Morphological convergence as on-line lexical analogy

Péter Rácz, Clay Beckner, Jennifer B. Hay, Janet B. Pierrehumbert

Language, Volume 96, Number 4, December 2020, pp. 735-770 (Article)

Published by Linguistic Society of America
DOI: *https://doi.org/10.1353/lan.2020.0061*

MORPHOLOGICAL CONVERGENCE AS ON-LINE LEXICAL ANALOGY

PÉTER RÁCZ
Central European University and
University of Canterbury

CLAY BECKNER
University of Warwick and
University of Canterbury

JENNIFER B. HAY
University of Canterbury

JANET B. PIERREHUMBERT
University of Oxford and
University of Canterbury

The English past tense contains pockets of variation, where regular and irregular forms compete (e.g. *learned*/*learnt*, *weaved*/*wove*). Individuals vary considerably in the degree to which they prefer irregular forms. This article examines the degree to which individuals may converge on their regularization patterns and preferences. We report on a novel experimental methodology, using a cooperative game involving nonce verbs. Analysis of participants' postgame responses indicates that their behavior shifted in response to an automated co-player's preferences, on two dimensions. First, players regularize more after playing with peers with high regularization rates, and less after playing with peers with low regularization rates. Second, players' overall patterns of regularization are also affected by the particular distribution of (ir)regular forms produced by the peer.

We model the effects of the exposure on participants' morphological preferences, using both a rule-based model and an instance-based analogical model (Nosofsky 1988, Albright & Hayes 2003). Both models contribute separately and significantly to explaining participants' pre-exposure regularization processes. However, only the instance-based model captures the shift in preferences that arises after exposure to the peer. We argue that the results suggest an account of morphological convergence in which new word forms are stored in memory, and on-line generalizations are formed over these instances.*

Keywords: morphology, convergence, computational modeling, language variation and change, generalized context model, minimal generalization learner

1. INTRODUCTION. Investigations of verbal inflection—with a particular focus on the English past tense—have been a mainstay in linguistics for several decades. Variation between the regular and irregular past tenses has provided the basis for long-standing debates over language acquisition and innateness, the nature of linguistic representation and generalization, and processes of language change (Bybee & Slobin 1982, Bybee & Moder 1983, Rumelhart & McClelland 1986, Plunkett & Marchman 1991, Hare & Elman 1995, McClelland & Patterson 2002, Albright & Hayes 2003, Seidenberg & Plaut 2014).

The regular past-tense form is more productive than any of the irregular past tenses. However, the irregular past tenses can be productive for novel verb stems to some extent. This extent depends on how many existing stems are similar to the irregular form, and in what ways; the details of this variation are precisely what different theories undertake to explain. Variation across individuals or contexts has been rather less explored. While numerous developmental studies explore the child's path toward the typical adult pattern, there is also a great deal of variability among adults. Some adults prefer regular forms more pervasively than others. This variability provides an opportu-

* This work was supported by a grant from the John Templeton Foundation (Award ID 36617) to JBP and JH, and a Royal Society of New Zealand Rutherford Discovery Fellowship (Grant No. E5909) to JH. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the funding agencies. The authors would like to thank Lisa Dawdy-Hesterberg, Chun Liang Chan, Robert Fromont, Adam Albright, Pat LaShell, Jacqui Nokes, Jacq Jones, Ryan Podlubny, our steadfast associate editors Linda Wheeldon and Megan Crowhurst, Marc Brysbaert, Volya Kapatsinski, and three anonymous referees. All faults remain ours.

nity to develop a connection between quantitative models of verbal inflection and a key phenomenon in sociolinguistics, namely convergence between speakers. Convergence occurs when speakers who begin an interaction with differences in their linguistic systems use more similar forms as the interaction progresses. By studying convergence, we can gain insights into how the cognitive system represents and manipulates variability in morphology.

Our article considers the degree to which preferences in verbal inflection are affected by interaction between speakers. For example, if your conversational partner says *wove* as the past tense of *weave*, are you more likely to use the irregular past tense of another variable stem (e.g. *dove* for *dive*)? Will you generalize to similar novel stems? To address such questions, we use an innovative experimental methodology to expose participants to different (ir)regularization patterns in a cooperative task. We then observe how these different patterns of exposure affect their subsequent past-tense preferences. Our results have broad relevance for the study of morphology and the lexicon.

The article is organized as follows: Section 2 provides an overview of the existing literature on linguistic convergence, morphological models, and the English past tense in particular. We outline our hypotheses in §3 and then turn to an outline of our experimental paradigm in §4, which involves exposing people to different versions of the English past-tense distribution using novel past-tense forms, and then observing the effects on their subsequent past-tense preferences. The results, discussed in §5, show that individuals are influenced not only by the overall level of regularization in the set of words they are exposed to, but also by the details of the particular distribution of these words. In §6 we consider how these shifts in preference can be accounted for by two different classes of model, one involving analogical generalization over instances (Nosofsky 1988), and one involving the inference of more abstract rules (Albright & Hayes 2003). We find that both make statistically independent contributions in explaining participants' initial preferences, indicating that English past-tense formation is governed by a combination of 'rules' and 'analogy'. However, only the analogical model is able to capture participants' shifts in preferences, based on the new forms to which they have been exposed. The consequences of these results are discussed in §7.

2. Background.

2.1. Morphological convergence. Interlocutors tend to converge. That is, as people use language to communicate with each other, they tend both to sound more similar to each other and to make more similar choices of words and phrases (Giles & Coupland 1991, Garrod & Pickering 2004, Pickering & Garrod 2004, 2006). Convergence has been attested in a wide variety of linguistic domains, including naming preferences (Brennan & Clark 1996, Roberts 2010), syntax (Estival 1985, Bock 1986, Gries 2005, Hall et al. 2015), and basic phonetic properties like speech rate (Webb 1972) and fundamental frequency (Gregory et al. 1993, Babel & Bulatov 2012). For instance, in a range of studies, Babel (2010, 2012) shows that, when a speaker has a range of phonetic realizations of a vowel available to them, they shift toward the realization displayed by an interlocutor. In a study of transcribed telephone conversations, Boulis and Ostendorf (2005) find that people modify their lexical choices depending on the gender of the person they are talking to.

Researchers in this area have been concerned with demonstrating that such shifts take place at all, and in examining the degree to which they may be socially mediated (see e.g. Gregory & Webster 1996). No previous work has examined how shifts are distributed across the range of lexical items produced by a speaker.

Morphology offers a domain in which generalizations sometimes compete, and multiple morphological variants can be available to speakers (e.g. *weaved* and *wove* as past-tense variants of *weave*) (Haber 1976, Fehringer 2004, Säily 2011, Thornton 2012). It is therefore feasible to investigate whether a speaker's morphological choices are influenced by the choices of interlocutors. Surprisingly, despite the considerable literature on convergence across multiple linguistic domains, convergence in morphology has been relatively understudied. Convergence in word selection has been the subject of active research (Boulis & Ostendorf 2005, Horton 2007, Horton & Brennan 2016, Ibarra & Tanenhaus 2016, Brandstetter et al. 2017), and, in general, word-formation processes are otherwise an area of considerable theoretical debate. Thus, convergence patterns in morphology may shed light on fundamental questions of representation and generalization.

Though relatively rare, a few exceptions in the literature exist, which indeed suggest that morphological convergence is amenable to research. Szmrecsanyi (2005, 2006) investigates English future marking, where an auxiliary verb (*will see*) competes with a lexicalized construction (*gonna see*). He finds that the choice between the two variants persists in discourse, even across speakers—the variant used in one instance is a significant predictor of which variant will be selected in the following instance. Beckner et al. (2016) study morphological imitation in an experimental setting, by asking participants to provide the past tenses of English verbs using a peer-pressure paradigm modeled on Asch 1951. Beckner et al. (2016) find that: (i) speakers can converge to a morphological pattern, generalizing observed behavior to new words; (ii) convergence is affected by the speakers' baseline behaviors, since speakers are less prone to produce variation on lexical items that do not otherwise vary in their own lexicon; and (iii) morphological convergence is a socially mediated process—speakers converge to humans, but not to humanoid robots.

2.2. MORPHOLOGICAL MODELS. Any explanation of morphological convergence effects is tied to a larger tradition that considers the extent to which the language system has access to the details of its input, and the degree to which these details influence said system and its outputs. This tradition goes back at least to Louis Hjelmslev (1961 [1953]) (advocating a radical abstractionist view) and Hermann Paul (1995 [1880]) (stressing the importance of individual words in language variation and change—see also Auer et al. 2015).

Broadly speaking, one set of theories, following Hjelmslev, is rule-based. It assumes the lexical system to be an inventory of words and a set of abstract rules. This approach has been modified and extended over time to accommodate variable and gradient effects. A key work of the rule-based approach is by Prasada and Pinker (1993), who posit a dual-route model of English past-tense formation. In this model, irregular behavior is determined by analogy based on similarity to existing forms. Regular behavior is driven by a single regular rule that has no phonological conditioning. Albright and Hayes (2003) propose a next-generation model in which a learning algorithm creates rules of variable scope over a lexical inventory. In their model, regular behavior is determined by a rule or rules that are sensitive to the phonological makeup of their target words.

The second view of the lexical system, following Paul, is analogy-based. It is more focused on variability in word behavior. Generalizations emerge from the inventory of words as consequences of the word frequencies and their similarities in form and function. Some instance-based models (Bybee 1995, Todd et al. 2019) implement this approach directly. A related group of connectionist models (Rumelhart & McClelland 1986, Plaut & Gonnerman 2000) advocates distributed representations for words, but

still shares the claim that generalization is based on experience with whole words and displays cumulative effects of similarity and frequency (Dell 2000).

To evaluate how this spectrum of theoretical approaches can capture our observed patterns of morphological convergence, we focus on three specific proposals—two based on rules and one based on analogy. Anchoring the analysis, we evaluate one prediction of a simple rule-based model (Prasada & Pinker 1993), namely that (a) convergence might occur by strengthening or weakening the context-independent regular affix rule (the default rule). This would predict that, whenever convergence in past-tense preferences occurs, this convergence would affect all past-tense forms to an equal degree. Any finding that convergence patterns are more word-specific would require a different explanation.

We also compare two proposals supporting whole-word updating of the lexical system: (b) the rule based-model of Albright and Hayes (2003) and (c) an analogy-based model in the tradition of Paul, which applies Nosofsky 1988 (Dawdy-Hesterberg & Pierrehumbert 2014). The predictions of these two models differ in complex ways, and we devote §6 to exploring them.

Models (b) and (c) both respond to experimental findings suggesting that all inflectional patterns arise from a stochastic architecture that generalizes over whole words (Bybee & Slobin 1982, Plunkett & Marchman 1993). It is widely agreed that clusters of related past-tense irregulars (such as *dive*/*dove*, *ride*/*rode*, *freeze*/*froze*) are productive to varying degrees, and that such lexical 'gangs' can attract new members based on similarity and type frequency (Bybee & Moder 1983, Stemberger & MacWhinney 1986, Hayes et al. 2009, Kapatsinski 2010, Cuskley et al. 2014). There is evidence that this picture also extends to regular patterns. Alegre & Gordon 1999a reports that regular past-tense forms are processed faster if they are high in token frequency, suggesting that these are stored in the lexical system. Albright & Hayes 2003 reports lexical gang effects for regular patterns.

In order to evaluate (b) and (c), we implement them to assess against our data. In doing so, we assume that all new words encountered are simply added to the lexicon; the lexical system is then updated according to the specific model. This treatment is clearly an idealization, which is not available in models that have no lexical inventory per se, such as Rumelhart & McClelland 1986 and Plaut & Gonnerman 2000. However, it is not as remote from such models as it might appear. In a seminal work, Marr and Poggio (1976) distinguish a number of levels of abstraction in cognitive models, two of which are relevant here. Computational models are defined on the most abstract level and specify overall patterns of outcomes. Algorithmic models in turn describe the processing mechanisms that give rise to these outcomes. Connectionist models tend to work at the algorithmic level, but can be considered at the computational level. Instance-based models tend to work on the computational level, but can be extended to work on the algorithmic level, as in Todd et al. 2019. Instance-based and connectionist models can even be equivalent on the algorithmic level (Ashby & Rosedahl 2017).

This study is carried out at the computational level in order to support comparisons between rule-based and analogy-based approaches. Further distinguishing between instance-based and connectionist approaches would also be valuable, but would require data and analysis that exceed the scope of the article.

2.3. Nonce words in the lexicon. If we assume whole-word updating of the lexical system, morphological convergence in the wild might involve either updating existing lexical items with new variants produced by an interlocutor or adding new lexical items to the inventory. In order to reduce the influence of individual existing lexical

forms, we use nonce forms. This enhances our ability to observe changes in the productivity of different morphological patterns. This follows a long tradition in morphology of using nonwords to test variability in morphological productivity (the WUG test; Berko 1958).

Does it really make sense to use nonwords to probe adjustments to the entire lexical system? Perhaps surprisingly, the literature shows that when individuals are exposed to nonce words, these can enter the lexicon and become nontrivially integrated. For example, seminal work by Gaskell and Dumay (2003) finds that if participants are exposed to nonce words in a learning task, these not only are rote-learned, but also participate in lexical priming and inhibition after what they call a period of LEXICAL INTEGRATION—a costly cognitive process that requires a period of sleep to be successful. A large body of subsequent research shows that degree and speed of integration varies depending on the context and the task, with at least some results showing immediate priming effects, without intervening sleep (Lindsay & Gaskell 2013, Coutanche & Thompson-Schill 2014, Kapnoula et al. 2015). In a highly relevant article, Lindsay et al. (2012) show that not only nonce words but also their inflected forms can take part in lexical priming. They show that nonce words can be interpreted and integrated as verbs by participants, and that the inflected regular past-tense forms of the nonce verbs show lexical priming, even if the participants never encountered those forms in training.

In short, nonce words join real words in inhibiting and priming word processing based on formal similarity, and are integrated into the lexical system with relative ease. This means that we can use them in an experimental task to probe the relationship between the lexicon and morphological preferences, and to study how this might predict patterns of morphological convergence.

3. RESEARCH QUESTIONS. This article explores three interrelated research questions:

(i) Does morphological convergence occur?
(ii) Is convergence sensitive to the overall distribution of lexical forms produced by the interlocutor?
(iii) Can rule-based and/or analogically based categorization models account for observed convergence patterns?

These questions are answered through an experiment using the English past tense.

In order to answer (i), we investigate whether individuals are affected by the overall regularization rate of a peer. If you are exposed to a peer whose rate of regularization is high, for example, does this lead you to increase your own regularization rate?

In order to answer (ii) we examine whether exposing individuals to different lexical distributions of regularization (while controlling for the overall rate) leads to different patterns of convergence. If the answer to (i) is 'yes', then this would provide evidence against a simple account based on the reweighting of a default affix (i.e. option (a) in §2.2), and in support of an account that relies on whole-word updating of the lexical system (options (b) or (c)).

The results outlined in §5 provide good evidence that the answer to (i) and (ii) is indeed 'yes'. This leads us to conclude that morphological convergence can occur, and that the underlying mechanism is lexically sensitive.

We then turn to research question (iii), which asks what kind of process might best account for our results. We assume that participants store the forms they are exposed to, and we attempt to model how these new forms would influence subsequent regularization preferences. We implement two candidate models—one analogical, and one rule-based—and test which model best predicts the overall pattern of results observed.

4. METHODS. We test our three research questions using a task, hosted on Amazon Mechanical Turk, in which participants play a word-matching game with a partner.[1] The task consists of three parts. In the PRETEST, the participant picks regular or irregular past-tense forms from nonce-verb prompts using a forced-choice button paradigm. In the ESP MATCHING TASK, the participant plays along with a 'peer' whose behavior is a controlled variation on the participant's pretest behavior. The behavior of this peer is the treatment in our experiment, as it differs across participants, and we expect participants to react to it. In the final part, the POSTTEST, the participant goes back to picking regular or irregular forms for prompt verbs without the presence of a peer. The term 'ESP' was proposed by von Ahn and Dabbish (2004), who originally developed this type of paradigm; we discuss this in more detail later in this section.

We expect participant behavior to change in the task, and that this change will be observable in the participant's posttest choices: how many verbs they regularize and, specifically, whether there is consistency in what these verbs look like. Analyzing the posttest data across the different types of peer behavior can address research questions (i) and (ii): whether convergence occurs and, if so, whether it is sensitive to the distribution of lexical forms. The use of categorization models to model participant behavior in the posttest allows us to test question (iii): whether rules or analogy (or both) describe participant behavior better.

In the following section (§4.1), we describe our stimuli. Then, in §4.2, we explain the experiment structure in more detail. Finally, in §4.3, we give a description of our participant sample.

4.1. STIMULI. We created the nonce-verb stimuli used in the ESP experiment based on the formal characteristics of existing, varying irregular verbs in English. The verbs are drawn from four different classes, representing English verb schemas that exhibit morphological variation. The classes are as follows:

- SANG: verbs that have a nasal coda and form the past tense by a vowel change from [ɪ] to [æ], such as *sing-sang*, *swim-swam* (e.g. *zim*, *gring*).
- BURNT: verbs that end in [ɛ]/[ɜ]/[ɪ] and a sonorant and that form the past tense by adding a [t], with no change in the vowel, such as *burn-burnt*, *learn-learnt* (e.g. *hurn*, *dwill*).
- KEPT: verbs that form the past tense by adding a final [t] and changing the stem vowel from [i] to [ɛ], as in *keep-kept*, *mean-meant* (e.g. *kreen*, *streel*).
- DROVE: verbs that form the past tense with a vowel change from [aɪ] or [i] to [oʊ], as in *drive-drove*, *weave-wove* (e.g. *strine*, *beeve*).

The classes were based on Bybee & Slobin 1982 and Moder 1992, with slight adjustments (see the online supplementary information at http://muse.jhu.edu/resolve/111, with additional details in Appendix A). To inventory the relevant forms, we used the CELEX lexical database (Baayen et al. 1993; based on the COBUILD corpus, Sinclair 1987). The aim was to capture the LEXICAL GANGS that show these specific irregular patterns in English. Lexical gangs can vary in their homogeneity, and different members of the same gang can resemble one another along different phonetic dimensions (see Bybee & Moder 1983, Stemberger & MacWhinney 1986, Alegre & Gordon 1999b). To give an example, the base form of *ride* shares a nucleus with *drive*, and *drive* shares a coda with *weave*. All three verbs can form the past tense with a vowel change

[1] All data and code are available at https://doi.org/10.5281/zenodo.4103379.

(to [oʊ]). They are members of the same lexical gang (along with *rise*, *freeze*, *speak*), connected through family resemblance. While such disjunctive classes increase the set of acceptable descriptions, their use has extensive motivation in studies of phonological data and diachronic processes (Mielke 2008).

This construction of the stimuli means that the experiment involves five different potential output types (four irregular types, plus the regular), grouped into two response categories, IRREGULAR and REGULAR.

A baseline experiment was used to establish quantitative rankings for nonce verbs used in the main ESP experiment. The baseline is a forced-choice task in which participants, completing the task alone, are presented with 316 nonce verbs (e.g. *spling*) and asked to choose a regular (*splinged*) or irregular (*splang*) past-tense form.[2] Stimuli are presented visually; participants see a prompt and have a choice between buttons showing the regular past form and the irregular form.

While forced-choice tasks may be different from open-choice tasks in important ways (cf. Treiman et al. 2015), they have been shown to be statistically sensitive and robust for well-formedness judgments (Sprouse & Almeida 2017). In the present study, they allowed us to analyze binary responses across several verb classes.

For the baseline experiment, we gathered data from 233 participants on Amazon Mechanical Turk. Overall, thirty-one participants were discarded, eleven for not being speakers of American English, and twenty for failing to meet attentiveness benchmarks during the experimental tasks. Responses from the remaining 202 participants provided us with a rich data set about the rate of regularization of our nonce verbs. These data were used to select three matched stimulus lists (randomized across participants for the three stages of the main ESP experiment). The resulting ranked lists are shown in Table 1, which shows the 156 verbs selected into each of three matched stimulus lists (see below for an explanation of the separation into three lists). The verbs are presented alongside their regularization rate from the baseline experiment, sorted such that the most-regularized verbs appear at the top.

Participants showed individual, structured variation in their preferences. Inspection of verb distributions across participants reveals a consistent hierarchy. That is, if a given individual regularized only ten verbs in the list, it would be very likely that these were among the top-ranked (most often regularized) verbs, and very unlikely that they would be among the bottom-ranked (least regularized) verbs. Additionally, more frequent regularizers extend regularization further down the ranked list than less frequent regularizers do. For example, the verb *fim* has a baseline mean of 0.772—it is regularized by 77% of the participants in the baseline task. The verb *spride* has a baseline regularization rate of 0.347. A participant who regularized *spride* was very likely to regularize *fim*, but a participant who regularized *fim* might or might not regularize *spride*.

In creating our stimulus lists, it was our intention to span a wide range of nonce verbs, showing a wide distribution of probability of regularization. We also wanted the same number of verbs to occur from each schema. Pilot work demonstrated that 156 verbs would be a reasonable size for a thirty-minute experiment, and thus we set out to cull the extra items from our baseline set. Each verb class (SANG, BURNT, KEPT, DROVE) was sorted according to verb baseline mean, and ties were removed at regularly spaced inter-

[2] Our baseline experiment included 256 verbs in the SANG, BURNT, KEPT, and DROVE classes (sixty, forty, eighty, and seventy-six items, respectively). We also created and tested sixty 'no change' verbs modeled after verbs like *cut*. However, this category showed much lower levels of variability, and was thus excluded from the main ESP experiment.

vals to yield thirty-nine nonce verbs per class. Trios of similarly scored verbs from the same category were grouped as a matching set, and each verb from that set was assigned (randomly) to list 1, list 2, or list 3. The items from each verb category (39/3 = 13 per list) were merged into a stimulus list, with 13 × 4 = 52 verbs, as shown in Table 1.

This process yielded three lists of fifty-two nonce verbs each, with each list showing a wide span of variation in regularization across native speakers. The assignment of lists to roles (pretest, ESP, or posttest) was balanced across different runs of the main ESP experiment.

	LIST 1			LIST 2			LIST 3		
RANK	VERB	BASELINE MEAN	CLASS	VERB	BASELINE MEAN	CLASS	VERB	BASELINE MEAN	CLASS
1	fim	0.772	s	strill	0.772	b	vrill	0.772	b
2	snurn	0.757	b	drurn	0.748	b	chim	0.743	s
3	cheem	0.733	k	cheen	0.743	k	skrill	0.733	b
4	skurn	0.733	b	zim	0.733	s	swurn	0.723	b
5	gim	0.728	s	rurn	0.731	b	jeem	0.723	k
6	murn	0.723	b	yill	0.718	b	jurn	0.718	b
7	trurn	0.718	b	pline	0.713	d	sneem	0.713	k
8	streen	0.718	k	hurn	0.713	b	gurn	0.708	b
9	prill	0.713	b	frim	0.708	s	schmine	0.703	d
10	surn	0.698	b	splurn	0.703	b	slurn	0.698	b
11	shreen	0.693	k	sneen	0.698	k	klill	0.693	b
12	slill	0.693	b	thurn	0.693	b	prurn	0.688	b
13	sprurn	0.688	b	geem	0.693	k	skeen	0.688	k
14	squine	0.678	d	trell	0.678	b	kreen	0.683	k
15	preem	0.678	k	feem	0.673	k	thrim	0.683	s
16	dwill	0.673	b	lell	0.668	b	glill	0.673	b
17	skrurn	0.668	b	smill	0.668	b	vurn	0.668	b
18	neen	0.663	k	zeem	0.663	k	neem	0.658	k
19	drell	0.653	b	jine	0.658	d	threll	0.653	b
20	snine	0.644	d	schmeem	0.649	k	kleem	0.653	k
21	spreem	0.639	k	stell	0.644	b	nink	0.647	s
22	zell	0.639	b	ming	0.639	s	blurn	0.639	b
23	schmim	0.634	s	dreen	0.634	k	squeen	0.629	k
24	greem	0.629	k	sprell	0.629	b	pite	0.629	d
25	vink	0.624	s	jink	0.614	s	strim	0.619	s
26	gline	0.619	d	trine	0.614	d	sline	0.614	d
27	skrine	0.599	d	geeve	0.599	d	glink	0.609	s
28	twink	0.594	s	gling	0.584	s	chite	0.579	d
29	skell	0.584	b	quink	0.569	s	smim	0.579	s
30	smink	0.579	s	greel	0.562	k	pleel	0.574	k
31	beeve	0.574	d	twell	0.559	b	brurn	0.554	b
32	kleel	0.554	k	strine	0.559	d	klite	0.554	d
33	sleel	0.55	k	cheel	0.545	k	breep	0.55	k
34	quing	0.53	s	twim	0.54	s	zite	0.54	d
35	snite	0.53	d	gring	0.515	s	sking	0.52	s
36	fleel	0.525	k	blide	0.515	d	squeep	0.52	k
37	vrink	0.515	s	streel	0.51	k	dwim	0.515	s
38	klide	0.495	d	splink	0.49	s	squite	0.505	d
39	dwing	0.485	s	slive	0.49	d	theel	0.495	k
40	fide	0.485	d	sweel	0.475	k	trink	0.49	s
41	skeep	0.485	k	quide	0.47	d	splide	0.475	d
42	grink	0.47	s	shring	0.46	s	pring	0.465	s
43	swite	0.46	d	sning	0.455	s	shreep	0.46	k
44	skrink	0.45	s	schmite	0.45	d	shing	0.45	s
45	dreep	0.441	k	zeep	0.441	k	twite	0.446	d

(TABLE 1. *Continues*)

	LIST 1			LIST 2			LIST 3		
RANK	VERB	BASELINE MEAN	CLASS	VERB	BASELINE MEAN	CLASS	VERB	BASELINE MEAN	CLASS
46	thide	0.436	d	dwink	0.431	s	strite	0.426	d
47	dweep	0.426	k	sneep	0.431	k	snink	0.421	s
48	strink	0.416	s	splive	0.421	d	theep	0.416	k
49	dwide	0.416	d	thring	0.411	s	squide	0.406	d
50	spling	0.401	s	yide	0.401	d	sping	0.401	s
51	shride	0.376	d	thride	0.376	d	brive	0.391	d
52	spride	0.347	d	vrite	0.366	d	swide	0.356	d

TABLE 1. Three lists of stimulus verbs, ordered according to regularization rate in the baseline experiment. All verbs are associated with one of four classes with respect to the irregular form presented: BURNT (b), KEPT (k), SANG (s), or DROVE (d).

4.2. THE DESIGN OF THE MAIN ESP EXPERIMENT. Our baseline experiment was used to estimate the rate of regularization for our nonce-word stimuli. Our ESP experiment, run subsequently and with different participants, was used to test our hypotheses on morphological convergence.

The participants completed the ESP experiment on their computer. The experiment uses orthographic stimuli and consists of three phases (see Figure 1). All three lists in Table 1 are used in the experiment, with one list per phase. The allocation of lists across phases is randomized per participant, and verbs from within each list are presented in random order during each phase.

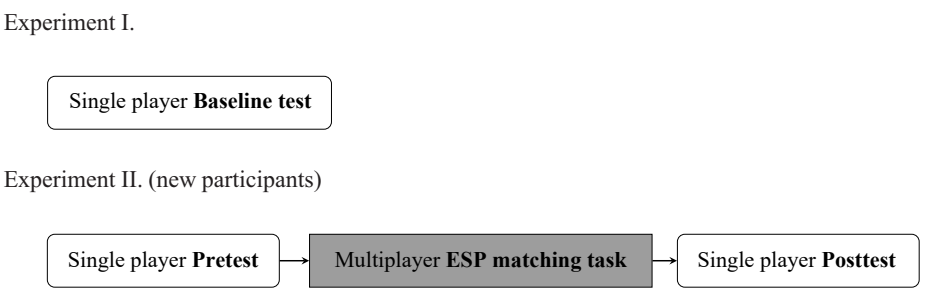

FIGURE 1. Structure of the ESP experiment. In the ESP matching task, participants play with a peer with one of nine possible behaviors, as outlined in Table 2.

The pretest is a forced-choice task. The player is presented with English nonce verbs (like *spling*) and has to pick either the regular past-tense or the irregular past-tense form (*splinged*/*splang*). The player responds to fifty-two targets and receives no feedback.

This is followed by the ESP matching phase. This is similar to the pretest, except that there are two players: the participant and a bot peer. This phase uses a simple interactive matching game based on the ESP paradigm (von Ahn & Dabbish 2004). ESP was originally developed as a crowdsourcing technique for image labeling. Our ESP task asks players to attempt to predict their co-player's responses over fifty-two trials with English nonce verbs. The ESP task thereby offers a controlled platform for manipulating morphological exposure.

Both the participant and the peer pick a past-tense form, with the additional instruction that the goal is to guess the other player's answer in advance. While the participant is making their choice, the bot peer is 'thinking'—its pick is revealed only after the player makes their selection. If the participant and the peer responses match, the participant is awarded a point. There are no deductions for mismatch.

The ESP task is forced-choice, which means that in every case, the participant sees both the regular and irregular forms on the screen. Influence during the ESP test there-

fore does not arise merely from word presentation, but is dependent on the effects of prediction and selection.

We do not explicitly tell the participant that the other player is a bot or a human. Playing against the computer in video games is extremely common and natural.

The last part of the task is the posttest. As in the pretest, the player has to choose the regular or irregular past-tense forms for fifty-two English nonce verbs, with no peer, and with no feedback.

		REGULARIZATION SHIFT		
		−40%	NO CHANGE	+40%
LEXICAL TYPICALITY	TYPICAL	$n - 0.4n$ highest ranked	n highest ranked	$n + 0.4n$ highest ranked
	RANDOM	$n - 0.4n$ random	n random	$n + 0.4n$ random
	REVERSED	$n - 0.4n$ lowest ranked	n lowest ranked	$n + 0.4n$ lowest ranked

TABLE 2. Across-participant conditions in the ESP matching task.

As summarized in Table 2, the experimental manipulation occurs in the ESP matching phase, which has two across-participant factors in a 3 × 3 crossover design. The first one is the bot peer's REGULARIZATION SHIFT in the ESP test. The peer's behavior is based on the human player's behavior in the pretest. Under different across-participants conditions, bot peers will regularize (i) the same percentage of forms as the human player in the pretest, (ii) 40% more, or (iii) 40% fewer. Given a human player's regular response count of k, bot regularization is $k = n$, $k = n + 0.4n$, or $k = n - 0.4n$. For example, if the player regularized ten verbs in the pretest, then in the ESP test, the bot peer will regularize (i) ten, (ii) fourteen, or (iii) six verbs. This dynamic element of the design presents the complication that, for some participants, the peer cannot increase or decrease regularization by 40%, due to ceiling or floor effects. We explain our solution to this problem below in §4.3.

The second factor is the bot peer's LEXICAL TYPICALITY. The peer may regularize the forms that human players are (i) most likely to regularize (TYPICAL PEER), (ii) least likely to regularize (REVERSED PEER), or (iii) random forms (RANDOM PEER). We could not customize peer lexical preferences for individual players because the sample of the pretest was small and intentionally wide. This did not allow us to generalize individual verb regularization preferences to a new list. Instead, peer lexical preferences were modeled on the TYPICAL human player, created using regularization rankings from the baseline data (as shown in Table 1). If peer behavior is typical, the list is sorted according to regularization ranking, and the peer regularizes the k forms at the list head—those that baseline participants were most likely to regularize overall. If the peer is reversed, it regularizes the k forms at the list tail—forms baseline participants were least likely to regularize. Finally, if it is random, the peer selects k verbs at random.

Consider the example in which a participant is playing with a peer who produces regular forms at a rate of 60%. If they are in a TYPICAL CONDITION, the peer during the ESP test will regularize the top 60% of forms from the relevant list (verbs 1–31). If they are in a REVERSED CONDITION, the peer will regularize the bottom 60% of forms from the list (verbs 21–52). And in a RANDOM CONDITION, the peer will regularize 60% of forms, choosing at random from across the list. Thus, it is only in the typical condition that participants are exposed to a distribution resembling typical human player behavior.

The Regularization shift factor allows us to test for the presence of morphological convergence in the ESP experiment, addressing research question (i). An observed shift in the posttest is likely due to the effect of the bot peer's behavior in the ESP matching task. The Lexical typicality factor allows us to look at research question (ii): whether

there is an effect of lexical distributions on morphological convergence. That is, this manipulation tests whether convergence goes beyond shifting rates of use (e.g. higher/lower regularization) and is sensitive to the specific morphological forms encountered.

The aim of this 'faux' ESP design is to create an interaction in which experimenters can carefully control the parameters. It may be compared to the MAP TASK, an established experimental method of studying linguistic convergence (Bard et al. 1989, Anderson et al. 1991, Pardo 2006). In the map task, two participants navigate a map together. The map, or maps, is designed to include specific objects, placed so that participants are steered toward a specific set of lexical choices. This ability to influence language use while allowing for relatively natural conversation is the main advantage of the map task over other methods.

While the ESP design is more rigid—conversation is replaced by clicking on online buttons, and the set of lexical choices is severely restricted—it allows for absolute control over the lexical choices presented to the participant and the variants the participant is exposed to. These are necessary in an experiment examining fine-grained morphophonological variation across a distribution of related forms. While this comes at the cost of naturalistic interaction, our results clearly show that participants rely on linguistic knowledge in making their decisions. Participants' pretest and posttest behavior are both strongly correlated with (i) predictions of categorization models trained on the English lexical inventory (see §6) and (ii) participant behavior in the simple baseline experiment (see below). Participants in the ESP experiment respond to treatments based on existing patterns of similarity in the English lexicon (see §5).

We also note that the ESP game instructions explicitly ask participants in the interactive round to copy the response patterns of their co-player. Given this, the interspeaker mechanisms at work here are not identical to those explored in studies of spontaneous convergence in conversation (e.g. Pardo 2006). However, any subsequent shifts in behavior during the posttest round will nevertheless be of interest; in the absence of ongoing feedback, shifts in preferences cannot be directly attributed to explicit instruction, but would represent persistent influence of the interaction. This dynamic parallels studies of real-world interactions which find that, in language acquisition, positive reinforcement (smiling, proximity) can prompt linguistic changes that persist beyond the interaction (Goldstein et al. 2003). Moreover, in second-language contexts, seeking of extrinsic rewards (in addition to other factors) can increase motivation and facilitate long-term language learning (Gardner & MacIntyre 1991).

4.3. PARTICIPANTS. For the main experiment, we collected data from a total of 331 participants on Amazon Mechanical Turk (AMT). We chose to collect data using AMT because it is known to provide large and reliable data sets (Snow et al. 2008), which are generally highly correlated with data from laboratory experiments, albeit more variable (Balota et al. 2001, Wurm et al. 2011, Crump et al. 2013). One possible reason for this variability is that, while the accessible AMT pool is not necessarily much larger than the subject pool at a large university, it is somewhat more diverse and closer to the general adult population (Ipeirotis 2010, Stewart et al. 2015). This diversity may address the limitations of data coming from university subject pools (Henrich et al. 2010).

All participants had to be native speakers of English from the United States, eighteen years or older. Participants were paid $3 for their participation. Initial analysis of participants required several individuals to be removed, as follows: three participants failed to complete the task; five players showed suspicious tendencies to linger on the same button for many trials in a row on the posttest, indicating inattentiveness; four par-

ticipants were discarded for learning English outside of the US. Removal of these individuals leaves a pool of 319 participants.

Participants regularize verbs to varying degrees in the pretest. We removed the left and right tail of the distribution of average participant regularization to keep the effect of the co-player consistent across participants, relative to the pretest. In the INCREASE CONDITION the peer regularizes 40% more verbs in the ESP task than the participant did in the pretest, and 40% fewer verbs in the DECREASE CONDITION. To ensure that every participant experiences a proportional increase/decrease, participants whose regularization rate was too high/low were removed in order to avoid ceiling/floor effects. For example, suppose participant A regularizes ten verbs in the pretest, participant B regularizes twenty verbs, and participant C regularizes fifty verbs. In the ESP increase condition, participant A sees fourteen regular verbs, participant B sees twenty-eight regular verbs, and participant C sees fifty-two regular verbs (the total number of verbs in the set). While A and B see a proportional increase of 40%, C does not experience the same change in rate of regularization. Consequently, in order to interpret our results consistently, participant C is removed from the final analysis. We removed over- and underregularizers in all conditions. Since participants, on average, are more likely to regularize verbs in the pretest (mean rate of regularization = 0.59), the distribution is skewed, and we needed to remove more overregularizers than underregularizers. In the analyses that follow, across all conditions, we include only participants with a participant pretest mean in the range between 0.06 and 0.70. These values represent thresholds at which the peer is at risk of encountering the floor/ceiling, due to the regularization shift manipulation (which applies multiplicatively, rather than additively). This filtering step removes two participants who regularize less than 6% in the pretest, and ninety-five participants who regularize more than 70% in the pretest,[3] resulting in a final pool of 222 participants.

Table 3 lists the number of participants per subcondition, after our data set has been filtered as described above. The mean age of participants is 32.93 (*SD* = 10.22); 107 participants are women, and 115 are men. In comparison, 117 women and 84 men took part in the baseline task, with a mean age of 34.04 (*SD* = 10.12). The mean duration of the experiment was 11.66 minutes (*SD* = 3.82). The trimmed data set has no overall bias toward regularization (the overall mean pretest regularization is 0.49).

	−40%	NO CHANGE	+40%
TYPICAL	20	22	29
RANDOM	19	22	37
REVERSED	20	23	30

TABLE 3. Participants per 3 × 3 subcondition.

5. RESULTS OF THE EXPERIMENT. As a precursor to analyzing the posttest responses, we first verify the balancing of pretest scores with respect to the experimental conditions. To do this, we fit a logistic mixed-effects regression model on the pretest data to see if there is any significant difference between participant pretest responses depending on across-participant conditions (including relevant interactions) to make sure that the prior states of the participants are balanced across the design. We find no significant

[3] One referee raises the possibility that higher posttest regularization rates could result from regression to the mean, due to the exclusion of participants based on their pretest preferences (particularly affecting the highest regularizers). However, the filtering process cannot, on its own, lead to spurious regularization effects, because the experiment conditions do not merely increase or decrease regularization. Participants in the +40 and −40 conditions are analyzed in comparison to participants in the 'no change' condition, and (as the referee further notes) the no-change condition cannot be affected by regression to the mean.

patterns, suggesting that participants do not differ significantly as a function of condition in the pretest—individuals, on average, regularize to different degrees, but the pretest distributions are the same for all conditions. Since participant means are distributed equally in the pretest, any differences in posttest distributions must be attributed to the effects of the ESP test.

One alternate approach would be to analyze pretest and posttest data in a single regression model; this approach would allow us to omit the participant's pretest regularization (see below) from among the predictors. However, such an approach would necessitate starting models with a four-way interaction, accompanied by complexities in visualization and difficulties in model convergence.

We fit a logistic mixed-effects regression model on the posttest data, with participant response (REGULAR VS. IRREGULAR) as the outcome variable. The predictor variables were VERB BASELINE MEAN (representing the average regularization for each nonce verb in our baseline study; see §4.1), PARTICIPANT PRETEST MEAN (a participant's average regularization rate in the pretest), LEXICAL TYPICALITY OF THE ESP PEER (typical, reversed, or random), and REGULARIZATION SHIFT of the ESP peer (+40, −40, or no-change). These factors are summarized for reference in Table 4. Additionally, the model includes random intercepts for participants and items, and a random slope for Verb baseline mean by participant. We build the model in a stepwise fashion, starting with all three-way interactions for these variables and removing interactions that are not significant. The model summary appears in Table 5.

VERB BASELINE MEAN	The mean rate of regularization for a given nonce verb in our baseline experiment
PARTICIPANT PRETEST MEAN	The mean rate of regularization for a given participant in the pretest of our main ESP experiment
LEXICAL TYPICALITY	An aspect of bot peer behavior: the typical peer regularizes those verbs that have higher verb baseline means, the reversed peer regularizes those that have lower means, and the random peer chooses verbs at random.
REGULARIZATION SHIFT	An aspect of bot peer behavior: the +40% peer regularizes 40% more verbs than the participant did in the pretest, the −40% peer regularizes 40% fewer verbs, and the no-change peer regularizes the same amount.

TABLE 4. Our terminology.

	COEFF	STD. ERROR	*z*-SCORE	SIG
(intercept)	−0.11	0.16	−0.67	
Verb baseline mean	8.26	0.69	12.04	***
Lexical typicality = random	0.04	0.17	0.25	
Lexical typicality = reversed	0.04	0.17	0.20	
Participant pretest mean	7.13	0.56	12.67	***
Regularization shift = −40%	−0.46	0.18	−2.47	**
Regularization shift = +40%	0.55	0.17	3.31	***
Verb baseline mean : Lexical typicality = random	−2.40	0.93	−2.59	**
Verb baseline mean : Lexical typicality = reversed	−2.13	0.96	−2.23	*

TABLE 5. Experimental data, posttest regression model summary. Model formula: Posttest regular response ~ Verb baseline mean × Lexical typicality + Participant pretest mean + Regularization shift + (1 + verb.baseline.mean | participant) + (1 | verb).

Note that the continuous predictors—Participant pretest mean and Verb baseline mean—are mean-centered in this model to counteract collinearity between main and interaction terms. Additional details about the model-selection procedure can be found in the supplementary information.

The average propensity of each participant to regularize verbs (as assessed by our pretest) and the average propensity for each verb to be regularized (as assessed by our

baseline study) are the strongest predictors of posttest behavior. This means that participants who are inclined to regularize in the pretest are also inclined to do so in the posttest. Verbs that are good candidates for (ir)regularization, as determined by the baseline test, tend to also be (ir)regularized in the posttest. Thus, the ESP task does not prompt participants to completely abandon their prior biases in posttest responses.

However, the experimental conditions in the ESP task (Regularization shift and Lexical typicality) do have the expected effects in the posttest. The peer's overall rate of regularization influences participant behaviors in the expected directions, as shown by the highly significant variable Regularization shift. An overall increase in peer regularization in the ESP test prompts an increase in participant regularization in the posttest. Likewise, a decrease in the ESP task prompts a decrease in the posttest. This main effect is illustrated in Figure 2.

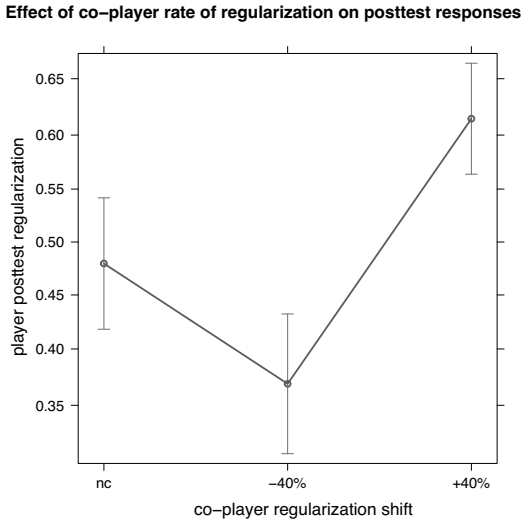

FIGURE 2. Effect of Regularization shift: model predictions of how participant responses in the posttest are affected by the regularization shift of the peer during the ESP task. Error bars represent the 95% confidence interval of model predictions for each regularization shift condition.

The results show that participant responses to specific verbs are also influenced by the peer behavior seen in the ESP test. Recall that a peer that exhibits typical behavior makes selections that respect baseline verb rankings—that is, the peer selects the *k* best verbs to regularize on the basis of verb baseline mean. A reversed peer reverses these rankings (by selecting the *k* worst verbs), and a random peer selects *k* verbs at random. In other words, typical 'peers' show a preference for regularizing forms with high baseline means. Reversed peers show a preference for regularizing forms with low verb baseline means.

As previously noted, participants' preferences for individual verbs are well predicted by the verb's rate of regularization in the baseline. However, the Verb baseline mean is a weaker predictor for participants exposed to a reversed or random peer, compared to the typical condition. Participants who played with the reversed or random peers are more likely to regularize low baseline forms and less likely to regularize high baseline forms, compared with participants who played with typical peers.

Reordering the factors indicates that there is no significant difference between the interaction of Verb baseline mean with random peers versus reversed peers. The relevant

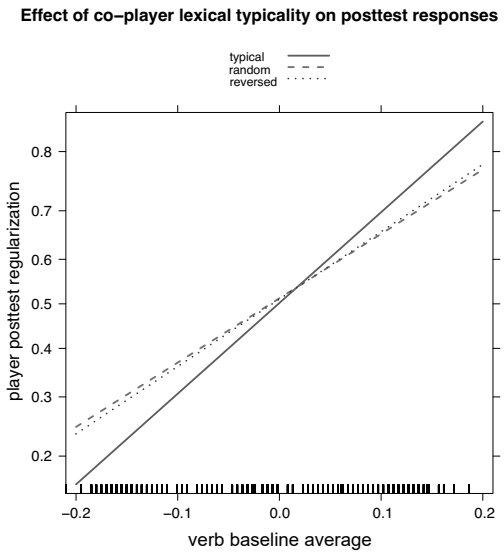

FIGURE 3. Effect plot of Verb baseline mean × peer Lexical typicality: model predictions of how posttest responses to verbs with different baseline means are affected by the lexical typicality of the peer choices.

interaction plot appears in Figure 3, showing that the adherence to Verb baseline mean biases is somewhat flattened out among participants who were paired with a reversed or random peer. In summary, ESP interaction with a typical peer reinforces baseline tendencies, whereas interaction with a less typical peer counteracts these tendencies.

Note that if participants were influenced only by the specific choices of the peer, we might expect the reversed line in Fig. 3 to slope in the opposite direction, with low baseline forms having high regularization. For example, if the peer selected the irregular form *snurnt* for *snurn* during the ESP task, this might cause the participant to select *drurnt* for *drurn* (an item with the same low rank) in the posttest. If the peer selected *sprided* for *spride*, the participant might select *swided* for *swide* (an item with the same high rank) in the posttest. The fact that this does not happen shows that participants' generalization performance in the posttest is influenced by general or prior knowledge as well as by the peer choices (similar to e.g. phonetic convergence; see Pardo 2006).

One might also expect the slope of the response to the random peer to sit between those for the typical and reversed peers, due to the random peer making some typical choices and some reversed choices. However, this expectation is not borne out; in our experiment, exposure to the random peer leads to the same type of distribution as exposure to the reversed peer.

In the discussion we speculate that this may relate to different words having different degrees of influence in the experiment. What remains clear is that, based on the above analysis, verb distributions do have an overall effect on participant behavior.

Additional questions may be considered regarding the roles of particular verb classes (SANG, BURNT, KEPT, DROVE) in the results described above. Verb classes are clearly relevant to participant behavior; for instance, BURNT verbs are more likely to be regularized in baseline than DROVE verbs (as can be seen from surveying Table 1). Thus we performed follow-up mixed-effects modeling to determine whether the results are driven by specific verb classes. These subsequent investigations show that there is no significant interaction between verb class and regularization shift, and no significant interaction for lexical.typicality * baseline * verb class. Thus, the postinteraction influ-

ence from regularization shift and peer item preferences are not isolated to just one or two verb classes. Of course, verb classes are still integral to participants' morphological generalizations in the experiment. We incorporate information about English verb classes into the learning models in the next section.

In sum, participant behavior in the posttest is affected by peer behavior in the ESP test. We set out to test for two possible effects of morphological convergence. First, we wanted to test whether an overall preference for regularization would be influenced by peer lexical choices. The significant effect of the regularization level (Regularization shift) shows that it is. Second, we wanted to test whether convergence would be sensitive to the overall distribution of forms produced by the ESP peer. The significant interaction between Verb baseline mean and Peer typicality shows that differing peer behavior influenced participants to deviate from baseline tendencies in different ways.

We have therefore demonstrated an effect of morphological convergence, which is highly influenced by the distribution of forms to which the individual is exposed. In sum, the answer to both research questions (i) and (ii) is YES. This supports a model of morphological convergence in which the lexical system is constantly updated on the basis of new experiences, and generalizations are formed over whole words in a rapid, on-line manner. As discussed in §2, lexical updating (when considered at the processing level) does not necessarily involve the rapid incorporation of novel word forms into the lexical inventory. However, we retain this idealization in order to go on and investigate the level of detail available for the convergence mechanism.

We now turn to research question (iii), examining the degree to which participant behavior can be modeled by two different categorization models, under the assumption of rapid lexical updating.

6. RULE-BASED AND ANALOGICAL CATEGORIZATION MODELS. We have now established that morphological convergence occurs, and that it does not arise from a straightforward reweighting of a default regular affix. Participants do not simply increase or decrease their rate of use of the regular past-tense affix. Rather, the nature of the particular lexical items that a participant is exposed to has an important effect on their tendency to be (ir)regularized, supporting accounts that involve lexical updating.

In order to attempt to pinpoint more precisely what the consequences of lexical updating are for subsequent past-tense regularization patterns, we now consider the process through which morphological generalizations arise from the lexicon. In this section, we explore our research question (iii): Can rule-based and/or analogy-based models account for the observed convergence patterns? This question takes us beyond the issue of morphological convergence per se, requiring us to address more fundamental questions about the relationship between the lexicon and morphological generalizations.

We focus on two stochastic models explored in detail by Albright and Hayes (2003)—the GENERALIZED CONTEXT MODEL (GCM; Nosofsky 1990) and the MINIMAL GENERALIZATION LEARNER (MGL; Albright & Hayes 2003). Both models have been previously used to explore patterns of long-term learning and generalization over an existing lexicon. The former relies on processes of item-based analogy, while the latter infers abstract morphological rules from the patterns in the lexicon. Crucially, these models both differ from earlier dual-mechanism accounts (Prasada & Pinker 1993, Clahsen 1999) in which regular forms are handled, in an all-or-nothing manner, by a regular rule. In contrast, these stochastic models are consistent with evidence that participant ratings of both regular and irregular forms are affected gradiently by phonological similarity to the base forms of other regular/irregular verbs in the lexicon. The two models treat regular forms differently as a result of the difference in their architectures. The MGL assumes broad

generalizations based on a large number of similar regular verbs in the lexical inventory, while the GCM focuses on narrow lexical gangs of irregulars defined by similarity and takes account of the degree to which regular verbs resemble each gang.

We provide more background on the GCM and MGL in §§6.1 and 6.2, respectively. In our analysis, we first fit these models to our baseline data (§6.3) to account for how well each of the models contributes to participants' preferences for the regular form, across different verb types. This is not unlike the types of data sets and long-term learning problems for which these models have been used in the past. We then explore, in §6.4, what predictions the models make for the posttest responses, under the assumption of lexical updating. If we assume that participants placed the nonce forms we exposed them to in their lexicons, and then updated their analogical generalizations (GCM) or abstract rules (MGL), respectively, what predictions would these models make about subsequent regularization patterns? Can changes in the model behavior following lexical updating account for the changes in participant preferences that we observe?

6.1. GENERALIZED CONTEXT MODEL. The generalized context model (Nosofsky 1988, 1990) is an instance-based analogical model. Instance-based models assume that people complete analogies using richly detailed categories composed of instances. For the GCM, the process of selecting a response category entails the comparison of a new instance to previously encountered ones. The number and type of instances in memory codetermine the outcome of categorization. The GCM is a highly successful instance-based analogical model of human categorization (McKinley & Nosofsky 1996, Maddox & Ashby 1998). It has been successfully adapted to explain the ways humans incorporate new versus old information in processing (Donkin & Nosofsky 2012, Nosofsky et al. 2014) and has been widely used in linguistic modeling (see e.g. Krott et al. 2001, Nakisa et al. 2001, Albright & Hayes 2003, Dawdy-Hesterberg & Pierrehumbert 2014).

To assign category membership to a novel instance, the GCM first calculates its similarity to instances in preexisting categories in a given training set. In morphophonology, the target instances are words. In our example, they are base forms assigned to past-tense categories (REGULAR/IRREGULAR; see Appendix A). (Our implementation considers these categories in the verb's morphological class; see below.) Calculating the overall similarity between two words depends on aligning their segments and then finding how similar the corresponding segments are. For example, the calculated similarity of *splive* to *strive* reflects the fact that the first segment and the last two segments are identical ([s], [aɪv]), while the second and third segments ([pl] vs. [tɹ]) are similar. The GCM selects as the output for a novel instance the category with the most members that are the most similar (Nosofsky 1990). For *splive*, the support for the regular outcome *splived* rests on the comparisons to existing regular verbs (e.g. *hived*, *signed*), and the support for the irregular outcome *splove* rests on the comparisons to existing irregular verbs (e.g. *strode*, *strove*, *smote*). The overall regularity score is based on which set of verbs offers more total support.

The particular implementation of the GCM we use is based on Dawdy-Hesterberg & Pierrehumbert 2014. The training inventory is based on the CELEX corpus (Baayen et al. 1993). The model's irregular class of comparison consists of irregular English verbs that display the target irregular alternation. The regular class of comparison consists of phonologically similar regular verbs as well as MISCELLANEOUS regular verbs outside of our schemata. Segmental similarity is calculated using the method developed by Frisch et al. (2004). Since the range of responses in our tasks is 0–1, we standardize GCM predictions to match this range. Further mathematical details are provided in Appendix A.

In our GCM baseline model, *splive* has a regularity score of 0.57—the GCM regards it as a relatively regular verb. Compare this to its actual rate of regularization in our baseline experiment: 0.42.

6.2. MINIMAL GENERALIZATION LEARNER. The minimal generalization learner (Albright & Hayes 2002, 2003) uses more abstract generalizations (referred to as RULES in the model's terminology), rather than the richly detailed instances used by the GCM. These rules are based on sets of forms in the training data that show similar behavior. The MGL builds on the work of Albright and Hayes, whose approach to bootstrapping morphological rules has been widely used in the statistical natural language processing literature. The MGL has been shown to be highly accurate in modeling the behavior of nonce forms in the English past tense. Albright and Hayes argue that the MGL outperforms the GCM in predicting participant behavior in a nonce-verb production task they conducted.

The MGL iterates over pairs of words in the lexical inventory, hypothesizing generalizations conservatively on the basis of any phonological features that are shared across the words. It then iterates over rules, attempting to collapse rules into a more general rule when possible. A rule is scored according to how many words it applies to in the inventory, weighted against cases in which the inferred phonological context is present but the rule fails to apply. The resulting system consists of a catalog of weighted natural class-based generalizations that compete with one another, and that are more or less likely to apply in various phonological contexts (for regular as well as irregular verbs).

The MGL is implemented here from materials made available by Albright and Hayes (2003), along with the segmental similarity metric of Frisch et al. (2004). The particular details of our implementation of the MGL are provided in Appendix B.

Let us take our previous example, the nonce verb *splive* in the DROVE class. The MGL recognizes a number of rules that could have this verb as their input. The two that are relevant are the ones that create the two choices in the task, regular *splived* and irregular *splove*. Following Albright and Hayes (2003), we can calculate the MGL regularity score of this verb by dividing the adjusted confidence of the regular rule by the summed adjusted confidence of both rules (for details, see Appendix B). When trained on CELEX, the MGL regularity score of *splive* is 0.73.

In our analysis, the training set for both the MGL and the GCM is the English lexical inventory—the set of existing English verbs. The GCM uses individual verbs to categorize new verbs, while the MGL creates overlapping rules for sets of verbs and assigns a relative strength to each rule. Both models impose restrictions on the training set. As a consequence, neither model uses all available verbs in this inventory for inference. The number of forms (with a token frequency of 10 or above) in our CELEX dictionary is 3,156. Across the four verb classes and the miscellaneous regular verbs, the GCM makes decisions on the basis of 1,427 forms. The MGL creates generalizations over and makes decisions on the basis of 401 forms.

The difference arises because the GCM operates on a large regular set by default and uses distance weighting to account for the attraction of a small set of highly similar irregulars. In contrast, the MGL builds rules from the bottom up and stops when these rules provide optimal coverage under starting parameters. (We discuss this in Appendices A–B and the supplementary information.)

6.3. MODELING PAST-TENSE PREFERENCES: MODEL FITS ON THE BASELINE DATA. In this section we assess the performance of the two models on our baseline data. This provides some insight into how well they can capture past-tense regularization preferences in the absence of peer influence.

To generate the model predictions, we assume that each individual's starting point is the same: complete familiarity with the same set of regular and irregular verbs listed in the CELEX corpus (Baayen et al. 1993). While this is a simplification, it is a reasonable approximation. English irregular verbs constitute a relatively small set and are all relatively frequent (Cuskley et al. 2014), so a native speaker is likely to know them all. The model also makes the assumption that all participants behave in exactly the same way, based on the verb types they (all) know. This is clearly not true, as participants vary considerably in their rate of regularization, both in the baseline task and in the pretest of the ESP task. This individual variation very likely comes, in part, from socially motivated preferences between individuals (e.g. variation by gender, age, and social class), which we have not considered in this study. It may also be due to broader cognitive factors: individuals vary widely in various learning tasks (Siegelman & Frost 2015), and individual behavior in a given task can be hugely affected by the individual's cognitive style (Lleras & Von Mühlenen 2004). Substantial across-participant variability has been found in other language-learning tasks as well (Schumacher et al. 2014, Rácz et al. 2017).

We abstract away from these factors and work with idealized participants in order to keep the number of parameters reasonable in our model. We train the models on sets of real verbs, derived from CELEX, and fit them on our 256 nonce verbs, the relevant forms in the baseline task. We then compare the models' regularization scores with participants' choices to regularize/irregularize the verbs in the baseline experiment, and use these data to calculate concordance indices (Harrell 2013). A concordance index, C, measures the probability of agreement between a gradient predictor and a binary outcome; that is, comparing over all pairs of items in the data, C represents the proportion of trials in which the continuous predictor correctly predicts the categorical response. In this case, the continuous predictor is the regularization score assigned to each nonce verb by the model of interest (GCM or MGL), and the binary outcome is the participant's choice of the regular or irregular verb form. The concordance index C of the MGL scores and participant responses is 0.59, and 0.58 for the GCM. On the face of it, the MGL slightly outperforms the GCM, but the accuracy of the two models is similar.

A separate question is whether the MGL and the GCM explain variation in our data INDEPENDENTLY (that is, whether either or both of them effectively predict participant behavior). To test this, we conducted mixed-effects logistic regression models of participants' choices in the baseline task and used a standard residualization procedure to test whether the MGL and the GCM have separate explanatory power. A residual is the amount of variation left over after we take a predictor's effect into consideration (see e.g. Gelman & Hill 2006). Using residualization, we can test whether the GCM accounts for variation not accounted for by the MGL and vice versa. We find that the MGL and the GCM have independent predictive power; they both explain different parts of how participants regularize verbs in the baseline experiment. This means that the MGL and the GCM contribute independently toward explaining variation in the baseline data. This, in turn, suggests that both instance-based analogy and more abstract generalization play a role in shaping English past-tense variation. (We discuss this analysis in detail in the supplementary information.)

6.4. MODELING CONVERGENCE: MODEL FITS ON THE POSTTEST DATA OF THE ESP EXPERIMENT. In this section we fit the categorization models on data from the ESP experiment. Rather than using the experimental manipulations to predict participant behavior directly (as we did in §5), we ask whether it is also possible to use shifts in the predictions of the GCM and/or the MGL to capture participants' changing behavior. Our analysis has shown that participant behavior changes on a lexical level due to exposure to the

ESP peer. If the basis of this is lexical updating, this should be reflected in the way the categorization models behave when exposed to the same stimuli as the participants.

In the baseline analysis described in the previous section, the models are trained on existing English verbs. Each model is fitted once, on the IDEAL participant responding in the baseline task. In the posttest analysis, the models are trained on existing English verbs and the responses of the bot peer in the ESP test—these verbs constitute the novel information to which the participant converges. We assume that individuals update their lexical inventories with the nonce words presented to them. We also assume that during the ESP test, the specific words seen by the player are added to the regular or irregular categories, depending on the peer's exposure. We thus add the specific words and categorizations the player encounters to a player-specific inventory alongside the real English verbs from CELEX. As every participant receives a unique set of stimuli, we are able to generate unique predicted regularization preferences for each participant, based on the MGL and GCM models generated from these updated inventories. We are able to compare two sets of predictions generated by the two models for the posttest forms.

The updated part of the inventory is different for every player, because word lists and verb forms vary across all players in all conditions. The contents of this new part of the inventory are affected by both the Regularization shift factor (whether the player is in the +40, −40, or no-change condition) and the Lexical typicality factor—that is, whether the distribution of the peer's regular forms is chosen to be TYPICAL, REVERSED, or RANDOM.

We refer to the models that are trained on existing English verbs from CELEX, and are shared for all participants, as the CELEX-MGL and CELEX-GCM models. We refer to the models that are trained on individualized post-ESP-test inventories as INDIVIDUAL-MGL and INDIVIDUAL-GCM models. These are unique to each participant.

This approach makes a number of simplifications. First, as noted above, our models proceed as if all participants started the experiment with the same baseline lexical inventory. While we know that variation will exist across individual lexicons, for this simulation we use corpus-derived data to represent the initial state of all speakers. Second, it is unlikely that the words encountered during the ESP test are stored in the lexical inventory right away with the same effect on category structure as existing verbs of the language. It remains true, however, that new words can be integrated into the lexical system relatively quickly (see §2). Third, learning is likely incremental in the ESP test, with regularization rates shifting dynamically as new stimuli are presented. In contrast, our analysis compares two discrete models—one in which there has been no exposure to the ESP words, and one in which there has been full exposure.

We inspected our data using four different techniques. Each technique points to the same pattern—the individual-GCM outperforms the CELEX-GCM in predicting individual posttest performance, while the individual-MGL offers no improvement over the CELEX-MGL. The specific details of the four techniques and the conclusions drawn from each are described below.

FIRST, by inspecting the individual-GCM and individual-MGL (which incorporate what the participant has seen in the ESP test), we find that the individual-GCM ($C = 0.68$) performs considerably better on the posttest than the CELEX-GCM does ($C = 0.6$), whereas the individual-MGL ($C = 0.63$) is not much better at predicting the ESP posttest data than the CELEX-MGL is ($C = 0.61$) (see the supplementary information). This suggests that the updated GCM model may be capturing some of the change that we see in participant behavior.

SECOND, we directly tested for any significant improvement provided by the individual models. If the categorization model is able to approximate the change in a partici-

pant's representation after the ESP treatment, the model should perform significantly better on the posttest data if it was trained on real English verbs plus information from the ESP test as compared to a version of the model that was trained only on real English verbs. We test for this the following way: for a given posttest response by a participant to a nonce verb, we take the CELEX-MGL/GCM score and subtract it from the individual-MGL/GCM score. (The CELEX score for nonce verbs will not be affected by a participant's exposure in the ESP test.) The resulting score represents the extra information gained by adding the ESP test to the model's training set. We can then use this score as a predictor in a mixed-effects regression model of responses in the ESP posttest to see if it explains any extra variation over the CELEX model. As outlined in the supplementary information, this analysis shows that the extra information from the individual-GCM is a significant predictor of posttest responses, beyond the CELEX-GCM score (EST = 2.61, *SE* = 0.47, ***). The individual-MGL also adds some information over and above the CELEX-MGL, but this effect is much weaker (EST = 0.30, *SE* = 0.14, *). This suggests that the participants' convergent behavior is largely influenced by analogical generalization over a lexicon that includes the nonce words that we have exposed them to, rather than by a more abstract, rules-based method of generalization. However, this is an indirect comparison of two estimates.

THIRD, we used regression to ask whether RULES play any role in predicting posttest behavior, once the updated GCM predictions are taken into account. This provides a direct comparison between predictions of the two learning models.

While we have no evidence that rules have been 'updated', it is still possible that the original, premanipulation rules continue to play a role in influencing participant choices. Indeed, an analysis using residualization of predictors, along the lines of the analysis conducted of the baseline data (see the supplementary information), reveals that the best models of the post-ESP data contain contributions from the CELEX-MGL and the updated individual-GCM scores, but not from the CELEX-GCM and the individual MGL scores. As the relevant predictors are not problematically correlated with one another, this leads us to a final overall best model in which no residualization is necessary (VIF < 2), as shown in Table 6.

Note that the CELEX-GCM and individual-MGL components are not significant predictors and are excluded from the model.

	COEFF	STD. ERROR	*z*-SCORE	SIG
(intercept)	−2.94	0.27	−10.90	***
CELEX-MGL (rescaled)	1.60	0.17	9.36	***
Individual-GCM (rescaled)	2.79	0.38	7.37	***
Participant pretest mean	5.44	0.58	9.30	***

TABLE 6. Regression model summary, with GCM and MGL predictors. Model starting formula: posttest regular response ~ CELEX-MGL + individual-GCM + participant pretest mean + (1 + CELEX-MGL + individual-GCM | participant) + (1 | verb).

FOURTH and finally, we inspected correlation patterns between different predictions for the same verbs to try to understand why the updated individual-GCM scores captured participant responses better than the updated individual-MGL scores did. While the GCM assigns an individual regularization score to each verb, the MGL assigns a set of rules with varying degrees of confidence to each verb, and sets of verbs share rules. This means that the regularization scores of given verbs will reflect the various rules they share with other verbs. According to the MGL, the probability that a verb will be regularized is determined by the strength of the regular and the irregular rules that have it as their input.

Given new information, the GCM will individually change the regularization score of every instance, resulting in incremental change in the rate of regularization. In contrast, the MGL will generate new rules that incorporate the new information as well as the old information. As a result, the rule structure can shift drastically, which results in drastically shifted regularization scores for the individual verbs. The MGL regular rules are not 'paired' with the irregular rules in any way to resemble the GCM categories. Regular rules, built on a higher number of regular forms, can be larger and more robust (and more inflexible) when exposed to new stimuli.

It appears that the behavior of participants in our experiment reflects gradual, instance-based adjustments in category space rather than abrupt, rule-based shifts.

The MGL predicts less variation than the GCM. This can be easily seen for the models trained on CELEX. For the 156 verbs that occur in the ESP experiment, the GCM generates one regularity score for each verb. In contrast, forty-four MGL rules cover these verbs, effectively grouping them together based on formal overlap.

It is remarkable that the MGL has the same accuracy as the GCM on the baseline verbs, given that it uses a more restricted set of predictors. This has an adverse effect, however, when individuals are exposed to orderly heterogeneous variation. Participants show subtle shifts in response to their exposure to the various verb distributions. The individual-GCM is able to capture the general patterns in the shift, whereas the individual-MGL will make rapid adjustments based on a changing rule structure. This can be clearly seen if we compare the CELEX model scores and post-ESP individual scores for the MGL and the GCM, and compare them to real data. These relationships are shown in Figure 4. This figure shows three relationships: the baseline and post-ESP average rates of regularization by our participants (top-left panel), and the baseline and post-ESP average regularity scores predicted by the GCM (top-right panel) and the MGL (bottom panel). The plot thus allows us to compare the effects of a typical versus reversed peer; differences in regularization shift conditions are ignored in this plot.

Averaging real data of binary responses should be interpreted cautiously. What is more, we average over three different rates of regularization—albeit the conditions are balanced in the experiment. At the same time, average plots provide a useful illustration of the underlying patterns of regularization. The actual model comparisons, discussed above, all use binary responses as outcome variables (and these models take stock of all the independent variables).

What Fig. 4 shows is that in the real data, there is a correlation between the baseline data and the posttest data, and that this is echoed in the relationship between the CELEX-GCM and the individual-GCM. This pattern is present for the MGL, but with more residual variation.

The new verbs introduced in the ESP test lead to adjustments in individual patterns in the posttest, but do not completely change them. As outlined in the analysis above, the adjustments predicted by the individual-GCM are significantly predictive of participants' actual behavior. For the MGL, the relationship between the CELEX predictions and the posttest individual predictions is more erratic. Some verbs that received high regularity scores in the CELEX model received low scores in the individual models, and vice versa. These changes in prediction are not significantly related to participants' actual behavior. The forms introduced during the ESP test reduce the individual-MGL's performance—even when those forms are produced by a typical peer. Finally, we can see from the figure that the individual-GCM captures the predicted difference in slope between the typical and the reversed condition. The MGL does so too, but, again, to a lesser extent.

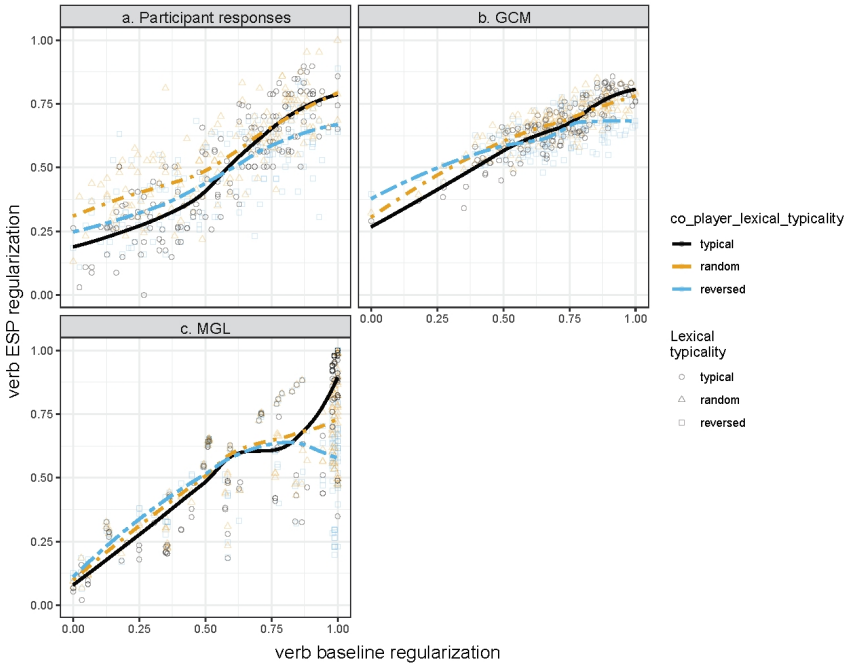

FIGURE 4. Baseline versus post-ESP aggregates of verbs, comparing (a) experimental participant data, (b) GCM predictions, and (c) MGL predictions. In (a), plot points represent cross-participant aggregates for each verb, averaged separately for the typical, random, and reversed conditions, irrespective of regularization shift differences. In (b) and (c), they represent regularity scores by the GCM and the MGL, respectively. In a given panel, a nonce verb has one baseline score and three post-ESP aggregate scores for the three lexical-typicality conditions in the ESP phase. LOESS lines show the trend by lexical-typicality condition. Axes were scaled for comparison.

In sum, both the MGL and the GCM contribute separately to predicting the baseline data, suggesting that both RULES and ANALOGY play a role in determining participant regularization preferences for individual verbs. However, by all of our metrics, the GCM is much more successful in capturing the rapid shifts we observe in these preferences, based on exposure to new lexical items. This suggests that the morphological convergence behavior we observe arises from rapid on-line generalization over lexical forms.

7. DISCUSSION. This article set out to test three research questions. The first question was whether morphological convergence occurs. The answer provided by our experiment was 'yes'. Participants increased or decreased their regularization rates in response to the behavior of a bot peer.

The second question related to the computational mechanism underpinning convergence. Is convergence sensitive to the overall distribution of lexical forms produced by the interlocutor? The answer to this is also 'yes'. Participants adjust not just their level of regularization, but also their distribution of regular forms across verbs. If convergence in this task involved the reweighting of a regular affix, or adjusting an overall baseline regularization rate to better reflect a peer, then we would see an adjustment to the rate of regularization that is constant across lexical items. However, this was not observed.

Convergence was also not narrowly item-specific. Our result cannot be caused by remembering the specific past-tense form produced by the peer and then reproducing that form when those items are reencountered. This explanation could not underpin the ob-

served behavior, because the experiment was set up so that no individual lexical item was seen twice.

Our third research question attempted to probe the relationship between the lexicon and morphological structure in order to understand exactly how updating the lexicon could affect morphological generalizations. We showed that both rules (implemented via the minimal generalization learner) and analogy (via the generalized context model) appear to shape people's overall preferences for past-tense forms, but only analogical effects are manifest when the participants rapidly update their lexicon in response to new items.

Despite the considerable existing literature on linguistic convergence, very few investigations of morphological convergence have been reported. This article provides strong evidence that morphological convergence can occur. Players exposed to different past-tense forms show different degrees and types of shifts in their past-tense preferences in a subsequent task. The convergent behavior extends beyond the immediate interactive task to a subsequent task in which there is no interaction with a peer.

This study, then, adds morphology to the list of linguistic levels on which convergent behavior has been observed. However, it also goes beyond past studies of linguistic convergence by closely examining how the convergent behavior is distributed across a range of different lexical items. This enables us to provide a greater level of insight into the mechanisms that underpin linguistic convergence.

Post-exposure, our participants' preferences appear to be driven by a combination of rules derived from 'real' English words and on-line analogy over existing forms in the lexicon, including the nonce forms to which they have just been exposed. While a rapidly updating GCM captures the convergent behavior elegantly, it is undoubtedly not the only model that could do so. What would be difficult, however, is to capture the results in any model that did not allow for ongoing experience with new lexical items to affect morphological generalizations in an on-line manner. Our results clearly favor an interpretation where morphological convergence is based on rapid on-line adjustments of the lexicon, based on ongoing language experience.

The use of nonce words in our experiment enabled us to detect changes in morphological preferences more easily than if we had used real words as targets. This is because for real words, changes in analogical or rule-based pressures would compete with the strength of stored item-specific representations. This does not mean that we would not expect to observe any changes in real words. However, more extensive exposure over a longer period of time would probably be needed to affect representations of already-known words. In order to uncover the subtle analogical effects that we have shown here, the analysis would need to disentangle the effects of the experimental exposure from the variable strength of particular stored forms. Beckner et al. (2016) have shown that individuals adjust their regularization rate for real verbs in a peer-pressure task where human peers show abnormally high regularization rates. This means that regularization patterns for existing verbs are at least somewhat malleable.

Interestingly, while the pattern of responses following the typical peer were different from the pattern of responses to both the random and reversed peer, the random and reversed peer did not have significantly different effects. The overall sensitivity to baseline shift was equivalent in all three cases—that is, the degree of increased or decreased regularization affected participants in the three conditions equally. But participants in the random and reversed conditions both showed an increased tendency to regularize typically irregular forms, and to irregularize typically regular forms. Our hypothesis, and indeed the GCM (see Fig. 4b), would predict this tendency to be stronger for the REVERSED participants than the RANDOM participants.

We do not have a definitive answer concerning the lack of difference for this distinction. One possibility is that different exemplars may influence participants' behavior differently. The random condition includes more nonexpected forms than the typical condition, and the reversed condition includes even more nonexpected forms. The research literature on human cognition shows that anomalous input can have different effects from more expected input for encoding, memory, and subsequent judgments or actions. But these effects are complex and can work in different directions. In priming experiments, unexpected forms can produce stronger priming (see e.g. Bock 1986, Jaeger & Snider 2013, Peter & Rowland 2019). However, anomalous forms can also be at a disadvantage in being stored in memory (see review: Todd et al. 2019) and as a result may even fail to produce priming at longer time scales (Clopper et al. 2016). Furthermore, even children tend to disregard input from a source that they have found to be generally unreliable (Yow & Li 2018). This complex situation suggests that trade-offs among multiple effects may be responsible for the lack of a significant difference between the random condition and the reversed condition. Exploring such effects is an important direction for future research. In particular, it would be desirable to dynamically track the trajectory of changing preferences over the course of exposure, and to vary the temporal relationship of the exposure and the posttest.

In terms of what the results might imply for convergent past-tense behavior in real interaction, most real interactions do not involve exposure to completely novel word forms. If our past-tense usage is influenced by that of our peers, it is less likely to be via new words, and more likely via changing probabilities for familiar words. We may have both *weeped* and *wept* in our lexical inventory, for example, and exposure to one or the other form might increase the probability of one at the expense of the other. This remains true even if we acknowledge that, given the malleability of the noun-verb distinction in English, speakers do frequently encounter novel denominal verbs, and, what is more, lexical gang effects are far more likely to influence the morphology in languages with rich (inflectional) morphology (see, for example, Daland et al. 2007).

As we point out in §2, our computational implementations of both the GCM and the MGL make the strong assumption that the lexical system is updated by storing and relying on nonce verbs encountered in the experiment. However, they use the information differently in the posttest, where the GCM accesses words and the MGL accesses only abstract generalizations, which are affected to a lesser degree by the novel words introduced in the experiment.

A number of algorithmic implementations of the GCM are possible, including connectionist implementations (Kruschke 1992, Ashby & Rosedahl 2017). This situation raises the possibility that the processing mechanism for the whole-word effects we have demonstrated is not the rapid addition of nonce words to the lexical inventory, but rather modification of the activation patterns for existing words. For example, after encountering *spride*/*sprode*, the processing mechanism might just increase the activation levels for similar verbs (e.g. *stride*/*strode*, *ride*/*rode*). This amounts to an explanation of our data based in whole-word lexical priming as a processing mechanism. Word-level priming of irregular past-tense patterns by nonce forms has been demonstrated in an experiment that manipulates semantic factors but not phonological factors (Ramscar 2002). It could be described in an instance-based processing model that uses attentional weighting, such as Kruschke 1992. A neural network model of structural priming such as Plaut & Gonnerman 2000 would be capable of modeling such a process without explicit reference to word-sized units.

Our data do eliminate the possibility of context-independent priming of the regular past-tense affix, because this process would produce global reweighting of the affix.

However, they do not effectively distinguish between instance-based models in which nonce words are added to the lexical inventory and instance-based or neural network models in which patterns of connections are facilitated or inhibited. It is entirely likely that such a processing model could be developed. However, deciding on and evaluating the specifics of the model would best be done using experimental data on lexical processing that exceed the scope of our study.

Another important area for future research is the nature of the relationship between the short-term convergent behavior we observed and long-term adaptation and learning. There is clearly some link between convergence in interaction and long-term adaptation. In sociophonetics, accent change over the course of the lifetime is claimed to result from the cumulative effects of short-term convergence in many interactions (see e.g. a review in Foulkes & Hay 2015, though see also Sonderegger et al. 2017). It would not be surprising for this phenomenon to occur in other linguistic domains as well. But when do short-term convergence effects feed into long-term learning, and when are they inherently short term and context-specific? Would repeated exposure to our bot-peer shift an individual's overall preferences in the 'real world', or are we simply seeing short-term contextual learning? Our posttest results indicate that the convergence has some effect beyond the direct interaction that triggers it; in a short-term adaptation task, on-line linguistic generalization over recently experienced forms appears to drive behavior. This is congruent with the observation that the ability to adapt to the interlocutor (discussed extensively in our introduction) is one of the hallmarks of human linguistic competence. What we cannot know from our results is what happens thereafter. Would a period of lexical embedding (e.g. with sleep) eventually cause the new forms to affect 'rules' at a more abstract level? For our experiment, given the hollow semantics of the nonce words, the time scale of the experiment, and the fact that the posttest takes place in a setting similar to the 'gamified' ESP test, it does seem likely that the convergence effects we observe are restricted in time and space.

It does not follow from our results that all short-term or context-specific learning will be purely analogical. In adapting to novel speech patterns, people are capable of rapid remapping of the relationship between an allophone and a phoneme in both production (German et al. 2013) and perception (Cutler et al. 2010). Our experiment differs from these studies both in target level of the linguistic system (morphology rather than allophony) and in the experimental manipulation in the training phase. Instead of an absolute shift in the outcomes for just one or two categories, our study had probabilistic shifts in the outcomes for a larger number of verb classes. It is possible that we do not observe the updating of abstract rules in our study because either of these factors, or both together, reduce the impact on the 'core' lexicon and on the abstractions that are built from that lexicon. Alternatively, it could be that the process of updating morphological abstractions is simply not so rapid, or requires a greater degree of genuine lexical embedding (McClelland et al. 1995, O'Reilly & Norman 2002, Gaskell & Dumay 2003, Kumaran et al. 2016; see also §2). The scope of the present study allows us to observe that short-term morphological convergence can arise from an analogical process over a rapidly updating lexicon. Establishing the precise relationship between such short-term convergent behavior and longer-term learning will be an important direction for future work.

The fact that our participants are sensitive to the distribution of regular past-tense forms over lexical items also informs the more general debate about the nature of inflectional morphology in general, and the English past tense in particular. The English past tense has been a testing ground for a wide range of theories and predictions regarding the nature of lexical representations, in particular the representation of inflectional patterns

as rules or as generalizations over specific items (Bybee & Slobin 1982, Rumelhart & McClelland 1986, Plunkett & Marchman 1991, 1993, McClelland & Patterson 2002, Albright & Hayes 2003, Seidenberg & Plaut 2014). Our results lend strong support to a view of past-tense formation as including both an abstract component and a component involving on-line generalization over specific items. Moreover, they suggest that the set of relevant items is constantly being updated, and that the nature of past-tense generalization is thus highly malleable. Abstract rules play a role, but these appear to be less malleable, and thus less likely to be implicated in morphological convergence. Just as hybrid models of phonology showing that multiple levels of representation and abstraction underpin sound systems (cf. Pierrehumbert 2016), our results indicate that morphological productivity, well-formedness, and variation are also governed by influences at multiple levels.

In sum, the morphological convergence we observed does not result from a simple adjustment of preference for one variant over another. Rather, it is the result of updating, in real time, the distributions of morphological variants that our generalizations rely on. This article thus provides a very clear case of morphological convergence, together with evidence that the convergent behavior emerges through an analogical process over a rapidly updating lexicon.

APPENDIX A: IMPLEMENTATION OF THE GENERALIZED CONTEXT MODEL

A1. OUTLINE. Our implementation of the GCM evaluates the competition between two categories, regular and irregular, for each nonce-verb base form. The framework of Nosofsky 1990 is adapted to morphophonology by using a segmental-similarity calculation based on natural classes (Frisch et al. 2004). The same treatment of segmental similarity is used in the implementations of the GCM in Albright & Hayes 2003 and Dawdy-Hesterberg & Pierrehumbert 2014. We build on Dawdy-Hesterberg & Pierrehumbert 2014 in that we define our categories based on formal similarity.

A2. TRAINING DATA. Participants are presented with a sequence of nonce-verb base forms and have to pick either a regular or an irregular past-tense form for each. The irregular past-tense form is predetermined by the class of the stem, so that, for a given verb, the participants can only choose between the regular past-tense form and the irregular past-tense form we assigned to the verb. (So, for instance, for *splive*, a verb in the DROVE class, they can choose either *splived* or *splove*, but not *splift* or *sploven*, etc.) For a given class (such as DROVE verbs), the GCM has a choice between two sets of verb types.

The irregular set consists of verb types in CELEX that form their past tense according to the pattern captured by the class (such as an {[aɪ],[i]} → [oʊ] alternation). The regular set consists of verb types that have base forms that are similar to these irregular forms but have a regular *-ed* past-tense form, as well as miscellaneous regular verbs—those that do not belong to any of our schemata. We narrow the regular set to monosyllabic forms. However, all polysyllabic irregular forms that could serve as a point of comparison are compounds based on monosyllabic forms (an example is *overwrite*, a compound form of irregular *write*). A compound form might be more regular than a simplex form, but CELEX will list both the regular and the irregular variant in both cases. Table A1 shows the descriptions of the verb classes using regular expressions.

	INPUT	ALTERNATION	
VERB CLASS	REGULAR EXPRESSION	IPA	
DROVE	[i2][zvdltnk]$	{i, aɪ} + {z, v, d, l, t, n, k} ##	{i, aɪ} → oʊ
SANG	I(m|N|Nk)$	{ɪ} + {m, ŋ, ŋk} ##	ɪ → æ
KEPT	i[lpnm]$	{i} + {l, p, n, m} ##	i → ɛCt
BURNT	[3EI]nl$	{ɜ, ɛ, ɪ} + {n, l} ##	{ɜ, ɛ, ɪ} → {ɜ, ɛ, ɪ}Ct

TABLE A1. Descriptions of verb classes in the GCM. 'C' marks any consonant.

Our starting point for the training set, following Albright and Hayes (2003), is the list of verbs in the CELEX corpus (Baayen et al. 1993, based on Sinclair 1987) with a token frequency of 10 or above, encompassing 3,156 forms. However, similarity requirements restrict the respective training sets. We use the DISC transcription in which each contrastive segment of English is represented by a unique character (<dr2v> equals [dɹaɪv]).

Table A2 shows the number of verbs in CELEX that were used as training sets for our verb classes. The irregular set consists of forms that match the schema and are irregular. The regular set contains schema matches

that are regular, in addition to miscellaneous regulars. The miscellaneous set consists of monosyllabic verbs that do not belong to any of the schemata and are regular. These are included in the regular training set of each verb class.

VERB CLASS	IRREGULAR SET	REGULAR SET	MISC. REGULAR VERBS
BURNT	6	42	1218
DROVE	14	83	1218
KEPT	12	31	1218
SANG	8	13	1218
count of unique forms	*40*	*169*	*1218*

TABLE A2. Number of forms in each verb class, GCM training data.

The model calculates the similarity of a given nonce verb to the regular and the irregular set. Comparisons to stems in other classes are not calculated, as past-tense markings for these classes were not available to participants in the forced-choice tasks.

A3. ESTIMATION. To calculate the similarity between two words, we first compute their dissimilarity. This is achieved using the string-edit (Levenshtein) distance, which is the smallest number of changes needed to transform one word into the other. For one unit of edit distance, these costs range from 0 (the corresponding segments are identical) to 1 (inserting or deleting an entire segment). Following Albright and Hayes (2003) and Dawdy-Hesterberg and Pierrehumbert (2014), costs between 0 and 1 are assigned to corresponding segments that are not identical, based on how much the segments differ.

All parts of the word are weighted equally, because despite evidence that past-tense formation in English is predominantly driven by overlaps in word endings, onsets also play a role (cf. the predominance of *s*(−cont) onsets in irregular verbs forming the past tense with a vowel change, e.g. *stink*, *sink*, etc.; see Bybee & Moder 1983).

The transformation in A1, originating with Nosofsky 1990, is used to convert dissimilarity into similarity.

$$\text{(A1)} \quad \eta_{ij} = \exp\left(-d_{ij}/s\right)^p$$

In equation A1, η_{ij} represents the similarity between form i and form j, while d_{ij} is the dissimilarity between the two forms. s and p are free parameters. We explored a range of parameter settings and use $s = 0.9$ and $p = 1$, which provide the best model fit on the baseline data. (In contrast, Albright and Hayes use $s = 0.4$ and $p = 1$.)

When p is set to 1, as here, the similarity function is an exponential, rather than a Gaussian, function of the dissimilarity. The weighting parameter s controls how quickly the similarity decreases as the difference (or distance) between the forms increases. When s is small, the behavior of the model will be dominated by the small group of instances that differ very little from any given novel form. As it becomes larger, instances that differ more increase their influence on the overall model behavior (Nosofsky 1990, Nakisa et al. 2001, Albright & Hayes 2003, Dawdy-Hesterberg & Pierrehumbert 2014). Thus, s effectively controls the size of the set of verbs that will be taken into account in determining the support for the regular versus the irregular outcome.

The overall similarity S_{iC_J} of a test form i to a set C_J is calculated by summing the similarity η_{ij} of each member j of class C_J to the test form i, and dividing by the summed similarity η_{ij} of each member k of class C_K (the class of all stored forms) to the test form i. This calculation is summarized in equation A2.

$$\text{(A2)} \quad S_{iC_J} = \frac{\Sigma_{j \in C_J} \eta_{ij}}{\Sigma_{k \in C_K} \eta_{ik}}$$

A4. OUTPUT FORMAT. The overall score used in our analyses is the REGULARITY SCORE, which is the complement to the IRREGULARITY SCORE and reaches a maximum of 1.0 when the output is certain to be regular. Unlike Dawdy-Hesterberg & Pierrehumbert 2014, there is no decision rule on top of the scoring, such that any form that is more likely than not to be regular is predicted to surface as regular all the time. This specific decision rule is statistically optimal and was imposed in Dawdy-Hesterberg & Pierrehumbert 2014 in order to determine the ceiling performance for a computational model. The present article, in contrast, analyzes data aggregated across human participants with differing decision thresholds. As discussed in Schumacher et al. 2014 and Schumacher & Pierrehumbert 2017, the input-output relationship in such aggregated data are typically reported to be nearly probability-matching.

We rescale the regularity score to match the range of participant responses: [0,1]. The modified score is interpretable as the probability that the outcome will be regular in aggregated data. It is also appropriate to attribute this type of gradience to people's initial expectations about other people's behavior, on the assumption that people realistically encode the variability they have encountered.

A5. EXAMPLE: *SPLIVE*. The nonce form *splive* belongs to the DROVE class in our model. The two past forms of *splive* in the experiment are regular *splived* and irregular *splove*. It is compared to 1,301 regular verbs—

eighty-three verbs that match the DROVE schema (e.g. *side*, *hive*, *line*), and 1,218 miscellaneous verbs. It is also compared to fourteen irregular verbs (e.g. *drive*, *stride*, *smite*) in this class. Overall, it is more similar to the regular set: its regularity score is 0.57.

APPENDIX B: IMPLEMENTATION OF THE MINIMAL GENERALIZATION LEARNER

B1. OUTLINE. The minimal generalization learner is an algorithm for forming input-output rules of varying generality, which then compete to generate the output. The MGL is implemented here from materials made available by Albright and Hayes (Albright & Hayes 2003). These include their SEGMENTAL SIMILARITY CALCULATOR, implementing the natural class-based similarity metric due to Frisch et al. (2004), also used in the GCM implementation. Due to issues with the MGL code, we had to fit the MGL using the graphical interface for each separate participant. The PYAUTOGUI library in Python was used to automate this; the code is available in the online repository at https://doi.org/10.5281/zenodo.4103379.

B2. TRAINING DATA. For our model fitted on our baseline nonce-word stimuli, the MGL is trained on regular and irregular English verbs with a minimum-frequency cutoff of 10 in CELEX (Baayen et al. 1993), encompassing 4,160 past/present verb transcriptions.

The MGL builds rules based on all verb forms in CELEX with a token frequency of 10 or above. However, the structural descriptions of the resulting rules do not cover all of these forms. Table A3 shows the number of unique forms covered by the structural descriptions of the REGULAR and IRREGULAR rules that are relevant to each class.

CATEGORY	RULE TYPE	RELATED FORMS	EXCEPTIONS
BURNT	irregular	24	41
DROVE	irregular	21	114
KEPT	irregular	10	9
SANG	irregular	11	77
BURNT	regular	38	21
DROVE	regular	29	27
KEPT	regular	67	41
SANG	regular	47	38

TABLE A3. Number of forms in each verb class, MGL training data.

The MGL generates multiple possible past-tense forms for each nonce verb. We consider to be relevant only those rules that generate the past-tense forms that appear in the experiment (e.g. *splive* : *splived*/*splove*). There is at most one relevant regular rule and one relevant irregular rule for one verb, but multiple rules can generate the (ir)regular forms for each verb class. We return to this in the next section.

Note that the sets of exceptions and related forms of each rule can overlap. As a consequence, the MGL rules apply to fewer forms than might appear from the table: 401 in total.

B3. ESTIMATION. The MGL begins by considering the relationship between each verb and its past tense as a RULE. For each pair of verbs in the training data, it then attempts to create a more general rule. It does so by aligning the word forms and analyzing shared phonetic features. For example, merging the word-specific rules for *ring*/*rang* and for *stink*/*stank* yields a more general rule that expresses the information they share: [ɪ] → [æ] / [+coronal] _ [ŋ]. Each rule inferred in this way is then further generalized on the basis of more comparisons; for instance, taking note of *swim*/*swam* expands the [ɪ] → [æ] rule to specify that it occurs before all [+nasal] consonants.

The structural description for each rule has a scope, which is the number of verbs conforming to the description, to which the rule might apply. The number of hits is the number of such verbs where the rule generates the correct output. In our example, *think* and *blink* fall in the scope of the rule, but they are not hits, because their past tenses display other patterns (*thought* and *blinked*). The RAW CONFIDENCE of the rule is the ratio of hits to scope.

$$\text{(A3)}\quad \text{Raw confidence} = \frac{\text{hits}}{\text{scope}}$$

The raw confidence is 1.0 if the rule applies to all forms that meet its structural description. It is less than 1.0 if some forms meeting its structural description have past tenses other than that predicted by the structural change. Raw confidence values of 0 are not found, because a rule needs to apply to two or more examples to be posited in the first place.

The MGL raw confidence metric is adjusted on the basis of user-specified confidence limits to generate an ADJUSTED CONFIDENCE SCORE that takes into account the amount and distribution of available data. The MGL's lower limit affects how much confidence is assigned to rules that have a small number of instances; generalizations that are based on a smaller number of word types are penalized. The MGL's upper limit cur-

tails the application of seemingly general rules that are in fact driven by a more specific rule (Albright & Hayes 2002). The MGL is implemented here with its default settings, with the exception of the algorithm's confidence limits. We implement the MGL with lower and upper confidence limits of 55% and 95%, respectively, since these values afford the best fit to English verb data in Albright & Hayes 2003.

Note that the MGL algorithm automatically groups together verbs on the basis of shared phonological properties; thus, verbs are most likely to form strong generalizations with other verbs that share the same onset or rhyme. Attempts to merge diverse word forms under a single generalization would be more likely to incur penalties (i.e. exceptions). This feature of the MGL is important for comparing with the methods of the GCM. Both algorithms allow for category-specific similarities to play a role in rule formation.

B4. OUTPUT FORMAT. Recall that in both our baseline and ESP experiments the trial task is prompted by a stem and offers a choice between a regular form and a specific irregular form, presented orthographically. In order to model this choice, we take the MGL rule for the stem that outputs the regular form (the relevant regular rule) and the rule that outputs the specific irregular form (the relevant irregular rule). If several regular/irregular rules generate the same form, we take the one with the highest adjusted confidence, following Albright & Hayes 2003. We use these rules to calculate the form's relative (adjusted) confidence.

Of 156 test verbs in the ESP posttest, the CELEX-trained MGL generates a relevant regular rule for every verb. It does not generate a relevant irregular rule for twenty-eight verbs. These are all nonce verbs in the KEPT category (see §4.1 in the main text). In this category, irregular forms are derived from the stem through a vowel change (e.g. *greel* → *grelt*). The verbs missing the relevant irregular rule all have bases ending in <m>, <n>, or <l>. In our implementation, there is an insufficient number of verb types in the training set to support the induction of irregular rules covering these bases. Decreasing the cutoff criterion for the model leads to the generation of more of the currently missing irregular rules, but the overall model fit becomes worse (cf. below). Therefore, we keep the cutoff criterion and assume that the adjusted confidence of the irregular rule for these twenty-eight verbs is zero.

For all forms, we then take the relative confidence of the regular rule as compared to the regular and the irregular rule for each verb and take this as the adjusted regular confidence of the given verb. (If the irregular rule is missing, the value of this adjusted regular confidence is 1.) This is given by equation A4.

(A4) Relative (adjusted) confidence =

$$\frac{\text{adjusted confidence of relevant regular rule}}{\text{adj.confidence of relevant regular rule} + \text{adj.confidence of relevant irregular rule}}$$

This relative adjusted confidence represents the MGL regularity score for an item, to be compared against the regularity score from the GCM (see Appendix A).

B5. EXAMPLE: *SPLIVE*. The two past forms of *splive* in the experiment are regular *splived* and irregular *splove*. The relevant regular rule that generates the regular past tense is '∅ → [d] / {ð, ʃ, θ, ʒ, f, s, v, z}'. The structural description indicates that this is a suffixation rule that can apply to forms that end in an anterior fricative (a natural class in our feature system). The raw confidence of this rule is 0.98. This is because this rule applies to most forms in its scope (698/712). The adjusted confidence is very similar: 0.968. This is because this rule applies to a large number of forms overall. The relevant irregular rule that generates the irregular form is '[aɪ] → [oʊ] / {ð, ʒ, dʒ, d, l, n, ɹ, z} _ v'. It applies to [aɪ] in the nucleus preceded by a voiced anterior consonant and followed by [v]. In CELEX, the rule applies to three forms (*drive*, *strive*, *dive*) and fails to apply to five (*arrive*, *thrive*, *contrive*, *rive*, *connive*). Its raw confidence is 0.375. Its adjusted confidence is slightly lower (0.366). This is because it applies to a smaller number of forms overall. The relative (adjusted) confidence of the predicted regularity of *splive* is 0.98 / (0.98 + 0.37) = 0.73. (As with the GCM predictions—see §A4—we rescale the MGL predictions for statistical analysis. The respective rescaled value is 0.363.)

Further notes on the two models can be found in the online supplementary information.

REFERENCES

ALBRIGHT, ADAM, and BRUCE HAYES. 2002. Modeling English past tense intuitions with minimal generalization. *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, 58–69. DOI: 10.3115/1118647.1118654.

ALBRIGHT, ADAM, and BRUCE HAYES. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90.119–61. DOI: 10.1016/S0010-0277(03)00146-X.

ALEGRE, MARIA, and PETER GORDON. 1999a. Frequency effects and the representational status of regular inflections. *Journal of Memory and Language* 40.41–61. DOI: 10.1006/jmla.1998.2607.

ALEGRE, MARIA, and PETER GORDON. 1999b. Rule-based versus associative processes in derivational morphology. *Brain and Language* 68.347–54. DOI: 10.1006/brln.1999.2066.
ANDERSON, ANNE H.; MILES BADER; ELLEN GURMAN BARD; ELIZABETH BOYLE; GWYNETH DOHERTY; SIMON GARROD; STEPHEN ISARD; JACQUELINE KOWTKO; JAN MCALLISTER; JIM MILLER; et al. 1991. The HCRC map task corpus. *Language and Speech* 34.351–66. DOI: 10.1177/002383099103400404.
ASCH, SOLOMON E. 1951. Effects of group pressure upon the modification and distortion of judgments. *Groups, leadership, and men: Research in human relations*, ed. by Harold Guetzkow, 177–90. Pittsburgh: Carnegie Press.
ASHBY, F. GREGORY, and LUKE ROSEDAHL. 2017. A neural interpretation of exemplar theory. *Psychological Review* 124.472–82. DOI: 10.1037/rev0000064.
AUER, PETER; DAVID FERTIG; PAUL J. HOPPER; and ROBERT W. MURRAY. 2015. *Hermann Paul's principles of language history revisited: Translations and reflections*. Berlin: De Gruyter.
BAAYEN, R. HARALD; RICHARD PIEPENBROCK; and HEDDERIK VAN RIJN. 1993. The CELEX lexical database on CD-ROM. Philadelphia: Linguistic Data Consortium.
BABEL, MOLLY. 2010. Dialect divergence and convergence in New Zealand English. *Language in Society* 39.437–56. DOI: 10.1017/S0047404510000400.
BABEL, MOLLY. 2012. Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics* 40.177–89. DOI: 10.1016/j.wocn.2011.09.001.
BABEL, MOLLY, and DASHA BULATOV. 2012. The role of fundamental frequency in phonetic accommodation. *Language and Speech* 55.231–48. DOI: 10.1177/0023830911417695.
BALOTA, DAVID A.; MAURA PILOTTI; and MICHAEL J. CORTESE. 2001. Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition* 29.639–47. DOI: 10 .3758/BF03200465.
BARD, ELLEN G.; A. J. LOWE; and GERRY T. M. ALTMANN. 1989. The effect of repetition on words in recorded dictations. *Eurospeech '89: Proceedings of the European Conference on Speech Communication and Technology*, vol. 2, 573–76.
BECKNER, CLAY; PÉTER RÁCZ; JÜRGEN BRANDSTETTER; JENNIFER B. HAY; and CHRISTOPH BARTNECK. 2016. Participants conform to humans but not to humanoid robots in an English past tense formation task. *Journal of Language and Social Psychology* 35.158–79. DOI: 10.1177/0261927X15584682.
BERKO, JEAN. 1958. The child's learning of English morphology. *Word* 14.150–77. DOI: 10 .1080/00437956.1958.11659661.
BOCK, J. KATHRYN. 1986. Syntactic persistence in language production. *Cognitive Psychology* 18.355–87. DOI: 10.1016/0010-0285(86)90004-6.
BOULIS, CONSTANTINOS, and MARI OSTENDORF. 2005. A quantitative analysis of lexical differences between genders in telephone conversations. *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics*, 435–42. DOI: 10.3115/1219 840.1219894.
BRANDSTETTER, JÜRGEN; CLAY BECKNER; EDUARDO BENITEZ SANDOVAL; and CHRISTOPH BARTNECK. 2017. Persistent lexical entrainment in HRI. *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, Vienna, 63–72.
BRENNAN, SUSAN E., and HERBERT H. CLARK. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22.1482–93. DOI: 10.1037/0278-7393.22.6.1482.
BYBEE, JOAN L. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes* 10.425–55. DOI: 10.1080/01690969508407111.
BYBEE, JOAN L., and CAROL LYNN MODER. 1983. Morphological classes as natural categories. *Language* 59.251–70. DOI: 10.2307/413574.
BYBEE, JOAN L., and DAN I. SLOBIN. 1982. Rules and schemas in the development and use of the English past tense. *Language* 58.265–89. DOI: 10.2307/414099.
CLAHSEN, HARALD. 1999. Lexical entries and rules of language: A multidisciplinary study of German inflection. *Behavioral and Brain Sciences* 22.991–1013. DOI: 10.1017/s014 0525x99002228.
CLOPPER, CYNTHIA G.; TERRIN N. TAMATI; and JANET B. PIERREHUMBERT. 2016. Variation in the strength of lexical encoding across dialects. *Journal of Phonetics* 58.87–103. DOI: 10.1016/j.wocn.2016.06.002.

Coutanche, Marc N., and Sharon L. Thompson-Schill. 2014. Fast mapping rapidly integrates information into existing memory networks. *Journal of Experimental Psychology: General* 143.2296–2303. DOI: 10.1037/xge0000020.

Crump, Matthew J. C.; John V. McDonnell; and Todd M. Gureckis. 2013. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLOS ONE* 8:e57410. DOI: 10.1371/journal.pone.0057410.

Cuskley, Christine F.; Martina Pugliese; Claudio Castellano; Francesca Colaiori; Vittorio Loreto; and Francesca Tria. 2014. Internal and external dynamics in language: Evidence from verb regularity in a historical corpus of English. *PLOS ONE* 9:e102882. DOI: 10.1371/journal.pone.0102882.

Cutler, Anne; Frank Eisner; James M. McQueen; and Dennis Norris. 2010. How abstract phonemic categories are necessary for coping with speaker-related variation. *Laboratory phonology 10*, ed. by Cécile Fougeron, Barbara Kuehnert, Mariapaola D'Imperio, and Nathalie Vallee, 91–112. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110224917.1.91.

Daland, Robert; Andrea D. Sims; and Janet B. Pierrehumbert. 2007. Much ado about nothing: A social network model of Russian paradigmatic gaps. *Proceedings of the 45th annual meeting of the Association for Computational Linguistics*, 936–43. Online: https://www.aclweb.org/anthology/P07-1118.pdf.

Dawdy-Hesterberg, Lisa, and Janet B. Pierrehumbert. 2014. Learnability and generalisation of Arabic broken plural nouns. *Language, Cognition and Neuroscience* 29.1268–82. DOI: 10.1080/23273798.2014.899377.

Dell, Gary. 2000. Commentary: Counting, connectionism, and lexical representation. *Papers in laboratory phonology V: Acquisition and the lexicon*, ed. by Michael B. Broe and Janet B. Pierrehumbert, 335–48. Cambridge: Cambridge University Press.

Donkin, Chris, and Robert M. Nosofsky. 2012. A power-law model of psychological memory strength in short- and long-term recognition. *Psychological Science* 23.625–34. DOI: 10.1177/0956797611430961.

Estival, Dominique. 1985. Syntactic priming of the passive in English. *Text & Talk* 5.7–22. DOI: 10.1515/text.1.1985.5.1-2.7.

Fehringer, Carol. 2004. How stable are morphological doublets? A case study of /ə/~Ø variants in Dutch and German. *Journal of Germanic Linguistics* 16.285–329. DOI: 10.1017/S1470542704040425.

Foulkes, Paul, and Jennifer B. Hay. 2015. The emergence of sociophonetic structure. *The handbook of language emergence*, ed. by Brian MacWhinney and William O'Grady, 292–313. West Sussex: Wiley-Blackwell. DOI: 10.1002/9781118346136.ch13.

Frisch, Stefan A.; Janet B. Pierrehumbert; and Michael B. Broe. 2004. Similarity avoidance and the OCP. *Natural Language and Linguistic Theory* 22.179–228. DOI: 10.1023/B:NALA.0000005557.78535.3c.

Gardner, Robert C., and Peter D. MacIntyre. 1991. An instrumental motivation in language study: Who says it isn't effective? *Studies in Second Language Acquisition* 13.57–72. DOI: 10.1017/S0272263100009724.

Garrod, Simon, and Martin J. Pickering. 2004. Why is conversation so easy? *Trends in Cognitive Sciences* 8.8–11. DOI: 10.1016/j.tics.2003.10.016.

Gaskell, M. Gareth, and Nicolas Dumay. 2003. Lexical competition and the acquisition of novel words. *Cognition* 89.105–32. DOI: 10.1016/S0010-0277(03)00070-2.

Gelman, Andrew, and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

German, James S.; Katy Carlson; and Janet B. Pierrehumbert. 2013. Reassignment of consonant allophones in rapid dialect acquisition. *Journal of Phonetics* 41.228–48. DOI: 10.1016/j.wocn.2013.03.001.

Giles, Howard, and Nikolas Coupland. 1991. *Language: Contexts and consequences*. Pacific Grove, CA: Thomson Brooks/Cole.

Goldstein, Michael H.; Andrew P. King; and Meredith J. West. 2003. Social interaction shapes babbling: Testing parallels between birdsong and speech. *Proceedings of the National Academy of Sciences* 100.8030–35. DOI: 10.1073/pnas.1332441100.

Gregory, Stanford W., Jr., and Stephen Webster. 1996. A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social sta-

tus perceptions. *Journal of Personality and Social Psychology* 70.1231–40. DOI: 10.1037/0022-3514.70.6.1231.

GREGORY, STANFORD W., JR.; STEPHEN WEBSTER; and GANG HUANG. 1993. Voice pitch and amplitude convergence as a metric of quality in dyadic interviews. *Language & Communication* 13.195–217. DOI: 10.1016/0271-5309(93)90026-J.

GRIES, STEFAN TH. 2005. Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research* 34.365–99. DOI: 10.1007/s10936-005-6139-3.

HABER, LYN R. 1976. Leaped and leapt: A theoretical account of linguistic variation. *Foundations of Language* 14.211–38. Online: https://www.jstor.org/stable/25170054.

HALL, MATTHEW L.; VICTOR S. FERREIRA; and RACHEL I. MAYBERRY. 2015. Syntactic priming in American Sign Language. *PLOS ONE* 10:e0119611. DOI: 10.1371/journal.pone.0119611.

HARE, MARY, and JEFFREY L. ELMAN. 1995. Learning and morphological change. *Cognition* 56.61–98. DOI: 10.1016/0010-0277(94)00655-5.

HARRELL, FRANK E. 2013. *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York: Springer Science & Business Media.

HAYES, BRUCE; PÉTER SIPTÁR; KIE ZURAW; and ZSUZSA LONDE. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85.822–63. DOI: 10.1353/lan.0.0169.

HENRICH, JOSEPH; STEVEN J. HEINE; and ARA NORENZAYAN. 2010. The weirdest people in the world? *Behavioral and Brain Sciences* 33.61–83. DOI: 10.1017/S0140525X0999152X.

HJELMSLEV, LOUIS. 1961 [1953]. *Prolegomena to a theory of language*. Trans. by Francis J. Whitfield. Madison: University of Wisconsin Press.

HORTON, WILLIAM S. 2007. The influence of partner-specific memory associations on language production: Evidence from picture naming. *Language and Cognitive Processes* 22.1114–39. DOI: 10.1080/01690960701402933.

HORTON, WILLIAM S., and SUSAN E. BRENNAN. 2016. The role of metarepresentation in the production and resolution of referring expressions. *Frontiers in Psychology* 7:1111. DOI: 10.3389/fpsyg.2016.01111.

IBARRA, ALYSSA, and MICHAEL K. TANENHAUS. 2016. The flexibility of conceptual pacts: Referring expressions dynamically shift to accommodate new conceptualizations. *Frontiers in Psychology* 7:561. DOI: 10.3389/fpsyg.2016.00561.

IPEIROTIS, PANAGIOTIS G. 2010. Demographics of Mechanical Turk. Ceder 10–01 working paper. New York: New York University.

JAEGER, T. FLORIAN, and NEAL E. SNIDER. 2013. Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition* 127.57–83. DOI: 10.1016/j.cognition.2012.10.013.

KAPATSINSKI, VSEVOLOD. 2010. Velar palatalization in Russian and artificial grammar: Constraints on models of morphophonology. *Laboratory Phonology* 1.361–93. DOI: 10.1515/labphon.2010.019.

KAPNOULA, EFTHYMIA C.; STEPHANIE PACKARD; PRAHLAD GUPTA; and BOB MCMURRAY. 2015. Immediate lexical integration of novel word forms. *Cognition* 134.85–99. DOI: 10.1016/j.cognition.2014.09.007.

KROTT, ANDREA; R. HARALD BAAYEN; and ROBERT SCHREUDER. 2001. Analogy in morphology: Modeling the choice of linking morphemes in Dutch. *Linguistics* 39.51–94. DOI: 10.1515/ling.2001.008.

KRUSCHKE, JOHN K. 1992. ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review* 99.22–44. DOI: 10.1037/0033-295X.99.1.22.

KUMARAN, DHARSHAN; DEMIS HASSABIS; and JAMES L. MCCLELLAND. 2016. What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in Cognitive Sciences* 20.512–34. DOI: 10.1016/j.tics.2016.05.004.

LINDSAY, SHANE, and M. GARETH GASKELL. 2013. Lexical integration of novel words without sleep. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 39.608–22. DOI: 10.1037/a0029243.

LINDSAY, SHANE; LEANNE M. SEDIN; and M. GARETH GASKELL. 2012. Acquiring novel words and their past tenses: Evidence from lexical effects on phonetic categorisation. *Journal of Memory and Language* 66.210–25. DOI: 10.1016/j.jml.2011.07.005.

Lleras, Alejandro, and Adrian Von Mühlenen. 2004. Spatial context and top-down strategies in visual search. *Spatial Vision* 17.465–82. DOI: 10.1163/1568568041920113.
Maddox, W. Todd, and F. Gregory Ashby. 1998. Selective attention and the formation of linear decision boundaries: Comment on McKinley and Nosofsky (1996). *Journal of Experimental Psychology: Human Perception and Performance* 24.302–22. DOI: 10.1037/0096-1523.24.1.301.
Marr, David, and Tomaso Poggio. 1976. From understanding computation to understanding neural circuitry. *Neurosciences Research Program Bulletin* 15.470–88.
McClelland, James L.; Bruce L. McNaughton; and Randall C. O'Reilly. 1995. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* 102.419–57. DOI: 10.1037/0033-295x.102.3.419.
McClelland, James L., and Karalyn Patterson. 2002. Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences* 6.465–72. DOI: 10.1016/S1364-6613(02)01993-9.
McKinley, Stephen C., and Robert M. Nosofsky. 1996. Selective attention and the formation of linear decision boundaries. *Journal of Experimental Psychology: Human Perception and Performance* 22.294–317. DOI: 10.1037//0096-1523.22.2.294.
Mielke, Jeff. 2008. *The emergence of distinctive features*. Oxford: Oxford University Press.
Moder, Carol Lynn. 1992. *Productivity and categorization in morphological classes*. Buffalo: State University of New York dissertation.
Nakisa, Ramin C.; Kim Plunkett; and Ulrike Hahn. 2001. A cross-linguistic comparison of single and dual-route models of inflectional morphology. *Models of language acquisition: Inductive and deductive approaches*, ed. by Peter Broeder and Jaap Murre, 201–22. Cambridge, MA: MIT Press.
Nosofsky, Robert M. 1988. Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14.54–65. DOI: 10.1037/0278-7393.14.1.54.
Nosofsky, Robert M. 1990. Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology* 34.393–418. DOI: 10.1016/0022-2496(90)90020-A.
Nosofsky, Robert M.; Gregory E. Cox; Rui Cao; and Richard M. Shiffrin. 2014. An exemplar-familiarity model predicts short-term and long-term probe recognition across diverse forms of memory search. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 40.1524–39. DOI: 10.1037/xlm0000015.
O'Reilly, Randall C., and Kenneth A. Norman. 2002. Hippocampal and neocortical contributions to memory: Advances in the complementary learning systems framework. *Trends in Cognitive Sciences* 6.505–10. DOI: 10.1016/s1364-6613(02)02005-3.
Pardo, Jennifer S. 2006. On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America* 119.2382–93. DOI: 10.1121/1.2178720.
Paul, Hermann. 1995 [1880]. *Prinzipien der sprachgeschichte*. Berlin: Walter de Gruyter.
Peter, Michelle S., and Caroline F. Rowland. 2019. Aligning developmental and processing accounts of implicit and statistical learning. *Topics in Cognitive Science* 11.555–72. DOI: 10.1111/tops.12396.
Pickering, Martin J., and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences* 27.169–90. DOI: 10.1017/S0140525X04000056.
Pickering, Martin J., and Simon Garrod. 2006. Alignment as the basis for successful communication. *Research on Language and Computation* 4.203–28. DOI: 10.1007/s11168-006-9004-0.
Pierrehumbert, Janet B. 2016. Phonological representation: Beyond abstract versus episodic. *Annual Review of Linguistics* 2.33–52. DOI: 10.1146/annurev-linguistics-030514-125050.
Plaut, David C., and Laura M. Gonnerman. 2000. Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language and Cognitive Processes* 15.445–85. DOI: 10.1080/01690960050119661.
Plunkett, Kim, and Virginia Marchman. 1991. U-shaped learning and frequency effects in a multi-layered perception: Implications for child language acquisition. *Cognition* 38.43–102. DOI: 10.1016/0010-0277(91)90022-V.

PLUNKETT, KIM, and VIRGINIA MARCHMAN. 1993. From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition* 48.21–69. DOI: 10.1016/0010-0277(93)90057-3.

PRASADA, SANDEEP, and STEVEN PINKER. 1993. Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes* 8.1–56. DOI: 10.1080/01690969308406948.

RÁCZ, PÉTER; JENNIFER B. HAY; and JANET B. PIERREHUMBERT. 2017. Social salience discriminates learnability of contextual cues in an artificial language. *Frontiers in Psychology* 8:51. DOI: 10.3389/fpsyg.2017.00051.

RAMSCAR, MICHAEL. 2002. The role of meaning in inflection: Why the past tense does not require a rule. *Cognitive Psychology* 45.45–94. DOI: 10.1016/S0010-0285(02)00001-4.

ROBERTS, GARETH. 2010. An experimental study of social selection and frequency of interaction in linguistic diversity. *Interaction Studies* (Special issue: *Experimental semiotics: A new approach for studying the emergence and the evolution of human communication*, ed. by Bruno Galantucci and Simon Garrod) 11.138–59. DOI: 10.1075/is.11.1.06rob.

RUMELHART, DAVID E., and JAMES L. MCCLELLAND. 1986. On learning the past tenses of English verbs. *Parallel distributed processing, vol. 2: Explorations in the microstructure of cognition: Psychological and biological models*, ed. by David E. Rumelhart, James L. McClelland, and the PDP Research Group, 216–71. Cambridge, MA: MIT Press.

SÄILY, TANJA. 2011. Variation in morphological productivity in the BNC: Sociolinguistic and methodological considerations. *Corpus Linguistics and Linguistic Theory* 7.119–41. DOI: 10.1515/cllt.2011.006.

SCHUMACHER, R. ALEXANDER, and JANET B. PIERREHUMBERT. 2017. Prior expectations in linguistic learning: A stochastic model of individual differences. *Proceedings of the 39th annual meeting of the Cognitive Science Society (CogSci 2017)*, 1059.

SCHUMACHER, R. ALEXANDER; JANET B. PIERREHUMBERT; and PATRICK LASHELL. 2014. Reconciling inconsistency in encoded morphological distinctions in an artificial language. *Proceedings of the 36th annual meeting of the Cognitive Science Society (CogSci 2014)*, 2895–2900. Online: https://cogsci.mindmodeling.org/2014/papers/500/paper500.pdf.

SEIDENBERG, MARK S., and DAVID C. PLAUT. 2014. Quasiregularity and its discontents: The legacy of the past tense debate. *Cognitive Science* 38.1190–1228. DOI: 10.1111/cogs.12147.

SIEGELMAN, NOAM, and RAM FROST. 2015. Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language* 81.105–20. DOI: 10.1016/j.jml.2015.02.001.

SINCLAIR, JOHN M. (ed.) 1987. *Looking up: An account of the COBUILD project in lexical computing*. London: Collins.

SNOW, RION; BRENDAN O'CONNOR; DANIEL JURAFSKY; and ANDREW Y. NG. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 254–63. Online: https://www.aclweb.org/anthology/D08-1027.

SONDEREGGER, MORGAN; MAX BANE; and PETER GRAFF. 2017. The medium-term dynamics of accents on reality television. *Language* 93.598–640. DOI: 10.1353/lan.2017.0038.

SPROUSE, JON, and DIOGO ALMEIDA. 2017. Design sensitivity and statistical power in acceptability judgment experiments. *Glossa: A Journal of General Linguistics* 2:14. DOI: 10.5334/gjgl.236.

STEMBERGER, JOSEPH PAUL, and BRIAN MACWHINNEY. 1986. Form-oriented inflectional errors in language processing. *Cognitive Psychology* 18.329–54. DOI: 10.1016/0010-0285(86)90003-4.

STEWART, NEIL; CHRISTOPH UNGEMACH; ADAM J. L. HARRIS; DANIEL M. BARTELS; BEN R. NEWELL; GABRIELE PAOLACCI; and JESSE CHANDLER. 2015. The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision Making* 10.479–91. Online: http://journal.sjdm.org/14/14725/jdm14725.pdf.

SZMRECSANYI, BENEDIKT. 2005. Language users as creatures of habit: A corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory* 1.113–50. DOI: 10.1515/cllt.2005.1.1.113.

Szmrecsanyi, Benedikt. 2006. *Morphosyntactic persistence in spoken English: A corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis*. Berlin: Walter de Gruyter.

Thornton, Anna M. 2012. Overabundance in Italian verb morphology and its interactions with other non-canonical phenomena. *Irregularity in morphology (and beyond)*, ed. by Thomas Stolz, Hitomi Otsuka, Aina Urdze, and Johan van der Auwera, 251–69. Berlin: Akademie.

Todd, Simon; Janet B. Pierrehumbert; and Jennifer Hay. 2019. Word frequency effects in sound change as a consequence of perceptual asymmetries: An exemplar-based model. *Cognition* 185.1–20. DOI: 10.1016/j.cognition.2019.01.004.

Treiman, Rebecca; Mark S. Seidenberg; and Brett Kessler. 2015. Influences on spelling: Evidence from homophones. *Language, Cognition and Neuroscience* 30.544–54. DOI: 10.1080/23273798.2014.952315.

von Ahn, Luis, and Laura Dabbish. 2004. Labeling images with a computer game. *CHI '04: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 319–26. DOI: 10.1145/985692.985733.

Webb, James T. 1972. Interview synchrony: An investigation of two speech rate measures in an automated standardized interview. *Studies in dyadic communication*, ed. by Aron Wolfe Siegman and Benjamin Pope, 115–33. New York: Pergamon.

Wurm, Lee H.; Annmarie Cano; and Diana A. Barenboym. 2011. Ratings gathered online vs. in person: Different stimulus sets and different statistical conclusions. *The Mental Lexicon* 6.325–50. DOI: 10.1075/ml.6.2.05wur.

Yow, W. Quin, and Xiaoqian Li. 2018. The influence of language behavior in social preferences and selective trust of monolingual and bilingual children. *Journal of Experimental Child Psychology* 166.635–51. DOI: 10.1016/j.jecp.2017.09.019.

[raczp@ceu.edu]
[clay.beckner@warwick.ac.uk]
[jen.hay@canterbury.ac.nz]
[janet.pierrehumbert@oerc.ox.ac.uk]

[Received 27 April 2016;
revision invited 1 March 2017;
revision received 22 February 2018;
revision invited 15 August 2018;
revision received 7 July 2019;
revision invited 1 November 2019;
revision received 18 December 2019;
accepted pending revisions 1 January 2020;
revision received 28 April 2020;
accepted 29 April 2020]