Multi-word-constructions and linguistic development in early foreign language classrooms: the role of input variability

Johannes Maximilian Schulz
M.Sc. Applied Linguistics and Second Language Acquisition, University of Oxford
B.Ed. German and English Language Studies, University of Tübingen

Thesis submitted in fulfilment of the requirements for the Degree of

Doctor of Philosophy in Education (Applied Linguistics)

University of Oxford
St Hugh's College
Supervisors: Professor Elizabeth Wonnacott and Professor Victoria Murphy

June 2024

An allem Unfug, der passiert, sind nicht etwa nur die schuld,
die ihn tun, sondern auch die, die ihn nicht verhindern.

Erich Kästner – Das fliegende Klassenzimmer (1933)

If there is mischief, not only are those doing the mischief to
blame, so are those who do nothing to stop it.

Erich Kästner – The flying classroom (1933)
translated by Anthea Bell (2014)

Acknowledgments

I'm under no illusion; I know that the acknowledgments and the abstract are perhaps the most widely read parts of most PhD theses. And I am not blaming the readers. After all, the acknowledgments are where the saucy personal details are, right?! And the rest? Well, there are thousands of more fun things to do than reading a PhD thesis. So, enjoy the acknowledgments and the abstract, and perhaps you scroll through the colourful plots or STRG/CMD+F something (when you know, you know!). In any case, I am glad you are here! And feel free to keep reading. It was a lot of work, after all.

I understand and appreciate that the PhD journey can often be stressful, and some candidates feel relieved when it's 'all over'. Fortunately, I enjoyed my work and had a positive relationship with my supervisors. And although I am obviously an extremely pleasant German to work with, most of the credit for this positive experience goes to my supervisors, Vicki and Liz, who did a brilliant job. In all earnestness, I would like to thank them both wholeheartedly. They always treated me and my work with respect and believed in my abilities. Admittedly, they may have doubted my work here and there (that's their job!), but they never gave me the feeling that they doubted *me*. Vicki paved the financial way for this DPhil and placed a lot of trust in me. Throughout the years, I could always rely on her to have my back. In general, she constantly gave me the reassuring feeling of: "Trust me, I know exactly what we are doing" – and oh boy, she does! This was extremely valuable. And Liz, what a researcher! She reliably spots mistakes that I didn't even know one could make, incredible! Her striking attention to detail, laudable commitment to integrity in science, and extensive knowledge and intelligence have always impressed me and made my work so much better. I learned a lot. Thank you both!

A big 'Thank you!' goes to the school I worked with during my research, specifically to the teacher who facilitated the collaboration. Unfortunately, I am not allowed to say their name, but they are an incredibly committed teacher, and they are one of the most generous and welcoming persons I have ever met. And of course, importantly, 'Thank you!' to the students who were patient, curious, and overall did a wonderful job. Without your input, well, this thesis would be empty.

So, I also did stuff outside of work. (Here, I said it!) Usually, being the predictable German I am, I spent my time playing football with the lads (LADS! LADS! LADS!). Let's say the quality wasn't much, but not for a lack of trying! And, yes, objectively speaking, our losses were generally the referee's fault, what can I say. I had a great time on the pitch and in the pub, thanks! Off the pitch (often injured in some way or another), I also enjoyed spending time with my friends over coffee, at lunches and dinners, at festivals, at pubs, at parties, on punts, on trips, wherever really. No matter whether you are old friends from school, friends from my undergrad in my beloved alma mater tubingensis (*Vivat, crescat, floreat, rex*!), or friends I met during my PhD from all around the world. I want to emphasize how deeply I appreciate each and every one of you. In addition to my friends, I would like to thank my family for their unwavering interest and support in what I do, and for their love. They are very important to me. Thank you!

Whoever knows me, knows that one person is still missing. I realize that being able to say this is the greatest gift. She sat in the same classroom with me in high school, and now she is still sitting here next to me. It's wonderful to have you in my life every day. Thank you, Anki.

Oh, one last thing. Although this might sound idealistic and corny, I would also like to tell my future self for when I am reading this in a couple of decades: You used to be smarter and faster, now you're (hopefully) wiser and (certainly) slower. In any case, you are doing great. Keep questioning, keep disagreeing, keep listening, keep asking, keep caring. And don't take yourself too seriously!

meinem Bruder

Table of Contents

Abstract

Upon finishing FL learning at primary school, many students lack productive knowledge, for example regarding structures like verb-argument-constructions (e.g., [Verb] about [Noun]), indispensable for increasing communicative agency, as expected by curricula. These curricula adopt a usage-based constructionist approach to language learning and consider 'large linguistic units' (e.g., routines (*How are you?*) or patterns (*My favourite _ is _*)) to be catalysts for students' development of a productive linguistic repertoire. However, despite the ubiquity of such units in primary curricula, evidence on the impact of teaching input consisting of such structures on primary FL students' linguistic development is scarce, as indicated by Schulz et al. (2023), reported in this thesis.

Across cognitive domains, increased initial input variability (the variation in our experience with different exemplars, e.g., [talk/think/rant/wonder] about [god/bicycles/dogs]) can improve generalization (i.e., [Verb] about [Noun]), and enhance learning. In controlled experiments, increased input variability proved beneficial for children's inductive generalization and extension of linguistic information from input structures to novel contexts (e.g., Wonnacott et al., 2012). Such findings drove the investigation into extending the benefits of input variability to real classrooms.

Following a usage-based constructionist approach to language learning, I report on two quasi-experimental teaching intervention studies (each lasting two weeks) with two British Year 2 classes learning German (age 6; 20 students/class). Experiment 1, comprising of a high (HV) and low (LV) input variability condition, focused on 16 German 'approach' event verb-argument-constructions (*Zum X [robbt/schleicht/rutscht/etc.] der/die/das Y*; *To the X [approach verb] the Y*), featuring one (LV) or four different verbs (HV) in the construction's verb slot. Post-tests indicated that children exposed to increased input variability demonstrated better generalization to novel verbs compared to controls. Experiment 2 focused on three sets of non-adjacent dependencies (cf. Gómez, 2002). The HV and LV conditions included 30 and five 'intervener positions', respectively. The learning of non-adjacent dependencies and the ability to generalise structural information and extend it to novel contexts (i.e., unknown interveners) was investigated in post-tests (grammaticality judgments), yielding ambiguous results.

The experiment 1 findings largely aligned with more 'controlled' experiments, suggesting that even in 'noisy' classroom environments, increased input variability can positively impact students' construction generalization and extension to novel verbs. However, the experiment 2 data did not support this variability effect. Instead, they underscored the potential benefits of LV input under specific pedagogical circumstances.

List of Tables

List of Figures

List of Abbreviations

ANCOVA	Analysis of Covariance
BF	Bayes Factor
EBSCO	Elton B. Stephens Company [database]
EFL	English as a Foreign Language
ERIC	Education Resources Information Centre
ESL	English as a Second Language
FL	Foreign Language
GIF	Graphics Interchange Format
GLMM	Generalized Linear Mixed Model
HV	High Variability
IDESR	International Database of Education Systematic Reviews
L1	Native Language
L2	Second Language
LLAB	Linguistics and Language Behaviour Abstracts
LV	Low Variability
MFL	Modern Foreign Language
MMAT	Mixed Methods Appraisal Tool
MWC	multi-word-construction
MWU	multi-word-unit
NAD	non-adjacent dependency
PICO	Participants, Intervention, Comparison, Outcomes
PIRLS	Progress in International Reading Literacy Studies
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PSYNDEX	Psychological Index
PVST	Picture Vocabulary Scale Test
RCT	randomised controlled trial
RR	Robustness Region
SLA	Second Language Acquisition
SR	Systematic Review
SUDOC	Système Universitaire de Documentation
TEFL	Teaching English as a Foreign Language
TESL	Teaching English as a Second Language
TESOL	Teaching English to Speakers of Other Languages
VAC	verb-argument-construction
WASI	Wechsler Abbreviated Scale of Intelligence
WoS	Web of Science

1 Literature Review

1.1 Introduction[1]

Since the 1980s, English as a Foreign Language (EFL) at primary level has been introduced worldwide (Enever, 2011; Graddol, 2006; Rixon, 2013). Europe joined this movement in the early 2000s (Council of Europe, 2001), and the UK made Foreign Language (FL) classes mandatory from Key Stage 2 (7-to-11-year-olds) in 2014 (Holmes & Myles, 2019). In primary school contexts, current FL curricula focus on communicative interactions and metalinguistic competences (KM-BW, 2004: 68–71), aiming to enable children to use the FL productively and to educate active language users (Bredenbroecker, 2018; Diehr, 2009; KM-BW, 2004; Sambanis, 2007). In addition to establishing communicative agency as a core aim (i.e., the ability to select and arrange language in a way that corresponds to the learner's intended meaning; Doyé & Lüttge, 1977), curricula have introduced the implementation of *multi-word-constructions* (MWCs) into early FL teaching (Council of Europe, 2018; Europarat, 2001), realized by phrasal structures like *chunks* and *formulaic sequences* (KM-BW, 2016) (see more detail on MWCs in section 1.2). The reason for this implementation is mainly that routinized MWCs, ubiquitous in many primary school classrooms worldwide (e.g., *How are you?*), equip learners with immediate linguistic agency to realise communicative goals. In turn, immediate communicative agency increases (or at least maintains) primary aged children's generally high language-learning motivation (Hempel, 2016; Lenzing & Roos, 2012). Curricula acknowledge that at this early stage, communicative agency mainly consists of imitated routinized MWCs. In contrast, curricula consider variable MWCs with fillable slots to provide learners with generative patterns that allow rule extraction and generalization, ideally leading to independent production (KM-BW, 2004). Despite this, FL communication skills by the end of primary school are often considered rudimental (Hempel, 2016) and many students, especially weaker ones, are reported to be stuck on a reproductive and *imitative plateau* (Engel et al., 2009; cf. Pienemann et al., 2006). Against this curricular background, this thesis investigates MWCs, linguistic development, and input variability in the context of instructed primary school FL learning.

[1] Note that parts of this chapter are adopted from Schulz et al. (2023). In that paper, the first author conceived and designed the analysis, collected the data, performed the analysis, and wrote the paper independently. The second author contributed to the double-blind screening and quality appraisal, as is good practice in systematic reviews, and reviewed and edited the draft. The third and fourth authors supported the conceptualization of the paper, and reviewed and edited the draft.

In the current Introduction chapter, firstly, I will begin by discussing in more detail the different purposes which individual types of phrasal structures are intended to serve in primary school FL contexts (section 1.2). Note that this discussion is framed within a curricular and pedagogical context. It is intended to explore the (perhaps simplified) pedagogical *purposes* of individual types of phrasal structures in curricula rather than contribute to a theoretical taxonomy of the many technical linguistic terms used to describe phrasal structures. In so doing, the discussion will touch upon some aspects of a usage-based constructionist approach to language learning and introduce the umbrella-term *multi-word-construction*. Those two notions will be integral to the remainder of this thesis.

Secondly, I provide a short overview of the usage-based constructionist approach to language learning which will guide the inquiry in the current thesis (section 1.3). Taking an example from Germany, I demonstrate how this approach to language learning has been incorporated into a real-world primary school FL curriculum.

Thirdly, against the background of constructionist views of language learning, I discuss a selection of research findings that demonstrate the important role of MWCs in learners' FL development from initial imitation to independent production, both in naturalistic as well as in instructed learning settings (sections 1.3.1, 1.3.2 and 1.4).

Fourthly, in the context of a usage-based constructionist approach to FL learning, I will zoom in on the important interplay of *generalization* and *input variability* in language learning (sections 1.5 and 1.6) and draw on research where the latter variable has been deliberately manipulated to benefit language learning (section 1.6). I will also explore the implications that this research might have for instructed primary school FL learning.

1.2 Phrasal structures – Multi-Word-Constructions (MWC)

Frequently used concepts in the literature and in primary school FL curricula referring to phrasal structures include *idioms*, *frozen phrases, chunks, routines, collocations, formulaic sequences*, *multiword strings, phrasal patterns,* and *verb-argument-constructions* (Christiansen & Arnon, 2017; Ellis, 2017; KM-BW, 2004, 2016; Wray, 2002; cf. Arnon & Christiansen, 2017, for an overview of technical terms). Different theoretical frameworks to systematize these concepts are proposed in the literature, including by structure (e.g., Becker, 1975), by function (e.g., Nattinger & DeCarrico, 1992), by length (e.g., Schwandtke, 2021), or by fixedness (e.g., Howarth, 1998). These theoretical frameworks are often based on corpus-research (cf. Czarnecka, 2011, for an overview). In general, the differences between terms such as *idioms*, *collocations*, or *phrasal patterns* can be characterized by a range of factors including their fixedness, their productivity, their expected frequency as well as their degree of idiomaticity and conventionality. Yet, as a 'least common denominator', the terms are all used to categorize different types of non-random and meaning-carrying 'strings of words' in a given language.

From a primary school FL curricular perspective, a basic difference between the different types of phrasal structures mostly lies in whether a structure possesses flexibly fillable slots, and if so, how many. The distinction whether a structure possesses fillable slots or not is an important one when accounting for the different ways of how curricula aim to establish and develop young learners' communicative agency. On the one hand, some phrasal structures -mostly without fillable slots- are viewed as instruments to establish young beginner FL students' initial basic communication. For example, curricula assert that *students develop a repertoire of words and phrases that refer to (basic) information about their person and individual concrete situations*, and *they can make themselves understood with the help of practiced formulaic expressions and short phrases* (KM-BW, 2016: 9, *translated from German*). With the intention to kick-start communication, a clear focus is put on *the availability of pre-formulated and holistically learned situational means of speech* (MSB NRW, 2012: 72, *translated from German*).

On the other hand, other phrasal structures -usually with flexibly fillable slots- are viewed as sources of linguistic knowledge extraction such that *in these construction patterns, students generalize their language experiences and make hypotheses about how patterns and sentences should be formed* (MSB NRW, 2012: 72, *translated from German*). It is stressed that phrasal structures are a means for students to *gain opportunities to separate language*

structures from the situation and to build up situation-independent language knowledge that they can use specifically for communication (KM-BW, 2004: 68, *translated from German*).

Importantly, as will become clear later (see section 1.6.3), research suggests that it is easier for learners to gradually extract linguistic information from structures that are flexible in the input. While this does not preclude that no information can be inferred from relatively fixed structures[2], flexible structures can be applied in a wider variety of contexts and are more conducive to introducing increased variability in teaching input compared to fixed structures. We will see in section 1.6 that this increased variability is beneficial for linguistic knowledge extraction, and ultimately for learning.

In primary school FL curricula in particular, a plethora of terms is used to describe slightly different types of phrasal structures. These terms include *chunks*, *formulaic sequences*, *frozen phrases*, *routines*, or *patterns*, among others (cf. Kostka, 2020). As pointed out above, some of those structures traditionally are relatively fixed (e.g., routines; *How are you?*) whereas others come with some flexibility through fillable slots (e.g., patterns; *My favourite _ is _*). The systematic review reported in section 2 focuses on these relatively routinized, formulaic, and context-dependent types of structures. In contrast, the experimental part of this thesis focuses on relatively abstract and compositional structures, such as verb-argument-constructions and long-distance dependencies. (More details regarding the reasons for this switch are provided in section 2.4). While these abstract and compositional structures might be considered 'higher order' compared to what is routinely taught in primary FL classrooms, it is important to note that relatively routinized and fixed patterns, such as *My favourite _ is _* are simply individual instantiations of a higher-order construction, for example to express one's favourite thing or person.

For the sake of using consistent terminology in the remainder of the current thesis, note that all terms mentioned above as well as those used more widely in curricula instantiate *constructions* in a constructionist view of language for they all "bind lexis, grammar, and meaning" (Ellis et al., 2016: 42) in one way or another and therefore represent phrasal form-function pairings (Goldberg, 1995, 2006; Jackendoff, 2002). This terminological 'summary' is

[2] Even highly routinized structures such as *How are you?* could serve as source of linguistic knowledge extraction. After all, the structure is not ungrammatical and follows the syntactic regularities of (in this case) the English language. However, in the context of primary school FL instruction, this structure would most likely rather be intended as a source for the immediate establishment of basic communicative agency through simple imitation. Other, more flexible structures such as *My favourite _ is _* (Kostka, 2020) would more likely be intended to provide students with a means to abstract linguistic information throughout exposure.

useful for the current thesis as it applies irrespective of whether an individual structure in a curriculum is primarily intended to serve imitation and the initial establishment of communicative agency (e.g., routines; *How are you?*) or the gradual extraction of linguistic information (e.g., patterns; *My favourite _ is _*). At the same time, the terminological 'summary' also applies to the 'higher order' abstract structures (e.g., verb-argument-constructions), or perhaps 'representations', of the previously mentioned relatively routinized and fixed structures. All three types of structures are *constructions* in the constructionist view of language (although some of them might be more formulaic and routinized or more compositional and abstract) since "any linguistic pattern is recognized as a construction as long as some aspect of its form or function is not strictly predictable from its component parts or from other constructions recognized to exist" (Goldberg, 2006: 5). And since the current work focuses on form-function correspondences at the phrasal level, that is, consisting of multiple words,[3] I will use the term *multi-word-construction*. Thus, for the remainder of this thesis, the different types of phrasal structures, ranging from relatively fixed structures like *routines* to more flexible ones like *patterns*, and extending to abstract structures such as *verb-argument-constructions* or *non-adjacent dependencies* which will become important in the experimental part of this work, will be regarded as instances of MWCs.[4]

Given the preceding discussion, one might rightly ask: *What is not an MWC*? From a theoretical constructionist perspective, the author agrees that, beyond single words, almost any non-random string of two or more words could be considered a MWC. Yet, as highlighted above, for readability, MWC serves as a common denominator or 'umbrella term' for a variety of terms that describe linguistic structures at a general phrasal level. Importantly, the term MWC can capture both the terminological scopes of the systematic review (section 2) and of the experiments (sections 3 and 4). The linguistic structures examined in these sections all represent form-meaning correspondences, ranging from formulaic and routinized to

[3] The concept of a linguistic construction extends beyond the syntactic level to encompass other linguistic domains, such as morphology.

[4] Note, as will become clear to the reader, the authors of the systematic review reported in Chapter 2 (Schulz et al., 2023) coined the range of terms describing phrasal structures with the umbrella-term *multi-word-units* (MWU) instead of *multi-word-constructions*, following suggestions by Arnon and Christiansen (2017). It was only after publication and following further discussion with colleagues at CALP4 (*Constructionist Approaches to Language Pedagogy* conference, Friedrich-Alexander Universität Erlangen-Nürnberg, Germany, March 2024) that it became clear that the umbrella-term *multi-word-construction* might have been a better fit since the more universal term *construction* by definition comprises *all* meaning-carrying 'strings of words' in the constructionist tradition of phrasal form-function correspondence (Goldberg, 1995, 2006). In the current thesis, only the term MWC is used.

compositional and abstract. Thus, the term MWC, despite its admittedly broad applicability, serves as a convenient tool for harmonizing the terminology used in the systematic review and the experiments. It helps avoid constant shifts between technical terms. This is particularly relevant since the linguistic nuances between these terms do not directly pertain to the educational subject under investigation in this thesis.

While the systematic review presented in section 2 focused on the kinds of relatively formulaic, routinized and context-dependent structures prevalent in primary FL classrooms, the experimental part of this thesis will focus primarily on the more abstract MWCs under the assumption that those might be more beneficial for learners' development of independent communicative agency since it is easier to systematically manipulate their structure in the input. This notion will be discussed in more detail in sections 1.6.1 and 1.6.3.

In the following section, I provide a short overview of the usage-based constructionist approach to language learning referenced in the preceding paragraphs. Taking an example from Germany, I demonstrate how this approach to language learning has been incorporated into a real-world primary school FL curriculum.

1.3 Usage-based constructionist approach to language learning

The usage-based constructionist approach to language learning makes input-driven predictions maintaining that through domain-general cognitive mechanisms learners induce structural information and communicative functions of linguistic structures from the input embedded in specific communicative situations (Behrens, 2009; Ellis & Wulff, 2015; Robinson & Ellis, 2008; Tomasello, 2003). Language development is driven by statistical learning of form-function pairings guided by the distributional properties of the structures in the input; learners *generalize* from the input across contexts (Ellis & Wulff, 2020). In other words, "learners acquire generalizations" from the input (Goldberg, 2009: 93). This important role of input is stressed in FL curricula too (cf. KM-BW, 2004: 6). Usage-based theory further maintains that across multiple encounters with specific constructions (i.e., form-function pairings), increasingly abstract construction-specific form-function mappings develop (also referred to as *constructional abstractions*; Robinson & Ellis, 2008). The theory therefore predicts a construction-based language development (Goldberg, 1995; Goldberg, 2006; Yuldashew et al., 2013; or *exemplar-based* Ellis, 2002). Via gradual pattern abstractions across input and usage-events, learners' linguistic inventory, additionally impacted by input characteristics such as salience and frequency, emerges along a continuum of construction-abstractness. This continuum can be thought of as a 'constructional spectrum' (Ibbotson, 2013). It varies from concrete, initially unanalysed units [I don't know], via semi-fixed formulas [I don't + verb], to abstract constructions like syntactic schemas [Subj Verb (Obj)] (Bannard et al., 2009; Ibbotson, 2011, 2013; Robinson & Ellis, 2008). In other words, what develops over time is "an abstraction across (memories of) usage events, that is, across exemplars of one constructional type" (Madlener-Charpentier, 2015: 32).[5] Importantly, the more abstract the mental representations of constructions become in this process, the more productive those representations can become as well (Bannard & Lieven, 2009; Eskildsen, 2009). Consequentially, learners' communicative agency increases along the constructional spectrum as they journey from imitation to production (Sambanis, 2007).

Usage-based Language Learning theory -an inherently constructionist approach to language learning- is often invoked to substantiate the importance of MWCs in driving learners' development from imitation to (independent) production. Although initially developed in L1 acquisition (e.g., Tomasello, 2003), usage-based approaches have been extended to FL learning

[5] A ‚constructional type' could be, for example, a ditransitive structure such as *X Verb Y to Z*.

(Eskildsen, 2009) and incorporated into curricula. For example, KM-BW (2004: 69, *translated from German*) states that FL *beginner learners initially make intuitive metalinguistic assumptions about language structures. Then, learners apply those assumed structures in communication, receive feedback from their environment, and subsequently draw metalinguistic inferences which potentially trigger an adaption of their initial assumptions. This way, learners explore linguistic elements and assign individual functions to them while establishing a classification of linguistic categories. Over time, these processes lead learners to develop linguistic structures equipping them with the necessary instruments to understand aural and written input and to produce utterances in a planned manner*. Other primary curricula[6] offer similar accounts of young students' language learning process, asserting that *students generalize their language experiences and make hypotheses about how shapes and sentences should be formed* (MSB NRW, 2012: 72, *translated from German*). Examples like these demonstrate that the usage-based constructionist approach to language learning has been adapted in current primary school FL curricula. Therefore, the experimental inquiry in this thesis will be guided by this theoretical approach.

Against the background of constructionist views of language learning, I will now introduce findings that demonstrate learners' FL language development from imitation to production along the constructional spectrum. The evidence supports the facilitating role of MWCs in naturalistic and instructed learning settings. Having reviewed only trace-back and observational studies that did not actively manipulate the MWC input either in naturalistic nor in instructed settings, I will then turn to the role of generalization and the impact of input manipulations through manipulated input variability in the context of a constructionist approach to language learning. After that, I will discuss findings from studies where the structure of the MWC input was actively manipulated through input variability, and explore the implications that these findings could carry for instructed primary school FL learning.

1.3.1 Naturalistic Foreign Language development

Regarding FL learning, as early as 1938, Kenyeres and Kenyeres tracked their Hungarian-speaking daughter's untutored FL French learning in Geneva, finding that she frequently imitated memorised MWCs to communicate. Following an argument with her mother after two months of French exposure, the child produced the following recombination of individual MWCs to clear the air: *Maman, // s'il vous plait // qu'est-ce que c'est // voulez-vous*? [Mummy,

[6] In Germany, each federal state has their own curriculum.

please, what is it, would you like (to)]? After three months, she began to modify elements in MWCs relative to the communicative context, such as changing 3rd person singular verbs to plural. Decades later, Fillmore (1976) tracked five Mexican children's FL English immersion development, reporting that they gradually extracted individual constituents from MWCs according to communicative context. The author argues that once all constituents of a MWC had been extracted, the children were left with an abstract grammatical structure which they could use generatively. Other traceback studies confirm that children's naturalistic FL acquisition heavily relies on MWCs (Hakuta, 1974; Huang & Hatch, 1978). These early studies demonstrate the key characteristics of MWC's role in FL learning and in FL pedagogy: (a) MWCs equip learners with early communicative agency through imitation and (b) constitute developmental 'starting points' for generalisation and abstraction processes.

1.3.2 Instructed Foreign Language development

Building on naturalistic FL development findings, research into instructed adult FL learning with necessarily limited input (given it is predominantly available only in the classroom) has become prominent in recent years (Ellis, 2015; Ellis & Ferreira-Junior, 2009; Ellis & Wulff, 2020; Eskildsen, 2009; Eskildsen & Cadierno, 2015; Mellow, 2006; Roehr-Brackin, 2014; Tyler, 2010). For example, Eskildsen (2009) traced the use of *can*-patterns in longitudinal data collected from one adult Spaniard (Carlos) learning English as a Second Language (ESL). The data included 120h aural recordings of Carlos in ESL classrooms over four years. Counting MWC occurrences in the participant's production data, Eskildsen distinguished a 'starting' pattern (*I can + verb*) based on which Carlos gradually developed more sophisticated constructions, such as *can you* / *can I* / *you can* + *verb*. These findings correspond to work in naturalistic FL settings, suggesting that new and more abstract patterns emerged from initially concrete units.

In classrooms with younger learners, Myles, Mitchell and Hooper (1999) investigated MWC's role in students' creative production of FL interrogatives. Over two years, they collected data from 60 beginning French FL learners in Year 7 (age 11) in England. Like Eskildsen (2009), Myles et al.'s data suggest that interrogative chunks such as *comment t'appelles-tu* [what's your name?] were catalysts for linguistic analysis and free production. For example, late in the study, students used structural information from an initial chunk (*comment t'appelles-tu*?) as a schematic facilitator for advanced utterances in a different context like *comment s'appelle le garçon/fille* [boy/girl]? Those findings suggest that younger learners too can learn a FL from initially unanalysed chunk input which they gradually segment and analyse across contexts.

Over time, this process leads to increasingly productive use of abstract structural knowledge (cf. Kersten, 2015).

1.4 Multi-Word-Constructions in young learners' Foreign Language classrooms

Since research suggests that MWCs are beneficial to FL learners in naturalistic and instructed settings, it appears logical to purposefully implement MWCs into FL classroom teaching to facilitate learning. Indeed, MWCs have been investigated in non-beginner adult FL instructed contexts (Boers, Dang & Strong, 2017; Boers, Eyckmans, Kappel, Stengers, & Demecheleer, 2006; Boers & Muñoz-Basols, 2021; Choi, 2017; Eyckmans & Lindstromberg, 2017; Thomson, 2020; Wood, 2009; cf. Boers & Lindstromberg, 2012, for an overview). However, such work has mainly focused on the relationship between instructing MWCs and developing fluency (Thomson, 2020), idiomaticity (Boers & Muñoz-Basols, 2021) and processing speed (Choi, 2017). Here, MWCs are not considered developmental 'starting points' because participants are often already proficient, or at least past the initial learning stages. The fact that the learners are not beginners emphasises an important distinction between MWCs as drivers of fluency and idiomaticity and as drivers of language acquisition (Aguado, 2002). Since MWCs are sources for grammatical knowledge extraction, they can arguably be catalysts of language learning at the *beginning* of learning trajectories (Kersten, 2015).

It follows then that young FL learners at the start of their FL learning journey might particularly benefit from targeted MWC input to launch their FL development. Primary school FL instruction has been introduced worldwide and some contexts have indeed implemented findings supporting MWC's facilitating role in language learning into primary FL curricula (see section 1.3). In parallel, results from some evaluations of primary school FL learning outcomes have begun to reflect MWC's important role in young students' FL learning along the constructional spectrum discussed in section 1.3. A relevant and typical example of such research is found within the context of Germany where primary FL curricula are designed in alignment with the broader educational policies of the European Union. The following paragraph, therefore, discusses evaluations in Germany, highlighting the role of MWCs in young learners' FL development.

Several studies in German primary schools report on children's FL attainment.[7] Engel, Groot-Wilken and Thürmann's (2009) large-scale assessment of FL attainment at the end of Year 4 (age 9–10) found that 42.2% of learners mainly reproduced MWCs in their communication (i.e., imitation), while 24.3% could sometimes vary individual units of MWCs.

[7] The foreign languages (FLs) in these studies are typically English and, occasionally, French, depending on the federal state where the individual study was conducted.

The so-called BIG-study (BIG-Kreis, 2015; Müller et al., 2016), another large-scale assessment at the end of Year 4, reports that 79.5% of utterances in student dialogues consisted of holistically learned MWCs. Other studies focused on learners' learning processes rather than assessing learning outcomes. Werlen (2008) tracked the FL progress of twelve primary school classes over four years (Years 1–4; beginner learners aged 6–10) through video-recordings of English/French classes, audio data transcriptions, and interviews with individual learners. She reports that learners gradually extracted MWCs from the input and modified them for personal communication goals. Sambanis (2007) reports similar longitudinal action research data from Year 1 and 2 English and French classes, maintaining that some learners could vary MWCs via slot-filling. Kahl and Knebler (1996) report that MWCs provided weak students in Years 3 and 4 with linguistic frames to rely on while strong students used them to advance to independent productions, for example via productive slot-filling through codeswitching: *The bird is green* → *The bird is auf dem Dach* (*on the roof*). Diehr (2009) analysed speech data of 216 children from ten classes in Years 3 and 4. She reports that by the end of Year 4, many students managed to variably fill slots in patterns, showing signs of a developing awareness for grammatical structure. Note that such findings correspond closely to what curricula (e.g., KM-BW, 2004) assert regarding learners' development.

Evaluation findings such as those presented in the previous paragraph suggest that MWCs are likely to play an important facilitating role throughout a young learner's instructed FL development. On the one hand, the evaluations suggest that MWCs promote the achievement of early communicative agency through imitation. On the other hand, the evaluations also conform to constructionist notions suggesting that MWCs constitute integral building blocks and catalysts for primary students' FL development. Corresponding to the mechanics of a usage-based constructionist approach to language learning and corroborated by evidence from FL studies, classroom evaluation data suggests that the imitative characteristic of fixed units establishes young learners' communicative agency while the productive characteristic of variable patterns with fillable slots facilitates the advancement of their communicative agency. However, several reasons make it difficult to discern cause-effect relationships between teaching input and the results of evaluation studies. Firstly, while trace-back studies such as Fillmore (1976) are situated in untutored immersion contexts, evaluations report on instructed settings. Critically, learning context is important since FL input in teaching settings cannot compare to the quantity and quality of input in immersion settings. Secondly,

even in contexts where MWCs are established in FL curricula, it is the individual teacher's decision as to how to implement them. Although recommendations about MWC teaching exist, such as advocating teaching high-frequency MWCs (Bredenbröcker, 2018), we typically do not yet know which MWCs were taught in which situations and how. This lack of detail makes it impossible to link evaluation findings to confident conclusions about MWCs' role in primary school FL development.

We have seen evidence from a range of traceback-, observation-, and evaluation-studies suggesting that FL development both in naturalistic and (adult and adolescent) instructed settings seems to follow the trajectory predicted by a usage-based constructionist approach to language learning. MWCs in the input appear to play an important role on this trajectory. However, it has remained unclear exactly how MWC input impacts *young* learners' FL development in instructed settings. Given the proliferation of FL instruction at younger and younger ages, this issue is increasingly important. From a pedagogic perspective, the question arises: How can the MWC input be structured to be most beneficial for the development of young learners' communicative agency? To address this question, I will now turn to a core learning mechanism in language learning, which I touched on before (see section 1.3), namely generalization, and explore evidence demonstrating how input variability can facilitate generalization processes, and ultimately support learning.

1.5 Generalization

At the heart of current primary school FL learning goals lies a high degree of communicative agency (see section 1.1). From a constructionist view on language learning, communicative agency beyond imitation is contingent upon the degree of generativity of a learner's linguistic repertoire. The degree of generativity relates to the degree of abstractedness of a learner's systematic mental network[8] of productive constructional abstractions such as verb-argument-constructions. Those constructional abstractions equip the learner with communicative flexibility. In turn, the development of such abstract linguistic representations is contingent upon a learner's ability to induce rules and systematicities across individual occurrences of specific linguistic structures in a variety of contexts (Goldberg & Casenhiser, 2008; Wulff & Ellis, 2015). A typical induction process where linguistic units are encountered in a variety of contexts could develop from concrete units [*I don't know*], via semi-fixed formulas [*I don't + verb*], to abstract constructions like syntactic schemas [*Subj Verb* (*Obj*)] (Bannard et al., 2009; Ibbotson, 2011; 2013). In the end, learners have acquired a generalization. Thus, at the heart of a usage-based constructionist approach to language learning and of the process of establishing communicative agency lies the ability to arrive at generalizations.

'Generalization' comprises non-language specific cognitive mechanisms such as pattern detection, analogy, schematization, and categorization (Goldberg et al., 2004; Tomasello, 2003, 2009). For example, categorization strategies include the ability to group linguistic experiences into classes based on shared individual-characteristics and distinctive group-characteristics. In all domains of our lives, including language learning, generalization is considered a core component of development (Mitchell, 1982). For example, we are unlikely to encounter and therefore cannot know about every shape a bicycle can possibly have. However, past experiences of seeing all sorts of bicycles helped us build a mental representation of which general features we expect a bicycle to have. That is, we induced bicycle-specific information across encounters. This process of acquiring a generalization allows us to apply past experiences to novel contexts, where we can make an educated guess about what a bicycle should look like. We are thus able to determine if some new apparatus we encounter is a bicycle or not, even if we have never seen that particular exemplar before. In essence, eventually we

[8] In constructionist theory, this mental lexicon is sometimes referred to as *construct-i-con* and can be thought of as a storage of schematic templates of constructions, that is, of abstract form-function pairings (cf. Hilpert, 2014).

come to learn what a ‘bicycle’ is and can extend this knowledge to novel exemplars (cf. Raviv et al., 2022).

1.6 Input variability

While generalization is a core mechanism of learning, input variability -simply the variation in our experience with different exemplars- influences generalization and impacts learning outcomes (Raviv et al. 2022). Research in cognitive sciences across many domains including motor learning (cf. Bortoli et al., 1992), computational modelling (cf. Tenenbaum & Griffiths, 2001), problem solving (Likourezos et al., 2019) and education (cf. Bjork & Bjork, 2011) has suggested that increased initial input variability can effectively improve generalization, and therefore enhance learning. Consider input variability regarding our bicycle example: On the one hand, if we always encountered the exact same type of bicycle (e.g., brand, colour, shape of the handlebars, etc.), we would quickly perfect our skill to recognize this particular object as a 'bicycle'. However, we might struggle to recognize other types of bicycles as 'bicycles' if we only relied on the idiosyncratic information we induced from the one type of bicycle we have previously encountered. On the other hand, if we encountered a *different* bicycle each time we encounter a 'bicycle' (i.e., different brands, colours, shapes of handlebars, etc.), it might initially be more difficult for us to recognize an unknown object as a 'bicycle' because we experienced many different cues as to which features resemble a 'bicycle'. However, with greater experience with bicycles we would become more skilled at categorizing a large variety of different exemplars of bicycles as 'bicycles' due to our ability to generalize as to which features most consistently constitute a bicycle, such as the presence of two wheels and two wheels only, or the presence of a handlebar and pedals. Vice versa, we learn to ignore the irrelevant features, such as the bicycle's colour, the brand, the wheel's specific type of tyres, or the specific shapes of different types of handlebars. In other words, the classification of novel (bicycle) items constitutes a gradual learning process facilitated by input variability (Wahlheim et al., 2012). Thus, initially, increased variability might slow down learning as it poses a more difficult cognitive challenge (cf. Viviani et al., n.d.), but ultimately it would pay off "in increased generalizability of what is learned" (Raviv et al., 2022: 462).

Recall that in section 1.4, I highlighted the crucial role of MWCs with fillable slots in enhancing students' learning beyond imitation, in contrast to fixed MWCs. It is precisely this advantageous impact of variability on learning that underscores the importance of fillable slots in MWC input, as variability in the input is contingent upon these fillable slots. I will explore the role of input variability in language learning in the following section.

1.6.1 Input variability in language learning

In the following, I extend the input variability logic of the bicycle example to language learning, specifically to the learning of form-function mappings inherent to MWCs. When learners encounter few distinctive types of a MWC, such as verb-argument-constructions like V(erb) *about* N(oun) or non-adjacent dependencies like *AXD, BXE, CXF*, the properties associated with these MWCs are highly restricted due to the invariable nature of the input. This low type-frequency in the input results in an insufficient supply of reliable cues about the construction's linguistic properties. Consequently, the properties remain restricted which means they can only apply to a limited range of linguistic items which, in turn, constrains the construction's productivity. This negative reciprocal interaction between a construction's low type frequency and its restrictive properties makes generalization difficult (Bybee, 1995, 2008).

In contrast, when learners encounter many different types of a construction, more linguistic properties associated with this construction can be induced because the variable input provides a larger number of cues about those properties. Specifically, the increased number of cues helps learners to distinguish irrelevant cues from core cues which help to reliably shape a construction's set of properties. Learners now have reliable data as to how to 'fission' the input structure (Peters, 1983). Consequently, the larger number of identified linguistic properties can apply to a wider range of items which enhances the construction's productivity (Bybee, 1995, 2008; Madlener-Charpentier, 2016). This positive reciprocal interaction between a construction's high type frequency and its versatile properties makes generalization easier. Note that for variability effects to work in linguistic contexts, the target construction (unlike fixed sequences such as *routines*) must possess generative features (i.e., productive slots) in the first place, otherwise variability would not be possible to any extent.[9] That is why variable patterns such as constructions with flexible slots like the verb-argument-constructions with productive verb-slots used in Wonnacott et al. (2012) (*V(erb) N(oun)*$_2$ *N(oun)*$_1$; see section 1.6.3) or the non-adjacent dependencies with productive intervener-slots used in Gómez (2002) (*AXD, BXE, CXF*; see section 1.6.3) are ideal for exploiting variability effects in language learning.

To conclude, as in other domains of learning, input variability is beneficial for generalization in language learning because variable linguistic input provides a larger number of cues which help learners to induce linguistic properties of target MWCs, such as verb-

[9] Vice versa, it is exactly this lack of variability which allows these structures to be learned as fixed phrases.

argument-constructions or non-adjacent dependencies. Having a wider range of cues at their disposal supports learners in distinguishing between relevant and irrelevant cues and, in turn, helps them develop a reliable structural frame of the MWCs from the input; generalization is promoted. In contrast, low variability in the input cannot provide enough distinctive cues for learners to distinguish between relevant and irrelevant information. This scarcity of cues makes it more difficult for learners to draw appropriate conclusions about the target MWC's linguistic properties and generalization is impeded.

1.6.2 Usage-based constructionist language learning, generalization, and input variability

Generalization is a crucial mechanism in a usage-based constructionist approach to language learning that facilitates the development from lexically specific exemplars to abstract structural knowledge (Ellis & Wulff, 2020). While usage-based theorists claim that both ends of this constructional spectrum are part of a learner's linguistic repertoire, critics have argued that the approach struggles to provide specific predictions as to which factors precisely impact generalization, and accordingly, what influences the development of constructional abstractions (Ibbotson, 2013). This is where the structure of the input comes into play, specifically the variability of the input (although other factors such as pragmatic constraints are important too; cf. Goldberg, 2009). Considering that 'language learning' in a usage-based constructionist view is inherently input-driven (Harrington & Dennis, 2002), research has suggested that input variability is an important contributing factor to explain and predict generalization, and thus, language development (e.g., Eidsvåg et al., 2015). Several studies (reviewed in section 1.6.3) suggest that the (high/low) degree of generalization (which determines a learner's developmental outcome on the constructional spectrum) is partly a result of the input's degree of variability. This research tallies with theoretical accounts of the role of input variability in language learning (cf. Bybee, 1995) discussed in section 1.6.1. Thus, language development along the constructional spectrum maintained by a usage-based constructionist approach to language learning can be approximated as an increasingly sophisticated inductive generalization process following domain-general psychological principles, where said generalization process is to some extent determinable by the input's degree of variability.

In the following section, I will review findings suggesting that input variability facilitates generalization in different domains of language learning. Based on this evidence, I will turn to

instructed settings and discuss how input variability in the context of teaching MWCs could have beneficial effects on young FL learners' development of communicative agency, thwarting claims suggesting that primary school students lack productive FL skills.

1.6.3 Input variability – Previous research

In an artificial language setting, Gómez (2002) investigated input variability effects on the learning of non-adjacent dependencies. Such dependencies are ubiquitous in language use; they describe a long-distance governing-relationship between two linguistic units that are separated by at least one third unit. For example, in the sentence *The girl in the garden plays tennis* the subject *the girl* governs the inflectional suffix of the verb *play* while the dependency is interrupted by the prepositional phrase *in the garden*. Gómez (2002) compared the performance of adults and 18-month-olds as prior related work had suggested that both infants and adults can segment continuous speech into word-like units based on the rapid and domain-general learning of transitional probabilities of adjacent syllables (e.g., Aslin et al., 1998). In her study, participants were exposed to three distinct three-element strings structured A-X-D (*pel-wadim-rud*), B-X-E (*vot-wadim-jic*), and C-X-F (*dak-wadim-tood*). 'A' (*pel)* appeared with 'D' (*rud)*, 'B' (*vot)* appeared with 'E' (*jic)*, and 'C' (*dak)* appeared with 'F' (*tood*), irrespective of the intervening X-element. Participants were supposed to learn these long-distance relationships without explicit instruction. The number of middle elements ('X') was systematically varied between participants, including set sizes of either 2, 6, 12, or 24 elements. Gómez (2002) reports that after exposure both age groups managed to differentiate successfully between correct (A-X-D, B-X-E, C-X-F) and incorrect (e.g., A-X-B) three-element structures only when they had experienced the highest degree of variability in the X-position during exposure. According to the author, this suggests that both groups -provided the input was highly variable- were able to learn the word order rules as demonstrated by their ability to successfully judge novel (incorrect) three-element strings. Gómez (2002) concludes that learning the invariant nonadjacent dependencies (e.g., 'A' (*pel*) appears with 'D' (*rud*)) was facilitated by high input variability because this decreased the predictability of *adjacent* dependencies, such as between A (*pel)* and X (*wadim/kicey/puser/fengle/etc)*. In turn, the unreliability of adjacent dependency cues increased the learners' reliance on a wider set of discriminatory cues, specifically in this case on *non*-adjacent dependencies. Gómez and Maye's (2005) replication study with 15-month-olds suggests that they too can induce generalizations about non-adjacent dependencies from variable input. Note that Gómez's work did not involve measurements

regarding the participants' ability to apply the acquired generalizations to novel contexts. While her data demonstrate that the participants learned the underlying structure of the non-adjacent dependencies, it remains unclear whether they would have been able to extend this knowledge productively to novel contexts. This observation shall become important later for experiment 2 (see section 4).

Turning from non-adjacent dependencies to word category learning, Twomey et al. (2014) suggest that learning from multiple exemplars of a category facilitates children's learning of word categories. In addition to several known categories (e.g., vehicles, animals), the authors exposed 2-year-olds to several unknown categories in an artificial language such as 'doff' (a category representing a set of artificial geometrical shapes which differ in colour). The children encountered either multiple different exemplars of an unknown category (e.g., the '*doff*-category' shape in several different colours) or repetitions of only one exemplar (e.g., the '*doff*-category' shape in repeatedly the same colour) during referent selection trials. When tested on novel objects of a learned (unknown) category (e.g., another new colour of the *doff* shape not previously encountered), only children in the high-variability group showed learning of the name-object associations across referent selection trials. The authors argue that distinguishing between relevant and irrelevant features of a word-category is impacted by category-specific exemplar variability in the input. Note, this category learning is similar to the 'bicycle' example in section 1.5.

Regarding argument-structure construction learning, Casenhiser and Goldberg (2005) exposed 6-year-olds to a construction involving the novel form NP_1NP_2V and a novel abstract event semantics. This construction, set within an English context using novel verbs, depicted the entity denoted by NP_1 appearing in or on the location denoted by NP_2. For instance, the sentence *The king the chair vakoed* illustrated a king appearing on a chair. Children learned these semantics by watching a series of 16 animated scenes accompanied by audio, presented in a single block lasting approximately 3 minutes. Results indicated that better generalization of the abstract construction to new vocabulary occurred when the input skewed toward one particular nonsense verb, as opposed to an equal distribution of examples with each novel verb. These findings were consistent with similar experiments involving adult learners (Goldberg, Casenhiser, & Sethuraman, 2004). Yet, while Casenhiser and Goldberg (2005) -unlike Gómez (2002)- demonstrate that young learners generalize form-function pairings in the input after minimal exposure, there was no measurement investigating participants' ability to extend (i.e.

active production) this knowledge productively to novel contexts. This means that the children only had to recognize target structures during testing whereas other tasks such as act-out or production tasks would require children to recall and productively apply individual features of the acquired form-function correspondences (Goldberg, 2007).

In a similar artificial language study by Wonnacott et al. (2012), the repertoire of outcome measurements is expanded to capture children's ability to extend the generalization which they acquired from the input to novel contexts. In addition, and importantly so, Wonnacott et al. (2012) investigated the role of input variability, which the previous studies did not address. The authors' data from 5-year-olds suggest that variability in the productive verb-slot of a novel construction impacts generalization of the construction to novel verbs. Children learned an artificial approach-event construction (VN_1N_2; N_2 approaches N_1 in the manner denoted by V), such as *Chadding rabbit gorilla.* ('Chadding' meaning (to) hop on one's head). During exposure, one group of children encountered the same verb in the verb slot on each trial (low variability; LV), while the other group encountered different (novel) verbs in each trial (high variability; HV), with four different verbs in total. During testing, all children were exposed to novel verbs. Critically, only children in the HV condition were able to both generalize and extend their knowledge, that is, they could understand and produce the target structure when the verb was unknown. Based on their findings, the authors conclude that type frequency (i.e., variability in the verb slot) facilitates generalization. Specifically, they argue that children's difficulties to generalize to novel contexts after having been exposed to only one type is a result of the input structure instead of input quantity (which was kept constant across conditions). This generalization-facilitating effect of type frequency in the input compared to token frequency has been attested elsewhere. Drawing on grammatical phenomena from English, German, Arabic and Hausa, Bybee (1995) suggests that increased type frequency leads to morphological categories being more productive and more generalizable because more cues become associated with a category, which increases the category's inclusivity to cues from non-attested items. In contrast, high token frequency was shown to prevent category generalization because it limited the number of features associated with a category, constraining its productivity regarding novel elements.

The studies reviewed so far in this section have all (partly) used artificial languages. One study that used natural languages was conducted by Eidsvåg et al. (2015) who demonstrate that input variability is beneficial for the learning of grammatical subcategories. They

investigated the learning of Russian noun gender by adult learners without prior knowledge of Russian in a HV and a LV condition, where 32x different nouns were repeated once, or 16x different nouns were repeated twice, respectively. Only the HV participants showed learning of grammatical gender categories. Similar to Gómez (2002) in an artificial language setting, Eidsvåg et al. (2015) argue in a natural language setting that the degree of input variability impacts generalization of grammatical information.

Another study which used natural language input was conducted by Viviani et al. (n.d., pre-registered, *in-principle accepted*). They conducted a study where 7-to-8-year-olds learned some Japanese. The English L1 children learned two Japanese postpositions (equivalent to English prepositions *above* and *below*) in either a HV or LV training condition. They were prompted to move different objects around a grid responding to a *N(oun)+N(oun)+P(ostposition)* structure. In the LV condition, children were exposed to four unique sentences (two including *above* and two including *below*) which were repeated 14 times each. In the HV condition, children were exposed to 56 unique sentences (28 including *above* and 28 including *below*), repeated once each. Thus, each child heard 56 Japanese sentences. The eight involved nouns (i.e., the objects on the grid) were English cognates and kept constant across conditions. Viviani et al.'s pilot data suggest that children in the LV group show better performance during training, whereas the HV group shows better performance during testing with sentences containing novel nouns. Although data collection is not completed at the time of writing, the preliminary analyses suggest that irrespective of the involved objects, variable input supports children's ability to generalize the invariant features of the postpositions' semantics to unattested contexts. Furthermore, their data which indicate superior performance of the LV group *during training* corroborate Raviv et al.'s (2022) argument that variable input might initially pose a more difficult cognitive challenge than repetitive input.

1.6.4 Input variability in the Foreign Language classroom

The studies discussed in the previous section highlight which input factors might be profitably manipulated to provide learners with the most effective environment to enhance their generalization. Those factors are: (a) The target construction has to have generative features (i.e., possess productive slots) to allow for versatile properties which in turn can be applied to a large(r) set of items. (b) The target construction requires high 'cue-strength' reflected by high type frequency in the input (i.e., high variability). In fact, although for each study in slightly different ways, all studies discussed in section 1.6.3 manipulated those two

factors (a) by using generative constructions with (artificially) productive slots (like many MWCs in primary school FL curricula) and (b) by manipulating the structure of the input between exposure conditions. The resulting variability effect appears to be a robust one (cf. Raviv et al., 2022, for non-language specific domains of learning). While those findings all come from controlled settings, there is no reason to assume that a similar manipulative (and arguably beneficial) system could not be applied to a primary school FL classroom context. Many MWCs in teaching contexts are generative as they come with productive slots. In fact, as discussed in section 1.3, this generativity is specifically intended by curricula to enhance generalization and communicative agency. In addition, type frequency in the input can be manipulated even in classroom settings. In sum, there might be a realistic potential for young FL learners to benefit from input variability effects in classrooms when exposed to MWC input.

1.7 Summary

The evidence from language learning studies, and also from other domains (cf. Raviv et al., 2022), suggests that the variability effect generally is a robust one. Specifically in language learning, input variability appears to facilitate generalization, and might thus be a promising factor to manipulate in the MWC input in primary school FL classrooms. In fact, targeted manipulations of type and token frequencies in the input have been considered to have important implications for FL teaching, as noted by Goldberg and Casenhiser (2008). Accordingly, Robinson and Ellis (2008: 509) raise the questions: "How do type and token frequency interact in learning SL [Second Language] constructions? […] Does increasing type variation in the verb slot lead to greater abstraction and generalizability of argument structure constructions?" Considering the worldwide rise in primary school FL instruction, those questions are certainly important to address in FL research. However, we have not yet encountered relevant primary school evidence in this regard (cf. Madlener-Charpentier, 2015, 2016, and Henk, 2019, for research on input variability in FL classrooms with adult learners). Given the significant role of MWCs in language learning and an understanding of language learning as being construction- or exemplar-based, the question arises whether the variability effect can be extended to MWC input in an authentic primary school FL classroom for the benefit of the learners. Put differently, increased MWC input variability might support instructed primary FL learners' ability to acquire generalizations about the MWCs' underlying form-function correspondences and extend this knowledge to novel contexts, thereby increasing their communicative agency.

1.8 Outlook

In the following chapter, and before introducing my own experimental research, I will report on a systematic review (Schulz et al., 2023) of the current state of the research regarding the impact of MWC input in early FL learning and teaching contexts. This systematic review, constituting the first part of empirical work of this thesis, aims to comprehensively scope the available evidence regarding the role of MWC input in instructed primary school FL learning. Its objective is to provide valuable insights to inform the subsequent classroom research detailed in Chapters 3 and 4.

There are two important issues to highlight: First, the systematic review encompasses various forms of MWC input in primary school FL settings, without a specific focus on input variability as a teaching manipulation. This is because the systematic review was conducted before I made the decision to centre this thesis's experimental inquiry on the potential of a 'variability manipulation' in MWC learning and teaching contexts. In fact, the decision to explore input variability in the experimental inquiry emerged later in this doctoral research, once it became apparent that manipulating input variability might be a promising strategy for targeting MWCs in the input, as suggested by empirical evidence presented in section 1.6.3. Second, conducted at the beginning of this doctorate with a slightly different research focus, the systematic review examines the kinds of structures already used in primary schools, such as *routines* (e.g., *How are you?*) or *patterns* (e.g., *My favourite _ is _*). While some of those structures include fillable slots, they remain mostly context-dependent and formulaic. Thus, terms like *verb-argument-construction*, or *non-adjacent dependency*, which become important later in this thesis, did not feature in the review's search string. If there was any classroom-based research targeting these more abstract and compositional structures, the search would not have detected it. Nevertheless, the systematic review remains a valuable and rigorous overview of the general state of research on the kind of MWC input already ubiquitous in primary school FL contexts. More information regarding the transition to the experimental studies reported in this thesis is provided in section 2.4.

Building on the results of the systematic review, and on other research discussed in the current chapter, I will then move on to report on two teaching intervention experiments in primary school FL classrooms where learners were exposed to targeted MWC input, and said input was manipulated in terms of its variability (see sections 3 and 4).

2 Systematic Review[10]

Selective literature reviews for teaching intervention studies focusing on the impact of any type of targeted MWC input in primary school FL settings were unfruitful. While many findings from FL studies suggest that MWCs can facilitate early FL learning, we lack robust classroom-based evidence that MWC input, already established in curricula, has a measurable effect on specific aspects of students' FL attainment, such as productive skills. Without that evidence it is not possible to offer evidence-based recommendations for FL teaching.

We therefore conducted a systematic review to provide a rigorous overview of the work that has been done in this research area and to critically evaluate and synthesise relevant results. Understanding the effects of purposefully focusing on MWC input in early FL teaching could be powerful for FL researchers, stakeholders, and practitioners. Firstly, it would provide researchers with evidence to better understand and explain children's FL development findings from evaluation studies, which in many respects tally with usage-based constructionist theory and naturalistic FL learning trajectories. Secondly, it would help policymakers, teacher educators and teaching material developers to justify and potentially readjust the inclusion of MWCs in curricula, teacher education and materials. Finally, it would support teachers in knowing to what extent certain types of focused MWC input impact learning outcomes and how those might be improved. However, without the requisite research, the potential impact of teaching MWCs to young FL learners in formal educational contexts remains guesswork. Consequently, this review addresses the following research questions:

RQ 1. What is the extent of original research investigating the role of MWC input in early FL learning and teaching contexts?

RQ 2. What is the extent of the evidence on the effectiveness of learning from/teaching MWCs in early FL contexts?

2a. What impact does MWC input have on young learners' grammatical and vocabulary skills and knowledge in the target language?

[10] Note that parts of this chapter are adopted from Schulz et al. (2023). In that paper, the first author conceived and designed the analysis, collected the data, performed the analysis, and wrote the paper independently. The second author contributed to the double-blind screening and quality appraisal, as is good practice in systematic reviews, and reviewed and edited the draft. The third and fourth authors supported the conceptualization of the paper, and reviewed and edited the draft.

2b. What impact does MWC input have on young learners' communication abilities (e.g., quantity and quality of spoken output; receptive/productive vocabulary knowledge)?

2.1 Methodology

2.1.1 Eligibility criteria

This review was pre-registered on IDESR (International Database of Education Systematic Reviews) and covered literature on teaching interventions with typically developing monolingual children aged 5 to 12 learning a FL in instructed settings. Detailed inclusion and exclusion criteria are provided in Table 1 below. After unfruitful selective literature searches, we included all languages of publication and grey literature.

Item	Inclusion criterion	Rationale
Bibliographic information	*INCLUDE:* Studies with a full reference or sufficient bibliographic information. *EXCLUDE:* Studies with insufficient bibliographic information.	Without sufficient bibliographic information, retrieval of works is unfeasible.
Date of publication	*INCLUDE:* all. *EXCLUDE:* none.	Preliminary scoping searches showed that research on this topic is scarce, therefore, all available work will be included.
Population	*INCLUDE A:* Children aged 5–12. *EXCLUDE A:* Children under the age of 5 and over the age of 12.	This research is concerned with young learners who are at primary school age. This is the age range where European foreign language curricula have targeted MWC input since the early 2000s.
	INCLUDE B: Learners of any foreign language (including but not limited to English as a Foreign Language) and artificial language, if evidence is available.	This research is concerned with general, non-language-specific foreign language acquisition.
	INCLUDE C: All target languages where the language is a taught, foreign language. *EXCLUDE C:* Minority language learning contexts.	Minority language learners are exposed to the target language outside of learning and teaching settings (e.g., in their homes or home communities). This research, however, is concerned with 'foreign' language learners who learn a new language outside the target language community, in taught, input-limited contexts.
	INCLUDE D: Studies on typically developing foreign language learners. Include studies even if no explicit reference is made to learning ability if reasonable assumption can be made that participants are comprised mainly of typically developing individuals. *EXCLUDE D:* Studies that exclusively target non-typically developing learners	This review seeks to assess the effectiveness of providing early foreign language learners with MWC input in teaching and learning settings as applies to typically developing populations. The findings for non-typically developing individuals may not hold for a larger population, and so these results should not be extrapolated, nor will they be included in this review.
Intervention	*INCLUDE A:* Studies involving interventions in teaching settings (such as classrooms or language clubs) with a focus on MWC input, addressing any (or all) of the four skills (i.e., reading, listening, writing, speaking), and studies involving correlational design with a focus on MWC	This research seeks to assess the effectiveness of providing early foreign language learners with MWC input in teaching and learning settings. Thus, studies where no intervention is reported, where no emphasis is put on MWCs in the input, or – at least – where no correlations between MWC input and other

	input, e.g., studies investigating the correlation between reading comprehension and MWC input. *EXCLUDE A:* Studies where no particular emphasis is put on MWC in the input and where no particular emphasis is put on correlations between the phenomenon of MWCs in the learning trajectory and other linguistic variables.	linguistic/outcome variables are reported are not applicable since they provide no information as to the usefulness of MWC input in early foreign language learning settings.
	INCLUDE B: Studies involving interventional and correlational designs in experimental laboratory settings (e.g., priming or eye-tracking experiments) with a focus on MWC input, addressing any (or all) of the four skills (i.e., reading, listening, writing, speaking). *EXCLUDE B:* Studies where no particular emphasis is put on MWC in the input and where no particular emphasis is put on correlations between the phenomenon of MWCs in the learning trajectory and other linguistic variables.	
	INCLUDE C: Studies that define 'linguistic units larger than one word' with terms other than 'MWC', including but not limited to 'formulaic sequences', 'formulas', 'multi-unit expressions', or 'lexicalised phrases'. *EXCLUDE C:* Do not exclude studies based on their terms used to define 'linguistic units larger than one word'.	MWCs are difficult to define linguistic phenomena, and much scientific work has slightly different views as to what constitutes MWCs. Part of the current research is to map out how different researchers operationalise MWCs in their work and why they do so, i.e., what are the expected learning effects (e.g., productivity, idiomaticity)
Outcomes	*INCLUDE A:* Primary research studies reporting any measure of MWC input effectiveness, including but not limited to language outcomes (e.g., vocabulary uptake, grammatical skills), productivity outcomes (e.g., enhanced communicative abilities), or segmentation-/abstraction-/generalisation-ability outcomes. Include studies reporting either quantitative or qualitative outcomes. *EXCLUDE A:* Systematic reviews and studies that provide narrative evaluation of an educational program but provide no measures of MWC input effectiveness.	A synthesis of empirical findings in this field of literature is impossible without the reporting and evaluation of concrete data.
	INCLUDE B: All types of study design. *EXCLUDE B:* Do not exclude studies based on the research design.	Research on this topic is scarce, and the exclusion of any one study design may provide an even narrower view of the research in this area.
Setting	Include all types of instructed settings, including but not limited to schools, after-school language clubs, psycholinguistic laboratories.	
Publication status	Do not exclude studies based on publication status. Include grey literature.	This paper seeks to offset potential publication bias by including a wider range of research, including grey literature.

Language of publication	Do not exclude studies based on the language of publication.	Limiting to studies written in English may result in a systematic neglect of a certain body of research.

Table 1 Eligibility criteria.

2.1.2 Information sources

The consulted databases covered education, linguistics, psychology, and multidisciplinary sources. Since between us (i.e., the authors) we speak German and French in addition to English, we increased search scopes by adding relevant German and French databases. All databases are shown in Table 2. Individual German and French databases are discussed below.

Discipline	**Database**		
	English	*German*	*French*
Education	ProQuest Education Collection (including ERIC), British Education Index EBSCO	Fachportal Pädagogik	n/a
Linguistics	ProQuest Linguistics Collection (including LLBA)	n/a	n/a
Psychology	PsychInfo	n/a	n/a
Multidisciplinary	Web of Science, Scopus	Humboldt University Berlin (university library catalogue)	SUDOC, Pascal-Francis
Grey literature	ProQuest Dissertations & Theses Global	n/a	n/a

Table 2 List of databases.

2.1.2.1 German

After consultation with German researchers, we conducted pilot searches on the *Fachportal Pädagogik*, *PSYNDEX*, and the Humboldt University zu Berlin's online catalogue. The latter was consulted because, unlike other university databases, it processes Boolean strings. Following pilot searches, we decided on the Humboldt University's catalogue and the *Fachportal Pädagogik*, run by the Leibniz Institute for Research and Information in Education. Being Germany's main Education Index, it processes Boolean strings and runs meta-searches in several databases such as *FIS Bildung Literaturdatenbank* and *BASE*. *PSYNDEX* was excluded because pilot searches mainly yielded results from psychology, such as psychometric tests.

2.1.2.2 French

After consultation with French colleagues, *Pascal-Francis, SUDOC* (Système Universitaire de Documentation), and *Theses.fr* were initially considered suitable. Following pilot searches, *Theses.fr* was subsequently excluded because it did not process Boolean strings.

Instead, we used *Pascal-Francis*, a database in 'exact, human and social sciences' run by the library of the Centre National de la Recherche Scientifique, and *SUDOC*, a meta-catalogue of French university libraries run by the Agence Bibliographique de l'Enseignement Supérieur.

In all languages, we conducted further searches using 'citation-chaining' (or 'snowball searching') which is considered best practice (Boland, Cherry & Dickson, 2017). In addition, forward citation searches were conducted in *Web of Science*.

2.1.3 Search strategy

To find all results available while avoiding unmanageable search outputs, we balanced sensitivity and specificity in search strings (Brunton, Stansfield, Caird & Thomas, 2017). The initial string was created with the support of our department's librarian. The terms describing the concept 'multi-word-construction' in relevant literature were of concern because research provides no uniform conceptualisation. To cast a wide net, we used Christiansen and Arnon's (2017) list of 18 terms as reference for a search field specifying the language input participants received during studies. Other terms from relevant literature were also added (cf. Wray, 2002; Council of Europe, 2001).

As this review focuses on young learners aged five to twelve, we added a 'target participants' search category. Multiple labels were included to represent the diverse terminology of formal teaching contexts (e.g., primary/elementary/junior school). Following pilot scoping searches on *ProQuest Education*, the category 'participants' was assigned to the search frame ABSTRACT. Assigning the category to ALL FIELDS resulted in too many results because terms like 'primary school' appear in too many publications. The search category specifying the target object (i.e., FL) was assigned to TITLE to obtain a manageable number of results. In addition, since MWCs are this review's key characteristic, we assigned NOFT (i.e., all text except full text) to the category specifying 'MWC'. All translated linguistic terms were double-checked in pertinent German/French publications. We used Boolean operators such as truncation in the search strings. Example strings are provided in Table 3.

Database	*Search string*

English	ProQuest Education	ti(efl OR esl OR tefl OR mfl OR tesol OR tesl OR "second language*" OR "foreign language*" OR "artificial language*" OR english* OR fl OR L2 OR SLA) AND noft(multiword* OR multi-word* OR multiunit* OR frame OR frames OR idiom* OR multi-unit* OR prefabricated* OR "pre-fabricated*" OR phras* OR collocation* OR formulaic* OR "fixed expression*" OR "semi-fixed*" OR listeme* OR mwu OR mwe OR formula* OR chunk* OR routine* OR "sentence pattern*") AND ab("primary school*" OR "elementary school*" OR child* OR kids OR "young learner*" OR "grade school*" OR "infant school*" OR "early year*" OR "elementary grade*" OR kindergar?en OR "junior school*") AND ab(productivity OR ability OR vocabulary OR competence* OR "linguistic resource*" OR skill* OR "language knowledge" OR proficiency OR creativity OR "language uptake" OR development)
German	Fachportal Pädagogik	((((Abstract: „Kuenstliche Sprache“ OR DAZ oder DAZ oder "DEUTSCH ALS ZWEITSPRACHE" oder DAF oder "DEUTSCH ALS FREMDSPRACHE" oder L2 oder SLA oder TEFL oder TESOL oder TESL oder ENGLISCH* oder FRANZOESISCH* oder FREMDSPRACH* oder ZWEITSPRACHE oder ZWEITSPRACHERWERB oder ZWEISPRACHIG*) und (Abstract: KONSTRUKTION oder FORMEL* oder WENDUNG oder PATTERN oder SATZFORMEL oder KOLLOKATION* oder IMITATION oder REPRODUKTION oder IMITIEREN oder AUTHENTISCH oder PHRASE oder AUTOMATISIEREN oder SCHEMA oder "VERBALE STEREOTYPE" oder CHUNK)) und (Abstract: SCHUELER* oder GRUNDSCHUL* oder JUNG* oder LERN* oder KIND* oder FRUEH* oder PRIMARSTUFE)) und (Freitext: PRODUKTIVITAET oder "KOMMUNIKATIVE FAEHIGKEIT" oder VOKABEL* oder SPRACHKOMPETENZ oder KOMPETENZ oder "LINGUISTISCHE RESSOURCE" oder KREATIVITAET oder WORTSCHATZ oder LEISTUNG)) und (Datenquelle: "FIS Bildung" oder "Library of Congress" oder "Casalini libri" oder ERIC oder "EBSCOhost ebooks" oder "BBF 1945-1993" oder "Online Contents" oder BASE)
French	Pascal-Francis	ti.*:(LVE OR L2 OR SLA OR TEFL OR TESOL OR TESL OR "non-natifs" OR "langue vivante" OR "langue moderne" OR "langue étrangère" OR "deuxième langue" OR "langue seconde" OR FLS OR FLE OR enseigne* OR anglais* OR allemand* OR langue* OR "français langue seconde") AND (collocation* OR congloméré* OR forme* OR lexème OR locution* OR métaphore* OR "mot composé" OR phrase* OR syntagme* OR "unité phraséologique" OR formul* OR "unité polylexicale" OR "expression polylexicale" OR séquence* OR SF OR phraséologi* OR "expression figée" OR idom*) AND ("École primaire" OR "école maternelle" OR "enseignment primaire" OR "école élémentaire" OR élève* OR apprenant* OR enfant* OR jeune*) AND (vocabulaire* OR productivité OR parole* OR compétence* OR créativité OR appropriation* OR acquisition* OR gramma*)

Table 3 Example search strings.

2.1.4 Selection Process

Following duplicate deletion, the first author (i.e., Johannes Schulz) screened each title and abstract, excluding studies which unambiguously violated one or more inclusion criteria. Being blind to the first author's decisions, the second author independently screened titles and abstracts of a randomly selected 10% sample of all records (87x English; 51x German; 33x French). Afterwards, results were unblinded and the two authors discussed every conflict until a conclusion was reached (conflicts: 6x English, $\kappa = 0.93$; 4x German, $\kappa = 0.90$; 0x French, $\kappa = 1$).

Having retrieved the records marked for full-text screening, the first author screened all full-texts and excluded studies which violated one or more inclusion criteria. Given the high kappa value from the first round of peer-screening, another round was deemed unnecessary. At this stage, difficult inclusion decisions were discussed among all authors.

2.1.5 Data collection process

Prior to the final searches, a data extraction form was created based on Boland et al. (2017) and the Cochrane Good Practice Guide (Cochrane Effective Practice and Organization of Care, 2017; completed forms in supplementary materials of Schulz et al., 2023). The form included all essential PICO items (Participants, Intervention, Comparison, and Outcomes; Petticrew & Roberts, 2006). Another section was added covering the type of MWC operationalisation used in the studies. In one case where data was reported insufficiently, the authors were successfully contacted for additional information.

One included study reported extensive qualitative evidence (~300 pages), rendering the original data extraction form unsuitable. Therefore, following approaches in other syntheses (e.g., Carlsen, Glenton & Pope, 2007), the qualitative report was read repeatedly to identify key concepts. Based on those, an additional qualitative data extraction sheet was created (completed form in supplementary materials of Schulz et al., 2023) which – corresponding to relevant Cochrane guidance – was aimed at synthesising "qualitative evidence within a stand-alone, but complementary, qualitative review to address questions on aspects other than effectiveness" (Noyes & Lewin, 2011: 8). Given the large volume of qualitative data, the outcomes most relevant to the review's research purpose were selected as recommended by Noyes and Lewin (2011) and Page et al.'s (2021) latest PRISMA guidance. This process resulted in an analysis approach extracting data in the following categories, closely following the original author's qualitative data analysis:

(a) Which slot-filling/insertion word(s) was/were used?

(b) Does the transcript example represent paradigmatic or syntagmatic variations of formulaic sequences?

(c) Are segmentations of formulaic sequences observed?

(d) Did the student self-actively recombine formulaic sequences?

(e) Could the student realise his/her personal communicative goals?

This extraction approach focused on any MWCs and slot-fillers taught during the intervention, tracing them back to the individual transcription examples. The extraction ignored

data irrelevant to the review's purpose, such as descriptions of students' degree of extroversion.

Following the first author's data extraction, the second author independently extracted data for one of the two included studies, resolving discrepancies through discussion. Since the other included work was an >800-page mixed-methods study written in German, the second author was only able to independently extract the quantitative data. An independent qualitative data extraction accounting for more than 300 pages of the study was deemed unfeasible because none of this review's other authors is a native German speaker. To establish the most objective and coherent data extraction process possible, the first and second author discussed the first author's approach to qualitative data extraction in detail. This process included the second author selecting sections of qualitative data at random and checking whether information was adequately represented on the data extraction form. In all cases, the data was documented comprehensibly and thoroughly.

2.1.6 Risk of bias

Since research on MWCs in primary school FL teaching contexts is scarce, we included all available research designs in the selection process, following Slavin's (1986) best-evidence synthesis which does not discriminate against any research designs during study selection.

The Mixed Methods Appraisal Tool (MMAT; Hong et al., n.d.; Pluye, Gagnon, Griffiths & Johnson-Lafleur, 2009) allowing researchers to appraise research quality in five methodology categories (qualitative, randomised controlled trials, non-randomised, quantitative descriptive, mixed methods) was employed as critical appraisal tool. This instrument has been updated and revised repeatedly (Hong, et al., 2019; Pace et al., 2012) and used in similar systematic reviews including quantitative and qualitative designs (Richter, 2021; Willis, Neil, Mellick & Wasley, 2019).

Usually, each study receives quality scores of 0 to 5 per assessment category which amounts to a global quality score, facilitating cross-study comparisons. However, the presentation of global quality scores has been discouraged in the past because metric scores cannot represent a study's problematic elements (Hong et al., n.d.). Since the current review only includes two studies, the individual risk of bias assessments for each study are reported instead of global ratings.

The second author independently completed the MMAT form for the included study published in English. Interrater reliability was high ($\kappa = 0.86$), and the sole discrepancy emerged

from missing information that the original author had only provided via e-mail. Since the second included study was published in German and consisted of >800 pages it was difficult for the co-authors (non-native speakers of German) to complete the MMAT independently. Therefore, the first and second author (German FL speaker) discussed the completed MMAT in detail, the first author providing evidence for each rating in the original report (including translations). Agreement was reached on all ratings.

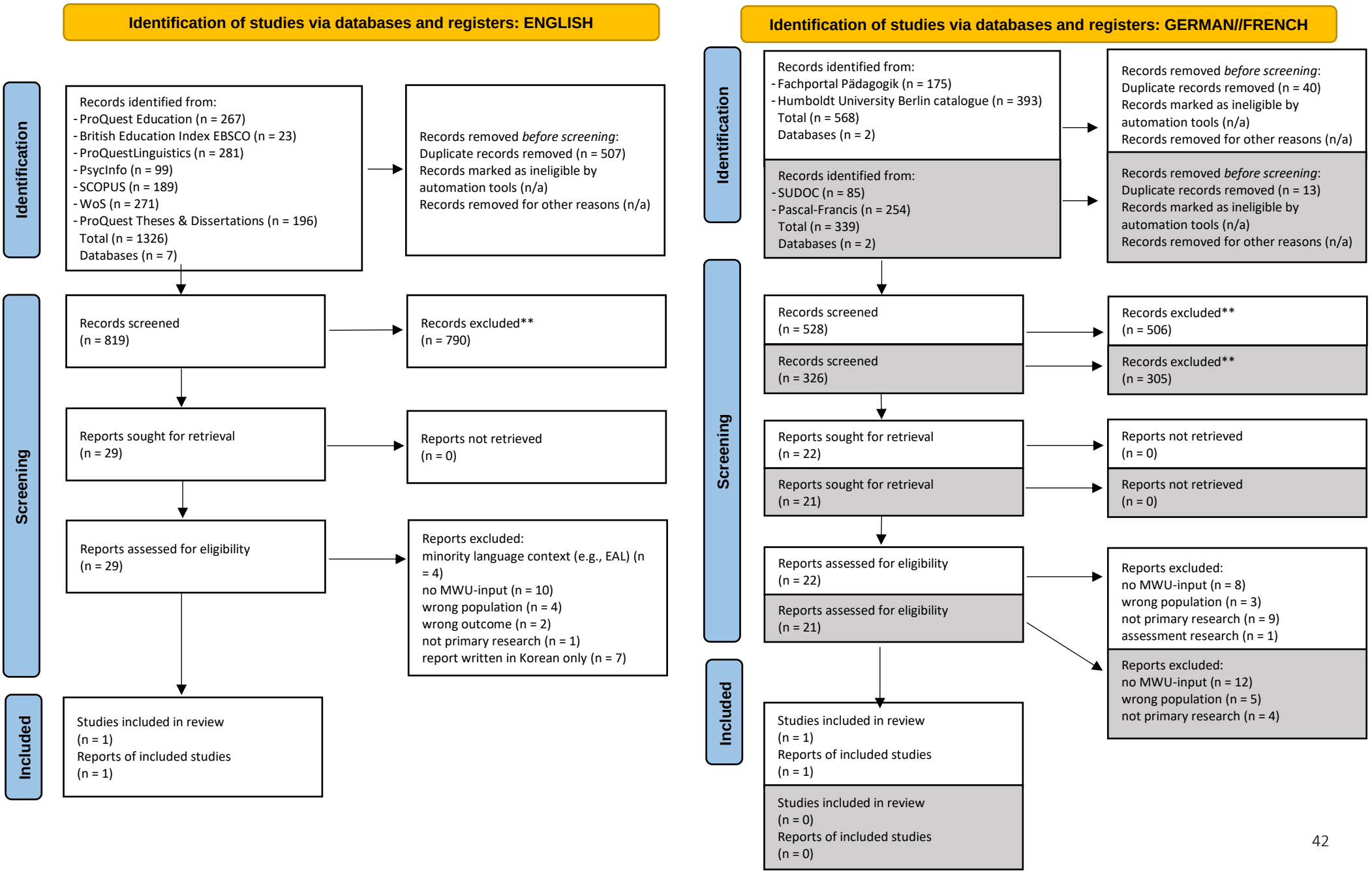
2.1.7 Synthesis methods

Since the number of included studies was small (n = 2) and they differed considerably regarding their methodology (quantitative vs. mixed-methods) and design (i.e., intervention vs. corpus analyses and participatory action research), there was not enough comparable quantitative data for statistical meta-analyses. Instead, following Thomas et al. (2004), we triangulated qualitative and quantitative data in a narrative synthesis of study quality and findings.

2.2 Results

Results of search and selection processes are presented in Figure 1. Of the 29 English reports selected for full-text screening, seven were published in Korean (Chae & Kim, 2019; Jeon & Kim, 2018; Jung & Shin, 2013; Kim & Kang, 2005; Kim & Lee, 2009; Lee & Jeong, 2010; Lim & Lee, 2012). Requests to each paper's authors for an English version were unfruitful. Numerous translation attempts showed that the works' reference lists included additional potentially relevant works. Eventually, all seven papers were excluded because no member of the authors' research groups was a Korean-speaking linguistics expert, and translations were considered excessively labour-intensive and unreliable for they could not be evaluated by proficient Korean speakers.

Figure 1 Selection Process. Left side: English. Right side: German/French.



2.2.1 Included studies

General study characteristics of the two included studies are provided in Table 4, followed by narrative summaries of the design, context, and findings of each included study.

Study	Balcı and Çakır (2012)	Kostka (2020)
Citation	Balcı, Ö. & Çakır, A. (2012). Teaching vocabulary through collocations in EFL Classes: The case of Turkey. *International Journal of Research studies in Language Learning* 1(1), 21–32.	Kostka, N. (2020). *Produktives Sprechen im Englischunterricht der Grundschule: Eine empirische Studie zur Bedeutung formelhafter Sequenzen.* Doctoral dissertation. Giessen: University Library Publications.
Publication Status	Journal Article	Published PhD thesis
Study Design	intervention	observational (*Prospective study*; cf. Petticrew & Roberts, 2006: 134)
Data type	quantitative	quantitative/qualitative
Country	Turkey	Germany
Sample Size	n = 59	Quantitative: n = 18 Qualitative: n = 7
Recruitment	State primary school in Konya	Three state primary schools near Frankfurt am Main.
Study duration	6 weeks	1 school year (October – July)
Non-MWC control group?	Yes	No
General outcomes	Vocabulary uptake, vocabulary retention	Productive speaking skills (e.g., paradigmatic variations of patterns)
Specific outcome measures	Non-standardised vocabulary tests (multiple choice and gap-filling)	Quantitative: token frequency of formulaic sequences (patterns/routines) in speech corpus; token/type frequency of patterns in speech corpus; token/type frequency of individual patterns in speech corpus Qualitative: communicative-functional analysis of transcription examples from speech corpus

Table 4 General characteristics of included studies.

2.2.1.1 Balcı and Çakır (2012)

Balcı and Çakır investigated whether teaching new vocabulary through collocations results in better learning than teaching through 'classical' techniques (original report's terminology) by conducting a teaching intervention study with two 7th grade classes at a Turkish state primary school. The participants' age ranged from 12 to 14 years. We included the study because it is conducted in primary school and the age range is within the scope of the review protocol's age range (5-12). Both the control and intervention group received four hours of English classes per week over a six-week period. Whilst the control group learned

vocabulary through techniques including direct translation and synonym/antonym tasks, the intervention group was presented with two reading passages per week (two hours per passage) including the same vocabulary items implemented in collocation contexts. After students had read the passages, the included collocations were discussed in class and presented in collocation networks on the blackboard. In both groups, the authors conducted a general vocabulary-based multiple-choice pre- and post-test on English proficiency, plus gap-filling vocabulary tests on the specific vocabulary used in the intervention at the end of each week (six in total). One week after the study, they conducted an additional retention test with both groups containing some of the gap-filling exercises from the previous weekly vocabulary tests. All tests were non-standardised.

The experimental group outscored the control in the last weekly test ($p < 0.01$, Cohen's $d = 0.87$) and in both the post-test and retention test (post-test: $p < 0.01$, Cohen's $d = 1.16$; retention test: $p < 0.01$; Cohen's $d = 0.78$). Since there were no such significant differences between the groups at pre-test, or in the first five weekly tests, the authors conclude that post-training differences are due to the intervention. Unfortunately, this conclusion is too strong, since a significance test on pre-test difference cannot provide evidence for a lack of differences between the groups at that time point (a non-significant p-value does not show whether there is evidence for the null-hypothesis, though it is routinely misinterpreted in this way – cf. Dienes, 2008, 2014; Lakens 2017). Similarly, where the authors do look at pre- to post-test differences (uptake over time) they look at this for each group separately, finding within-group (week 1 vs. week 6) significant improvement in test scores for the experimental group ($p < 0.01$, Cohen's d = not reported) but not the control group. Again, however, inferring from this that gains in the experimental group are larger than those in the control is to overinterpret the null result in the control condition. Without a direct comparison of the gains in the two groups (e.g., using a time-by-group interaction) it is difficult to draw strong conclusions from this data. Thus, though the patterns are certainly suggestive, the authors' conclusion that the students benefitted from learning new vocabulary through collocations, and that this effect specifically showed up after five weeks of training, must be treated with some caution.

2.2.1.2 Kostka (2020)

Kostka (2020) investigated the role of MWC input (which the author refers to as *formulaic sequences*) in German primary students' productive English skills. Supported by three

participating primary school English teachers, Kostka designed and initiated a planned, controlled, gradual reconfiguration (original report's terminology) of the English classes of three Year 3 classes (8- to 9-year-olds) in schools around Frankfurt am Main, Germany, over a period of one school year. Her work was not intended to conform to classic intervention designs as no control group was involved. The teaching reconfiguration consisted of increased teaching foci on a set of formulaic sequences subclassified into 'patterns' (with open slots) such as *My favourite _ is _* and 'routines' (fixed constructions) such as *What's your name*? Considering previous MWC research, Kostka anticipated that this input focus would improve students' productive skills over time. The reconfiguration was divided into three teaching phases and an end-of-year assessment phase. During the school year, Kostka collected students' oral speech data from video recordings of classroom interactions and of targeted end-of-year assessment situations. While Kostka transcribed and analysed quantitatively the speech data of 18 students, she analysed qualitatively the speech data of seven students. In both cases, the selected students covered a range from weak to strong students (see section 5.2.1 for a discussion of 'weak' versus 'strong' students).

Overall, Kosta reports that taught formulaic sequences accounted for about 50% of all articulated words. Based on token frequencies, her quantitative corpus analyses revealed that irrespective of their proficiency level, students used *patterns* more frequently than *routines* in all three teaching phases, the assessment phase, and the entire year's corpus. The overall use of *formulaic sequences* decreased significantly from teaching phase one to the assessment phase ($p < .0001$). According to Kostka, this decrease supports the assumption of a parallel running process of students' increasing segmentation and analysis abilities. The conclusion that students' productive speaking skills improved over the year is supported by that fact that Giraud's Index (a measure of lexical diversity) of the 'content areas' that students drew from to fill slots in *patterns* increased steadily over the year. That is, the quantitative data suggest that students gradually inserted words from an increasing variety of contexts into empty slots of *patterns*.

In her qualitative analyses, Kostka conducted communicative-functional analyses of transcribed speech examples of seven students' productive-dialogic speaking over the course of all three teaching phases and the assessment phase. In her analyses, the author considered individual student characteristics (data collected during classroom observations) both on a

personal level (e.g., extroversion) and a subject-related level (e.g., language awareness). Overall, the qualitative analyses revealed that all seven students could self-actively implement formulaic sequences into spontaneous dialogues to reach their personal communication goals. To varying degrees, both weak and strong students could apply learned lexical knowledge to new linguistic contexts. At the beginning of the year, weaker students relied on a smaller set of formulaic sequences while stronger students integrated a comparatively large spectrum of routines and patterns into their dialogues. Paradigmatic variations of patterns played a central role in the gradually increasing productive use of formulaic sequences throughout the school year. This qualitative observation corresponds to the quantitative findings. However, weaker students relied on patterns longer than strong students, who started showing an increasing ability to segment and analyse patterns and reconfigure individual segments earlier. Syntagmatic variations also occurred more often among strong students than weak students. Regarding the stronger students, one exercise format in the assessment phase required them to dismantle the learned routines and patterns, initiating linguistic transfers of structural knowledge to new contexts, thereby increasing their segmentation abilities.

2.2.2 Risk of bias assessments for included studies

Detailed risk of bias assessments are provided in the supplementary materials of Schulz et al., 2023). We consider Kostka's (2020) overall level of risk of bias to be low, confounded only by the non-consideration of students' language biographies. Balcı and Çakır's (2012) overall level of risk of bias we consider high, for example because all employed measurements were non-standardised.

2.2.3 Summary statistics

Summary statistics for Balcı and Çakır (2012) are provided in the supplementary materials of Schulz et al. (2023). Kostka (2020) used no classic intervention design (i.e., no control group), therefore, summary statistics for each group and effect estimates cannot be reported. Descriptive statistics relevant to this review's research questions and corresponding within-group differences of mean difference between the individual teaching/assessment phases are provided in the supplementary materials of Schulz et al. (2023). Summaries of Kostka's (2020) qualitative findings are provided in the supplementary materials of Schulz et al. (2023) as well.

2.2.4 Results of syntheses

RQ1: What is the extent of original research investigating the role of MWCs in early FL learning and teaching contexts?

The empirical research output on MWC's role in early instructed FL learning is scarce. While many works that were excluded in the screening phase provided theoretical accounts of MWC's important role in language education, and some works reported empirical work from other age groups, only two works were retrieved that report relevant empirical data from primary school contexts. Seven works were excluded because they were only available in Korean. They could not be excluded based on their English abstracts and therefore may well have been eligible for inclusion in the current review.

RQ2: What is the extent of the evidence on the effectiveness of learning from/teaching MWCs in early FL contexts?

2a. What impact does MWC input have on young learners' grammatical and vocabulary skills and knowledge in the target language?

Since only two studies were identified in our review, the extent to which these studies can address our research questions is limited.

Only Balcı and Çakır (2012) provide test data from vocabulary tests. However, since the authors do not report time-by-group interactions it is difficult to discern whether there was indeed a between-group difference in pre- to post gains. We cannot say whether the intervention influenced vocabulary uptake, although this is consistent with the descriptive statistics.

Direct measures of grammatical knowledge are not reported in the included studies. Kostka (2020) provides indirect accounts of students' grammatical skills as she describes a gradual development of rudimentary top-down processes by some strong students (e.g., student S1a). Which exact grammatical patterns those students have acquired remains unclear.

2b. What impact does MWC input have on young learners' communication abilities (e.g., quantity and quality of spoken output; receptive/productive vocabulary knowledge)?

Balcı and Çakır (2012) do not report data on their learners' communication abilities as their vocabulary tests measure single-word knowledge. Kostka (2020) finds MWC use has positive effects on students' communication abilities. Given the small number of participants

and the lack of control group, Kostka (2020) only tentatively ascribes the students' linguistic development to the implemented teaching changes. Nonetheless, she takes into consideration the possibility of a relationship between the teaching focus on formulaic sequences and the findings from the corpus data. In her study, MWC input gave students relatively strong communicative agency enabling them to realise personal communication goals in their first year of FL study. The reported quantitative and qualitative findings complement one another in this respect. While the corpus data showed that the overall number of formulaic sequences gradually decreased over time, Giraud's Index of 'content areas' and the qualitative analyses indicate that the decrease was due to an uptake of segmentation and analysis abilities. Based on her analyses, Kostka concludes that the variety of communication goals the students managed to realise throughout the year via (minimal) formal-linguistic variations of learned formulaic sequences indicates that productive speaking can already be initiated in the first year of primary school FL classes. Importantly, Kostka reports differences between weak and strong students regarding the impact of MWC input on communication abilities. While weaker students rely longer on formulaic sequences, stronger students start developing segmentation abilities and applying abstracted pattern-knowledge productively earlier.

With regards to the general result of this systematic review, note that Balcı and Çakır (2012) received low quality ratings, weakening the reliability of their results, and Kostka (2020) did not include a control group in her study. Rightly, she emphasises that her findings cannot unambiguously be traced back to the implemented teaching changes. Although the reported findings are promising, considering the paucity of available evidence, this systematic review cannot report trustworthy evidence of the effectiveness of teaching MWCs to young FL learners.

2.3 Discussion

2.3.1 Available research

The review's results demonstrate that research into the effectiveness of MWC input in early FL teaching is scarce and available evidence is inconclusive. These results are striking since MWCs are considered vital for FL development in linguistics, psycholinguistics, cognitive linguistics, and language pedagogy (Siyanova-Chanturia, 2017). From a pedagogical perspective, the lack of research is startling since FL education is part of many primary school curricula worldwide, and in Europe, MWCs are already an integral part of said curricula (see section 1.1). Scientific evidence supporting the effectiveness of MWC input in early FL teaching is crucial to (a) legitimate and reinforce the importance of MWCs in curricula, teaching materials and teacher education, and (b) provide important groundwork facilitating the implementation of targeted changes in curricula and teaching to further improve learning outcomes.

The review may be biased due to the inaccessibility of seven Korean papers we excluded because English versions were unavailable. The inaccessibility limits the review because the methodologies and results reported in the English abstracts were promising, indicating that MWC input improved students' FL abilities (Lee & Jeong, 2010), that control groups were used (Lim & Lee, 2012), and that interventions positively impacted students' language learning motivation and self-esteem (Jung & Shin, 2013). Unfortunately, with English being the lingua franca of scientific publishing, these papers are difficult to retrieve, rendering their contents mostly inaccessible. However, even if each of the seven Korean language studies had been included, that would still amount to only nine available pieces of original research addressing what we believe to be an important theoretical and practical question. Therefore, despite this potential bias our conclusion that there is a vanishingly small amount of research in this area still holds.

Importantly, as stressed by Gray (2021), identifying only two studies in this review is nonetheless a valuable finding for current FL research because we can now be confident that there is little relevant research available; researching whether MWCs in curricula and teaching have beneficial effects on young learners' FL development is a much-needed area of future inquiry.

In addition to the scarcity of research is the issue of the quality of available research. On the one hand, Kostka (2020) reports meticulous work, but her findings cannot be tied to the teaching because a control group was lacking. On the other hand, Balcı and Çakır's (2012) study includes a control, but there are other limitations such as non-standardised testing. Considering the limited and mixed evidence-base, the following discussion of the second research question should therefore be regarded as tentative.

2.3.2 Effectiveness of MWC input

Balcı and Çakır's (2012) results seem to suggest that the teaching intervention was successful, yet their statistical analyses were inconclusive (see section 0). Nonetheless, their data evoke consideration of the appropriateness of different measurements in MWC research contexts. On the one hand, Balcı and Çakır (2012) rightly used 'traditional' vocabulary measurements such as single-word gap-filling exercises corresponding to their research aim, namely, to investigate single-word vocabulary uptake. On the other hand, if the role of MWCs in language development in future work is considered in constructionist learning frameworks, then researchers should introduce different types of measurement. That is, future research aiming at communication-related dependent variables such as productivity would benefit from the use of measurements that can capture MWC knowledge. In fact, researchers have proposed ways of measuring MWC knowledge, such as using mutual information scores (MI) in speech output (Polio & Yoon, 2020). Others have introduced tests such as the multi-word phrase test (MPT) measuring children's multi-word vocabulary knowledge based on *verb + object* phrase knowledge (Smith & Murphy, 2015). Such tasks are reminiscent of outcome measurements used in studies like Wonnacott et al. (2012), where tasks such as 'act out' or 'production' tasks tapped into children's ability to infer linguistic information from input and apply it to novel contexts. Since the purpose of early MWC input is (a) the provision of early communicative agency and (b) the catalysis of abstraction and generalisation processes, future research must avoid unreliable measurements which might misrepresent young FL learners' language knowledge and blur our understanding of the effectiveness of MWC input. If the aim of teaching is to introduce pre-made MWCs to initiate immediate early communication, tasks focusing on imitation abilities may be more suitable. In contrast, if the goal is to develop students' abstract understanding of linguistic structures through the teaching of flexible MWS, tasks that encourage generalization abilities may be more appropriate.

In general, instead of specifically asking about learners' vocabulary and grammar gains, perhaps more relevant questions concerning the impact of MWC input are about learners' communication abilities as in Kostka (2020). Such inquiries correspond to main goals of current primary school FL teaching to make children active language users. Beginning FL primary learners rely heavily on MWCs in the first stages of their development as such linguistic units are their first and only means for successful FL communication (Diehr & Polte, 2009). Having immediate linguistic agency to realise personal communication goals increases (or at least maintains) children's generally high language-learning motivation (Hempel, 2016; Lenzing & Roos, 2012). And while MWCs are important linguistic agents for weak learners to communicate in the first place, they are catalysts for stronger learners to abstract patterns and advance to free productions (Kahl & Knebler, 1996; Sambanis, 2007; Kostka, 2020). Congruously, Kostka (2020) did not simply administer vocabulary tests but analysed her qualitative data relative to the students' ability to realise personal communicative goals. Supported by her quantitative findings, the qualitative work confirms what other research (Myles et al., 1999; Kahl & Knebler, 1996; Sambanis, 2007) has suggested and what curricula (e.g., KM-BW, 2016) have expected, namely that students rely heavily on MWCs to communicate successfully. In addition, her data show that weaker learners relied longer on formulaic sequences than stronger learners, who started dismantling phrases and using abstracted structural knowledge productively earlier. Concerning communication abilities, both proficiency levels benefitted from the use of MWCs enabling them to realise basic communicative goals or to fill slots in MWCs and rearrange and combine patterns to realise advanced goals. Therefore, following Kostka's approach, measurements of the effectiveness of MWC input should be regarded against the background of children's linguistic generativity instead of solely concentrating on their vocabulary and grammar uptake. Unfortunately, the lack of a control group in Kostka (2020) makes it impossible to discern to what extent the uptake in students' communication ability was a direct result of the focus on MWCs in the teaching input.

Kostka's (2020) data show that cognitive processes such as rule abstraction are slow and rely on large amounts of input and on context conditions, such as exercise type, which is unsurprising (cf. Eskildsen, 2009; Nattinger & DeCarrico, 1992; Wulff, 2018). In this context, researchers have repeatedly warned about establishing realistic FL outcomes among primary school learners (Jäger, 2012). The time spent in FL classrooms varies widely across contexts

and for many children is less than one hour per week. Language teaching often plays a subordinate role in schools' curricula and in some countries, such as Germany, FL teaching only starts in Year 3. Therefore, some researchers argue that young learners will only ever be able to achieve rudimentary communication skills (Jäger, 2012). Although results from longitudinal large-scale FL evaluations in German primary schools have indicated that learners benefit from MWCs and gradually start breaking them down, realistically, communication skills by the end of Year 4 are still rudimental among most students (Hempel, 2016). For example, Engel et al. (2009) and Pienemann, Keßler and Roos (2006) report that despite high motivation levels among learners, they were often not able to realise their communicative goals fully and by the end of Year 4 many students, especially weaker ones, were stuck on a reproductive and imitative plateau. This corresponds to Kostka's (2020) finding that only stronger students were able to segment patterns and produce speech freely. Therefore, considering limited teaching time, it appears sensible to support students' learning from MWC input by targeted manipulations of that input, for example by manipulating its variability (see section 1.6), to try and 'get the most out of' the limited input.

2.4 Transition to experimental work

2.4.1 The impact of targeted MWC input on young learners' Foreign Language development

Kostka (2020) presents tentative findings substantiating what constructionist theory and previous non-classroom FL research has suggested. While the lack of a control group in her study makes it impossible to tie her findings to the teaching intervention, her data from an authentic teaching setting is valuable in that it underlines the importance of MWCs in young learners' FL development, both from the perspective of imitation and of productivity. To exemplify this, in the following, I discuss a concrete example demonstrating an individual student's gradual development from imitation to tentative independent production.[11] At the beginning of the school year, the student utters: *My favourite pets are horse and a cat*. The pattern *My favourite _ are _* had been part of the taught input, as were the inserted slot fillers of the content area 'pets'. By the end of the year, without being prompted to do so, the same students utters: *My favourite star – I have don't no a favourite star*. Again, the pattern *My favourite star is _* as well as the routines *Yes, I have* and *No, I don't* had been part of the taught input. We see that the student successfully managed to extract the nominal phrase *my/a favourite star* from the initially taught larger pattern *My favourite star is _* and subsequently attempted to recombine various separate elements in a correct, apparently rule- or structure-governed, fashion. The ungrammaticality of the sentence aside, it demonstrates that the student has developed metalinguistic awareness beyond imitation: (a) The student detached the confirmatory particle *Yes* from the routine *Yes, I have*, making the possessive phrase *I have* available for further use. (b) The student negated the extracted structure *I have* with the auxiliary negation structure *don't* as well as with the negation particle *No*, both negations being extracted from another taught structure (*No, I don't*). (c) The student realized (perhaps without conscious awareness) that the extracted nominal phrase *my/a favourite star* can precede the self-constructed negated possessive structure *I have don't know*, creating the global negation construction *S(ubject) + $_{Neg}$Aux + have + ($_{neg}$Part) + O(bject)*. Importantly, by making productive use of extracted structural information, the student considerably increased her/his

[11] Example taken from student *S1a* (Kostka, 2020: 381-388 and 403-406). English translation and summary of transcript sections can be found in the supplementary materials of Schulz et al. (2023).

communicative agency; s/he was able to break out of the expected communicative frame (i.e., the students *having* a favourite star) and to adequately react to her/his personal context requirements (in fact, *not* having a favourite star) in an independent and non-prescribed manner. This meaning could not have been conveyed solely through the means of the taught input.

The example tentatively demonstrates a learner's journey along the constructional spectrum from imitation towards increasingly abstract linguistic knowledge and indicates that this process parallels her/his increasing communicative scope (see section 1.3). Examples from instructed settings like this combined with Kostka's (2020) general findings support ubiquitous claims in curricula and research regarding MWCs' potential for early communicative agency development. The data tentatively suggest that MWC input supports learners' inductive learning of structural rules which leads to increasingly sophisticated paradigmatic variations, gradually increasing students' communicative agency. Yet, overall, the systematic review indicates that the evidence about the impact of targeted MWC input in primary school FL settings is limited.

2.4.2 Increasing productivity - From imitation to independent production

The systematic review could not report trustworthy evidence regarding the impact of MWC input on students' FL development in instructed primary school FL settings. Unsurprisingly, the few reported classroom studies focused mainly on linguistic sequences on the relatively lexicalized side of the constructional spectrum since such fixed structures like routines are dominant in curricula. Yet, while such lexicalized MWCs appear to boost early communicative agency through imitation, they often lack the potential to serve as catalysts for the development of abstract structural knowledge due to their inherent fixedness (i.e., they lack productive slots in the instructed setting). As a result, we observe that students are often stuck on 'imitative plateaus', although curricula prioritize communicative agency and support constructionist approaches to language learning (see section 1.3). Essentially, learners struggle to bridge the gap between the lexical and grammatical system. While the former is essential to initiate learning and communication, the latter is crucial to build a generative linguistic repertoire, fostering flexible communicative agency in the long run.

Considering the scientific 'blank slate' revealed by the systematic review (and by traditional reviews of the literature), researchers are now tasked with making a well-reasoned decision 'where to begin their investigation'. From a pedagogical perspective, policymakers,

practitioners and, crucially, young learners have a strong interest in increasing communicative agency. To address this goal, the following intervention experiments depart from the context of the systematic review in two major ways:

First, having found almost no research in this area, we planned to conduct a teaching intervention study focusing on the relatively context-dependent, routinized, and formulaic multi-word-constructions (MWCs) already ubiquitous in primary school classrooms. However, while designing an experimental-control group study, we realized that a control group with 'no such MWC input' would result in almost no teaching or only single word teaching, as most conversations in beginner FL primary classrooms are highly patterned and formulaic. From pedagogical, theoretical, and ethical perspectives, this approach was unworkable, especially considering the research aim to increase the communicative agency of primary school FL students. Thus, we decided to move from an input versus no-input design to one where the input itself is manipulated. After reviewing the literature, it seemed promising to investigate the impact of manipulated input variability on students' ability to infer underlying linguistic structures and extend those to novel contexts. Given that the variability effect is generally robust, this inquiry aligned with our goal of enhancing students' communicative agency.

Second, the systematic review focused on the kinds of relatively context-dependent and formulaic MWCs already prevalent in primary school classrooms, while the experiments reported in this thesis examine two more abstract and compositional structures, as will become clear later. This decision was made for two reasons: (a) To the best of the author's knowledge, these experiments are the first in this research area to be conducted in a classroom, despite the systematic review's search string limitations. Therefore, we adopted structures used in previous controlled experiments to model our work on earlier research. (b) Gaining an in-depth understanding of non-formulaic and context-independent structures, such as verb-argument-constructions used in studies like Wonnacott et al. (2012), was considered more beneficial to students' long-term communicative agency than gaining knowledge of relatively formulaic structures, despite their fillable slots.

In sum, considering both the pedagogical and the research perspectives and to respond to calls for increasing communicative agency, my experimental inquiry shifts focus from the relatively lexically oriented and formulaic outcomes of the systematic review to more abstract and compositional MWCs, such as *verb-argument-constructions* and *non-adjacent dependencies*. On the one hand, this decision was primarily based on research focus and design

considerations, rather than theoretical reasons. On the other hand, more abstract types of MWCs -due to their paradigmatic flexibility- are perhaps most likely to support students in building a generative linguistic repertoire and their structure can be heavily manipulated in the teaching input (see section 1.6.3). In addition, the learning of these two types of MWCs has been successfully manipulated in previous works in more controlled settings (Gómez, 2002; Wonnacott et al., 2012) through a manipulation of their input variability.

3 Experiment 1

The current thesis adopts a constructionist approach to language learning and representation, where 'learning' is considered a function of input characteristics, context information, and non-language specific cognitive mechanisms (Wulff & Ellis, 2015; Tomasello, 2009). Embedded in this theoretical framework, the central hypothesis of this thesis is that input variability can impact young FL learners' generalization of structural properties of linguistic constructions in instructed settings.

As discussed in section 2.3.2, upon finishing FL learning at primary school, many students lack productive vocabulary knowledge, specifically regarding structures like verb-argument-constructions (e.g., [*Verb*] *about* [*Noun*]), indispensable for increasing communicative agency, as required by curricula. Across cognitive domains, increased initial input variability (the variation in our experience with different exemplars, e.g., [*talk/think/rant/wonder*] *about* [*god/bicycles/dogs*]) can improve generalization (i.e., [*Verb*] *about* [*Noun*]), and enhance learning. In controlled experiments, increased input variability proved beneficial for children's generalization of linguistic information to novel contexts (e.g., Wonnacott et al., 2012). Such findings inspired the current investigation into examining whether these benefits of input variability can be extended to real classrooms.

In experiment 1, the effect of increased input variability on students' ability to infer underlying structural patterns and generalize them to novel contexts was investigated by targeting a verb-argument-construction (VAC) in a classroom-based teaching intervention experiment. VACs were used since they are ubiquitous structures in our language use and have high productive potential. In the past, a positive variability effect on children's learning of VACs' underlying structures and their ability to extend those inferred structures to novel contexts has been demonstrated in controlled settings (Wonnacott et al., 2012). In the current classroom experiment, students were split into two groups and received in-class teaching of the target VAC over a period of two weeks either with four verbs in the verb slot (high variability; HV) or with only one verb in the verb slot (low variability; LV). Specifically, it was investigated whether increased variability in the target VAC's verb-slot would be beneficial for the HV students' ability to infer the underlying structural pattern of the VAC and generalize this abstract knowledge to novel contexts with unknown verbs, as compared to the LV students. The

experiment took place in the context of British primary students learning German as a Foreign Language in an instructed setting. The overarching research question for experiment 1 was:

> *Does input variability in a German verb-argument-construction's verb slot impact primary school FL students' ability to generalize the construction to unattested verbs?*

Based on the research results reported in the psycholinguistic literature (see section 1.6.3) as well as the wider educational literature reporting on the positive effect of increased input variability on learning across domains (Raviv et al., 2022), a positive effect of increased input variability on the learning of underlying structural patterns was expected, even in the 'noisy' environment of a classroom.

In the remainder of this chapter, the methodology of experiment 1 is presented, including an overview of how the participating school was recruited and how approval by the school and consent by the students and their parents was obtained. Then, various pre-testing measures are described, including details regarding the pre-testing materials and the pre-testing procedures in school. Moving on to the main experiment, the process of generating the teaching input and details regarding the outcome measurements are presented. After that, the specific teaching arrangement during the intervention is presented. In the end, an overview of the final outcome measurement procedure is provided, before I move on to reporting the results.

3.1 Methodology

3.1.1 Research Design

This classroom intervention experiment was conducted in a British primary school with two intact classes. The participants were Year 2 students learning German in their second year of study. There was a low variability and a high variability condition. Thus, there were two manipulated variables: input group and time. The two classes were allocated randomly at group level to either condition.[12] It is important to acknowledge that this experimental design came with limitations caused by its quasi-experimental character and its sample size which is effectively N = 2 (i.e., two classes) considering the clustering of the data in each class. Therefore, the experiment should be regarded as 'proof of concept' trial providing initial insights into the potential effectiveness of the proposed treatment. It is hoped that in the future, larger randomized-controlled trials could further explore the findings of the current work with higher statistical power.

Experiment 1 targeted VACs like Wonnacott et al. (2012). More details regarding the VAC materials are provided below. The contact time with all participants was approximately fifteen minutes daily for two weeks which exceeds that in similar intervention studies with comparable age-levels (e.g., Hopp & Thoma, 2021).

In addition to the low sample size discussed above, some further limitations of this experiment's methodology are presented in section 5.6.

3.1.2 Participants

The current experiment followed a quasi-experimental design as the participating school and students were selected using convenience sampling. The detailed sampling process is described in appendix 7.1.

The participating school was an independent pre-preparatory school in the South-East of England, UK, and the participating classes were two Year 2 classes who have had German instruction since Year 1 for one lesson each day (30 minutes) with a German native-speaking teacher. The decision to work with the Year 2 students was made based on experiences from multiple visits to the school prior to the study, where I audited several German classes and received a first impression of the students' learning habitus and skills. The one-year prior

[12] Across the two empirical studies reported in this thesis, the random allocation of classes to the conditions was balanced, so each class was in the high variability condition once across experiments.

exposure to German at the time of intervention (German teaching started in Year 1) was considered a useful and necessary basis for the introduction of more complex linguistic structures as planned in the current study. The students in Year 2 were between 6 to 7 years of age.

Both classes had a size of 20 students which resulted in a total of 40 potential participants. Although parents and children were not asked directly, both the Headmistress and the Head of Pre-Prep Languages confirmed that they had no knowledge of any cognitive impairments or learning disabilities among their Year 2 students.

During the numerous testing periods on different days, it was unavoidable that some individual children were unavailable (e.g., sickness), although every effort was made to try and test each child and collect as much data as possible. In the end, participant numbers were 19 and 20 across both classes for the outcome measures of experiment 1. Numbers of available participants in pre-tests and during outcome measures are presented in Table 5. Of all the children who were absent during testing sessions including both pre-tests and outcome measures, only two missed numerous test sessions due to longer periods of illness. One child missed the WASI *and* the Language Magician.[13] They were from class 2.

group	N	PVST	WASI	Language Magician	Exp 1 (VACs)
class 1	20	20	19	19	19
class 2	20	19	16	18	20

Table 5 Numbers of available participants across pre-tests and outcome measure during experiment 1.

After an initial visit to the school in early autumn 2022 which included a meeting with the headmistress of the pre-preparatory school, the school was officially invited to participate in the study via e-mail in October 2022. The e-mail contained a letter to the principal of the entire school and the headmistress of the pre-preparatory school, information sheets for parents and children, as well as opt-out forms for parents and opt-in forms for children. Both the principal and the headmistress accepted the invitation and approved the forms. All relevant documents, including the letter to the principal, information sheets, and opt-out and opt-in forms, are attached in appendix 7.2.

[13] Another child missed the WASI, the Language Magician, *and* the outcome testing in experiment 2, see section 4.1.1.

3.1.3 Ethics approval and consent

The study was approved by the Central University Research Ethics Committee (CUREC) of the University of Oxford (Ethics Approval Reference: [CIA – 22TT - 135]). Once the principal and the headmistress had approved of the study, of the relevant parent and child information sheets, and of the parent opt-out and children opt-in forms, the school disseminated the parent information sheets and the opt-out forms to the Year 2 students' parents in early 2023, well before the start of the study to allow parents enough time to ask questions and potentially opt-out. No parent chose to opt-out. Prior to pre-testing, the children's German teacher and I allocated one lesson per class to discuss the study with the students, to provide them with a copy of and talk through the information sheet for students, to respond to any questions or concerns and to obtain written consent (i.e., opt-in) from each individual student. Together, we discussed the study in detail and by the end of the lesson, each student provided their written consent on the consent form for students. All students opted-in to participate in the study.

Prior to piloting the post-test task with Year 3 students (more detail below in appendix 7.7), an additional information sheet for their parents was created, approved by the headmistress, and disseminated to parents via the school. No parent chose to opt-out. On the day of piloting, each Year 3 student was provided with and talked through the information sheet for students and given the opportunity to ask questions. All piloting students opted-in to participate in the study and provided written consent. The host school kept hold of all the Year 2 and Year 3 children's consent forms.

3.1.4 Pre-Tests

Similar to previous FL teaching intervention studies in primary schools (e.g., Busse et al., 2021; Hopp & Thoma, 2021), a cognitive abilities test and German and English language tests were administered prior to the intervention as baseline measurements to rule out systematic linguistic or cognitive differences between the groups.

3.1.4.1 Wechsler Abbreviated Scale of Intelligence for Children (WASI) Matrix reasoning subtest

As a measure of non-verbal intelligence, the Wechsler Abbreviated Scale of Intelligence for Children (WASI) Matrix reasoning subtest (Wechsler, 1992) was administered in February 2023. Using the WASI as pre-test helped to determine whether all participants demonstrated

typical and comparable general cognitive skills. The measure has been used in previous work with primary school children (e.g., Smith & Murphy, 2015). The developers of the Matrix reasoning subtest report a high split-half reliability coefficient for 6-year-olds (r = .89), suggesting that the subtest is largely unaffected by measurement errors (Wechsler, 2011).[14] While the authors also report strong evidence for the general validity of the WASI-II, they do not report subtest-specific validity results.

Note that the additional administration of a short-term memory digit span task was proposed in earlier examinations of this thesis (i.e., Transfer of Status). However, after consultation with colleagues from Experimental Psychology during research group meetings, the digit span task was deemed unnecessary as the WASI test already provided an approximate measure of cognitive ability.

3.1.4.1.1 WASI Materials

The WASI measure is a pen- and paper-based test including 24 trials. Per trial, participants are presented with an unfinished matrix of six items and asked to choose one of several provided options to complete the set (Wechsler, 1992). As per the test instructions, the WASI should be conducted with each child individually. Test administration takes approximately 15 minutes per child. Since 40 children had to be tested, this would have amounted to at least ten additional hours of removing students from their classes to test them which was not possible. Despite consultations with several researchers in Education and Experimental Psychology, no suitable digital alternative to the WASI could be found that would have accelerated the testing process. After careful consideration and discussion with colleagues, it was decided to stick to the WASI's pen-and-paper format yet have all children in each class be tested simultaneously.[15] Therefore, printouts of the WASI containing six matrices on each page were created, following the original testing order. As a result, the entire matrix reasoning subtest, consisting of 24 items, fit on four DIN-A 4 pages. Care was taken to ensure good readability and all copies were made in original colours. Examples are provided in appendix 7.3.

[14] The split-half method is a technique to assess the reliability of a test. Here, the subtest is roughly divided into two equal parts, matched for difficulty. The split-half reliability coefficient represents the correlation between the total scores of the two half-tests (Wechsler, 2011).

[15] Changing the WASI test format was deemed appropriate since the WASI results were regarded as a simple baseline measure to rule out differences between groups. They were not supposed to be integrated as predictors in multi-level models.

3.1.4.1.2 WASI Procedure

Both classes were tested on the same day during lesson time, one after the other. Prior to testing, the host-teacher and I prepared the room in such a way that each child would sit on their own, facing away from other children, to minimize distraction. Each child was provided with a pen and with the exact same colour copy of the test (i.e., four pages with six matrices per page). As a whole class, the children received instructions as per the original test instructions, including an example on the Whiteboard. Instead of pointing to the individual items in each trial, we asked the students to circle the item they thought matches the matrix. Students had the opportunity to ask clarification questions. Repeatedly, the teacher and I put a strong emphasis on the importance of the children completing the test by themselves. We instructed the class to be silent throughout testing and each child to stick to their own sheet. Once started, the host-teacher and I stayed in the room to guide them through the test and to make sure students would remain quiet and stick to their own work. During trials, we waited for each child to circle one option before the whole class moved on to the next trial. Only once everyone had finished a trial the class moved on to the next trial. In fact, prior to testing, students were instructed to wait patiently for their peers who might require more time for their answers. The sound of a triangle was used to signal the students to move on to the next trial. Compliance with our instructions was high. The testing procedure including the initial instructions took approximately 45 minutes. The tests were collected from each child, and they were thanked for their collaboration and patience.

3.1.4.1.3 WASI Scoring

Scoring followed the original WASI instructions. Children received one point for each correct response. As a result, each child received a WASI score between 0 and 24. Scoring stopped immediately once a child had given three incorrect responses consecutively. In this case, all remaining responses were ignored, even if they were correct.

3.1.4.2 Picture Vocabulary Size Test (PVST)

In earlier examinations of this thesis (i.e., Transfer of Status), the initial proposal was to use the pen-and-paper-based British Picture Vocabulary Scale-II (BPVS) to measure English vocabulary skills. Like the WASI, this vocabulary test takes approximately 15 minutes to complete and is conducted with each child individually. Again, given the organizational constraints of taking 40 children out of their lessons for 15 minutes each, in addition to the

unavoidably time-intensive one-to-one outcome measures for each experiment (see section 3.1.8), the school requested if it would be possible to test all students simultaneously on their English vocabulary. After careful consideration and consultation with colleagues, Paul Nation's and Laurence Anthony's digital *Picture Vocabulary Size Test* was considered suitable for testing this age group in the context of a FL classroom (Nation & Anthony, 2016). The test is intended primarily for young pre-literate native speakers up to eight years of age. It assesses receptive vocabulary size by measuring whether the test-taker can match a given partly contextualized word form with a suitable meaning illustrated by a picture (Anthony & Nation, 2021). The PVST utilizes the 6000 most common word families in English, tailored for young native-speaking children. Each trial represents 62.5 words in the source lists, adding up to 96 trials.

The test developers report that they trialled the test repeatedly to improve validity (Anthony & Nation, 2021). Data regarding the test's reliability is not reported. The authors provide the program as freeware to download from the internet. Since it was not feasible to install the program on each school computer individually, an exact copy of the original test was created, including all stimuli kindly provided in raw format by the creators, and implemented into the Gorilla testing environment (www.gorilla.sc; cf. Anwyl-Irvine et al., 2020). This way, the test could be accessed via a simple link on a computer, and the whole class could take the test simultaneously during a computer lesson.

3.1.4.2.1 PVST Materials

All materials were downloaded from the test creators' website and no changes were made. Each trial consisted of four pictures and one audio file. During trials, pictures were always arranged in a rectangular grid with one picture in each corner of the screen. Like in the original program, the order of trials and the arrangement of pictures on the screen were not randomized between or within participants. Thus, each child received the same input during testing. A 'Play' button was positioned in the middle of the screen. Children had to click on 'Play' for the audio to start in each trial to enable them to complete the test at their own pace. Per picture matrix, the audio consisted of a short sentence preceded by the target word, such as: *Behind. He is behind the car*. To avoid having the students click through the experiment without paying attention to the audio, they had to listen to the full audio at least once before the pictures became 'clickable'. As in the original program, participants had the possibility to repeat the audio once. Halfway through the experiment, a pause screen with a message

popped up congratulating the children for having finished half of the experiment and encouraging them to stay concentrated and commence with the second part.

The entire PVST can be accessed in Gorilla (Note to examiners: Enter any ID). All visual and aural stimuli are provided online on Laurence Anthony's website.

3.1.4.2.2 PVST Procedure

Before conducting the PVST, I made sure on a separate occasion that all computers and headphones worked. On the day of testing, the PVST was administered to each class separately during computer lessons. Each child took the test on their own computer while wearing headphones. Prior to testing, the entire class received instructions and had the opportunity to ask clarification questions. Again, they were instructed to stick to their own computer and remain concentrated and quiet. During the test, in each trial children heard an English audio file and were presented with a matrix of four pictures. They were asked to click on the picture they thought corresponded to what they heard. Before the main testing started, there were five training trials to help everyone get acquainted with the task. Instructions were repeated in writing again in the experiment. The host-teacher and I stayed in the room during testing to encourage the class to remain quiet and concentrated. Compliance with our instructions and concentration levels throughout testing were high. Children who had finished the test were allowed to complete another task unrelated to the current study. The entire procedure took approximately 45 minutes per class.

3.1.4.2.3 PVST Scoring

Participants received one point for each correct response. According to the test creators' instructions, to calculate a participant's estimated vocabulary size, represented as word families, their raw score is multiplied by 62.5 (since each trial represents 62.5 words in the source lists, see section 3.1.4.2). Thus, the potential high score in vocabulary size is 6,000 (*96 trials x 62.5*) which corresponds to the maximum vocabulary size the PVST is suitable to be used for. The formula to estimate each child's approximate vocabulary size (in word families) is *size = n(correct)*62.5* (Anthony & Nation, 2021).

3.1.4.3 Language Magician

As a German language proficiency test, the children played the digital *Language Magician* Game. This language assessment game for primary school students was developed by the European Union and research institutes across Europe. It is a response to the recent rise

in FL instruction at primary school level and provides teachers with a child-appropriate tool to measure FL ability in various languages, including German (Language Magician website). The assessment taps into reading comprehension, listening comprehension, and writing skills. Since the game was not designed as a scientific tool for data collection, no reports evaluating the game's reliability and validity are available. Although the ongoing large-scale validation of the program has not yet been published in peer-reviewed journals, the consensus from personal communication with Professor Suzanne Graham (Reading University, UK), one of the scientific leads of the project, was that the game provides a solid instrument to measure general proficiency (cf. Klein, 2018). The participating children were already familiar with the game's format having completed it in the past year with their German teacher. Like the game creators had intended, the school used the game once a year to get a rough idea of their students' annual progress in FL German. For the current study, the last time the students had played the game had been one year prior to their participation in this study. The game is specifically designed for classes to play it annually to monitor children's progress. To ensure that progress can be tracked successfully, the Language Magician is designed in a way that having played it before does not have an impact on subsequent performance.

3.1.4.3.1 Language Magician Materials

The Language Magician can be accessed online. Teachers need to sign up to the website to get access to the game for their students. The game contains visual and aural input and is animated in cartoon-style. Together with some animals, children go on a quest in an old medieval tower to defeat an evil magician. They gain points (i.e., 'magic stars') for completing language assignments and successively move up in the tower. With each additional floor, the difficulty of the game increases. The game begins with simple Reading and Listening exercises and ends at the most difficult level with Writing exercises. There are 90 tasks in total and children have two attempts per task.

An English version of the game's manual can be accessed here.

3.1.4.3.2 Language Magician Procedure

The students required few instructions as they were already well acquainted with the game format having played it before. As part of their normal curriculum, they would have played the game anyway to assess their FL German progress over the year. Like with the PVST, we checked the technical equipment prior to testing. Each class played the game separately

during German lesson. In the computer room, each child received a pair of headphones and completed the game on their computer. They were given instructions and asked to stick to their own game and keep quiet. They also had the opportunity to ask clarification questions. During the game, the students progressed through the tower independently and the game provided them with additional instructions along the way. The host-teacher and I stayed in the room throughout the entire game. The children enjoyed the game and concentration levels were high.

3.1.4.3.3 Language Magician Scoring

The Language Magician is an online game and provides automatically calculated outcome scores for each participant. It provides a percentage score of 'correct' responses across all trials per participant. For example, if a participant has a score of 56, that means that 56% of their responses (across the different levels of difficulty in the game) were correct.

3.1.5 Teaching intervention

All pre-testing was completed over several days during a period of three weeks in early February 2023. Two weeks after that, experiment 1 started. It took two weeks, following the temporal set-up shown in Table 6. Note that experiment 2 (see section 4) commenced in the week after the end of experiment 1.

	Pre-Intervention	Intervention							Post-Tests		
Day	N/A	1	2	3	4	5	6	7	8	9	10
Class 1	pre-tests	back-up day	High Variability condition						Testing		
Class 2	pre-tests	back-up day	Low variability condition						Testing		

Table 6 Time frame experiment 1. The back-up day was intended for potential travel disruptions caused by industrial strike actions.

Similar to Wonnacott et al. (2012), experiment 1 introduced variability in a VAC's verb slot to investigate whether input variability in this slot impacts students' learning of global construction semantics and linking rules. The target construction described an 'approach event' (*To the Y goes the X*), denoting that X approaches Y in a certain manner: *To the gorilla jumps the camel* or, in German: *Zum Gorilla huepft das Kamel*. (A more detailed discussion of the markedness of these constructions caused by the order of subject and object can be found

in appendix 7.4). The construction consists of a prepositional phrase featuring the object (*zum Gorilla*) and of a verb phrase split into a movement-verb (*huepft/jumps*) and the subject (*das Kamel*). While all object-, subject- and verb-positions are fillable slots, the verb slot was the target taught with high/low input variability. To ensure students understood the rest of the construction, all accompanying nouns in object- and subject-position were German-English cognates, such as *elephant-Elefant*, *bear-Bär*, *camel-Kamel.*

3.1.6 Materials

3.1.6.1 Input – Exposure sentence sets

Prior to the experiment, the host-teacher and I met privately to plan the teaching for the upcoming weeks. The 'animals' context (e.g., camel, giraffe) was chosen as it provided a good basis for sentence generation and would have been covered in German lessons anyway during this teaching period in Year 2, according to the school's curriculum. In addition, stuffed toy animals could be used during input sessions as well as during outcome testing. According to the host-teacher's experience, such a context would be engaging for the children.

The target sentence structure was *To the Y goes the X*. To generate sentences, suitable animals meeting several conditions had to be found. First, they had to have cognate names in English and German. Second, the animals had to be available to buy as toy animals and the toys had to be of a similar size and make. This avoided the possibility that any observed learning effects were the result of some animals being fluffier or larger than others. The aim was to find a group of six animals similar to Wonnacott et al. (2012). The final set included camel (*Kamel*), gorilla (*Gorilla*), bear (*Bär*), elephant (*Elefant*), zebra (*Zebra*), and giraffe (*Giraffe*).

The verbs for the VACs' verb slots had to fulfil several conditions as well. First, they had to denote an approach event. Second, they had to be similar in length (i.e., two syllables). German compound verbs had to be avoided. This was to control that word length did not impact students' learning. Third, the verbs must not have been part of the teaching/curriculum yet. Fourth, while it was necessary for all verbs to denote approach events, their semantic meanings needed to be distinct enough from each other. This differentiation was crucial to ensure that I could unambiguously enact each verb's meaning using toy animals without any ambiguity across verbs. For example, it is difficult to unambiguously enact with toy animals the difference between *bouncing* and *jumping*. While the different verbs' enactments had to be distinctive from one another, they also had to be relatively easy to mimic, as the students

would have to enact sentences themselves during testing. Note that verbs such as *walking* or *galloping* were intentionally not used, as they represent some animals' natural movements. If those 'natural' verbs had been combined with 'unnatural' verbs such as *sliding*, it would have been impossible to tell whether the 'unnatural' verbs were semantically more salient to the children (considering their prior experience of how animals such as gorillas or camels move) and thus easier to remember. In fact, it was decided that a verb could only be included if it did *not* represent the normal way of moving for *any* of the included animals (e.g., *climbing* would not have been suitable since gorillas climb but elephants do not). This way, it was assumed that if a verb was 'unnatural' it would be equally 'unnatural' for all included animals and therefore each verb would pose an equally difficult learning challenge for the children. After careful consideration of all conditions and discussion with colleagues and the children's German teacher, it was decided to use the following seven German verbs: jump (*huepfen*), 'slide on your belly' (*rutschen*), scoot (*robben*), tumble or 'doing a somersault' (*purzeln*), roll (*rollen*), sneak (*schleichen*), and 'slide on your back' (*schlittern*). Example videos of each movement can be found in appendix 7.5. It is important to keep in mind that while it might certainly be debatable from a semantic perspective whether the enacted movements for verbs such as *schlittern* or *rutschen* are absolutely accurate (e.g., one could *schlittern* on one's feet instead of on the back, or one could *rutschen* on one's bottom instead of on the belly), what is important in the context of the current study is that the different movements can unambiguously be discriminated from one another and be associated with one and only one of the included verbs. As the children did not know the German verbs anyway (and would not be taught about them explicitly at any point throughout the study), the details of the German semantics were considered less important than clear enactments.

The choice to use seven verbs, instead of, say, six or eight, was based on previous work by Wonnacott et al. (2012). In their study, they used a low variability to high variability verb-ratio of one to four and successfully detected a variability effect in a similar experimental context. Specifically, children in their low variability group were exposed to one verb type during exposure, children in the high variability group were exposed to four verbs. In addition, the plan in the current study was to administer three different outcome measures. Therefore, three additional verbs were required as children were tested on both familiar and unfamiliar verbs. Thus, in sum, seven verbs were required.

Prior to the intervention, 16-sentence exposure sets were randomly generated for each class. This set size was used in Wonnacott et al. (2012), and the host-teacher and I deemed it realistic to expose each class to 16 sentences per day, adding up to 96 sentences across 6 teaching days. Using balanced randomisation, 16 different 'animal-combinations' were drawn from the pool of six animals. This resulted in all animals being drawn five times each, except for giraffe and elephant being drawn six times each (*(4x5+2x6)/2=16*). From this pool of 32 animals, sets of two different animals each were drawn randomly. Then, for each set of two, the linking-order (i.e., which animal approaches which) was randomized as well. After that, four random verbs for the HV condition and one random verb for the LV group were drawn from the pool of verbs. For the HV group, the 16 animal-combinations and the four randomly drawn verbs were combined randomly to generate 16 complete VACs sentences. For the LV group, the same 16 animal-combinations and the one randomly drawn verb were combined as well. This resulted in two sets of 16 sentences shown in Table 7. For both the LV and HV condition, six versions of their set of 16 sentences were created, each version with randomized sentence order. The resulting six exposure sets per condition were used as input over the 6-day teaching intervention in each condition.

HV	LV
Zum Baer purzelt die Giraffe	Zum Baer huepft die Giraffe
Zum Baer robbt das Zebra	Zum Baer huepft das Zebra
Zum Elefant huepft der Baer	Zum Elefant huepft der Baer
Zum Elefant purzelt das Zebra	Zum Elefant huepft das Zebra
Zum Elefant purzelt der Gorilla	Zum Elefant huepft der Gorilla
Zum Elefant robbt der Gorilla	Zum Elefant huepft der Gorilla
Zum Elefant schlittert der Baer	Zum Elefant huepft der Baer
Zum Gorilla huepft die Giraffe	Zum Gorilla huepft die Giraffe
Zum Kamel huepft der Gorilla	Zum Kamel huepft der Gorilla
Zum Kamel schlittert der Baer	Zum Kamel huepft der Baer
Zum Zebra huepft der Elefant	Zum Zebra huepft der Elefant
Zum Zebra robbt die Giraffe	Zum Zebra huepft die Giraffe
Zum Zebra schlittert der Gorilla	Zum Zebra huepft der Gorilla
Zur Giraffe purzelt das Kamel	Zur Giraffe huepft das Kamel
Zur Giraffe robbt das Kamel	Zur Giraffe huepft das Kamel
Zur Giraffe schlittert das Kamel	Zur Giraffe huepft das Kamel

Table 7 Example input sentence sets for high variability (HV) and low variability (LV) conditions in experiment 1.

3.1.6.2 Outcome measurements – Test sentences preparation

Outcome testing consisted of three separate tasks: act out comprehension, production, and a forced choice task, similar to Wonnacott et al. (2012). Each task included familiar and unfamiliar (i.e., novel verbs) test trials. In the act out comprehension task, children heard a sentence and had to enact its meaning using the toy animals. In the production task, children saw an enacted sentence and had to produce the sentence. In the forced choice task, children saw two videos each showing the same action with reversed linking rules and had to choose the correct one relative to a sentence they heard. Detailed task procedures are provided in section 3.1.6.2.1.

The test sentences for the outcome measurements featured the same six animals as the input sentences during exposure. Each task included four trials and was completed twice, once with familiar verbs, and once with unfamiliar verbs. Thus, in total, each child was tested on 12 familiar and 12 unfamiliar trials spread across three tasks (i.e., sentences containing familiar/unfamiliar verbs). Two 12-set animal-combinations (balanced randomisation both for frequency of each animal and order of animals) were generated. Then, for each child in the HV exposure group, one random verb was drawn from the group of four already familiar verbs from the exposure phase and combined with the first set of 12 animal-combinations (3 tasks x 4 sentences) to generate the *familiar* test sentences. For children in the LV condition featuring only one verb during exposure anyway, this specific verb was used in all 12 familiar trials across the three tasks. Regarding the unfamiliar trials, for each child regardless of whether they came from the LV or HV exposure condition, the three verbs not already used during exposure in the HV condition were randomly combined with the remaining set of 12 animal-combinations.[16] Note that those unfamiliar verbs were distributed at task level. Thus, each of the three tasks during testing with unfamiliar sentences was associated with only one unfamiliar verb. Appendix 7.6 shows the test sentence output generated for each one exemplary child in either the LV or HV exposure condition. The entire described procedure was repeated for each child across both input conditions.

[16] The same three remaining verbs were used for testing of children coming from the LV and the HV exposure condition, although, in theory, I could have drawn from 6 remaining unfamiliar verbs for children from the LV condition. However, to avoid verb effects during testing of unfamiliar verbs, it was decided to use the exact same unfamiliar verbs across children who experienced high and low input variability.

3.1.6.2.1 Forced choice tasks

While the act out comprehension and production tasks featured in-person enactment during testing, the forced choice task had to be prepared in advance in the Gorilla testing environment. Children would see two videos above each other simultaneously on the screen. In both videos, the same approach event featuring the same two animals was enacted. The only difference between the videos was that one video showed the correct linking rule while the other showed the reverse, that is the wrong linking rule (i.e., which animal approaches which). Whether the correct video was on top or on the bottom was randomized for each trial in Gorilla.

Since each child had to complete the forced choice task twice, once featuring a familiar verb and once an unfamiliar verb, and the relevant test sentences were generated randomly (see section 3.1.6.2), videos of every possible combination of nouns and verbs had to be prepared. Example videos are included in appendix 7.5. Once all videos were prepared and all familiar and unfamiliar test sentence sets were generated for each child, an individual spreadsheet associating a set of test sentences with the respective videos (two videos per sentence; correct linking rule and incorrect (i.e., reversed linking rule) was generated for each child and implemented in Gorilla.

3.1.7 Teaching Arrangement – General considerations

Prior to experiment 1, the host-teacher and I planned the detailed teaching arrangement. Specifically, we planned how I would implement the target sentences into the daily German lessons. In doing so, we made every effort to find a good balance between authenticity and controllability of input. On the one hand, authenticity meant teaching the input sentences in the most natural way possible resembling typical classroom processes. Authenticity was crucial from a pedagogical perspective since the crux of the current experiment was to test findings from controlled environments in the 'noisy' environments of authentic classrooms. On the other hand, controllability of input and comparability of teaching settings across conditions was crucial from an experimental perspective. While we did aim to find the right balance between authenticity of teaching and careful experimental control, we acknowledge that we put a greater emphasis on control given this experiment acts as a proof of concept that this kind of input manipulation can be effective in FL classroom teaching. While the initial plan had been to implement sentences individually throughout each lesson as part of a larger teaching topic, we decided to stick to an approximately 15-minutes window per

lesson to introduce the entire set of sentences for each day at once. This decision was made for two major reasons: First, students regularly left the classroom for various reasons (e.g., music lessons) during German lesson time. As it was not possible to have all students stay in the classroom for the entire German lesson every day for the entire duration of the experiment, we found a compromise with the headmistress which involved concentrating all input teaching in a period of 15-minutes per lesson where the host-teacher and I could make sure that all students were in the room. Second, during a test-teaching lesson in early 2023 (covering topics unrelated to the current study to get to know the class) we realized that naturally occurring activities in the classrooms, such as questions or misbehaviour, made it extremely difficult to teach the exact same lesson in both classes while making sure to articulate the exact same input in both classes. As a result of those two issues, it was decided to structure the teaching arrangements for the benefit of controllability and comparability across classes. Specifically, teaching in experiment 1 was concentrated in a 15-minute period in each lesson, always at the beginning. Although the student's German teacher was always in the room, it was only me teaching during those 15 minutes. The students and I sat in a circle on the carpet. I wore a small lapel microphone (i.e., a microphone attached to the collar) to record what I was saying. Although I strictly followed a printed-out script of input sentences in both experiments, the microphone was intended as an instrument to check after each lesson whether I had adhered to the script. The recordings showed that I had consistently followed the script in each lesson without omitting any input.

3.1.7.1 Teaching arrangement - Details

Prior to experiment 1, both classes were introduced to the six toy animals in a separate lesson to make sure that all children were well acquainted with the animals' names. During that lesson, the students were engaged in role-plays where they bought the stuffed animals from a local toy store and haggled for the price. During the intervention, on days 1 to 6, I spent a couple of minutes each day with the children in both classes to explore topics related to the animals. This was supposed to keep up engagement and included topics such as '*Where do the animals come from*' (using an inflatable globe), '*What do the animals weigh*?' (using a scale), '*In which movie/book did you encounter the animals before*' (on national book day), or '*What do the animals eat*?' (using some hay and grass). Those introductions were given in English and would typically take around 5 minutes. The target constructions were never part of the introductions.

Following the introduction, I would make a transition to animals visiting each other, for example along the lines of '*We explored where the animals live, now they visit each other*' or '*We explored what the animals eat, now they invite each other over for dinner*'. The host teacher and I tried our best to have the transitions not too far-fetched. Importantly, the introductions and transitions were always the same in both classes. After the transition, the daily exposure phase featuring 16 sentences began. First, I read out a sentence while simultaneously acting out its meaning with the toy animals. Then, the students and I repeated the sentence together as a choir while I enacted the sentence again. After receiving praise for their good work, we moved onto the second sentence. While this exposure setting had a slightly artificial character it allowed for a high degree of controllability and still resembled an authentic call-and-response teaching situation as much as possible. Once all 16 sentences were finished, the children had a short refreshment break, and their usual German teacher took over the teaching. The animals did not feature in the remainder of the German lessons, again, to control input frequency in both classes.

3.1.8 Outcome measurement procedure

Note that all pre-tests as well as the outcome measures were piloted. Details are provided in appendix 7.7.

Outcome testing started on day 7 of the experiment, following 6 days of exposure. In contrast to the teaching sessions, these tests were conducted individually. It was decided to conduct the tests exclusively in the morning due to the school's afternoon program, which involved numerous out-of-class activities such as PE lessons. These activities would have caused significant disruptions to the testing process. The testing took place in the corridor in front of the classrooms during lesson time. Before testing, two chairs and a table were set up for playing with the toy animals. In addition, a laptop was set up for the Forced Choice task. All the relevant teachers were notified and supportive of the arrangement that students would undergo individual testing sessions throughout the morning. Following no specified order, the teachers sent out a new student each time the previous student had returned to the classroom. Following each lesson, we alternated between classes. Specifically, for one lesson, only students from the HV condition were tested, and for the next lesson, only students from the LV condition were tested, and so forth. Although this rotation made logistics more difficult, it was implemented to prevent students from one input condition being tested one or two days ahead of students from the other input condition. The testing of each student took

approximately 15 to 20 minutes. Overall, testing was completed over the course of three mornings with the vast majority of students being tested on either the first or the second day of testing. Only two students were tested on the third day due to their absence during the initial two days of testing.

Before the outcome measurements were administered to each child, we spent some time talking about how they were doing and things like which animal was their favourite animal to make each student feel comfortable with the situation. Then, the testing began, always following the same script. First, the Act Out Comprehension task, the Production task, and the Forced Choice task were conducted with one familiar verb each. Following that, the same tasks were repeated in the same order, featuring one unfamiliar verb each. This specific task order was selected with the anticipation that children would excel in familiar tasks which would initiate testing on a positive note and sustain their motivation throughout. In addition, act out comprehension tasks were completed first since they were considered most engaging as children would interact with the animals. To avoid having to change back and forth between the animals and the laptop too often during the testing procedure, the production task was completed second, followed by the forced choice task. The order of tasks was the same for each child. What varied between children were (a) the three familiar verbs that were randomly drawn from the pool of four familiar verbs taught during the exposure lessons, (b) which task each familiar verb was associated with, and (c) which task each of the three remaining unfamiliar verbs was associated with. Note that (a) and (b) did not apply to children from the LV condition since there was only one familiar verb from exposure.

The detailed testing procedure is outlined below:

Familiar trials:

Act Out Comprehension task – Familiar verb (4x test trials)

1. The child was instructed to act out the sentence I read out to them.
2. I read out a test sentence featuring a familiar verb.
3. The child acted out the sentence.

Production task – Familiar verb (4x test trials)

1. The child was instructed that we would now swap roles and they had to say the sentence which I would enact.
2. The child was told that in order to support them a little bit I would give them one word (i.e., the verb).

3. I enacted an approach event featuring a familiar verb and began the description by saying the verb.
4. The child completed the sentence.

Forced Choice task – Familiar verb (4x test trials)

Two video clips were played simultaneously, and at the same time, I read out the test sentence. This procedure was repeated once if the child asked me to.

1. The child was informed that they would hear a sentence and simultaneously see two short video clips on the screen. They were instructed to indicate the clip they believed matched the sentence (NB: the arrangement (top/bottom) of the correct/incorrect clip was randomized for each trial). The child was told that it would not matter whether they pointed to the clip or verbally indicated which one they believed was correct; the method of response did not matter.
2. The child was told that they could request one repetition of the procedure per trial.
3. I played the clips simultaneously and read out the sentence.
4. The child indicated which clip they believed matched the sentence.

Unfamiliar trials

Act Out Comprehension task – Unfamiliar verb (4x test trials)

1. The child was given instructions that they would have to enact scenes using the toy animals again. However, this time, I showed them a new (i.e., unfamiliar) word before they enact the scene.
2. I illustrated the meaning of the unfamiliar verb:
    a. I reminded the child of a familiar sentence encountered during exposure. I said the sentence while enacting it as an example (e.g., *Zum Gorilla huepft das Kamel*).
    b. Then, I used a different animal and demonstrated through enactment that the verb from the preceding example sentence can be used intransitively: "Now this is *huepft*", while the animal is '*huepfing'* back and forth (e.g., the zebra jumping back and forth without an antecedent).
    c. Then, I introduced the unfamiliar verb in an intransitive structure using yet another toy animal: "This is *rutscht*", while making the giraffe '*rutsch'* (slide) on its own.
3. I read out a test sentence featuring the unfamiliar verb.
4. The child acted out the sentence.

Production task – Unfamiliar verb (4x test trials)

1. The child was instructed that we would now swap roles again and they had to say the sentence which I would enact.
2. The child was told that in order to support them a little bit I would give them one word (i.e., the verb).

3. I warned the child that they would see an action that they are unfamiliar with.
4. I enacted an approach event featuring the unfamiliar verb and began the description by saying the verb.
5. The child completed the sentence.

Forced choice task – Unfamiliar verb (4x test trials)

Two video clips were played simultaneously, and at the same time, I read out the test sentence. This procedure was repeated once if the child asked me to.

1. The child was informed that they would hear a sentence featuring an unfamiliar verb or action, and simultaneously see two short video clips on the screen. They were instructed to indicate the clip they believed matched the sentence (NB: the arrangement (top/bottom) of the correct/incorrect clip was randomized for each trial). The child was told that it would not matter whether they pointed to the clip or verbally indicated which one they believed was correct; the method of response did not matter.
2. The child was told that they could request one repetition of the procedure per trial.
3. I played the clips simultaneously and read out the sentence.
4. The child indicated which clip they believed matched the sentence.

As the forced choice task was supposed to tap into the children's understanding of linking rules, it was deemed unnecessary to introduce them to yet another new verb prior to the 'unfamiliar' testing.

Throughout the entire testing process, children were not provided with feedback as to the accuracy of their responses. No matter the response, the children were praised and encouraged to keep up the good work. After testing, each child was thanked and sent back to their classroom.

3.2 Results

I begin by presenting the outcomes of the pre-tests. Then, the analysis plan is presented, followed by the descriptive statistics for experiment 1 results. Additional error analyses are introduced, followed by statistical analyses to answer the research question: *Does input variability in a German verb-argument-construction's verb slot impact primary school FL students' ability to generalize the construction to unattested verbs?* (see section 3).

3.2.1 Pre-Tests

The mean scores per class of all three pre-tests including standard deviations and medians are provided in Table 8. The data indicates that the means, medians, and standard deviations for all three tasks show a comparable pattern across the groups, suggesting a consistent distribution of participant performance.

			WASI			PVST			Language Magician		
class	Exp 1	Exp 2	M	SD	MD	M	SD	MD	M	SD	MD
1	HV	LV	14	6.1	16	4,142	537	4,094	56	8.8	56
2	LV	HV	15	4.3	15	4,329	355	4,375	53	6.6	52

Table 8 Pre-tests descriptive statistics. PVST expected score for this age group is approximately 4,000 (Anthony & Nation, 2021). The expected age equivalent raw scores for the WASI Matrix Reasoning subtest range from 8 points for children aged 6 years and 2 months to 12 points for children aged 7 years and 6 months (Wechsler, 2011).[17] Expected scores for the Language Magician are not available on the Language Magician website. HV = High Variability. LV = Low Variability.

Although the descriptive statistics suggest that mean performance in the tests was consistently distributed across groups, additional independent samples t-tests were conducted to test for any between group differences per pre-test. After confirming that participants'

[17] Notably, the expected age equivalent raw scores for the Matrix Reasoning subtest are lower compared to the collected data in the current study. Potentially, being unsupervised in a group decreased the attention and pressure on individual children, thus, increasing performance. Another reason for the unusually high average scores could be that children might have gone back and forth between test items (although they were not supposed to do so) which -unlike in the one-to-one administration- gave them the chance to reconsider their responses on earlier test items.
While it might be the case that the unusual mode of administration improved performance, note that the concept of 'age equivalents' -although easy for teachers and parents to comprehend- is criticized in the relevant literature since said equivalence scores are derived from raw scores which do not establish consistent intervals or evenly spaced units along the scale (Wechsler, 2011). Ideally, standardized T-scores are used to compare performances. In the case of the current study, considering standardized scores instead of raw scores slightly decreases the differences in performance between age equivalents and the current sample. For example, the corresponding T-score for 7:6-year-old children's expected age equivalent raw score (12 points) is 51, while the corresponding T-score for a raw score of 14 (i.e., class 1, current experiment) in this age group is 55 (T-score scale from 20 to 80).

mean scores on each pre-test met the normality assumption, three separate t-tests were conducted. Results provided no evidence for between group differences (all $p > .0.5$), corresponding to the descriptive statistics. (Note that this does not equal evidence for no between-group differences).

The results of the pre-tests suggest that the children in the two participating classes performed similarly across pre-tests, meaning that there were no detectable underlying group differences regarding language-related abilities which could have had a systematic effect on the experiments' outcomes. It was also considered whether there was evidence that the pre-test results were correlated with performance in either of the experiments. In that case, pre-test scores could be included as covariates in the statistical models. Plots and correlation analyses are reported in appendix 0. There was no clear evidence of associations between the pre-tests and outcome measurement performance, thus, it was decided not to include pre-test results in the models to avoid having overcomplex models.

3.2.2 Analysis plan

Separate analyses were conducted for each outcome measurement task. All analyses were conducted in the R Computing Environment (R Core Team, 2021). The data and script are provided in appendix 7.9. For each data set, a generalized Linear Mixed Model (GLMM) of the binomial family with a logit link function predicting response accuracy was run. Considering the current work's pedagogical background, a response was scored 'correct' in the act out comprehension and production tasks only if a participant produced an 'entirely' correct approach event, including the correct linking rule and the correct verb semantics. All models had fixed effects of condition (i.e., HV group vs. LV group) and familiarity at test (familiar vs. unfamiliar) as well as their interaction. An additional independent variable for 'test day', reflecting whether the child took the test either on day 1, day 2 or day 3 after exposure, was also included since this generally contributed to the model. However, 'test day' was considered a control variable (acting like a covariate in ANCOVA) and no interactions with this factor were included. All predictors were centred such that the intercept represents the grand-mean and the fixed effects can be interpreted as main effects. Also included were random intercepts for participants, a random slope by participant, a random slope for familiarity at test and correlations between the two. In each task, additional models over sub-setted data were run for each of the familiar and unfamiliar verb trials. These models were equivalent but without familiarity, interaction, and the random slope for familiarity.

Statistics taken from the coefficients for the fixed effects were used to test a set of hypotheses which came from the theory that variability promotes generalization while high token frequency promotes stronger 'whole unit' learning with those items. These hypotheses and how they relate to the effects in the different models are given in Table 9.

Each coefficient that was extracted provides an estimate of the difference in question (β) and its associated standard error (in log odds space). The associated z-scores and p-values are also automatically provided, however, in the current work it was decided instead to compute and interpret a Bayes Factor (BF) for each of these effects. The Bayes Factor (BF) serves as a powerful tool to quantify the strength of evidence in favour of one hypothesis (H1) over another (H0), or vice versa. This method allows researchers to precisely measure the degree of support for a given hypothesis (H1) when compared to the null hypothesis (H0). This stands in contrast to p-values, which lack the capability to assess the level of evidence in favour of the null hypothesis ($p > .05$ results do not signify increased confidence in the null hypothesis). In addition, Bayes Factors offer the added advantage of providing a continuous measure of the strength of evidence.

A Bayes factor tests if the data is more likely under H1 or H0, and thus its computation requires a model of the plausibility of different effects under H1 and a model of H0. Following Dienes (2014), H1 was modelled as a half-normal distribution (one tailed, because predictions are directional) with a mean of 0 and the standard deviation (SD) set to a rough estimate of the predicted difference for the hypothesis in question; H0 was modelled as a single point null. This computation requires three numbers: (1) an estimated mean difference for the effect in the data (β), (2) the associated standard error (SE), and (3) a rough estimate of the predicted mean difference (s) for that effect under H1. For each BF calculation, β and SE were drawn from the relevant coefficients in the respective mixed effects models.[18] For the predicted mean differences, ideally, values from previous similar studies are used. These were thus taken from Wonnacott et al. (2012) wherever possible, that is, wherever an equivalent effect was indeed found in the equivalent task. In cases where previous values were not available, a plausible maximum was used to determine a plausible estimate of H1. The details are presented in Table 9.

[18] This method assumes normality. This assumption is met since the calculations were in log-odds space, (Breklemans et al., 2022; Silvey et al., under review).

Although BFs are defined on a continuous scale, they are frequently interpreted according to discrete evidential categories introduced by Jeffreys (1961) and expanded by Dienes (2014) whereby BF>10 indicates strong evidence for H1, BF>3 indicates substantial/moderate evidence for H1 (and often lines up with $p< 0.05$, though this is not guaranteed), BF<1/3 indicates moderate/substantial evidence for H0, BF<1/10 indicates strong evidence for H0, and otherwise the evidence is ambiguous (i.e., the data is insensitive to test the hypothesis). Since the choice of estimate of expected effect to inform H1 is subjective, Robustness Regions (RR) for each BF were calculated, which show the range of estimates of H1 that could have been used for which the data would support the same conclusion (i.e., to accept H1/H0 or inconclusive) based on the cut-offs of BF>3 or BF<1/3. This RR is noted as [x1,x2] with x1 being the smallest standard deviation and x2 the largest standard deviation (Dienes, 2021).

Since they are more familiar to the reader, the p-values associated with the hypotheses are also reported, though they are not interpreted.

	Prediction	Motivation	Model from which coefficient statistics are extracted (Models provided in appendix 7.10)	Relevant coefficient (from which beta and SE are extracted)	Model of H1 Estimate of predicted effect size in log odds This will be set as SD of half normal with mean of 0. Odds ratio in parentheses.
1 – act out	There is an interaction between verb familiarity and variability in the direction of better performance for children from the HV condition on unfamiliar verbs and no such effect (or a reverse effect) for familiar verbs.	Higher input variability will lead to greater generalization and thus greater performance in children from the HV condition than from the LV condition for unfamiliar items. Specifically, this benefit isn't relevant for LV items, where instead there could be a benefit of LV due to higher token frequency.	*model 1*: predicting accuracy in act out comprehension task	input group by familiarity	3.93 (50.91)[19]
1a – act out	In unfamiliar trials, children from the HV condition show better performance in the act out comprehension task than do children from the LV condition	Higher type frequency of structures during exposure will lead to better generalization of those structures, and give an advantage when enacting unfamiliar structures	*model 1a*: predicting accuracy in unfamiliar trials of the act out comprehension task with 'unfamiliar' as reference level	input group at reference level	1.52 (4.57)[20]

[19] Extracted from beta value in Wonnacott et al. (2012) for two-way interaction of Input Group (high variability vs low variability) and Verb Familiarity (unfamiliar vs familiar) in the direction of the HV input group having a higher likelihood of providing correct responses in unfamiliar test trials compared to the LV group. The logit mixed model included 'Input Group', 'Verb Familiarity', 'Test Day', and the interaction of 'Input Group x Verb Familiarity' as fixed factors, and 'Participant' (intercept), 'Verb Familiarity', 'Day' and the interaction of 'Verb Familiarity x Day' as random factors.

[20] No relevant (significant at $p < 0.05$) value from previous research was available. Thus, average proportions 'correct' on day 3 taken from Table 1 in Wonnacott et al. (2012) were used to calculate beta (i.e., beta = logodds(average HV on day 3) – logodds(average LV on day 3). The direction of the subtraction was guided by the direction of the hypothesis that HV outperforms LV in unfamiliar trials.

1b – act out	In familiar trials, children from the LV condition show better performance in the act out comprehension task than do children from the HV condition	Higher token frequency of structures during exposure will lead to better learning of those structures, and give an advantage when enacting familiar structures	*model 1b*: predicting accuracy in familiar trials of the act out comprehension task with 'familiar' as reference level	input group at reference level	2.82 (16.81)[21]
2 – production	Children from the HV condition show a greater familiarity effect in the production task than do children from the LV condition	Higher input variability will lead to greater generalization and thus greater performance in children from the HV condition than from the LV condition for unfamiliar items. Specifically, this benefit isn't relevant for LV items, where instead there could be a benefit of LV due to higher token frequency.	*model 2*: predicting accuracy in production task	input group by familiarity	3.93 (50.91)[22]
2a – production	In unfamiliar trials, children from the HV condition show better performance in the production task than do children from the LV condition	Higher type frequency of structures during exposure will lead to better generalization of those structures, and give an advantage when producing unfamiliar structures	*model 2a*: predicting accuracy in unfamiliar trials of the production task with 'unfamiliar' as reference level	input group at reference level	1.52 (4.57)[20]
2b – production	In familiar trials, children from the LV condition show better performance in the production task than do children from the HV condition	Higher token frequency of structures during exposure will lead to better learning of those structures, and give an advantage when producing familiar structures	*model 2b*: predicting accuracy in familiar trials of the production task with 'familiar' as reference level	input group at reference level	10.69 (44026)[23]

[21] No relevant (significant at $p < 0.05$) value from previous research was available. Thus, beta was simply the intercept of the model, following the motivated-maximum approach (Silvey et al., under review). The motivated maximum approach entails deriving the predicted effect size from a maximum value (i.e., intercept or *grand mean*), followed by setting the predicted effect size at double this value. In the current case, the intercept was not doubled, because H1 was modelled as a one-tailed half-normal distribution since the prediction was directional.

[22] No relevant (significant at $p < 0.05$) value from previous research was available. Therefore, since the act out comprehension task and the production task are very similar and tap into the same type of knowledge, the same prior was used for the production analysis as for the act out comprehension analysis.

[23] No relevant (significant at $p < 0.05$) value from previous research was available. Thus, beta was simply the intercept of the model, following the motivated-maximum approach (Silvey et al., under review). H1 was modelled as a one-tailed half-normal distribution since the prediction was directional.

3 – forced choice	Children from the HV condition show a greater familiarity effect in the forced choice task than do children from the LV condition	Higher input variability will lead to greater generalization and thus greater performance in children from the HV condition than from the LV condition for unfamiliar items specifically; this benefit isn't relevant for LV items, where instead there could be a benefit of LV due to higher token frequency.	*model 3*: predicting accuracy in forced choice task	input group by familiarity	5.1 (164)[24]
3a – forced choice	In unfamiliar trials, children from the HV condition show better performance in the forced choice task than do children from the LV condition	Higher type frequency of structures during exposure will lead to better generalization of those structures, and give an advantage when assessing the accuracy of unfamiliar structures	*model 3a*: predicting accuracy in unfamiliar trials of the forced choice task with 'unfamiliar' as reference level	input group at reference level	1.9 (6.69)[25]
3b – forced choice	In familiar trials, children from the LV condition show better performance in the forced choice task than do children from the HV condition	Higher token frequency of structures during exposure will lead to better learning of those structures, and give an advantage when assessing the accuracy of familiar structures	*model 3b*: predicting accuracy in familiar trials of the forced choice task with 'familiar' as reference level	input group at reference level	3.2 (24.53)[25]

Table 9 Contrasts to be tested. Presented are the predictions that were tested, the motivation for testing this prediction, and the estimated effect size for each prediction. The logistic mixed effects model, from which the summary data originated, is specified for each hypothesis tested. Table adapted from Brekelmans et al. (2022).

[24] The baseline for the forced choice task was defined as 0 since in this task chance was at 50% (log odds (0.5) = 0). Thus, beta was simply calculated by multiplying the intercept of the model by 2, following the motivated-maximum approach (Silvey et al., under review). H1 was modelled as a two-tailed half-normal distribution (hence the multiplication by 2) since the prediction was non-directional.

[25] Beta was simply the intercept of the model, following the motivated-maximum approach (Silvey et al., under review). H1 was modelled as a one-tailed half-normal distribution since the prediction was directional.

3.2.3 Descriptive data

All three outcome measures yielded binary data (correct/incorrect) and were coded with '0' for incorrect and '1' for correct responses. As testing was conducted in person at school and video recordings were not part of this study, all coding across all three outcome measures was completed by the main researcher. Recall that trials were counted as 'correct' when the child enacted an approach event, correctly applied the linking rule, and used the correct verb semantics. Note, use of incorrect grammatical gender (i.e., wrong articles) was not considered an 'error' throughout the entire experiment since the use of incorrect articles, albeit sounding 'odd', would have no effect on meaning comprehension in German.

A visual overview of each participants' mean response scores across all three task types is provided in Figure 2.

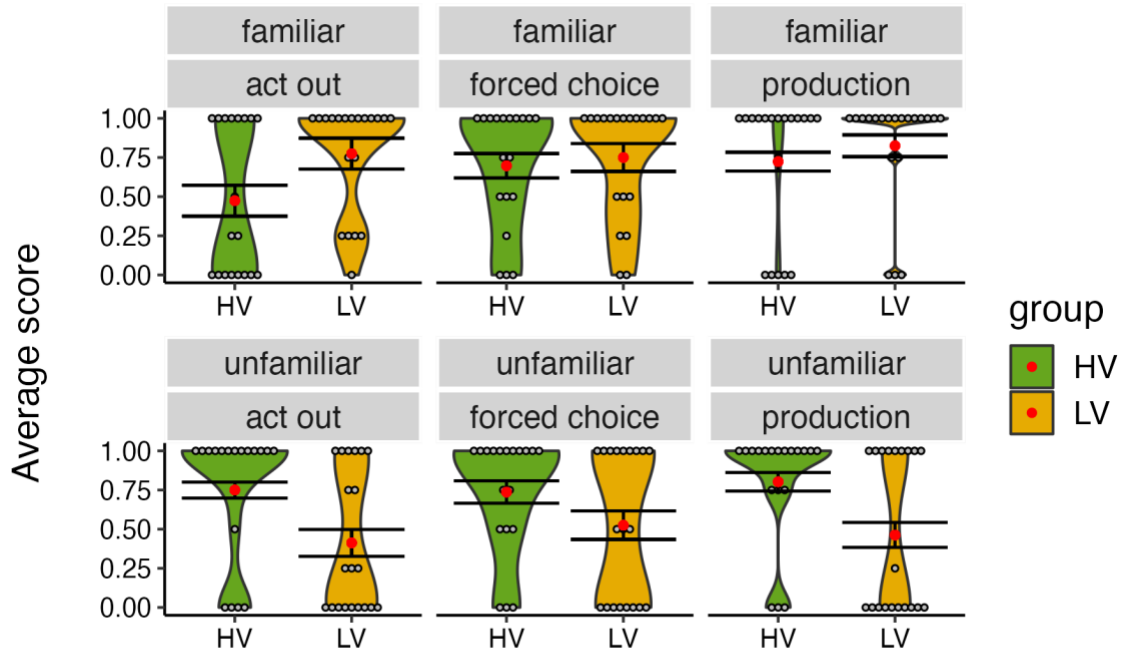


Figure 2 Violin Plots displaying participant mean proportion correct in each of the three task types during VACs (experiment 1) outcome measurements ('correct' scores are responses where the global construction semantics, the linking rule and the verb semantics are applied correctly). HV = high variability. LV = low variability. Error bars present 95% Confidence Intervals. The same visualization over trials scored 'correct' regardless of verb accuracy (i.e., only global construction semantics and linking rule correct) is presented in appendix 7.11

The plots indicate that the LV group outperformed the HV group during testing with familiar items, but the reverse is true for unfamiliar trials. This was confirmed statistically (section 3.2.5 below).

3.2.4 Error analysis for Act out and Production tasks

For act out comprehension and production tasks, the types of errors that participants were making were further investigated. They were classified based on the coding scheme outlined in section 3.2.2. The categories were:

(Reponses coded as 'correct' in main analyses)

- (1) fully correct

(Responses coded as 'incorrect' in main analyses)

- (2) correct order, wrong verb (correct linking rule but wrong verb semantics)
- (3) wrong order and wrong verb (linking rule and verb semantics wrong)
- (4) incorrect approach event (approach event but wrong animal(s) and/or wrong order and/or wrong verb)
- (5) no approach event (child produced/acted out an unidentifiable action)

Figure 3 provides a visual overview of the distribution of the different errors. Error types are only reported for act out comprehension and production tasks since the forced choice tasks by virtue of being a binary judgment allowed no specific error types.

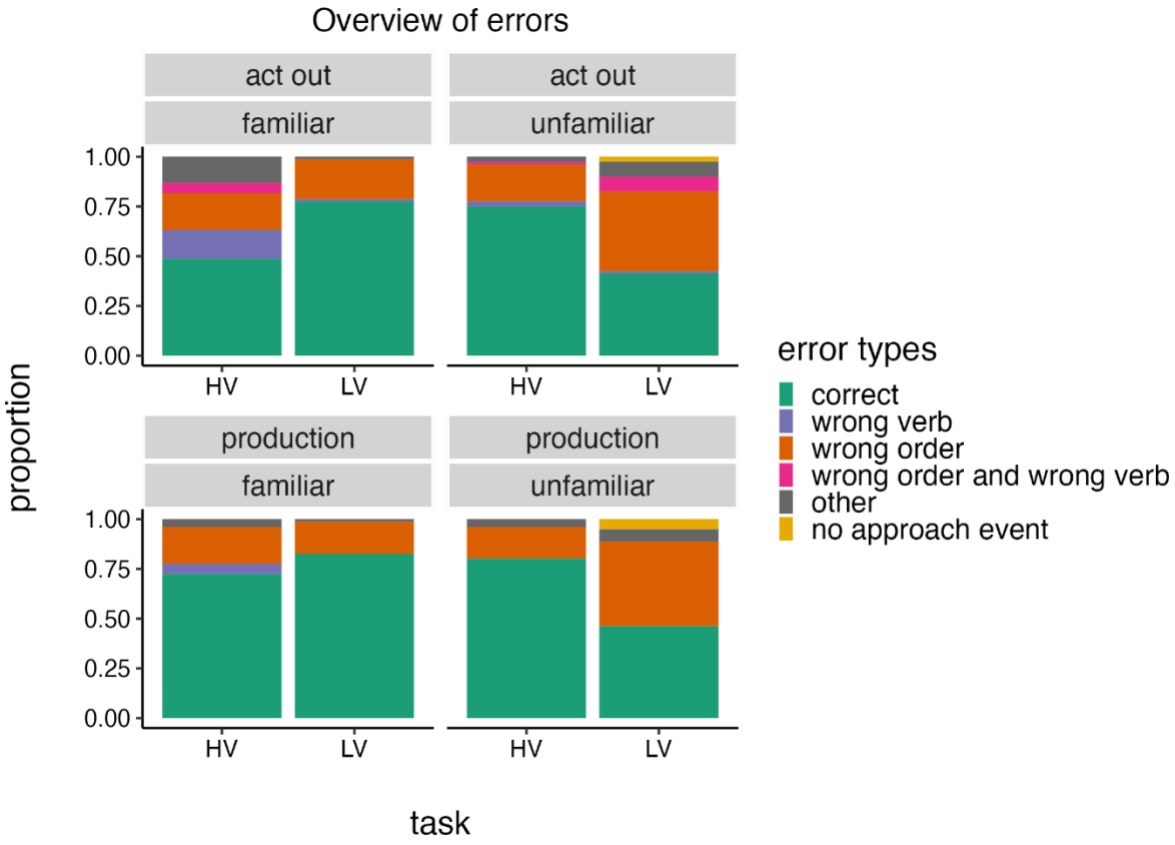


Figure 3 Distribution of different response types made during VAC outcome measurements displayed by task type, condition, and familiarity. HV = high variability. LV = low variability.

Unsurprisingly, the analysis revealed that most errors stemmed from using the wrong linking rule as most errors made by participants in both groups and across both types of tasks and familiarities were 'wrong order' errors (orange). This error type is most likely caused by transfer from English, which is to be expected given the reverse syntactical structure *The X [approach verb] to the Y* is ubiquitous in English. Other error types only account for a smaller amount of the general distribution of error types. An important observation is that children generally understood the construction's global 'approach' semantics, contrasting results from work with younger children where 'no approach' event errors (yellow) were more frequent (e.g., Wonnacott et al., 2012). In the current study, these errors made up less than 1% of all responses in act out comprehension and production tasks, respectively. Therefore, unlike in that previous work, no statistical analyses designed to capture differences in 'no approach event' responses between the conditions were conducted.

Interestingly, the largest proportion of 'wrong verb' responses (purple) comes from participants in the HV group tested on familiar verbs in the act out comprehension task, potentially reflecting the fact that this group encountered each individual verb less frequently

than the LV group. This difference between conditions is not captured in the main analyses since these trials were coded as 'incorrect' in the main analyses. (Additional analyses on 'verb accuracy' only are provided in appendix 7.12). The phenomenon of getting the linking rule right but using the wrong verb and vice versa -getting the verb right but applying the wrong linking rule- is investigated further in the discussion of experiment 1 (section 3.3).

3.2.5 Statistical analyses

Children from each group were tested on familiar and unfamiliar verbs in each of the three outcome measurement tasks. Recall that in the following main analyses, 'correct' responses are responses where the *global construction semantics*, the *linking rule* and the *verb semantics* were applied correctly. Equivalent additional analyses over trials scored 'correct' regardless of verb accuracy (i.e., only *global construction semantics* and *linking rule* correct) are presented in appendix 7.13. Although the priors in the current main analyses were derived from *linking rule correct* data, irrespective of verb accuracy (Wonnacott et al., 2012, Table 1), they were deemed appropriate (i.e., in the right ballpark) for the current analyses due to the striking similarities in the descriptive data for mean 'correct' between *linking rule correct* and *entirely correct* in the current study, as demonstrated in Figures 2 and 8 (in appendix 7.11).

Below, only statistics relevant for testing the hypotheses are reported. Full models for each task type are provided in appendix 7.10.

3.2.5.1 Act out comprehension

Recall that Figure 2 indicates that the HV group outperformed the LV group during testing with unfamiliar items, but the reverse was true for familiar trials. The statistics relevant for testing the hypotheses (see Table 9) confirmed that there was strong evidence for an interaction between input group and familiarity with BF = 15 (predicted effect = 3.93, RR = [1.6; 330], β = 12.36, SE = 3.98, z = 3.11, p = 0.0019) indicating that the data is 15 times more likely under H1 than under H0. This interaction was broken down statistically by testing evidence for (a) the hypothesis that there would be better performance in the LV group for familiar verbs and for (b) the hypothesis that there would be better performance in the HV group for unfamiliar verbs. Evidence for the former (a) was substantial with BF = 3.66 (predicted effect = 2.82, RR = [0; 2.1], β = 5.91, z = -2.04, p = 0.04). Evidence for the latter (b) was ambiguous with BF = 1.49 (predicted effect = 1.52, RR = [0; 6.3], β = 9.50, z = 1.79, p = 0.07).

3.2.5.2 Production

Recall that Figure 2 indicates that the HV group outperformed the LV group during testing with unfamiliar items, but the reverse was true for familiar trials. The statistics relevant for testing the hypotheses (see Table 9) confirmed that there was strong evidence for an interaction between input group and familiarity with BF = 1557 (predicted effect = 3.93, RR = [0.83;7549], β = 21.75, z = 5.37, p < 0.001) indicating that the data is more than 1,500 times more likely under H1 than under H0. This interaction was broken down statistically by testing evidence for (a) the hypothesis that there would be better performance in the LV group for familiar verbs and for (b) the hypothesis that there would be better performance in the HV group for unfamiliar verbs. Evidence for the former (a) was substantial for H0 with BF = 0.29 (predicted effect = 10.69, RR = [0; 10.65], β = 0.51, z = -0.18, p = 0.86). This means that there was evidence of no difference between the two groups. Evidence for the latter (b) was substantial with BF = 4.53 (predicted effect = 1.52, RR = [1.2; 6626], β = 20.95, z = 4.47, p < 0.001). Thus, the data are more than 4 times more likely under H1.

3.2.5.3 Forced choice

Recall that although Figure 2 indicates that the HV group outperformed the LV group during testing with unfamiliar items and that the reverse was true for familiar trials, this was the only task where the 95%-CIs overlapped. The statistics relevant for testing the hypotheses (see Table 9) also provided only ambiguous evidence for an interaction between input group and familiarity with BF = 1.3 (predicted effect = 5.1, RR = [0; 25.2], β = 3.81, z = 1.83, p = 0.07). The ambiguous result corresponds to Figure 2 which demonstrates that the forced choice task was the only task where the 95%-CI of both groups overlapped. This observation was further investigated statistically by testing evidence for (a) the hypothesis that there would be better performance in the LV group for familiar verbs and for (b) the hypothesis that there would be better performance in the HV group for unfamiliar verbs. Evidence for the former (a) tended towards supporting the null but was ambiguous with BF = 0.42 (predicted effect = 3.2, RR = [0; 4.6], β = 0.11, z = 0.07, p = 0.95). Evidence for the latter (b) was ambiguous as well with BF = 2.2 (predicted effect = 1.9, RR = [0; 90], β = 5.34, z = 1.76, p = 0.08).

3.2.6 Summary

For all three tasks, the data pattern in the predicted direction of stronger performance in the LV group for familiar items and stronger performance in HV group for unfamiliar items.

The main statistical analyses of the act out comprehension and production task data (i.e., using the coding where participants had to both apply the correct linking rule and get the verb semantics correct) provided strong evidence for an interaction between input condition and verb familiarity. In the production task, further analyses of unfamiliar test trials specifically showed that there was evidence for a positive effect of increased input variability on students' response accuracy in novel contexts. In the act out comprehension task, evidence for this effect in unfamiliar trials was in the same direction but remained ambiguous. Note from results provided in appendix 7.13, when running analyses over data where responses are scored 'correct' if the phrase's *structure* is correct, regardless of verb accuracy (i.e., equivalent to the coding reported in Wonnacott et al., 2012), both the act out comprehension and the production tasks yield substantial evidence for stronger performance in HV than LV.[26] Thus, the data provide evidence for the hypothesis that in this specific experimental context, students' generalisation of linguistic *structures* to novel verbs seems to have benefited from increased input variability. Verb accuracy is discussed in the Discussion for experiment 1 (see section 3.3).

Analyses results for the forced choice task data were in the same direction of an LV benefit for familiar items and an HV benefit for novel items, but the evidence for the interaction was ambiguous. Looking at unfamiliar trials only, the evidence was in the direction of a positive effect of increased input variability on students' response accuracy in terms of correct linking rule application but was also ambiguous.

[26] While the difference in act out comprehension tasks from ambiguous evidence (structure + verb coding) to substantial evidence (structure coding only) is the result of BF analyses, note from Figure 3 (error analyses) that in unfamiliar test trials only three responses in the act out comprehension task by HV students were in fact incorrect. In other words, the difference in the overall statistical results between the two types of response-coding is caused by a vanishingly small amount of incorrect verb responses.

3.3 Discussion

The results in experiment 1 indicate that students' ability to produce the correct linking rules and verb semantics was impacted by the interaction of input condition and verb familiarity. However, when the interaction is broken down, only the production task data yield substantial evidence for the hypothesis that there is better performance in the HV group for unfamiliar verbs (i.e., in novel 'context'). Regarding the act out comprehension task, evidence for this hypothesis is only substantial when verb accuracy is disregarded (see analyses in appendix 7.13). In other words, there is more evidence for the positive impact of increased input variability on accuracy at test across the two tasks when 'accuracy' disregards *verb* accuracy and only focusses on *structural* accuracy.

In the following, I will only briefly discuss the results of the forced choice task since the statistical outcomes from this task were ambiguous. I shall then turn to discuss why it is that the HV group seemingly struggled with verb accuracy in the other two tasks while at the same time the input manipulation with increased variability seems to have had the desired positive effect on the HV group's learning of the VAC's underlying structure. I will then turn to the LV group and explore what it was that they learned in the current experiment. Then, I will explore how difficulties with verb learning in HV contexts can be avoided, and touch on the role of explicit input and feedback. Lastly, I will summarise the types of learning that took place across the input conditions.

3.3.1 Forced choice task

The means in the forced choice task patterned in the predicted direction of stronger performance in the HV group for generalization items and stronger performance in the LV group for familiar items. However, although the evidence patterned in the predicted directions, Bayes Factor analyses found that it was in the ambiguous category. Thus, any discussion of this data must be tentative.

Why might the variability effect not be as strong in this task as in the other two tasks? One potential reason might be the fact that the forced choice task was always the last task during testing. The act out comprehension and the production task are already challenging and require high levels of concentration and effort from the children. By the time they reached the forced choice task they might have had a decrease in concentration. Another perhaps more likely reason might be that the forced choice task, despite not being anticipated to be

particularly challenging or confusing, is extremely difficult and possibly even confusing, especially because the preceding tasks had already introduced the students to numerous unfamiliar test sentences. Seeing the same approach action in two reverse orders in two videos on top of each other on one screen while hearing the input sentences (half of them with unfamiliar verbs) and being asked to point to (or otherwise indicate) the correct video demands very high levels of attention. In addition, as a form of encouragement, students had been praised for *all* their responses in the previous tasks, regardless of accuracy. It remains unclear if they used this praise to adjust what they had inferred from the input during the previous exposure days and if this potential readjustment impacted their performance in the last task. Nevertheless, the fact that the means are in the same 'directions' as the means of the other tasks is at least consistent with the claim that the forced choice task was still tapping in the same learning mechanism (i.e., learning of the linking rule) as the other tasks, though the lack of statistical support means it should be treated with caution.

Note that Wonnacott et al. (2012) report similar outcomes in their forced choice task. They account for this result by suggesting an 'agent-first bias' (Wonnacott et al., 2012: 474) among English L1 speakers which corresponds to the typical structure in their L1. This bias might be competing with the learned linking-rules, resulting in ambiguous data. In fact, the authors confirm this bias in a second experiment and show that the failure of the initial forced choice task to pick up on the children's linking rule learning (clearly evidenced in other tasks) was due to the 'agent-first bias' overshadowing the linking rule learning. Clearly, this bias might have impacted the results in the current experiment as well. While this complicates the interpretation of the forced choice data, it is crucial to consider biases like this when analysing data from FL teaching interventions that involve natural language input. In authentic teaching situations, FL learning is often influenced by crosslinguistic transfer effects (e.g., Hopp et al., 2019; Madlener-Charpentier, 2015: 322).

3.3.2 The variability effect in the 'noisy' classroom

When the ability to infer the underlying linking rule and generalize it to novel contexts is considered, then the data in the current study provide substantial evidence across tasks that students from the HV group are better at inferring and subsequently generalizing this structural information to novel contexts compared to the LV group (see appendix 7.13). This finding then is in line with previous research in less 'noisy' environments (e.g., Wonnacott et al., 2012) and corresponds to evidence reported elsewhere (section 1.6.3) suggesting that

increased variability in the input is beneficial for the learning of underlying structural patterns. Corresponding to Raviv et al.'s (2022) work on the robustness of the positive effect of increased input variability on learning across domains, the current data provide preliminary evidence that this effect can also hold in the noisy environment of an FL classroom. This is good news because it is precisely young FL learners' apparent lack of structural representation of language that has been criticised in the literature (see section 2.3.2), resulting in calls to find ways to equip students with a productive repertoire of linguistic structures and help them overcome the 'imitative plateau'.

Unsurprisingly, it seems that learning of underlying structural patterns comes at the cost of verb accuracy as demonstrated by the HV students' struggle with verb accuracy in familiar trials compared to students in the LV group. Although increased input variability is considered to have a positive effect on learning throughout the current work, it is important to acknowledge that it also comes with caveats, the most important one being the resulting decreased token frequency in the input.[27] Individual verbs are encountered considerably less frequently in the HV group compared to the LV group and as a result, HV students might struggle with verb accuracy due to a frequency effect.

In contrast, students in the LV group seem to have benefited from the high token frequency of the single verb that they were exposed to, which resulted in strong verb learning of that verb. This is underlined by the observation that in familiar trials students from the LV group almost never got the verb wrong even though they frequently got the linking rule (i.e., the structure) wrong (see Figure 3).

Noteworthy is also the observation that in familiar trials (see Figure 2) the data generally trended in the direction that the LV group consistently outscored the HV group across all tasks. This trend, however, crossed the threshold for substantial evidence only in the act out comprehension task and only when using the type of response coding where only entirely correct sentences are considered 'correct'. In production tasks, employing the same response coding, there was substantial evidence for no difference between groups. When the data is coded so that having the VAC structure correct regardless of verb accuracy is sufficient for a

[27] Another caveat is that increased input variability tends to make learning more difficult in the very beginning (cf. training difficulties and skewed input in Viviani et al. n.d.; see section 1.6.3).

'correct' response, the act out comprehension task yields ambiguous evidence, and the production task yields substantial evidence for no between group differences.

Thus, at least the act out comprehension task data tentatively suggest that the LV group might be better in simply re-producing the entire VAC which they, most likely, learnt in its entirety without fully breaking it down to its individual functional-structural parts; otherwise, they should have been able to generalize the structure to novel verbs, as measured in unfamiliar trials. In sum, on the one hand, the LV students seem to have struggled to disentangle the input verb (i.e., *huepft*; 'jump') from the wider structural characteristics of the VAC which limited their ability to infer the underlying structure independent of the verb and generalize said structure to novel verbs. Put differently, they might have learned the VAC as one 'whole' MWC.[28] On the other hand, on 'familiar terrain' the LV students were well able to correctly *re*-produce (or *imitate*) entire phrases from the input which included getting the verb correct – to some extent better so than students in the HV group. However, note that this 'verb advantage' of the LV group, evident in the act out comprehension task, is not present in the production task.[29] The production task data, yields evidence for *no* between-group differences in familiar trials. It is reasonable to assume that the HV group may have partially caught up with the LV group in terms of verb learning by the time of the outcome measurement. This is discussed in more detail in the following paragraph.

From a pedagogical perspective, as the students' German teacher pointed out during the study, there does not seem to be a point for the HV group in gaining structural knowledge without getting the verb correct. (Vice versa, there also seems to be no point in getting the verb correct but lacking structural knowledge). In the context of the current experiment, how could the lack of verb accuracy with familiar items in the HV group have been avoided? As touched on above, I argue that the potentially detrimental impact of increased variability on verb learning in the HV condition could have been further mitigated against simply by increasing overall exposure. In the HV group, the four individual verbs are more difficult to

[28] Note that I do not mean that students verbatim learned each VAC in the input by heart. There were too many possible animal combinations to do so. Rather, what I am referring to is that the students learned *PATIENS + huepft + AGENS* as one unit without disassociating the intervening verb (i.e., *huepft*) from the underlying linking rule.

[29] In the production task, there was evidence for no between-group differences regardless of the type of response coding used. Here, in the context of discussing verb accuracy, I refer to the coding where only entirely correct responses (comprising both verb and structure) are deemed 'correct'. The coding type where a correct VAC structure alone warrants a 'correct' response does not apply in this context

keep track of as their individual frequency is low, and students must track the four different movements (i.e., types of approach) associated with these four different verbs. By increasing the overall amount of exposure, however, the frequency of each individual verb increases as well, and over time, learners should be able to learn their meaning simply through repeated encounters of the verb and its associated meaning. Crucially, the equivalent cannot be said of learning of the *structure* in the LV . That is, even if the LV input was frequently repeated, the underlying structure of the VAC would be unlikely to be learned in the LV group because the reliability of the cues about the characteristics of the underlying structure does not change and remains comparatively weak (i.e., compared to HV condition). In the LV condition, there is only weak evidence for the link between the Object-Subject structure and *Patiens-Agens* semantics which is independent from the intervening verb. This is because with only one intervening verb (i.e., *huepft*), structure-meaning is never dissociated from this verb. This restrictiveness most likely suggests to learners that this property can only be applied to a limited number of verbs, or potentially only to this individual verb (cf. Bybee, 1995; see section 1.6.1). Crucially, an increase in exposure does not change anything about this. In contrast, verb learning in the HV group, although potentially problematic with too little input, can improve over time with more input because a basic frequency effect would most likely kick in eventually and at least to some extent make up for the 'worse' verb learning. Simultaneously, the increased input variability still supplies learners with relatively reliable cues about the VAC's underlying structure.

Of course, there are constraints as to how many times input can reasonably be repeated in instructed settings. Simply increasing exposure in a scenario with high input variability is not viable in most teaching contexts. Most likely, this is where explicit input and teacher feedback would naturally come into play in pedagogical contexts. In an authentic teaching situation, teachers would perhaps explain the structural 'rule' underlying the construction at hand and they would most likely provide more direct teaching of verb semantics by repeating the individual verbs or by providing translations. It is also likely that teachers would provide feedback, for example by correcting individual students if they make mistakes. The role of explicit input and feedback has been deliberately avoided in the current thesis since it is investigating the variability effect on young learners' ability to *infer* linguistic information from context. It explores if this variability effect reported in previous more controlled research can be detected in a real classroom. However, future research in this area should undoubtedly incorporate explicit input and feedback since this more authentically reflects the reality in FL

classrooms. In fact, explicit input and feedback and high input variability might complement each other. This shall be discussed in more detail in the General Discussion, considering the wider context of this thesis (see section 5).

3.3.3 Types of learning

Based on the discussion of experiment 1 so far, one might summarise that three different types of 'learning' happened in the classrooms: (1) The first type could be referred to as *structure learning*. In a pedagogical view, this type might be considered the most sustainable and productive type of learning, most beneficial for students' communicative agency in the long term. Previous research and the data at hand suggest that increased input variability is beneficial for this first type of learning. Yet, having learned the underlying structure of the VACs does not necessarily entail verb accuracy at test with familiar items, at least in the current experiment, as indicated by the data. (2) The second type of learning could be referred to as *verb learning.* By this, I am referring to the learning of the 'pure semantic content' of individual verbs, disregarding their structural features which verbs necessarily carry (e.g., transitive verbs) in order to be used in context. For example, for the verb *slide,* what I am referring to is the actual *sliding* action, not the structural features which the verb carries to function in context (e.g., the verb's characteristic of being used intransitively (*somebody slides*) or transitively (*somebody slides towards something/somebody;* and other potential forms of usage). Due to high token frequency in the input, this type of learning seems to have taken place particularly with the LV group. (3) The third type of learning could be referred to as *'whole unit' learning*. As discussed above, the LV students outscored the HV students in familiar trials -at least numerically- across tasks getting *both* the structure and the verbs correct. Yet, it was the same students who struggled with applying the underlying VAC structure to novel verbs. This suggests that getting the structure (and the verb) correct in familiar trials was most likely not a consequence of having correctly inferred the underlying structure of the VAC, but rather of having retained the VACs as 'whole units'. As a result, the LV students could reliably *re*-produce them when prompted to do so in familiar tasks. The fact that these three types of learning were observed in the current experiment is unsurprising as they are the consequence of an input-driven approach to language learning. What they suggest, in the end, is that the structure of the input impacted the structure of what has been learned, which is an important general observation.

The observations regarding the 'three types of learning' made above need to be considered with caution since some of them are based on visual inspections of the data only (see Figure 2). The current data only yielded substantial evidence to confirm that in the production task the HV students outperformed the LV students in unfamiliar trials when 'correct' responses entailed both correct structure and correct verb semantics. Nonetheless, the observations discussed above are still important in pedagogical contexts since they lead to the important question of what it is that we want young FL learners to learn, and how to achieve that. It is important to acknowledge that learning took place in *both* input conditions. Therefore, is increasing input variability going to be beneficial for everyone? Might there be a benefit in low variability input under certain circumstances? What should students ideally take away from their limited FL instruction? I shall return to these questions in more detail in the General Discussion, considering the wider context of this thesis (see section 5).

4 Experiment 2

The second teaching intervention experiment reported in this thesis draws on the same constructionist approach to language learning and representation as experiment 1. Again, 'learning' is considered a function of input characteristics, context information, and non-language specific cognitive mechanisms (Tomasello, 2009; Wulff & Ellis, 2015). Embedded in this theoretical framework, the central hypothesis remains that input variability can impact young FL learners' generalization of structural properties of linguistic constructions in instructed settings.

The current experiment focuses on non-adjacent dependencies, adopting a similar approach like Gómez (2002). Non-adjacent dependencies are ubiquitous in sequential syntactic patterns. At the phrasal level, they describe syntactic relations that hold between non-adjacent words, such as verb agreement in the sentence *The child that ate the apples is angry.* Here, the conjugation of the auxiliary verb *is* depends on the non-adjacent subject *child*, rather than on the more 'local' -i.e., adjacent- object *apples*. And while a body of research has suggested that adults and infants alike can identify statistical regularities between adjacent elements in the input (cf. Bulf et al., 2011; Kirkham et al., 2002; Saffran et al., 1996; Saffran, 2001; Estes et al., 2007; Teinonen et al., 2009), learning non-adjacent dependencies seems to pose challenges to learners (cf. Bulf et al., 2011; Estes et al., 2007; Kirkham et al., 2002; Saffran et al., 1996; Saffran, 2001; Teinonen et al., 2009), learning non-adjacent dependencies seems to pose challenges to learners. This might be the case because humans might *by default* be relying on adjacent dependencies during learning (Gómez, 2002). In turn, this reliance on adjacent dependencies might hinder the learning of non-adjacent dependencies. Consequently, as argued by Gómez (2002) and others (e.g., Sandoval & Gómez, 2013; Wilson et al., 2020), if the reliability of adjacent dependencies is decreased (i.e., the transitional probability between adjacent elements is decreased), learners should pick up on the next, 'higher-order' reliable structure, that is, non-adjacent dependencies. This notion that aides learning of non-adjacent dependencies was confirmed by Gómez's (2002) data (discussed in section 1.6.3) which suggests that learners' default reliance on adjacent dependencies appears to switch to a reliance on non-adjacent dependencies as soon as the set size of the intervener becomes too large, and the adjacent dependencies become sufficiently unreliable. In essence, experiments like Gómez's demonstrate the effect of input variability on the learning of

invariant non-adjacent dependencies without explicit instruction. Given the ubiquity of non-adjacent dependencies in language and the fact that those structure have been the subject of prior research with children in the context of input variability, the current teaching experiment also draws on non-adjacent dependencies and aims to test the same input variability effect in an authentic classroom with young beginner FL learners. In addition to testing the learning of the underlying non-adjacent dependency (as tested in Gómez, 2002), the current intervention experiment also tests students' ability to extend this knowledge to novel contexts.

Participants were split into two groups and received in-class teaching of a target NAD over a period of two weeks with either high type frequency in the NAD's intervener slot (high variability; HV) or with high token frequency in the NAD's intervener slot (low variability; LV). Specifically, following research results such as those reported by Gómez (2002; see section 1.6.3), it was investigated whether increased variability in the target NADs intervener slot would enhance the HV students' ability to learn the underlying non-adjacent dependency. A novel aspect of the current study was the further exploration of whether input variability also affects students' capacity to generalize this abstract knowledge to novel contexts with unknown interveners. The resulting overarching research question for experiment 2 was:

> *Does input variability in German dependent clause constructions featuring a non-adjacent dependency impact primary school FL students' ability to learn the non-adjacent dependency and to generalize it to unattested contexts?*

Based on Gómez' (2002) work, a positive effect of increased input variability on students' learning of NADs was expected, even in the 'noisy' classroom environment. In addition, based on work reported elsewhere (see section 1.6.3) suggesting a positive variability effect on generalization, a positive effect on their ability to generalize NADs to novel contexts was expected as well.

In the following, the methodology of experiment 2 is presented. Since the same classes participated in experiment 1 and experiment 2, the main focus of the methodological details will be on reporting the process of generating the teaching input and details regarding the outcome measurement. In addition, the specific teaching arrangement during the intervention is presented. In the end, an overview of the final outcome measure procedure is provided. I will then move on to presenting and discussing the results.

4.1 Methodology

This classroom intervention experiment was conducted in a British primary school with two intact classes. The participants were Year 2 students learning German in their second year of study. They were the same individuals as in experiment 1. The experiment was conducted one week after the end of experiment 1 reported earlier in this thesis (section 3). Like in experiment 1, experiment 2 featured a low variability and a high variability condition. Thus, there were again two manipulated variables: input group and time. The two classes were allocated randomly at group level to either condition. Note that across the two empirical studies reported in this thesis, the random allocation of classes to the conditions was balanced, so each class was in the high variability condition once across experiments. Like in experiment 1, it is important to acknowledge that this experimental design came with limitations caused by its quasi-experimental character and its sample size which was effectively N = 2 (i.e., two classes) considering the clustering of the data at class level. Therefore, experiment 2 should also be regarded as 'proof of concept' trial providing initial insights into the potential effectiveness of the proposed treatment. It is hoped that in the future, larger randomized-controlled trials could further explore the findings of the current work with higher statistical power. In addition to the low sample size, some further limitations of experiment 2's methodology are presented in section 5.6.

Experiment 2 targeted NADs, adapted from Gómez (2002). The contact time with all participants was approximately fifteen minutes daily for two weeks which exceeds that in similar intervention studies with comparable age-levels (e.g., Hopp & Thoma, 2021).

4.1.1 Participants

Experiment 2 adhered to a quasi-experimental design as the participating school and students were selected using convenience sampling. The sampling process was the same as for experiment 1 and is described in appendix 7.1.

The participating school and students were the same as in experiment 1 (section 3.1.2). The students in the two participating Year 2 classes were between 6 to 7 years of age. Both classes had a size of 20 students which resulted in a total of 40 potential participants. Although parents and children were not asked directly, both the Headmistress and the Head of Pre-Prep Languages confirmed that they had no knowledge of any cognitive impairments or learning disabilities among their Year 2 students.

As in experiment 1, during the numerous testing periods on different days, it was unavoidable that some individual children were unavailable (e.g., sickness), although every effort was made to try and test each child and collect as much data as possible. In the end, participant numbers were 20 and 18 across both classes for the outcome measures of experiment 2. Numbers of available participants in pre-tests and during outcome measures are presented in Table 10. Of all the children who were absent during testing sessions including both pre-tests and outcome measures, only two missed numerous test sessions due to longer periods of illness. One child missed the WASI *and* the Language Magician, and another child missed the WASI, the Language Magician, *and* the outcome testing in experiment 2. Those participants were both from class 2. Since there were no outcome measure data for those children, their pre-test data were not used in the analyses.

group	N	PVST	WASI	Language Magician	Exp 2 (NADs)	group
class 1	20	20	19	19	20	LV
class 2	20	19	16	18	18	HV

Table 10 Numbers of available participants across pre-tests and outcome measures during experiment 2.

4.1.2 Ethics approval and consent

The entire research project presented in this thesis (i.e., both experiments) was approved by the Central University Research Ethics Committee (CUREC) of the University of Oxford (Ethics Approval Reference: [CIA – 22TT - 135]). The information sheets and opt-out forms which parents were provided with prior to experiment 1 covered information and parental consent for the entire project, that is, for both experiment 1 and experiment 2. Similarly, the German lesson prior to the pre-testing which we had allocated to explaining the project to the children and obtaining their consent covered information and consent on both experiments as well. As reported in section 3.1.3, no parent chose to opt-out and all children opted in to participate in the study.

Experiment 2 was piloted with the same group of Year 3 students as in experiment 1. The process of obtaining their and their parents' consent was the same as for experiment 1 (see section 3.1.3).

4.1.3 Pre-Tests

No additional pre-tests were conducted for experiment 2. See section 3.1.4 for a description of pre-tests administered for this entire research project.

4.1.4 Teaching intervention

Experiment 2 commenced in the week after the end of experiment 1. The time frame of experiment 2 was equivalent to the time frame of experiment 1 (Table 11).

	Intervention							Post-Tests		
Day	1	2	3	4	5	6	7	8	9	10
Class 1	back-up day	High Variability condition						Testing		
Class 2	back-up day	Low variability condition						Testing		

Table 11 Time frame experiment 2. Back-up days were intended for potential travel disruptions caused by industrial strike actions.

In experiment 2, adapted from Gómez (2002), the investigation focused on students' acquisition of NADs as measured by their ability to judge the accuracy of (incorrect) NADs and generalize the underlying NAD structure to new contexts. The target structure was a 'reporting event', realized by a German subordinate clause construction featuring a NAD: **that Anna in the forest eats*; in German: […] *dass Anna im Wald isst* ([…] *that Anna eats in the forest*). Usually, in German subordinate clauses like *[…] dass Anna im Wald isst*, the inflectional congruency between noun (*Anna*) and finite verb in verb-last position (*isst/eats*) constitutes the NAD. It is primarily a grammatical dependency in that the noun's grammatical characteristics such as number govern the verb inflection. However, recall that the children's German level, despite having had a year of prior German FL instruction, was basic. Therefore, it was decided that straightforward idiosyncratic constraints on a grammatical structure should constitute the NAD in the current experiment. It was expected that idiosyncratic constraints would be much more salient for the children compared to the difficult to spot inflectional non-adjacent dependencies which often feature in German grammar and mostly depend on individual letters only. Specifically, the idiosyncratic rules, that is, the non-adjacent dependencies, to be inferred were that *person A* always does *action A*, *person B* always does *action B*, and *person C* always does *action C*.[30] Thus, the dependencies were created on a semantic rather than a grammatical level. Students were not expected to pick up on any grammatical features of this input, but rather on the non-adjacent person-action associations. The dependent clause featuring this

[30] In linguistic terms, one could say that *person X* governs *action X*.

non-adjacent association, in turn, was embedded in a main clause as this is the only grammatical way to realize a dependent clause structure featuring a finite verb in verb-last position in German. The entire main clause communicated that somebody conducts a certain type of action in a certain location, reported by an external interlocutor: *[Mama sagt], dass Anna im Wald isst* (*[Mum says] that Anna eats in the forest)*. The intervening prepositional phrase (e.g., *im Wald/in the forest*) is interchangeable and represented the target variable slot in this experiment, filled with different prepositional phrases, such as *am Bahnhof (at the train station), am Strand (at the beach), am Markt (at the market).* Like Gómez (2002), three different dependency structures were used, AXD, BXE, CXF, where X was the prepositional phrase slot (e.g., *in the forest*) (see Table 12 below).

dependency structure	subject (A - C)	intervener	verb (D - F)	English translation
AXD	$Klara_A$	[prepositional phrase]$_X$	$huepft_D$	jumps
BXE	$Anna_B$	[prepositional phrase]$_X$	$isst_E$	eats
CXF	Tim_C	[prepositional phrase]$_X$	$singt_F$	sings

Table 12 Non-adjacent dependency structures in experiment 2. As in Gómez (2002), the A-, B-, C-, D-, E- and F-positions were always filled with the same items. Only the X-position varied.

In the HV condition, the intervener position was filled with 30x different prepositional phrases. In the low variability condition, only 5x of those 30x items were used. It was decided to use 30x and 5x interveners based on results of Gómez (2002), Gómez and Maye (2005) and Grunow et al. (2006) whose analyses suggested that the gap between high- and low-variability must be considerable to detect treatment-effects.

4.1.5 Materials

4.1.5.1 Input – Exposure sentence sets

Like the VAC input sentences for experiment 1 (see section 3.1.6), the NAD input sentences were prepared as part of the teaching planning with the host-teacher. The sentence structure had to follow a German dependent clause structure to generate a NAD. We were aware that this structure would exceed the children's level of German proficiency. However, we could not find an 'easier' NAD structure (which was grammatically correct), and it was important to introduce a structure that the children did not know yet. To make the learning

worthwhile for the children, sentences containing prepositional interveners were created, such as *in the forest*. Although those prepositional phrases would not have been part of the children's regular curriculum, they were considered beneficial to learn about for the children as they contained prepositions ubiquitous in the German language (i.e., *im*, *am*) and useful common nouns such as *forest*, *beach*, *train station*, or *pool*. The full list of 30 intervening prepositional phrases was compiled in collaboration with the host-teacher and is presented in appendix 7.14.

The three names, *Klara*, *Anna*, and *Tim* were chosen as they are all used regularly in both German and English, and they had not been part of any previous teaching materials to the best of the host-teacher's knowledge. In addition, no children in the participating classes were called *Klara*, *Anna*, or *Tim*. The two latter conditions ensured that none of the names carried an increased saliency for the children which might have impacted their learning process.

The three associated verbs *jump*, *eat*, and *sing* had to represent actions that were so distinctive to one another that they could be depicted unambiguously with cartoons (in video format). In addition, it was ensured that none of the verbs would be semantically or phonologically more salient than the others. For example, a verb such as *fly*, which would have been ideal for unambiguous depiction, was deemed unsuitable as it was considered more salient than verbs such as *jump* or *sing* since it would contradict the children's encyclopaedic knowledge that humans cannot fly. Regarding the verbs' phonological features, they were all monosyllabic when conjugated to 3rd person singular in German and thus had a similar length during speaking.

Input sentence generation began with a random allocation of verbs (*jump*, *eat*, *sing*) to nouns (*Klara*, *Anna*, *Tim*). This resulted in the following three combinations: *Klara jumps*, *Anna eats*, *Tim sings*. Throughout the entire exposure phase, across both the HV and LV conditions, those associations represented the NADs and did not change.

90 short GIFs (Graphics Interchange Format) were created representing all possible noun, prepositional phrase and verb combinations (i.e., three NADs x 30 different intervening prepositional phrases). Examples are provided in appendix 7.15. As the ratio of prepositional phrases between LV and HV was 5/30, five random prepositional phrases were selected and assigned to the LV condition. As shown in appendix 7.16, the students in both conditions were exposed to 15 input sentences each day. In the LV condition, the five prepositional phrases

were repeated once with each of the three NADs each day. The order of all sentences was randomized. In the HV condition, each of the thirty total prepositional phrases appeared once with each of the three NADs over the 6-day exposure period, adding up to 90 phrases in total. Each day included 15 prepositional phrases, spread across the three NADs. The order of all sentences was randomized each day. Individual prepositional phrases appeared only once each day across all three NADs.

4.1.5.2 Outcome measurements – Test sentences preparation

The test sentences featured the same three people (*Klara*, *Anna*, *Tim*) and verbs (*jumps*, *eats*, *sings*) as in the exposure phase. Students were tested on three different types of sentences: (1) familiar sentences, (2) sentences containing an incorrect dependency (e.g., **Tim* […] *jumps*), and (3) sentences featuring a novel intervener (e.g., *cinema*). Test sentence types (1) and (2) were adapted from Gómez (2002). Type (3) was introduced to test students' generalisation abilities.

During testing, participants heard a sentence and had to indicate whether they thought the sentence was correct or not (more details on procedure in section 4.1.7). No visual input was provided during testing. The five prepositional phrases used in the LV group during exposure were used as prepositional phrases in the 'familiar' sentence category across all three NADs (*3x5=15* 'familiar' sentences) since children across both input conditions were familiar with those. For 'wrong dependency' sentences, three re-arranged noun-verb associations were randomly generated: (1) *Klara eats*, (2) *Anna sings*, and (3) *Tim jumps*. Each of those associations was paired with the same five prepositional phrases that were used for the 'familiar' sentences (*3x5=15* 'wrong dependency' sentences). For 'novel intervener' sentences, the three familiar NADs were each paired with five novel interveners (*3x5=15* 'novel intervener' sentences). In total, there were 15 sentences per test category (45 in total), each category containing five sentences for each of the three NADs. To generate test sentence sets for each child, two random sentences were drawn from each of the following nine subsets: *Klara familiar, Klara wrong dependency, Klara novel intervener*; *Anna familiar*, *Anna wrong dependency*, *Anna novel intervener*; *Tim familiar*, *Tim wrong dependency*, *Tim novel intervener*. As a result, each child was tested on 18 sentences randomly selected from the larger pool of 45 total available test sentences (see appendix 7.17) for LV and HV example test sets). The number of test sentences was determined based on experiences from piloting (see appendix 7.18). The order of test sentences was randomised for each child.

4.1.6 Teaching arrangement

The teaching arrangements in experiment 2 were similar to experiment 1 and followed the same considerations regarding authenticity and controllability (see section 3.1.7). Approximately 15 minutes of teaching time were allocated to the input sentences at the beginning of each lesson. Both classes were already well acquainted with a toy bee which their teachers had been using for teaching for a couple of months. Students were introduced to the curious bee who flies around in the world and reports back to the beehive what she experienced during her flights. This arrangement enabled me to implement the German dependent clause structure naturally into the input. Specifically, the input sentences were always preceded by '*Die Biene sagt, dass Tim am Bahnhof singt*' (*The bee says that Tim sings at the station*). Prior to exposure every day, the class and I would explore the bees' daily life. For example, we would discuss how they make honey or how they communicate with each other. This was intended to keep the students engaged and did not feature any target input. The curious bee would return to the beehive in the evening and tell everyone about what she saw that day. The transitions to the NADs were particularly challenging in this context, as we could not use individual places featuring in the target dependent clauses to create engaging contexts (e.g., *pollinating flowers in the forest while Anna is there eating a honey bread*) because doing so could have rendered some prepositional phrases and actions more salient than others. To make the content not too far-fetched, we introduced the toy bee as a particularly 'annoying' bee who is liked by everyone but also tells the other bees about everything she saw during the day no matter whether the other bees want to hear about it or not. This way, we could legitimate why the annoying bee would repeat similar sentences every day. As expected by the host-teacher, the collective 'alliance' against the annoying bee who forces everyone to repeat the sentences kept engagement at a steady level throughout the exposure period. As in experiment 1, the introductions and the transitions to the NADs were always the same in both classes. After the transition, the daily exposure phase began. First, I read out an input sentence while simultaneously showing the corresponding GIF on a Whiteboard. Then, the students and I repeated the sentence together as a choir while we viewed the GIF again. After receiving praise for their good work, we moved onto the next sentence.

4.1.7 Outcome measurement procedure[31]

Outcome measurements started on day 7 of experiment 2, one day after the final input lesson. The testing set-up was the same as in experiment 1. Again, teachers were informed about the procedure and sent out children one by one. As in experiment 1, we varied children from the HV condition and the LV condition between lessons to prevent students from one input condition being tested one or two days ahead of students from the other input condition. Again, testing was completed over the course of three mornings with the vast majority of students being tested on either the first or the second day of testing. Only two students were tested on the third day due to their absence during the initial two days of testing. The testing of each student took approximately 15 minutes.

Based on the test sentence generation process described in section 4.1.5.1, a spreadsheet was prepared for each child containing all 18x test trials (6x familiar trials, 6x novel intervener trials, 6x wrong dependency trials). The order of trials was randomized on each spreadsheet. In those spreadsheets, it was recorded whether the child gave a correct or incorrect response.

Before the outcome measurements with each child started, I spent some time making small talk with each student to make them feel comfortable with the situation. Then, the testing started. Students were instructed to listen to each sentence and indicate whether they thought the sentence was correct or not. Throughout the testing process, children were verbatim asked the following question on each trial: *Is it correct that ...?* The question was then followed by the respective test sentence. This specific interrogative main clause structure, which introduces a subordinate clause, was crucial as it seamlessly paved the way for the insertion of the German subordinate clause test sentences. In addition, the semantic structure of the interrogative clause prompted children to provide a binary yes or no response. Students were told that it did not matter how they indicated agreement and disagreement, they could shake or nod their heads or say their response out loud. Throughout testing, students were not provided with any feedback regarding the accuracy of their responses. After testing, they were praised for their good work, thanked for their collaboration, and sent back to their classrooms.

[31] The outcome measurements for experiment 2 were piloted. Details are presented in appendix 7.18.

4.2 Results

Results of the pre-tests are presented elsewhere in section 3.2.1. They suggest that the children in the two participating classes performed similarly across pre-tests, meaning that there were no detectable underlying group differences regarding language-related abilities which could have had a systematic effect on the experiments' outcomes. It was also considered whether there was evidence that the pre-test results were correlated with performance in experiment 2. Plots and correlation analyses are reported in appendix 7.19. In sum, there was no clear evidence of associations between the pre-tests and performance in experiment 2. Thus, it was decided not to include pre-test results in the statistical models to avoid having unnecessarily complex models.

4.2.1 Analysis Plan

Separate analyses were conducted for each trial type of the outcome measurements (i.e., familiar, novel intervener, wrong dependency). All analyses were conducted in the R Computing Environment (R Core Team, 2021). The data and script are provided in appendix 7.9. For each trial type, a generalized Linear Mixed Model (GLMM) of the binomial family with a logit link function predicting the proportion of 'Yes' responses was run. All models included a fixed effect of condition (i.e., HV group vs. LV group). An additional independent variable for 'test day', reflecting whether the child took the test either on day 1, day 2 or day 3 after exposure, was not included since this did not generally contribute to the model. The predictors were centred such that the intercept represents the grand-mean and the fixed effect can be interpreted as a main effect. Also included were random intercepts for participants and sentence.

Statistics taken from the coefficients for the fixed effect were used to test a set of hypotheses which came from the theory that variability promotes (a) detection of correct non-adjacent dependencies and (b) generalization with unfamiliar items while high token frequency (i.e., as found in our low variability condition) promotes stronger learning with familiar items. These hypotheses and how they relate to the effects in the different models are given in Table 13.

Each coefficient that was extracted provides an estimate of the difference in question (β) and its associated standard error (in log odds space). The associated z-scores and p-values are also automatically provided, however, in the current work it was decided instead to

compute and interpret a Bayes Factor (BF) for each of these effects. Details about the reasons for using BF analyses are provided above in section 3.2.2.

Since they are more familiar to the reader, the p-values associated with the hypotheses are also reported, though they are not interpreted.

	Prediction	Motivation	Model from which coefficient statistics are extracted (Models provided in appendix 7.20)	Relevant coefficient (from which beta and SE are extracted)	Model of H1 Estimate of predicted effect size in log odds This will be set as SD of half normal with mean of 0. Odds ratio in parentheses.
1 – familiar	In familiar trials, children from the LV condition show better performance (i.e., more 'Yes' responses) than do children from the HV condition (cf. Gómez, 2002)	Higher token frequency of structures (irrespective of them featuring NADs) during exposure will lead to better learning of those structures, and give an advantage when judging the accuracy of familiar structures	*model 1*: predicting proportion of 'Yes' responses in familiar trials	input group at reference level 'familiar'	5.91 (369)[32]
2 – novel intervener	In trials featuring novel interveners, children from the HV condition show better performance (i.e., more 'Yes' responses) than do children from the LV condition	Higher type frequency of structures featuring NADs during exposure will lead to better generalization of those NADs, and give an advantage when judging the accuracy of NADs featuring novel interveners	*model 2*: predicting proportion of 'Yes' responses in trials featuring novel interveners	input group at reference level 'novel intervener'	0.71 (2.03)[33]
3 – wrong dependency	In trials featuring wrong dependencies, children from the HV condition show better performance (i.e., fewer 'Yes' responses) than do	Higher type frequency of NADs during exposure will increase participants' focus on invariant structures and thus promote better learning of those NADs (cf. Gómez, 2002),	*model 3*: predicting proportion of 'Yes' responses in trials featuring wrong dependencies	input group at reference level 'wrong dependency'	1.55 (4.70)[34]

[32] No relevant (significant at $p < 0.05$) value from previous research was available. Thus, the beta value from the 'entirely correct' analyses of the familiar act out comprehension task trials from experiment 1 (see section 3.2.5) was used as prior.

[33] No relevant (significant at $p < 0.05$) value from previous research was available. Therefore, the mean difference in 'Yes' response scores from the current experiment's pilot were used to calculate beta (i.e., beta = logodds(0.67) – logodds(0.50). The direction of the subtraction was guided by the direction of the hypothesis that HV provides more 'Yes' responses than LV in trials featuring novel interveners.

[34] No relevant (significant at $p < 0.05$) value from previous research was available. Thus, average proportions of 'Yes' responses in 'untrained' (i.e., wrong dependency) trials taken from Table 2 in Gómez (2002) were used to calculate beta (i.e., beta = logodds(mean percentage 'Yes' response LV) – logodds(mean percentage 'Yes' response HV). Values for HV were taken from Gómez' set size = 24, and values for LV were taken from her set size = 6. This proportion between HV and LV (i.e., 24/6) was closest to the set sizes used in the current experiment (30/5). The direction of the subtraction was guided by the direction of the hypothesis that LV provides more 'Yes' responses than HV in trials featuring wrong dependencies. Note that the data in Table 2 by Gómez (2002) is from young adults (undergraduate students). Data from 18 months-old infants (experiment 2 in Gómez, 2002) could not be used since only mean listening time data is reported.

	children from the LV condition (cf. Gómez, 2002; Table 2).	leading to an advantage when judging the accuracy of wrong dependencies.	

Table 13 Contrasts to be tested. Presented are the predictions that were tested, the motivation for testing this prediction, and the estimated effect size for each prediction. The logistic mixed effects model, from which the summary data originated, is specified for each hypothesis tested. Table adapted from Brekelmans et al. (2022).

4.2.2 Descriptive Data

The outcome measure tasks yielded binary data and were coded with '0' for 'No' and '1' for 'Yes' responses. As testing was conducted in person at school and video recordings were not part of this study, all coding across all three tasks was completed by the main researcher. A visual overview of participants' mean response scores is provided in Figure 4.

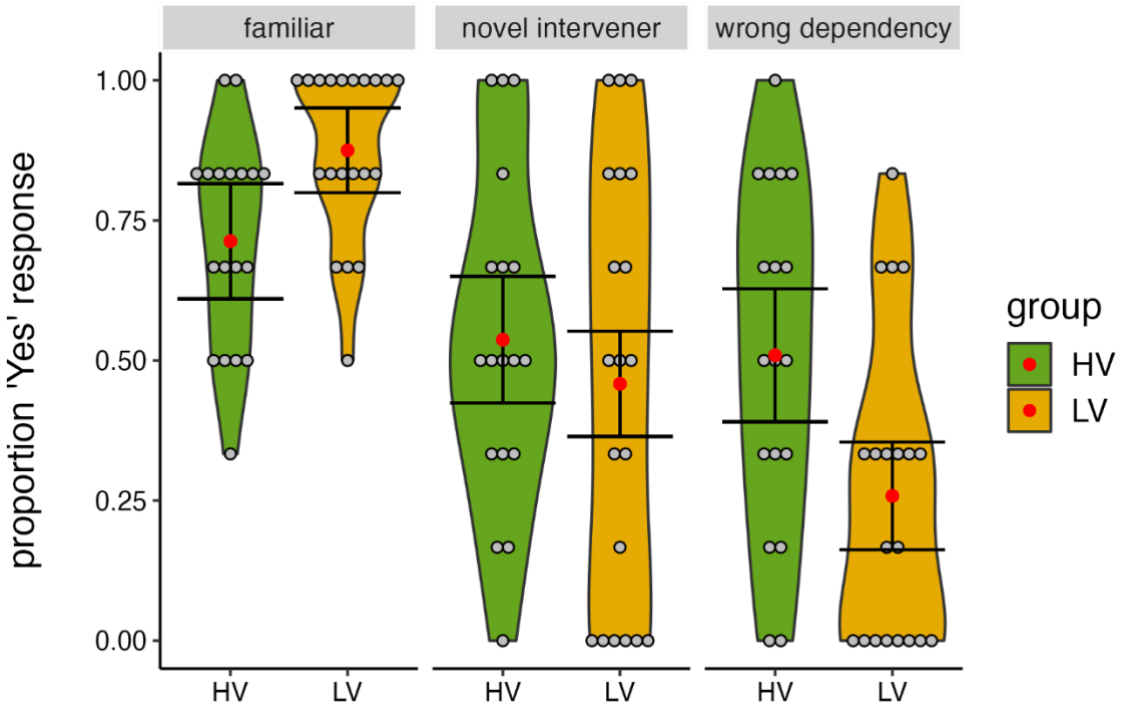


Figure 4 Violin plots displaying participant mean proportion 'Yes' responses during NADs (experiment 2) outcome measurements. HV = high variability. LV = low variability. Error bars present 95% Confidence Intervals.

In Figure 4, note that a higher proportion of 'Yes' responses indicates stronger performance with familiar and novel intervener items, but weaker performance with wrong dependency items. Similar to experiment 1, the plots indicate that the LV group outperformed the HV group in familiar trials. However, both groups seem to perform similarly in trials featuring novel interveners. In trials featuring wrong dependencies, the LV group seems to provide fewer 'Yes' responses than the HV group – indicating stronger learning.

4.2.3 Statistical Analyses

Each child was tested on each of the three trial types. Order of trials was randomized for each child. Recall that the following main analyses are conducted over 'Yes' responses in each trial type.

Only statistics relevant for testing the hypotheses are reported. Individual full models are provided in appendix 7.20. Below, analyses on each trial type separately are reported first.

4.2.3.1 Familiar trials

Recall that Figure 4 in line with the hypothesis indicates that the LV group outperformed the HV group during testing with familiar items. This observation was tested statistically by testing evidence for the hypothesis that there would be better performance in the LV group than in the HV group (i.e., more 'Yes' responses) for familiar trials. Evidence was substantial with BF = 9.01 (predicted effect = 5.91, RR = [0.16; 18.04], $\beta = 1.09$, $z = -2.93$, $p < 0.05$).

4.2.3.2 Novel intervener trials

Figure 4 suggests that there was no considerable between-group difference in performance in 'novel intervener' trials. The initial hypothesis that there would be stronger performance in the HV group than in the LV group (i.e., more 'Yes' responses) was tested which resulted in ambiguous evidence with BF = 1.29 (predicted effect = 0.71, RR = [0; 5.76], $\beta = 0.65$, $z = 0.91$, $p > 0.05$).

4.2.3.3 Wrong dependency trials

Contrary to the initial hypothesis that the HV group would provide fewer 'Yes' responses than the LV group in trials featuring wrong dependencies (cf. Gómez, 2002), Figure 4 indicates that the HV group provided *more* 'Yes' responses than the LV group.[35] Evidence for the *initial* hypothesis was tested which resulted in substantial evidence for H0 that the HV group does *not* provide fewer 'Yes' responses than the LV group in trials featuring wrong dependencies with BF = 0.10 (predicted effect = 1.55, RR = [0.49; ∞], $\beta = -1.69$, $z = 2.74$, $p < 0.05$).

Given that the means trended in the opposite direction to the prediction, it was also tested if the data provide evidence for the alternative one-tailed prediction that the LV group shows stronger performance (i.e., fewer 'Yes' responses) in wrong dependency trials. There was no value from previous research to base the estimate of the prior on, therefore, the motivated-maximum approach -calculated over rank scores- was used (Silvey et al., under review). Details of the calculation of the estimate of the prior are provided in appendix 7.21.

[35] Recall that a 'Yes' response in this scenario is technically an inaccurate response in the context of the NAD. See more detailed discussion of this issue in section 4.3.5.

Testing evidence for the hypothesis that there would be *more* 'Yes' responses in the HV group than in the LV group in trials featuring wrong dependencies resulted in substantial evidence with BF = 5.49 (predicted effect = 9.5, RR = [0.28; 17.62], $\beta = 1.69$, $z = 2.74$, $p < 0.05$).

4.2.4 Summary

Regarding response behaviour in familiar trials, the LV group outscored the HV group on average. This was confirmed statistically with substantial evidence for an effect of token frequency on the learning of familiar structures.

In trials featuring novel interveners, the data did not support the initial hypothesis of a positive effect of increased type frequency (i.e., variability) on the generalization of the underlying NAD-structure to novel contexts.

Contrary to the initial hypothesis, the analyses of response behaviours in trials featuring wrong dependencies yielded substantial evidence for the hypothesis that the HV group provided *more* 'Yes' responses than the LV group in trials featuring wrong dependencies, apparently suggesting a detrimental effect of increased input variability on the learning of NADs.

4.3 Discussion

The results in experiment 2, specifically data from trials featuring novel interveners, did not indicate that students' ability to generalize the underlying NAD structure to novel intervener verbs was impacted by the input condition. In fact, their response behaviour in those trials was not different from chance.[36] Turning from the ability to generalise information to novel contexts to the 'basic' ability to distinguish a correct NAD from an incorrect NAD, data from the wrong dependency trials (resembling Gómez' (2002) 'untrained' trials) provided substantial evidence that the LV group was better at rejecting incorrect NADs compared to the HV group. This is surprising and stands in contrast to results reported elsewhere (e.g., Gómez, 2002; Gómez & Maye, 2005). Regarding the familiar trials, there was substantial evidence that the LV group outperformed the HV group.

In the following Discussion, I will begin by exploring reasons for the unexpected results in the wrong dependency trials. I will put an emphasis on the beneficial impact of the visual and semantic context in the current experiment, particularly so for the LV group. I will then shortly turn to familiar trials, before zooming in on the HV group's struggles in wrong dependency and novel intervener trials. In the end, I will present several accounts as to why we seem to see some *learning* but no *generalization* in the current experiment. To conclude the discussion, I will contextualize the results of the present experiment within the findings of previous works.

4.3.1 Unexpected results in wrong dependency trials

Recall from sections 1.6.3 and 4 that there is a widespread notion in the literature regarding the competition between adjacent and non-adjacent dependencies, namely, that a decrease in the reliability of lower order dependencies (i.e., adjacent) can support learning of higher-order dependencies (i.e., non-adjacent) (for an overview of research on NADs, see Sandoval & Gómez, 2013). This also guided the hypothesising in the current study, where HV participants were expected to be better at picking up on the underlying NAD structure and thus outperform LV participants in wrong dependency trials. However, the opposite pattern was observed in the current study. Why?

[36] Additional statistical analyses to compare performance in the novel intervener task to chance-level were conducted. There was substantial evidence for H0 that performance in the novel intervener task was not different from chance with BF = 0.06 (predicted effect = 5.91, RR = [1.2; ∞], β = -0.01, z = -0.04, p = 0.97). For more information, please refer to the analysis script in Appendix 7.9.

One possible explanation for this pattern of results might be that the HV condition in the current study was simply too 'difficult'. After all, students in the HV condition were exposed to 30 intervening prepositional phrases, whereas the LV group was only exposed to 5 interveners. It might be the case that this relatively high number of 30 interveners 'stretched' the variability effect to such a degree where it did not make a difference anymore. As a result, perhaps the input was too confusing. However, the same phenomenon could be explained from the opposite angle where the 'confusion' is likely to lead to better learning of the NAD. In fact, one could argue that the high amount of 30 intervener phrases makes the reliability of the lower-order dependencies in the input structure so vanishingly small, that the NAD becomes extremely salient and thus easily picked up. Another factor that speaks against the notion of the input being too overwhelming for the HV group is the fact that the HV group did reasonably well in familiar trials. This performance in familiar trials seems to suggest that at least to some extent the HV group may have picked up on the underlying NAD, and that the input was not generally too difficult for them to process. However, note that they could also complete this task without picking up on the NAD by knowing the whole phrase – although this would be more difficult to achieve given the high number of types in the input. In addition, there might also be a bias to providing 'Yes' responses.

Interestingly, the LV group seems to have been more 'confident' in their knowledge about the NAD -or about the whole phrase- as indicated by the 'wrong dependency' data. When confronted with wrong dependency trials, the LV group on average demonstrates a solid ability to correctly judge a wrong dependency to be in fact 'wrong'. In comparison, the HV group, although potentially having demonstrated reasonable knowledge of the underlying NAD -or about the whole phrase- in the familiar trials, shows a wide spread of mean responses when prompted to judge an incorrect NAD. This is surprising when we consider that precisely the NAD was the only reliable and thus an extremely prominent structure in the HV input. In sum, the input most likely was not too difficult for the HV group and learning seems to have happened in both groups. Yet, the LV group outperformed the HV group in wrong dependency trials. Why?

4.3.1.1 Visual support

Of course, the LV participants could have simply based their responses on having memorized the 15 different phrases (see section 4.1.5.1) from the input. Yet, this seems like a difficult task given the students' generally low German proficiency and the fact that the

intervener phrases must have sounded alien to them. Let us take a step back and consider what other types of knowledge the LV participants might have also drawn upon to judge sentences featuring wrong dependencies so successfully. Apart from potentially having memorized whole phrases, what else could they have learned that the HV group apparently did not learn sufficiently in comparison? In essence, all participants had to learn was the pattern that *person A* was exclusively associated with *action A*, *person B* was exclusively associated with *action B*, and *person C* was exclusively associated with *action C* - no matter in which of the five locations the respective action took place. Specifically, to be successful during testing -which was done aurally only-, participants had to learn that the sound /tˈɪm/ was always associated with the sound /zˈɪŋt/, the sound /ˈanɑː/ was always associated with the sound /ˈɪst/, and the sound /klˈɑrɑː/ was always associated with the sound /hˈypft/.

In light of this, I argue that it might be the case that the additional visual input during exposure is a decisive factor in accounting for the unexpected results. In fact, while students would have had to *infer* the NAD's *syntactic structure* from the aural input (since they were not provided with explicit input regarding the structure), they did not have to *infer* the NAD's *semantic realisation* from the visual input. The pattern was simply obvious from the visual input (see example videos in appendix 7.15). For example, *Tim*'s constant *singing* action is unambiguously associated with *Tim* because participants always saw the person *Tim* at the same time while that person was also *singing*. Thus, based on the visuals, there is no underlying structure to be *inferred*. Instead, it is straightforward to understand that it is *Tim who sings* (and *Anna who eats*, and *Klara who jumps*). There is not even a need to memorise the whole phrase. Consequently, it seems like the LV students had an intrinsic advantage compared to the HV students regarding their learning of the NAD structure: While both groups received visual input (which most likely strongly supported their learning of the association between *person* and *action*), only the LV group received less variable input which potentially decreased their attention to the intervener phrases (and to the actual visible intervener *locations* in the visual input). In essence, the LV group encountered less visual and phonological distraction between the *person* sound and the *action* sound. This and the fact that the visual input made the association between *person* and *action* already abundantly clear might have helped the LV group to really solidify the associations between individual *persons* and *actions*, both visually and phonologically. As a result, when asked to judge sentences where this association was disrupted, they performed well, even when visual support was absent.

While the visual support might have played a decisive role in learning the *person* and *action* associations, note that the visuals do not entail any information regarding the fact that these associations are generalisable beyond the intervening prepositional phrases (i.e., visual locations) featuring in the input. Thus, while the visual input might have helped to learn the NAD itself, it does not provide additional support for the assumption that this NAD is generalisable to novel contexts. In conjunction with the fact that the LV group received only low variability input which provided little cues as to the NAD's generalizability, this might be the reason why we do not see strong performance by the LV group in the novel intervener trials, despite their strong performance in wrong dependency trials.

4.3.2 Familiar trials

Regarding familiar trials, when we consider the additional semantic information which both groups received through the visual input, the pattern of relatively strong performance in both groups as well as the better performance by the LV group makes sense: The LV group received little input variability (i.e., only 15 different sentences in fact) and strong visual support of the *person* and *action* association. As a result, they performed well on familiar sentences. On the other hand, the HV group received strong visual support as well, yet they encountered many more tokens of the NAD structure (90 different sentences in total) which might have negatively impacted their 'whole unit' knowledge, thus leading to slightly worse (but still mostly accurate) response behaviour in familiar trials.

4.3.3 HV group – Wrong Dependency

Let us take a closer look at the HV group's performance in the outcome measurement. In their case, the 'wrong dependency' data suggest that they did not quite learn the NAD structure. In fact, it seems like they did not fully pick up on the *person* and *action* association, despite the helpful visual input. As discussed above, it might be the case that the large number of intervener phrases (and thus the large number of visual locations where the persons and actions took place) in the input redirected a considerable part of the HV students' attention away from the *action* plus *person* association resembling the NAD. Instead, they might have been distracted by exploring the different visual 'locations' which were indeed prepared in a child-friendly and engaging way.

In addition, it has been argued elsewhere that if the transitional probabilities among adjacent dependencies are low (as in the HV condition), the learners' focus shifts to the next

higher-order reliable dependency, in the current case represented by the NAD (e.g., Sandoval and Gómez, 2013). Recall that the current experimental hypotheses were based on this notion. However, the visuals pre-empt the need to infer the NAD since at least its semantic realisation was simply *visible*. As a result, it is not quite clear what the HV students really focused on during exposure. Potentially, they had the more 'interesting' input (because it featured more locations) and they focused more on the locations in the visuals and the corresponding prepositional phrases, thereby paying somewhat less attention to the sound-pattern of, for example, */tˈɪm […] zˈɪŋt/* [*Tim […] singt; Tim […] is singing*]. As a result, when presented with wrong dependencies and -importantly- no supporting visuals during testing, it might be that they struggled to identify the wrong dependencies solely based on aural input. Thus, in the wider context of the current experimental design where visual input was present as is often the case in schools, high input variability might potentially have had a detrimental effect on NAD learning. This is explored further in the General Discussion (section 5).

4.3.4 HV group – Novel Intervener

The HV group's performance in the novel intervener trials seem like a direct consequence of what has been discussed above regarding the wrong dependency trials. On the one hand, the increased variability in the intervener slot might well have provided HV participants with stronger evidence as to the context-independence of the NAD structure. On the other hand, they seem to have failed to learn the (phonological) pattern of the NAD from the visual and aural input during exposure as indicated by their performance in the wrong dependency trials. As a result, it seems plausible that they were also not able to generalise this apparently not fully learned structure to novel contexts.

4.3.5 No generalization – despite learning?

It is indeed surprising that students across both groups seem to have struggled in the 'novel intervener' condition even though at least some learning seems to have taken place in both groups. In fact, I will lay out a tentative argument as to why it should have been quite *easy* for them to master the novel intervener trials once they had learned the NAD (as demonstrated in the other tasks). Recall that Gómez (2002) used an *artificial* language which provided learners with no semantic cues about the structure and productivity of the pattern. In contrast, the current study, by virtue of using *natural* language input, did to some extent provide learners with both structural and semantic cues. (In addition, there was visual input which

corresponded to the semantics of the natural language input.) On the one hand, the structural cues were expressed through the NADs' repetitive structure across input items (i.e., Subject + [Prepositional Phrase] + Verb), similar to previous works. On the other hand, the semantic context itself (i.e., describing that a person is performing an action in a certain location) -which has unambiguously been demonstrated by the visual input regardless of potential comprehension issues- provides learners with the basic information that the structure can be used in other contexts as well. This is simply because learners *know* from their encyclopaedic knowledge that one can *jump/sing/eat* in locations other than those presented in the (visual) input. There is no reason to expect why learners should assume the opposite. In fact, the three target verbs can be applied in all 'locations' provided throughout the experiment, both during the exposure and the test phase.[37] So, while participants in the current study were exposed to approximately the same ratio of high and low variability input across conditions compared to earlier works, the task of inferring that the target structure can be used relatively context-independently was guided by both structural and semantic rather than exclusively structural criteria. The obvious consequence of this reasoning would be that both groups, perhaps the HV group even more so than the LV group due to a variability effect, should have no problem to extend the NAD structure to novel interveners. Yet, the data suggests otherwise. Thus, if we assume (a) that a reasonable number of students had picked up on the NAD structure (perhaps mainly through visual input) and (b) that there are no obvious semantic restrictions regarding the productivity of the NAD, there must have been other factors at play blocking the productive generalization of the NAD structure to novel contexts in the novel intervener trials.

One important factor that might have impacted response behaviour in the novel intervener trials leading to these unexpected results relates to how students were *prompted* to respond. Recall that students in both groups were always asked (verbatim): *Is it correct that [test sentence]?*. This question is ambiguous: Does it target the NAD structure only or does it target the entire sentence, including the intervener? Depending on how participants interpret the question, they can provide two types of 'correct' responses. In the former case targeting the NAD structure only (which was the intended meaning), a 'Yes' response to sentences

[37] This means that there were no semantically anomalous sentences such as "[…] Klara im Ozean hüpft" (*[…] Klara in the ocean jumps*"). Such a sentence would have worked structurally in German, but not necessarily semantically. (In German, the contracted *im* preposition (i.e., *inside of*) is a locative preposition, unlike English where it could also be interpreted directionally (i.e., *into*)).

featuring a novel intervener is correct. In the latter case targeting the entire sentence, a 'No' response to a sentence featuring a novel intervener is correct since the sentence does not match the sentences from the input during exposure. For example, participants might realize that *Tim* did never *sing* in *the museum* during exposure. (Instead, he might have sung elsewhere.) As a result, participants reject the sentence although they might well be aware that the NAD is still correct. With this in mind, perhaps it is the participants' interpretation of the prompt that contributed to the wide spread of responses in the novel intervener trials.

Another factor that might account for the results in the novel intervener trials is the switch from the *aural plus visual* exposure mode to the *aural-only* test mode. For familiar sentences, this switch might be unproblematic precisely because the sentences are familiar and the phonological association between *person* and *action* is uninterrupted. For wrong dependency trials, the switch might be problematic, yet the ubiquitous and input-prominent *person* and *action* sound patterns are so blatantly disrupted (and not concealed by novel interveners) that they might be relatively easy to pick up on, even in the aural-only mode. Regarding the novel intervener trials, recall that the participants had limited German FL knowledge as they were beginner learners. Most prepositional phrases, if they had been provided without visual context during exposure, would have made no sense to many students. Now consider that in the novel intervener trials, the participants were suddenly exposed to several unknown prepositional phrases without any visual context. This might have simply led to confusion. In turn, this might be another contributing factor as to why the performance pattern in the novel intervener trials is so unclear.

In sum, the novel intervener data reflect a considerable spread in mean response behaviour across both groups (see Figure 4). Some participants gave only 'Yes' or only 'No' responses to sentences featuring novel interveners, whereas others constantly changed their response behaviour during testing, resulting in mid-range mean responses (around 0.5). Although I presented various explanations for the ambiguous pattern observed in the data, the evidence is not strong enough to draw clear conclusions about students' ability to generalize the NAD structure to novel contexts at this stage.

4.3.6 The variability effect and non-adjacent dependencies

Based on the data at hand, it is difficult to draw conclusions regarding the impact of input variability on the NAD learning process, and on the ability to generalise this knowledge to novel contexts. In response to the preceding discussion on potential reasons for this outcome, one

might argue that while some linguistic structures benefit from the variability effect, others may not to the same extent. However, in more 'controlled' studies, if observed, a variability effect is typically observed regardless of the target structure (see section 1.6.3). The variability effect appears to be relatively robust. It is important to note, however, that in the noisy context of a classroom, there is no other evidence available featuring a wider range of linguistic structures since the current study is the first of its kind. Nonetheless, I suggest that the variability effect might indeed vary across target linguistic structures, and that these potential differences are most likely mediated by other context variables particularly influential in real-world classroom settings, such as the provision of visual context.

In the study at hand, the widely discussed competition-mechanism between adjacent and non-adjacent dependencies might have been diluted by the introduction of semantic context through natural language input, and the addition of visual context (section 4.3.1.1). Of course, it might well have been the case that the LV group relied more on adjacent dependencies compared to the HV group, yet the results, particularly in the wrong dependency trials, do not point in this direction. And in any case, keep in mind that those results are most likely impacted by the presence of semantics and visual context. Somewhat akin to the chicken-and-egg scenario, there remains the circular question of whether the observed results are actual evidence against the notion that there is a competition between adjacent and non-adjacent dependencies -which would be interesting news-, or whether they are the consequence of the experimental context. I argue that the latter is more likely to be the case since both the 'familiar' results (i.e., where HV *should* have performed worse) as well as the 'wrong dependency' results (i.e., where LV *should* have performed worse) point towards the important role of semantics and visual support in the current experiment. As a result, discussing the influence of input variability on the competition between adjacent and non-adjacent dependencies, based on the available data, would remain speculation due to the what are -in retrospect- flaws in the design. Therefore, the discussion ends here. Note that the issues arising when *natural* language input (as is always the case in FL classrooms) meets artificially introduced input variability shall be explored further in the General Discussion (section 5).

5 General Discussion

The empirical work reported in this thesis aimed to shed some initial light on the role of input variability in the context of MWC input in primary school students‘ FL development. Specifically, the two reported teaching intervention experiments aimed to test the effect of input variability on children’s ability to generalize structural information of MWCs to novel contexts, thereby enhancing their communicative agency. This final chapter puts the main findings of this thesis in a broader theoretical and educational context. Limitations, including methodological flaws, of the current work are discussed as well, and avenues for future research are identified.

Broadly, the two teaching intervention experiments reported in this thesis tested whether increased input variability in a verb-argument-construction’s verb slot (experiment 1; modelled on Wonnacott et al., 2012) and in a non-adjacent-dependency’s intervener slot (experiment 2; modelled on Gómez, 2002) impacts young beginner FL students’ ability to infer the MWC’s underlying structural information and extend this knowledge (i.e., generalize) to novel contexts. Existing research with young learners in this area has primarily focused on how input characteristics influence learning conceptually within a usage-based constructionist approach to language learning, rather than on direct pedagogical applications. As a result, such research has typically been conducted in ‘controlled’ settings (i.e., not in FL classrooms) and with artificial language input (e.g., Casenhiser & Goldberg, 2005; Gómez, 2002; Gómez & Maye, 2005; Wonnacott et al., 2012). Based on results of a systematic review of the research area discussed in Chapter 2 (Schulz et al., 2023) and to the best of the author’s knowledge, the current work can be considered the first ‘proof of concept’ teaching intervention study featuring both control and experimental groups to report data on targeted and manipulated MWC input in an instructed primary school FL setting.

At the most fundamental level, the results showed that across both experiments and across all conditions, learning took place. Mappings between form and function developed rapidly, in fact, after only six sets of exposure in extremely ‘distracting’ input settings in an authentic classroom. Like in Wonnacott et al. (2012), the development of those mappings followed the quantity and structure of the input. Specifically, a positive effect of increased input variability on young FL students’ ability to not only infer but also *extend* the inferred structural knowledge to novel verbs was detected in experiment 1 featuring verb-argument-

constructions. Somewhat surprisingly, experiment 2, featuring non-adjacent-dependencies, failed to provide enough evidence to detect this variability effect on the ability to generalise inferred structures to novel contexts. Perhaps, this was due to a flaw in the experimental design which I touched on in section 4.3.5 and shall return to later.

In the following, I will start by discussing the simple yet important observation that learning took place in an authentic context with authentic input in the current study. Then, having based most parts of this thesis on the benefit of high input variability, I will consider the usefulness of low input variability against a pedagogical background. I will also turn to the role of visual and of explicit input in construction learning in early FL contexts. In subsequent sections, I will sketch out desirable future research, highlight limitations of the current study, and present some pedagogical implications that could be drawn from the current work.

5.1 Learning in an authentic setting

As mentioned above, a key point to draw from the current findings (and one that has rightly been highlighted by colleagues at conferences and other opportunities where the current work was presented) is that students in both experiments and across conditions learned *something* about the target structures in the input without receiving any explicit information about those structures. Even in experiment 2, where the evidence remains mostly ambiguous as to the effect of the intervention, the two groups' overwhelmingly correct performance on familiar trials suggests that they were at least able to recognize the items which they had encountered in the input. Despite the current works' focus on a targeted input manipulation, specifically on variability, the fact that such learning took place regardless of the input condition is a crucial observation. It does not only underline the perhaps obvious notion that input is key for any kind of progress in language learning. Rather, it also corroborates what curricula have begun to touch on recently (see section 1.2), namely, that young beginner FL learners can pick up on 'large' linguistic units in the input and at the very least recognize or imitate those units already after a relatively small amount of exposure. Essentially, although with varying degrees of input variability, the current experiments simulated situations of input flooding of specific target MWCs with little but targeted input at an intense dose. This successfully led to learning. Importantly, in comparison to other more controlled experimental studies (see section 1.6.3), the current experiments used natural language input for those floods and the children had to navigate input that featured authentic -and potentially challenging- linguistic characteristics, such as German phonological characteristics. This is a main strength of the current work compared to earlier studies which investigated input variability using artificial languages that featured characteristics which matched the L1 English in many aspects, such as the target items' morphological and phonological characteristics (e.g., Wonnacott et al., 2012). Despite the potential learning obstacles which the children faced, caused by the classroom context and the authenticity of the natural language FL input, the current experiments show with improved ecological validity that nonetheless learning took place across input conditions. In addition, at least in experiment 1, there was substantial evidence for the variability effect on the learning and generalization of underlying structural patterns. Therefore, the current experiments contribute valuable data derived from authentic

language input in an authentic learning situation, notwithstanding the inherent limitations of such a context, such as low statistical power (see section 5.6).

5.2 The usefulness of low input variability

The input variability context of the current thesis might give the impression that increased variability is a panacea for language learning and that input should necessarily be structured in a way that gives preference to high type frequency compared to token frequency. The data at hand, along with the findings of related studies, suggest that this might not necessarily be the case. In the following, I discuss why the students in the LV conditions learned what they learned, and I explore, from a pedagogical perspective, reasons why low input variability might in fact sometimes be the preferred mode of instruction.

When we disregard for a moment the 'higher order' ability to *extend* inferred knowledge to novel contexts, we can see that particularly the experiment 2 data suggest that students in the LV condition tended to better *recognize* and *imitate* familiar structures compared to students in the HV condition (see section 3.2.5). The 'familiar verb' data from experiment 1, to some extent, tended in this direction as well (see section 3.2.5). As discussed in section 4.3, these patterns are likely the result of high token frequency in the input -at the expense of high type frequency- which led to better learning of the fewer remaining types. In a usage-based constructionist perspective, this make sense: The low variability in the input provided students with little or no cues as to the extendibility of the structural frame they were presented with in the input. As a result, they might have inferred that the verb slot in the verb-argument-construction or the intervener slot in the non-adjacent dependency is (relatively) inflexible, and thus failed to learn that "the construction can be disassociated from a particular verb" or prepositional phrase (Wonnacott et al., 2012: 475). The LV students are not to blame; given the 'inflexibility' of these slot, there was not much to be inferred for the LV group in terms of the extendibility of these slots because the input did not feature the relevant information that might have pointed students in that direction. And while it would in theory be possible to extract the relevant information from only one recurring type in the input, it would constitute an uninformed guess and speak against the 'evidence' provided by the input as a whole. From this type of low variability input, it appears relatively straightforward to simply rote-memorize the MWC, without any higher order inferencing processes. One could consider this 'whole unit' learning of 'low variability' structures to be analogous to the acquisition of fixed phrases and routines that are already prevalent in primary FL classrooms and were the focus of investigation in the systematic review reported in section 2.

In this context, researchers might still argue -perhaps rightly so- that this rote-memorization seems to mirror what others have labelled an *imitative plateau* (Engel et al., 2009). One might question to what extent students benefit from being able to simply imitate the input upon receiving repetitive input with low variability. I argue that there are two main reasons why input with low variability is valuable too and, in some cases, might perhaps constitute the more suitable pedagogical option.

5.2.1 Individual differences

The first reason concerns students' FL proficiency, or more broadly, their individual differences. Note that this notion was not specifically targeted in the current study. Also, recall that correlations between outcome measurements and the English and German pre-tests as well as the WASI matrix reasoning subtest in the current study were negligible (see section 3.2.1). Yet, there is some evidence elsewhere in the literature that factors such as FL proficiency moderate the extent to which FL students are able to extract structural information from MWC input. For example, as discussed in more detail in section 2.2.1.2, Kostka (2020) reports that weaker students (i.e., lower English FL proficiency) in her study relied longer on whole, unanalysed structures which they kept imitating directly from the input. Although this is precisely what constitutes the aforementioned *imitative plateau*, importantly, the weaker students' communicative scope was *still* slightly increased by virtue of imitating pre-fabricated MWCs from the input. On the other hand, stronger students were able to gradually extract structural information from the MWCs in the input and extend this knowledge to novel contexts (i.e., *paradigmatic variations* in Kostka, 2020), thereby increasing their independent communicative agency – admittedly, more so than the weaker students. Yet, from a pedagogical view, the MWCs in this case fulfil *both* of their crucial functions at two different student proficiency levels: On the one hand, MWCs provide all students, but particularly weaker ones, with a (long-term) means to reach a basic degree of communicative agency over and above the repetition of single words. As rightly highlighted by curricula, MWCs, even if they are only imitated, play an important role in equipping beginner students with basic communicative agency. On the other hand, MWCs serve stronger students as sources for structural knowledge extraction and as a basis for increased independent communicative agency. Importantly, as rightly noted by Kostka (2020), given the unavoidably heterogeneous character of any primary school classroom in terms of FL proficiency and other individual

differences, both types of learning progress are crucial when we want to bear the development of *all* students in mind.

One issue with the observation that apparently 'weaker' students rely longer on unanalysed MWCs than 'stronger' students, as suggested in Kostka (2020), is that it remains somewhat unclear what made a student 'weak' or 'strong'. In Kostka's work, FL proficiency was based on teacher reports, teacher interviews, and the researcher's own classroom observations (cf. Kostka, 2020: 318). Yet, her study was conducted with primary school students at the beginning of their first year of English as a Foreign Language instruction. Thus, it remains questionable if FL proficiency was a sensible measurement of individual differences in the first place since, in theory, the students' proficiency should be close to zero. Of course, the students' experiences with extramural English input should not be neglected (e.g., Sylvén & Sundquist, 2016), and some students might in fact already have some knowledge of English. Nevertheless, Kostka's (2020) categorization of her students into weak and strong students remains somewhat unclear. However, this does not dismiss the possibility that other language-related individual differences, such as L1 ability, or cognitive abilities such as working memory capacity, or metalinguistic awareness, could significantly influence individual students' ability to extract linguistic information from the input without explicit instruction. Undoubtedly, it is likely that such factors were at play in Kostka's (2020) research and the current work as well. In fact, a plethora of previous research has shown these abilities to be predictors of FL development (L1 ability's impact on FL development: Duran-Karaoz & Tavakoli, 2020; Pae, 2019; Sparks et al., 2023; van Koert et al., 2023; for overviews of working memory's impact on FL development: In'nami, 2022; Mitchell et al., 2015; for an overview of metalinguistic awareness's impact on FL development: Roehr-Brackin, 2018). For example, in the context of the current work, it would be reasonable to assume that children with higher working memory capacity might find it easier to divert some of their attention to the relationship between a MWC's structure and its usage context (i.e., form-meaning mapping) and 'unpick' the MWC, while students who are weaker in this respect might struggle to do so and resort to rote memorization. Exploring the impact of such individual differences on the effect of MWC input on early FL development constitutes an interesting avenue for future research.

In general, in the context of MWC input manipulations via variability, it seems likely that 'weaker' students (for example in terms of working memory capacity) might in certain situations benefit more from low input variability compared to high input variability, potentially

rendering low input variability the preferred pedagogical choice. In essence, similar to various other types of input encountered across different school subjects, it appears unlikely that a one-size-fits-all approach to input variability would be equally effective for all students. This issue has been highlighted in the context of similar classroom studies with older FL learners as well (cf. Madlener-Charpentier, 2015, 2016; Henk, 2019). Instead, it seems more plausible that stratifying the level of input variability based on students' individual characteristics is essential to ensure maximal benefit from MWC input for everyone, and in some situations, low variability might well be the preferred pedagogical choice. Note that even in the current study, some HV students failed entirely to pick up on the MWCs and to extend the underlying structure to novel contexts (see Figures 2 and 4). While this might have been caused by chance or by unsystematic contextual factors inherent to the testing situation (e.g., distractions in the hallway during testing), there might also be systematic underlying factors related to individual differences that rendered the HV input simply too challenging for some students, effectively resulting in no learning progress.

5.2.2 Learning goals

The second reason why input featuring low variability might sometimes be preferred over high variability input is not necessarily related to the students' abilities but rather to the specific learning goals in a given teaching situation. Although this might seem obvious, at times, the primary aim of a lesson could be to facilitate rote-memorization of a particular structure, without directing attention to any underlying linguistic information. For example, this might be the case in situations where structures that are likely to exceed even the strongest students' FL proficiency are required to introduce a new topic area. In such a situation, it is not necessary for students to understand the underlying structure but rather it might be desirable to simply have everyone memorise the MWC as quickly and with as little distraction as possible. For the sake of maintaining a smooth lesson flow, it might sometimes even be desirable to actively avoid having students 'unpick' the MWC to prevent confusion over structures they already know. While the LV groups' performances in familiar trials in both experiments of the current thesis suggest that the LV students -simply through high token frequency in the input- were successful at imitating the target MWC, high input variability would most likely be an obstacle for such a straightforward memorisation learning goal, as suggested by the data yielded by HV students in familiar trials in the current study (see sections 3.2.5 and 4.2.3). This is reminiscent of findings elsewhere suggesting that increased input variability is -at least initially- more

challenging for learners than low input variability (Viviani et al., n.d.; Raviv et al., 2022), as discussed in section 1.6.3. In sum, in specific scenarios, educators may intentionally opt to sidestep the challenge that high variability can pose and choose low input variability as the preferred mode of instruction.

5.3 Visual Input (experiment 2)

As discussed in section 4.3.1.1, the provided visual input in experiment 2 seems to have pre-empted the expected positive effect of HV variability on the learning of non-adjacent dependencies. In fact, when combined with obvious visual input that unambiguously illustrates the non-adjacent dependencies (i.e., the association between a person and an action), HV input may have even had a negative impact on learning these dependencies due to the perhaps overwhelming variability of respective visual input. This raises two important questions: First, from a pedagogical perspective, might it be preferable to adopt a LV approach to learning non-adjacent dependencies in instructed settings? And second, is having semantic support reasonable in the context of learning non-adjacent dependencies?

Regarding the first question, the answer depends on the presence of visual support or any input which emphasizes the conceptual relationship in a non-adjacent dependency. If a non-adjacent dependency that refers to a (idiosyncratic) semantic association between two or more entities is immediately obvious from visual or other relevant types of input, then HV linguistic input is probably detrimental because it distracts from this association, which is already clear from the visuals. However, typically the difficult-to-learn long-distance associations in language feature more abstract NADs, such as subject-verb agreement. These syntactic associations might be more difficult (or impossible) to unambiguously illustrate visually or otherwise. In such cases, increased variability might still be the more beneficial mode of input.

The second question warrants a similar response: Yes, semantic support, such as that provided by visual input, is probably reasonable for learning non-adjacent dependencies. In fact, visual input might be the simplest and most preferable option for exemplifying and helping students learn a non-adjacent dependency, provided the dependency can be visually illustrated.

Yet, the crucial point regarding both questions is that visual input with no or low variability might support learning but does not provide students with any cues about the generalizability of the target structure. In a pedagogical context, educators risk having students get stuck on the aforementioned *imitative plateau*. Students may memorize the non-adjacent dependency from the visual input but may not infer that the structure is generalizable to other

contexts, thus only minimally extending their communicative agency in a relatively unproductive way.

A pedagogic compromise might be to introduce non-adjacent dependencies with visual support, if possible. Then, without or at least with less visual or other support, increase the variability during subsequent exposure to give students the opportunity to infer that the structure is extendable while avoiding the potentially detrimental effect of too much and too variable visual and therefore distracting input.

5.4 Explicit input

In keeping with previous more 'controlled' research, the current study focused on children's ability to *infer* structural information from the input without providing them with any explicit external information about structural regularities. Clearly, this situation does not fully resemble an authentic teaching context. While lengthy explicit rule input is often discouraged in primary school FL classrooms for reasons related to motivation, attention, and student engagement, teachers usually provide children with several forms of explicit input, for example in the form of some low-dosed rule teaching, feedback, or L1 translations (Legutke et al., 2012). In fact, research suggests that explicit instruction plays an important facilitating role in instructed FL development, though this is mediated by factors such as practice type and duration, and delivery method (for overviews see e.g., Akakura, 2011; Li & Sun, 2024). Explicit input in the form of corrective feedback has also been shown to be impactful in FL development (for an overview see e.g., Lyster et al., 2013). Regarding the current study, it would have been interesting to know if and to what extent additional explicit input would have impacted learning outcomes across input conditions. For example, in the context of supporting students' inferencing, I argue that alerting their attention to the structure of the input at the beginning of the exposure phase would have benefited learning outcomes in both groups. The HV group in particular would have most likely benefited from some introductory prompts along the lines of *Pay attention to what is repetitive in those sentences and to what changes between them*. On the other hand, in the context of straightforward explicit rule instruction, the students could have been explicitly taught that the first animal is always the animal that is being approached (experiment 1), and that *Tim is always singing*, *Anna is always eating*, and *Klara is always jumping*, respectively (experiment 2). Corresponding to similar questions raised elsewhere (e.g., Madlener-Charpentier, 2015), it would be interesting to see whether and to what degree subsequent input variability manipulations impact on the long-term fostering and development of this explicitly introduced structural knowledge foundation. In addition, it would be interesting to identify whether and to what degree such input manipulations would enhance students' ability to utilize this explicitly introduced knowledge productively in context-independent communication in the long-term. Finally, it would be useful to investigate the additional impact of corrective feedback during students' production on the development of their mental representations of form-function pairings.

These questions could build on existing literature. While the current study did not feature any explicit input, there is some evidence elsewhere which suggests that it might be beneficial for MWC input too to be supported with corresponding rule-input and training of metalinguistic awareness (McDonough & Trofimovich, 2013; Tode, 2003). For example, as demonstrated in Tode's (2003) study on copula-acquisition among beginning adult Japanese EFL learners, the 'chunk-rule continuum' in instructed settings, that is, learners' development from exemplar-based to abstract knowledge, does not only develop based on inductive rule-abstraction but also 'normal' deductive rule-input is crucial to learning. From a pedagogical perspective, Edmondson (1995) and Legutke et al. (2012) argue that grammar and lexis must be purposefully connected in the FL curriculum, and Wulff (2018: 19) maintains that "inductive construction learning […] complements deductive, rule-based learning processes". It is important to note though, that in such potential intervention studies it would become increasingly difficult to disambiguate the individual amounts of impact that the different input types (i.e., input flood with manipulated structure versus various forms of explicit input) have on students' attainment at test. Yet, this disambiguation might not be of primary importance in the context of pedagogical research with the goal of improving learning outcomes in mind, although it is clearly of theoretical relevance.

5.5 Future Research

Several issues that would be interesting to investigate in the future have already been touched on, such as the role of individual differences or explicit input in the context of manipulated MWC input or the precise ratio of type versus token frequency in the input. The latter is concerned with the perhaps single most important question for educators in the context of input variability: "[If variability is introduced], how much type variation is enough [in the context of a FL classroom]? How much variation is needed for learners to detect the target pattern and to productively extend it to novel items?" (Madlener-Charpentier, 2015: 301). I argue that it is likely that there are specific type-token frequency 'sweet spots' depending on the target structure, the students' individual differences, their current learning progress, and the input context (i.e., classrooms versus more controlled settings). The current work cannot differentiate between the benefits of different type-token ratios because it adapted its ratios from previous studies. Yet, others have introduced varying type-token ratios (in controlled settings), like Gómez (2002) who reports that there was evidence for the variability effect only in the condition with the highest amount of input variability (24x intervener items). At the same time, excessive input variability is considered detrimental to learning (Madlener-Charpentier, 2015). Problems that arise from excessive input variability make sense from a theoretical perspective because the input becomes too confusing and the structural cues become too unreliable to allow informed inferences about the input's structure, especially if there is no feedback to correct those inferences once voiced in communication. Also, there are limits to the amount of type variation that can reasonably be introduced in a classroom. And certainly, while teachers already aim to stratify their input as much as possible, there are limits to the amount of stratification that can be achieved in a classroom. Here, I would argue that educational technologies, especially those generative AI based technologies that can 'learn' alongside their learners' progress and continuously adapt the input accordingly, could play a crucial role in capitalising on the variability effect to boost learning. At the time of writing, the author is not aware of any published scientific investigations into such technologies. However, based on personal communication, the author knows that these technologies are currently in development

As suggested elsewhere (e.g., Madlener-Charpentier, 2015), from a pedagogical perspective, it would also be interesting to investigate whether the temporal ordering of LV

and HV input, specifically LV first followed by HV input, would be beneficial for young FL learners in classrooms. The idea would be that students pick up on the familiar, highly repetitive tokens first, and once they are confident with this knowledge base, more variable input is introduced to help them infer the more abstract underlying structure and extend it to novel contexts. This research avenue also pertains to the role of visual input which might, in initial learning stages, aide students to recognize the target structure.

Similarly, another interesting issue to examine concerns the role of different input modalities. The current study -like many previous studies in more controlled environments- used aural and some visual input. Yet, written input would be interesting to investigate too, since even young FL students are typically confronted with at least some written input, for example in the form of textbooks. In orthographic form, it might be easier to highlight the slots of MWCs and illustrate the fact that those can be filled flexibly. Of course, this must be implemented carefully, not least because it is not guaranteed that every student in the relevant primary school class levels is literate, as repeatedly suggested by the Progress in International Reading Literacy Studies (PIRLS) (e.g., European Commission, 2021).

A final potentially worthwhile endeavour for future intervention work is the introduction of skewed input. While this might not be the most important issue for educators, it is certainly of interest from a theoretical perspective as natural language input is often skewed and its structural features more often than not follow a Zipfian distribution (Zipf, 1935). For example, in the context of verb-argument-constructions this means that their verb slots are often (but not always; see Sethuraman & Goodman, 2004) filled with a prototypical verb (e.g., Ellis, 2009; Ellis & Ferreira-Junior, 2009; Ninio, 1999, 2006; O'Donnell & Ellis, 2009, 2010). One example would be the English ditransitive structure which is associated with the verb *to give* in many usage cases (Goldberg, 2006; cf. Boyd & Goldberg, 2009). Thus, introducing skewed input in the context of MWC input in early FL instruction would in most cases closely mimic the input conditions of naturalistic FL acquisition. However, it is important to note that the evidence regarding the benefit of skewed input over high and low variability input is not entirely clear, with some studies reporting a null effect of skewed input (e.g., Year, 2009; Year & Gordon, 2009; but note Boyd & Goldberg's (2009) comments on the statistical analyses of Year & Gordon, 2009), and others reporting a benefit of skewed input (Goldberg & Casenhiser, 2005) or, in contrast, better learning outcomes with balanced (i.e., low variability) input (e.g., McDonough & Nekrasova-Becker, 2014). In addition, input conditions (e.g., frequency),

learning trajectories, and (expected) learning outcomes in pedagogical contexts are vastly different from those in natural language development and it remains to be seen whether the taught input structure necessarily needs to follow the structure it would have in natural settings.

5.6 Limitations

Needless to say, the current work has a number of limitations. First, there was -what is in retrospect- a flaw in the design of experiment 2. The question which was used to prompt students' responses during outcome testing could be interpreted in two opposite ways and was thus ambiguous, specifically so in trials featuring a novel intervener. In these trials, the question *Is it correct that … ?* can be interpreted either as targeting the non-adjacent dependency itself (e.g., *Tim* is the one who is *singing*) or in a more global way, that is, whether the student encountered *this particular* test sentence as a whole in the input before. In trials featuring novel interveners, the former interpretation warrants a 'Yes' response because the non-adjacent dependency is correct. On the other hand, the latter interpretation warrants a 'No' response because the sentence as it is did not feature in the input. Although the first interpretation targeting the non-adjacent dependency was the intended interpretation by the researcher, both responses would be 'correct' from the students' perspective. This oversight was only noticed after data collection and might have contributed to the ambiguous and somewhat unexpected findings in trials featuring novel interveners (i.e., the trials targeting generalization) in experiment 2. Fortunately, in familiar trials, the question is not ambiguous. However, in wrong dependency trials, the picture is less clear. On the one hand, both interpretations of the question warrant 'No' responses: the non-adjacent dependency is violated, and the sentence did not feature in the input. On the other hand, it remains unclear on which grounds the participants made their decision here. They could have rejected wrong dependency sentences because they had picked up on the non-adjacent dependency and noticed that it was violated, or they could have rejected the sentences on the grounds that they had not encountered them in the input – without specific attention to the non-adjacent dependency. In sum, experiment 2 still demonstrates that students learned something from the input, but this limitation dilutes our understanding of the grounds on which students made their response decisions.

Two further limitations concern the sample which was used in the current experiments. First, as mentioned elsewhere, participants were not sampled randomly. As a result, the sample size of the reported experiments is effectively N=2 because the data is nested at class-level. Thus, the current experiments are underpowered, despite the use of Bayes Factor statistics in an effort to mitigate against this to some extent. That is why the current work has

been introduced as a 'proof of concept' study throughout this thesis. Having found promising tentative evidence for the variability effect, future investigations might be able to conduct work at a larger scale than an individual's PhD work which is constrained both financially and logistically. Such work should ideally follow an RCT (randomized controlled trial) structure. Nonetheless, the current work contributes to a small set of much-needed primary school FL intervention studies and might inform larger future studies.

Second, background data on the students' language biographies should have been collected, as has been done elsewhere (e.g., Viviani et al., n.d.). Collecting such data would have been important because in this way the behaviour of students who are proficient in a language (other than English) that closely follows the introduced German structures could have been monitored closely and their data could have been potentially excluded from analysis. However, given that the entire study only consisted of roughly 40 children, the statistical findings remain tentative either way. Yet, for future works with larger sample sizes this might be an important point to keep in mind. In addition, data on students' language biographies might offer interesting insights into the potentially mediating impact of L1 transfer effects on the variability effect. However, such L1 effects are not consistently observed in instructed FL learning (e.g., Hopp et al., 2019).

Furthermore, while the decision to change the procedures of the WASI and the PVST was a logistic one, the administration procedures did not fully correspond to the tests' original instructions and therefore violated the parameters of the tasks. It is recognized that this could have introduced threats to the validity of the assessments. Yet, since the pre-tests were supposed to constitute background measures and were not essential to the answering of the research question, this risk was considered worth taking as the alternative would have been not to conduct the tests at all

5.7 Pedagogical implications

As underlined previously, the current study is not sufficiently powered and, importantly, it is the first work of its kind in a real primary school FL classroom. Consequently, pedagogical implications are necessarily tentative. Nonetheless, the current proof of concept study offers several valuable pedagogical implications which can be pursued in future research. First and foremost, the current data suggest that even the youngest FL learners in 'noisy' contexts with limited input can pick up on structural features of the input. They can do so even without being explicitly prompted to these features and without receiving any feedback as to their performance. From a pedagogical perspective, this is reassuring news since it offers a positive response to previous somewhat resigned claims that young FL learners often fail to pick up on structural features in the target language, such as verb-argument knowledge (Engel et al., 2009). In contrast to such claims, the current data suggest that the input structure in early instructed FL settings can be manipulated in a way that supports students' learning of crucial form-function mappings. This is especially true considering that an authentic teaching situation would most likely feature explicit input, for example in the form of corrective feedback or rule explanation, which would most likely contribute even more to students' learning. In sum, although it is early days, the current study can tentatively confirm that some form of variability in the input -in appropriate situations- is likely to be beneficial to FL learners' generalization abilities in early instructed settings. And while this is not a specific pedagogical recommendation per se, it is still a valuable and perhaps encouraging piece of information for educators as it offers an actionable and pedagogically realistic approach to help students overcome the *imitative plateau*, pending crucial further research (see section 0).

Second, and perhaps banally so, the current study is another testimony to the immense language learning capacity of children. Even under 'noisy' classroom conditions, students across both input groups learned something. Certainly, the variability manipulation may have impacted their respective ability to infer and subsequently extend structural knowledge to novel contexts. However, overall, the current study adds data from a real classroom and featuring natural language input to the already available evidence that children are able to learn constructional meaning even from minimal input, as repeatedly suggested elsewhere under more controlled conditions (e.g., Casenhiser & Goldberg, 2005; Goldberg et al., 2004; Gómez, 2002; Gómez & Maye, 2005; Wonnacott et al., 2012). While this observation might not

necessarily bear a specific pedagogical implication, it does challenge the somewhat pessimistic claims made elsewhere in the literature that we as researchers and educators need to be realistic about young primary school learners' 'necessarily rudimentary' learning outcomes by the end of primary school (Jäger, 2012). Certainly, there are many structural and contextual challenges to primary school FL instruction internationally. Undoubtedly, only so much can be taught and learned in sometimes merely one lesson of FL input per week (Holmes & Myles, 2019). Yet, the current study offers some encouraging classroom data corroborating what might come as no surprise to the research community: In principle, the children bring the necessary cognitive prerequisites to the table to overcome *imitative plateaus* and foster communicative agency. For FL instruction at primary level to be more successful, it is the responsibility of educational stakeholders to create improved context-conditions. Simultaneously, it is up to researchers and policymakers to collaborate closely and provide evidence-based recommendations to educators in FL classrooms, enabling them to effectively leverage each child's FL learning capacities, for example through targeted input variability.

5.8 Conclusion

Given the results of a systematic review (Schulz et al., 2023), to the best of the author's knowledge, the teaching interventions reported in the current thesis constitute the first investigation into the impact of increased input variability on young FL learners' learning and generalization of underlying structural patterns of two target MWCs in an authentic FL classroom. Acquiring such generalizations is considered crucial for young FL learners' development of context-independent, productive communicative agency. However, FL learning outcomes in input-limited instructed settings are often suboptimal, with students frequently leaving primary school possessing relatively limited and imitative language abilities. The results of the current teaching interventions tentatively confirm the positive impact of increased input variability on students' ability to infer and generalize underlying structural patterns – even in a 'noisy' classroom context. Specifically, in experiment 1 featuring verb-argument-constructions, the data from unfamiliar trials in the production task yield substantial Bayes Factor evidence in favour of the hypothesis of improved performance by the HV group in contexts with unfamiliar verbs. This variability effect corresponds to previous 'more-controlled' research in psycholinguistics but also across other cognitive domains. However, the current data are also in line with previous research in the way they indicate that HV input is no panacea and may, in certain situations and for certain learners, even be detrimental or at least not beneficial to learning. In fact, the data suggest that LV input facilitates imitation and may therefore (continue to) play an important role in early FL pedagogy that should not be overlooked.

More generally, the current work illustrates that even in a 'noisy' classroom context, young FL learners can pick up on complex form-function correspondences from extremely limited input provided it is sufficiently repetitive and purposefully structured. Tentatively, the data provides researchers and educators alike with the important confidence that when certain input parameters are set correctly, young FL students' communicative agency may be improved even in severely input limited instructed contexts. Of course, the current thesis illustrates a classroom-based research avenue in its early stages. Yet, it already highlights several important issues for the future where researchers need more data to provide educators and stakeholders with evidence-based recommendations as to which input-parameters to tweak and how to tweak them to best support young FL students in increasing

their communicative agency. Those parameters include the exact type-token ratios that teachers should ideally adopt in their input, the role of visual and explicit input in the context of high or low input variability, the role of individual differences, and the potentially mediating impact of different input modalities on the strength of the variability effect.

6 References

Aguado, K. (2002). Formelhafte Sequenzen und ihre Funktionen für den L2-Erwerb. *Zeitschrift für angewandte Linguistik*, *37*, 27–49.

Akakura, M. (2012). Evaluating the effectiveness of explicit instruction on implicit and explicit L2 knowledge. *Language Teaching Research*, *16*(1), 9–37. https://doi.org/10.1177/1362168811423339

Anthony, L. and Nation, I.S.P. (2021). PVST (Version 1.2.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Arnon, I., & Christiansen, M. H. (2017). The Role of Multiword Building Blocks in Explaining L1–L2 Differences. *Topics in Cognitive Science*, *9*(3), 621–636. https://doi.org/10.1111/tops.12271

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of Conditional Probability Statistics by 8-Month-Old Infants. *Psychological Science*, *9*(4), 321–324. https://doi.org/10.1111/1467-9280.00063

Balcı, O., & Çakır, A. (2012). Teaching vocabulary through collocations in EFL classes: The case of Turkey. *International Journal of Research Studies in Language Learning*, *1*(1), 21–32.

Bannard, C., & Lieven, E. (2012). Formulaic Language in L1 Acquisition. *Annual Review of Applied Linguistics*, *32*, 3–16. https://doi.org/10.1017/S0267190512000062

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modelling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences*, *106*(41), 17284–17289. https://doi.org/10.1073/pnas.0905638106

Becker, J. D. (1975). The Phrasal Lexicon. In B. L. Nash-Webber & R. Schank (Eds.), *Theoretical Issues in Natural Language Processing*. https://aclanthology.org/T75-2013

Behrens, H. (2009). Usage-based and emergentist approaches to language acquisition. *Linguistics*, *47*(2), 382–411.

BIG-Kreis (Ed.). (2015). *Der Lernstand im Englischunterricht am Ende von Klasse 4: Ergebnisse der BIG-Studie*. Domino Verlag.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society* (pp. 56–64).

Boers, F., Dang, T. C. T., & Strong, B. (2017). Comparing the effectiveness of phrase-focused exercises: A partial replication of Boers, Demecheleer, Coxhead, and Webb (2014). *Language Teaching Research*, *21*(3), 362–380. https://doi.org/10.1177/1362168816651464

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006a). Formulaic sequences and perceived oral proficiency: Putting a Lexical Approach to the test. *Language Teaching Research*, *10*(3), 245–261. https://doi.org/10.1191/1362168806lr195oa

Boers, F., & Lindstromberg, S. (2012). Experimental and Intervention Studies on Formulaic Sequences in a Second Language. *Annual Review of Applied Linguistics*, *32*, 83–110. https://doi.org/10.1017/S0267190512000050

Boers, F., & Muñoz-Basols, J. (2021). Acquisition of idiomatic language in L2 Spanish. In J. Barcroft & J. Muñoz-Basols (Eds.), *Spanish vocabulary learning in meaning-oriented instruction* (pp. 62–88). Routledge, Taylor & Francis Group.

Boland, A., Cherry, M. G., & Dickson, R. (Eds.). (2017). *Doing a systematic review: A student's guide* (2nd edition). SAGE.

Bortoli, L., Robazza, C., Durigon, V., & Carra, C. (1992). Effects of Contextual Interference on Learning Technical Sports Skills. *Perceptual and Motor Skills*, *75*(2), 555–562. https://doi.org/10.2466/pms.1992.75.2.555

Boyd, J. K., & Goldberg, A. E. (2009). Input Effects Within a Constructionist Framework. *The Modern Language Journal*, *93*(3), 418–429. https://doi.org/10.1111/j.1540-4781.2009.00899.x

Bredenbröcker, M. (2018). *A corpus-based approach to English as a foreign language at primary school level: Collocations for early beginners*. Verlag Dr. Kovač.

Brekelmans, G., Lavan, N., Saito, H., Clayards, M., & Wonnacott, E. (2022). Does high variability training improve the learning of non-native phoneme contrasts over low variability training? A replication. *Journal of Memory and Language*, *126*, 104352. https://doi.org/10.1016/j.jml.2022.104352

Brunton, G., Stansfield, C., Caird, J., & Thomas, J. (2017). Finding relevant studies. In D. Gough, S. Oliver, & J. Thomas (Eds.), *An introduction to systematic reviews* (2nd ed., pp. 93–122). SAGE.

Bulf, H., Johnson, S. P., & Valenza, E. (2011). Visual statistical learning in the newborn infant. *Cognition*, *121*(1), 127–132. https://doi.org/10.1016/j.cognition.2011.06.010

Busse, V., Hennies, C., Kreutz, G., & Roden, I. (2021). Learning grammar through singing? An intervention with EFL primary school learners. *Learning and Instruction*, *71*, 101372. https://doi.org/10.1016/j.learninstruc.2020.101372

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, *10*, 425–455.

Bybee, J. (2008). Usage-based grammar and second language acquisition. In P. J. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 216–236). Routledge, Taylor & Francis Group.

Carlsen, B., Glenton, C., & Pope, C. (2007). Thou shalt versus thou shalt not: A meta-synthesis of GPs' attitudes to clinical practice guidelines. *British Journal of General Practice*, *57*(545), 971–978. https://doi.org/10.3399/096016407782604820

Casenhiser, D., & Goldberg, A. E. (2005). Fast mapping between a phrasal form and meaning. *Developmental Science*, *8*(6), 500–508. https://doi.org/10.1111/j.1467-7687.2005.00441.x

Chae, J., & Kim, H. (2019). The effects of receptive and productive collocation learning task types on elementary school Students' English vocabulary acquisition. *The Journal of Education*, *39*(2), 399–422.

Choi, S. (2017). Processing and learning of enhanced English collocations: An eye movement study. *Language Teaching Research*, *21*(3), 403–426. https://doi.org/10.1177/1362168816653271

Christiansen, M. H., & Arnon, I. (2017). More Than Words: The Role of Multiword Sequences in Language Learning and Use. *Topics in Cognitive Science*, *9*(3), 542–551. https://doi.org/10.1111/tops.12274

Cochrane Effective Practice and Organisation of Care (EPOC). (2017). *Data collection form.* https://epoc.cochrane.org/sites/epoc.cochrane.org/files/public/uploads/Resources-for-authors2017/good_practice_data_extraction_form.doc

Council of Europe. (2001). *Common European framework of reference for language learning, teaching, assessment.* Cambridge University Press.

Council of Europe. (2018). *Common European framework of reference for languages. Learning, teaching, assessment. Companion volume with new descriptors.* https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Czarnecka, M. (2011). Formelhafte Sequenzen in der Erst- und Zweitsprache: Versuch einer Begriffsbestimmung aus psycholinguistischer Perspektive. *Germanica Wratislaviensia*, *133*, 189–199.

Diehr, B. (2009). Young learners' use of English. Imitation or production? In T. Stewart (Ed.), *Insights on teaching speaking in TESOL* (pp. 53–66). TESOL.

Diehr, B., & Polte, L. (2009). Zur Entwicklung diskursiver Fähigkeiten im Englischunterricht der Grundschule. Eine vergleichende Untersuchung von Sprechern des Englischen als Erst- und Fremdsprache. *Zeitschrift Für Fremdsprachenforschung*, *20*(2), 147–174.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference* (1. publ). Palgrave Macmillan.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00781

Dienes, Z. (2019). How Do I Know What My Theory Predicts? *Advances in Methods and Practices in Psychological Science*, *2*(4), 364–377. https://doi.org/10.1177/2515245919876960

Dienes, Z. (2021). How to use and report Bayesian hypothesis tests. *Psychology of Consciousness: Theory, Research, and Practice*, *8*(1), 9–26. https://doi.org/10.1037/cns0000258

Doyé, P., & Lüttge, D. (1977). *Untersuchungen zum Englischunterricht in der Grundschule: Bericht über das Braunschweiger Forschungsprojekt 'Frühbeginn des Englischunterrichts', FEU* (1. Aufl). Westermann.

Duran-Karaoz, Z., & Tavakoli, P. (2020). Predicting L2 Fluency from L1 Fluency Behavior: The Case of L1 Turkish and L2 English Speakers. *Studies in Second Language Acquisition*, *42*(4), 671–695. https://doi.org/10.1017/S0272263119000755

Edmondson, W. (1995). Wortschatzerwerb und Spracherwerb. In K.-R. Bausch & H. Christ (Eds.), *Erwerb und Vermittlung von Wortschatz im Fremdsprachenunterricht* (pp. 55–62). Narr.

Eidsvåg, S. S., Austad, M., Plante, E., & Asbjørnsen, A. E. (2015). Input Variability Facilitates Unguided Subcategory Learning in Adults. *Journal of Speech, Language, and Hearing Research: JSLHR*, *58*(3), 826–839.

Ellis, N. (2015). Cognitive and social aspects of learning from usage. In T. Cadierno & S. W. Eskildsen (Eds.), *Usage-Based Perspectives on Second Language Learning:* (pp. 49–73). de Gruyter Mouton. https://doi.org/10.1515/9783110378528

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188. https://doi.org/10.1017/S0272263102002024

Ellis, N. C. (2009). Optimizing the Input: Frequency and Sampling in Usage-Based and Form-Focused Learning. In M. H. Long & C. J. Doughty (Eds.), *The Handbook of Language Teaching* (1st ed., pp. 139–158). Wiley. https://doi.org/10.1002/9781444315783.ch9

Ellis, N. C. (2017). Chunking in Language Usage, Learning and Change: I Don't Know. In M. Hundt, S. Mollin, & S. E. Pfenninger (Eds.), *The Changing English Language* (1st ed., pp. 113–147). Cambridge University Press. https://doi.org/10.1017/9781316091746.006

Ellis, N. C., & Ferreira–Junior, F. (2009). Construction Learning as a Function of Frequency, Frequency Distribution, and Function. *The Modern Language Journal*, *93*(3), 370–385. https://doi.org/10.1111/j.1540-4781.2009.00896.x

Ellis, N., Römer, U. & O'Donnell, M. B. (2016) Constructions and Usage-based Approaches to Language Acquisition. (2016). *Language Learning*, *66*(S1), 23–44. https://doi.org/10.1111/lang.1_12177

Ellis, N. C., & Wulff, S. (2015). Usage-based approaches to SLA. In B. van Patten & J. Williams (Eds.), *Second language acquisition research series: Theories in second language acquisition* (pp. 75–94). Routledge, Taylor & Francis Group.

Ellis, N., & Wulff, S. (2020). Usage-based approaches to L2 acquisition. In B. VanPatten, G. D. Keating, & S. Wulff (Eds.), *Theories in second language acquisition: An introduction* (3rd ed., pp. 63–82). Routledge, Taylor & Francis Group.

Enever, J. (Ed.). (2011). *Early Language Learning in Europe*. British Council.

Engel, G., Groot-Wilken, B., & Thürmann, E. (Eds.). (2009). *Englisch in der Primarstufe - Chancen und Herausforderungen: Evaluation und Erfahrungen aus der Praxis* (1. Aufl., 1. Dr). Cornelsen Verlag.

Eskildsen, S. W. (2009). Constructing another Language—Usage-Based Linguistics in Second Language Acquisition. *Applied Linguistics*, *30*(3), 335–357. https://doi.org/10.1093/applin/amn037

Eskildsen, S. W., & Cadierno, T. (2015). Advancing usage-based approaches to L2 studies. In T. Cadierno & S. W. Eskildsen (Eds.), *Usage-Based Perspectives on Second Language Learning:* (pp. 1–16). de Gruyter Mouton. https://www.degruyter.com/document/doi/10.1515/9783110378528/html

Estes, K. G., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can Infants Map Meaning to Newly Segmented Words?: Statistical Segmentation and Word Learning. *Psychological Science*, *18*(3), 254–260. https://doi.org/10.1111/j.1467-9280.2007.01885.x

Europarat (Ed.). (2001). *Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen, lehren, beurteilen: Niveau A1, A2, B1, B2, C1, C2*. Langenscheidt.

European Commission, Directorate-General for Education, Youth, Sport and Culture, (2023). *Children's reading competence and well-being in the EU: an EU comparative analysis of the PIRLS results*, Publications Office of the European Union. https://data.europa.eu/doi/10.2766/820665

Eyckmans, J., & Lindstromberg, S. (2017). The power of sound in L2 idiom learning. *Language Teaching Research*, *21*(3), 341–361. https://doi.org/10.1177/1362168816655831

Fillmore, L. W. (1976). *The second time around: Cognitive and social strategies in second language acquisition* [Doctoral dissertation]. Stanford University.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press.

Goldberg, A. E. (2007). Learning linguistic patterns. In A. B. Markman (Ed.), *Categories in use* (pp. 33–63). Elsevier.

Goldberg, A. E. (2009). The nature of generalization in language. *Cognitive Linguistics*, *20*(1), 93–127. https://doi.org/10.1515/COGL.2009.005

Goldberg, A., & Casenhiser, D. (2008). Construction Learning and Second Language Acquisition. In P. J. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 197–215). Routledge, Taylor & Francis Group.

Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. (Eds.). (2004). Learning argument structure generalizations. *Cogl*, *15*(3), 289–316. https://doi.org/10.1515/cogl.2004.011

Gómez, R. L. (2002). Variability and Detection of Invariant Structure. *Psychological Science*, *13*(5), 431–436. https://doi.org/10.1111/1467-9280.00476

Gómez, R. L., & Maye, J. (2005). The Developmental Trajectory of Nonadjacent Dependency Learning. *Infancy*, *7*(2), 183–206. https://doi.org/10.1207/s15327078in0702_4

Graddol, D. (2006). *English Next*. British Council. https://www.teachingenglish.org.uk/publications/case-studies-insights-and-research/english-next

Gray, R. (2021). Empty systematic reviews: Identifying gaps in knowledge or a waste of time and effort? *Nurse Author & Editor*, *31*(2), 42–44. https://doi.org/10.1111/nae2.23

Grunow, H., Spaulding, T. J., Gómez, R. L., & Plante, E. (2006). The effects of variation on learning word order rules by adults with and without language-based learning disabilities. *Journal of Communication Disorders*, *39*(2), 158–170. https://doi.org/10.1016/j.jcomdis.2005.11.004

Hakuta, K. (1974). Prefabricated patterns and the emergence of structure in second language acquisition. *Language Learning*, *24*(2), 287–297. https://doi.org/10.1111/j.1467-1770.1974.tb00509.x

Harrington, M., & Dennis, S. (2002). Input-driven language learning. *Studies in Second Language Acquisition*, *24*(2), 261–268.

Hempel, M. (2016). Förderung produktiver Sprachkompetenzen – Sind Lehrwerke Teil des Problems oder Teil der Lösung? In H. Böttger & N. Schlüter (Eds.), *FFF - Fortschritte im frühen Fremdsprachenlernen: Tagungsband zur 4. FFF-Konferenz 2014 in Leipzig* (pp. 124–133). Westermann.

Henk, K. (2019). *Strukturerwerb im Französischunterricht: Hypothesen aus gebrauchsbasierter Sicht und ihre empirische Überprüfung*. Verlag Empirische Pädagogik.

Hilpert, M. (2014). *Construction grammar and its application to English*. Edinburgh University Press.

Holmes, B., & Myles, F. (2019). *White Paper: Primary languages policy in England—The way forward.* RiPL. http://www.ripl.uk/policy/

Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., & Vedel, I. (2019). Improving the content validity of the mixed methods appraisal tool: A modified e-Delphi study. *Journal of Clinical Epidemiology*, *111*, 49-59.e1. https://doi.org/10.1016/j.jclinepi.2019.03.008

Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., & Vedel, I. (n.d.). Mixed methods appraisal tool (MMAT), version 2018. Registration of Copyright (#1148552), Canadian Intellectual Property Office, Industry Canada. http://mixedmethodsappraisaltoolpublic.pbworks.com/w/file/fetch/127916259/MMAT_2018_criteria-manual_2018-08-01_ENG.pdf

Hopp, H., Steinlen, A., Schelletter, C., & Piske, T. (2019). Syntactic development in early foreign language learning: Effects of L1 transfer, input, and individual factors. *Applied Psycholinguistics*, *40*(05), 1241–1267. https://doi.org/10.1017/S0142716419000249

Hopp, H., & Thoma, D. (2021). Effects of Plurilingual Teaching on Grammatical Development in Early Foreign-Language Learning. *The Modern Language Journal*, *105*(2), 464–483. https://doi.org/10.1111/modl.12709

Howarth, P. (1998). Phraseology and Second Language Proficiency. *Applied Linguistics*, *19*(1), 24–44. https://doi.org/10.1093/applin/19.1.24

Huang, J., & Hatch, E. M. (1978). A Chinese child's acquisition of English. In E. Hatch (Ed.), *Second language acquisition: A book of readings* (pp. 118–131). Newbury House.

Ibbotson, P. (2011). Abstracting Grammar from Social–Cognitive Foundations: A Developmental Sketch of Learning. *Review of General Psychology*, *15*(4), 331–343. https://doi.org/10.1037/a0025609

Ibbotson, P. (2013). The Scope of Usage-Based Theory. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00255

In'nami, Y., Hijikata, Y., & Koizumi, R. (2022). Working Memory Capacity and L2 Reading: A Meta-Analysis. *Studies in Second Language Acquisition*, *44*(2), 381–406. https://doi.org/10.1017/S0272263121000267

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press.

Jäger, A. (2012). Die Förderung kommunikativer Fähigkeiten im Englischunterricht der Grundschule. In H. Böttger (Ed.), *Englisch: Didaktik für die Grundschule* (pp. 112–122). Cornelsen.

Jeffreys, H. (1961). *Theory of probability* (3 ed.). Oxford: Oxford University Press.

Jeon, J.-H., & Kim, J.-R. (2018). Reflection on self-instruction in elementary English based on the frequency of new vocabulary by activity of COLT part 2. *Asia-Pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology*, *8*(9), 353–366.

Jung, D.-W., & Shin, C.-O. (2013). The effects of chunk-focused learning on Korean elementary school Students' English skills and interest. *Primary English Education*, *19*(1), 219–239.

Kahl, P. W., & Knebler, U. (1996). *Englisch in der Grundschule - und dann? Evaluation des Hamburger Schulversuchs Englisch ab Klasse 3* (1. Aufl). Cornelsen.

Kenyeres, A., & Kenyeres, E. (1938). Comment une petite hongroise de sept ans apprend le français. *Archives de Psychologie*, *26*, 321–366.

Kersten, S. (2015). Language development in young learners: The role of formulaic language. In J. Bland (Ed.), *Teaching English to young learners: Critical issues in language teaching with 3-12 year olds* (pp. 129–145). Bloomsbury Academic.

Kim, G., & Lee, S. (2009). A study on vocabulary teaching through English stories. *English21*, *22*(3), 157–186.

Kim, H.-R., & Kang, W.-S. (2005). Using English cartoons for children's literacy development in the elementary school. *Primary English Education*, *11*(1), 209–240.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35–B42. https://doi.org/10.1016/S0010-0277(02)00004-5

Klein, J. (2018). *Latest research results from the EU project The Language Magician.* https://www.thelanguagemagician.net/latest-research-results-from-the-eu-project- the-language-magician/

(KM-BW) Kultusministerium Baden-Württemberg. (2004). *Bildungsplan 2004. Grundschule.* http://www.bildungsplaene-bw.de/site/bildungsplan/get/documents_E-1204433591/lsbw/Bildungsplaene/Bildungsplaene-2004/Bildungsstandards/Grundschule_Bildungsplan_Gesamt.pdf

(KM-BW) Kultusministerium Baden-Württemberg. (2016). *Bildungsplan der Grundschule. Englisch (ab Klasse 3/4).* https://www.bildungsplaene-bw.de/site/bildungsplan/get/documents/lsbw/export-pdf/depot-pdf/ALLG/BP2016BW_ALLG_GS_E34.pdf

Kostka, N. (2020). *Produktives Sprechen im Englischunterricht der Grundschule—Eine empirische Studie zur Bedeutung formelhafter Sequenzen*. Universitätsbibliothek.

Lakens, D. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Lee, Y.-J., & Jeong, D.-B. (2010). The effect of collocation-based English vocabulary learning on Children's speaking abilities. *Studies in English Language and Literature*, *36*(3), 293–334.

Legutke, M., Müller-Hartmann, A., & Schocker- von Ditfurth, M. (2012). *Teaching English in the primary school* (4. Aufl). Klett Lerntraining.

Lenzing, A., & Roos, J. (2012). Die sprachliche Entwicklung und Ausdrucksmöglichkeiten von Grundschülerinnen und Grundschülern im Englischunterricht. In M. Bär, A. Bonnet, H. Decke-Cornill, A. Grünewald, & A. Hu (Eds.), *Globalisierung—Migration—Fremdsprachenunterricht: Dokumentation zum 24. Kongress für Fremdsprachendidaktik der Deutschen Gesellschaft für Fremdsprachenforschung (DGFF) Hamburg* (pp. 207–220). Schneider.

Li, F., & Sun, Y. (2024). Effects of different forms of explicit instruction on L2 development: A meta-analysis. *Foreign Language Annals*, *57*(1), 229–255. https://doi.org/10.1111/flan.12726

Likourezos, V., Kalyuga, S., & Sweller, J. (2019). The Variability Effect: When Instructional Variability Is Advantageous. *Educational Psychology Review*, *31*(2), 479–497. https://doi.org/10.1007/s10648-019-09462-8

Lim, E., & Lee, K. (2012). Elementary students' use of English collocation and its influence on written language learning and learning attitudes. *Journal of British & American Studies*, *26*, 265–298.

Lyster, R., Saito, K., & Sato, M. (2013). Oral corrective feedback in second language classrooms. *Language Teaching*, *46*(1), 1–40. https://doi.org/10.1017/S0261444812000365

Madlener-Charpentier, K. (2015). *Frequency effects in instructed second language acquisition*. De Gruyter Mouton.

Madlener-Charpentier, K. (2016). Input optimization: Effects of type and token frequency manipulations in instructed second language learning. In H. Behrens & S. Pfänder (Eds.), *Frequency effects in language: What counts in language processing, acquisition and change.* (pp. 133–173). De Gruyter Mouton.

McDonough, K., & Nekrasova-Becker, T. (2014). Comparing the effect of skewed and balanced input on English as a foreign language learners' comprehension of the double-object dative construction. *Applied Psycholinguistics*, *35*(2), 419–442. https://doi.org/10.1017/S0142716412000446

Mcdonough, K., & Trofimovich, P. (2013). Learning a novel pattern through balanced and skewed input. *Bilingualism: Language and Cognition*, *16*(3), 654–662. https://doi.org/10.1017/S1366728912000557

Mellow, J. D. (2006). The Emergence of Second Language Syntax: A Case Study of the Acquisition of Relative Clauses. *Applied Linguistics*, *27*(4), 645–670. https://doi.org/10.1093/applin/aml031

Mitchell, A. E., Jarvis, S., O'Malley, M., & Konstantinova, I. (2015). Working Memory Measures and L2 Proficiency. In Z. (Edward) Wen, M. Borges Mota, & A. McNeill (Eds.), *Working Memory in Second Language Acquisition and Processing* (pp. 270–284). Multilingual Matters. https://doi.org/10.21832/9781783093595-019

Mitchell, T. M. (1982). Generalization as search. *Artificial Intelligence*, *18*(2), 203–226. https://doi.org/10.1016/0004-3702(82)90040-6

(MSB NRW) Ministerium für Schule und Bildung des Landes Nordrhein-Westfalen (Ed.). (2012). *Richtlinien und Lehrpläne für die Grundschule in Nordrhein-Westfalen*. Ritterbach Verlag. https://www.schulentwicklung.nrw.de/lehrplaene/upload/klp_gs/LP_GS_2008.pdf

Müller, T., Böttger, H., Schlüter, N., Kierepka, A., Börner, O., Legutke, M., Kronisch, I., & Lohmann, C. (2016). Der Lernstand im Englischunterricht am Ende von Klasse 4: Erste Ergebnisse der BIGStudie. In H. Böttger & N. Schlüter (Eds.), *FFF - Fortschritte im frühen Fremdsprachenlernen: Tagungsband zur 4. FFF-Konferenz 2014 in Leipzig* (pp. 8–44). Westermann.

Myles, F., Mitchell, R., & Hooper, J. (1999). Interrogative chunks in French L2: A basis for creative construction? *Studies in Second Language Acquisition*, *21*(1), 49–80. https://doi.org/10.1017/S0272263199001023

Nation, P., & Anthony, L. (2016). Measuring vocabulary size. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning. 3* (pp. 355–368). Routledge Taylor & Francis Group.

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching* (Nachdr.). Oxford University Press.

Ninio, A. (1999). Model learning in syntactic development: Intransitive verbs. *International Journal of Bilingualism*, *3*(2–3), 111–130. https://doi.org/10.1177/13670069990030020301

Ninio, A. (2006). *Language and the Learning Curve*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199299829.001.0001

Noyes, J., & Lewin, S. (2011). Extracting qualitative evidence. In J. Noyes, A. Booth, K. Hannes, J. Harden, S. Harris, S. Lewin, & C. Lockwood (Eds.), *Supplementary guidance for inclusion of qualitative research in Cochrane systematic reviews of interventions. Version 1 (updated august 2011)*. Cochrane Collaboration Qualitative Methods Group. http://cqrmg.cochrane.org/supplemental-handbook-guidance

O'Donnell, M., & Ellis, N. C. (2009). Measuring formulaic language in corpora from the perspective of Language as a Complex System. *5th Corpus Linguistics Conference*, University of Liverpool, July, 2009.

O'Donnell, M., & Ellis, N. C. (2010). Towards an Inventory of English Verb Argument Constructions. *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, Los Angeles, California, June 2010.

Pace, R., Pluye, P., Bartlett, G., Macaulay, A. C., Salsberg, J., Jagosh, J., & Seller, R. (2012). Testing the reliability and efficiency of the pilot Mixed Methods Appraisal Tool (MMAT) for systematic mixed studies review. *International Journal of Nursing Studies*, *49*(1), 47–53. https://doi.org/10.1016/j.ijnurstu.2011.07.002

Pae, T. (2019). A Simultaneous Analysis of Relations Between L1 and L2 Skills in Reading and Writing. *Reading Research Quarterly*, *54*(1), 109–124. https://doi.org/10.1002/rrq.216

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, n71. https://doi.org/10.1136/bmj.n71

Peters, A. M. (1983). *The units of language acquisition*. New York: Cambridge University Press.

Petticrew, M., & Roberts, H. (2006). *Systematic Reviews in the Social Sciences: A Practical Guide* (1st ed.). Wiley. https://doi.org/10.1002/9780470754887

Pienemann, M., Keßler, J.-U., & Roos, E. (Eds.). (2006). *Englischerwerb in der Grundschule: Ein Studien- und Arbeitsbuch*. Ferdinand Schöningh.

Pluye, P., Gagnon, M.-P., Griffiths, F., & Johnson-Lafleur, J. (2009). A scoring system for appraising mixed methods research, and concomitantly appraising qualitative, quantitative and mixed methods primary studies in Mixed Studies Reviews. *International Journal of Nursing Studies*, *46*(4), 529–546. https://doi.org/10.1016/j.ijnurstu.2009.01.009

Polio, C., & Yoon, H.-J. (2020). Exploring multi-word combinations as measures of linguistic accuracy in second language writing. In B. Le Bruyn & M. Paquot (Eds.), *Learner corpus research meets second language acquisition* (1st ed., pp. 96–121). Cambridge University Press.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raviv, L., Lupyan, G., & Green, S. C. (2022). How variability shapes learning and generalization. *Trends in Cognitive Sciences*, *26*(6), 462–483. https://doi.org/10.1016/j.tics.2022.03.007

Richter, K. (2021). *Educational outcomes in multilingual CLIL school settings: A systematic review.* [Master's thesis, University of Oxford]. https://ora.ox.ac.uk/objects/uuid:52221dda-7655-4771-ad7b-66f8c3ee23ca

Rixon, S. (2013). *British Council Survey of Policy and Practice in Primary English Language Teaching Worldwide*. British Council.

Robinson, P., & Ellis, N. (Eds.). (2008). *Handbook of cognitive linguistics and second language acquisition*. Routledge, Taylor & Francis Group.

Roehr-Brackin, K. (2014). Explicit Knowledge and Processes From a Usage-Based Perspective: The Developmental Trajectory of an Instructed L2 Learner. *Language Learning*, *64*(4), 771–808. https://doi.org/10.1111/lang.12081

Roehr-Brackin, K. (2018). *Metalinguistic awareness and second language acquisition*. Routledge, Taylor & Francis Group.

Saffran, J. R. (2001). The Use of Predictive Dependencies in Language Learning. *Journal of Memory and Language*, *44*(4), 493–515. https://doi.org/10.1006/jmla.2000.2759

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Sambanis, M. (2007). *Sprache aus Handeln: Englisch und Französisch in der Grundschule*. Empirische Pädagogik.

Sandoval, M., & Gómez, R. L. (2013). The development of nonadjacent dependency learning in natural and artificial languages. *WIREs Cognitive Science*, *4*(5), 511–522. https://doi.org/10.1002/wcs.1244

Schulz, J., Hamilton, C., Wonnacott, E., & Murphy, V. (2023). The impact of multi-word units in early foreign language learning and teaching contexts: A systematic review. *Review of Education*, *11*(2). https://doi.org/10.1002/rev3.3413

Schwandtke, K. (2021). *Die Grammatik- und Wortschatzkenntnisse von Englischlernenden am Ende der Jahrgangsstufe 4: Eine sprachdidaktische Auswertung der im Rahmen der BIG-Studie erhobenen dialogischen Sprachdaten* (1. Auflage). Cuvillier Verlag.

Sethuraman, N., & Goodman, J. (2004). Children's mastery of the transitive construction. In E. Clark (Ed.), *Online proceedings of the 32nd session of the Stanford Child Language Research Forum* (pp. 60–67). CSLI Publications.

Silvey, C., Dienes, Z., & Wonnacott, E. (under review). *Bayes factors for logistic (mixed effect) models*. https://doi.org/10.31234/osf.io/m4hju

Siyanova-Chanturia, A. (2017). Researching the teaching and learning of multi-word expressions. *Language Teaching Research*, *21*(3), 289–297. https://doi.org/10.1177/1362168817706842

Slavin, R. E. (1986). Best-Evidence Synthesis: An Alternative to Meta-Analytic and Traditional Reviews. *Educational Researcher*, *15*(9), 5–11. https://doi.org/10.3102/0013189X015009005

Smith, S. A., & Murphy, V. A. (2015). Measuring productive elements of multi-word phrase vocabulary knowledge among children with English as an additional or only language. *Reading and Writing*, *28*(3), 347–369. https://doi.org/10.1007/s11145-014-9527-y

Sparks, R., S. Dale, P., & M. Patton, J. (2023). Individual differences in L1 attainment and language aptitude predict L2 achievement in instructed language learners. *The Modern Language Journal*, *107*(2), 479–508. https://doi.org/10.1111/modl.12841

Sundqvist, P., & Sylvén, L. K. (2016). *Extramural English in Teaching and Learning: From Theory and Research to Practice*. Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-46048-6

Teinonen, T., Fellman, V., Näätänen, R., Alku, P., & Huotilainen, M. (2009). Statistical language learning in neonates revealed by event-related brain potentials. *BMC Neuroscience*, *10*(1), 21. https://doi.org/10.1186/1471-2202-10-21

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *The Behavioral and Brain Sciences*, *24*(4), 629–640; discussion 652-791. https://doi.org/10.1017/s0140525x01000061

Thomas, J., Harden, A., Oakley, A., Oliver, S., Sutcliffe, K., Rees, R., Brunton, G., & Kavanagh, J. (2004). Integrating qualitative research with trials in systematic reviews. *BMJ*, *328*(7446), 1010–1012. https://doi.org/10.1136/bmj.328.7446.1010

Thomson, H. (2020). *Developing fluency with multi-word expressions.* [Doctoral dissertation]. Victoria University, NZ. https://researcharchive.vuw.ac.nz/handle/10063/9373

Tode, T. (2003). From unanalyzed chunks to rules: The learning of the English copula be by beginning Japanese learners of English. *IRAL - International Review of Applied Linguistics in Language Teaching*, *41*(1). https://doi.org/10.1515/iral.2003.002

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Tomasello, M. (2009). The usage-based theory of language acquisition. In E. L. Bavin (Ed.), *The Cambridge Handbook of Child Language* (pp. 69–88). Cambridge University Press.

Twomey, K. E., Ranson, S. L., & Horst, J. S. (2014). That's More Like It: Multiple Exemplars Facilitate Word Learning: Multiple Exemplars Facilitate Word Learning. *Infant and Child Development*, *23*(2), 105–122. https://doi.org/10.1002/icd.1824

Tyler, A. (2010). Usage-Based Approaches to Language and Their Applications to Second Language Learning. *Annual Review of Applied Linguistics*, *30*, 270–291. https://doi.org/10.1017/S0267190510000140

Van Koert, M., Leona, N., Rispens, J., Tijms, J., Molen, M. V. D., Grunberg, H. L., & Snellings, P. (2023). English Grammar Skills in Dutch Grade 4 Children: Examining the Relation Between L1 and L2 Language Skills. *Journal of Psycholinguistic Research*, *52*(5), 1737–1753. https://doi.org/10.1007/s10936-023-09968-x

Viviani, E., Ramscar, M., & Wonnacott, E. (n.d.). Stage 1 registered report: *Go above and beyond: Does input variability affect children's ability to learn spatial adpositions in a novel language?* https://osf.io/3q4yg/

Wahlheim, C. N., Finn, B., & Jacoby, L. (2012). Metacognitive judgments of repetition and variability effects in natural concept learning: Evidence for variability neglect. *Memory & Cognition*, *40*, 703–716.

Wechsler, D. (1992). *Wechsler individual achievement test.* Psychological Corporation.

Wechsler, D. (2011). *Wechsler Abbreviated Scale of Intelligence – Second Edition* (WASI-II) [Database record]. APA PsycTests. https://doi.org/10.1037/t15171-000

Werlen, E. (2008). *Schlussbericht der Wissenschaftlichen Begleitung WiBe der Pilotphase Fremdsprache in der Grundschule.* (KM-BW) Kultusministerium Baden-Württemberg. https://s.schulamt-bw.de/site/pbs-bw-new/get/documents/KULTUS.Dachmandant/KULTUS/kultusportal-bw/zzz_pdf/Internentfassung.pdf

Willis, S., Neil, R., Mellick, M. C., & Wasley, D. (2019). The Relationship Between Occupational Demands and Well-Being of Performing Artists: A Systematic Review. *Frontiers in Psychology*, *10*, 393. https://doi.org/10.3389/fpsyg.2019.00393

Wilson, B., Spierings, M., Ravignani, A., Mueller, J. L., Mintz, T. H., Wijnen, F., Van Der Kant, A., Smith, K., & Rey, A. (2020). Non-adjacent Dependency Learning in Humans and Other Animals. *Topics in Cognitive Science*, *12*(3), 843–858. https://doi.org/10.1111/tops.12381

Wonnacott, E., Boyd, J. K., Thomson, J., & Goldberg, A. E. (2012). Input effects on the acquisition of a novel phrasal construction in 5year olds. *Journal of Memory and Language*, *66*(3), 458–478. https://doi.org/10.1016/j.jml.2011.11.004

Wood, D. (2009). Effects of focused instruction of formulaic sequences on fluent expression in second language narratives: A case study. *Canadian Journal of Applied Linguistics*, *12*(1), 39–57.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press.

Wulff, S. (2018). Acquisition of formulaic language from a usage-based perspective. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding formulaic language: A second language acquisition perspective* (pp. 19–37). Routledge, Taylor & Francis Group.

Wulff, S., & Ellis, N. (2015). Usage-based approaches to SLA. In *Theories in Second Language Acquisition: An Introduction* (pp. 75–93). Routledge, Taylor & Francis Group.

Year, J. (2009). *Korean speakers' acquisition of the English ditransitive construction: The role of input frequency and distribution* [Doctoral dissertation]. Columbia University.

Year, J., & Gordon, P. (2009). Korean Speakers' Acquisition of the English Ditransitive Construction: The Role of Verb Prototype, Input Distribution, and Frequency. *The Modern Language Journal*, *93*(3), 399–417. https://doi.org/10.1111/j.1540-4781.2009.00898.x

Yuldashev, A., Fernandez, J., & Thorne, S. L. (2013). Second Language Learners' Contiguous and Discontiguous Multi-Word Unit Use Over Time. *The Modern Language Journal*, *97*, 31–45. https://doi.org/10.1111/j.1540-4781.2012.01420.x

Zipf, G. K. (1935). *The psycho-biology of language.* Houghton Mifflin Harcourt.

7 Appendix

7.1 Sampling Process

Given my intent to conduct the teaching during the intervention myself, the languages involved in the experiment were restricted to English and German. In addition, the participating students should already have had some prior instructed exposure to the target FL since the input structures I intended to target required some level of prior FL knowledge on the part of the participating students. Therefore, British primary schools that provided German FL instruction were considered. Searches showed that only one school in England does so, and it turned out that I was already acquainted with their Head of Pre-Prep Languages in charge of German teaching through the Goethe Institute in London. This teacher of German as a Foreign Language (GFL), as well as the pre-preparatory school's headmistress and the whole school's principal, liked the idea of conducting FL research with their pre-prep students.

7.2 Recruitment documents

Name – Principal

Name – Headmistress of the Pre-Preparatory School

Name – Head of Pre-Prep Languages

School Name

Pre-Preparatory School & Nursery

Address Line 1

Address Line 2

Address Line 3

Address Line 4

12.10.2022

The impact of multi-word units in early foreign language learning and teaching contexts

Ethics Approval Reference: [CIA – 22TT - 135]

Dear Mr X, dear Mrs X, dear Mrs X,

I am writing to enquire about conducting some research in your school in the academic year 2022/2023. I am a German doctoral research student at the University of Oxford's Department of Education, supervised by Prof Dr Victoria Murphy and Dr Elizabeth Wonnacott, and funded by the Department of Education. In my research study, *The impact of multi-word units in early foreign language learning and teaching contexts*, I explore how targeted teaching-manipulations in foreign language classes affect primary school students' foreign language development.

The research would take place with your Year 2 German students, in collaboration with your Head of Pre-Preparatory School Languages, X. I visited Mrs X and her classes during an informal school visit in June 2022. During that visit, I also met with Mrs X. Your school's Year 2 classes would be ideal for our research purposes because they have already had consistent German instruction at a young age, and they are very enthusiastic about learning German.

The commitment from the school would be to allow me to teach the Year 2's German lessons over a period of 6 weeks. I am a qualified German and English language teacher (B. Ed., 2020, University of Tübingen, Germany) and Mrs X would be always present during lessons. As part

of this teaching intervention, we will implement a few changes to the teaching approach while following your school's normal curriculum. We as researchers from the University of Oxford can assure you that no child will be at an educational disadvantage, as approved by the University of Oxford's Central Research Ethics Committee. By participating in the research, your school would be contributing to research that will advance current language learning theory and contribute to foreign language pedagogy and policy development.

Before and after the 6-week teaching period, the Year 2 students would be asked to complete some language assessment tests, such as vocabulary tests or metalinguistic skills tests. For the students, those tests will be presented as fun language games in an engaging and child-appropriate manner.

The University of Oxford has strict ethical procedures on conducting ethical research with teachers and students, consistent with current British Educational Research Association guidelines. Before beginning the research and in collaboration with your school, I would inform parents/guardians about the research and offer the students and parents/guardians the opportunity to refuse to participate. In this context, it would be your choice whether we use an 'opt-in' procedure (i.e., parents have to actively agree) or an 'opt-out' procedure (i.e., parents have to actively disagree). Throughout the research, students and parents/guardians will be able to withdraw from the study at any time.

All participants, including students, teachers and the school, would be made anonymous in all research reports. We would not collect any personal data about the students. The data collected would be kept strictly confidential, available only to my supervisors, to researchers associated with their research groups, and to myself and not used other than specified without the further consent of all involved being obtained.

I have requested an enhanced DBS check via your school's administration. Mrs X from your administration team has provided me with access to all teaching seminars that your school requires their teachers to undertake. I will complete these courses prior to the study.

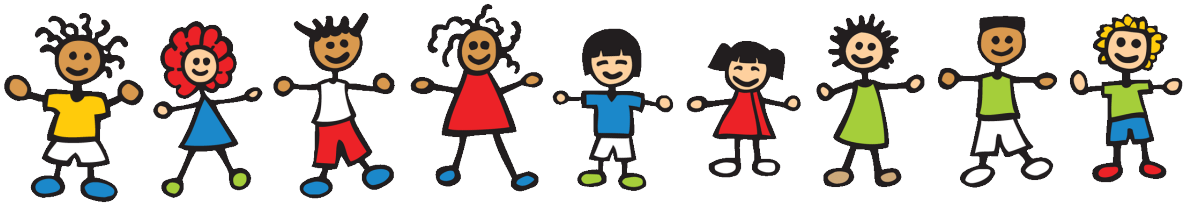
Please find enclosed copies of the information for parents/guardians and students as well as different options of 'opt-in' and 'opt-out' forms for the parents.

If your school would like to take part in the study, or you need more information about what is involved, please contact me or my supervisors.

Thank you for your time and attention. I look forward to hearing from you.

Yours Sincerely,

Johannes Schulz, MSc, B.Ed



Learning German at school

INFORMATION SHEET FOR STUDENTS AGED 6 TO 10 YEARS

Ethics Approval Reference: [CIA – 22TT - 135]

Guten Tag liebe Kinder!

My name is Johannes Schulz, and I am a researcher. Researchers work at universities and in laboratories, and we try to find out the answers to complicated questions. Currently, we are trying to find out how it can be that you all learn German so well although you haven't lived in Germany before. To find out more about this very interesting question, I am doing some research and would like you, together with the rest of your German class, to join in this study. Together, we can come one step closer to answering this question, and perhaps help future generations of students like you to learn as much German as you!

I have also written to your parents to tell them about this research and asked them to think about whether you should be included. Please talk to your parents about what you would like to do and, if you are unhappy, let me or Mrs X know.

Why have I been asked?

I am asking you if you would like to take part in my study because you are in Year 2, studying German. I am asking all the students in your class to help us. We want to learn from all of you how it is that you learn German so well!

Do I have to join in?

No, you don't have to if you don't want to. You can ask questions before choosing whether you want to join in.

You can change your mind at any time by telling Mrs X, myself, or your parents. You don't have to say why, and this will not affect your education.

If you decide to stop, no one will be upset with you, and if you do not wish to join in, you will attend a different class during German lesson time.

What will happen?

For four weeks, I will take over the German teaching from Mrs X. We will continue with your classes as usual, and Mrs X will always be around. You will study in the same way as you are used to, but with two teachers in the room. At the beginning and at the end of the four weeks, you will complete some fun language tasks in German and play a few language games.

Will anything about the research upset me?

No, nothing about this research should upset you. Your German teacher, Mrs X, will always be around and available for you if you have any concerns. Other researchers and your headmistress Mrs X have checked this research and they all said it's ok for you to participate.

Will joining in help me?

For a few weeks, you will have two native German speakers teaching you German, that's a wonderful opportunity to learn even more German! And in addition to that, the research might help other students in the future.

Will anyone else know I'm doing this?

The people in our research team and the school will know you are taking part. No one will know that you have helped us with this research - unless, of course, you tell them yourself!

What happens to what the researchers find out?

When I collect information from you, I will keep it in a safe place and only the people doing the research, or helping with the research, can look at it. I will use the information to discuss with other researchers and teachers, some of it in written form. Ideally, we will all contribute to helping future generations of students to learn as much German and other foreign languages as you do!

Is this study OK to do?

Before any research involving people happens, it has to be checked by a group of people known as a *Research Ethics Committee* to make sure that the study is fair. 'Ethics' means that something is ok and nice to do, and nobody is being harmed. A 'Committee' is a group of smart people who discuss a lot and make decisions about whether something is ok. So, basically, the Research Ethic Committee's job is it to look after you and keep you save. They checked this study in detail and said it's ok for you to participate.

What do I do now?

Please tell your parents or Mrs X whether you are happy to take part. I hope you will agree to take part in my study.

What if there is a problem or something goes wrong?

If you are not happy because of something that happened in the study, please talk to your parents, to Mrs X, or to any other member of staff in school.

They will talk to me.

Thank you! Dankeschön!

Thank you for reading – please ask me any questions.

ASSENT FORM/ORAL SCRIPT FOR CHILDREN UNDER 16

Ethics Approval Reference: [CIA – 22TT - 135]

The impact of multi-word units in early foreign language learning and teaching contexts

Student (or if unable, parent/researcher/teacher on their behalf) to circle all they agree with:

- Has your teacher explained this project to you? Yes / No
- Do you understand what this project is about? Yes / No
- Have you asked all the questions you want? Yes / No
- Have you had your questions answered in a way you understand? Yes / No
- Do you understand it's OK to stop taking part at any time? Yes / No
- Are you happy to take part? Yes / No

If any answers are "no" or you don't want to take part, that's OK! No one will be cross with you.

If you do want to take part, please write your name below.

Your name _____

Date _____

The researcher who explained this project to you needs to sign too:

Print Name _____

Signature _____

Date _____

The impact of multi-word units in early foreign language learning and teaching contexts

INFORMATION SHEET FOR PARENTS / GUARDIANS

Ethics Approval Reference: [CIA – 22TT - 135]

In partnership with researchers at the University of Oxford, School name Pre-Preparatory School has agreed to take part in a research study. We would like to invite your child, along with the rest of their Year 2 German class, to be involved in this study, which is focussed on the general language performance of the entire class, not on individual students. We very much hope you would like your child to be involved, but before you decide, it is important that you understand why the study is being done.

What are we trying to find out?

Current Applied Linguistics research strongly suggests that so-called 'multi-word units' (also referred to as *phrases*, *formulas*, *chunks*) play a vital role in native and foreign language acquisition. Common multi-word units include conventionalized phrases such as "How are you?" and more abstract yet well-known constructions such as "X gives the Y to Z". Large parts of everyday speech in any language are made up of such multi-word units and the scientific consensus is that they function as 'catalysts' for language learning.

Primary school foreign language teaching is on the rise worldwide since proficiency in more than one language is essential in today's societies. It is therefore crucial to focus research efforts on multi-word units as 'engines' of language development to make teaching most effective and improve learning outcomes. However, most research in this area with children has been done in artificial settings, for example in psycholinguistic laboratories. Together with your children, we are conducting the first real-world classroom-based study on this topic to investigate the phenomenon in a context where actual language learning is happening on a daily basis.

More information about the study can be obtained by contacting the main researcher of this project, Johannes Schulz (see contact details below). We are happy to answer any questions.

Why has my child been invited to be involved in this research?

We are inviting your child because s/he is in Year 2 at School name studying German. All

students in her/his class are invited to be involved in this research.

Does my child have to be involved?

You can ask questions about the study before deciding whether to allow your child to be involved. If you do not agree to their involvement, you may withdraw your child at any time, without giving a reason and without any effect on their education, by advising the school accordingly.

What will my child be asked to do?

A member of our research team, Johannes Schulz, will take over the German lessons in Year 2 for a period of four weeks in March 2023. Johannes is a qualified German and English language teacher from Germany, and a doctoral student in Applied Linguistics at the University of Oxford. He has worked in a British primary school in the past. Together with Mrs X, your child's current German teacher, Johannes will prepare four weeks of German teaching with an increased focus on multi-word units, following the school's normal curriculum. He will be delivering the teaching while Mrs X will be in the room at all times.

Before and after the four-week teaching period, we will ask your child to play some language games and complete some language-related tasks, delivered in a child-appropriate and engaging manner. For example, the students will be playing the 'Language Magician' game, a computer-based language game for primary school students developed by the European Union and the University of Reading, UK. In addition, we will conduct a cognitive aptitude test with the class before the study.

In February 2023, Johannes Schulz will visit the Year 2 German lessons regularly to support students getting used to a new face and to establish a good rapport with them. During that time, he will also pilot with some students some of the games and tasks used in the main study, to ensure everything will run smoothly.

What are the advantages / disadvantages of taking part?

Taking part in our study does not put your child at any educational risk, as approved by the University of Oxford's Central Research Ethics Committee. We follow the school's normal curriculum, and your child will keep learning German as usual. It is likely that your child will actually benefit from having two native German teachers in class at once.

What happens to the data provided?

Any information your child provides during the study is the research data. No personal data about your child will be collected at any point. Each child will be assigned a randomly generated ID number which we will use to administer the games and tasks. Only the school will have access to a combined list of both the pupils' names and random ID numbers. We, the researchers, will not have access to this list.

Opt-out forms will be retained by the school for the duration of the study, and for as long as the school determines appropriate after research activities have concluded at the school.

Only Johannes Schulz, his supervisors Prof Dr Victoria Murphy and Dr Elizabeth Wonnacott, and members of their research groups will have access to the research data. Responsible members of the University of Oxford may be given access to data for monitoring and/or audit of the research.

We will send a brief report on the research to your child's school at the end of the study, and you are welcome to see this. We will not be able to provide you with feedback on individual results of your child as the data we collect is anonymized. Also, we will not identify the school, teacher or any students in any reports of the research.

Will the research be published?

The research may be published, for example in peer-reviewed academic journals. It will also be made available as part of a doctoral thesis. On successful submission of the thesis, it will be deposited both in print and online in the University archives to facilitate its use in future research. If so, the thesis will be openly accessible. The University of Oxford is committed to the dissemination of its research for the benefit of society and the economy and, in support of this commitment, has established an online archive of research materials. This archive includes digital copies of student theses successfully submitted as part of a University of Oxford doctoral degree programme. Holding the archive online gives easy access for researchers to the full text of freely available theses, thereby increasing the likely impact and use of that research.

Who is conducting this research?

The research is organised by Johannes Schulz of the University of Oxford, who is a doctoral research student supervised by Prof Dr Victoria Murphy and Dr Elizabeth Wonnacott. He has a current enhanced DBS certificate. The research is funded by the Department of Education.

Ethics

This study has been reviewed by, and received ethics clearance through, the University of Oxford's Central University Research Ethics Committee [reference number: CIA – 22TT - 135]. Any research with students will be conducted with care and sensitivity as some students might feel shy about having a researcher in the classroom.

What if there is a problem?

If you have a concern about any aspect of this study, please contact Johannes Schulz (johannes.schulz@education.ox.ac.uk) or Prof Dr Victoria Murphy (victoria.murphy@education.ox.ac.uk) or Dr Elizabeth Wonnacott (elizabeth.wonnacott@education.ox.ac.uk) and we will do our best to answer your query. We will acknowledge your concern within ten working days and give you an indication of how it will be dealt with. If you remain unhappy or wish to make a formal complaint, please contact the Chair of the Research Ethics Committee at the University of Oxford who will seek to resolve the matter as soon as possible:

Chair, Social Sciences & Humanities Inter-Divisional Research Ethics Committee; Email: ethics@socsci.ox.ac.uk; Address: Research Services, University of Oxford, Wellington Square, Oxford OX1 2JD

Data Protection

No personal or identifiable data about your child will be collected at any point during the study.

Further information about your rights with respect to personal data is available from http://www.admin.ox.ac.uk/councilsec/compliance/gdpr/individualrights/.

What should I do next?

You do not have to take any action if you agree for your child to participate in this research. If you would not like your child to take part in this study, please let Frau X know in writing.

If you would like to discuss the research with someone beforehand (or if you have questions afterwards), please contact Johannes Schulz.

Thank you very much for considering to let your child take part in our research.

Yours sincerely,

Johannes Schulz

Department of Education

University of Oxford

15 Norham Gardens

Oxford

OX2 6PY

johannes.schulz@education.ox.ac.uk

The impact of multi-word units in early foreign language learning and teaching contexts

OPT-OUT FORM PARENTS

Ethics Approval Reference: [CIA – 22TT - 135]

If you DO NOT want your child to be included in the above-named research study please fill out the form below and return it to the school by [dd/mm/yyyy].

If we do not receive an opt-out form from you by this date, your child may be included in this study, as described in the accompanying information sheet.

-

I, the undersigned, hereby DO NOT give permission for my child to be included in the study titled *The impact of multi-word units in early foreign language learning and teaching contexts*.

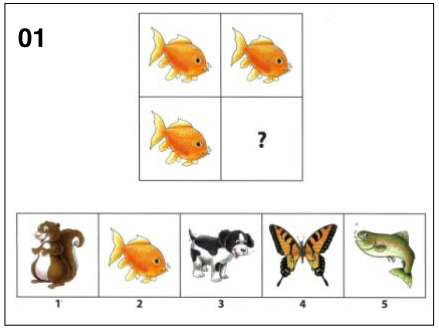
Name of child: _____

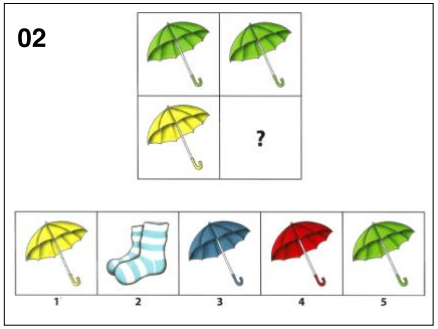
_____

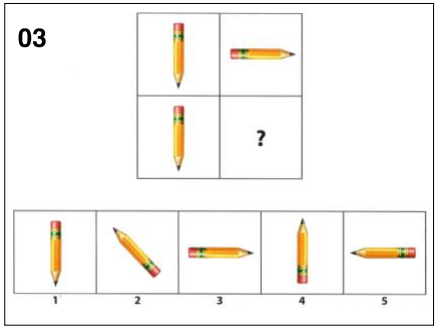
Name of parent/guardian: _____

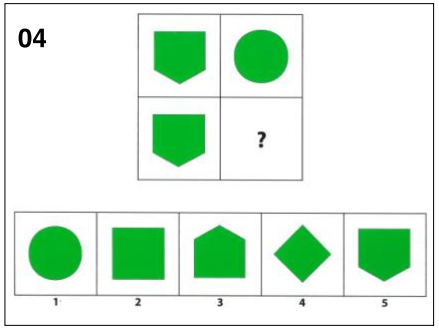
Signature: _____ Date: _____

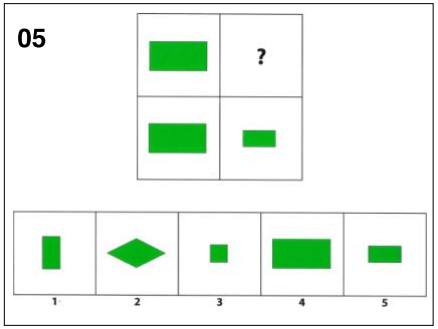
7.3 WASI examples

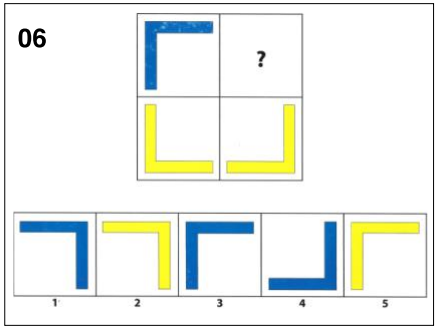












7.4 Experiment 1 – Details on word order of target structure

The subject (noun1) and object (noun2) in the discussed VAC are switched around from an unmarked structure (N_1VN_2; *The camel jumps to the gorilla*) to a marked one (N_2VN_1; *To the gorilla jumps the camel*). Albeit felicitous, this syntactic structure is rare in English, potentially limiting transfer effects. Similar reasoning to limit transfer-effects can be found in Viviani et al. (2022), Gómez (2002) and Wonnacott et al. (2012). While Gómez (2002) and Wonnacott et al. (2012) used artificial language input, Viviani et al. (2022) used Japanese structures that do not exist in English. In contrast, the current approach-structure is not entirely unknown to English speakers. German and English are both West-Germanic languages, so constructions with entirely different global construction semantics are extremely rare. While this makes avoiding transfer-effects challenging, it resembles a natural and unavoidable phenomenon that simply mirrors English L1 children's everyday learning experience when learning German. In fact, in many FL classroom contexts, input structures co-exist in the L1 and the FL, especially if both languages share similar origins. In the current study, if there were potential transfer effects, they would be expected in both groups.

7.5 Experiment 1 – Example videos of animal movements

Seven VACs example videos can be found here. Alternatively, the files have been submitted to the University via the Research Theses Digital Submission application.

schleicht (sneak)

rutscht ('slide on your belly')

huepft (jump)

rollt (roll)

schlittert ('slide on your back')

purzelt ('doing somersaults')

robbt (scoot)

7.6 Experiment 1 – Example sets of test sentences (Outcome measure)

		HV			LV		
task	familiarity	N1	N2	verb	N1	N2	verb
ao	familiar	zebra	baer	schlittert	kamel	baer	huepft
ao	familiar	baer	gorilla	schlittert	baer	elefant	huepft
ao	familiar	gorilla	kamel	schlittert	elefant	gorilla	huepft
ao	familiar	kamel	elefant	schlittert	gorilla	zebra	huepft
p	familiar	elefant	giraffe	schlittert	zebra	giraffe	huepft
p	familiar	giraffe	zebra	schlittert	giraffe	kamel	huepft
p	familiar	baer	elefant	schlittert	baer	elefant	huepft
p	familiar	kamel	baer	schlittert	kamel	giraffe	huepft
fc	familiar	zebra	kamel	schlittert	gorilla	zebra	huepft
fc	familiar	giraffe	zebra	schlittert	elefant	baer	huepft
fc	familiar	gorilla	giraffe	schlittert	giraffe	kamel	huepft
fc	familiar	elefant	gorilla	schlittert	zebra	gorilla	huepft
ao	unfamiliar	zebra	baer	rollt	gorilla	zebra	rutscht
ao	unfamiliar	gorilla	kamel	rollt	kamel	giraffe	rutscht
ao	unfamiliar	kamel	elefant	rollt	baer	elefant	rutscht
ao	unfamiliar	giraffe	zebra	rollt	elefant	baer	rutscht
p	unfamiliar	gorilla	giraffe	schleicht	kamel	baer	schleicht
p	unfamiliar	baer	gorilla	schleicht	giraffe	kamel	schleicht
p	unfamiliar	giraffe	zebra	schleicht	zebra	giraffe	schleicht
p	unfamiliar	baer	elefant	schleicht	baer	elefant	schleicht
fc	unfamiliar	elefant	gorilla	rutscht	giraffe	kamel	rollt
fc	unfamiliar	zebra	kamel	rutscht	zebra	gorilla	rollt
fc	unfamiliar	kamel	baer	rutscht	elefant	gorilla	rollt
fc	unfamiliar	elefant	giraffe	rutscht	gorilla	zebra	rollt

Table 14 Example test sentences for each one child coming from either the LV or the HV condition. N1 and N2 represent the noun order in the test sentences, i.e., Zum N1 V(erb) the N2. Regarding the tasks, 'ao' refers to act out comprehension, 'p' refers to production, and 'fc' refers to forced choice tasks.

7.7 Experiment 1 – Pilot and pilot results

7.7.1 Pre-Tests

Although the pre-tests were already established measurements, two of them were piloted, the WASI and the PVST, because their testing format had been adapted for the current study. There was no need to pilot the Language Magician as the teacher and the children had already played the game in the past and were well acquainted with the procedure. The WASI and the PVST, in their adapted forms, were piloted with three 10-year-old children of my friends and families, thus, they were slightly older than the target population. The main aim of the pilot was to see whether the children could navigate the tests and to see which instructions they needed to do so successfully. In addition, as the PVST was a computer-based test, the pilot was an opportunity to check whether the vocabulary test was programmed correctly without bugs.

Regarding the WASI, it was found that it would be beneficial to have all children move onto the next item simultaneously to keep an overview of everyone's progress and keep the procedure structured. In addition, it was found that the children required some sort of signal to tell them when to move on to the next item as a group. As a result, a triangle signal was introduced for the real pre-testing.

For the PVST, the pilot children expressed that they required a short break halfway through the test to keep up motivation and concentration. Therefore, a break was implemented into the testing environment where children were praised for their efforts and asked to try and remain concentrated for the remainder of the test.

7.7.2 Outcome measurement

Children from the two participating Year 2 classes could not be used for piloting the outcome measurement because those participants would have had to be excluded from the main experiment. Given the already low number of participants, the outcome measurements were piloted with Year 3 children who were one year older than the target population. Another obstacle regarding the piloting of outcome measurements was the amount of input. Children in experiment 1 should encounter 96 sentences during the 6 exposure days. It was deemed unfeasible to expose the Year 3 pilot children to 96 sentences within a couple of minutes and immediately test them afterwards. To circumvent this problem, the four strongest performing children in Year 3 were recruited and exposed to only a subset of the original input sentences.

It was clear that this piloting procedure might not capture sustainable learning effects since the exposure phase was too short and included too much input. Yet, the focus of the piloting was mainly on the testing procedure and environment, specifically, how long the testing would take per child, whether the testing with toy animals was engaging for the children, whether children responded to the test sentences at all, whether they understood the tasks, and whether they could concentrate in the busy testing environment which was a corridor in the school building.

On the day of piloting, four Year 3 children were taken out of class individually. The testing took place on a table in the corridor during lesson time. In the beginning, each child was introduced to the six animals and received instructions for the piloting. Then, I read and acted out a set of 16 exposure sentences. The procedure was the same as during the main experiment.

The set of 16 sentences was the same for each child, however, the number of verbs featuring in the sentences was different between children. Two children received LV input featuring one verb only (*huepft*), and two children received HV input featuring four verbs (*huepft*, *schlittert*, *robbt*, *purzelt*). Each of the four HV verbs appeared four times across the 16-sentences exposure set. As in the main experiment, the distribution of the six animals across the exposure sentences followed a balanced randomisation.

The total exposure consisted of the set of 16 sentences repeated for four times. Note that the order of sentences was randomized in each set. Thus, each child encountered each individual sentence four times during exposure, and 64 sentences in total. This number of sentences was deemed a reasonable compromise, balancing the objective of not subjecting the children to the entirety of 96 sentences in a single session (i.e., equivalent to the number of exposure sentences during the main experiment) with the need to provide them with a sufficient number of sentences to observe potential learning effects.

After exposure, the children took a short one-minute break to have a drink, before continuing with the outcome measurements. Each child had to complete four familiar act out comprehension trials, four familiar production trials, four familiar forced choice trials, four unfamiliar act out comprehension trials, four unfamiliar production trials, and four unfamiliar forced choice trials. Like in the main experiment, familiar test sentences contained the one verb that was familiar to both input groups (i.e., *huepft*) and the unfamiliar test sentences contained the three remaining unfamiliar verbs, always one verb per task type.

The piloting showed that using toy animals was engaging for the children. They enjoyed participating and playing with the animals. Although there was some unavoidable disturbance by other students or staff who passed through the hallway during exposure and testing, concentration levels remained high throughout the entire procedure. All children were able to understand the task instructions and act accordingly during testing.

During piloting, each test sentence was scored with 1 (correct) or 0 (incorrect) on a spreadsheet. The different types of errors the children made were recorded as well, including acting out or articulating the wrong noun order, or acting out or articulating the wrong verb, or different combinations of both. Importantly, the piloting showed that it was crucial to have pre-defined error categories on hand as it was challenging to keep track of the type of error while listening to the child's response and preparing the next test sentence without losing track of what the child was doing or saying. To circumvent this problem in the main experiment, a comprehensive list of potential error categories was devised, each category was assigned a unique identifier, and this list was employed for recording outcomes during the main experiment. Consequently, when a child erred, only the corresponding error code had to be noted down instead of composing a full description of the mistake.

The results from the piloting are presented in Table 15 below. They are not discussed as the data is not sufficiently meaningful given the constraints on the exposure phase and the low number of participants. Interestingly, the general trend of the limited data already conformed to the predictions in that high input variability seemed to have been beneficial to generalization.

7.7.3 Pilot Results

condition	task type	familiarity	mean	median
HV	ao	familiar	0.88	1.0
HV	p	familiar	1.00	1.0
LV	ao	familiar	1.00	1.0
LV	p	familiar	1.00	1.0
HV	ao	unfamiliar	0.62	1.0
HV	p	unfamiliar	1.00	1.0
LV	ao	unfamiliar	0.50	0.5
LV	p	unfamiliar	0.50	0.5

Table 15 Experiment 1 (VACs) piloting results. N = 3.

7.8 Experiment 1 – Correlation analyses between pre-tests and outcome measurement

Before modelling the experiment 1 data, it had to be determined whether there was any indication in the data to justify including pre-test scores as fixed effects in the statistical models. Correlation analyses between results of the individual pre-tests and the outcome measurements were conducted. Plots of the correlations between pre-test raw scores and outcome measure results including linear regression lines and SE are presented in Figures 5-7 below. Additional correlation analyses for 'structure' correct responses were conducted as well (i.e., where responses were coded 'correct' regardless of verb accuracy). Those are not presented, as the qualitative results did not change.

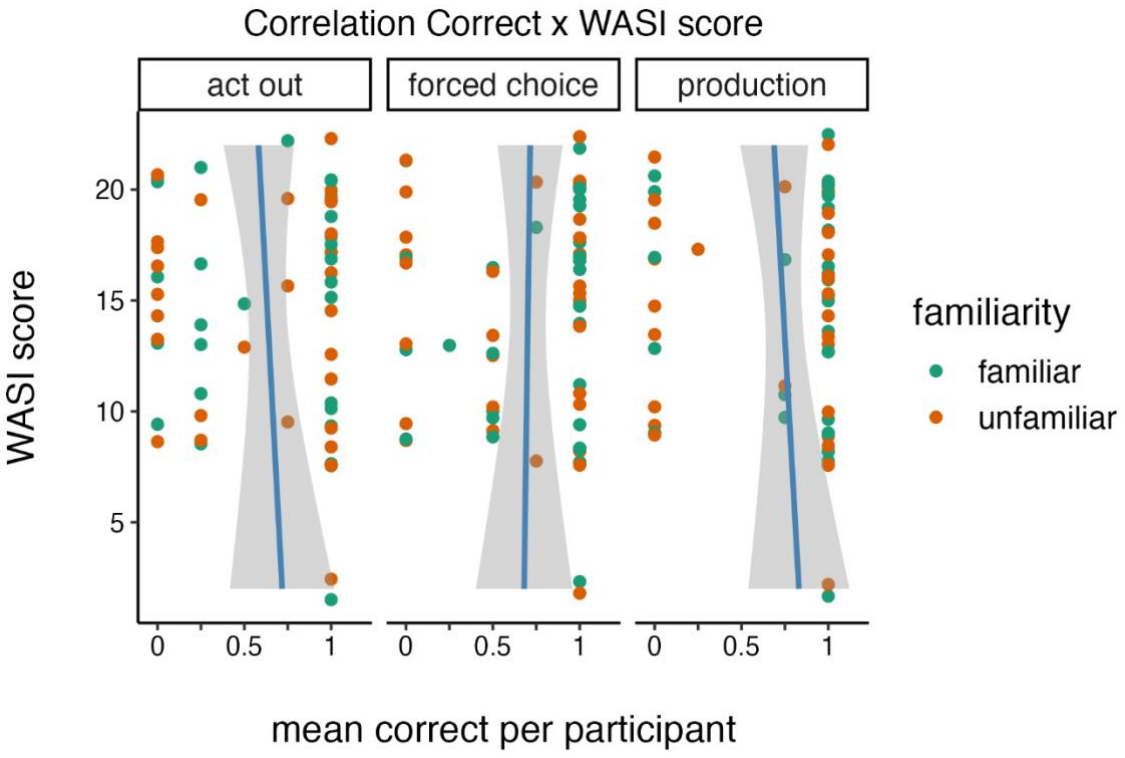


Figure 5 Correlations between participant mean scores on the WASI (pre-test) and participant mean scores on the three outcome measurements (subdivided by familiarity). 'Correct' responses in outcome measurements entail responses where both the linking rule and the verb were correct.

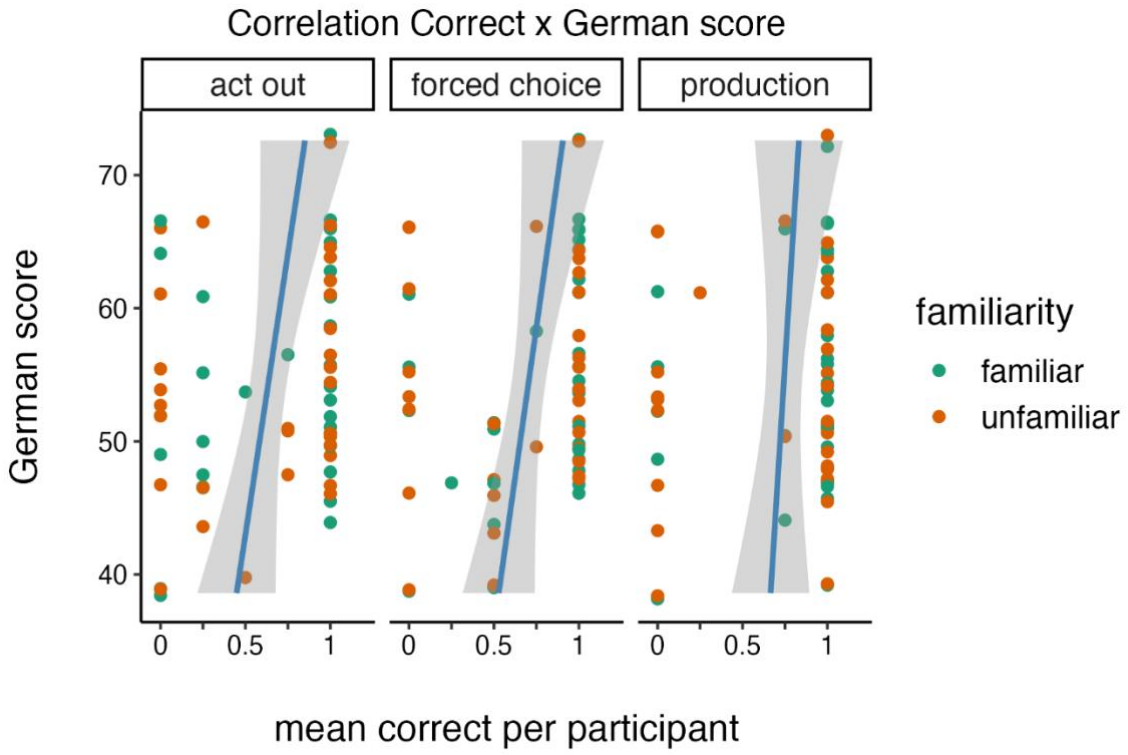


Figure 6 Correlations between participant mean scores on the Language Magician (pre-test) and participant mean scores on the three outcome measurements (subdivided by familiarity). 'Correct' responses in outcome measurements entail responses where both the linking rule and the verb were correct.

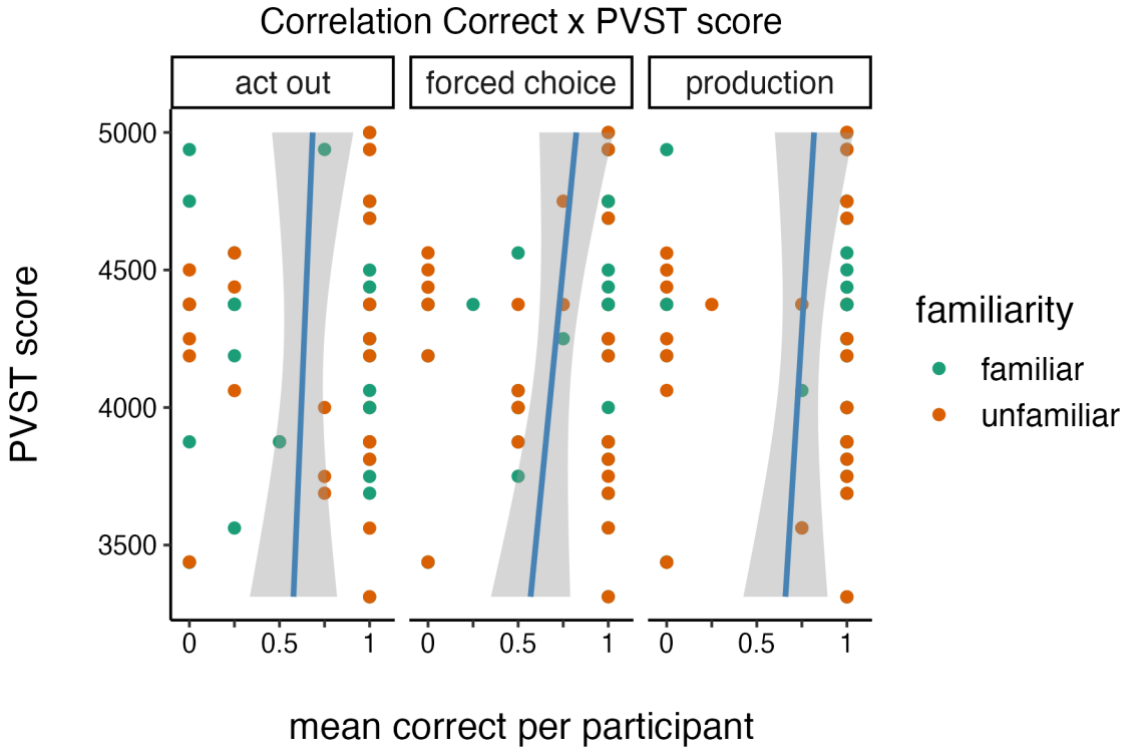


Figure 7 Correlations between participant mean scores on the PVST (pre-test) and participant mean scores on the three outcome measurements (subdivided by familiarity). 'Correct' responses in outcome measurements entail responses where both the linking rule and the verb were correct.

Visual inspection of the plots suggested no obvious associations between any of the pre-test and outcome measurement variables. Additional statistical correlation analyses were conducted (computed over the z-scores of participants' responses in the three outcome measurements). The correlations with performance on familiar verb and unfamiliar verbs were analysed separately (noting that only the latter tap generalization). To reduce the number of analyses computed, scores reflect the average performance on the act out comprehension task and the production task by averaging the z-scores for act out production and comprehension since these conceptually tapped into the same constructs, that is, whether participants inferred the target structure's global construction semantics, linking rule, and verb semantics. Thus, presented below are correlations between the pre-test scores and this averaged z-score and the pre-test scores and z-scores reflecting performance on the forced choice task.

	outcome measurement responses (z-scores)			
	act out + production *familiar*	act out + production *unfamiliar*	forced choice *familiar*	forced choice *unfamiliar*
WASI score	-0.09	-0.08	0.14	-0.04
German score	0.23	0.16	0.36*	0.13
PVST score	0.01	0.18	0.29	0.06

Table 16 Spearman rank correlation coefficients (ρ) of correlations between z-sores of outcome measure responses and raw pre-test scores. Spearman rank correlation ranges from -1 to 1. Coefficients marked with * are statistically significant at $p < .05$. For familiar and unfamiliar act out comprehension and production values (first two columns), the z-scores of participants' responses in act out comprehension and production tasks were added together and averaged. 'Correct' responses in outcome measurements entail responses where the linking rule was applied correctly, irrespective of whether the verb was correct.

Analysis results (Table 16) revealed weak negative and positive correlations. However, only the correlation between German scores and z-scores of responses to familiar trials in the forced choice task ($r = 0.36$) reached the threshold for statistical significance ($p < .05$) (and this would not be the case if the analyses were adjusted for multiple comparisons). Although the absence

of evidence of associations does not equal evidence of no associations between the variables, the generally weak correlation coefficients was not considered a strong enough evidence base to justify the inclusion of pre-test scores in the statistical modelling.

7.9 Analysis scripts and data

The R scripts for the analyses of the experiment 1 and experiment 2 data as well as the relevant data sets can be found online on OSF, here. Alternatively, the documents have been submitted to the University via the Research Theses Digital Submission application.

The script for experiment 1 works with two files:

- data_exp_VACs.csv
- data_ctrl_VACs.csv

Those files contain the raw VACs outcome measure data as well as each child's pre-test scores.

The script for experiment 2 works with one file:

- total_NADs_processed.csv

This file contains the raw NADs outcome measure data as well as each child's pre-test scores.

Note: The file loadFunctionsGithub.R needs to be stored in the same working directory.

7.10 Experiment 1 – Detailed model outcomes

7.10.1 Act out comprehension

	Estimate	SE	Z	p
Fixed effects				
(intercept)	2.23	1.15	1.94	0.052
test day = 1	-5.08	1.86	-2.72	0.0065**
input group = HV	0.52	2.33	0.22	0.822
familiarity = unfamiliar	-1.91	1.86	-1.02	0.306
input group x familiarity	12.36	3.98	3.11	0.0019**
Random effects				
	SD			
participant (intercept)	5.13			
familiarity	7.83			

Table 17 *MODEL 1* Act out comprehension task mixed model. In this model, the dependent variable denotes 'entirely' correct responses, including accurate verb semantics. The estimated coefficients (Estimate) represent the log odds of the event ('correct') occurring for a one-unit change in the predictor, relative to the mean of the centred variable.

familiar	Estimate	SE	Z	p	unfamiliar	Estimate	SE	Z	p
Fixed effects					***Fixed effects***				
(intercept)	2.82	1.36	2.07	0.038	(intercept)	2.02	2.37	0.85	0.39
test day = 1	-0.78	1.88	-0.41	0.68	test day = 1	-8.69	4.54	-1.92	0.06
input group = HV	-5.91	2.89	-2.04	0.04	input group = HV	9.50	5.31	1.79	0.07
Random effects					***Random effects***				
	SD					SD			
participant (intercept)	5.99				participant (intercept)	5.72			

Table 18 *MODELS 1a and 1b* Act out comprehension task simple effect mixed models. In these models, the dependent variable denotes 'entirely' correct responses, including accurate verb semantics. The estimated coefficients (Estimate) represent the log odds of the event ('correct') occurring for a one-unit change in the predictor, relative to the mean of the centred variable.

7.10.2 Production

	Estimate	SE	Z	p
Fixed effects				
(intercept)	14.85	1.71	8.70	< 0.001***
test day = 1	-20.46	3.40	-6.01	< 0.001***
input group = HV	10.56	3.21	3.29	< 0.001***
familiarity = unfamiliar	-10.50	2.55	-4.12	< 0.001***
input group x familiarity	21.75	4.05	5.37	< 0.001***
Random effects				
	SD			
participant (intercept)	30.2			
familiarity	19.1			

Table 19 MODEL 2 Production task mixed model. In this model, the dependent variable denotes 'entirely' correct responses, including accurate verb semantics. The estimated coefficients (Estimate) represent the log odds of the event ('correct') occurring for a one-unit change in the predictor, relative to the mean of the centred variable.

familiar	Estimate	SE	Z	p	unfamiliar	Estimate	SE	Z	p
Fixed effects					*Fixed effects*				
(intercept)	10.69	2.11	5.06	<0.001***	(intercept)	8.71	1.88	4.64	<0.001***
test day = 1	-0.80	2.53	-0.32	0.75	test day = 1	-20.80	4.31	-4.83	<0.001***
input group = HV	-0.51	2.90	-0.18	0.86	input group = HV	20.95	4.68	4.47	<0.001***
Random effects					*Random effects*				
	SD					SD			
participant (intercept)	21.4				participant (intercept)	24.8			

Table 20 MODELS 2a and 2b Production task simple effect mixed models. In these models, the dependent variable denotes 'entirely' correct responses, including accurate verb semantics. The estimated coefficients (Estimate) represent the log odds of the event ('correct') occurring for a one-unit change in the predictor, relative to the mean of the centred variable.

7.10.3 Forced Choice

	Estimate	SE	Z	p
Fixed effects				
(intercept)	2.53	0.82	3.07	0.002
test day = 1	-5.00	1.48	-3.38	< 0.001
input group = HV	2.06	1.41	1.46	0.14
familiarity = unfamiliar	-1.88	1.19	-1.59	0.11
input group x familiarity	3.81	2.08	1.83	0.07
Random effects				
	SD			
participant (intercept)	3.72			
familiarity	4.24			

Table 21 MODEL 3 Mixed model of the likelihood of scoring 'correct' (i.e., applying linking rule correctly) in the forced choice task of experiment 1. The estimated coefficients (Estimate) represent the log odds of the event ('correct') occurring for a one-unit change in the predictor, relative to the mean of the centred variable.

familiar	Estimate	SE	Z	p	unfamiliar	Estimate	SE	Z	p
Fixed effects					*Fixed effects*				
(intercept)	3.19	1.13	2.82	0.005	(intercept)	1.91	1.32	1.44	0.15
test day = 1	-3.72	1.61	-2.32	0.02	test day = 1	-7.37	3.53	-2.09	0.04
input group = HV	0.11	1.56	0.07	0.95	input group = HV	5.34	3.04	1.76	0.08
Random effects					*Random effects*				
	SD					SD			
participant (intercept)	4.09				participant (intercept)	4.98			

Table 22 MODELS 3a and 3b Simple effect mixed models of the likelihood of scoring 'correct' (i.e., applying linking rule correctly) in the forced choice task of experiment 1. The estimated coefficients (Estimate) represent the log odds of the event ('correct') occurring for a one-unit change in the predictor, relative to the mean of the centred variable.

7.11 Experiment 1 – Violin plots with 'structure' correct responses

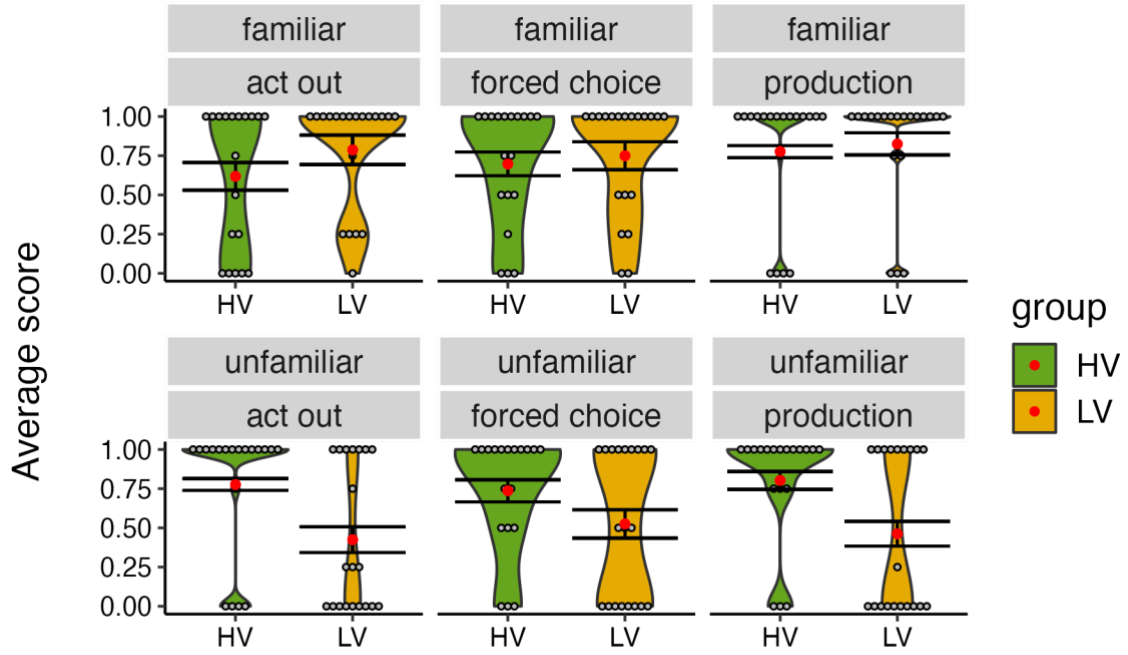


Figure 8 Violin Plots displaying participant mean proportion correct in each of the three task types during VACs (experiment 1) outcome measurements ('correct' scores are responses where the global construction semantics and linking rule are applied correctly, regardless of verb accuracy). HV = high variability. LV = low variability. Error bars present 95% Confidence Intervals.

7.12 Experiment 1 – Additional analyses on 'verb correct' responses

As demonstrated by the error analyses in section 3.2.4, the descriptive data suggested that high input variability might have been detrimental to learning the individual verb's semantics. To further investigate 'verb learning' in the context of the current study, additional analyses were conducted over trials were children got the verb correct (regardless of whether the linking rule was correct, or whether the correct animals were used). This was done only on familiar trials, since during unfamiliar trials, children were presented with the verb semantics only seconds before they had to enact the verbs. As a result, there were almost zero errors in unfamiliar trials across all tasks and input groups. Furthermore, the production task is not part of the analysis since children were given the correct verb during testing (see section 3.1.8), making it challenging to determine if they learned the verb or simply repeated it from hearing it moments ago. The forced choice task is not analysed as it did not require children to enact/produce the verb. The structures of the mixed model reported below was identical to the main analyses. BF calculation was identical as well, except that the prior stemmed from Table 5 in Wonnacott et al. (2012) where the authors report proportions of verb 'correct' (instead of approach 'correct').

For the act out comprehension task, the hypothesis that there would be better verb accuracy in the LV group for familiar verbs was tested (act out comprehension mean proportion 'correct': LV = 0.98, HV = 0.71). Model details are provided in Table 23. The evidence tended towards H0 but was ambiguous (BF = 0.58; predicted effect = 3.10; RR [0; 7.8], β = -4.62, SE = -0.96, z = 0.07, p = 0.33).

act out	Estimate	SE	Z	p
Fixed effects				
(intercept)	9.05	2.71	3.34	< 0.001
test day = 1	2.54	0.67	-2.32	0.50
input group = HV	-4.62	-0.96	0.07	0.33
Random effects				
	SD			
participant (intercept)	9.16			

Table 23 Mixed model of the likelihood of enacting/producing the correct verb in familiar trials irrespective of other potential error types in the act out comprehension task of experiment 1. The estimated coefficients (Estimate) represent the log odds of the event ('correct') occurring for a one-unit change in the predictor, relative to the mean of the centred variable.

After checking the descriptive statistics of the 'verb correct' data (Table 24), it became clear that the result from the mixed model might stem from the fact that only 1 out of 80 responses of children in the LV group were incorrect in the act out comprehension task. This lack of variability in responses within the LV group potentially made it difficult for the model to estimate a meaningful effect (which formed the basis of BF analyses).

	act out comprehension task	
group		
	incorrect	*correct*
HV	22	54
LV	1	79

Table 24 Descriptive statistics (count) of verb accuracy (irrespective of other error types) in the familiar test trials of the act out comprehension task.

The analyses should be interpreted with caution given the lack of response variability, specifically in the LV group, which could have impacted the analyses. Considering descriptive statistics only, see Table 24, it seems like the LV group was better at getting the verb correct which would align with the fact that they experienced higher overall token frequency during exposure which would be expected to positively impact learning of those tokens. Although this is an interesting finding, given the low sample size and low confidence in the analysis results, the analysis should be considered with caution in this regard.

7.13 Experiment 1 – Additional analyses on 'structure'-correct responses (regardless of verb accuracy)

The analyses in the main text were run over 'entirely correct' trials where both the linking rule *and* the verb were correct. The current appendix presents additional analyses over trials which were scored correct if the linking rule was applied correctly, regardless of verb accuracy. Apart from the change in the dependent variable, the analyses were identical to the main analyses. Regarding the BF analyses, the same prior as in the main analyses (see Table 9) were used.

7.13.1 Act out comprehension

Tables 25 and 26 provide the full output from the logistic mixed effects models from which the beta and SE coefficients were extracted to compute Bayes Factors. Here, only statistics relevant for testing the hypotheses are reported (see Table 9). These confirmed that there was substantial evidence for an interaction between input group and familiarity with BF = 7.6 (predicted effect = 3.93, RR = [2; 110], $\beta = 11.59$, SE = 4.28, $z = 2.71$, $p = 0.0068$) indicating that the data is more than 7 times more likely under H1 than under H0. This interaction was broken down statistically by testing evidence for (a) the hypothesis that there would be better performance in the LV group for familiar verbs and for (b) the hypothesis that there would be better performance in the HV group for unfamiliar verbs. Evidence for the former (a) tended in the predicted direction but was ambiguous (i.e., BF not lower than 0.33) with BF = 0.72 (predicted effect = 6.08, RR = [0; 16.8], $\beta = -1.9$, $z = -0.72$, $p = 0.47$). Evidence for the latter (b) was substantial with BF = 3.59 (predicted effect = 1.52, RR = [1.4; 5132], $\beta = 18.43$, $z = 3.96$, $p < 0.001$). Thus, the data is more than 3 times more likely under H1.

	Estimate	SE	z	p
Fixed effects				
(intercept)	3.77	1.42	2.65	0.0081**
test day = 1	-6.29	2.3	-2.73	0.0063**
input group = HV	3.08	2.36	1.30	0.1922
familiarity = unfamiliar	-2.72	2.09	-1.30	0.1933
input group x familiarity	11.59	4.28	2.71	0.0068**
Random effects				
	SD			
participant (intercept)	5.88			
familiarity	7.94			

Table 25 MODEL 1 Act out comprehension task mixed model. In this model, the dependent variable denotes responses where the linking rule was applied correctly, irrespective of verb accuracy. The estimated coefficients (Estimate) represent the log odds of the event ('correct') occurring for a one-unit change in the predictor, relative to the mean of the centred variable.

familiar	Estimate	SE	Z	p	unfamiliar	Estimate	SE	Z	p
Fixed effects					*Fixed effects*				
(intercept)	6.08	4.92	1.23	0.22	(intercept)	7.11	1.88	3.79	< 0.001
test day = 1	-2.28	3.32	-0.69	0.49	test day = 1	-18.07	4.45	-4.06	< 0.001
input group = HV	-1.85	2.57	-0.72	0.47	input group = HV	18.43	4.66	3.96	< 0.001
Random effects					*Random effects*				
	SD					SD			
participant (intercept)	7.89				participant (intercept)	14.4			

Table 26 MODELS 1a and 1b Act out comprehension task simple effect mixed models. In these models, the dependent variable denotes responses where the linking rule was applied correctly, irrespective of verb accuracy. The estimated coefficients (Estimate) represent the log odds of the event ('correct') occurring for a one-unit change in the predictor, relative to the mean of the centred variable.

7.13.2 Production

Analyses of production task data yielded the same qualitative outcomes as the analyses presented in the main text. Tables 27 and 28 provide the full output from the logistic mixed effects models from which the beta and SE coefficients were extracted to compute Bayes Factors. Here, only statistics relevant for testing the hypotheses are reported. These confirmed that there was substantial evidence for an interaction in the predicted direction between input group and familiarity with BF = 368 (predicted effect = 3.93, RR = [0.9; 6641], β = 20.02, z = 4.83, $p < 0.001$) indicating that the data is more than 300 times more likely under H1 than under H0. This interaction was broken down statistically by testing evidence for (a) the

hypothesis that there would be better performance in the LV group for familiar verbs and for (b) the hypothesis that there would be better performance in the HV group for unfamiliar verbs. Evidence for the former (a) was substantial for H0 with BF = 0.28 (predicted effect = 10.71, RR = [0; 9.9], β = -0.22, z = -0.08, p = 0.94). This means that there was evidence of no difference between the two groups. Evidence for the latter (b) was substantial with BF = 4.53 (predicted effect = 1.52, RR = [1.2; 6627], β = 20.95, z = 4.47, p < 0.001). Thus, the data are more than 4 times more likely under H1.

	Estimate	SE	Z	p
Fixed effects				
(intercept)	14.24	1.73	8.22	< 0.001
test day = 1	-19.78	3.35	-5.90	< 0.001
input group = HV	10.39	2.99	3.48	< 0.001
familiarity = unfamiliar	-9.85	2.37	-4.15	< 0.001
input group x familiarity	20.02	4.14	4.83	< 0.001
Random effects				
	SD			
participant (intercept)	26.4			
familiarity	15.5			

Table 27 MODEL 2 Production task mixed model. In this model, the dependent variable denotes responses where the linking rule was applied correctly, irrespective of verb accuracy. The estimated coefficients (Estimate) represent the log odds of the event ('correct') occurring for a one-unit change in the predictor, relative to the mean of the centred variable.

familiar	Estimate	SE	Z	p	unfamiliar	Estimate	SE	Z	p
Fixed effects					*Fixed effects*				
(intercept)	10.71	2.20	4.87	< 0.001	(intercept)	8.71	1.88	4.64	< 0.001
test day = 1	-1.17	2.58	-0.45	0.65	test day = 1	-20.80	4.31	-4.83	< 0.001
input group = HV	-0.22	2.95	-0.08	0.94	input group = HV	20.95	4.68	4.47	< 0.001
Random effects					*Random effects*				
	SD					SD			
participant (intercept)	20.2				participant (intercept)	24.8			

Table 28 MODELS 2a and 2b Production task simple effect mixed models. In these models, the dependent variable denotes responses where the linking rule was applied correctly, irrespective of verb accuracy. The estimated coefficients (Estimate) represent the log odds of the event ('correct') occurring for a one-unit change in the predictor, relative to the mean of the centred variable.

7.14 Experiment 2 – List of 30 intervening 'places'

in Aegypten (egypt)
am Bahnhof (station)
am Bett (bed)
im Buero (office)
am Damm (damm)
am Fluss (river)
im Gewaechshaus (green house)
am Golfplatz (golf court)
am Hafen (harbour)
am Haus (house)
am Kollosseum (colloseum)
im Labor (laboratory)
am Markt (market)
im Museum (museum)
am Palast (palace)
im Parkhaus (car park)
im Schwimmbad (swimming pool)
am Sofa (couch)
am Spielplatz (playground)
im Stadion (stadium)
am Strand (beach)
am Tisch (table)
im Tor (goal)
im Verkehr (traffic)
am Vulkan (vulcano)
im Wald (forest)
am Wasserfall (waterfall)
am Weihnachtsbaum (christmas tree)
am Zelt (tent)
am Zirkus (circus)

7.15 Experiment 2 – Example videos

Three NADs example videos can be found here. Alternatively, the files have been submitted to the University via the Research Theses Digital Submission application.

…, dass Anna am Zelt isst.

…, dass Klara im Museum huepft.

…, dass Tim in Aegypten singt.

7.16 Experiment 2 – Teaching input overview

LV	HV					
days 1-6	**day 1**	**day 2**	**day 3**	**day 4**	**day 5**	**day 6**
Klara_Bahnhof_huepft	Klara_Aegypten_huepft	Klara_Fluss_huepft	Klara_Kollosseum_huepft	Klara_Parkhaus_huepft	Klara_Strand_huepft	Klara_Wald_huepft
Klara_Parkhaus_huepft	Klara_Bahnhof_huepft	Klara_Gewaechshaus_huepft	Klara_Labor_huepft	Klara_Schwimmbad_huepft	Klara_Tisch_huepft	Klara_Wasserfall_huepft
Klara_Tisch_huepft	Klara_Bett_huepft	Klara_Golfplatz_huepft	Klara_Markt_huepft	Klara_Sofa_huepft	Klara_Tor_huepft	Klara_Weihnachtsbaum_huepft
Klara_Vulkan_huepft	Klara_Buero_huepft	Klara_Hafen_huepft	Klara_Museum_huepft	Klara_Spielplatz_huepft	Klara_Verkehr_huepft	Klara_Zelt_huepft
Klara_Zelt_huepft	Klara_Damm_huepft	Klara_Haus_huepft	Klara_Palast_huepft	Klara_Stadion_huepft	Klara_Vulkan_huepft	Klara_Zirkus_huepft
Anna_Bahnhof_isst	Anna_Fluss_isst	Anna_Kollosseum_isst	Anna_Parkhaus_isst	Anna_Strand_isst	Anna_Wald_isst	Anna_Aegypten_isst
Anna_Parkhaus_isst	Anna_Gewaechshaus_isst	Anna_Labor_isst	Anna_Schwimmbad_isst	Anna_Tisch_isst	Anna_Wasserfall_isst	Anna_Bahnhof_isst
Anna_Tisch_isst	Anna_Golfplatz_isst	Anna_Markt_isst	Anna_Sofa_isst	Anna_Tor_isst	Anna_Weihnachtsbaum_isst	Anna_Bett_isst
Anna_Vulkan_isst	Anna_Hafen_isst	Anna_Museum_isst	Anna_Spielplatz_isst	Anna_Verkehr_isst	Anna_Zelt_isst	Anna_Buero_isst
Anna_Zelt_isst	Anna_Haus_isst	Anna_Palast_isst	Anna_Stadion_isst	Anna_Vulkan_isst	Anna_Zirkus_isst	Anna_Damm_isst
Tim_Bahnhof_singt	Tim_Kollosseum_singt	Tim_Parkhaus_singt	Tim_Strand_singt	Tim_Wald_singt	Tim_Aegypten_singt	Tim_Fluss_singt
Tim_Parkhaus_singt	Tim_Labor_singt	Tim_Schwimmbad_singt	Tim_Tisch_singt	Tim_Wasserfall_singt	Tim_Bahnhof_singt	Tim_Gewaechshaus_singt
Tim_Tisch_singt	Tim_Markt_singt	Tim_Sofa_singt	Tim_Tor_singt	Tim_Weihnachtsbaum_singt	Tim_Bett_singt	Tim_Golfplatz_singt
Tim_Vulkan_singt	Tim_Museum_singt	Tim_Spielplatz_singt	Tim_Verkehr_singt	Tim_Zelt_singt	Tim_Buero_singt	Tim_Hafen_singt
Tim_Zelt_singt	Tim_Palast_singt	Tim_Stadion_singt	Tim_Vulkan_singt	Tim_Zirkus_singt	Tim_Damm_singt	Tim_Haus_singt

Table 29 Overview of entire input during 6-day teaching intervention in experiment 2 (NADs). Fields highlighted in colour exemplify the distribution of interveners across subjects (Klara, Anna, Tim) and days. Each intervener appeared only once per day in the HV condition. Across all six days, each intervener appeared three times, once with each subject. In both the LV and the HV condition, order of sentences was randomized each day.

7.17 Experiment 2 – Example sets of test sentences (Outcome measure)

type	example child from HV	example child from LV
familiar	dass Tim am Zelt singt	dass Anna am Tisch isst
familiar	dass Tim am Tisch singt	dass Anna im Parkhaus isst
familiar	dass Anna im Bahnhof isst	dass Klara am Vulkan huepft
familiar	dass Anna am Vulkan isst	dass Klara im Bahnhof huepft
familiar	dass Klara am Tisch huepft	dass Tim am Vulkan singt
familiar	dass Klara im Parkhaus huepft	dass Tim im Bahnhof singt
novel intervener	dass Tim im Badezimmer singt	dass Anna am Fenster isst
novel intervener	dass Anna im Badezimmer isst	dass Anna im Badezimmer isst
novel intervener	dass Tim am Eingang singt	dass Klara im Badezimmer huepft
novel intervener	dass Klara am Fenster huepft	dass Klara im Getümmel huepft
novel intervener	dass Anna am Fenster isst	dass Tim im Badezimmer singt
novel intervener	dass Klara am Eingang huepft	dass Tim im Kino singt
wrong dependency	dass Klara am Tisch isst	dass Anna am Vulkan singt
wrong dependency	dass Tim am Tisch huepft	dass Anna im Bahnhof singt
wrong dependency	dass Tim im Parkhaus huepft	dass Klara am Tisch isst
wrong dependency	dass Anna am Zelt singt	dass Klara am Vulkan isst
wrong dependency	dass Anna am Vulkan singt	dass Tim am Vulkan huepft
wrong dependency	dass Klara im Bahnhof isst	dass Tim im Bahnhof huepft

Table 30 Example test sentence sets for each one child from the LV and the HV condition. Sentence order was randomized during testing.

7.18 Experiment 2 – Pilot and pilot results

7.18.1 Outcome measurement

The piloting of experiment 2 took place immediately after the piloting of experiment 1 with the same four Year 3 children. The testing procedure closely followed the procedure in the main experiment. Again, each child was tested individually in the corridor during lesson time. In the beginning, each child was introduced to *Klara*, *Anna*, and *Tim* and received instructions regarding the exposure phase and the testing. Then, each child was exposed to the same randomly generated 90 sentences, the only difference between children being that two of them were exposed to 5 different intervening prepositional phrases while the other two children were introduced to 30 different interveners. Those 90 sentences were identical to the input used during the main experiment (6 days x 15 sentences). The order of sentences was randomised for each child. I clicked through the individual short videos on a laptop and said the corresponding sentences out loud on each trial. Then, in each trial, the child and I repeated the sentence while re-watching the scene. After this exposure phase which lasted approximately ten minutes, the children had a short refreshment break before we continued with testing.

Anticipating that the two consecutive and demanding pilots for experiments 1 and 2 would be mentally taxing for the students, we took steps to mitigate potential fatigue by limiting the number of testing sentences to 18 per child. These 18 sentences were distributed as three sets, each containing six sentences featuring either *Klara*, *Anna*, or *Tim*. In each of those sets featuring one cartoon person, there were two familiar sentences, two sentences with wrong dependencies, and two sentences with novel interveners. For each child, the two familiar sentences were selected randomly from the larger pool of five sentences per cartoon character, all of which they had become familiar with during the exposure phase. Similarly, for the two sentences with incorrect dependencies, interveners were randomly selected from the set of five familiar interveners that had been presented during the exposure phase.

The pilot for the second experiment showed that children found the cartoons engaging and stayed focused throughout exposure. Of course, 90 exposure trials in a row turned out to be challenging for the children. However, this was for piloting purposes only and children would only be exposed to 15 sentences per day during the main experiment. Overall, there

were some distractions caused by students or staff passing by, yet the piloting went as smoothly as it did for the first experiment.

Regarding the outcome measure procedure, it became apparent that the number of test sentences had to be limited to 18 sentences for the main experiment as well. This limitation stemmed from feedback provided by students, who reported that they would have experienced confusion if they had encountered additional 'incorrect' (i.e., novel) sentences repeatedly during testing, particularly those with wrong dependencies. This comprehension challenge is quite understandable, as the sentences, whether familiar, featuring incorrect dependencies, or introducing novel interveners, share an extremely similar structural pattern. Listening to the test sentences consecutively within a short timeframe (and without visual input during testing) can naturally pose difficulties and potential confusion. However, it was decided not to reduce the number of test sentences, even though even younger children would be assessed in the main experiment. Two primary reasons underpinned this choice. Firstly, there is plenty of evidence in the literature that children, including younger ones, have the capacity to discern patterns within linguistic structures that may initially appear highly perplexing. Secondly, for the sake of robust statistical analyses, it was considered advantageous to include two instances of each test sentence type per cartoon character, rather than just one. Descriptive results of the experiment 2 pilot are presented in Table 31 below.

condition	familiarity	mean	median
HV	familiar	0.85	1.0
LV	familiar	0.82	1.0
HV	novel intervener	0.50	0.5
LV	novel intervener	0.67	1.0
HV	wrong dependency	1.00	1.0
LV	wrong dependency	0.70	1.0

Table 31 Experiment 2 (NADs) piloting results. N = 3.

7.19 Experiment 2 - Correlation analyses between pre-tests and outcome measurement

Before modelling the experiment 2 data, it had to be determined whether there was any indication in the data to justify including pre-test scores as fixed effects in the statistical models. Correlation analyses between results of the individual pre-tests and the outcome measurement were conducted. Plots of the correlations between pre-test raw scores and outcome measure results including linear regression lines and SE are presented in Figures 9-11 below.

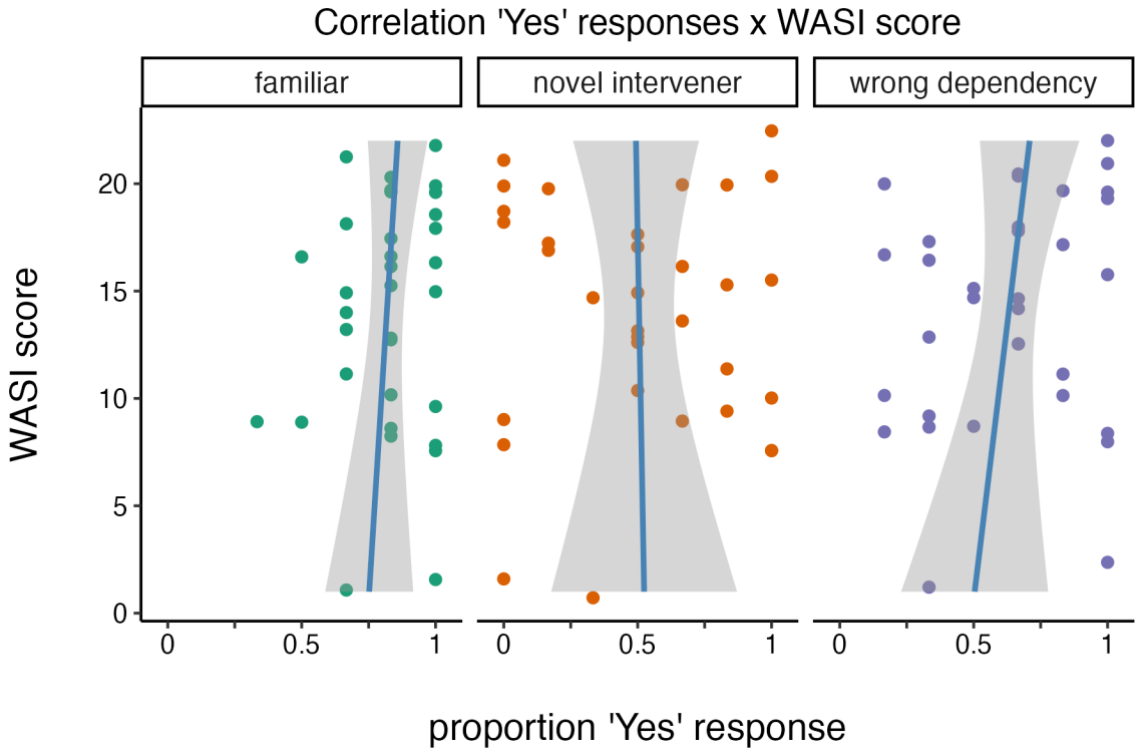


Figure 9 Correlations between participant mean scores on the WASI (pre-test) and participant mean proportion of 'Yes' responses in the outcome measurement (subdivided by familiarity).

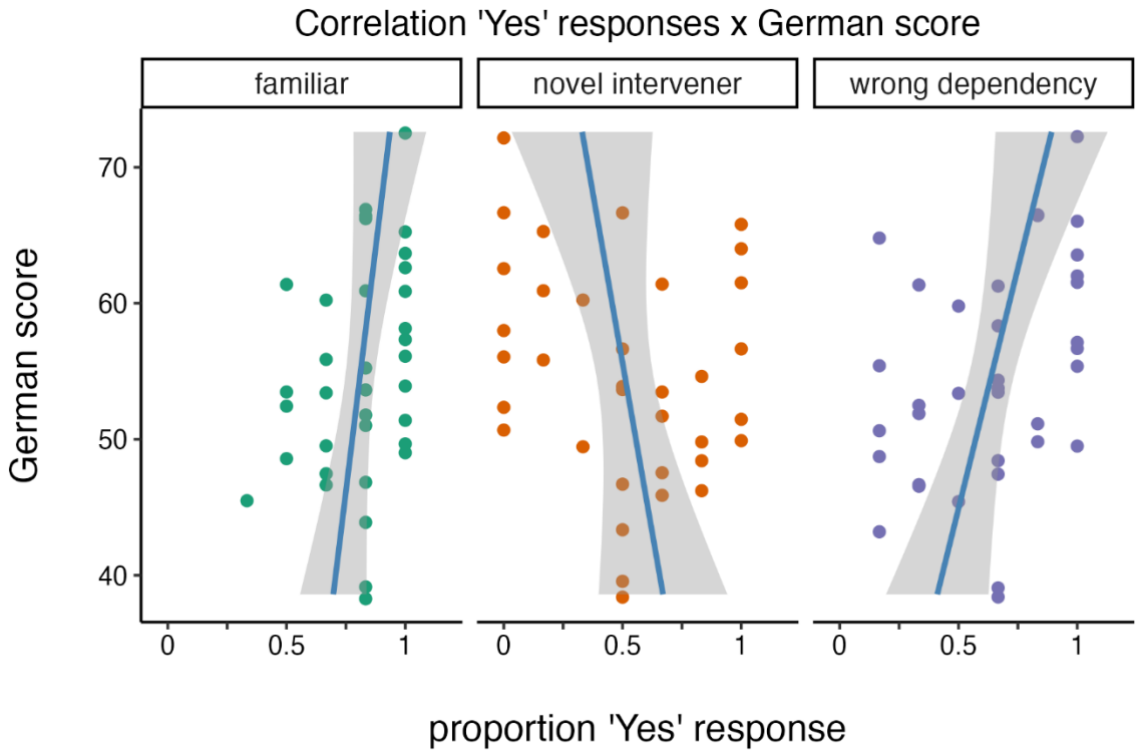


Figure 10 Correlations between participant mean scores on the Language Magician (pre-test) and participant mean proportion of 'Yes' responses in the outcome measurement (subdivided by familiarity).

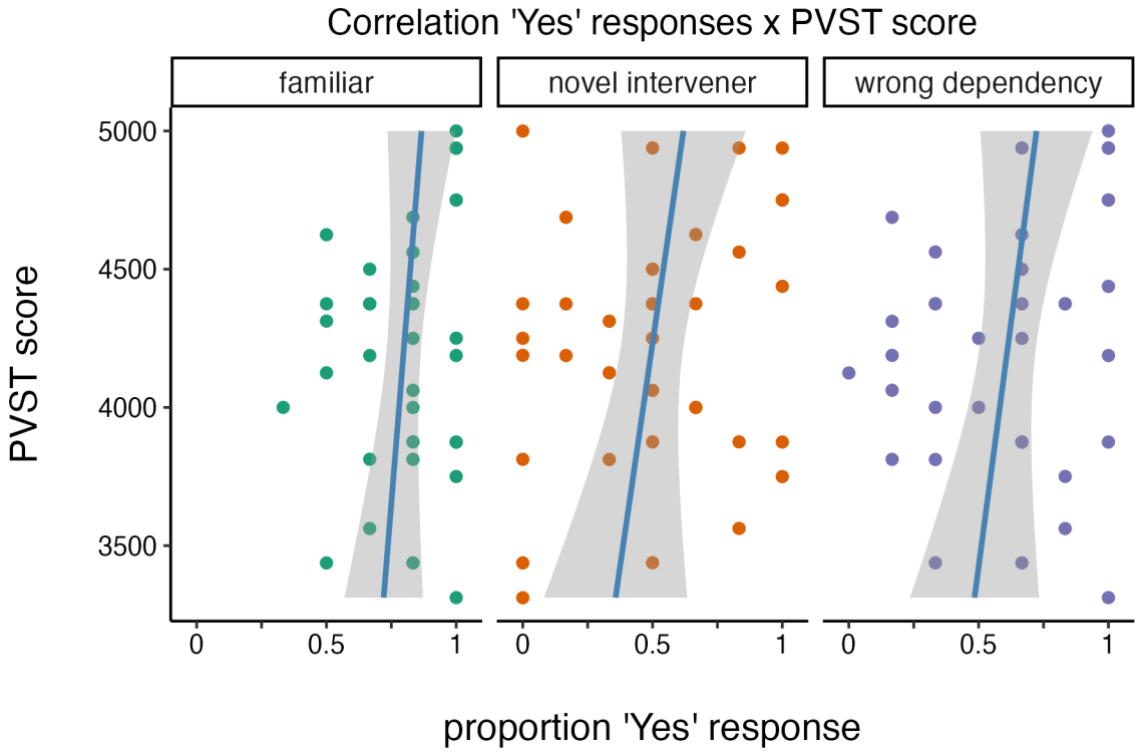


Figure 11 Correlations between participant mean scores on the PVST (pre-test) and participant mean proportion of 'Yes' responses in the outcome measurement (subdivided by familiarity).

Visual inspection of the plots suggested no obvious associations between any of the pre-test and outcome measurement variables. Additional statistical correlation analyses were conducted (computed over the z-scores of participants' responses in the outcome measurements). The correlations with performance on familiar, novel intervener, and wrong dependency trials verbs were analysed separately. Thus, presented below are correlations between the raw pre-test scores and the z-scores of outcome measure scores (i.e., proportion 'Yes' responses) in each of the three individual trial types.

	outcome measurement responses (z-scores)		
	familiar	novel intervener	wrong dependency
WASI score	0.16	-0.02	0.18
German score	0.30	-0.23	0.38*
PVST score	0.20	0.20	0.20

Table 32 Spearman rank correlation coefficients (ρ) of correlations between z-sores of outcome measure scores (i.e., proportion 'Yes' responses) and raw pre-test scores. Spearman rank correlation ranges from -1 to 1. Coefficients marked with * are statistically significant at $p < .05$.

Analysis results (Table 32) revealed weak negative and positive correlations. However, only the correlation between German scores and z-scores of response scores in wrong dependency trials ($r = 0.38$) reached the threshold for statistical significance ($p < .05$) (and this would not be the case if the analyses were adjusted for multiple comparisons). Although the absence of evidence of associations does not equal evidence of no associations between the variables, the generally weak correlation coefficients was not considered a strong enough evidence base to justify the inclusion of pre-test scores in the statistical modelling.

7.20 Experiment 2 – Detailed model outcomes

7.20.1 familiar

	Estimate	SE	Z	p
Fixed effects				
(intercept)	1.54	0.24	6.32	< 0.001
input group = HV	-1.10	0.38	-2.93	0.003
Random effects				
	SD			
participant (intercept)	0.29			
sentence	0.42			

Table 33 Familiar trials mixed model for predicting 'Yes' responses. The estimated coefficients (Estimate) represent the log odds of the event ('Yes') occurring for a one-unit change in the predictor, relative to the mean of the centred variable.

7.20.2 novel intervener

	Estimate	SE	Z	p
Fixed effects				
(intercept)	-0.02	0.38	-0.06	0.95
input group = HV	0.65	0.71	0.91	0.36
Random effects				
	SD			
participant (intercept)	1.88			
sentence	0.54			

Table 34 Novel intervener trials mixed model for predicting 'Yes' responses. The estimated coefficients (Estimate) represent the log odds of the event ('Yes') occurring for a one-unit change in the predictor, relative to the mean of the centred variable.

7.20.3 wrong dependency

	Estimate	SE	Z	p
Fixed effects				
(intercept)	-0.87	0.38	-2.31	0.021
input group = HV	1.70	0.62	2.74	0.006
Random effects				
	SD			
participant (intercept)	1.42			
sentence	0.81			

Table 35 Wrong dependency trials mixed model for predicting 'Yes' responses. The estimated coefficients (Estimate) represent the log odds of the event ('Yes') occurring for a one-unit change in the predictor, relative to the mean of the centred variable.

7.21 Experiment 2 – Estimate of the prior calculation (wrong dependency trials)

To derive an estimate of the predicted mean difference under H1 for the hypothesis that the HV group provides *more* 'Yes' responses than the LV group (corresponding to Figure 4), the motivated maximum approach was used (cf. Silvey et al. under review, and 'related room to move' hypothesis in Dienes, 2019). Recall that this technique was used for deriving priors in experiment 1 as well (see section 3.2.2). The same statistical approach was also used in an experimental study by Schulz and Wonnacott (under review). The motivated maximum approach entails deriving the predicted effect size from a maximum value, followed by setting the predicted effect size at double this value. Note that for calculating priors for experiment 1, the grand means (i.e., intercepts) were used to derive a 'motivated maximum effect'. In the case at hand, the relevant intercept is negative -corresponding to Figure 4- and cannot be used. Given that H1 is directional and that therefore the model of H0 follows a half-normal distribution with a mean of 0, the maximum value is roughly equivalent to twice the standard deviation. The calculation involves using rank scores, thus, we can compute the logically possible maximum difference between the HV group (N = 18) and the LV group (N = 20) if no participant in the LV group scored a higher mean proportion 'Yes' responses than any participant in the HV group. This possible maximum difference is equal to (18+20)/2, resulting in 19. Thus, the rough predicted effect size is set to be equal to half of this value, resulting in 9.5.[38]

[38] Note that although this approach limits the potential range of effects, H1 is conceptualized as a distribution. The adoption of a half-normal distribution with a mean of 0 reflects the anticipation that smaller values are more likely.