Bilingualism: Language and Cognition

cambridge.org/bil

Learning second language morphosyntax in dialogue under explicit and implicit conditions: An experimental study with advanced adult learners of German

Eva M. Koch[1], Johanna F. de Vos[2], Alex Housen[1], Aline Godfroid[3] and Kristin Lemhöfer[2]

[1]Vrije Universiteit Brussel, Brussels, Belgium; [2]Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands and [3]Michigan State University, East Lansing, MI, USA

Research Article

Cite this article: Koch EM, de Vos JF, Housen A, Godfroid A, Lemhöfer K (2023). Learning second language morphosyntax in dialogue under explicit and implicit conditions: An experimental study with advanced adult learners of German. *Bilingualism: Language and Cognition* **26**, 984–997. https://doi.org/10.1017/S1366728922000797

Received: 13 December 2020
Revised: 12 October 2022
Accepted: 28 October 2022
First published online: 21 March 2023

Keywords:
morphosyntax; awareness; incidental second language learning; explicit and implicit instruction; allomorphy

Address for correspondence:
Eva M. Koch,
Vrije Universiteit Brussel, Department of Linguistics & Literary Studies, Office 5B.28
Pleinlaan 2
1050 Brussels
Belgium
Email: koch.eva.marie@gmail.com

© The Author(s), 2023. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial licence (http://creativecommons.org/licenses/by-nc/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original article is properly cited. The written permission of Cambridge University Press must be obtained prior to any commercial use.

Abstract

We investigate the role of awareness in learning non-salient grammar features in a second language during oral interaction. We conducted a learning experiment during which forty-eight adult Dutch-speaking advanced learners of German and a native German-speaking experimenter engaged in a scripted oral dialogue game. The experimenter and learner in turn produced sentences based on pictures eliciting German strong verbs with stem-vowel alternations, a morphosyntactic feature that represents a persistent learning difficulty. While learners in the implicit condition were merely instructed to focus on sentence meaning, learners in the explicit condition were encouraged to also pay attention to and learn from the target structure in the experimenter's input. Although the explicit group achieved higher accuracy scores overall, both groups had similar (absolute) learning gains, showing that oral input provided during interactive exchanges can lead to substantial learning not only under explicit, learning-targeted conditions, but also without an explicit directive to learn.

Introduction

Even when a second language (L2) is learned in an instructional context, the learning process is likely to be also shaped by incidental exposure while interacting with first-language (L1) speakers or more proficient learners. However, it is still an open issue if such learning can occur without the intention to learn or without the awareness that one is learning, and to what extent this depends on the nature of the linguistic target structure. The role of awareness for learning, and particularly for grammar learning, has been much debated in Second Language Acquisition (SLA) (e.g., Leow & Hama, 2013; Schmidt, 1995; Williams, 2009). Most grammar learning experiments on this issue have been conducted under controlled laboratory conditions, including the use of (semi-)artificial languages (e.g., Rebuschat & Williams, 2012), and relied on exposing beginning learners to various types of one-directional, monologic, and often written L2 input (e.g., Robinson, 1996). However, our understanding of grammar learning cannot be complete without the investigation of acquisition processes in more advanced learners, and without considering the role of output, and how input and output come together during interaction (Gass, 2003).

The present study investigates the L2 learning of morphosyntax through an experimentally controlled, yet interactive design integrating exchanges of native-speaker input and learner output. The target structure is verb-stem allomorphy in German strong verb inflection (e.g., *tragen*, "to carry"; *er trägt*, "he carries"), a subregular grammatical feature that represents a persistent learning difficulty (Godfroid, 2016). The participants in our study were adult Dutch-speaking intermediate-to-advanced learners of German who possessed prior knowledge of this structure. They performed a meaning-focused dialogue game with the experimenter, an L1 speaker of German, based on picture combination. Participants in the explicit condition were additionally encouraged to pay attention to the target structure in the input, whereas participants in the implicit condition did not know that grammar learning was targeted.

Incidental second language learning

INCIDENTAL LEARNING refers to learning without the intention to learn and therefore represents the opposite of INTENTIONAL LEARNING (Ortega, 2009). As intentionality can be hard to measure, some researchers (e.g., de Vos, Schriefers, ten Bosch & Lemhöfer, 2019; Hulstijn, 2003) prefer to operationalize incidental learning as learning without a learning instruction. Importantly, incidental learning is not necessarily IMPLICIT LEARNING (i.e., unconscious learning; see

Williams, 2009), as it allows for a certain degree of fleeting awareness (Ortega, 2009) of the target structure. Although it remains a matter of debate whether implicit L2 grammar learning is possible (e.g., Andringa, 2020; Godfroid, 2016; Leow & Hama, 2013; Leung & Williams, 2012), to date, most SLA researchers agree that adult L2 learning can occur incidentally (for a review, see Hulstijn, 2003). Note that LEARNING in this context refers to the type of learning process, and is not to be confounded with KNOWLEDGE, which is the outcome resulting from this process and which can be EXPLICIT (conscious and verbalizable) or IMPLICIT (unconscious and non-verbalizable; see Williams, 2009).

SLA research also distinguishes between different EXPOSURE CONDITIONS for learning: EXPLICIT CONDITIONS are those where learners are provided with specific metalinguistic information, or with the instruction to extract rules from linguistic input; under IMPLICIT CONDITIONS, any metalinguistic information or rule-search instructions remain absent (Norris & Ortega, 2001). When implemented in classroom contexts, researchers also refer to such conditions as explicit versus implicit types of INSTRUCTION. Typically, under implicit conditions, learners carry out a meaning-focused language task during which they are exposed to a target structure. The question is whether learners, while focusing on meaning, will (incidentally, or even fully implicitly) learn the target structure from the input (Godfroid, 2016). While explicit learning conditions usually entail intentional, explicit (i.e., conscious) learning, implicit conditions typically entail implicit or incidental learning. Nonetheless, implicit conditions may also yield intentional learning, depending on how aware a learner has become of the task's actual language learning goal and the relevant grammatical rules or features.

An extensive body of research has compared the effectiveness of explicit versus implicit conditions for L2 learning, mostly consisting of classroom studies. Overall, cumulative evidence suggests explicit conditions to be more effective than implicit conditions (for meta-analytic reviews, see Goo, Grañena, Yilmaz & Novella, 2015; Norris & Ortega, 2000; Spada & Tomita, 2010). The more recent meta-analysis by Kang, Sok and Han (2019), however, nuances this picture as it found no significant difference between explicit and implicit types of instruction at the level of immediate learning outcomes; in addition, the authors even reported an advantage of implicit instruction over explicit instruction on delayed posttests. Importantly, the results of such comparative studies need to be interpreted with care as there have been inconsistencies in the operationalization of explicit and implicit conditions across studies (see R. Ellis, 2009). Moreover, measurement practices of early comparative studies were often biased towards explicit conditions (for discussions, see Andringa, de Glopper & Hacquebord, 2011; R. Ellis, 2009; Norris & Ortega, 2001). In their classroom study, Andringa et al. (2011) attempted to prevent this bias by using a more balanced design, i.e., by providing the same amount of exposure in the implicit and explicit conditions, and by using both free response and grammaticality judgement tasks as outcome measures. The explicit group outperformed the implicit group on the judgment task, but not on the free response task, suggesting that explicit conditions are not always superior in terms of learning.

Interesting contributions also come from a line of research that explicitly aims to study natural language learning under implicit, naturalistic conditions – enhancing ecological validity – whilst maintaining a high degree of experimental control. De Vos et al. (2019), a study that is relevant to us from a methodological perspective, introduced a novel research paradigm to approximate naturalistic vocabulary learning in the laboratory. Their study was presented as a psychological experiment and took the design of a dialogue between participant and experimenter, the former being a learner of Dutch, the latter a native speaker. In turn, they orally compared objects by their price. Unknown to the participants, they were exposed to and tested on a set of unknown Dutch words. The results revealed significant incidental word learning (modulated by number of exposures and cognate status but not retention interval).

Dialogue-like techniques have also been applied to L2 grammar learning (Brandt, Schriefers & Lemhöfer, 2021; Conroy & Antón-Méndez, 2015; McDonough & Mackey, 2008). Brandt et al. (2021) investigated the incidental acquisition of Dutch grammatical gender by advanced German-speaking learners, a population known to be prone to persistent errors on gender-marked determiners for Dutch nouns with incompatible gender in German. Two experiments – an alternating picture description task and a memory game – combined audio-recorded native-speaker input and oral learner productions of gender-marked determiners. Importantly, participants were not made aware of the learning aspect of the study; yet, in the picture description experiment, participants became aware of the learning aim to a greater degree than in the memory game. Both experiments yielded comparable amounts of incidental learning, regardless of the differences concerning the transparency of the tasks' learning aim. In brief, the studies just discussed provide a promising approach to explore the incidental learning of new target structures in a more naturalistic manner (i.e., involving input alternating with output in form of a dialogue) than what is often the case in laboratory studies.

The effectiveness of particular learning conditions might further be modulated by the nature of the target structure (DeKeyser, 1995; N. C. Ellis, 1993; Housen, 2020; Robinson, 1996; Spada & Tomita, 2010). Few studies have directly investigated the extent to which the learnability of grammatical features under explicit and implicit conditions is mediated by properties that are often subsumed under the rubric of their 'difficulty'. Such properties include, among others, allomorphy, regularity, productivity, communicative redundancy, and frequency, all of which conspire to determine a feature's transparency and salience (Housen & Simoens, 2016). Most studies on L2 grammar learning under implicit conditions have used grammatical structures with a relatively high saliency and/or clear communicative value, i.e., rather 'easy' structures (e.g., English regular past tense: R. Ellis, Loewen & Erlam, 2006). However, incidental L2 learning research should include a wider array of difficult morphosyntactic structures that are non-salient, opaque, redundant and/or infrequent; these structures, too, play a role in fine-tuning L2 competence to approach native-like levels. The question is whether such structures can also be learned – or improved upon, when initial acquisition has failed – under implicit exposure conditions, and what the role of awareness is for learning.

As a showcase for such an opaque and 'difficult' grammatical feature, we will investigate the learning of allomorphy-based verb inflection. Inflectional verb morphology in general has been identified as a source of enhanced difficulty for adult L2 learners, both at the level of comprehension and production (for reviews, see DeKeyser, 2005; N. C. Ellis, 2006b; Larsen-Freeman, 2010). The majority of studies in this domain have addressed verb affixation and suppletion; less is known about the L2 processing and learning of verb allomorphy (Krause, Bosch & Clahsen, 2015). We therefore chose the stem-vowel alternations that occur in German strong verbs as our target structure, further building upon research (Godfroid, 2016; Godfroid & Uggen, 2013; Krause

https://doi.org/10.1017/S1366728922000797 Published online by Cambridge University Press

et al., 2015) that found this allomorphic structure to be well-suited to gain insights on the acquisition and processing of subtle, difficult linguistic features.

Stem-vowel alternations in German strong verbs

The German conjugation system distinguishes between weak, strong, and irregular verbs. In weak conjugation – an unmarked, regular and productive paradigm –, information of grammatical person, number, tense, and mood is encoded exclusively through affixation. Strong verbs, however, are considered marked because morphosyntactic information is provided through allomorphy, in addition to affixation: their stem vowels undergo morphophonemic alternations, leading to the co-existence of different stem variants ('allomorphs'). Although they represent a closed group of verbs with low type frequency, most strong verbs occur frequently in everyday language use (Köpcke, 1998). In the present study, we focus on German strong verbs in the third person singular in the present indicative tense (3SG PRES). These forms have the same *–t* suffix as weak verbs, but in addition, they undergo stem-vowel alternations. The two main types are *a→ä* and *e→i(e)* changes (see Table 1 for examples).

The changing vowel represents an L2 learning difficulty (Godfroid, 2016; Godfroid & Uggen, 2013). The alternations in the stem are restricted to only one phoneme and might therefore be hard to perceive. Another complicating factor is information redundancy (DeKeyser, 2005; N. C. Ellis, 2006a): the allomorph encodes number and person, but this information is also provided through the *–t* suffix and the subject noun phrase. Moreover, the alternations are highly unpredictable in contemporary German (Bybee & Newman, 1995): the infinitive alone does not provide any cues about whether a verb is strong or weak. Therefore, to which type a verb belongs has in principle to be learned together with the lexical item itself, similar to German word gender (e.g., Hopp, 2013). The strong paradigm is, however, not entirely unpredictable and irregular. Different theories refer to it as a 'subregular' system in that the different items form clusters corresponding to certain phonemic and semantic regularities (DeKeyser, 1995; Godfroid, 2016; Köpcke, 1998; Krause et al., 2015).

Learning difficulties related to stem-vowel alternations can be expected to generalize to other non-transparent morphosyntactic features. Yet, only few experimental studies have investigated this structure so far. Godfroid and Uggen (2013) used eye-tracking to measure attention directed to the changed stems of strong verbs, presented in their written forms, by English-speaking beginning learners of German without prior knowledge of this structure. Longer fixation times on the changed stems and visual comparisons between changed and unchanged stems (by looking back and forth) predicted the learners' performance on a written posttest, emphasizing the facilitative role of attention for learning. Godfroid (2016) examined whether advanced English-speaking learners of German would become more sensitive to the alternating vowels after a session of auditory input flooding, an implicit learning condition that exposed learners intensively to strong verbs in the context of a sentence-picture matching task. The author found significant implicit learning, operationalized as an increase in sensitivity during exposure as reflected by longer reaction times on ungrammatical trials that were presented toward the end of the exposure phase. Two pre- and posttests revealed that implicit learning had led to the development of implicit but not explicit knowledge. Both studies illustrate that stem-vowel changes represent a difficulty for learners at the levels of production and learning, arising from weak conjugation being the default system in these learners' interlanguages, but also that learning of this structure under incidental or implicit conditions is possible.

We extend this line of research by investigating learning under explicit and implicit conditions in L1 speakers of Dutch, a Germanic language closely related to German. A brief online pilot study (see Materials) confirmed that the vowel change represented a learning difficulty in our population (average of 33% of failures to apply the stem-vowel change). Many Dutch verbs are cognates with German verbs, yet alternating stem vowels in PRES do not exist in Dutch (see Table 1; such alternations only occur in the irregular past tense and in participles in Dutch), possibly causing negative L1-L2 transfer effects that may add up to the learning difficulties that are inherent to non-salient, communicatively redundant forms. Moreover, learners may suffer from BLOCKING effects (N. C. Ellis, 2006a, 2006b, 2006c); that is, learners first learn to associate the weak *-t* suffix with 3SG PRES, making it harder for them to learn later that a strong verb's stem vowel also provides the same morphosyntactic information. On this account, we would predict explicit conditions to be more effective than implicit conditions, because they may help to bypass such blocking effects (see Cintrón-Valentín & N. C. Ellis, 2015).

The present study

We investigated the learning of German strong verb allomorphy in intermediate to advanced L2 learners under naturalistic, interactive conditions. We achieved this by using a meaning-focused

Table 1. Examples of German weak and strong verb conjugation in the present tense

	German					
	Weak verbs (*a*-stem)	Weak verbs (*e*-stem)	Strong verbs (*a-ä* type)	Strong verbs (*e-i/ie* type)	For comparison: Examples of Dutch cognates	
Infinitive	jag-*en* ("to hunt")	leg-*en* ("to put")	fang-*en* ("to catch")	sprech-*en* ("to speak")	vang-*en* ("to catch")	sprek-*en* ("to speak")
Present 1SG	jag-*e*	leg-*e*	fang-*e*	sprech-*e*	vang	spreek
Present 2SG	jag-*st*	leg-*st*	fäng-*st*	sprich-*st*	vang-*t*	spreek-*t*
Present 3SG	jag-*t*	leg-*t*	fäng-*t*	sprich-*t*	vang-*t*	spreek-*t*

Note. *a* is realized as /a/, *ä* as /ɛ/, *i* as /ɪ/, *ie* as /i/. Depending on the phonological context, *e* is pronounced /e/ or /ɛ/. A minority of strong verbs also requires changes from diphthong *au* /aʊ̯/ to *äu* /ɔʏ̯/, or *o* /o/ to *ö* /ø/.

https://doi.org/10.1017/S1366728922000797 Published online by Cambridge University Press

learning task consisting of a dialogue game between participant and experimenter (e.g., de Vos et al., 2019), including elements of natural conversations, such as turn-taking involving an output component. Moreover, we explored the role of awareness for learning by manipulating the task instructions given to participants before this dialogue.

The intention of introducing a dialogic, yet strictly controlled paradigm was to obtain a better balance between experimental control and ecological validity than what is often the case in traditional laboratory studies. This paradigm also allowed us to conceal the learning aim of the study from participants in the implicit condition. Finally, in this paradigm, the classical components of a learning study (pretest, treatment, posttest – typically representing separate phases of traditional designs, which augments the risk of revealing the actual goal of the study) were now all embedded within the same dialogue game, to the benefit of the naturalness and incidental character of the study.

Our manipulation of learning condition (explicit vs. implicit) consisted of splitting participants into two groups who received diverging task instructions before the dialogue part: all learners were instructed to focus on sentence semantics during the task, but only the learners in the explicit condition were made aware of the target structure and the learning purpose, on top of the focus on meaning. The experimenter, a German native speaker and the participants' dialogue partner, provided correct input as a 'natural' part of her utterances. Learning was measured as improvements in accuracy between participant productions before and after this input. As an additional control condition, input was provided only for half of all items, to verify that any improvements were indeed input-induced. To assess our participants' awareness status and the type of learning they had engaged in, we conducted retrospective interviews immediately after the learning task.

The participants were Dutch native speakers who were intermediate-to-advanced learners of German with prior knowledge of the target structure. Thus, we did not investigate the entirely novel acquisition of a conjugation paradigm, but the further development of a known, difficult morphosyntactic feature. Put differently, our participants did not have to learn a new 'rule', but the study's purpose was to learn, item-wise, to which verbs the vowel alternation applies. To date, few other studies have looked at already consolidated erroneous production patterns, and how they 'clash' with correct input (but see Brandt et al., 2021; Godfroid, 2016; Lemhöfer, Schriefers & Indefrey, 2020; McDonough & Mackey, 2008).

Research questions

The following research questions guided the study:

- RQ1: Do the L2 learners in our study show learning of the stem-vowel alternations in German strong verbs from oral native-speaker input during a scripted dialogue?
- RQ2: Does exposure condition (explicit vs. implicit), and the participants' awareness status resulting thereof, influence the learning rate?

Methods

Participants

Fifty-five non-dyslexic L2 learners of German, mostly students or academics living in Brussels, Belgium, participated in our study. They received monetary compensation or course credits in reward. The experiment was initially presented to all participants as a study on the relationship between the language one speaks and one's way of thinking – a 'cover story' meant to conceal the study's actual focus on grammar and learning. Participants were randomly assigned to either the 'explicit' condition in which the participants received information about the target structure and the experiment's learning purpose (see Appendix S1 in the supplementary materials: https://osf.io/938ye/), or the 'implicit' condition in which they did not receive this information.

We excluded two participants for not speaking Dutch as their L1, one for having taken psychoactive medication shortly before the experiment, and two for failing to apply weak verb conjugation in 3SG PRES (i.e., they only produced infinitives during the learning task, suggesting they were unfamiliar even with the basics of German inflection). We also excluded one participant from the implicit condition who had become fully aware of the learning purpose, and one from the explicit condition for the apparent use of a default strategy (for details, see Materials).

The final sample consisted of 48 L1 Dutch speakers (33 females) aged between 17 and 37 years (mean age 24) with an intermediate-to-advanced level of German (see Table 2 for learner background variables). All participants (except for three who were later in the removed 'unaware' subgroup; see next two paragraphs) were familiar with the target structure, mainly from prior German formal language instruction. Thirty participants had been enrolled in language-related university programs, and fifteen out of them had studied German at university level. Nine participants reported another L1 in addition to Dutch and all participants also spoke other foreign languages, in particular English, French and Spanish, none of which require stem-vowel alternations in the same conditions as German strong verbs (note that in English, vowel changes occur only in past tense forms; in certain irregular French and Spanish verbs, it occurs in first-, second- and third-person singular and third-person plural of the simple present tense).

Immediately after the learning task, we interviewed the participants to assess their awareness of the presence of the target structure (TS; i.e., noticing the presence of vowel-changing strong verbs in the learning task) and, in case they had, additional awareness of the task's learning purpose (LP; i.e., knowing that the task's aim was to learn the vowel change). The interviews confirmed that all participants of the explicit condition were aware of both target structure and learning purpose; we denoted these participants as [+TS, +LP] and will henceforth refer to them as the INTENTIONAL GROUP ($n = 21$). All participants in the implicit condition (except for the excluded one) turned out to be unaware of the learning purpose; yet, the majority showed awareness of the target structure, except for some participants who remained completely unaware. Therefore, we reassigned the implicit participants to a [+TS, -LP] group, henceforth INCIDENTAL GROUP ($n = 21$), and a [-TS, -LP] UNAWARE GROUP ($n = 6$).

We did not perform an a priori power analysis to determine the number of participants because at the time of designing our experiment, there were no comparable experimental studies which could indicate an expected effect size. Instead, we recruited as many participants as possible, an enterprise that was limited by our target population being a rather restricted and not easily accessible group. We tested participants until both the intentional and incidental groups included at least 20 participants whose data could be used for analysis. The six unaware participants were excluded from all analyses due to the extremely small sample

https://doi.org/10.1017/S1366728922000797 Published online by Cambridge University Press

Table 2. Descriptive statistics and difference tests on variables related to the participants' language background in L2 German

	Intentional group $n = 21$	Incidental group $n = 21$	Group comparisons	
	Mdn (IQR)	*Mdn* (IQR)	Wilcoxon rank-sum	*r*
Age	23.00 (6.00)	23.00 (7.00)	$W = 185.5$, $p = .38$	−.14
Years of instruction (school)	3.00 (0.75)	3.00 (2.00)	$W = 180.5$, $p = .30$	−.16
Years of instruction (university)	0.08 (1.00)	0.08 (2.00)	$W = 225.5$, $p = .91$	−.02
Number of other L2s	2.00 (1.00)	2.00 (1.00)	$W = 207.5$, $p = .73$	−.05
Frequency of usage*	2.00 (2.00)	2.00 (1.00)	$W = 187.0$, $p = .38$	−.14
General proficiency*	3.00 (1.00)	3.00 (1.00)	$W = 206.5$, $p = .72$	−.06
Speaking*	3.00 (1.00)	3.00 (2.00)	$W = 176.5$, $p = .25$	−.18
Listening*	4.00 (1.00)	4.00 (1.00)	$W = 199.5$, $p = .57$	−.09
Writing*	2.00 (1.00)	3.00 (1.00)	$W = 181.0$, $p = .30$	−.16
Reading*	4.00 (1.00)	4.00 (0.00)	$W = 179.5$, $p = .25$	−.18
Premeasure critical items	25.00 (50.00)	0.00 (25.00)	$W = 256.0$, $p = .35$	−.15
Premeasure control items	100.00 (25.00)	100.00 (0.00)	$W = 205.0$, $p = .62$	−.08
	M (*SD*)	*M* (*SD*)	Welch *t*-test	*d*
Vocabulary size (LexTALE)	65.77 (8.08)	65.42 (6.35)	$t(37.90) = 0.16$, $p = .87$	0.05

Note. *Mdn* = median; IQR = interquartile range. Welch *t*-tests were used for normally distributed variables, Wilcoxon rank-sum tests for ordinal or not normally distributed interval variables. Asterisks mark variables based on self-ratings on a 1-5 Likert scale (1 = *very low*, 5 = *very high*). Premeasures and LexTALE scores are percentages. Years of instruction at school reflects German instruction at secondary school or through evening classes; years of instruction at university covers both German as a main field of study or as an elective course.

size and because the only three participants reporting no prior knowledge of the vowel change were all part of this group, representing a possible confound (for a description of their results, see Appendix S2 in the supplementary materials).

Table 2 compares the intentional and incidental groups on a set of individual-difference variables that might influence the acquisition of German strong-verb conjugation. These include a measure of German verb-conjugation performance before the experiment, consisting of four stem-vowel changing strong verbs (critical items) and four non-vowel-changing verbs (control items). Vocabulary size was measured with the German version of the LexTALE (www.lextale.com; see Lemhöfer & Broersma, 2012). All other variables in Table 2 were gathered through a background questionnaire (which also included some distractor questions that were supposed to support the study's cover story, asking about personal preferences regarding traveling, food and the environment, topics that would return in the learning task). The tests revealed no significant group differences for any of the variables (all $p \geq .25$, all $r \leq .18$).

Materials

We used 90 German verbs, including 32 vowel-changing strong verbs as critical items, 32 non-vowel-changing verbs as control items, and 26 non-vowel-changing verbs as filler items. Critical items were those we would analyze to assess learning of the vowel change, control items were included to detect over-application of the vowel change to verbs that did not require it, and fillers were used to reduce the proportion of vowel-changing verbs in the task but would not be analyzed (see below for details about items). The item list is available in the supplementary materials.

The selection of 32 suitable strong verbs as critical items was based on an online pilot study, testing 71 Dutch-speaking learners of German (who did not participate in the main study) on their spontaneous written production of 41 strong verbs inflected in 3SG PRES in a fill-in-the-blanks task. Verbs that yielded average stem-vowel production accuracy close to 100% or 0% were considered too easy or too difficult, respectively, and discarded. The remaining 32 final items, mostly cognates with German as many Dutch verbs are, were divided into two sets of 16 items that were matched for the verbs' average accuracy on the pilot study (see item list), type of stem-vowel alternation, word length, and transitivity. During the learning task, participants would receive input only for one of the two sets. Which of the sets this was, was counterbalanced across participants.

As control and filler items, we chose mainly high-frequency verbs and cognates with Dutch to ensure familiarity. The 32 control items contained mostly weak but also strong and irregular verbs that had the same stem vowels as the critical items, yet did not require any vowel alternations in PRES. They allowed us to detect possible overapplication of the vowel change: learners who knew/suspected the true study aim might start to apply the vowel change as a default strategy, giving rise to low accuracy rates on control items. Indeed, one participant produced more than 50% of incorrect (i.e., changed) vowels on control items and was thus excluded, because such a default strategy presumably leaves little room for item-wise learning (see Brandt et al., 2021). The control items were also divided into two equivalent sets and participants would receive input only on one set. The 26 non-vowel-changing filler verbs covered a more diverse range of stem vowels in an effort to conceal the high frequency of *a* and *e* in verb stems.

During the verb knowledge assessment at the end of the experimental session, the participants received a sheet listing all critical and control verbs, and were asked to indicate all verbs of which they did not know the meaning (prior to the

https://doi.org/10.1017/S1366728922000797 Published online by Cambridge University Press

experiment) by crossing them out. Unknown verbs were later excluded from the analysis for the respective participant (see Data preparation).

Procedure

Participants were tested individually in a quiet room. They first completed a language background questionnaire, followed by the LexTALE. Next, they received either the implicit or explicit task instructions before they performed the main learning task, i.e., the dialogue game. This game was immediately followed by the awareness interview. We also administered a short phonemic discrimination task that we later decided not to analyze for the present study. Next came a 15-minute delayed posttest during which the participants orally conjugated the critical and control items from the main task; because of its explicit nature, this task was however not found to serve as a valid posttest measure and therefore not reported (but see Appendix S3 in the supplementary materials for a discussion of these data). Lastly, the participants completed the verb knowledge assessment (see Materials). The total session duration was approximately ninety minutes. The interview and production tasks were audio-recorded. All supplementary materials (appendices, scripts, data) are available at https://osf.io/938ye/).

Learning task: dialogue game

The main task was a picture-based, meaning-centered sentence-formation task which mimicked a dialogic learning situation in German between an L2 and an L1 speaker (the experimenter). Participant and experimenter sat behind opposite computer screens displaying, per trial, a set of six pictures and the corresponding German labels and determiners (nominative, singular); in addition, a(n) (in)transitive verb in the infinitive form appeared at the screen top, and prepositions were given if applicable (Figure 1 provides an example). The first two pictures on the left side represented potential sentence subjects; the following two pairs of pictures constituted (in)direct objects. The speaker – alternately the participant or the experimenter – selected three pictures from left to right to create the most semantically plausible and 'typical' sentence that was possible given the pictures, using the given verb. The 'typicality' of the sentence meaning was to be defined subjectively; we suggested that the participants might base themselves on the first scenario that would come to their minds when looking at the pictures. When the speaker had spoken the sentence out loud, the other person was required to silently indicate their (dis)agreement that this was the most typical sentence one could form using the given words, by pushing a yes/no-button. Unbeknownst to the participants, the experimenter did not make such judgments; instead, she would code whether the participant had correctly conjugated the verb. The typicality judgments were untimed and not analyzed.

The task, presented with PsychoPy (Peirce, 2009), consisted of 280 trials and took approximately 45 minutes. As items, we used 90 different German verbs (see Materials) that were always presented along with different picture combinations, and that always needed to be conjugated in 3SG PRES. This last aspect was not made explicit to the participants, but was apparent from the task context, i.e., the pictures of singular agents and the experimenter's example utterances (note that participants almost never produced verb forms other than 3SG PRES, showing that this worked fine).

The first ten trials, involving ten verbs not present in the main task, were presented as practice and did not yet involve turn-taking. In fact, we used these trials as a short production premeasure of German conjugation knowledge (see Participants; see Table 2 for results).

Immediately after the premeasure, participants first received identical general instructions about the upcoming turn-taking and judgments. Remember that the study had been presented to all participants, irrespective of condition, as a study 'on the relationship between language and one's way of thinking'. Participants in the implicit condition received no further instructions. Those in the explicit condition, however, received one extra instruction page, informing them about the true study purpose (see Appendix S1): the central role of the vowel-changing strong verbs in the task, that they should try to conjugate these verbs correctly, and that they would have the possibility to learn from the experimenter's utterances.

During the 270 trials of the learning task disguised as a dialogue game, participant and experimenter took turns in producing sentences, with breaks after approximately every 70 trials. The participants produced all critical and control items twice during this task. Between the first (T1) and the second (T2) participant production, there were two input trials in which the experimenter produced sentences containing the correct verb form (see Figure 2 for an illustration), providing the participant the possibility to learn. To disentangle any input-based learning from accuracy changes that were unrelated to input (e.g., practice effects), the experimenter provided input only for half of the critical (and control) items.

Trial lists were constructed such that, for a given critical or control item, there were always 13 trials between T1 and T2. For items with input, there were two trials between T1 and the first input moment, five between first and second input, and four between second input and T2. To prevent excessively long trial lists and to conceal the systematicity of this recurring pattern, all items were nested: for instance, a T1 for one item could immediately be followed by an input trial for another, which could then be followed by a T2 for yet another item. Filler items did not follow this pattern, but they occurred two or four times each across participant and experimenter trials, filling empty trial 'slots' that were not occupied by T1, T2 or input trials. There were never more than two trials in a row that contained critical items.

Awareness interview

After the learning task, we conducted interviews to assess whether participants had developed awareness of the presence of the target structure and of the task's learning purpose (note that awareness of TS does not necessarily mean that there is also awareness of LP), serving also as a manipulation check. In case of a successful manipulation, participants in the explicit condition should show +TS and + LP (intentional group). Participants in the implicit condition should show -LP and may or may not have awareness of TS (+TS, -LP; incidental group, vs. -TS, -LP; unaware group).

The interview questions (see supplementary materials) were identical for all participants, regardless of condition. Following the guidelines of Rebuschat (2013), the experimenter started by asking general questions, inviting the participants to report what they thought the study was about or whether they had noticed anything special, and then passed on to increasingly specific questions. All participants who were still unaware of LP or TS

https://doi.org/10.1017/S1366728922000797 Published online by Cambridge University Press

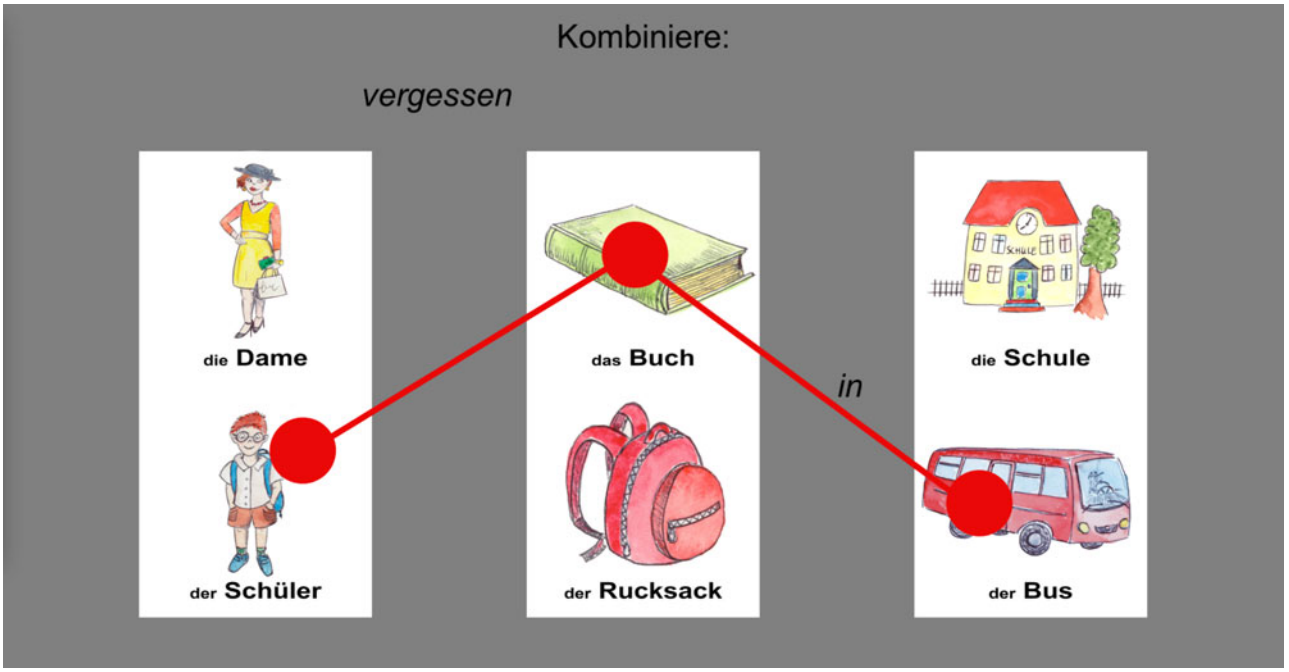


Figure 1. Illustration of an experimental trial of the dialogue game. The dots and lines represent a possible selection of three pictures out of six. Based on a horizontal combination of the selected pictures, the verb (*vergessen*) and the preposition (*in*), a sentence can be formed: *Der Schüler vergisst das Buch im Bus* ("The pupil forgets the book in the bus"). Due to copyright reasons, the pictures differ from the ones used in our experiment.

Trial nr.	Speaker	Produced sentence	Trial type
1	Participant	"Die Dame ***vergesst** das Buch im Bus." *The lady forgets the book in the bus.*	**T1**
2	Experimenter	(other trial with different item...)	
3	Participant	(other trial with different item...)	
4	Experimenter	"Die Frau **vergisst** die Brille im Café." *The woman forgets the glasses in the bar.*	**Input 1**
5	Participant	(other trial with different item...)	
6	Experimenter	(other trial with different item...)	
7	Participant	(other trial with different item...)	
8	Experimenter	(other trial with different item...)	
9	Participant	(other trial with different item...)	
10	Experimenter	"Der Radfahrer **vergisst** den Helm im Hotel." *The cyclist forgets the helmet in the hotel.*	**Input 2**
11	Participant	(other trial with different item...)	
12	Experimenter	(other trial with different item...)	
13	Participant	(other trial with different item...)	
14	Experimenter	(other trial with different item...)	
15	Participant	"Der Jäger **vergisst** das Gewehr im Wald." *The hunter forgets the gun in the forest.*	**T2**
16	Experimenter	(other trial with different item...)	
17	Participant	...	

improvement?

Figure 2. Illustration of trial order. The verb (marked in bold) used as an example is *vergessen* ("to forget"), a critical item requiring an *e-i* change in 3SG PRES. Input is provided twice, but remember that this was only true for half of all test items.

https://doi.org/10.1017/S1366728922000797 Published online by Cambridge University Press

at the end of the learning task would be informed about both at some point during the interview.

Analysis

Data preparation

We removed all verbs from the learning task dataset for which the participants had indicated that they did not know the meaning, resulting in a data loss of 4.76% for critical and 3.94% for control items. There were thus 5142 data points (2560 for critical items; 2582 for control items) from 42 participants left in the final dataset.

Scoring

Immediately after each participant production during the learning task, the experimenter registered the verb form's accuracy by assigning an error code (Table S1 in the supplementary materials), which was afterwards recoded as 0 for incorrect and 1 for correct stem vowels. Initially missing data were corrected with the help of the audio recordings. A second coder recoded 25% of the data. The first and second coder agreed on 97.87% of the trials. The remaining 2.13% were revisited by an independent third coder who agreed with the first coder in 64.71% of the cases, suggesting that 99.25% of the data can be estimated to have been coded correctly.

The binary scores at T1 and T2 represent the dependent variable for the mixed-effects analysis. For the descriptive statistics, we calculated percentages of correctly produced stem vowels based on the binary scores.

Modeling

We analyzed the data of the critical items by means of mixed-effects binomial logistic regression modeling, using maximum likelihood for model estimation. Effects are reported as significant at $p < .05$. We applied dummy coding (also called treatment coding), meaning that all parameter estimates need to be interpreted against a chosen reference category (explained below; note that we used binary factors only). We performed the analysis in *R* (R Core Team, 2018), using the *lme4* package (Bates, Mächler, Bolker & Walker, 2015; version 1.1-12). All data and scripts can be found in the supplementary materials.

We modeled the binary accuracy scores of the learner productions, including the following fixed effects: the within-participant factors Test moment (T1, T2) and Input (input, no input), the between-participants factor Group (intentional, incidental), and their interactions. We operationalized input-based learning (RQ1) as a significant accuracy increase from T1 to T2 for critical items with input, minus any increase for critical items without input. Therefore, we were mainly interested in the interaction between Test moment and Input. To assess differences in performance and learning between the two groups, we planned to investigate whether the interaction between Test moment and Input would be modulated by Group (RQ2).

As default random effects, we added random intercepts for items and for participants. We explored all additional random effects by adding the random slopes of the fixed effects and their interactions over items and/or participants to the model one by one, each time assessing whether model fit improved significantly, accompanied by an AIC decrease.

The data of the control items were not analyzed; however, we present them descriptively. They served merely as an 'alarm system' for systematic overgeneralization (see Materials), both on the participant level and for the whole group, which should become visible as unexpectedly low scores on control items.

Results

Interview outcomes

Based on the interview outcomes, the participants were reassigned to intentional, incidental, and unaware groups (see Participants). Although the intentional group had awareness of both TS and LP, these participants reported having difficulties in remaining concentrated on the verbs' stem vowels, as they also had to pay attention to semantics and case marking. Meaning and case were also the incidental group's main focus according to the interviews. These learners' awareness of TS was weak and fleeting: they reported having noticed the presence of strong verbs and having occasionally been thinking about the correct conjugation; furthermore, most of them remembered a few instances of noticing discrepancies between their own and the experimenter's productions, sometimes followed by the intention to learn. Still, these participants remained fully unaware of LP.

The interviews also revealed that almost all participants had received formal instruction about strong verb inflection in the past; nevertheless, the majority reported that they were still struggling with the correct application of the vowel change. Two participants of the incidental group and one of the intentional group reported having learned the vowel change only through immersion contexts. Three participants – all part of the unaware group – reported no prior knowledge.

Descriptive statistics

Table 3 provides descriptive statistics over all factor levels and combinations. The means are plotted in Figure 3.

For the critical items, we see the expected interaction between Test moment and Input for both groups: there was an increase in accuracy from T1 to T2 on items for which the experimenter provided input; without input, the mean scores did not change. The slope of the accuracy increase looks similar in both groups, suggesting comparable learning effects. In addition, we can observe a main effect of Group: overall, the intentional group obtained higher scores than the incidental group.

As for the control items, both groups had high scores overall, which remained fairly steady across T1 and T2. In comparison, the scores on critical items were substantially lower. We can thus be confident that there was no substantial degree of sample-level strategic overgeneralization.

Model comparisons

For the analysis of the critical items, the best model we could identify (for the modeling procedure, see Modeling) took the following form: (Stem-vowel accuracy) ~ 1 + Input*Test moment*Group + (1 + Input|Item) + (1|Participant). The first term in this notation represents the dependent variable; the remaining terms are the model terms. The asterisks mark an interaction. The random-effects terms are those which include the bar symbol (|); 1 represents an intercept. This model differs from the initial model by its inclusion of the random slopes of Input over items, which significantly improved model fit ($p = .03$). All model comparisons and an evaluation of model fit

https://doi.org/10.1017/S1366728922000797 Published online by Cambridge University Press

Table 3. Descriptive statistics of the percentage of correctly produced stem vowels

	Input			No input		
	T1 (%)	T2 (%)	T2-T1 (pp)	T1 (%)	T2 (%)	T2-T1 (pp)
Critical items						
Intentional group						
Mean	63.52	79.40	15.88	62.47	61.52	−0.95
SD	26.51	25.01	15.85	22.00	24.16	8.12
95% CI	51.41–73.42	65.63–87.63	8.98–22.54	52.35–70.90	51.24–71.03	−4.72–2.19
Incidental group						
Mean	37.04	54.80	17.76	41.25	41.66	0.41
SD	28.06	29.29	12.04	26.75	28.69	9.22
95% CI	26.29–49.55	41.90–66.48	12.90–22.95	30.38–52.84	29.97–54.09	−3.98–3.79
Control items						
Intentional group						
Mean	85.09	92.14	7.05	87.08	87.42	0.34
SD	7.33	7.21	6.65	8.60	8.88	7.37
95% CI	82.13–88.25	88.85–94.86	4.63–10.41	83.30–90.41	83.33–90.78	−2.98–3.12
Incidental group						
Mean	93.15	92.65	−0.50	91.47	91.65	0.17
SD	7.10	6.93	5.68	7.59	5.90	7.36
95% CI	89.58–95.83	89.31–95.24	−2.95–1.79	87.96–94.35	88.93–93.91	−2.86–3.31

Note. T1 = Test moment 1; T2 = Test moment 2; T2-T1 = difference score, expressed in percentage points (pp); CI = 95% confidence interval of the mean (10,000 samples BCa bootstrapping).

are provided in the supplementary materials (Appendices S4 and S5).

Inferential statistics

Interpretation of model output

Table 4 provides the outcomes of type III tests of fixed effects for our model, which indicate which of the main and interaction effects were significant. To interpret these effects, we use Table 5, providing the model's parameter estimates that are on the logit scale. We also report the estimated probabilities, since they are most intuitive to interpret. They express the probability of the strong verbs' stem vowels being produced correctly, given specific factor-level combinations.

As dummy coding was used, all fixed effects in Table 5 need to be interpreted against the model intercept, which represents the following reference level combination: the intentional group, tested at T1, on critical items for which no input would be provided. The logit is 0.69 and corresponds to a predicted 0.67 probability (or 67%) for intentional learners at T1 to produce a correct stem vowel for no-input items. We report the simple and

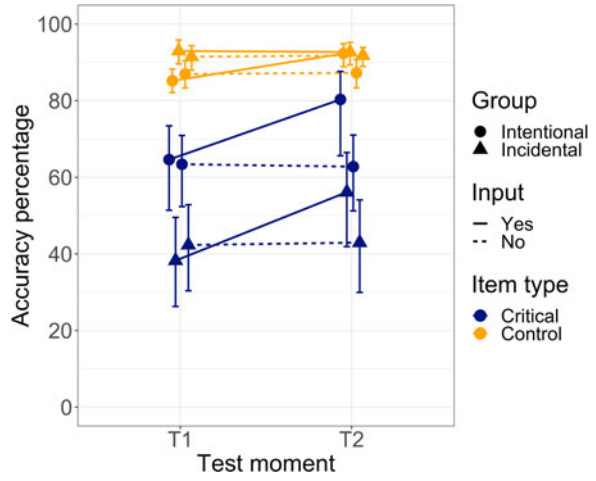


Figure 3. Mean test scores on the different test moments for the intentional and incidental groups. Error bars represent 95% confidence intervals (10,000 samples BCa bootstrapping).

Table 4. Outcomes of the type III Wald chi-square tests of fixed effects

Fixed effects	χ^2	*df*	*p*
(Intercept)	2.85	1	.091
Input	0.24	1	.624
Test moment	0.04	1	.837
Group	6.79	1	**.009**
Input: Test moment	16.54	1	**<.001**
Input: Group	1.18	1	.278
Test moment: Group	0.08	1	.771
Input: Test moment: Group	0.06	1	.805

Note. Significant *p*-values are printed in bold. Colons (:) represent interactions.

https://doi.org/10.1017/S1366728922000797 Published online by Cambridge University Press

Table 5. Parameter estimates of the mixed-effects model

Fixed effects	Logit	Probability	Odds ratio
(Intercept)	0.69 [−0.11, 1.49]	0.67 [0.47, 0.82]	1.99 [0.90, 4.43]
Input = yes	0.10 [−0.32, 0.52]	0.69 [0.40, 0.88]	1.11 [0.73, 1.69]
Test moment = T2	−0.04 [−0.44, 0.35]	0.66 [0.37, 0.86]	0.96 [0.65, 1.42]
Group = incidental	−1.32 [−2.32, -0.33]	0.35 [0.08, 0.76]	0.27 [0.10, 0.72]
Input = yes: Test moment = T2	1.21 [0.63, 1.79]	0.88 [0.44, 0.98]	3.35 [1.87, 6.00]
Input = yes: Group = incidental	−0.31 [−0.87, 0.25]	0.30 [0.03, 0.87]	0.73 [0.42, 1.28]
Test moment = T2: Group = incidental	0.08 [−0.48, 0.64]	0.36 [0.03, 0.90]	1.09 [0.62, 1.90]
Input = yes: Test moment = T2: Group = incidental	−0.10 [−0.91, 0.71]	0.58 [0.01, 1.00]	0.90 [0.40, 2.02]
Random effects	**Variance**	***SD***	
Participant (Intercept)	2.26	1.50	
Item (Intercept)	1.22	1.11	
Input = yes	0.18	0.42	

Note. The intercept represents the following combination of factor levels: Test moment = T1, Input = no; Group = intentional. The reported logits and odds ratios, but not the probabilities, need to be interpreted against the intercept. Squared brackets contain 95% confidence intervals. Equal signs (=) indicate the level of categorical predictors; colons (:) represent interactions.

interaction effects along with the factor level that is to be compared with the reference level of that factor. For instance, 'Input = yes' refers to the effect of input compared to no input.

We used odds ratios (ORs) as an effect size. They express the change in the odds of producing a correct vowel, associated with the change from the reference factor level to another factor level. ORs of 1 represent equal probabilities for both factor levels (i.e., no effect). When close to 0 or considerably higher than 1, ORs signal large effects. A more detailed explanation about logits and ORs is available in Appendix S6 (supplementary materials).

Outcomes

RQ1 and RQ2 addressed whether there was learning from input, and whether this was different in the two groups. Table 4 shows that the observed Input x Test moment interaction for critical items was significant. Table 5 shows that at T2 as compared to T1, the intentional group was 3.35 times more likely to correctly apply the vowel change for input items as compared to no-input items, a medium-sized effect reflecting learning. The absence of a significant three-way interaction between Input, Test moment and Group indicates that the learning gains were similar in the incidental group. At T2 as compared to T1, the incidental learners were 3.03 times more likely to produce a correct vowel for input items than for no-input items; this corresponds to an estimated 58% accuracy at T2 after input, as compared to only 36% for no-input items at T2, and 30% for input items and 35% for no-input items at T1 (values obtained through model releveling with 'incidental' as reference level; all other estimated values can be read directly from Table 5). The absence of significant main effects of Input and Test moment indicates equal scores on input and no-input items at T1, and stable scores on no-input items across T1 and T2 for the intentional group. The absence of significant interactions between these factors and Group indicates this was similar in the incidental group. Taken together, the analysis revealed significant and comparably large learning effects in both groups (RQ1, RQ2): participants became better at producing correct stem vowels from T1 to T2 during the learning task, but only after hearing input. The (observed) mean accuracy then increased by 15.88 percentage points (pp) in the intentional group and by 17.66 pp in the incidental group (Table 3).

There was also a significant main effect of Group (see Table 4): at T1 for no-input items, the intentional group was 3.70 (i.e., 1/0.27; medium-to-large effect) times more likely to produce correct stems than the incidental group (see Table 5); put differently, these learners started the task with scores that were on average 23.84 pp higher than the incidental group's scores (RQ2). This difference remained stable throughout the task, as there was no interaction between Group and Test moment.

Unaware group and posttest

As mentioned above, we decided not to report and discuss parts of our collected data in the main text due to validity issues; however, analyses of these data can be found in the supplementary materials. This concerns the data of the unaware subgroup (Appendix S2), consisting of only six participants (rendering reliable statistical analyses impossible), half of which had no prior knowledge of the target structure, unlike the remainder of the sample. No trends suggesting learning could be detected in this small group.

Furthermore, the supplementary materials also include the data of the 15-minute delayed posttest (Appendix S3): in the incidental group, we observed a sudden increase in accuracy at this test for both input and no-input items (without having provided further input). This was probably due to the awareness of the target structure after the debriefing in the awareness interviews (which were, for good reasons, conducted immediately after the learning task) and to the posttest's explicit task format (requiring the inflection of verb infinitives in isolation). This change in awareness in the incidental group and the different task format undermined the valid use of this posttest as a pure longer-term learning assessment.

Discussion

The current study used a lab-based learning treatment, embedding oral learner output trials before and after native speaker input trials all in one dialogue game, to examine the further learning of stem allomorphy in strong German verbs under explicit and implicit conditions. The conditions were implemented through diverging task instructions that caused the participants to be aware (intentional group) or unaware (incidental group) of the task's learning purpose. We ensured that all participants received exactly the same, meaning-based learning task; the only difference was the additional instruction to focus on the vowel change and to learn from the input in the explicit condition. The interviews proved that the learning purpose remained successfully hidden from the incidental group, confirming that our manipulation had been effective.

Despite the brevity of the intervention, the statistical analysis revealed that engaging in dialogue-based learning, including

https://doi.org/10.1017/S1366728922000797 Published online by Cambridge University Press

exposure to two instances of correct input for each verb, helped to improve learners' production accuracy by 16.82 pp (equaling a 33.45% increase). As there was no improvement on items without input, we are confident that these results reflect learning from exposure. Crucially, awareness of the task's learning purpose did not make a difference for the (absolute) learning gains: the increase in accuracy was statistically indistinguishable in the intentional and incidental groups. However, the intentional group showed higher overall performance across all productions (before and after input) alike. This was presumably a direct effect of our task-instruction manipulation, and highly unlikely – though not completely impossible – to be a consequence of a priori group differences, given that group equivalence had been statistically accounted for (see Participants).

Medium-sized overall learning effect

We found significant, robust learning of the vowel change across the entire sample (RQ1). Although the treatment consisted of only two instances of exposure, the effect was medium-sized.

Our participants had prior knowledge of stem-vowel alternations, but may have had insufficient exposure and opportunities for input- and output-based practice. What we observe as learning in this study therefore equals a reactivation and expansion of existing but limited knowledge of the conjugation paradigm. In the absence of a suitable delayed posttest, the learning effect we found does not guarantee sustained learning. Thus, the learning we observed can be seen as micro-steps of L2 learning that may or may not lead towards the development of sustainable knowledge available for active language use, and that can occur during conversation by picking up features of the interlocutor's speech.

At the representational level, one possibility to account for the observed learning in our study is that it involves the co-existence of two competing morphological representations of the same verb stem (e.g., *sprech-* versus *sprich-*, see Table 1). Prior to our experiment, the unchanged stem (*sprech-*) is likely to have been the erroneous, default representation for most strong verbs in our learners' minds, a finding that was also observed in Godfroid (2016), Krause et al. (2015), and in our pilot study (see above). The provision of input may then have triggered the creation of a new, competing representation (*sprich-*: correct, changed stem) which started to prevail over the old representation. For some verbs, a correct but weak representation may already have been in place – given the advanced learners' prior knowledge – which was then further boosted by the input.

Some researchers might prefer the term 'priming' to the term 'learning' in this context. However, we do not regard this distinction as relevant for our study, nor can we, on the grounds of our data, make such a distinction. Developmental psychologists see priming in children as a source of learning, including also long-term learning effects (e.g., Peter, Chang, Pine, Blything & Rowland, 2015). The same may of course hold for adults (Jackson & Hopp, 2020): in principle, priming has effects that vary continuously from short- to long-lasting, with more repetition having more long-lasting effects, which is learning (also see Chang, Dell & Bock, 2006; Pickering & Ferreira, 2008). In this respect, priming becomes a part of the acquisition process.

The implementation of intentional vs. incidental learning

Our task-instruction manipulation effectively influenced our participants' awareness. The interview results led us to distinguish between an intentional, an incidental, and an unaware group, reflecting the type of learning that the participants had predominantly engaged in. While the intentional group was aware of the LEARNING PURPOSE – entailing an explicit, goal-directed type of learning – this awareness was absent in the incidental group.

Both groups were aware of the presence of the TARGET STRUCTURE in the learning task, yet this awareness appeared to be much weaker and more fleeting in the incidental group. However, our interviews were not suitable to exactly quantify this difference in the extent and strength of awareness. The presence of awareness of the target structure in the incidental group may be a result of the task's output component (see Izumi, 2002; also see de Vos et al., 2019): the output trials prior to input trials may have prompted the learners to reflect on the verb conjugation; subsequently, the learners may have realized they were uncertain about the correct form to use; this may then have led them to be more attentive to the target form in the input, promoting learning.

We operationalized 'incidental' learning as learning without being aware that one is engaging in a language learning activity (de Vos et al., 2019; Hulstijn, 2003). Note that this does not necessarily exclude intentionality: in fact, most incidental learners manifested a fleeting form of intentional learning. Unlike with the intentional group though, this intentionality was not externally induced as would be typical for learning in classroom environments. Rather, it can be interpreted as a spontaneous and voluntary intention to learn after noticing uncertainty or a mismatch between one's own and the interlocutor's productions, not unlike what may happen during a natural conversation with a more proficient speaker (Gass & Varonis, 1994; Long, 1996).

These outcomes also show that our study was a successful extension of the naturalistic learning paradigm implemented by de Vos et al. (2019) and Brandt et al. (2021) to a new morphosyntactic feature and to the comparison of explicit and implicit learning conditions. Recall that Brandt et al. (2021) also found incidental learning of a difficult morphosyntactic structure (gender markings in determiners), but after only one instance of input and without comparing incidental learning to learning under explicit conditions. It goes without saying that the dialogue games implemented in these studies and in ours only approximate natural dialogue – as close as one can get when the exact to-be-produced forms are previously determined by the experimental design; nevertheless, we may conclude that this paradigm is well suited to investigate incidental learning.

The unaware group – which was excluded due to an extremely small sample size – showed neither awareness of the target structure nor of the learning goal. Any attested learning would thus reflect implicit learning; yet, no evidence of learning could be detected. Whether this null effect is 'real' or a result of insufficient statistical power is uncertain.

No influence of task awareness on degree of learning

The analyses revealed similar learning effects in the intentional and incidental groups (RQ2). Note that we refer here to absolute learning gains, which is the standard way to analyze pre-post data. At first glance, this finding deviates from numerous previous studies that provided evidence for the beneficial role of explicit instruction for grammar acquisition (e.g., Doughty, 1991; Robinson, 1996), but is in line with the findings by Andringa et al. (2011) and Kang et al. (2019). It should be borne in mind that the learning conditions implemented in our study differ

https://doi.org/10.1017/S1366728922000797 Published online by Cambridge University Press

considerably from how explicit and implicit conditions are typically operationalized in more traditional, often classroom-based quasi-experimental instruction studies, usually involving an extensive instruction component providing metalinguistic descriptions of the target structure in combination with structured, focused exercises (explicit condition), or input flooding and/or input enhancement (implicit condition). Moreover, while most comparative studies focus on the beginning stages of L2 learning – especially those using (semi-)artificial languages (e.g., Morgan-Short, Sanz, Steinhauer & Ullman, 2010) – our study used an advanced learner group and a target structure known to be a lasting source of errors for this population.

These results can, however, also be interpreted from a percentual perspective, taking into account the room for improvement that is left in both groups after T1. As the intentional group had a higher initial performance level than the incidental group (see next section), there was less room for improvement left in the former group. According to this view, initial error rates were reduced by 44% after the provision of input in the intentional group, compared to 28% in the incidental group (for similar considerations of learning gains, see Lemhöfer, Schriefers & Hanique, 2010): the results would thus be in line with prior research showing higher learning gains in explicit conditions. On the other hand, the two groups were comparable in terms of HOW MANY verbs (out of all critical verbs) they improved on, which is the standard way to look at differences between conditions based on our essentially additive analysis methods. Thus, while we acknowledge that learning gains may differ between our groups in relative terms, they did not do so in absolute numbers, suggesting that awareness of the learning purpose was surely not a prerequisite for learning. Rather, the incidental group learned comparable amounts despite remaining naïve of the learning purpose; this is in itself a remarkable finding that nicely illustrates how learning can happen (or at least begin to happen) incidentally during interaction.

Positive influence of task awareness on overall performance

Although it did not affect the learning gains, awareness of the task's learning purpose (and possibly also the larger degree of awareness of the target structure) considerably influenced overall task performance: already at T1, before exposure to the target forms, the intentional group's accuracy scores were on average 23.84 pp higher than those of the incidental group, a medium-to-large effect. This finding suggests that the intentional group benefitted from reliance on explicit knowledge and controlled processing, having been reminded of the vowel change and trying to apply it correctly. By contrast, the incidental group's productions were less accurate, probably because they were more spontaneous and less premeditated. Note that the control items (that did not require vowel changes) did not show such group differences, which rules out the possibility that the effect was merely due to a 'blind' overapplication of the vowel change in the intentional group. It is also noteworthy that the 15-minute delayed posttest, which was conducted under explicit conditions (see above), showed a large accuracy improvement compared to T2 in the incidental group only (see above, see supplementary materials), supporting the large benefits of awareness of the task's learning goal. Translated to real-life L2 experiences, the group effect suggests that the conscious effort of trying to produce correct linguistic structures can indeed increase accuracy.

Our experimental design made it possible to disentangle this direct effect of the task-instruction manipulation on production accuracy from any input-related effects, because we measured T1 AFTER the task instructions. If we had used a traditional pretest (T1) PRIOR to our instruction manipulation, both task instruction AND input would have taken place in between T1 and T2 (posttest). The learning gains between T1 and T2 could thus have been a result of the instruction, the input, or both. It would, in other words, not have been possible to distinguish between effects of the task instructions and those of input, and effects of the instructions may have been misinterpreted as higher input-based learning gains. Moreover, in traditional pretest-treatment-posttest designs, participants often know that they are being tested, having a potential impact on awareness and reducing the difference between incidental and intentional conditions. That our incidental group remained unaware of the learning purpose demonstrates that our design – in which T1, input and T2 trials were nested in one dialogue game and thus all took place under the same conditions (either explicit or implicit) – constitutes a promising alternative to such traditional set-ups.

German strong verbs as a target structure

In line with prior research (Godfroid, 2016; Godfroid & Uggen, 2013; Krause et al., 2015), the stem-vowel alternations in German strong verbs clearly represented a learning difficulty for our Dutch-speaking participants. Despite their prior knowledge, they produced a considerable number of errors throughout the experiment. The weak conjugation unmistakably represented the default paradigm in their interlanguages, which was confirmed by the high scores almost at ceiling on the non-vowel-changing control verbs. The stem-vowel alternations represent a good showcase of robust errors in L2 speakers on a non-transparent, difficult morphosyntactic feature that has to be learned for each lexical item individually. Other examples for such features are word gender or plural inflections in languages where these features are (semi-)opaque (e.g., German).

Our research extends Godfroid and Uggen (2013) and Godfroid (2016) to a new participant group with a different L1 and to a learning treatment comprising an oral production component in addition to the comprehension component. Our study found incidental learning gains of 17.66, which were similar to those of Godfroid and Uggen (2013, p. 308) who compared written production pre- and posttests and observed an average 17.33 pp increase. In contrast, the gains in Godfroid (2016, p. 200; measured with oral production pre- and posttests) were only 7.67 pp and attributed to a test-retest effect because they did not significantly differ from those observed in a no-treatment control group. Although it is impossible to tell which factor exactly caused these gain differences, the fact that the learning task in Godfroid (2016) had the most implicit, least awareness-raising format might have played a role. In our study, incidental learning was possibly enhanced by the interactive nature of the treatment. As for Godfroid and Uggen (2013), awareness levels may have been higher due to the use of the written modality, promoting learning. Such an interpretation would again point towards the facilitatory role of awareness for learning.

Conclusions

The present study extended natural language learning research by applying a lab-based interactive learning treatment, comprising

https://doi.org/10.1017/S1366728922000797 Published online by Cambridge University Press

oral learner output and native speaker input, to stem allomorphy in German strong verbs, which represents a source of persistent errors even in advanced learners. The results showed considerable degrees of morphosyntactic, item-wise learning that were independent of whether or not an explicit instruction to learn had been provided. However, awareness of the task's learning purpose did increase the overall accuracy rates during the task (already before input), reflecting a beneficial effect of a focus on form on language performance.

The mere fact that we found learning effects of a target structure as difficult (unproductive, redundant, non-salient) as ours and following such a brief learning treatment (two learner productions and two exposures to native input per item) is noteworthy. However, because of the indispensable post-experimental interview that gave away the learning purpose of the study, we were unable to assess long-term learning effects while preserving the awareness status of the two groups. In a next step, it would be interesting to explore long-term learning effects in order to assess whether the dialogue-based learning treatment does actually lead towards the development of sustainable knowledge. It seems conceivable though that the current findings concerning immediate learning gains will likely be preserved, or even enlarged, when longer and more intensive input is given.

The study also demonstrates the advantages of an innovative dialogue-based experimental design in which pre- and posttest trials (L2-learner output) and treatment trials (L1-speaker input) are all embedded in one dialogue game. It resembles natural learning situations outside the classroom more closely than traditional pretest/posttest designs and is, in our view, more suited to conceal the learning goal of the study. The latter seems essential for a fair comparison of learning with and without a learning intention. We therefore think that the paradigm is a valuable addition to existing designs, especially when it comes to the study of 'truly' incidental learning.

In sum, our findings show not only that advanced L2 speakers can learn to improve on difficult and opaque morphosyntactic structures after naturalistic input, and to comparable degrees under intentional and incidental conditions, but also that both the learning paradigm and the target structure we used could be of great potential use to future experimental studies investigating language learning under more natural conditions.

Data Availability Statement. Data, analysis scripts and additional supplementary materials are publicly available via the Open Science Framework site for this project: https://doi.org/10.17605/OSF.IO/938YE

Acknowledgements. This research has been supported by an FWO grant (Research Foundation – Flanders) awarded to Alex Housen and Aline Godfroid, and by a VIDI grant awarded to Kristin Lemhöfer by the NWO (the Dutch Research Council). The authors would like to thank Lara Stas and Bram Bulté for their advice with respect to the statistical analyses, and Bastien De Clercq and three anonymous reviewers for their helpful suggestions and comments on earlier versions of the manuscript.

References

Andringa S (2020) The emergence of awareness in uninstructed L2 learning: A visual world eye tracking study. *Second Language Research* **36**, 335–357.

Andringa S, de Glopper K and Hacquebord H (2011) Effect of explicit and implicit instruction on free written response task performance. *Language Learning* **61**, 868–903.

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

Brandt AC, Schriefers H and Lemhöfer K (2021) A laboratory study of naturalistic second language learning: Acquiring grammatical gender from simple dialogue. *Journal of Experimental Psychology: Learning, Memory, & Cognition*. Advance online publication.

Bybee J and Newman J (1995) Are stem changes as natural as affixes? *Linguistics* **33**, 633–654.

Chang F, Dell GS and Bock K (2006) Becoming syntactic. *Psychological Review* **113**, 234–272.

Cintrón-Valentín M and Ellis NC (2015) Exploring the interface: Explicit focus-on-form instruction and learned attentional biases in L2 Latin. *Studies in Second Language Acquisition* **37**, 197–235.

Conroy MA and Antón-Méndez I (2015) A preposition is something you can end a sentence with: Learning English stranded prepositions through structural priming. *Second Language Research* **31**, 211–237.

DeKeyser RM (1995) Learning second language grammar rules: An experiment with a miniature linguistic system. *Studies in Second Language Acquisition* **17**, 379–410.

DeKeyser RM (2005) What makes learning second-language grammar difficult? A review of issues. *Language Learning*, **55**, 1–25.

de Vos JF, Schriefers H, ten Bosch L and Lemhöfer K (2019) Interactive L2 vocabulary acquisition in a lab-based immersion setting. *Language, Cognition and Neuroscience* **34**, 916–935.

Doughty CJ (1991) Second language instruction does make a difference: Evidence from an empirical study of SL relativization. *Studies in Second Language Acquisition* **13**, 431–469.

Ellis NC (1993) Rules and instances in foreign language learning: Interactions of explicit and implicit knowledge. *European Journal of Cognitive Psychology* **5**, 289–318.

Ellis NC (2006a) Language acquisition as rational contingency learning. *Applied Linguistics* **27**, 1–24.

Ellis NC (2006b) Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics* **27**, 164–194.

Ellis NC (2006c) The associative-cognitive CREED. In VanPatten B and Williams J (eds), *Theories in second language acquisition: An Introduction*. Mahwah, NJ: Erlbaum, pp. 77–96.

Ellis R (2009) Implicit and explicit learning, knowledge and instruction. In Ellis R, Loewen S, Elder C, Erlam R, Philip J and Reinders H (eds), *Implicit and explicit knowledge in second language learning, testing and teaching*. Bristol: Multilingual Matters, pp. 3–25.

Ellis R, Loewen S and Erlam R (2006) Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in Second Language Acquisition* **28**, 339–368.

Gass SM (2003) Input and interaction. In Doughty CJ and Long MH (eds), *The handbook of second language acquisition*. Malden, MA: Blackwell Publishing Ltd., pp. 224–256.

Gass SM and Varonis EM (1994) Input, interaction, and second language production. *Studies in Second Language Acquisition* **16**, 283–302.

Godfroid A (2016) The effects of implicit instruction on implicit and explicit knowledge development. *Studies in Second Language Acquisition* **38**, 177–215.

Godfroid A and Uggen MS (2013) Attention to irregular verbs by beginning learners of German. *Studies in Second Language Acquisition* **35**, 291–322.

Goo J, Grañena G, Yilmaz Y and Novella M (2015) Implicit and explicit instruction in L2 learning: Norris & Ortega (2000) revisited and updated. In Rebuschat P (ed.), *Implicit and explicit learning of languages*, Vol. **48**. Amsterdam/Philadelphia: John Bejamins, pp. 443–482.

Hopp H (2013) Grammatical gender in adult L2 acquisition: Relations between lexical and syntactic variability. *Second Language Research* **29**, 33–56.

Housen A (2020) Complexity and difficulty of language features and second language instruction. In Chapelle CA (ed.), *The concise encyclopedia of applied linguistics*. Malden, MA: John Wiley & Sons Ltd., pp. 388–396.

Housen A and Simoens H (2016) Introduction: Cognitive perspectives on difficulty and complexity in L2 acquisition. *Studies in Second Language Acquisition* **38**, 163–175.

Hulstijn JH (2003) Incidental and intentional learning. In Doughty CJ and Long MH (eds), *The handbook of second language acquisition*. Oxford: Blackwell, pp. 349–381.

Izumi S (2002) Output, input enhancement, and the noticing hypothesis. *Studies in Second Language Acquisition* **24**, 541–577.

Jackson CN and Hopp H (2020) Prediction error and implicit learning in L1 and L2 syntactic priming. *International Journal of Bilingualism* **24**, 895–911.

Kang EY, Sok S and Han ZH (2019) Thirty-five years of ISLA on form-focused instruction: A meta-analysis. *Language Teaching Research* **23**, 428–453.

Köpcke KM (1998) Prototypisch starke und schwache Verben der deutschen Gegenwartssprache. *Germanistische Linguistik* **141**, 45–60.

Krause H, Bosch S and Clahsen H (2015) Morphosyntax in the bilingual mental lexicon. *Studies in Second Language Acquisition* **37**, 597–621.

Larsen-Freeman D (2010) Not so fast: A discussion of L2 morpheme processing and acquisition. *Language Learning* **60**, 221–230.

Lemhöfer K and Broersma M (2012) Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods* **44**, 325–343.

Lemhöfer K, Schriefers H and Hanique I (2010) Native language effects in learning second-language grammatical gender: A training study. *Acta Psychologica* **135**, 150–158.

Lemhöfer K, Schriefers H and Indefrey P (2020) Syntactic processing in L2 depends on perceived reliability of the input: Evidence from P600 responses to correct input. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **46**, 1948–1965.

Leow RP and Hama M (2013) Implicit learning in SLA and the issue of internal validity. Studies in *Second Language Acquisition* **35**, 545–557.

Leung JHC and Williams JN (2012) Constraints on implicit learning of grammatical form-meaning connections. *Language Learning* **62**, 634–662.

Long MH (1996) The role of the linguistic environment in second language acquisition. In Ritchie WC and Bhatia TK (eds), *Handbook of second language acquisition*. San Diego, CA: Academic Press, pp. 413–468.

McDonough K and Mackey A (2008) Syntactic priming and ESL question development. *Studies in Second Language Acquisition* **30**, 31–47.

Morgan-Short K, Sanz C, Steinhauer K and Ullman MT (2010) Second language acquisition of gender agreement in explicit and implicit training conditions: An event-related potential study. *Language Learning* **60**, 154–193.

Norris, JM and Ortega L (2000) Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning* **50**, 417–528.

Norris JM and Ortega L (2001) Does type of instruction make a difference? Substantive findings from a meta-analytic review. *Language Learning* **51**, 157–213.

Ortega L (2009) *Understanding second language acquisition*, Vol. **44**. London/New York: Routledge.

Peirce JW (2009) Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics* **2**, 1–8.

Peter M, Chang F, Pine JM, Blything R and Rowland CF (2015) When and how do children develop knowledge of verb argument structure? Evidence from verb bias effects in a structural priming task. *Journal of Memory and Language* **81**, 1–15.

Pickering MJ and Ferreira VS (2008) Structural priming: A critical review. *Psychological Bulletin* **134**, 427–459.

R Core Team (2018) *R: A language and environment for statistical computing* [Computer software]. Vienna: R Foundation for Statistical Computing. Available from https://www.R-project.org/.

Rebuschat P (2013) Measuring implicit and explicit knowledge in second language research. *Language Learning* **63**, 595–626.

Rebuschat P and Williams JN (2012) Implicit and explicit knowledge in second language acquisition. *Applied Psycholinguistics* **33**, 829–856.

Robinson P (1996) Learning simple and complex second language rules under implicit, incidental, rule-search and instructed conditions. *Studies in Second Language Acquisition* **18**, 233–247.

Schmidt RW (1995) Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. In Schmidt RW (ed.), *Attention and awareness in foreign language learning*. Honolulu, HI: National Foreign Language Resource Center, pp. 1–63.

Spada N and Tomita Y (2010) Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning* **60**, 263–308.

Williams JN (2009) Implicit learning in second language acquisition. In Ritchie WC and Bhatia TK (eds), *The new handbook of second language acquisition*. Bingley, UK: Emerald Press, pp. 319–353.

https://doi.org/10.1017/S1366728922000797 Published online by Cambridge University Press