analysis means that we do not need to generalise (or abstract) so much over the data, this in fact leads us to another problem. The less we generalise, the more we increase the number of categories on the response variable. However, the more levels on the categorical response variable, the more complex it is to fit and interpret the model. Lastly, the more categories we allow ourselves, the more likely we are to end up with small categories that contain few datapoints (an example of this is the perfective structure in Chapter 5). Small categories are troublesome because there are simply not enough datapoints to inform the statistical model. In short, the categorisation approach still leads to a dilemma: either we abstract more over the variability to have a manageable analysis, or we end up with many structural categories and a higher risk of results that are obscure or inconclusive.

The second problem relates to the control that is needed in order to achieve this structural grouping. That is, regardless of how many categories we allow, we still need participants to produce single sentences which fit into these structural classes or categories. As before, non-target structures must be rejected. As discussed in Chapter 3, we typically see high levels of data loss in these experiments. In particular, levels of data loss are higher when participants respond more exuberantly (in the sense of Bock, 1996). Exuberance is in turn more likely when we exert less control over how participants respond to the task (in the form of instructions or training, cf. Section 2.2.5). This situation is, in turn, amplified as we aim to incorporate more diverse languages into our research. The more diverse our sample, the more likely we are to (a) encounter more exuberance *and* (b) have less possibility to exert control over the way the participants respond to the task. This is because the wider range of languages studied increases the likelihood of working with participants that are not "socialized into being compliant responders" (Speed et al., 2017:192). The overall result is that, as we work with a wider range of languages, we are more likely to be confronted with higher rates of data loss. On top of this, we must consider the impact of high rates of data loss. Rejecting one third of the data may be tolerable when working in the lab with easily accessible participants, but is less tolerable when we are making special trips to conduct experiments in the field (as mentioned in Chapters 2 and 4). This consideration compounds the issue even further.

Ultimately, these concerns lead us to question what other ways there may be to approach the coding and analysis of *simply describing* data. Firstly, can we code and analyse the data without needing to categorise each response at the structural level? Can we avoid the need for the researcher to classify the responses along the dimension of abstract structural categories? Can we also sidestep the dilemma of whether to generalise across the variability vs. whether to face problems of small categories? Secondly, how can we get more out of our data? Can we gain relevant insights about sentence production variables – particularly grammatical function assignment and linear ordering – from the kind of data that we would typically reject? In what follows, I report a study which investigates the effects of referent foregrounding on grammatical encoding in Tarifiyt Berber, but in a way which addresses these methodological issues.

6.1.2 Towards a non-classificatory analysis of *simply describing* data

In the previous section, I described a number of problems that arise from the decision to classify sentential responses in a holistic way; that is, as falling into structural categories. Thinking of this in terms of the timeline of executing a study, the stage we are concerned with here is the process of going from fluent, exuberant raw data, to a coded variable that can be entered into a statistical analysis. At this stage in Chapter 3, the questions for coding process were, for Dutch: is this a passive or an active structure? and for Tarifiyt: is this an object-initial or a subject-initial structure? In Chapter 5, the larger number of levels on the categorical response variable meant that the question was among more options: is this an active, a di-passive, an adversity passive, or a perfective? In other words, each response in the raw data is given an overall label, which describes a general property of the response.

This approach treats each response in the experiment as a single token or datapoint. However, it is also possible to look within the responses, at the way that specific elements are realised *within* the structure. Recall the views on sentence production model discussed in Chapter 1. There are two ways of viewing the process of sentence production: namely, as structurally-driven or element-driven. The structural approach sees sentence production as essentially a choice of which overall structure is appropriate for a given situation. Meanwhile, the element-driven approach views the holistic structure (e.g. 'active' or 'passive') as the result of different rates of information retrieval for the elements within the sentence. Under a purely element-driven view, the abstract structure (e.g. passive) may even be viewed as epiphenomenal (van Nice & Dietrich, 2003). So, it is possible to view sentence production as an overall choice of structure (e.g. passive, object-topicalisation) based on the relative animacy of two referents, but it is also possible to view sentence production at an elemental level, as resulting from the choice of which element to put first, or which element to assign to subject function.

In parallel with these two views of the sentence production process – as a choice between holistic structures, or as a set of decisions about how to encode the elements within the sentence – there are also two ways to frame our research questions. Consider the pair of images in Figure 6.1, used to manipulate relative animacy of agent and patient in the studies reported in Chapters 3 and 5.

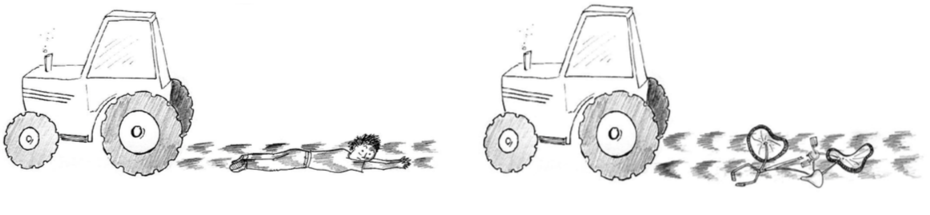


Figure 6.1: Example of stimuli pair used in *simply describing* experiments in Chapters 3 and 5, manipulating the relative animacy of agent and patient.

Here, the right-hand image depicts an equal-animacy scenario, while the left-hand image depicts a scenario where the patient is more animate than the agent. A typical research question for a *simply describing* experiment could be: does the picture on the left relate to *a higher rate of passives* than the right-hand picture? However, we can also ask this question slightly differently: does the patient in the left-hand picture *more often get assigned subject function* than the patient in the right-hand picture? On the one hand, the first question concerns the overall structure that is chosen; on the other hand, the second type of question concerns the choice about how to realise the patient itself. Similarly, we can reframe our questions about linear ordering. Previously, we have asked: does the left-hand picture lead to a higher rate of object-initial structures? However, we can also ask: does the patient more often get assigned preverbal position in descriptions of the left-hand picture? In other words, rather than asking about the overall structure, we can ask about the realisation of individual elements.

A parallel can be drawn here with research into another key variable in the study sentence production, namely pronominalisation. Pronominalisation is another important feature of sentence production that is sensitive to accessibility (Ariel, 1991). Note, however, that this variable *by definition* concerns the way elements are realised within the sentence, rather than properties of the structure as a whole. When investigating pronominalisation, we are interested in how our experimental manipulation affects whether individual referents are realised as full noun phrases or attenuated forms. Attenuated forms may be free pronouns (as in English, *he* rather than *a man*); clitic pronouns (such as in Berber); or even zero mention (sometimes termed 'pro-drop'), where a referent may be realised not in a separate morphological form, but rather through agreement morphology on the verb (Berber), or even completely absent at the morphological level and understood purely from the context (as, for example, in Mandarin Chinese, cf. Li & Thompson, 1981).

In the same way that we ask how the form of reference of an element is affected by its accessibility, we can ask how the likelihood of assigning subject function to a specific referent is affected by its accessibility, and how the linear position of a specific referent is affected by its accessibility. In other words, rather than asking whether a more accessible patient argument leads to a higher rate of passives, we can ask whether a more accessible patient argument is more likely to be assigned subject function. Although it seems to be a subtle difference, reframing the question in this way could potentially open up extra flexibility in analysing sentence production data, especially for languages that are typologically unfamiliar in psycholinguistics. In the remainder of this section, I reflect on these potential benefits.

Firstly, by reframing the question to be about sentence elements, we do not need to go through the stage of classifying structures in order to answer the question. This is because our outcome variable concerns the realisation of a specific referent rather than a descriptive label of the sentence as a whole. The result is that we avoid the issues that arise when the researcher must decide which overall structure is represented by each response (as discussed above). The coding process need only go as far as labelling individual sentence elements in terms of relevant properties, such as which grammatical function has been assigned, and where that element appears in relation to the verb, or to the other referent, and so on.

Secondly, since our outcome variable is no longer a categorical description of the holistic structure, there is less pressure for the data to fall into internally homogeneous groupings. This has the effect that we can relax the amount of participant control and data exclusion required to constrain the data into the relevant groupings, or to prepare it for meaningful analysis. Generally speaking, without the need to group responses into homogeneous structural categories, we have more freedom to let the responses be heterogeneous, that is, to vary along a wider range of dimensions – in other words, the situation we are faced with as a result of exuberant responding.

Thirdly, aside from resolving these issues raised from previous chapters, there is perhaps a more compelling benefit from investigating grammatical encoding at the level of referent realisation. Namely, important sentence production variables like animacy and givenness may be manifested in ways that affect referent realisation *without affecting the structural category*. One example of this has already been given: pronominalisation. However, this could also be argued to apply to grammatical function assignment. In particular, it is possible that sentences differ in the assignment of grammatical functions, without necessarily differing in *structural category* (e.g. active or passive). As mentioned in Chapter 1, the study of grammatical encoding has been largely concerned with the passive vs. active alternation, and how the roles of agent and patient are mapped to subject and object function. The correspondence between active and passive forms has tended to dominate the study of grammatical encoding, being seen as the archetypal example of the same proposition realised in different form (Gleitman et al., 2007). However, it is possible for the speaker to take different perspectives on the same event by using different predicates. In studies of sentence production in English, Lila Gleitman and colleagues (Gleitman, 1990; Gleitman et al., 2007) investigate alternations between pairs of so-called ‘perspective’ predicates. Perspective predicates include pairs such as *give* and *take*, or *chase* and *flee*. This is illustrated by the three sentences below:

(1) a. The policeman chases the thief.
b. The thief is chased by the policeman.
c. The thief runs away from the policeman.

All three sentences can be used to describe the same situation. The difference between (1a) and (1b) is one of voice, or diathesis: the thematic roles are the same, but the grammatical function assignment is different. In (1a), the policeman (agent) is subject; in (1b), the thief (patient) is subject. However, the difference between (1a) and (1c) is not one of voice (or diathesis) but rather a change in the perspective of the predicate. Sentence (1c) contrasts with sentence (1a) in that the situation is described from the perspective of the thief rather than the policeman; however, both are active, subject-initial sentences. In other words, even though sentences (1a) and (1c) are both active, there *is* nonetheless a difference in which referent is assigned which grammatical function.

If we simply label the structure as a whole, as active or passive, we will consider sentences like (1b) and (1c) as falling into the same category. In fact, it is quite likely that a sentence like (1c) would be rejected from many *simply describing* datasets on the basis that the thief and policeman are not expressed with exactly the intended thematic relationship (i.e. policeman realised as agent of the verb, thief as patient). That is, even though this is a perfectly felicitous description, it may be ruled out by the exclusion criteria (such as those used for the experiment reported in Chapter 3). However, if we approach such data in terms of the realisation of the referents – that is, tracking individual referents in production – we can simply ask what grammatical function they are assigned, without the issues that arise from structural classification.

Furthermore, sensitivity to the variation between examples such as (1c) and (1a) may prove fruitful as we move towards a more diverse linguistic sample. As discussed in Chapters 1 and 3, many languages mark grammatical relations, but do not make use of voice alternations (or diathesis). A key example of this is Tarifiyt Berber: while passive is marginal in this language, there is still a distinction between grammatical subject and object, with verbal agreement for number and gender, and case marking differentiating the two in postverbal positions (for a more in-depth description of the

Tarifiyt system, see Chapter 2.3.1). As described in Chapter 3, this profile can be captured under the notion of topic-prominence. In topic-prominent languages, subject function and topic role are distinguished, with the result that passivisation is not required as a strategy to encode prominence for patient arguments. Nonetheless, subject function still plays a role in sentence organisation. If there is a cross-linguistic hierarchy of grammatical functions as described in Chapter 1, it is still possible that subject function in such a language is more readily assigned to referents that rank highly in accessibility or prominence. For example, we could hypothesise that in Tarifiyt Berber, speakers *do* in fact assign subject function more readily to more accessible referents, even though this does not manifest in the use of the passive.

In addition to subject assignment, the elemental approach can be used to investigate linear ordering effects by investigating the likelihood of a certain character to take a certain position in the structure. That is, rather than asking if we are more likely to encounter a subject-initial, object-initial or verb-initial structure, we can ask whether being first seen relates to a higher likelihood of being given a certain position in the sentence structure, regardless of grammatical function. The positioning of a noun phrase relative to the verb is particularly relevant for Tarifiyt Berber. As discussed in Chapter 3, the preverbal position in Tarifiyt is associated with topic (Lafkioui, 2014); appearance in this position is not dependent on grammatical function (Mourigh & Kossmann, to appear). Therefore, with an element-oriented approach, we can directly ascertain how the status of a referent relates to the likelihood of its assignment to the preverbal topic position, as opposed to appearing in the postverbal sentence core.

6.2 The current study

This study, as in the other studies in this thesis, has dual aims. From a theoretical perspective, the aim is to investigate the effects of referent prominence on grammatical function assignment and linear ordering in Tarifiyt Berber. In Chapter 3, this same aim was approached using a design that is typical of *simply describing* experiments. Here, I investigate the same variables within the same language, but instead I probe the effects through referent realisation rather than response classification. Firstly, following on from the above discussion, I examine whether patterns of grammatical function assignment are present in the absence of diathesis: does a more accessible referent get assigned subject function more often? Secondly, I aim to investigate linear ordering: does the accessibility of an element affect its sentence position? In the case of Tarifiyt Berber, we are interested in whether the noun phrase referring to that referent appears in the preverbal position (cf. Chapter 3). Note that, given the possibility of verb-initial order in Tarifiyt sentences, we must also consider the possibility that all noun phrases can occur postverbally.

Alongside the questions of grammatical function assignment and linear order, I will examine the forms of reference used; namely, whether a referent is realised as a full noun phrase or attenuated form, and how this relates to the experimental variable (accessibility). This is required, because the design that I use means that participants can respond exuberantly and are given free rein to use pronominal reference. For Tarifiyt, it is important to recall from Section 2.3.1 that attenuated reference in Tarifiyt is frequently realised through agreement morphology on the verb, for subjects, and clitic pronouns attached on the verb for objects (in other words, for Tarifiyt, attenuated forms encompass more than just free pronouns, unlike English or Dutch).

From a practical (or methodological) perspective, the aim is to design and implement a study of accessibility effects on grammatical encoding phenomena through the realisation of referents rather than through holistic response classification. As discussed above, a key driver of this approach is that it should improve the feasibility of working with exuberant data, without the pressure for individual picture descriptions to conform to structural categories. Therefore it is vital that in implementing this approach, we must do so with data that is collected in a less controlled fashion. In other words, it is also important to assess how well this approach allows us to handle response exuberance. The design here is intended to be practicable for data collection in a non-lab setting, with participants that are not acquainted with the aims and conventions of experimental psychological research.

6.2.1 Study design

As discussed above, an element-level approach is customary as regards the study of pronominalisation, which necessarily refers to the encoding of individual elements. However, it is less usual to approach questions of grammatical function assignment and linear ordering from this perspective. One study that does adopt such an approach, however, is Vogels, Krahmer, and Maes (2013). In that study, the authors aimed to investigate the effect of contextually manipulated animacy on the linguistic realisation of characters in scene descriptions in Dutch. The experimental manipulation was the lexical and perceptual animacy of the on-screen characters.[1]

In that study, participants viewed the interaction between the on-screen characters in three consecutive animations. Firstly, the character of interest was presented alone on screen. Secondly, additional characters appeared on-screen, and an interaction with the first character was depicted. This was followed by a third stage where the character of interest then engaged in some further (intransitive) action. Participants were recorded describing these animations, and their responses were transcribed. Then, in terms of the processing and analysis of the data, responses were not classified into structural categories, but instead labelled in terms of the realisation of the character of interest. The questions here were: was the character in question assigned subject or not, and: was it pronominalised? In order to analyse the results, the authors used mixed logit modelling. They modelled the likelihood of the character of interest being referred to as subject, and separately, the likelihood of the character of interest being realised as a full noun phrase versus an attenuated form.

I propose to adopt a similar design to Vogels et al. in order to probe questions about sentence production in Tarifiyt Berber (as described above); that is, in order to investigate effects of accessibility on the realisation of referents. One referent will be made more accessible through visual foregrounding, by presenting it first in the sequence of pictures. This is envisaged along the lines of Prat-Sala and Branigan (2000)'s 'salience' manipulation. In Prat-Sala and Branigan's study, the salience of one or other referent in a scene was manipulated by use of a preamble. For example, participants were asked to describe a scene depicting a swing knocking over a scooter, and this picture was paired with a preamble which focused on one of the two referents. The preamble to make the swing more salient was *There was this old rusty swing standing in a playground near a scooter, swaying and creaking in the wind*, while the preamble to make the scooter more salient was *There was this old red scooter standing in a playground near a swing, with rusty wheels and scratched paint.*[2] The key difference in

[1] In short, the authors manipulated animacy by guiding participants either to perceive the on-screen entities as inanimate shapes, or to imbue the entities with animacy (i.e. to anthropomorphise them). For a fuller description of the experimental manipulation the reader is referred to Vogels et al. (2013), since this aspect is somewhat tangential to the current discussion.

[2] Although Prat-Sala and Branigan's study is conducted in both English and Spanish, the article only provides the English preambles.

the current study is that this ‘preamble’ is provided in the visual domain, requiring participants themselves to describe the foregrounded referent. Overall, this seems a more natural task for participants than continuing a story that another speaker has started.

The storytelling design is also advantageous for the current aims. In this design, the task for participants is not to describe isolated images, but relate a story on the basis of the sequence of visual stimuli. Telling a story is likely to be a more accessible task for a wider range of participants than just describing isolated pictures, in that it can be related to the common cultural practice of storytelling (cf. Section 2.2.3). However, the design of the current study will diverge from Vogels et al. (2013) in certain key ways. For example, in the study of Vogels et al., the authors only investigated the realisation of one character per story, whereas I wish to investigate the effects relating to the different characters that interact in a story, contrasting how these different characters are realised in the responses. In addition, participants in the study of Vogels et al. were encouraged to use a single sentence per story segment; this was intended to control the exuberance of the data, but as discussed above, one aim of the current study is to allow for exuberant responding.

Furthermore, given the importance of this issue as we aim to incorporate a diverse range of languages in our research, the current study aims to exert minimal control of response exuberance. Since the responses will be exuberant, the data is likely to be messy. Therefore, it is necessary to first find a way to extract meaningful measures from messy data (an issue that may be considered one of the perennial issues of sentence production research; Bock, 1996). The fact that we are asking questions about referents, rather than about holistic structures, will also have consequences for how the data is organised. In specific terms, the result is that each observation in the dataset will be an observation about a particular referent, rather than an observation about the holistic response. Concretely, each row of the dataset therefore needs to relate to one referent, rather than one picture.

In summary, the design of the current study will involve a storytelling task, implementing a manipulation of referent accessibility, in order to investigate how accessibility of referents relates to the way that they are realised, in terms of their grammatical function and their linear position. The aim is not to look at this in terms of choice of structure, but in the linguistic realisation of the individual referents. In particular, I will aim to draw a comparison between how the foregrounded character is realised, versus how the non-foregrounded character is realised. As in Vogels et al. (2013), the dependent variable in the current study is not the structure of the response, but how a given referent is realised in the response. However, unlike Vogels et al. (2013), I will look at more than one referent per response, in order to draw the necessary contrast between the foregrounded and the non-foregrounded character. Overall, the coding and organisation of the data differs significantly from the approach used in the *simply describing* studies reported in Chapters 3 and 5.

6.3 Story-telling task in Tarifiyt Berber

6.3.1 Experimental design

The study reported in this chapter was designed to elicit fluent picture description responses. The aim is to investigate the way that depicted characters are realised in the responses. Participants are asked to describe a sequence of four pictures that form a story. Each story involves two characters, engaged in an everyday interaction involving an inanimate object. In each story sequence, one of the two characters is foregrounded. The aim is to explore the relationship between the (visual) foregrounding in the picture stories and the (linguistic) realisation of that character of the participants' spoken stories. This foregrounding is intended as an operationalisation of accessibility, as described above (however, note that it could potentially also be construed in discourse terms as presenting that referent as topic; cf. Hwang & Kaiser, 2015).

The experimental manipulation in this task is, as just described, the foregrounding of the one or the other character. This is achieved as follows: in the first picture of the story, only one of the two characters is depicted; this means that speakers first must describe one of the two characters, before continuing to incorporate the other character in their story from picture two onward (i.e. similar to the setup of Vogels et al., 2013). Three such picture sequences, or 'storyboards' were created. These can be found in the Appendix 7.2. One of the storyboards is also in depicted in Figure 6.2. Each of the three storyboards depicts a short interaction between a pair of characters, involving an inanimate object. In one story the interaction is between an older and a younger male character (and a balloon), in another story it is between a man and a woman (and a bowl of soup), and finally another story depicts an interaction between a human and an animal (and a loaf of bread).

Two versions of each storyboard were created, where the only difference between the two versions was which of the two characters was presented in the first picture. The second, third and fourth pictures were identical between the two versions of the storyboard. The design is illustrated in Figure 6.2. The result is that the same story, i.e. the same interaction between the two characters, is described in both conditions. Of course, the same story should not be told twice by the same participant; therefore, the two conditions are counterbalanced so that each participant only responds to one version of each storyboard. For example, taking the sequence in Figure 6.2, some participants described the version where the little boy was foregrounded, while other participants described the version where the man was foregrounded. The intention here is to be able to control for the effects of inherent properties on referent realisation (such as gender and animacy). Note that this approach parallels the design of Vogels et al., who also used story pairs, differing only in the animacy of the character of interest (i.e. the pairs that contrasted purely in terms of the independent variable).

The task for participants, then, is simply to tell a story based on the pictures, viewing them one by one. As discussed above, this is intended to be a free description task that is accessible for many types of speakers, with the notion that story-telling is likely to be a much more familiar activity for most people than describing isolated scenes. The scenarios are purposely designed to be accessible for participants from a wide variety of cultural settings. Crucially, there is no control imposed on participants in terms of metalinguistic instructions about how many sentences they should produce, nor what form those sentences should have or which lexical items they should contain.

Hypotheses

For this study it is possible to identify specific hypotheses regarding the key outcome variables, as described below.

In Chapter 3, I implemented the animacy manipulation from Prat-Sala and Branigan (2000), investigating how relative animacy of referents affected sentence production in Tarifiyt Berber. I found that animacy affected **linear ordering** in Tarifiyt, in that the highly animate referent was more likely to appear in preverbal position. In the study of Prat-Sala and Branigan (2000), foregrounded referents showed a similar pattern of realisation to highly animate referents. Taken together, these findings lead us to hypothesise that in the current Tarifiyt study, a preverbal position will be more likely for foregrounded referents than for non-foregrounded referents.

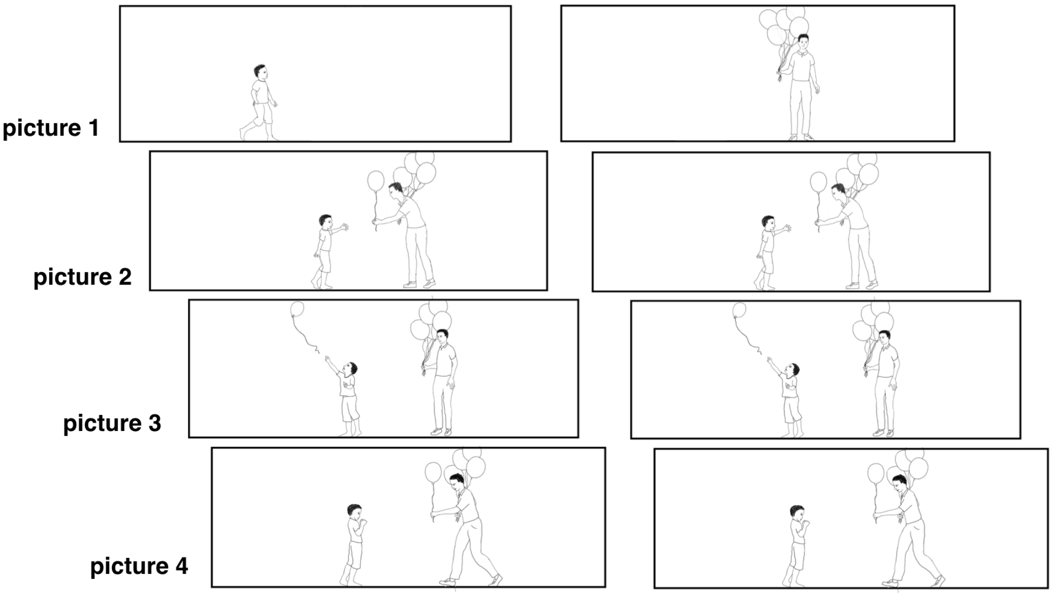


Figure 6.2: Example of storyboard stimulus design. Each participant sees one of the two versions, viewing images one by one on a laptop screen.

In Chapter 3, I coded and analysed the data in terms of holistic structures. As mentioned above, when we classify holistic structures, differentiation in **grammatical function assignment** will only be visible as an active vs. passive voice alternation. In Chapter 3, the active vs. passive alternation was found to be absent from the Tarifiyt data. Nonetheless, as discussed above, Tarifiyt does realise a grammatical subject, and it is possible that the decision to assign a referent to subject function can still be affected by whether or not it is foregrounded. If subject assignment in Tarifiyt is sensitive to accessibility, I expect to see that foregrounded referents more often show up as grammatical subject than non-foregrounded referents.

As for the **form of reference**, previous studies have found that more accessible referents are more often realised in attenuated form (Ariel, 1990, 1991). Therefore, the hypothesis is that the foregrounded referent in a story will be more often realised in attenuated form than the non-foregrounded referent.

Lastly, it should also be noted that, due to the methodological aims of this study (to implement a more accessible task (storytelling) while allowing for exuberant responding), an overall aim of the study is to assess whether we can identify accessibility effects in the current data, despite the lack of experimental control.

6.3.2 Participants and procedure

Twenty-eight speakers of Tarifiyt Berber participated in the study, a subset of those who participated in the study in Chapter 4. Participants were all undergraduate students at the Faculté Pluridisciplinaire de Nador. Recording sessions were conducted at the university in quiet study areas. Each participant indicated their consent to take part. Testing was kindly facilitated by the invaluable support of Prof. Mostafa Ben-Abbas and native-speaker assistant Hanae Boudihi.

Participants were informed about the task verbally in Tarifiyt. They were instructed that they would see four pictures which together formed a story, and that they should tell that story by describing each picture in turn. Pictures were presented individually on slides using Microsoft PowerPoint, meaning that the storyboard was only visible to participants one image at a time. Participants narrated the story at their own pace by pressing a key to advance through the slides. Presentation was counterbalanced in that each participant responded to one version of each storyboard, which were assigned pseudorandomly. Participants' responses were recorded and transcribed.

6.3.3 Dependent measures

To recap, I wished to extract measures relating to:

- grammatical function assignment (specifically, which character was assigned to subject)
- linear order (i.e. where in the order of the sentence the characters were realised)
- form of reference (here, whether characters were referred to with full noun phrases or in attenuated form)

As mentioned above, one key way in which this study departs from the design of Vogels et al. is that I extract data regarding both characters, and not only the foregrounded character. Moreover, I do not only restrict the data to one segment of the storyboard (cf. Vogels et al., 2013:5) but seek to extract data relating to the realisation of characters across the whole story. This not only facilitates a more in-depth exploration of the relationship between non-linguistic properties and linguistic form, but also makes maximal use of the dataset.

Indeed, although the dataset is expected to be 'exuberant', a key aim is to maximise the data yield. In other words, I wish to minimise data loss through exclusions, and to extract the maximum extent of information on the features of interest. The challenge, therefore, is to come up with a way of extracting these relevant measures across the whole dataset, in face of the large amount of variability. In the next section, I outline how the data is processed, and in particular how the measures are extracted. Overall, the approach to extracting measures is inspired by natural language processing techniques, where algorithms are scripted (written in a language such as Python) in order to iterate over bodies of text in an automated fashion, and compile information about the features encountered in the text (Bird, Klein, & Loper, 2009).

6.3.4 Data processing

Exclusions

Each participant described three storyboards with four pictures in each. In line with the foregoing discussion regarding data yield, it is considered important to only reject data if participants have misunderstood the task, or if they have perceived an entirely different scenario from the one that was intended by the picture.

This minimal exclusion criteria led to rejection of 11% of the data. These were cases where the participant was not telling a story and instead described each picture as if it was isolated: this constituted a deviation from the task. To illustrate the nature of the data collected, Example (2) gives an example of a story told by a participant. For comparison, Example (3) provides an example of an excluded response set, where the participant did not tell a story.

(2) a. řexxu ijj n waryaz yeqqim ... ɣar missa
now one of as.man 3sg.m.sit.perf ... at table
‘now a man is sitting ... at a table’
b. qa teggur ɣars ijj n temɣart, tiksi ijj n ...
presrel 3sg.f.walk.impf at.3sg one of as.woman 3sg.f.carry.perf one of ...
teggua ɣars ijj n temɣart tiksi afenjař
3sg.f.walk at.3sg one of as.woman 3sg.f.carry.perf fs.bowl
‘a woman is walking towards him, she has brought a ...
a woman is walking towards him she has brought a bowl’
c. tiweḍ ɣars ɣar missa yewḍa-yas
3sg.f.arrive at.3sg at table 3sg.m.fall.perf-3sg.io
‘she reached him at the table and it fell’
d. ikkar netta isɣuy xas
3sg.m.arise.perf he 3sg.m.scream.perf on.3sg
‘he got up he screamed at her’

(3) a. ijj n waryaz qa yeqqim
one of as.man presrel 3sg.m.sit.perf
‘a man is sitting’
b. ijj n waryaz tiwy-as-d ijj n temɣart lmakla
one of as.man 3sg.f.bring.perf-3sg.io-hither one of as.woman food
‘a man has been brought food by a woman’
c. ijj n temɣart tzelleɛ x waryaz nnes aman
one of as.woman 3sg.f.spill.perf on as.man of.3sg fs.water
‘a woman has spilled water on her husband’
d. ijj n waryaz imneɣ ak temɣart nnes
one of as.man 3sg.m.fight.perf with as.woman of.3sg
‘a man fought with his wife’

Coding

In this study, the decision has been taken to avoid constraining the number of sentences per description. However, it is still the ultimate aim to be able to say things about the production of sentence-like units. Therefore, the first step in processing the data is to identify sub-units that each represent a sentential core. With this, we can then compare the properties of different ‘sentences’ with each other and between conditions. To make clear that we are not dealing with the same kind of responses as in typical sentence production experiments, I will coin the term “verb unit”, indicating that each unit contains one overall predication, typically including a verb, its arguments and its adjuncts (including sub-clauses).

As discussed above, I wished to extract measures regarding (a) whether a given character was assigned to subject, (b) the position in which a character was realised in the sentence (particularly with relation to the verb), and (c) the manner in which a given character was referred to (i.e. full noun phrase or attenuated form). The aim is to then relate these measures to whether or not the character in question was the foregrounded (first-seen) character or the non-foregrounded character (second-seen). Therefore, it is necessary to first code the responses, producing a labelled version of each verb units, in a manner that facilitates subsequent extraction of the measures described.

To illustrate this, consider Example (4) which shows a response to one picture. The speaker is describing the second picture of the storyboard that involves the balloon (see the Appendix). In this case, the first picture that the speaker saw was the one with only the little boy. Below the original response, the 'coded' version is displayed. The response itself comprised two verb units: these are separated in the coded version using the €€ character sequence.

(4)

yemmerqa	ag	ijj	n	waryaz	yewc-as	glubu
3sg.m.meet.perf	with	one	of	as.man	3sg.m.give.perf-3sg.io	balloon

'he meets a man he gives him [a] balloon'

Coded version:
(boy)Sverb PREP ijjn(man)NP €€ (man)Sverb-(boy)IO (balloon)NP

The coding of responses is done here in a way that reflects the glossed version, but with some fundamental differences. In particular, the coded version does not include all of the grammatical information from the gloss (such as the state of the noun, cf. Section 2.3.1). The coding can be understood as a form of glossing that facilitates extracting the features that are relevant for the analysis. Key aspects are as follows:

- each element of the response is coded with a category label, e.g. verbs are glossed with *verb*, noun phrases are marked with *NP*, and so on.
- each reference to a character is also labelled with the name of the character. Taking the boy character from above as an example, we have '(boy)Sverb' when the boy is subject; '(boy)NP' when the boy is a full noun phrase, 'ijjn(boy)NP' for an indefinite noun phrase with *ijj n* ('one' or 'a'), and finally, '-(boy)DO' and '-(boy)IO' for direct object and indirect object clitics, respectively.
- The inanimate objects in the stories (e.g. the balloon) are treated in the same way as the animate characters (e.g. '-(balloon)DO' for a direct object clitic referring to a balloon).
- there is not necessarily a one-to-one relation between the coded elements and lexical items. For example, the phrase *ijj n waryaz* is tagged as 'ijjn(man)NP' in order to capture that the man is being referred to using a full noun phrase that is indefinite.

Since the responses were unrestricted and 'exuberant', participants sometimes embedded sub-clauses within their descriptions. In order to straightforwardly analyse the principal grammatical function assignments and ordering, I took the decision to only consider main verbs for further analysis.[3] These issues (and other limitations) will be reviewed again in the discussion in Section 6.5.

[3] This was achieved by adding a marker at the beginning and end of sub-clauses, whereby they could be skipped during parsing in Python. This means that information about sub-clauses could be included again by simply adjusting this code.

Extracting dependent measures

I am not looking at holistic properties of responses, but rather focusing on how the two characters in the story are encoded within the responses; therefore, rather than classifying a response as 'active' or 'passive', the approach taken here is to identify each individual reference to a character, and log the features of each reference (whether the character has subject function, whether it is a full noun phrase or an attenuated form, and where in the sentence it occurs). The point is to then be able to relate these features to the foregrounding manipulation – i.e. whether that character was the one presented first (henceforth 'first seen') or the one presented second (henceforth 'second seen').

This approach has a number of implications for how the dataset should be organised. Firstly, in typical sentence production studies, such as the ones reported in Chapters 3 and 5, each row of the dataset represents one response to one picture stimulus. This makes sense because in such a design, each datapoint is one holistic structure. In the current design, however, each row in the dataset is an observation about how a character is encoded in a given response. For example, a single row of the dataset contains information about whether a character is encoded as subject, the form of reference used, and the position of the element in the verb unit. Secondly, in a typical dataset, participants are restricted to produce only one sentence per picture. However, in the current design, more exuberant responses are permitted. This means that a single picture might elicit more than one sentence-like unit or 'verb unit'. The consequence for the current dataset is that each row of the dataset corresponds to one of these verb units, but there may be more than one verb unit per picture seen.

This organisation of the dataset is illustrated by the excerpt provided in Appendix 7.2. This excerpt contains all responses from a single participant.[4] The first table provides the original set of responses, along with the translations and coded versions of the responses. The second table demonstrates the way that this data was used to obtain variables for the analysis. In practical terms, these measures were extracted by using code written in Python to loop through the elements of the coded responses, in order to extract information relating to each of the features in the second table provided in 7.2. In what follows, I briefly outline each of the key variables for the analysis, including how the dependent measures were extracted. This description relates to the second table in Appendix 7.2.

Firstly, each row of the dataset contains information regarding general properties of responses, in particular, which story is being told, and which picture in that story is being described. Following that, the CHARACTER column indicates which character we are extracting information about. For example, in the *kitchen* story, we want to extract information regarding the *man* character, and information regarding the realisation of the *woman* character. The column CHARACTER:FIRST OR SECOND relates to the independent variable for this study, namely whether the character in question is shown in the first picture (*first seen*) or introduced in the second picture (*second seen*). In this excerpt, we see that for this participant, the first-seen character in the *kitchen* story was the *man*.

The column CHARACTER:SUBJECT/NON-SUBJECT indicates whether the character in question was the subject of the verb unit or not. In order to extract this information, the script was programmed to look for any instances of *(x)Sverb*, and to cross-reference x with the name of the character currently being assessed (i.e. in the CHARACTER column). If the two matched, then the character in question was logged as subject. The column CHARACTER:FORM OF REFERENCE refers to whether the character in question is realised with a full noun phrase, or an attenuated form (empty cells here indicate that the character was not referred to at all in the response). In order to extract this information, the script was programmed to cross-reference any instances of character reference with the name of the character

[4] For the purposes of exposition here, the excerpt chosen is from one of the less exuberant participants.

currently being assessed. Where *(x)NP* was found, the script logged an *NP* reference; otherwise, the reference was logged as *attenuated*.

If the character is realised as a noun phrase, the column np:prepostv then indicates whether the noun phrase is preverbal or postverbal. This linear ordering feature was extracted as follows. First, each coded response was converted into an ordered list in Python, meaning that each element of the list could be assigned a number according to its position on the list (i.e. its 'index'). The relative position of noun phrase and verb could then be easily extracted, simply by subtracting the numerical index of the verb from the numerical index of the noun phrase. If the noun phrase precedes the verb, the outcome is a negative number (e.g. noun phrase at position 1, verb at position 2, would mean $1-2 = -1$). Conversely, a positive number indicates that the noun phrase occupies a postverbal position.

6.4 Results

As a whole, the dataset comprises 1028 individual cases (rows). This total arises from 514 usable verb units, each of which was processed for measures relating to *each of the two characters* in the story, as described above (i.e. 514 x 2 = 1028). The number of (usable) verb units elicited per picture of each storyboard is displayed below in Table 6.1.

Storyboard	Unique verb units (total = 514)			
	pic 1	pic 2	pic 3	pic 4
balloon	30	41	30	37
bread	38	44	42	49
kitchen	46	50	43	64
Total:	*114*	*135*	*115*	*150*

Table 6.1: Number of unique verb units per picture of each storyboard (of 514 unique verb units in the entire dataset, after exclusions).

In what follows, I describe and analyse the data with respect to subject assignment, form of reference and linear ordering. I use binary mixed effects logistic regression to model each of the measures as a function of whether a referent was the first seen or second seen. In each model, the random effects structure includes random intercepts for participants and storyboards. As in previous chapters, model comparison is in each case conducted using ANOVA, the result of this is also reported in each of the model summary tables.

Subject assignment

I aimed to assess how being the first seen or second seen character relates to being assigned subject function during the ensuing story. This is plotted in Figure 6.3. In this plot, the counts (on the Y-axis) indicate the number of verb units where the given character (the first or the second character seen) is assigned the subject function. This is plotted against the order of picture presentation on the X-axis (i.e. picture 1, 2, 3 and 4). The first-seen character is plotted in red, the second-seen character in blue. The figure shows a trend for the first-seen character to be assigned subject function more often, even in the pictures where both characters are depicted.

Note that these are raw counts and so care should be taken in the interpretation. In particular, the total number of verb units per picture is not constant. This is because participants were not restricted to producing one sentence. The total number of verb units per picture can be observed in Table 6.1.

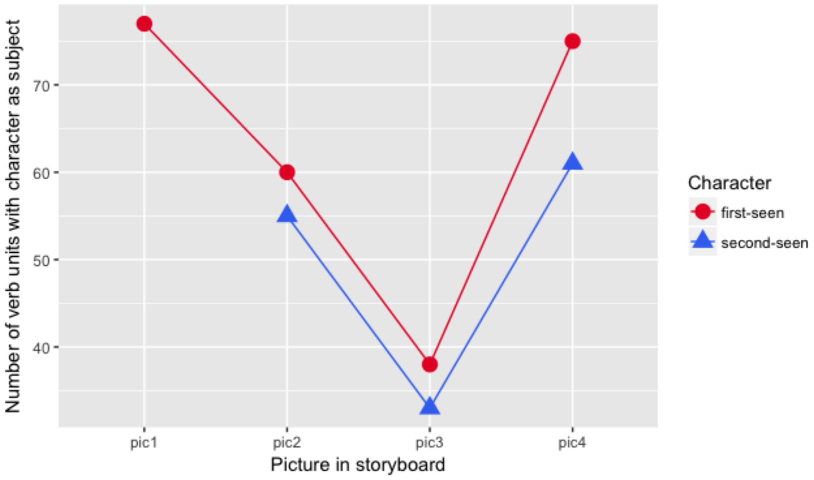


Figure 6.3: Subject-realisation of characters: how often the first-seen (red) and second-seen (blue) characters were realised as subject, per picture of the story. Note that this excludes units without a verb marked for subject agreement, such as copula constructions (cf. Section (6)).

There appears to be a trend for the third picture to have fewer instances of *either* first seen or second seen characters being assigned subject. This could be for one of three reasons: (i) fewer verb units produced overall, (ii) a tendency for the subject of the verb unit to be assigned to a different referent – usually the inanimate object in the story, or (iii) more nonverbal predications (typically copula constructions, of the type *There is a man*). Table 6.2 demonstrates that (ii) is likely to be the reason for this trend: in Picture 3 we see that the subject assignment to first-seen character, the second-seen character or to another entity (typically the inanimate object) is fairly evenly distributed, in stark contrast to Pictures 1, 2 and 4.

	Unique verb units (total = 514)			
Subject of verb unit	pic 1	pic 2	pic 3	pic 4
First-seen	77	60	38	75
Second-seen	0	55	33	61
Other	1	2	36	4
None	36	18	8	10

Table 6.2: Subject assignment by picture

The question is whether the trend for subject to be more often assigned to the first seen than the second seen represents a significant effect. In other words, does being the first seen relate to a statistically significant increase in the likelihood of being assigned subject? In Table 6.3 I report a logistic regression model examining the relationship between firstness and subject assignment. This model concerns the second, third and fourth pictures of the sequence (i.e. the pictures where there is a contrast possible between first and second seen).

	intercept only	subject model
(Intercept)	-0.40^{***}	-0.52^{***}
	(0.07)	(0.10)
first-seen		0.25
		(0.14)
ANOVA (model comparisons)	$\chi^2 = 3 \;\, (p = 0.08)$	
AIC	1084.42	1083.42
BIC	1098.47	1102.16
Log Likelihood	-539.21	-537.71
Num. obs.	800	800
Num. groups: Participant	26	26
Num. groups: Storyboard	3	3
Var: Participant (Intercept)	0.00	0.00
Var: Storyboard (Intercept)	0.00	0.00

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 6.3: Logistic regression model of subject assignment, in the second, third and fourth pictures of the story. The analysis models how the likelihood of a character being realised as subject is affected by whether it is first seen rather than second seen.

The model indicates that while there is a numerical trend for first-seen referents being more likely to be assigned subject, the effect does not reach statistical significance. The intercept-only model is improved only marginally by inclusion of fixed effect for first seen vs. second seen ($p = 0.08$).

Forms of reference

Next, we can turn to the question of whether characters are realised as full noun phrases, or attenuated forms, and how this relates to being first seen or second seen. The plot in Figure 6.4 illustrates this situation. This plot depicts the number of verb units in the whole dataset where each of the two characters are realised, and breaks this down into when the character is expressed using a full noun phrase (dark blue), versus when the character is only realised in attenuated form (light blue). Again, values on the Y-axis indicate raw counts of verb units.

Overall, it can be observed that the pattern of using full noun phrases or attenuated forms varies as the story progresses – as would be expected in such a task. The plots suggest that full noun phrases are used to introduce characters in the first two pictures: there is a higher rate of full noun phrases for the first seen in the first picture, and a high rate of full noun phrases in the second picture for the

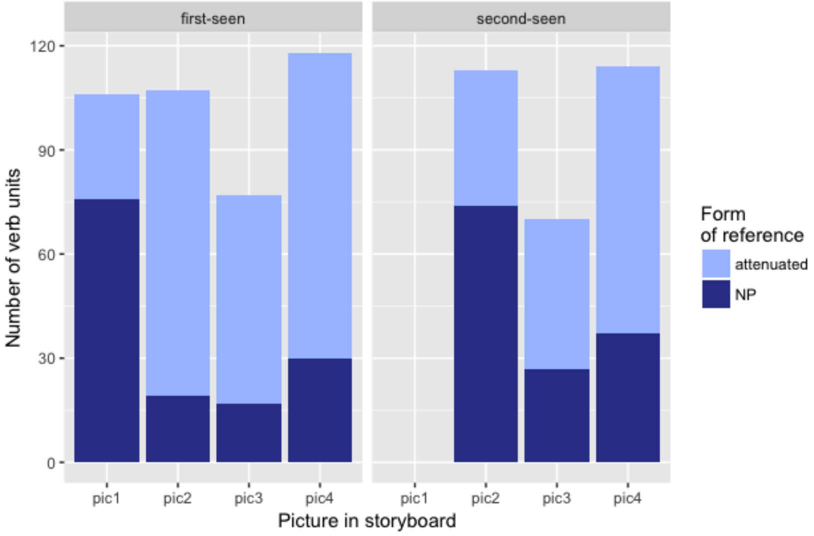


Figure 6.4: Forms of reference for first-seen and second-seen characters, per picture. The left panel shows the number of verb units where the first-seen character was realised as full noun phrase (dark blue) or attenuated form (light blue); the right panel shows this information for the second-seen character.

second-seen character, as would be expected.

For the statistical analysis of this data, it is more interesting to explore the relationship between firstness and forms of reference after both characters have already been introduced, i.e. in the third and fourth pictures of the story. That is, we may ask whether firstness has a *lasting* effect on the frequency of mention using a full noun phrase, beyond the introductions in the first two pictures. In line with our hypotheses, if the first-seen character has higher accessibility overall, it may be the case that speakers are overall less likely to refer to it using a full noun phrase, and more likely to refer to it in attenuated form, even in the third and fourth pictures of the story.

Table 6.4 reports a binary logistic regression model, examining the effect of character firstness on the likelihood of realisation of that character in full noun phrase form or attenuated form. Importantly, this model only assesses pictures three and four of the sequence, in order to assess whether there is a continued effect of firstness on noun phrase realisation (i.e. after the images where characters are introduced). The reference level for the FORM OF REFERENCE variable is full noun phrase. In other words, the model concerns the likelihood of encountering an attenuated form rather than a full noun phrase, depending on whether the character in question was the first one seen.

Here we see that whether a character is first seen or second seen relates significantly to whether it is realised in attenuated or full noun phrase form in the last two pictures of the story. The intercept-only model is improved significantly by inclusion of a fixed effect for first seen vs. second seen. The analysis indicates that when the character is first seen, there is a significantly higher likelihood of it being expressed in attenuated form rather than as a full noun phrase. Conversely, in the last two

	intercept only	FORM OF REFERENCE model
(Intercept)	0.97^{***}	0.69^{*}
	(0.27)	(0.30)
first-seen		0.57^{*}
		(0.25)
ANOVA (model comparisons)		$\chi^2 = 5.15\ (p < 0.03)$
AIC	434.91	431.76
BIC	446.73	447.51
Log Likelihood	-214.46	-211.88
Num. obs.	379	379
Num. groups: Participant	26	26
Num. groups: Storyboard	3	3
Var: Participant (Intercept)	1.11	1.13
Var: Storyboard (Intercept)	0.03	0.03

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 6.4: Logistic regression model of attenuated vs. full noun phrase realisation, in the third and fourth pictures of the story. The analysis models how the likelihood of a character being realised as attenuated form rather than full noun phrase is affected by whether it is first seen rather than second seen.

pictures of the story, when a full noun phrase is used, this is more likely to refer to the second-seen character.

Order of noun phrase(s) and verb

Figure 6.5 depicts the raw frequency of preverbal and postverbal realisation of noun phrases, related again to whether the character in question was seen first or second in the story.

The figure depicts the situation for the whole dataset. First, we can look at the outcomes for picture 1, where the first-seen character is introduced. We see here that first-seen characters are more often introduced with a preverbal noun phrase. Secondly, we can look at the outcomes for picture 2, where the second-seen character is introduced. We see here that the second-seen characters are more often introduced as postverbal noun phrases.

In terms of the analysis, we can investigate whether full noun phrases typically precede or follow the verb, and whether this relates to whether a referent was the first seen or second seen.[5] Table 6.5 reports a binomial logistic regression analysis that models the relationship between first/second seen and preverbal/postverbal position of a noun phrase. The model examines the effect of character firstness on the likelihood of realisation of that character as a preverbal noun phrase rather than a postverbal noun phrase (reference category for the position variable is postverbal). This model con-

[5] Note that this includes observations where only one referent was realised as a noun phrase, as well as the observations where both were realised as noun phrases (i.e. those described in Table 6.6).

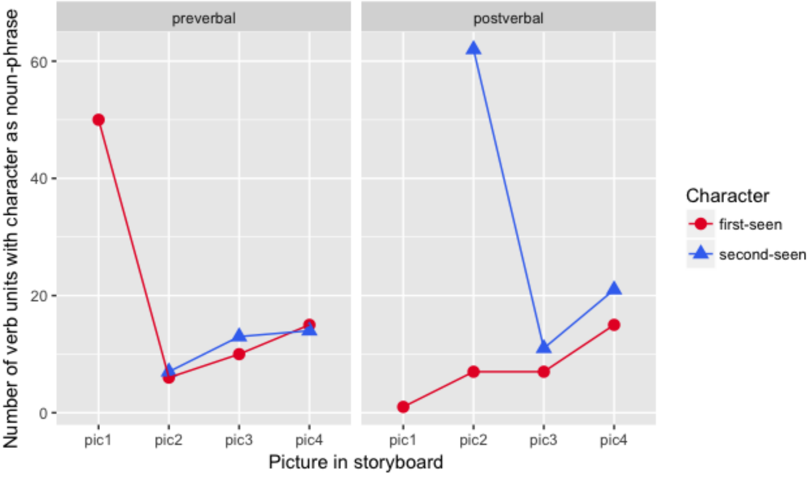


Figure 6.5: Distribution of preverbal vs. postverbal noun phrases. The left panel shows how often first-seen (red) and second-seen (blue) characters were realised as preverbal noun phrases; the right panel shows how often they were realised as postverbal noun phrases. This figure relates to the subset of data where one or both of the characters were realised as full noun phrase.

siders all pictures in the sequence (but, of course, restricts the data to verb units where a full noun phrase is found).

We see that when the character in question was the first seen, a noun phrase that refers to it is significantly more likely to be found in the preverbal position; conversely, a noun phrase referring to the second-seen character is more likely to occur in postverbal position. The intercept-only model is significantly improved by the inclusion of a fixed effect for first seen vs. second seen. This corroborates the observations based on Figure 6.5; based on the figure we can see that the effect is driven by the first two pictures of the sequence.

We may immediately wonder about the relationship of preverbal/postverbal position of noun phrases with grammatical function. For example, it could be the case that first-seen characters are introduced as subject, while second-seen characters are introduced as object, and that the preverbal/-postverbal pattern here is simply a side-effect of grammatical function assignment patterns. In other words, perhaps the relationship between linear position and first-seen vs. second-seen character is actually just down to a preference for certain grammatical functions to occupy certain linear positions. If this were the case – that is, if there were a confound between subject assignment and preverbal/postverbal positioning – we would expect to see the same pattern reflected in the subject assignment data (above). In particular, we would expect to see that in the second picture, the second-seen character is much less often expressed as subject of the verb unit. However, this is not the case. Although the first seen overall has a slightly higher incidence of subject assignment, it is not a big discrepancy. In conclusion, it does not seem that the preverbal or postverbal positioning of noun phrases is underpinned by the grammatical function of those noun phrases.

	intercept only	ORDER OF NP AND V model
(Intercept)	-0.07	-1.02^{***}
	(0.13)	(0.22)
first-seen		2.02^{***}
		(0.30)
ANOVA (model comparisons)		$\chi^2 = 53.41$ $(p < 0.001)$
AIC	336.98	285.57
BIC	347.41	299.47
Log Likelihood	-165.49	-138.78
Num. obs.	239	239
Num. groups: Participant	26	26
Num. groups: Storyboard	3	3
Var: Participant (Intercept)	0.00	0.00
Var: Storyboard (Intercept)	0.00	0.02

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 6.5: Logistic regression model of argument position relative the the verb, across all pictures of the story. The analysis models how the likelihood of a character being realised in preverbal position rather than postverbal is affected by whether it is first seen rather than second seen.

Verb units with both characters as noun phrase

Before concluding this section, it is important to note that there are only 24 verb units in the entire dataset where *both* first-seen and second-seen characters are realised as full noun phrases. In other words, in the vast majority of verb units where both characters are mentioned, at least one of the characters is realised in attenuated form. Table 6.6 provides a closer examination of noun phrase positioning relative to the verb in this small subset of cases where both characters are mentioned as a full noun phrase.

	preverbal (as subject)	postverbal (as subject)
first-seen	14 (14)	10 (1)
second-seen	8 (8)	16 (1)

Table 6.6: Position of noun phrases relative to the verb when both characters are mentioned as full noun phrases (n=24). Figures in brackets indicate the number of observations in each cell where the character in question was assigned subject.

Overall, when both characters are mentioned, they tend to also be split across the verb (in only four of these 24 cases did both noun phrases appear postverbally). It is interesting to observe that this small subset of data suggests a strong correlation of preverbal position with subjecthood; that is, *all* preverbal noun phrases here relate to the grammatical subject (which, as seen above, was more often the first-seen character). However, this correlation of preverbal position with subject function does not seem to be borne out by the dataset as a whole.

6.5 Discussion

Firstly, I aimed to assess whether **subject assignment** would be affected by whether a character was first seen or second seen in the story; in other words, whether the foregrounding of a referent meant that it would be significantly more likely to take subject function in the ensuing story (the second, third and fourth pictures). My hypothesis here was that first-seen referents would more often show up as grammatical subject than second-seen referents. The plot and the analysis indicated an overall trend for this to be the case; however, model comparisons revealed that the effect of first seen rather than second seen was at most marginally significant. Therefore, although there is a numerical trend for first seen as subject throughout the story, this non-significant result does not give us a strong basis on which to say that subject assignment in Tarifiyt is affected by foregrounding.

Secondly, I examined the impact of foregrounding on **forms of reference** once both characters had been introduced. Specifically, I wished to assess whether being first seen affected the likelihood of a character being realised using a full noun phrase or just attenuated form, in the third and fourth pictures of the sequence. My hypothesis here was that the first-seen referent would be more often realised in attenuated form than the second-seen referent. This hypothesis was borne out by the data, even in the later pictures of the story. The results indicated that even after both referents have been introduced, first-seen characters are significantly more likely to be realised in attenuated form, and second-seen characters are more likely to be referred to with full noun phrases.

Thirdly, I aimed to assess whether the foregrounding manipulation impacted the **linear ordering** of the responses, in terms of the position of noun phrases in the descriptions produced. In particular, I examined the position of noun phrases relative to the verb, and how this related to whether the referent was first seen or second seen. My hypothesis was that a preverbal position would be more likely for the first-seen referent than for the second-seen referent. Visualisation and analysis of the data indicated that when the first character is introduced, it is more likely to be mentioned as a preverbal noun phrase, and when the second character is introduced, it is more likely to be mentioned as a postverbal noun phrase. In the first two pictures, there was a strong relationship of firstness and preverbal/postverbal position of noun phrase; however, in the third and fourth pictures the incidence of preverbal versus postverbal noun phrases was comparable for both characters. Therefore, while the hypothesis is borne out where the characters are described for the first time, this effect does not appear to be a lasting one throughout the story.

Overall, we see the impact of foregrounding on the realisation of referents in the stories told by participants, despite the exuberance and minimal control. In general, the data supports the hypotheses that the foregrounded referent is prioritised in terms of linear order (preverbal position) and that the foregrounded referent is more likely to be referred to in attenuated form, and there is some indication that it is also prioritised in terms of realisation as subject. The question is then how we account for these findings. That is, does the data support the idea that speed (or ease) of information retrieval drives the realisation of referents?

Under the Accessibility Hypothesis, the proposal would be that the first-seen referent stays more active in memory throughout the story, and that this accounts for the difference in how it is realised compared to the second-seen referent. With regard to **grammatical function assignment**, under this account the first-seen referent is the first one assigned to a grammatical function, resulting in its assignment to subject (as described in Section 1.2.2). The data lends support to this idea; although the higher rate of subject assignment was not statistically significant, the trend was carried throughout the story and was not isolated, for example, to the first picture. With regard to **form of reference**, it is argued that more accessible referents are more likely to be referred to using attenuated forms such as pronouns or zero anaphora (Ariel, 1991). Here, the data align with an accessibility account, since reference to the first-seen character is more often in attenuated form. In terms of **linear ordering**, information retrieval accounts view the order of referents in a sentence as arising from the relative ease (or speed) with which they can be retrieved from memory (cf. Section 1.2.2). That is, a more easily retrieved referent occupies an earlier position because it was retrieved earlier, and therefore output earlier. On the face of it, the finding that the foregrounded referent appears more often in sentence-initial (preverbal) position appears to align with this view. However, when we consider that it is possible in Tarifiyt for both noun phrases to appear postverbally, the pattern of data becomes more difficult to account for.

Generally speaking, it is difficult to account for verb-initial vs. argument-initial variation through recourse to information retrieval, because in verb-initial languages, position relative to the verb can vary independently of position relative to other referents (for example, both NP^1-V-NP^2 and V-NP^1-NP^2 orders are possible, as shown in the current dataset). Previous authors have, however, suggested a possibility to account for preverbal/postverbal position in cognitive terms. In particular, the preverbal placement of a referent could serve to facilitate processing of subject-object relations, especially when both referents are animate (cf. Gennari, Mirković, & MacDonald, 2012). In a *simply describing* study of sentence production and planning in a verb-initial language, Tzeltal, Norcliffe and colleagues (2015) found that participants produced verb-initial sentences more often when subject was human and object was non-human, while the highest rate of subject-initial sentences was found when *both* referents were human. The authors interpreted this finding as evidence that speakers preferred to separate similar arguments to avoid interference (Norcliffe, Konopka, et al., 2015:16). In other words, avoidance of interference between arguments was taken to be the mechanism underlying the variation between subject-initial and verb-initial responses.

This account of word order variation in verb-initial languages is compelling, in that it offers a cognitive basis for the variation between verb-initial and argument-initial orders. In other words, it is fairly straightforward to incorporate an account based on interference in argument processing with the overarching picture of information retrieval as the driver of sentence form phenomena. Unfortunately, it is difficult to account for the current data in these terms. In the first place, the highest rate of preverbal subjects in the Tarifiyt study was found when a single referent was being introduced. Here, referent differentiation could not be playing a role in the choice between verb-initial and subject-initial forms. When introducing the second character – a point where differentiation could become important – postverbal subjects were often used to encode the second-seen character, even though in the majority of cases a human-human interaction was being described. A caveat is in order: in the Tzeltal study, participants described isolated pictures, encoding *both characters as full noun phrases* in their descriptions. In the current study, most responses displayed use of attenuated forms of reference. However, it is clear that argument differentiation cannot completely account for preverbal/postverbal ordering phenomena, at least in Tarifiyt.

How else could we account for the pattern of data seen in the current study? One possibility is that information status of the referent is important in deciding the relative placement of noun phrase and verb. To explore this idea further, we can begin by drawing on a descriptive account of verb-initial vs. subject-initial orders in Berber, provided by Kossmann's study of Figuig Berber (2016). Kossmann suggests that subject-initial order is typical of sentences where all information is new (so-called 'thetic' sentences). By contrast, when lexical subjects provide new information in the discourse, Kossmann notes that they typically occur postverbally. This again corroborates the finding from the current study: we saw that the second-seen character was more often introduced into the discourse as subject, but with a postverbal noun phrase. In addition, it is interesting to observe that the situation where all information is new – corresponding to 'thetic' sentences – is exactly the situation that was represented in the study in Chapter 3 on Tarifiyt. In each trial, participants responded to an image that they had not seen before, in an isolated context. Correspondingly, in that experiment, the number of verb-initial responses was extremely low.

When Chapter 3 is considered alongside the current study, we can suggest that in order to gain insight into the psycholinguistics of verb-initial ordering, it will be important to conduct (controlled) studies that contrast the linguistic form of responses between conditions where one, both, or neither of the referents have been introduced in prior discourse. Moreover, this line of research would benefit from comparative studies between different languages with verb-initial orders, such as Tzeltal and (Tarifiyt) Berber. In doing so, it is also important to be mindful when interpreting results from isolated picture descriptions. In particular, if we consider that isolated pictures represent a specific kind of information structural context (as suggested by the discussion of 'thetic' sentences), then discourse context should really be borne in mind as a relevant variable in any *simply describing* experiment, even when using isolated pictures. Essentially, there is still a discourse context represented by such pictures: a context where everything is new.

Before concluding, it is important to reflect on the methodological aims of this study. On the one hand, this study demonstrated that exuberant data can also be a source of insight into the manner in which message-related variables impact on linguistic form. On the other hand, despite the alternative manner of extracting the relevant measures, the exuberance (or messiness) of the data still poses difficulties. An example of this is the issue of how to deal with dependent clauses (sub-clauses) in the data – here, the decision was taken to ignore them in order to be able to focus on main clauses. However, such a decision might seem to contradict the fundamental aim of inclusiveness.

Furthermore, this approach offers a way to reduce bias surrounding the researcher's expectations of seeing certain 'structures'; however, it is not the case that all possibility of bias is removed. It is true that the coding procedure no longer requires the judgement about whether responses conform to structures, or indeed, which structures are relevant to the analysis. However, there is of course still scope for subjective bias in the decisions taken in coding and extracting measures. For example, the decision was taken here to treat subject agreement as a form of attenuated reference, in line with Galand (1964) (cf. Section 2.3.1); however, an alternative would be to instead contrast full noun phrases with pronominal forms (clitics and free pronouns), viewing subject agreement as a form of pro-drop. Ultimately, all forms of coding will be at some level prone to influence from bias; after all, even the most fine-grained of annotation involves interpretation (Lüdeling, Ritz, Stede, & Zeldes, 2016:608).

Overall, this approach leans more towards the methodology used in corpus-based approaches, particularly task-based corpora (Lüdeling et al., 2016). When viewed in this way, it is not parsimonious to approach the coding of the data with a completely new set of labels, given the range of annotation and tagging approaches already described in the literature. Nonetheless, similar caveats apply as before, regarding the application of existing labels to a 'new' language. In any case the coding procedure used here is advantageous in that, being written in Python, the annotation proceeds automatically, meaning it can be easily applied to a larger dataset (although this of course does not dispense with the need for the researcher to check the output for consistency and appropriateness; Lüdeling et al., 2016:610). Such an ‘experimental corpus’ approach would seem to be most useful for beginning work with understudied languages, where reduction of bias effects and maximisation of data yield is a priority. It can offer a bridge between more descriptive and more cognitive-theoretical work, by providing a semi-controlled insight into the data, which can be used to generate hypotheses to be tested in a more controlled setup. For example, based on the findings from this study, one could generate hypotheses for an experiment focusing on the production and perception of verb-initial and subject-initial sentences in a controlled set of contexts, manipulating animacy and discourse status. It would be hard to see how appropriate hypotheses for such an experiment could be generated purely on the basis of a more traditional study, such as reported in Chapter 3.

Perhaps the most pressing limitation of this study is that there are only a few stimuli and they are rather heterogeneous: for example, the animacy of referents and the types of actions differs between stories. Nonetheless, it should be noted that the experimental manipulation in this experiment concerned the contrast between two versions of the same story, featuring the same referents. Perhaps more troublesome is that the stimuli in this study depicted interactions of two referents with an inanimate object, thereby adding another referent into the mix. In designing the stimuli, this setup was chosen in order to have the freedom to design scenes that would be clear and straightforward for participants to describe, thereby minimising between-participant variation in how the scenes were perceived (i.e. to address Bock’s “comprehension contamination” problem; Bock 1996). Nonetheless, in order to develop this approach further, it would be advantageous to revisit the design of the stimuli, to see how a simpler set of stories can be created without increasing the risk of comprehension contamination.

6.6 Concluding remarks

In this chapter, I first gave an overview of some methodological issues encountered in previous chapters, including problems surrounding the (pre-)processing of *simply describing* data, exuberance and experimental control, and data loss. I argued that solving these problems is particularly important in order to facilitate the movement towards the inclusion of a wider range of diverse languages and speaker communities in psycholinguistic research. I then implemented a study in Tarifiyt Berber, with the aim of gaining further insight into the impact of referent foregrounding on grammatical encoding (grammatical function assignment, linear order and forms of reference), while not being hindered by the issues described. While foregrounded referents were encoded in a way that aligned with information retrieval accounts, the findings regarding the preverbal or postverbal encoding of referents were difficult to account for under purely processing terms. Overall, further research is needed to develop a full account of the cognitive underpinnings of the placement of noun phrases relative to the verb in languages which permit verb-initial ordering. While the method used in this chapter partially achieved its aims of minimising data loss and control while allowing exuberance, there is still much more to be done to fully address these issues.

CHAPTER 7

Conclusion

Across all human languages, speakers are able to organise their conceptual messages into ordered, structured phrases that can be successfully comprehended by listeners in real time. To understand the mechanisms by which this process works, it is important to understand what the process *is*. Given the wide range of variation between languages, it is important to be able to identify the aspects of the process that are constrained by language form, and the parts of the process which are influenced or guided by general cognitive principles. Clearly, if we wish to understand how humans formulate sentences in the service of communicating conceptual content, we must take care to design our research to engage with, rather than generalise over, linguistic variation. However, when investigating language production and perception from such a perspective, practical and methodological challenges arise from the fact that psycholinguistic research has focused on a small fraction of the world's linguistic diversity. It is for this reason that the current thesis has presented research questions that have both theoretical and methodological components. In this concluding chapter, I summarise the findings of the foregoing studies, in terms of these two threads of the thesis.

7.1 Theoretical considerations

The psycholinguistic account of the way that speakers organise elements into felicitous grammatical sentences has paid particular attention to the linguistic dimensions of grammatical function assignment and linear order. The input variables that appear to influence these output dimensions include topicality, givenness and animacy. With regard to these variables of interest, the psycholinguistic study of sentence production and the descriptive realm of information structure overlap to some degree. However, unlike information structure, the study of sentence production aims to explain these phenomena in terms of cognitive variables. In this, accounts based on information retrieval, such as the Accessibility Hypothesis, form the basis for psycholinguistic accounts. One key distinction between psycholinguistic and descriptive approaches is that the notion of topic, which plays a fundamental role in descriptive theories of information structure, has fallen out of favour in cognitive explanations of sentence production phenomena. Leading accounts of sentence form variation propose that formulation is fundamentally a reflection of information retrieval processes. Under this account there is nonetheless some flexibility in viewing sentence formulation as element-driven, that is, falling out from the relative properties or status of the verbal arguments to be encoded, or as structure-driven, meaning that the production process also considers the felicity of different structural forms.

In terms of the standard model of sentence production, it is not yet clear about exactly how language-specific differences should factor into the process. For example, it is not clear how the production system assesses which sentence elements are viable as starting points in different languages under different circumstances. It is also not clear how we can account for the fact that different languages favour different grammatical phenomena (passives, object-topicalisations) under the same communicative conditions. Therefore, the cross-linguistic perspective of this thesis centred on these issues.

In **Chapter 3**, I began by revisiting an existing dataset from a *simply describing* study in Tarifiyt Berber and Dutch, presenting a new analysis using appropriate techniques, and a new cross-linguistic perspective on the insights provided by the data. This study was designed to investigate the effect of patient animacy (relative to the agent) on the form of transitive picture descriptions. The findings were that in both languages, there was an overall predominance of the subject-initial active, coupled with an increased probability of patient-prominent structures in the animate patient (IA) condition compared with the inanimate patient condition (II). In addition, cross-linguistic comparison indicated the potential for a difference in magnitude of the effect, but this difference was not statistically supported.

The discussion in this chapter concerned the complementary profiles of the two languages, in terms of the linguistic forms used to encode patient prominence. Although these two languages displayed linguistic phenomena that are familiar in the study of sentence production (passivisation and object-topicalisation), the same communicative circumstances seem to lead to complementary outcomes in terms of form in these two languages. The classic model of sentence production (Bock & Levelt, 1994) does not make explicit mention of how cross-linguistic differences like these arise; however, a couple of suggestions have been made by previous authors. Ultimately the account which is best motivated on cross-linguistic grounds is one that draws on a well-known dimension of typological variation, namely subject-prominence and topic-prominence (Li & Thompson, 1976). How exactly we may integrate this account into the existing model of sentence production is something that requires further comparative experimental studies to clarify. However, a suggestion was made as to how we could design future studies to evaluate the usefulness of this account. Regardless of whether this account proves to be a source of insight in the future, it is clear that cross-linguistic work in this area would benefit from refining the sentence production model to clarify how general cognitive principles interface with language-specific properties.

In **Chapter 4**, questions regarding the interplay of linguistic form and cognitive principles were approached from the point of view of the listener. In particular, I investigated the way in which listeners are sensitive to linear order and animacy, as they interpret thematic structure. The sentences presented in this experiment were the kind of sentences produced by participants in Chapter 3. Moreover, this was studied in the same two languages, Tarifiyt Berber and Dutch, which exhibit different profiles as regards grammatical function assignment, another important variable for the study of sentence form.

A key question of this study was whether listeners would simply combine prominence features (early sentence position, highly animate) to identify the agent, or whether their predictions about thematic structure would involve weighing up available information against probabilities in the input. The findings supported the latter account. On the one hand, I found strong evidence for the preference of listeners to assign agent role to an ambiguous sentence-initial noun phrase in both Tarifiyt Berber and Dutch, in line with previous research. However, when these agents were also animate, this did not increase the likelihood of being interpreted as agent. In fact, it had no effect on the likelihood of agent interpretation in Tarifiyt and even reduced the likelihood of agent interpretation in Dutch. These findings align with accounts of comprehension where the parser has early access to a range of available information, and these sources interface with knowledge of the linguistic input, as part of a probabilistic process of thematic interpretation. Findings from cross-linguistic, cross-modality comparisons were interpreted as providing further evidence for a predictive account where listeners demonstrate sensitivity to the possibilities for felicitous sentence completion in their respective languages.

In **Chapter 5**, I investigated the effects of animacy on sentence form choice in the case of more than two possible categories of structure, in Pondok Tinggi. This was a replication of the study in Chapter 3, in a language where there exists more than one structure providing linguistic prominence for the patient argument. Although these different structural categories all offered prominence to the patient, the likelihood of choosing these structures was not uniformly impacted by animacy. If the choice of patient-prominent structures is attributed to accessibility, it is difficult to see how we can account for discrepancies in choices *among* patient-prominent structures. The discussion in this chapter raised the possibility that such variations in sentence form reflect other variables that are not reducible to information retrieval speed, such as the affectedness of the patient and the volition of the agent. Of course, the immediate issue with such variables is the question of how we can conceive of them in cognitive terms. A possible way to proceed here is to examine whether such variables show effects in languages already familiar to psycholinguistic research (such as English, Dutch, German and Spanish) as well as more languages with typological profiles similar to Pondok Tinggi. A *simply describing* experiment that holds known accessibility variables constant while varying the degree of adversity for the patient, or the degree of volition for the agent, could form a first step for the controlled study of such variables.

The study in **Chapter 6** was again a *simply describing* study in Tarifiyt Berber, but approached from a different methodological perspective (summarised again here in Section 7.2). This alternative approach to coding and analysis opened the way to ask questions that were not possible in the study in Chapter 3. The overall aim was to investigate the effects of foregrounding on linguistic form, and whether the findings would conform with the expectations from information retrieval accounts of sentence formulation. In particular, I wished to examine whether subject function was still assigned to more contextually prominent (foregrounded) referents in Tarifiyt Berber, despite the fact that this would not manifest in passivisation (as revealed by the study in Chapter 3). In addition, I wished to examine whether foregrounded status would lead to a referent being more often realised in preverbal position. Given the fluent nature of the responses, it was also expedient to examine the forms of reference used. Generally speaking, the findings aligned with accessibility accounts; however, on looking more closely, it proved difficult to account for the linear ordering findings in purely information retrieval terms. In particular, preverbal position seemed to be associated with a specific discourse status, rather than falling out from processing effects.

In summary, these studies corroborated previous findings related to the impact of variables such as animacy and foregrounding on the linguistic form of sentences, including grammatical function assignment and linear order. However, in terms of the theoretical accounts for these phenomena, the studies raised some questions. Firstly, the Accessibility Hypothesis view of sentence production, whereby information retrieval is the essential driver of sentence form variation, was not always sufficient to account for the findings. Ultimately, the limitation of the information retrieval account is that it seems to require that we account for the vast complexity of cross-linguistic variation with reference to a single latent variable. One possibility to alleviate this issue is to propose that there are different 'sources' of accessibility, along the lines suggested by Christianson and Ferreira in their study of sentence form in a structurally complex language, Odawa (2005; see also Section 1.2.2). However, if different 'sources of accessibility' have qualitatively different effects, this ultimately raises the question of whether these are facets of a single dimension of accessibility, after all.

Secondly, the findings of the studies in Tarifiyt Berber provide compelling evidence to re-evaluate the status of topic in the psycholinguistic account of sentence production. While this concept is notoriously tricky, the observation that some languages seem to demonstrate a formal encoding of topic means that we may need to revisit the status of topicality in cognitive terms. Particularly the typological generalisations of subject-prominence and topic-prominence could provide a framework within which to generate cross-linguistic hypotheses with a view to elucidating this area further.

7.2 Methodological considerations

In **Chapter 2**, I described previous work that has proposed the benefits of diversity in psycholinguistic research, demonstrating that the role of linguistic diversity in the field has been a topic of discussion more than just recently. I argued that there a number of challenges involved in doing psycholinguistic research on a wider range of languages, and that these mostly relate to the the fact that we have only minimal research on most of the world's languages, and that in order to research these languages we need to find means to conduct research within the communities where those languages are spoken. I suggested that discussions of this topic in the field have tended to focus on the practical and logistical issues associated with field-based work, but demonstrated that this only scratches the surface. In particular, the weight of extra challenges (and the investment needed to engage with these challenges) acts as a deterrent to researchers from pursuing this kind of research, particularly when results may have unclear bearing on theory. In addition, the skew of prior research towards a handful of familiar languages leads to bias which can permeate through all stages of study design, influencing not only the languages we choose to study, but also the manner in which we plan, conduct and interpret our experimental studies.

In terms of the former set of problems, I followed authors such as Norcliffe, Harris, and Jaeger (2015) in underscoring the need for investment, not only regarding the extra time and budgetary costs associated with this research, but also regarding the training and support that psycholinguistic researchers require in order to engage with a diverse linguistic sample. In terms of the latter set of problems, I aimed to distil and explore specific questions relating to the issues of study design, data processing and analysis in Chapters 3 to 6 of the thesis.

In **Chapter 3**, I highlighted the choices required in preprocessing and analysing raw sentence production data from *simply describing* experiments. Participants are instructed to produce just one sentence to describe the action depicted in a transitive scene. The responses are screened as to whether they fulfil certain criteria; in such experiments, there is known to be a high rate of data loss. In this study, the rate of data loss was 39% for Tarifiyt and 34% for Dutch. In addition, the exclusion and coding process is completed by hand, with the researcher (or assistants) assessing each response for its criteria fulfilment and relevance to the question under study. As part of this process, the fluent responses are streamlined to achieve dichotomous classification(s), rendering them amenable to statistical analysis with mixed effects binary logistic regression. This preprocessing stage, especially the process of classification, requires us to generalise or abstract in order to bring the data into an analysable form. This is true of all data coding procedures to some extent, but *simply describing* stands out in that the raw data may be extremely heterogeneous and exuberant (Bock, 1996), and yet it is typically reduced down to a (set of) binary variable(s). Clearly there is scope here for the researcher's choices to shape the processing of data. Overall, this chapter served to exemplify some of the issues raised in Chapter 2. It therefore concluded by sketching out a more refined set of questions to be taken forward for examination in the subsequent chapters.

The methodological value of **Chapter 4** lay in its aim to investigate online sentence processing, using techniques that are appropriate for non-lab based work. This study was conducted within the speaker community in Morocco, with the interaction with participants being conducted entirely in the target language. The approach to recruitment, informed consent and compensation was carefully thought out with attention to the cultural setting. Given that one of the two languages studied was predominantly oral (Tarifiyt), the written modality was eliminated from the entire study as far as practicable. The data collection technique employed used minimal equipment, which was relatively cheap, portable, and familiar for the participants. Nonetheless, this method yielded rich data, with a range of measures providing a range of data to address the research questions. However, this method of course faced a number of limitations: although care was taken to design an accessible study, the task was still rather circumscribed (in the sense of S. Chung 2012). Moreover, there may be other forms of equipment that can be used to acquire similar measures, with better ecological validity and temporal resolution.

In **Chapter 5**, I reflected on the binary approach to analysis used in previous *simply describing* studies. In the past, binary analysis techniques have been deployed to gain understanding of how speakers select structural forms in languages where the choice that speakers face is not truly dichotomous. The practice of analysing such data in a binary form may relate to the desire for continuity with previous studies, but likely also relates to the complexity of analysis required to consider multiple outcome categories. Binary analysis is certainly appropriate in a range of situations, but there are a number of limitations that we should bear in mind. These concern loss of information through dichotomising the data, the scope for bias due to the need to generalise, and the potential for spurious results. Overall, these issues form a strong case to consider other forms of analysis, especially when the motivation for using binary analysis is practically rather than theoretically motivated. In this chapter I therefore employed multi-categorical regression to answer research questions where sentence production is viewed as multiple rather than binary choice. I found that through applying this technique, it was possible to progress beyond simply ascertaining whether one type of structure was more likely than another, to modelling how choice between multiple structures varied as a function of the experimental manipulation.

In **Chapter 6**, I reflected on the difficulties of working with fluent production data, especially when participants respond exuberantly. I suggested that the high rates of data loss in *simply describing* experiments arise in part from the need for fluent responses to conform to a constrained target form, in order to have data that is amenable to analysis. I argued that these issues will come ever more to the forefront as we work with more diverse languages and speaker communities. For this reason, I aimed to execute a study, outside the lab, in which I investigated the variables familiar from previous chapters, but where I allowed for response exuberance, rather than attempting to rein it in. In order to do so, I explored how we could obtain dependent measures without the need for response classification, meaning that we bypass the need to judge whether the response conforms to a certain 'template'. This has the additional benefit that we also bypass the need to pre-specify the kinds of structure (or 'template') we are looking for, thus reducing the scope for researcher bias to shape what the study shows us. However, it must be noted that this approach does not fully circumvent the bias issue – this is probably ultimately impossible, since every form of data analysis requires some form of coding and therefore some form of abstraction. This point can be taken to additionally underscore the importance of typological (linguistic) training and support for psycholinguists, in overcoming issues of bias and skew.

In summary, the limited linguistic diversity in psycholinguistic research is widely recognised as problematic, but the specific challenges involved in this work may have the effect of slowing or preventing the necessary changes in the field. I have argued that addressing the logistical and practical issues of work outside the lab is only part of the story. It is of fundamental importance to evaluate the choices we make in experimental design, data processing and statistical analysis. Over time, a skewed sample affects all stages of the research cycle, as follows. Our theory is based on our data. When we generalise across human language from an unrepresentative sample then our theory runs the risk of bias. We derive research questions and hypotheses based on our theory: if our theory is biased, the same bias will also colour the questions we think of asking. In turn, we design our experimental paradigms in order to test our hypotheses. If our hypotheses derive from biased theory, it is likely that our experimental paradigms will have similar limitations. In order to fully address issues of bias, it is important to encourage critical reflection regarding research methodology, while also emphasising the value of typological linguistic study for the field of psycholinguistic research, with an important role for interdisciplinary dialogue and collaboration.

Appendix to Chapter 6: stimuli and data excerpt

Storyboard stimuli used in the experiment reported in Chapter 6.

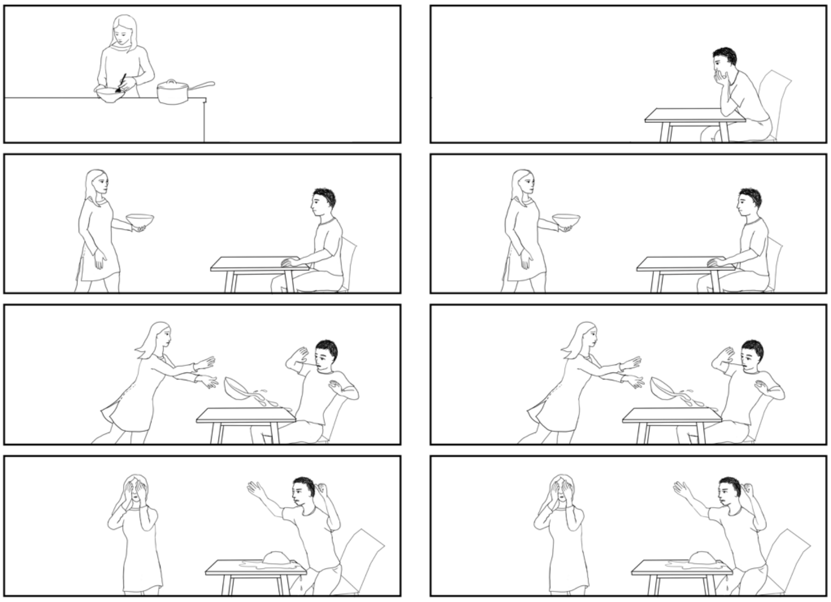


Figure 1: kitchen story

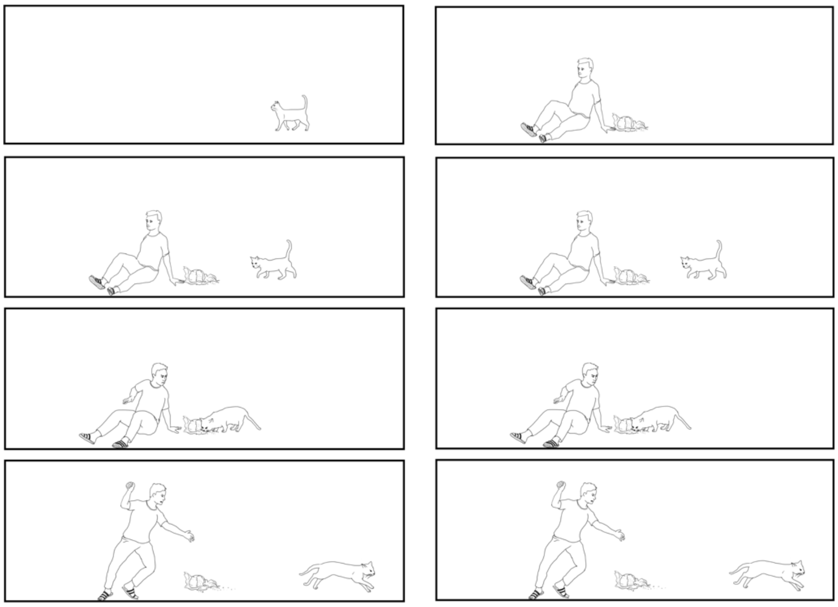


Figure 2: bread story

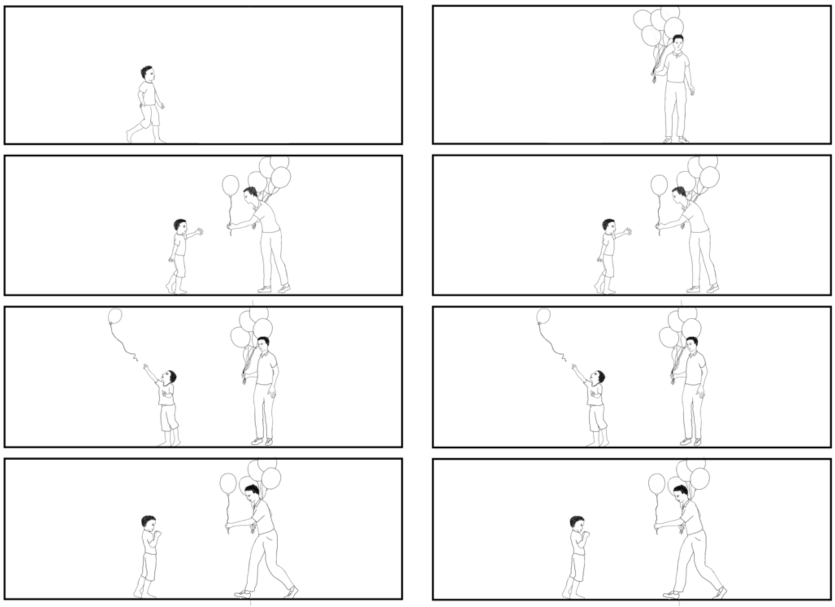


Figure 3: balloon story

Participant ID	Storyboard	Picture #	Verb-unit #	Response	Broad translation	Glossed response
018	kitchen	pic1	1	ijj n waryaz iqqim x řkursi	a man is sitting on [a] chair	[’ijjn(man)NP’, ’(man)Sverb’, ’PREP’, ’location’]
018	kitchen	pic2	1	tiwy-as-d ijj n temɣaaṯ afenjar	a woman brings him a bowl	[’(woman)Sverb-(man)IO-D’, ’ijjn(woman)NP’, ’(food)NP’]
018	kitchen	pic3	1	tamɣart-nni tzedjeɛ afenjar x waryaz	the woman spills the bowl on the man	[’(woman)NPnni’, ’(woman)Sverb’, ’(food)NP’, ’PREP’, ’(man)NP’]
018	kitchen	pic4	1	aryaz-nni itmenɣa ak temɣarṯ-nni	the man is fighting with the woman	[’(man)NPnni’, ’(man)Sverb’, ’PREP’, ’(woman)NPnni’]
018	balloon	pic1	1	ijj n waryaz iksi glubut, aṭṭas n glubuyat	a man carries balloons, a lot of balloons	[’ijjn(man)NP’, ’(man)Sverb’, ’(balloon)NP’, ’QUANT’, ’(balloon)NP’]
018	balloon	pic2	1	yiwc-as ijj n glubu i ijj uḥenjir	he gives a balloon to a boy	[’(man)Sverb-(boy)IO’, ’ijjn(balloon)NP’, ’INDIRECT’, ’ijjn(boy)NP’]
018	balloon	pic3	1	aḥenjir-nni iḍwa-s glubu-nni	the balloon flies away from the boy	[’(boy)NPnni’, ’(balloon)Sverb-(boy)IO’, ’(balloon)NPnni’]
018	balloon	pic4	1	yeɛni yiwc-as ijj n glubu nneḍni	so, he gives him another balloon	[’DISC’, ’(man)Sverb-(boy)IO’, ’ijjn(balloon)NP’, ’ADJ’]
018	bread	pic1	1	ijj n waryaz ṭarf nnes lmakla	a man [with] food next to him	[’ijjn(man)NP’, ’PREP(man)’, ’(food)NP’]
018	bread	pic2	1	ijj umucc yus-d ɣar lmakla-nni	a cat approaches the food	[’ijjn(cat)NP’, ’(cat)Sverb-D’, ’PREP’, ’(food)NPnni’]
018	bread	pic3	1	mucc-nni itett lmakla-nni	the cat is eating the food	[’(cat)NPnni’, ’(cat)Sverb’, ’(food)NPnni’]
018	bread	pic4	1	aryaz-nni yuzzeř x umucc	the man runs after the cat	[’(man)NPnni’, ’(man)Sverb’, ’PREP’, ’(cat)NP’]

Table 1: Illustration of glossing procedure in Chapter 6 (excerpt represents one participant’s response set).

Participant ID	Storyboard	Picture #	Verb-unit #	Character	Character:First/Second	Glossed response	Character:subject/non-subject	Character:form of reference	NP:prepostV
018	balloon	pic1	1	man	first-seen	[’ijjn(man)NP’, ’(man)Sverb’, ’(balloon)NP’, ’QUANT’, ’(balloon)NP’]	subject	NP	preverbal
018	balloon	pic2	1	man	first-seen	[’(man)Sverb-(boy)IO’, ’ijjn(balloon)NP’, ’INDIRECT’, ’ijjn(boy)NP’]	subject	attenuated	–
018	balloon	pic3	1	man	first-seen	[’(boy)NPnni’, ’(balloon)Sverb-(boy)IO’, ’(balloon)NPnni’]	non-subject	–	–
018	balloon	pic4	1	man	first-seen	[’DISC’, ’(man)Sverb-(boy)IO’, ’ijjn(balloon)NP’, ’ADJ’]	subject	attenuated	–
018	balloon	pic1	1	boy	second-seen	[’ijjn(man)NP’, ’(man)Sverb’, ’(balloon)NP’, ’QUANT’, ’(balloon)NP’]	non-subject	–	–
018	balloon	pic2	1	boy	second-seen	[’(man)Sverb-(boy)IO’, ’ijjn(balloon)NP’, ’INDIRECT’, ’ijjn(boy)NP’]	non-subject	NP	postverbal
018	balloon	pic3	1	boy	second-seen	[’(boy)NPnni’, ’(balloon)Sverb-(boy)IO’, ’(balloon)NPnni’]	non-subject	NP	preverbal
018	balloon	pic4	1	boy	second-seen	[’DISC’, ’(man)Sverb-(boy)IO’, ’ijjn(balloon)NP’, ’ADJ’]	non-subject	attenuated	–
018	bread	pic1	1	man	first-seen	[’ijjn(man)NP’, ’PREP(man)’, ’(food)NP’]	non-subject	NP	–
018	bread	pic2	1	man	first-seen	[’ijjn(cat)NP’, ’(cat)Sverb-D’, ’PREP’, ’(food)NPnni’]	non-subject	–	–
018	bread	pic3	1	man	first-seen	[’(cat)NPnni’, ’(cat)Sverb’, ’(food)NPnni’]	non-subject	–	–
018	bread	pic4	1	man	first-seen	[’(man)NPnni’, ’(man)Sverb’, ’PREP’, ’(cat)NP’]	subject	NP	preverbal
018	bread	pic1	1	cat	second-seen	[’ijjn(man)NP’, ’PREP(man)’, ’(food)NP’]	non-subject	–	–
018	bread	pic2	1	cat	second-seen	[’ijjn(cat)NP’, ’(cat)Sverb-D’, ’PREP’, ’(food)NPnni’]	subject	NP	preverbal
018	bread	pic3	1	cat	second-seen	[’(cat)NPnni’, ’(cat)Sverb’, ’(food)NPnni’]	subject	NP	preverbal
018	bread	pic4	1	cat	second-seen	[’(man)NPnni’, ’(man)Sverb’, ’PREP’, ’(cat)NP’]	non-subject	NP	postverbal
018	kitchen	pic1	1	man	first-seen	[’ijjn(man)NP’, ’(man)Sverb’, ’PREP’, ’location’]	subject	NP	preverbal
018	kitchen	pic2	1	man	first-seen	[’(woman)Sverb-(man)IO-D’, ’ijjn(woman)NP’, ’(food)NP’]	non-subject	attenuated	–
018	kitchen	pic3	1	man	first-seen	[’(woman)NPnni’, ’(woman)Sverb’, ’(food)NP’, ’PREP’, ’(man)NP’]	non-subject	NP	postverbal
018	kitchen	pic4	1	man	first-seen	[’(man)NPnni’, ’(man)Sverb’, ’PREP’, ’(woman)NPnni’]	subject	NP	preverbal
018	kitchen	pic1	1	woman	second-seen	[’ijjn(man)NP’, ’(man)Sverb’, ’PREP’, ’location’]	non-subject	–	–
018	kitchen	pic2	1	woman	second-seen	[’(woman)Sverb-(man)IO-D’, ’ijjn(woman)NP’, ’(food)NP’]	subject	NP	postverbal
018	kitchen	pic3	1	woman	second-seen	[’(woman)NPnni’, ’(woman)Sverb’, ’(food)NP’, ’PREP’, ’(man)NP’]	subject	NP	preverbal
018	kitchen	pic4	1	woman	second-seen	[’(man)NPnni’, ’(man)Sverb’, ’PREP’, ’(woman)NPnni’]	non-subject	NP	postverbal

Table 2: Illustration of measures extracted from fluent response data in Chapter 6 (same excerpt as Table 1).

Bibliography

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley-Interscience.

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). New York: Wiley-Interscience.

Agresti, A., Booth, J. G., Hobert, J. P., & Caffo, B. (2000). Random-effects modeling of categorical response data. *Sociological Methodology*, *30*(1), 27–80.

Allum, P. H., & Wheeldon, L. R. (2007). Planning scope in spoken sentence production: The role of grammatical units. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(4), 791–810.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

Anand, P., Chung, S., & Wagers, M. (2011). *Widening the net: Challenges for gathering linguistic data in the digital age.* [White Paper 121]. National Science Foundation.

Ariel, M. (1990). *Accessing noun-phrase antecedents*. London; New York: Routledge.

Ariel, M. (1991). The function of accessibility in a theory of grammar. *Journal of Pragmatics*, *16*(5), 443–463.

Arnold, J. E., Kaiser, E., Kahn, J. M., & Kim, L. K. (2013). Information structure: Linguistic, cognitive, and processing approaches. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(4), 403–413.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX Lexical Database. Release 2 (CD-ROM).* Linguistic Data Consortium, University of Pennsylvania.

Barca, L., & Pezzulo, G. (2015). Tracking second thoughts: Continuous and discrete revision processes during visual lexical decision. *PLoS ONE*, *10*(2), e0116193.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bates, E., McNew, S., MacWhinney, B., Devescovi, A., & Smith, S. (1982). Functional constraints on sentence processing: A cross-linguistic study. *Cognition*, *11*(3), 245–299.

Bender, E. M. (2016). Linguistic typology in natural language processing. *Linguistic Typology*, *20*(3), 645–660.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: Analyzing text with the natural language toolkit*. Beijing: O'Reilly.

Bock, J. K. (1982). Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review*, *89*(1), 1–47.

Bock, J. K. (1986a). Meaning, sound, and syntax: Lexical priming in sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(4), 575.

Bock, J. K. (1986b). Syntactic persistence in language production. *Cognitive Psychology*, *18*(3), 355–387.

Bock, J. K. (1987). An effect of the accessibility of word forms on sentence structures. *Journal of Memory and Language*, *26*(2), 119–137.

Bock, J. K. (1996). Language production: Methods and methodologies. *Psychonomic Bulletin & Review*, *3*(4), 395–421.

Bock, J. K., & Ferreira, V. S. (2014). Syntactically speaking. In M. Goldrick, V. S. Ferreira, & M. Miozzo (Eds.), *The Oxford Handbook of Language Production.* Oxford: Oxford University Press.

Bock, J. K., & Irwin, D. E. (1980). Syntactic effects of information availability in sentence production. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 467–484.

Bock, J. K., & Levelt, W. J. M. (1994). Language production. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 945–984). San Diego: Academic Press.

Bock, J. K., & Warren, R. K. (1985). Conceptual accessibility and syntactic structure in sentence formulation. *Cognition*, *21*(1), 47–67.

Bohnemeyer, J. (2009). Linking without grammatical relations in Yucatec: Alignment, extraction, and control. In Y. Nishina, Y. M. Shin, S. Skopeteas, E. Verhoeven, & J. Helmbrecht (Eds.), *Issues in functional-typological linguistics and language theory: A Festschrift for Christian Lehmann on the occasion of his 60th birthday* (pp. 185–214). Berlin: Mouton de Gruyter.

Bonin, P., Peereman, R., Malardier, N., Méot, A., & Chalard, M. (2003). A new set of 299 pictures for psycholinguistic studies: French norms for name agreement, image agreement, conceptual familiarity, visual complexity, image variability, age of acquisition, and naming latencies. *Behavior Research Methods, Instruments, & Computers*, *35*(1), 158–167.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2009). The role of prominence information in the real-time comprehension of transitive constructions: A cross-linguistic approach. *Language and Linguistics Compass*, *3*(1), 19–58.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2013). Neurotypology: Modeling crosslinguistic similarities and differences in the neurocognition of language comprehension. In M. Sanz, I. Laka, & M. K. Tanenhaus (Eds.), *Language down the garden path: The cognitive and biological bases for linguistic structures* (pp. 241–252). Oxford: Oxford University Press.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2015). Scales in real-time language comprehension: A review. In I. Bornkessel-Schlesewsky, A. L. Malchukov, & M. Richards (Eds.), *Scales and hierarchies* (pp. 321–352). Berlin: De Gruyter.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2014). Competition in argument interpretation: Evidence from the neurobiology of language. In B. MacWhinney, A. Malchukov, & E. Moravcsik (Eds.), *Competing motivations in grammar and usage* (pp. 107–126). Oxford: Oxford University Press.

Boudelaa, S. (2013). Psycholinguistics. In J. Owens (Ed.), *The Oxford Handbook of Arabic Linguistics* (Vol. 1). Oxford: Oxford University Press.

Bowerman, M. (2010). Linguistic Typology and First Language Acquisition. In J. J. Song (Ed.), *The Oxford Handbook of Linguistic Typology* (pp. 591–617). Oxford: Oxford University Press.

Branigan, H. P., & Feleki, E. (1999). Conceptual accessibility and serial order in Greek language production. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the Twenty First Annual Conference of the Cognitive Science Society* (pp. 96–101). New Jersey; London: Lawrence Erlbaum Associates.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*(421), 9.

Brookshire, B. (2013). Social science is WEIRD, and that's a problem. *Slate Magazine*.

Brown-Schmidt, S., & Konopka, A. E. (2008). Little houses and casas pequeñas: Message formulation and syntactic form in unscripted speech with speakers of English and Spanish. *Cognition*, *109*(2), 274–280.

Brustad, K. (2000). *The syntax of spoken Arabic: A comparative study of Moroccan, Egyptian, Syrian, and Kuwaiti dialects*. Washington, DC: Georgetown University Press.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

Burdick, A. (2018). Why nouns slow us down, and why linguistics might be in a bubble. *New Yorker*.

Bürkner, P.-C. (2017). *Brms R package for Bayesian generalized non-linear multilevel models using Stan*. https://github.com/paul-buerkner/brms.

Butler, L., Jaeger, T. F., & Bohnemeyer, J. (2012). Animacy is mediated by topicality in the production of word order in Yucatec Maya and Spanish. Poster presented at the International Workshop on Language Production, New York University, NY.

Butler, L., Jaeger, T. F., & Bohnemeyer, J. (2014). *Effects of animacy on sentence production: Ease of retrieval or topicality?* [Unpublished Manuscript]. University of Rochester.

Cadi, K. (2005). *Transitivité et diathèse en tarifit : Analyse de quelques relations de dépendance lexicale et syntaxique*. Publications de l'Institut Royal de la Culture Amazighe.

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLOS ONE*, *5*(6), e10729.

Calabrese, A. (1987). Focus structure in Berber: A comparative analysis with Italian. In *Studies in Berber Syntax* (pp. 103–120). Cambridge, Mass: MIT Press.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32.

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, *10*(7), 335–344.

Christianson, K., & Ferreira, F. (2005). Conceptual accessibility and sentence production in a free word order language (Odawa). *Cognition*, *98*(2), 105–135.

Chung, S. (2008). How much can understudied languages really tell us about how language works? *Invited Plenary Talk at the Annual Meeting of the Linguistic Society of America. Chicago, IL.*

Chung, S. (2012). Bridging methodologies: Experimental syntax in the Pacific. *Presidential Address at the Annual Meeting of the Linguistic Society of America. Portland, OR.*

Chung, S.-F. (2005). Kena as a third type of malay passive. *Oceanic Linguistics*, *44*(1), 194–214.

Comrie, B. (1989). *Language universals and linguistic typology: Syntax and morphology*. University of Chicago Press.

Costa, A., Alario, F.-X., & Sebastián-Gallés, N. (2007). Cross-linguistic research on language production. In *The Oxford Handbook of Psycholinguistics* (1st ed.). Oxford: Oxford University Press. Accessed online (DOI: 10.1093/oxfordhb/9780198568971.013.0032).

Cowles, H. W., & Ferreira, V. S. (2011). The influence of topic status on written and spoken sentence production. *Discourse Processes*, *49*(1), 1–28.

Cutler, A. (1985). Cross-language psycholinguistics. *Linguistics*, *23*(5), 659–668.

Dale, R., & Duran, N. D. (2011). The cognitive dynamics of negated sentence verification. *Cognitive Science*, *35*(5), 983–996.

Dutton, E. M. (2012). *Animacy effects on sentence planning and production: Evidence from an experimental study in Tarifiyt Berber* (ResMA Thesis). Leiden University.

Ernanda. (2015). Phrasal alternation in the Pondok Tinggi dialect of Kerinci: An intergenerational analysis. *Wacana*, *16*(2), 355–382.

Ernanda. (2017). *Phrasal alternation in Kerinci* (PhD Thesis). Leiden University.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*(05), 429–448.

Fauconnier, S. (2012). *Constructional effects of involuntary and inanimate Agents: A cross-linguistic study* (PhD Thesis). KU Leuven.

Fernández-i-Marín, X. (2016). Ggmcmc: Analysis of MCMC samples and Bayesian inference. *Journal of Statistical Software*, *70*(9).

Ferreira, F. (1994). Choice of passive voice is affected by verb type and animacy. *Journal of Memory and Language*, *33*(6), 715–736.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, *47*(2), 164–203.

Ferreira, V. S. (1996). Is it better to give than to donate? Syntactic flexibility in language production. *Journal of Memory and Language*, *35*(5), 724–755.

Ferreira, V. S., & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, *40*(4), 296–340.

Ferreira, V. S., & Yoshita, H. (2003). Given-new ordering effects on the production of scrambled sentences in Japanese. *Journal of Psycholinguistic Research*, *32*(6), 669–692.

Freeman, J. B. (2014). Abrupt category shifts during real-time person perception. *Psychonomic Bulletin & Review*, *21*(1), 85–92.

Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, *42*(1), 226–241.

Freeman, J. B., & Dale, R. (2013). Assessing bimodality to detect the presence of a dual cognitive process. *Behavior Research Methods*, *45*(1), 83–97.

Galand, L. (1964). L'énoncé verbal en berbère: Étude de fonctions. In *Cahiers Ferdinand de Saussure* (Vol. 21, pp. 33–53). Geneva: Librairie Droz.

Ganushchak, L. Y., Konopka, A. E., & Chen, Y. (2014). What the eyes say about planning of focused referents during sentence formulation: A cross-linguistic investigation. *Language Sciences*, *5*, article 1124.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Gennari, S. P., Mirković, J., & MacDonald, M. C. (2012). Animacy and competition in relative clause production: A cross-linguistic investigation. *Cognitive Psychology*, *65*(2), 141–176.

Gil, D. (2001). Escaping Eurocentrism: Fieldwork as a process of unlearning. In P. Newman & M. S. Ratliff (Eds.), *Linguistic fieldwork* (pp. 102–132). Cambridge: Cambridge University Press.

Gleitman, L. R. (1990). The structural sources of verb meanings. *Language acquisition*, *1*(1), 3–55.

Gleitman, L. R., January, D., Nappa, R., & Trueswell, J. C. (2007). On the give and take between event apprehension and utterance formulation. *Journal of Memory and Language*, *57*(4), 544–569.

Griffin, Z. M., & Bock, J. K. (2000). What the eyes say about speaking. *Psychological Science*, *11*(4), 274–279.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, *33*(2), 1–22.

Haeseryn, W., Romijn, K., Geerts, G., de Rooij, J., & van den Toorn, M. C. (1997). *Algemene Nederlandse Spraakkunst* (2; version 1.3 ed.). Groningen: M. Nijhoff, Wolters Plantyn.

Halliday, M. A. K. (1967). Notes on transitivity and theme in English: Part 2. *Journal of Linguistics*, *3*(2), 199–244.

Hartigan, J. A., & Hartigan, P. M. (1985). The Dip test of unimodality. *The Annals of Statistics*, *13*(1), 70–84.

Hehman, E., Stolier, R. M., & Freeman, J. B. (2015). Advanced mouse-tracking analytic techniques for enhancing psychological science. *Group Processes & Intergroup Relations*, *18*(3), 384–401.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83.

Hwang, H., & Kaiser, E. (2015). Accessibility effects on production vary cross-linguistically: Evidence from English and Korean. *Journal of Memory and Language*, *84*, 190–204.

Ibrahim, R., & Aharon-Peretz, J. (2005). Is Literary Arabic a second language for native Arab speakers?: Evidence from semantic priming study. *Journal of Psycholinguistic Research*, *34*(1), 51–70.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

Jaeger, T. F., & Norcliffe, E. J. (2009). The cross-linguistic study of sentence production. *Language and Linguistics Compass*, *3*(4), 866–887.

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 824.

Keenan, E. L., & Comrie, B. (1977). Noun phrase accessibility and Universal Grammar. *Linguistic Inquiry*, *8*(1), 63–99.

Kelly, M. H., Bock, J. K., & Keil, F. C. (1986). Prototypicality in a linguistic context: Effects on sentence structure. *Journal of Memory and Language*, *25*(1), 59–74.

Kempen, G., & Harbusch, K. (2004). A corpus study into word order variation in German subordinate clauses: Animacy affects linearization independently of grammatical function assignment. *Trends in Linguistics Studies and Monographs*, *157*, 173–182.

Kempen, G., & Hoenkamp, E. (1987). An incremental procedural grammar for sentence formulation. *Cognitive Science*, *11*(2), 201–258.

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650.

Kiss, K. É. (1995). *Discourse Configurational Languages*. Oxford: Oxford University Press.

Knoeferle, P., Crocker, M. W., Scheepers, C., & Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. *Cognition*, *95*(1), 95–127.

Konopka, A. E. (2012). Planning ahead: How recent experience with structures and words changes the scope of linguistic planning. *Journal of Memory and Language*, *66*(1), 143–162.

Konopka, A. E., & Bock, J. K. (2009). Lexical or syntactic control of sentence formulation? Structural generalizations from idiom production. *Cognitive Psychology*, *58*(1), 68–101.

Konopka, A. E., & Brown-Schmidt, S. (2014). Message encoding. In M. Goldrick, V. S. Ferreira, & M. Miozzo (Eds.), *The Oxford Handbook of Language Production.* Oxford: Oxford University Press.

Konopka, A. E., & Meyer, A. S. (2014). Priming sentence planning. *Cognitive Psychology*, *73*, 1–40.

Kossmann, M. G. (2009). Loanwords in Tarifiyt, a Berber language of Morocco. In U. Tadmor & M. Haspelmath (Eds.), *Loanwords in the World's Languages : A Comparative Handbook* (pp. 191–214). Berlin, Germany: De Gruyter Mouton.

Kossmann, M. G. (2016). On word order in Figuig Berber narratives: The uses of pre- and postverbal lexical subjects. In M. Jursa, M. Köhbach, R. Lohlker, S. Procházka, & C. Berlinches Ramos (Eds.), *Wiener Zeitschrift für die Kunde des Morgenlandes* (Vol. 106). Vienna: Institut für Orientalistik.

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (Edition 2 ed.). Boston: Academic Press.

Kuno, S. (1973). *The structure of the Japanese language* (No. 3). Cambridge, Mass: MIT Press.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59.

Lafkioui, M. (2010). La topicalisation en berbère: formes et structures. In H. Stroomer, M. G. Kossmann, D. Ibriszimow, & R. Vossen (Eds.), *Études berbères V: essais sur des variations dialectales et autres articles: actes du 5. Bayreuth-Frankfurt-Leidener Kolloquium zur Berberologie* (Vol. 28, pp. 121–132). Köln: Köppe Verlag.

Lafkioui, M. (2011). Intonation et topicalisation en berbère. In A. Mettouchi (Ed.), *Parcours berbères. Mélanges offerts à Paulette Galand-Pernet & Lionel Galand pour leur 90ème anniversaire* (pp. 387–397). Köln: Köppe Verlag.

Lafkioui, M. (2014). Topicalization in Berber: A typological perspective. *STUF - Language Typology and Universals*, *67*(1), 97–112.

Lafkioui, M. (2017). Le rifain (tarifit): Linguistique et sociolinguistique. *Encyclopédie Berbère*, *41*, 6916–6956.

Lambrecht, K. (1996). *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents*. Cambridge: Cambridge University Press.

Leifeld, P. (2013). Texreg: Conversion of statistical model output in R to LaTeX and HTML tables. *Journal of Statistical Software*, *55*(8), 1–24.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, Mass: MIT press.

Li, C. N., & Thompson, S. (1976). Subject and topic: A new typology of language. In C. N. Li (Ed.), *Subject and topic* (pp. 457–489). New York: Academic Press.

Li, C. N., & Thompson, S. (1981). *Mandarin Chinese: A functional reference grammar*. Berkeley: University of California Press.

Lockwood, H. T., & Macaulay, M. (2012). Prominence hierarchies. *Language and Linguistics Compass*, *6*(7), 431–446.

Loop, J., Hamilton, A., & Burnett, C. (Eds.). (2017). *The teaching and learning of Arabic in Early Modern Europe*. Leiden: Brill.

Lüdecke, D. (2018). *sjPlot – data visualization for statistics in social science*. R package version 2.4.0: https://CRAN.R-project.org/package=sjPlot.

Lüdeling, A., Ritz, J., Stede, M., & Zeldes, A. (2016). Corpus linguistics and information structure research. In C. Féry & S. Ishihara (Eds.), *The Oxford handbook of information structure* (1st ed., pp. 598–617). Oxford: Oxford University Press.

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, *4*.

MacWhinney, B. (1987). The competition model. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 249–308).

MacWhinney, B. (2001). The competition model: The input, the context, and the brain. In P. Robinson (Ed.), *Cognition and Second Language Instruction* (pp. 69–90). Cambridge: Cambridge University Press.

MacWhinney, B., & Bates, E. (1978). Sentential devices for conveying givenness and newness: A cross-cultural developmental study. *Journal of Verbal Learning and Verbal Behavior*, *17*(5), 539–558.

Maechler, M. (2016). *Diptest: Hartigan's Dip test statistic for unimodality – corrected.* https://cran.r-project.org/web/packages/diptest/index.html.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190.

McDonald, J. L., Bock, J. K., & Kelly, M. H. (1993). Word and world order: Semantic, phonological, and metrical determinants of serial position. *Cognitive Psychology*, *25*(2), 188–230.

Mettouchi, A., & Fleisch, A. (2010). Topic-focus articulation in Taqbaylit and Tashelhit Berber. In I. Fiedler & A. Schwarz (Eds.), *The expression of information structure: A documentation of its diversity across Africa.* (pp. 193–232). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Mourigh, K., & Kossmann, M. G. (to appear). *An introduction to Tarifiyt Berber (Nador, Morocco)*. Münster: Ugarit Verlag.

Myachykov, A., Garrod, S., & Scheepers, C. (2012). Determinants of structural choice in visually situated sentence production. *Acta Psychologica*, *141*(3), 304–315.

Myachykov, A., Thompson, D., Scheepers, C., & Garrod, S. (2011). Visual attention and structural choice in sentence production across languages. *Language and Linguistics Compass*, *5*(2), 95–107.

Myachykov, A., & Tomlin, R. S. (2008). Perceptual priming and structural choice in Russian sentence production. *Journal of Cognitive Science*, *6*(1), 31–48.

Nettle, D., & Romaine, S. (2000). *Vanishing voices: The extinction of the world's languages*. Oxford; New York: Oxford University Press.

Norcliffe, E., Harris, A. C., & Jaeger, T. F. (2015). Cross-linguistic psycholinguistics and its critical role in theory development: Early beginnings and recent advances. *Language, Cognition and Neuroscience*, *30*(9), 1009–1032.

Norcliffe, E., Konopka, A. E., Brown, P., & Levinson, S. C. (2015). Word order affects the time course of sentence formulation in Tzeltal. *Language, Cognition and Neuroscience*, *30*(9), 1187–1208.

Onishi, K. H., Murphy, G. L., & Bock, J. K. (2008). Prototypicality in sentence production. *Cognitive Psychology*, *56*(2), 103–141.

Osgood, C. E. (1971). Where do sentences come from? In D. D. Steinberg (Ed.), *Semantics: An interdisciplinary reader in philosophy, linguistics and psychology* (Reprinted ed., pp. 497–529). Cambridge: Cambridge University Press.

Ouhalla, J. (1991). *Focussing in Berber and Circassian and the V2 Phenomenon* [Unpublished Manuscript]. University College London.

Oulad Saddik, A. (2013). *Tudunin war itizghen*. Zutphen: Wöhrmann.

Paczynski, M., & Kuperberg, G. R. (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.

Palfreyman, D., & al Khalil, M. (2003). "A funky language for teenzz to use": Representing Gulf Arabic in instant messaging. *Journal of Computer-Mediated Communication*, *9*(1).

Palmer, F. R. (1994). *Grammatical roles and relations*. Cambridge, UK; New York: Cambridge University Press.

Pickering, M. J., Branigan, H. P., & McLean, J. F. (2002). Constituent structure is formulated in one stage. *Journal of Memory and Language*, *46*(3), 586–605.

Pickering, M. J., & Ferreira, V. S. (2008). Structural priming: A critical review. *Psychological Bulletin*, *134*(3), 427–459.

Prat-Sala, M., & Branigan, H. P. (2000). Discourse constraints on syntactic processing in language production: A cross-linguistic study in English and Spanish. *Journal of Memory and Language*, *42*(2), 168–182.

Protopapas, A. (2007). Check Vocal: A program to facilitate checking the accuracy and response time of vocal responses from DMDX. *Behavior Research Methods*, *39*(4), 859–862.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Sauppe, S. (2017). *The role of voice and word order in incremental sentence processing. Studies on sentence production and comprehension in Tagalog and German* (PhD Thesis). Max Planck Institute for Psycholinguistics, Nijmegen.

Sauppe, S., Norcliffe, E., Konopka, A. E., & Levinson, S. C. (2013). Dependencies first: Eye tracking evidence from sentence production in Tagalog. In *CogSci 2013: The 35th annual meeting of the Cognitive Science Society* (pp. 1265–1270). Cognitive Science Society.

Siewierska, A. (1984). *The Passive: A comparative linguistic analysis*. London; Canberra: Croom Helm.

Siewierska, A. (2013). Passive constructions. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Silverstein, M. (1976). Hierarchy of features and ergativity. In R. M. W. Dixon (Ed.), *Grammatical categories in Australian languages* (pp. 112–171). Canberra: Australian National University.

Simons, G. F., & Fennig, C. D. (Eds.). (2017). *Ethnologue: Languages of the world* (20th ed.). Dallas, Texas: SIL International.

Slevc, L. R. (2011). Saying what's on your mind: Working memory effects on sentence production.

Journal of Experimental Psychology: Learning, Memory, and Cognition, *37*(6), 1503–1514.

Smith, M., & Wheeldon, L. (1999). High level processing scope in spoken sentence production. *Cognition*, *73*(3), 205–246.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 174.

Sorensen, T., Hohenstein, S., & Vasishth, S. (2016). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. *The Quantitative Methods for Psychology*, *12*(3), 175–200.

Speed, L. J., Wnuk, E., & Majid, A. (2017). Studying psycholinguistics out of the lab. In A. M. B. de Groot & P. Hagoort (Eds.), *Research Methods in Psycholinguistics and the Neurobiology of Language: A Practical Guide* (pp. 190–207). John Wiley & Sons.

Spivey, M. J., Dale, R., Knoblich, G., & Grosjean, M. (2010). Do curved reaching movements emerge from competing perceptions? A reply to van der Wel et al. (2009). *Journal of Experimental Psychology: Human Perception and Performance*, *36*(1), 251–254.

Sridhar, S. N. (1988). *Cognition and sentence production: A cross-linguistic study*. Springer Science & Business Media.

Stallings, L. M., MacDonald, M. C., & O'Seaghdha, P. G. (1998). Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-NP shift. *Journal of Memory and Language*, *39*(3), 392–417.

Stan Development Team. (2017a). *Shinystan: Interactive Visual and Numerical Diagnostics and Posterior Analysis for Bayesian Models.* R package version 2.4.0: http://mc-stan.org.

Stan Development Team. (2017b). *Stan Modeling Language Users Guide and Reference Manual, version 2.17.0.* http://mc-stan.org.

Steinhauer, H., & Usman, A. H. (1978). Notes on the morphemics of Kerinci (Sumatra). In S. A. Wurm & L. Carrington (Eds.), *Second international conference on Austronesian Linguistics: Proceedings, Fascicle 1, Western Austronesian (Pacific Linguistics, C-61)* (pp. 483–502). Canberra: Research School of Pacific Studies, The Australian National University.

Tanaka, M. N., Branigan, H. P., McLean, J. F., & Pickering, M. J. (2011). Conceptual influences on word order and voice in sentence production: Evidence from Japanese. *Journal of Memory and Language*, *65*(3), 318–330.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Tomlin, R. S. (1995). Focal attention, voice, and word order: An experimental, cross-linguistic study. In P. Downing & M. Noonan (Eds.), *Word order in discourse* (pp. 517–554). Amsterdam; Philadelphia: J. Benjamins.

Toyota, J. (2011). *The grammatical voice in Japanese: A typological perspective*. Newcastle-upon-Tyne: Cambridge Scholars Publisher.

Tsegaye, M. T. (2017). *Plural gender: Behavioral evidence for plural as a value of Cushitic gender with reference to Konso* (PhD Thesis). Leiden University.

van der Wel, R. P. R. D., Eder, J. R., Mitchel, A. D., Walsh, M. M., & Rosenbaum, D. A. (2009). Trajectories emerging from discrete versus continuous processing models in phonological competitor tasks: A commentary on Spivey, Grosjean, and Knoblich (2005). *Journal of Experimental Psychology: Human Perception and Performance*, *35*(2), 588–594.

van Gompel, R. P. G., & Pickering, M. J. (2007). Syntactic parsing. In M. G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics.*

van Nice, K. Y., & Dietrich, R. (2003). Task sensitivity of animacy effects: Evidence from German picture descriptions. *Linguistics*, *41*(5), 825–849.

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190.

Verhoeven, E. (2014). Thematic prominence and animacy asymmetries. Evidence from a cross-linguistic production study. *Lingua*, *143*, 129–161.

Vogels, J., Krahmer, E., & Maes, A. (2013). When a stone tries to climb up a slope: The interplay between lexical and perceptual animacy in referential choices. *Frontiers in Psychology*, *4*, article 154.

Wagner, M. (2016). Information structure and production planning. In C. Féry & S. Ishihara (Eds.), *The Oxford Handbook of Information Structure* (1st ed., pp. 541–561). Oxford: Oxford University Press.

Wagner, V., Jescheniak, J. D., & Schriefers, H. (2010). On the flexibility of grammatical advance planning during sentence production: Effects of cognitive load on multiple lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 423–440.

Weyerts, H., Penke, M., Münte, T. F., Heinze, H.-J., & Clahsen, H. (2002). Word order in sentence processing: An experimental study of verb placement in German. *Journal of Psycholinguistic Research*, *31*(3), 211–268.

Whalen, D. H., & McDonough, J. (2015). Taking the laboratory into the field. *Annual Review of Linguistics*, *1*(1), 395–415.

Wickham, H. (2009). *Ggplot2: Elegant graphics for data analysis*. New York: Springer.

Samenvatting in het Nederlands

Deze dissertatie houdt zich bezig met de processen die sprekers gebruiken om hun ideeën te organiseren om zich te kunnnen uiten in passende, welgevormde zinnen. Bijzondere aandacht wordt er geschonken aan de manier waarop dit verloopt, met bijzondere aandacht voor het feit dat talen op talloze manieren van elkaar verschillen. De vragen die ik heb geprobeerd te beantwoorden zijn vragen die die voortvloeien uit onze steeds toenemende interesse in onderzoek naar typologisch verschillende talen. Deze vragen zijn zowel theoretisch als methodologisch van aard.

De theoretische vraag die dit proefschrift onderzoekt is hoe sprekers en luisteraars met hun taalspecifieke hulpmiddelen navigeren om conceptuele inhoud te communiceren. Om deze kwestie aan te pakken, is het belangrijk om te begripen hoe verschillende talen, die in formele (typologische) parameters verschillen, kunnen dienen om dezelfde communicatieve doelen van sprekers te bereiken. De huidige inzichten op dit gebied worden beperkt door het feit dat de psycholinguïstiek slechts een heel beperkt deel van de talen van de wereld in enige diepte heeft onderzocht. Bovendien blijkt dat bij het werken met voor dit onderzoeksgeboed nieuwe types talen en sprekersgemeenschappen, uitdagingen ontstaan bij het toepassen van bestaande methoden en paradigma's. Dit brengt ons daarom ook bij methodologische vragen.

Hoofdstuk 1 geeft de theoretische achtergrond voor de studie van grammaticaal coderen. Dit onderzoeksgebied houdt zich bezig met hoe de pre-linguïstische boodschap wordt gecodeerd in grammaticale structuren. De rol van grammaticaal coderen is om niet alleen stukjes informatie, maar ook hun relaties in een talige vorm te gieten. Wat deze talige vorm betreft, houden we ons hier primair bezig met hoe grammaticale functies worden toegewezen, en wat de relatieve volgorde van de zinselementen is. Cross-linguïsisch gezien is het duidelijk dat de toewijzing van grammaticale functies en de manier waarop elementen in de zin worden geordend in talen op verschillende manieren worden gerealiseerd. We gaan er echter over het algemeen van uit dat het niet-talige aspect van zinsproductie bij alle sprekers van de wereld hetzelfde is. Dit maakt duidelijk dat we cross-linguïstische variatie zullen moeten bestuderen om volledig te begrijpen hoe dit systeem werkt.

Hoofdstuk 2 onderzoekt de kwestie taalvariatie in de psycholinguïstiek, met een focus op de praktische en methodologische vragen die zich voordoen. De typologische dekking van het veld is scheef en wordt gedomineerd door een klein aantal West-Europese, met name Germaanse, talen. Deze situatie duurt nog steeds voort, ondanks de roep om meer taalverscheidenheid in de psycholinguïstiek, van vooraanstaande wetenschappers in het veld – al sinds tientallen jaren. In dit hoofdstuk stel ik voor dat de hardnekkigheid van deze stand van zaken niet te wijten is aan een gebrek aan bewustzijn, maar aan onopgeloste uitdagingen die onderzoekers ervan weerhouden om cross-linguïstisch onderzoek te omarmen. Dit hoofdstuk bespreekt een aantal van dergelijke uitdagingen, met name in het kader van studies naar de variatie in zinsvormen. Sommige uitdagingen hebben betrekking op logistieke of praktische moeilijkheden, maar veel meer nog zien we moeilijkheden bij het toepassen van bestaande methodologie of het integreren van resultaten in de huidige theorie. Deze tweede reeks problemen is ook nauw verbonden met de zogenaamde bias die onderzoekers kunnen hebben en die gemakkelijk optreedt bij het toepassen van bestaande benaderingen op onbekende taaltypen.

Om een grotere taalverscheidenheid in het veld te bereiken is het daarom nodig een open discussie te voeren en tot een evaluatie te komen van methodologische keuzes bij het werken met typologisch uiteenlopende talen. De studies die worden gepresenteerd in hoofdstukken 3-6 hebben tot doel hieraan bij te dragen, door het belichten en onderzoeken van de keuzes die we moeten maken bij het voorbewerken, coderen en analyseren van gegevens. De studies in hoofdstukken 3-6 betreffen het Nederlands, het Tarifiyt Berber en het Pondok Tinggi: drie talen die een breed scala aan contrasten bieden over relevante dimensies, zowel theoretisch als praktisch. Het einde van hoofdstuk 2 geeft een overzicht van deze talen en hun relevantie voor de onderzoeksvragen van dit proefschrift.

Hoofdstuk 3 herbeziet een bestaande dataset van een 'simply describing' (plaatjes beschrijven) studie in het Tarifiyt Berber en het Nederlands, en presenteert een nieuwe analyse en een nieuw cross-linguïstisch perspectief op de data. Deze studie onderzoekt het effect van animacy (bezieldheid; ook wel levendheid) van de *patiens* (ten opzichte van de *agens*) op de taalkundige vorm van beschrijvingen van afbeeldingen van transitieve gebeurtenissen. Zowel in het Tarifiyt als in het Nederlands laten de resultaten zien dat proefpersonen vooral kiezen voor subject-initiële actieve constructies. Er is echter een verhoogde kans op patiens-prominente structuren in beide talen wanneer de patiens meer bezield (meer levend of 'animate') is dan de agens. Een cross-linguïstische vergelijking geeft het potentieel aan voor een verschil in grootte van dit effect, maar dit verschil kon niet statistisch worden onderbouwd.

De discussie in dit hoofdstuk betreft de complementaire typologische profielen van de twee talen, in termen van de taalkundige vormen die worden gebruikt om de prominentie van de patiens te coderen. Hoewel deze twee talen taalverschijnselen vertonen die goed bekend zijn in de studie van zinsproductie (passivering en object-topicalisatie), is het opvallend dat tussen de twee talen dezelfde communicatieve omstandigheden lijken te leiden tot complementaire resultaten in termen van de taalkundige zinsvorm. Het klassieke model van zinsproductie (Bock & Levelt, 1994) maakt niet expliciet duidelijk hoe cross-linguïstische verschillen zoals deze ontstaan. Enkele suggesties van eerdere auteurs worden hier geïntroduceerd en geëvalueerd. De verklaring die het best gebaseerd is op cross-linguïstische gronden is er een die gebruikmaakt van 'subject and topic prominence' (subject-prominentie en topic-prominentie; Li & Thompson, 1976), een algemeen erkende dimensie van typologische variatie die desalniettemin weinig aandacht heeft gekregen in zinproductiestudies. Toekomstige vergelijkende experimentele studies zullen ongetwijfeld waardevol blijken om verder te verduidelijken hoe we dit kunnen integreren in het klassieke model van zinsproductie.

Hoofdstuk 3 vestigt ook de aandacht op de methodologische keuzes die een rol spelen bij het voorbewerken en analyseren van de ruwe zinproductiedata van 'simply describing' experimenten. In het bijzonder vereisen de voorbereidende en classificerende stadia bij dergelijke experimenten dat we generaliseren of abstraheren om de gegevens in een analyseerbare vorm te brengen. Dit geldt tot op zekere hoogte voor alle datacoderingsprocedures, maar bij 'simply describing' valt op dat de onbewerkte gegevens extreem heterogeen en uitbundig kunnen zijn (i.e. exuberant responding, in de zin van Bock, 1996), terwijl ze toch meestal worden gereduceerd tot een (set van) binaire variabe(len). Het is duidelijk dat de onderzoeker hier ruimte kan nemen om de verwerking van gegevens vorm te geven. Dit hoofdstuk sluit af met een aangescherpte reeks vragen die in hoofdstuk 4-6 worden behandeld.

Hoofdstuk 4 onderzoekt het samenspel van talige vorm en cognitieve principes vanuit het perspectief van de luisteraar. Gebruikmakend van *mouse-tracking* om begripprocessen te onderzoeken, onderzoek ik de manier waarop luisteraars gevoelig zijn voor lineaire volgorde en bezieldheid (*animacy*) bij het interpreteren van thematische structuur. Grammaticale functie-toewijzing wordt hierbij meegenomen door deze studie in het Tarifiyt Berber en het Nederlands uit te voeren, talen met verschillende typologische profielen vertonen bij grammaticale functie-toewijzing. Deze studie is opgezet als een tegenhanger van de productiestudie van hoofdstuk 2, wat een vergelijking tussen begrip en productie mogelijk maakt.

Een belangrijke vraag van dit onderzoek is of luisteraars eenvoudigweg prominentie-kenmerken (vroege zinspositie, zeer animate) combineren om de agens te identificeren, of dat hun voorspellingen over thematische structuur inhouden dat beschikbare informatie wordt afgewogen tegen waarschijnlijkheden in de input. De bevindingen ondersteunden het laatste. Deze bevindingen komen overeen met theorieën van van zinsbegrip, waarbij de *parser* (ontleder) vroege toegang heeft tot een reeks beschikbare informatie, en waarbij deze bronnen samenwerken met kennis van de taalkundige input, als onderdeel van een probabilistisch proces van thematische interpretatie. Wat de cross-linguïstische, cross-modaliteitsvergelijkingen betreft, boden de bevindingen verder bewijs en verdere nuance voor de voorspellende verklaring: de antwoorden die de luisteraars tijdens de test gaven lijken de cross-linguïstische tendensen en verschillen die we in de productiegegevens zagen te weerspiegelen.

Hoofdstuk 4 draagt ook bij aan methodologische doelen om online zinsverwerking in niet-laboratoriumomgevingen te onderzoeken. Deze studie werd uitgevoerd in sprekersgemeenschappen in Marokko en Nederland, waarbij de interactie met de deelnemers in beide gevallen volledig in de doeltaal werd uitgevoerd. De aanpak van werving, het vragen van een weloverwogen toestemming en het geven van compensatie was zorgvuldig doordacht met aandacht voor de culturele setting. Aangezien een van de twee bestudeerde talen een overwegend orale traditie kent (Tarifiyt), werd de schriftelijke modaliteit voor zover praktisch mogelijk uit de studie verwijderd. De gebruikte gegevensverzamelingstechniek (mouse-tracking) is een voorbeeld van een minimale uitrustingopstelling en -uitgaven met een rijke gegevensopbrengst.

Hoofdstuk 5 presenteert een zinproductiestudie met hetzelfde ontwerp als in hoofdstuk 3, waarin de effecten van animacy (bezieldheid) op de zinsvorm in het Pondok Tinggi worden onderzocht. In deze taal bestaat er meer dan één structuur om taalkundige prominentie voor het patiens-argument uit te drukken. Hoewel deze keuzes verschillende structurele categorieën vertegenwoordigen, geven ze allemaal prominentie aan de patiens. Huidige theorieën over structurele keuze verklaren patiens-prominentie alleen in termen van snelheid van het vinden van informatie (*accessibility*), en doen geen specifieke voorspellingen voor specifieke soorten patiens-prominente structuren. Dit leidt ertoe dat we voorspellen dat alle categorieën op dezelfde manier worden beïnvloed door bezieldheid. De resultaten hier geven echter aan dat niet alle patiens-prominente categorieën evenzeer worden beïnvloed door animacy (bezieldheid). In de discussie in dit hoofdstuk opper ik de mogelijkheid dat dergelijke variaties in zinsvorm andere variabelen weerspiegelen die niet herleidbaar zijn tot de snelheid van het vinden van informatie, zoals de *affectedness* van de patiens en de *volition* van de agens. Hoe deze begrippen in de theorie kunnen worden geïntegreerd, is een onderwerp voor toekomstige experimentele studies.

Hoofdstuk 5 werpt een belangrijk methodologisch punt op met betrekking tot de binaire benadering van analyse van zinsproductiedata, zoals standaard wordt gebruikt bij de analyse van *simply describing* studies. Binaire analysetechnieken worden doorgaans gebruikt om inzicht te krijgen in hoe sprekers structurele vormen selecteren, zelfs in talen waar de keuze voor sprekers niet echt tweevoudig is. Om dit te doen, moeten we de gegevens reduceren tot (een) dichotome variabele(n) voor men overgaat tot analyse. Deze praktijk kan voortkomen uit een wens voor continuïteit met eerdere studies, maar heeft ongetwijfeld ook te maken met de complexiteit van de analyse die nodig is om meerdere uitkomstcategorieën in ogenschouw te nemen. Hoewel binaire analyse in bepaalde situaties geschikt is, zijn er een aantal beperkingen waarmee rekening moet worden gehouden voor een cross-linguïstische studie van zinsproductie. In algehele zin betoog ik dat binaire analyse geen neutrale standaard is, maar juist de verschijnselen die we willen bestuderen misschien wel verduistert. De huidige studie toont aan dat door een multicategorische benadering van zinskeuzedata in plaats van een binaire het mogelijk is om verder te komen dan de vraag of het ene type structuur waarschijnlijker is dan het andere, om te modelleren hoe de keuze tussen meerdere structuren varieert als functie van de experimentele manipulatie.

Hoofdstuk 6 presenteert nog een *simply describing* studie in het Tarifiyt Berber, maar dan benaderd vanuit een ander methodologisch perspectief. Het hoofdstuk is een poging om het hoofd te bieden aan de moeilijkheden van het werken met *fluent* (vloeiende) productiedata, vooral wanneer deelnemers uitbundig reageren, zoals bias en gegevensverlies. Ik betoog dat deze kwesties steeds meer naar voren zullen komen als we met meer verschillende talen en sprekersgemeenschappen gaan werken. Daarom onderzoekt deze studie de variabelen die bekend zijn uit eerdere hoofdstukken, maar zorgt het juist voor responsuitbundigheid, in plaats van deze in proberen te perken. Deze methode omzeilt de behoefte aan antwoorden die voldoen aan een bepaald *template*. Daarmee verminderen de mogelijkheden voor bias van de onderzoeker om vorm te geven aan wat de studie ons laat zien. Het algemene doel is om de effecten van *foregrounding* op de taalkundige vorm te onderzoeken, in het bijzonder of de bevindingen in overeenstemming zijn met de verwachtingen die voortkomen uit het idee dat het zinsformulering gebeurt op basis van *accessibility* (het vinden van informatie). Bovendien opent deze alternatieve benadering van codering en analyse ook de weg om theoretische vragen te stellen die in het type onderzoek als in hoofdstuk 3 is gepresenteerd niet mogelijk zijn.

Over het algemeen komen de bevindingen overeen met de zogenaamde *Accessibility* account; bij nader inzien blijkt het echter opnieuw moeilijk om de lineaire ordeningen te verklaren in termen van louter snelheid van informatie-vinden. In het bijzonder lijkt de preverbale positie geassocieerd te kunnen worden met een specifieke discoursstatus, en is het niet alleen maar een neveneffect van verwerkingseffecten. Hoewel deze studie bedoeld is om problemen van bias te omzeilen, is het ook duidelijk dat we nooit alle bias kunnen uitsluiten, omdat iedere vorm van data-analyse codering en dus een zekere mate van abstractie vereist. Dit laat eens te meer zien hoe belangrijk typologische (taalkundige) training en ondersteuning voor psycholinguïsten om problemen van vertekening en scheefheid te overwinnen (zoals aan de orde gesteld in hoofdstuk 2).

Samenvattend bevestigden deze studies eerdere bevindingen met betrekking tot de impact van variabelen zoals *animacy* (bezieldheid) en foregrounding op de taalkundige vorm van zinnen, inclusief grammaticale functietoewijzing en lineaire volgorde. Wat de theoretische verklaringen voor deze fenomenen betreft, roepen de studies echter nieuwe vragen op. Ten eerste is de Accessibility account van zinsproductie, waarbij het vinden van informatie de essentiële drijfveer is voor variatie van de zinsvorm, niet altijd voldoende om de bevindingen te verklaren. Uiteindelijk is de beperking van de theorie van Accessibility dat het lijkt te vereisen dat we de enorme complexiteit van cross-linguïstische variatie met behulp van een enkele latente variabele verklaren. Een mogelijkheid om dit probleem te verlichten is door te stellen dat er verschillende ‘sources’ zijn voor Accessibility. Dit roept dan echter de vraag op of dit nog steeds facetten zijn van één enkel begrip van Accessibility.

Ten tweede laten de bevindingen over het Tarifiyt Berber overtuigend zien dat de status van *topic* in de psycholinguïstische theorie van zinsproductie opnieuw moet worden geëvalueerd. Hoewel het concept van *topic* notoir lastig is, moeten we toch de status van topicalisatie in cognitieve termen opnieuw bekijken. Dit wordt bevestigd door het feit dat een aantal andere talen een formele codering van het topic lijken te tonen, analoog aan de formele codering van het subject in talen die beter onderzocht zijn in het gebied van zinsproductie. De typologische generalisaties van subject-prominentie en topic-prominentie kunnen een kader bieden waarbinnen cross-linguïstische hypotheses kunnen worden gegenereerd om dit onderzoeksgebied verder te ontwikkelen.

Wat de methodologie betreft, hadden de studies in dit proefschrift tot doel om de uitdagingen die samenhangen met het werken met typologisch diverse, minder bestudeerde talen te identificeren en aan te pakken. Na verschillende praktische problemen aan het licht te hebben gebracht die niet breed uitgemeten zijn in de literatuur, wilde ik alternatieve benaderingen verkennen die de problemen die voortkomen uit restrictieve analyse of bias bij onderzoekers kunnen verlichten. Het is van groot belang om onze standaardtechnieken te heroverwegen wanneer zij de taalkundige variatie verontachtzamen. In dit proefschrift worden alternatieve benaderingen toegepast op typologisch diverse talen, en dit biedt meer genuanceerde inzichten in grammaticale codering dan mogelijk zou zijn geweest met standaardbenaderingen. Een belangrijke conclusie van dit proefschrift is daarom dat het niet alleen belangrijk is om een breder scala aan talen te selecteren om te bestuderen, maar ook om voortdurend de geschiktheid van geaccepteerde methoden en technieken voor een taalkundig diverse steekproef te beoordelen.

Curriculum Vitae

Eleanor Dutton was born in Kent, United Kingdom. She gained a first-class BA (hons) degree in Linguistics from University College London in 2007. After graduating, she worked as a project manager, translator, editor and broadcast subtitler in London. In 2010 she moved to Leiden, Netherlands to follow the Research Master's programme in Linguistics at the Leiden University Centre for Linguistics (LUCL). On receiving her Research Master's degree in 2012 (cum laude) she was awarded a position at the Leiden University Centre for Linguistics to carry out her proposal to investigate sentence production in Tarifiyt Berber, which she had begun researching with her Master's thesis. This line of research evolved to reflect more generally on theoretical and methodological challenges for the inclusion of understudied languages in the psycholinguistic study of sentence form variation. Alongside her PhD, she co-mananged the LUCL Experimental Linguistics labs from 2015-2018. This included the co-development of a teaching programme for experimental methods for students and researchers at LUCL.

This thesis is concerned with the processes by which speakers organise their ideas into felicitous, well-formed sentences, with particular focus on how this proceeds given that languages vary in myriad ways. The questions that this thesis aims to address are those that are raised as we aim to conduct more cross-linguistic work on typologically distinct languages – questions that are both theoretical and methodological in nature.

Universiteit
Leiden