



Universiteit
Leiden
The Netherlands

Language diversity in the psycholinguistic study of sentence form variation

Dutton, E.M.

Citation

Dutton, E. M. (2019, December 12). *Language diversity in the psycholinguistic study of sentence form variation*. Retrieved from <https://hdl.handle.net/1887/81786>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/81786>

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/81786> holds various files of this Leiden University dissertation.

Author: Dutton, E.M.

Title: Language diversity in the psycholinguistic study of sentence form variation

Issue Date: 2019-12-12

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 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 experi-

mental 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 Tarifit 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 Tarifit 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 Tarifit

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 intuitions 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 *worden*). 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

Tarifyt 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 Tarifyt 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 Tarifyt (cf. Kossmann, 2009).

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).

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 sentence¹ 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).

- 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 Tarifyt 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 non-verbal 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 ufedjah*:

- (8) a. *ijj n weyyuř yarceř ij j ufedjah*
 one of AS.donkey 3SG.M.kick.PERF one AS.farmer
 ‘a donkey kicks a farmer’
 b. *ijj ufedjah yarceř-it ij j n weyyuř*
 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 ufedjah 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 *nettāt* ‘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 Tarifit. 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.