This is a preprint of a published article which can be found at: https://www.sciencedirect.com/science/article/pii/S0749596X15000790

Please use the full citation as below:

Cuskley, C., Colaiori, F., Castellano, C., Loreto, V., Pugliese, M., and Tria, F. (2015). The adoption of linguistic rules in native and non-native speakers: Evidence from a Wug task. Journal of Memory and Language, 84, 205-223.

Linguistic rule adoption in native and non-native speakers: Evidence from a Wug task

Christine Cuskley[1b,a], Francesca Colaiori[b], Claudio Castellano[b], Vittorio Loreto[c,a,d], Martina Pugliese[c], Francesca Tria[a]

[a]*Instititute for Scientific Interchange, Social Computation Unit, Turin, Italy*
[b]*Istituto dei Sistemi Complessi (ISC-CNR), Consiglio Nazionale delle Ricerche, Rome, Italy*
[c]*University of Rome La Sapienza, Department of Physics, Rome Italy*
[d]*SONY-CSL, Paris, France*

Abstract

Several recent theories have suggested that an increase in the number of non-native speakers in a language can lead to changes in morphological rules. We examine this experimentally by contrasting the performance of native and non-native English speakers in a simple Wug-task, showing that non-native speakers are significantly more likely to provide non *-ed* (i.e., irregular) past-tense forms for novel verbs than native speakers. Both groups are sensitive to sound similarities between new words and existing words (i.e., are more likely to provide irregular forms for novel words which sound similar to existing irregulars). Among both natives and non-natives, irregularisations are non-random; that is, rather than presenting as truly irregular inflectional strategies, they follow identifiable sub-rules present in the highly frequent set of irregular English verbs. Our results shed new light on how native and non-native learners can affect language structure.

Keywords: regularity, morphology, sociolinguistics, language evolution, language dynamics

[1]email: ccuskley@gmail.com

1. Introduction

Learnability is a core property of language (Hockett, 1960; Christiansen and Chater, 2008), and therefore *who* is learning and using a language has the potential to shape and change the language itself. Structural traits which learners internalise and reproduce more accurately will proliferate and persist within a language (Kirby et al., 2008; Cornish, 2010). Thus, *how* learners internalise and reproduce linguistic input can shape the structure of a language.

While the body of research tying language acquisition and language evolution has expanded considerably in the past few decades (Monaghan, 2014), it tends to focus on the role of a specific type of learner: the child. However, historically - and perhaps increasingly in modern times - non-native adult learners have become a force in many languages, most notably in English, where about 70% of speakers are non-native (Dryer et al., 2005). Despite this, little is known about exactly how this shifted learner profile might affect language structure, although differences between child and adult language learners are well-documented (Clahsen et al., 2010).

English is not atypical in this regard, but represents one example of a language in contact, which occurs whenever a population or subset of a population uses more than one language (Thomason, 2001; Weinreich, 1963; Thomason, 2001; Hickey, 2010; Bakker and Matras, 2013). The notion of language contact is broad, including high rates of bilingualism, situations in which a *lingua franca* is needed, and also more extreme cases where entirely new languages are born (e.g., Creoles and Pidgins; Michaelis et al., 2013). This paper will aim to illuminate how the specific case of language contact wherein a language has a high rate of non-native learners may change linguistic structure.

In short time scales, high rates of non-native learners could potentially effect a language system by changing the nature of the "corpus" of a language; in other words, if 70% of speakers are non-native, than some sizeable proportion of written and spoken English will be the direct product of non-native learners. These effects can also span longer timescales, with new learners - both native and non-native - learning at least in part from non-native production. To examine this, we experimentally contrast how native and non-native adults apply simple past-tense inflection to novel English non-verbs, providing a specific experimental investigation of the individual mechanisms underlying patterns of change in languages which undergo prolonged periods of contact. First, we provide a brief overview of what previous research indicates about the effects of language contact on language structure, and the fundamental differences between child and adult language learners.

Generally, an influx of non-native learners in a language seems to lead to a reduction in morphological complexity (Dale and Lupyan, 2012; Lupyan and Dale, 2010; Wray, 2007; Trudgill, 2010), often referred to as *deflexion* (the loss or reduction of morphological marking, often in favour of lexical strategies; Allen, 2003). Broadly, this is analogous to a simplification or elimination of rules, though the mechanisms which cause this type of change are not well understood. As an example, while some languages use complex morphological paradigms to inflect verbs, others have partially collapsed inflections where differences are only retained in the written form, or lack distinct inflections altogether. Table 1 contrasts present tense verb inflection in Italian, French and English, three languages which although typologically close, exhibit differences in their inflectional strategies.

		Italian	French	English
Singular	1st person	io cammin**o**	je march**e**	I walk
	2nd person	tu cammin**i**	tu march**es**	you walk
	3rd person	egli/ella cammin**a**	il/elle march**e**	s/he walk**s**
Plural	1st person	noi cammin**iamo**	nous march**ons**	we walk
	2nd person	voi cammin**ate**	vous march**ez**	you walk
	3rd Person	essi/esse cammin**ano**	ils/elles march**ent**	they walk

Table 1: Present tense verb inflection in Italian, French, and English provides an instructive example of the varying degrees to which different languages utilise inflectional strategies. Both Italian and French derive from Proto-Romance, which made distinctions between each subject type much like Italian; this indicates that these have been lost in French over time (and note that the *tu* form, although it retains a final *-s* in written form, is pronounced identically to the *je* and *il/elle* forms). Likewise, Old English had more specified verb inflection than Modern English, indicating collapse over time.

English verbs are almost completely deflected for person and number, retaining marked inflection only for the third person singular. Italian, on the other hand, has distinct inflections for each subject type. French lies somewhere in the middle, with the *je, tu* (*I ,you(sg)*) and *il/elle* (*he/she*) forms phonologically, if not orthographically, collapsed. Each of these languages has seen varying levels of contact in terms of adult learners: Italian has relatively few non-native speakers, while French spent a long period as a major *lingua franca* (Wright, 2006). As the final extreme, English is considered a modern *lingua franca* on the rise (Seidlhofer, 2001); current estimates indicate considerably more non-native speakers of English than native speakers, and this number is likely growing (Dryer et al., 2005).

This example provides an illustrative anecdote, but stronger signals of this pattern abound throughout natural language (Roberts and Winters, 2012;

Lupyan and Dale, 2010). Many of the changes in English since the Old English period are thought to have been a result of contact (Trudgill, 2010), including the loss of the case system and complex adjectival markers (Lass, 1992). German has seen historically variable levels of contact, which has been reflected in different rates of morphological change over time, particularly for the past tense (Carrol et al., 2012).

Lupyan and Dale (2010) presented one of the first studies to quantify this on a large, cross-linguistic scale. By measuring the degree to which inflectional strategies were employed in thousands of languages, Lupyan and Dale (2010) found that languages with smaller and more isolated populations tend to use more complex morphological inflection, while languages with larger population sizes tend towards lexical strategies. They interpret this result specifically in terms of contact, assuming that languages with larger population sizes are by definition more prone to contact, and therefore, have more non-native adult learners. In a more recent experimental study, Dale and Lupyan (2012) showed that native speakers of American English living in areas with a larger non-native speaker populations preferred regularly inflected verb forms (e.g., *sneaked*) to irregularly inflected ones (e.g., *snuck*). This study demonstrates that non-native learners have the potential to affect a language both through direct production and influencing the preferences of native speakers.

However, specific, concrete experimental evidence for the effect of learner profiles on natural language structure is lacking. Lexical decision and priming studies indicate that a key difference in processing between native and non-native users is in the level of rule application: non-natives never attain the automaticity and accuracy at implementing grammatical and morphological rules that comes naturally to native speakers (Clahsen et al., 2010). In other words, this evidence might predict that non-natives process primarily on a lexical level, by simply memorising different word forms (contrasted with a morphological level, where rules are applied to roots to realise word forms). Non-native adult learners display an imperviousness to internalising and correctly applying rules automatically, while a native child learners do so effortlessly (Clahsen and Felser, 2006).

Artificial language learning (ALL) studies in children and adults can also inform hypotheses about mechanisms underlying the relationship between population structure and social structure. In these studies, participants are tasked with learning and reproducing small, artificial vocabularies. This gives a controlled set of input/output which allows for measures of (i) accurate learning, (ii) qualitative details regarding failures to reproduce structure present in the input, and/or (iii) the generalisation of structure present in the input or innovation of entirely new structure.

Hudson Kam and Newport (2005) trained both children and adults on partially rule-governed (compositional) artificial languages, and found that children eliminate variation and engage in regularization as part of reproducing the artificial languages. Adult learners, on the other hand, were more adept at reproducing input more accurately, perhaps as a result of more completely learning the system. In other words, adults reproduced variation present in their input more faithfully, rather than generalising over items in a rule-like way (see also Kam and Newport, 2009). This perspective is reinforced in part by the U-shaped learning curve observed in children, wherein they engage in a period of production where over-regularisation is particularly prevalent (e.g., *goed* instead of *went*; Maslen et al., 2004; Gershkoff-Stowe and Thelen, 2004). Wonnacott et al. (2013) also found that children engage in more over-generalisation than adults, particularly for low-frequency items.

But other ALL studies have found results showing the opposite: that adults prefer regularity more than children, and thus, predict that adult non-native learners would prefer regularly inflected forms. Boyd and Goldberg Boyd and Goldberg, 2012 show that young children (approximately 5 years old) are more conservative than older children and adults when it comes to extending rules to novel constructions. Other studies have shown that adults are adept at generalisaton as long as they are able to "start small" (i.e., observe only a small subset of the language prior to test; Kersten and Earles, 2001). Moreover, iterated ALL studies have shown that adults do generalise and introduce structure, but that this may not be measurable on an individual time scale, since cultural transmission is key to amplifying structure and eliminating unpredictable variation over time (Kirby et al., 2008; Cornish, 2010; Smith and Wonnacott, 2010; Reali and Griffiths, 2009). These results indicate that adults are at least as adept as young children, if not more, at extracting rules from minimal input data and generalising these rules to novel constructions.

In summary, there is ample evidence showing that children and adults internalise, process, and reproduce linguistic input differently. Yet, exactly how differences between child native learners and adult non-native learners may drive changes in linguistic structure is less clear. In other words, what do adult, non-native speakers *do* to a language? Some evidence indicates that adults prefer regularity while child learners cope more readily with irregularities (Wray, 2007), even to the point of introducing irregularity in a highly regular language (e.g., in Esperanto; Bergen, 2001). Some ALL studies show that children preserve irregular variation (Boyd and Goldberg, 2012), but other ALL studies seem to indicate that children eliminate irregular variation while adults preserve it (Hudson Kam and Newport, 2005; Wonnacott et al., 2013), although this could merely be an artefact of more effective adult learning in the ALL context. Evidence from non-native language processing shows that rather than having a preference for rules, non-native speakers tend not to use rules when realizing inflected word forms, and this strategy is frequency sensitive (Clahsen et al., 2010).

While many theories predict a trend of simplification and regularisation in language as a result of non-native adult learners, some evidence from non-native language processing and ALL studies predicts that adult non-native learners may preserve or even introduce irregular variation. This may occur because adult learners are heavily influenced by the generally high token frequency of irregular verbs (Bybee, 2001; Cuskley et al., 2014; Lieberman et al., 2007), and/or because they treat each past-tense form as a new lexical item, rather than generalising across forms and applying rules.

To address the broad question of exactly how the alteration of learner profiles through contact may contribute to change in language structure, we aim to contrast the behaviour of natives and non-natives in a simple experiment involving past-tense inflection, modelled after the well-known Wug-task (Berko, 1958). This task centers around providing participants with a nonsense word to elicit an inflected form. In contrast with many other reports of Wug-style experiments, we focus here on the *irregularisation* behaviour of participants. Historically, Wug tasks have been used to demonstrate how learners generalise rules, and thus focused primarily on regularisation behaviour Berko (1958). However, several notable studies have also shown irregularisation behaviour to some extent, indicating that irregularisation is not exactly rare - in some studies, native English speakers irregularise certain non-verbs at rates up to 40% Albright and Hayes (2003). These studies provide some experimental evidence of irregular groups or quasi-regularity (Bybee and Moder, 1983) : often, when participants provide irregular forms, they are modelled after existing irregulars in English (e.g., *dize/doze*, Albright & Hayes, 2003)

To examine (ir)regularization behaviour in light of nativeness, Experiment 1 asks if there are differences between natives and non-natives regarding the rate at which they inflect novel words irregularly (i.e., using a non *-ed* form), and if the phonological form of a novel verb has different effects on irregularisation rates in natives and non-natives. Experiment 2 extends this by examining in more detail how both natives and non-natives irregularise, demonstrating that although irregular forms do not follow “the” regular rule by definition, they are also not entirely random; rather, they are extensions of irregular sub-rules already present in English.

2. Natives versus non-natives in a past-tense Wug task

2.1. Experiment 1

2.1.1. Methods & Materials

Non-word prompts for the Wug-task were selected based primarily on their phonological similarity to existing English verbs (170 irregular verbs and the 500 most frequent regular verbs in the Corpus of Contemporary American English; Davies, 2014). Both previous past-tense Wug-style experimental studies (Albright and Hayes, 2003; Prasada and Pinker, 1993; Bybee and Moder, 1983) and corpus data (Cuskley et al., 2014) indicate that the phonological properties of non-words can have a crucial impact on regularity. We used phonological feature-based distance (using an 12 feature vector adapted from Nerbonne and Heeringa, 1997) to choose our non-word stimuli; further details are provided in Appendix A.

Using the phonological segments contained in the 670 verbs mentioned earlier and a consonant-vowel-consonant (CVC) syllable template, we generated an exhaustive list of non-words with each consonant onset and coda in the C position and each vowel in the V position. Many of these automatically generated words were immediately discarded as either existing verbs or phonotactically impossible non-words (e.g., anything with /ŋ/ in the onset position). The remaining words were assigned a phonological distance from their closest real regular and irregular verb. These words were categorised as either close to an irregular verb and distant from the closest regular (*irregular* non-words), close to a regular verb and distant from the closest irregular (*regular* non-words), or equally close to both a regular and irregular real word (*intermediate* non-words). Of this exhaustive list, 68 verbs (29 regular non-words, 29 irregular non-words, and 10 intermediate non-words; provided in Appendix B) qualified and were used for Experiment 1. Words were presented both in their written forms and as audio files generated using text-to-speech software (further details provided in Appendix B).

Participants were recruited through Amazon's Mechanical Turk, shown to be an effective tool for conducting psychological experiments (Paolacci and Chandler, 2014). The task involved completing a simple Wug-task (after Berko, 1958) through an online JavaScript applet, hosted on the Xtribe experimental platform (Cicali et al., 2011). Participants were provided with a link to the applet through Mechanical Turk , and after completion, were provided with a code to enter on Mechanical Turk itself to ensure both honest participation and timely compensation. In this experiment, participants were paid $0.15. For this fee, the participant had to complete at least one word, but could complete additional words if they chose (as in e.g., Cuskley, 2013). This means that for the preliminary experiment, each participant responded to anywhere between 1 and 68 words. Figure 1 shows the distribution of the number of responses across participants.

The task was briefly described on a splash page where participants completed an audio captcha to ensure their sound was functioning, and consented to continue to the experiment itself. The XTribe platform generates a random identifying string for each participant based on their IP address. This string allowed for the experimenters to prevent duplicate participation without having access to participants' identifying information. This random ID string was the only information stored and used to identify participants, simultaneously ensuring privacy and efficient handling of data.

Following correct completion of the captcha and consent, participants answered two simple questions about their language knowledge: (1) "is English your first language?", and (2) "is English the only language you speak?". If participants were monolingual English speakers (i.e., the answer to both questions was "yes"), they provided no further information[2]. If English was *not* the participant's first language, they provided a self-report of the age at which they learned English (Age of Acquisition, hereafter AoA) as well as a self-rated measure of English proficiency on a sliding scale of 0-100 (0=Beginner, 100=Fluent). If the participant also spoke other languages in addition to English (always true for non-natives, but only the case for a minority of bilingual natives), they provided up to two other languages they spoke best along with self-rated proficiency for each.

After answering these preliminary questions, participants were directed to the task which included detailed instructions (optionally hidden and pulled up at any time during the task). These instructions detailed that the words were ones the participants had never seen before, and that it did not matter what they meant, but the experimenters were interested in their intuitions on the past tense. In the instructions, particular stress was placed on the fact that "there was no right or wrong answer" and their "personal intuitions about language" were of particular interest. For each non-word, participants were given the prompt "Every day we [non-verb]", and asked to complete the sentence "Yesterday, we..." using the non-word provided. Non-words

[2]Note that this means we did not distinguish between different varieties of English (e.g., British, American, etc.). There is some evidence that different varieties of English differ subtly in verb regularity (e.g., see Michel et al. (2011)), but our aim here was to consider the broader metric of nativeness.

were presented in written form as well as a synthesised audio file using even stress and a male voice. Participants could not respond to a non-verb prompt without first listening to the audio file of the infinitive form of the non-word at least once. Items were presented in a random order to all of the participants, and participants could complete as many words as they wished. After task completion, a link was provided for all participants to debrief on the purpose of the experiment and contact the experimenters directly with any queries.

Each response was coded as either regular, irregular, or entirely invalid. The criteria for an invalid response was either that participants provided an existing past-tense form (e.g., *swin-swam*), or provided an existing word that clearly ignored the prompt (e.g., *swin-play*). Existing past-tense forms were eliminated to ensure that participants were responding to the non-word prompt, rather than directly to the non-word's closest real word neighbour. Regular forms were any *-ed* form with no other change to the word, with the exception of stem-final consonant gemination, which is a standard form of inflecting for the past tense in English orthography (e.g., *step-stepped*). In other words, forms such as *queted* and *quetted* were both considered regular. In some cases, the absence of consonant gemination could be interpreted as an irregularisation; for example, *swin-swined* could be assumed to involve a vowel change, while *swin-swinned* would be the "correct" regular form. However, in an effort to code responses conservatively without pronunciation directly from participants, these forms were considered regulars.

2.1.2. Participants

A total of 589 participants contributed to Experiment 1, giving a total of 1811 responses. Of these, 103 responses were considered invalid; 29 of the invalid responses were the only response given, leaving a total of 560 valid respondents (406 native and 154 non native) and 1708 responses for analysis (1196 native responses and 512 non-native responses). Figure 1 shows the distribution of the number of responses per participant, with the large majority of participants responding to between 1 and 5 non-verbs.

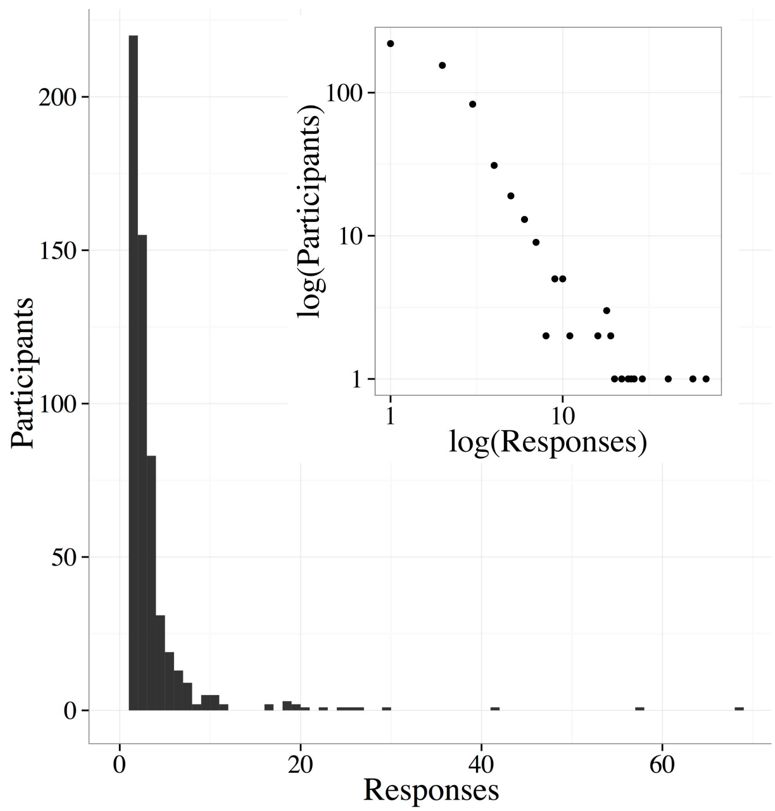


Figure 1: Number of responses per participant in Experiment 1.

Figure 2 provides a breakdown of the first languages (L1s) represented among the participants, as well as how many responses were contributed by speakers of each language in total.

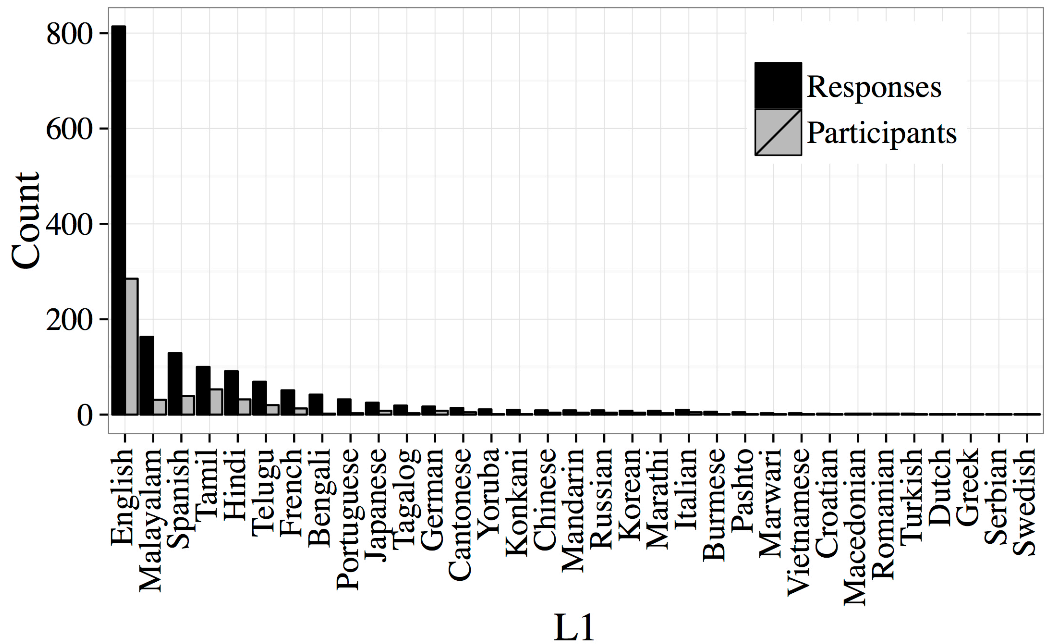


Figure 2: First languages represented among participants in Experiment 1.

2.1.3. Results & Discussion

Table 2.1.3 summarises the overall results in terms of nativeness, stimuli type, and word type.

The table provides both raw counts and percentages of natives and non-natives, defined as the proportion of irregular responses in a group relative to the total number of responses. Overall, the irregularisation rate of non-natives (35.7%) was almost twice that of natives (21.6%). In other words, non-natives were more likely to provide an irregular form than natives. The type of non-word also had some effect on the irregularisation rate. Figure 3 shows irregularisation rates in terms of both nativeness and stimuli type.

Nativeness	Stimuli Type Regular	Intermediate	Irregular	Total
Native	544 [468, 76]	170 [142, 28]	482 [328, 154]	1196 [938, 258]
Non-Native	227 [162,65]	68 [42, 26]	217 [125, 92]	1317 [781, 536]

Table 2: Results from Experiment 1. Total number of responses for each category are indicated with the total count and breakdown of Regular, Irregular] in brackets. Proportions are displayed in Figure 3.

Irregular stimuli - non-words phonologically close to an existing irregular form, and also far from frequent regulars - had the highest irregularisation rate. Regular stimuli had the lowest irregularisation rate, while intermediates fell somewhere in between for both groups. Results for these different non-verb categories replicate earlier studies which show that irregular-like non-words have the potential for higher rates of irregularisation (Albright and Hayes, 2003; Prasada and Pinker, 1993; Bybee and Moder, 1983), showing no drastic differences in this regard between natives and non-natives.

To assess the potential significance of these differences, a mixed-effects logit model was performed on the results. Mixed logit regression models are well-suited to analysing categorical data which generalises beyond subjects and items (Baayen et al., 2008), and are also equipped to deal with unbalanced designs (in this case, the optional number of items completed by each participant and the differing number of native and non-native participants; Jaeger, 2008). Mixed logit regression models return a coefficient estimating the log-odds for each contrast in the model, eliminating the need for post-hoc tests and planned contrasts (Arnon, 2010). In the following models, significant positive log-odds coefficients show that a regularisation is more likely

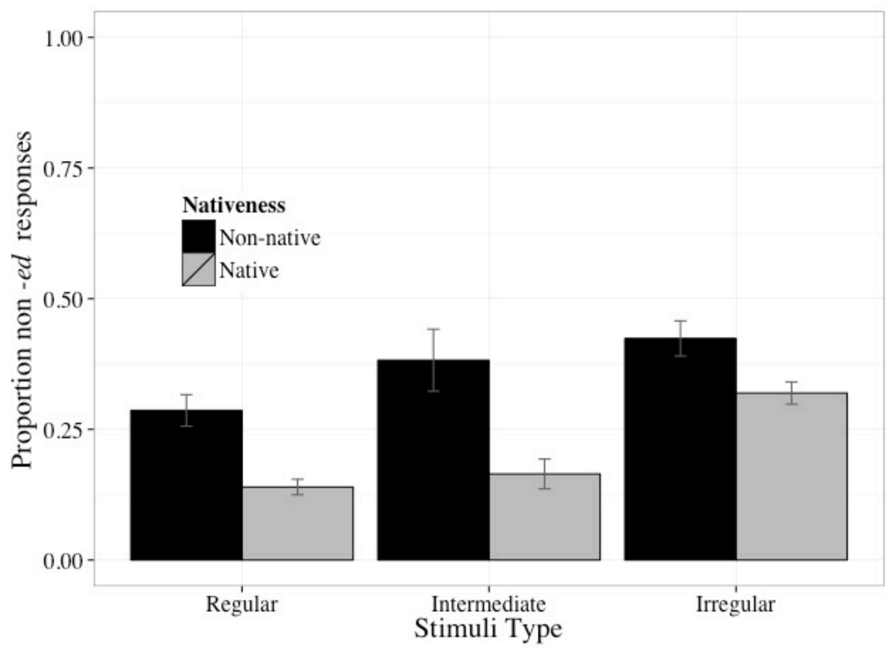


Figure 3: Irregular and regular responses in different stimuli categories for native and non-native particpants. Bars represent Standard Error.

in one relevant level of an independent variable than another. For example, a positive log-odds coefficient for natives shows that they are more likely to provide a regular form in response to non word than non-native participants; on the other hand, a negative log-odds coefficient would indicate that natives are less likely to provide a regular form than non-natives.

A mixed effects logit regression model with regularity of the response (regular [-*ed*] form vs irregular [non-*ed*]) as the outcome variable and nativeness (native or non-native) and non-word category (regular, irregular, or intermediate) included as fixed effects (with participant, response number, and non-word as random effects) was run, and results are presented in Table

3. For details on model selection, see Appendix C.

Table 3: Summary of fixed effects in mixed logit model for Experiment 1 (N=1708, log likelihood = -868.9). The intercept represents the log-odds of an irregular response for the reference values (in this case, native participant with an regular item). The estimate or β coefficient represents the increased (positive) log-odds or decreased (negative) log-odds of an irregular response relative to the reference values. SE and CI represent the standard error and confidence interval of the β value. The Wald's Z and p-values are obtained by dividing the β estimate over the SE, providing a normal distribution from which the p-values are derived. These values represent the probability of obtaining the observed estimate or a more extreme one, given the true estimate is 0 (i.e., given the null hypothesis that a change in nativeness or item category has no effect on the regularity or irregularity of response). The OR column indicates the Odds Ratio, an exponent of the β coefficient.

Predictor	β Coef.	SE	CI (95%)		Wald's Z	p	OR
			2.5%	97.5%			
Intercept	-2.43	(0.234)	-3.29	-1.57	-10.4	<0.001	0.08
Non-native	1.22	(0.242)	0.32	2.13	5.07	<0.001	3.40
Irregular	1.20	(0.237)	1.14	1.25	5.07	<0.001	3.32
Intermediate	0.225	(0.332)	-0.67	1.12	0.678	0.498	1.25

The model shows that both nativeness of the respondent and the type of stimuli are significant predictors of whether a non-word will be regularised or irregularised: non-native speakers are significantly more likely to irregularise novel non-verbs than native speakers. The stimuli type was also a significant predictor of irregularisation: irregular type non-words (closer in phonological form to existing irregulars) were more likely to be irregularised than regular type non-words across both participant groups. Although the odds of irregularising an intermediate item are slightly higher than for a regular item, this difference is not significant. In other words, intermediates seem to act much like regulars, while irregular items increase the overall odds of irregularisation significantly. A model which included interaction terms did not provide significantly better fit, and did not result in any meaningful interactions between nativeness and word type (see Appendix C).

To examine the effect of self-reported proficiency among non-natives, we ran another model with native participants removed, with non-word category and self-reported proficiency as predictor variables[3] (N=512, log likelihood = -315.8). Non-word category remained a significant predictor of whether a regular or irregular form was provided (see Appendix Appendix C for details), but this model also showed that self-reported proficiency is a good predictor of the likelihood of irregularisation among non-natives ($\beta = -0.025$, SE = 0.007, CI = -0.08 – -0.03, Wald's Z = -3.616, $p < 0.001$, OR = 0.97). Given the continuous nature of the proficiency predictor, the OR means that with every unit increase in proficiency, the odds of an irregular response decrease slightly[4]. In other words, participants who provided irregular forms were likely to have lower self-rated proficiency than those who provided regular forms (Figure 4).

In summary, this experiment revealed two main findings. First, as earlier studies have also shown, the phonological character of non-words is important: across both natives and non-natives, words which are phonological neighbors with existing irregulars are much more likely to elicit irregular forms than forms which also have a close regular neighbor, or only have a

[3]We also tested a model with age of acquisition as a predictor, but this was not significant on its own or combined with proficiency, and resulted in a significantly inferior fit in either case. See Appendix Appendix C for details

[4]Note that this means that while still significant, the β and OR values are much lower than for significant categorical predictors. This is because for a continuous predictor or fixed effect, change in odds applies to every *unit* of change in the predictor. In other words, since a single unit of increase in proficiency decreases the log odds of an irregular response by almost 1, five units increase in proficiency would decrease the log odds of an irregular response almost five fold.

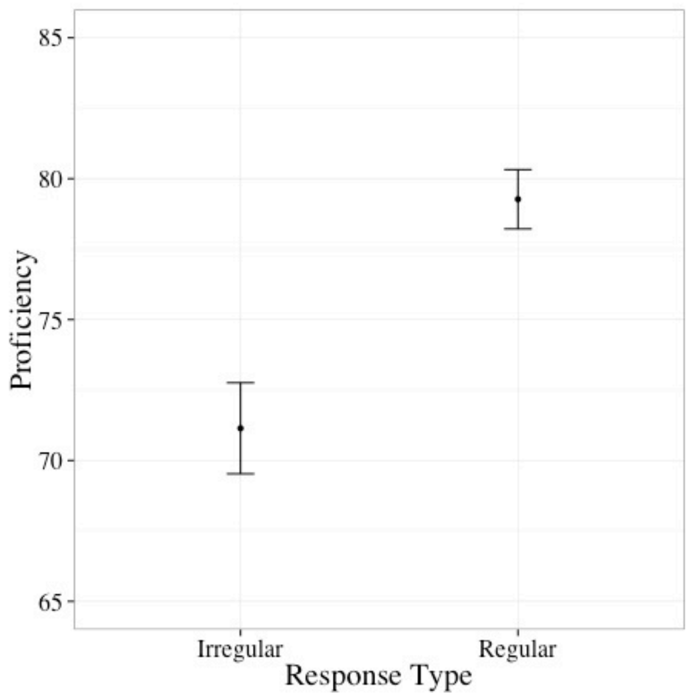


Figure 4: Proficiency and response type among non-natives. Regular responses were associated with higher proficiency overall, indicating that an increase in proficiency decreases the likelihood of a regular response. Bars represent standard error.

regular neighbor. Accordingly, non-words which have close, highly frequent regular neighbors are more likely to elicit a regular past tense form. Natives and non-natives did not act significantly different in this regard.

Second, there is an evident difference in regularisation behaviour between native and non-native speakers of English. Native speakers show an overall preference for the regular *-ed* rule, while non-natives are more likely overall to provide irregular forms. Furthermore, among non-natives, self-reported proficiency is a good predictor of whether a participant is more likely to provide a regular or an irregular form, with less proficient speakers being more likely to irregularise. This result runs contrary to what many theories of contact might predict: the high contact environment of English should result in a growth of the regular rule driven in particular by over-regularisation of non-native speakers. However, our results show the opposite for a set of novel words: non-natives are more likely to provide irregular forms while natives show higher odds of regularisation.

These results suggest that by irregularising more, non-natives are expanding or complexifying the rule set, rather than collapsing it or simplifying it as contact-deflexion theories might predict. Experiment 2 will examine in more detail exactly *how* non-natives are irregularising. By using a confined set of items, we show not only that the results of Experiment 1 generalise to a different sampling method, but we are able to examine in greater detail exactly how both natives and non-natives irregularise.

2.2. Experiment 2

2.2.1. Methods & Materials

Experiment 2 used the same methodology as Experiment 1, only each participant completed a total of fifteen non-word items: five words from each non-word category (regular, intermediate, and irregular). The subset of items used for Experiment 2 are highlighted in Appendix B. Items were presented in a random order for each participant, and their progress during the task was shown using a percentage bar at the bottom of the screen. There was no time limit to the task, but participants generally completed all 15 items within 5-10 minutes. After task completion, a code was shown for Mechanical Turk participants to enter in the Mechanical Turk interface, and a link was provided for all participants to debrief on the purpose of the experiment and contact the experimenters directly with any queries.

2.2.2. Participants

In order to widen the participant base for Experiment 2, participants were recruited both through Amazon's Mechanical Turk (paid $1 to complete all 15 items) and through volunteers on social networks such as Facebook and Twitter. A total of 210 participants completed the task[5]: 102 from Mechanical Turk (87 native and 15 non-native), and 108 volunteers (34 natives and 74 non-natives)[6].

Figure 5 provides a breakdown of the first languages (L1s) represented among the participants in Experiment 2.

2.2.3. Results & Discussion

A total of 3150 responses were collected. A total of 38 responses were invalid (for natives, 18 invalid responses total with [3,5,10] for [regular, intermediate, irregular] stimuli types; for non-natives, 20 invalid responses total, [6,7,7]), leaving a total of 3112 responses. The criteria for regular, irregular, and invalid were the same as for Experiment 1. Table 2.2.3 shows responses in terms of nativeness, stimuli category, and regularity.

As with Experiment 1, non-natives showed a higher irregularisation rate than natives (Table 2.2.3), and different stimuli types also resulted in markedly different irregularisation rates. Figure 6 shows irregularisation rates by na-

[5]If a participant provided more than three invalid responses of the 15, all of their responses were automatically removed from the sample.

[6]Using the XTribe participant ID, we were able to ensure that participants from this round had not completed the first experiment or any associated pilots, that volunteers had not also completed the task on Mechanical Turk (or vice versa), and that volunteers did not complete the task multiple times.

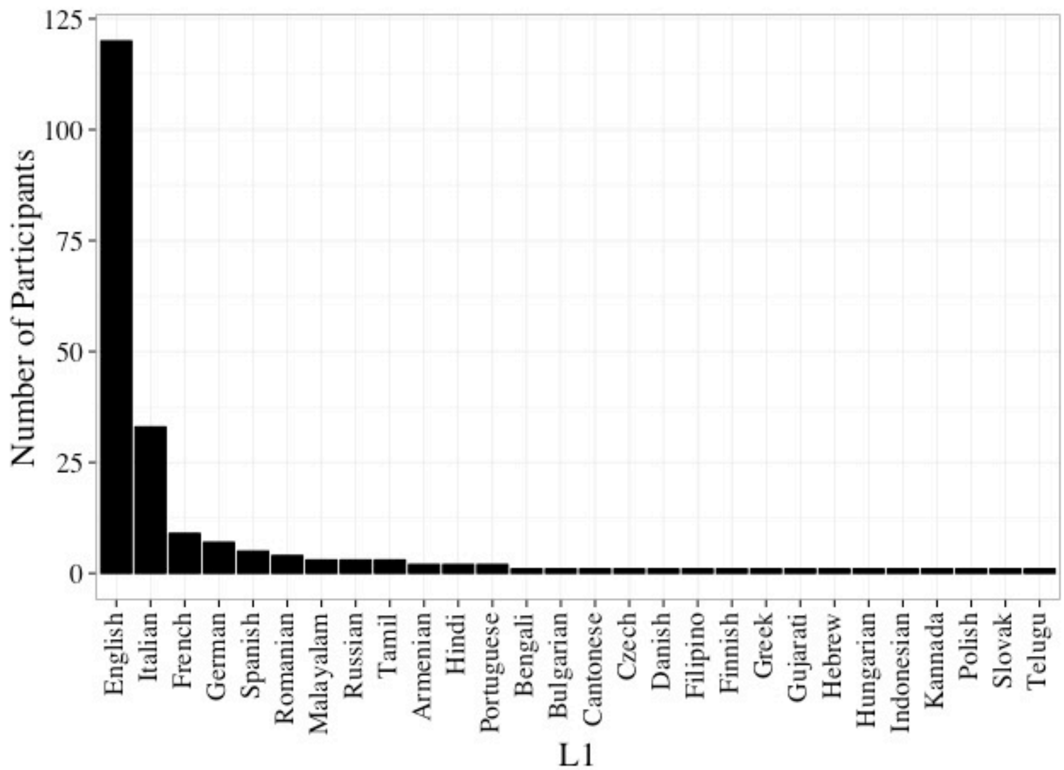


Figure 5: First languages represented among participants in Experiment 1.

tiveness and stimuli type. Relative to Experiment 1, overall irregularisation rates were higher, with non-natives irregularising well over 50% of irregular type items.

A mixed effects logit regression model with regularity of the response as the outcome variable and nativeness and non-word category (regular, irregular, or intermediate) included as fixed effects (with participant and non-word item as random effects) provided the best fit (see Appendix C for details on model selection). Table 5 shows the coefficients of the fixed effects, their 95%

	Stimuli Type			
Nativeness	Regular	Intermediate	Irregular	Total
Native	599 [538, 61]	598 [484, 114]	598 [340, 258]	1795 [1362, 433]
Non-Native	442 [348, 94]	435 [276, 159]	440 [157, 283]	1317 [781, 536]

Table 4: Results from Experiment 2. Total number of responses for each category are indicated with the count and breakdown of [Regular, Irregular] in brackets. Proportions are displayed in Figure 6.

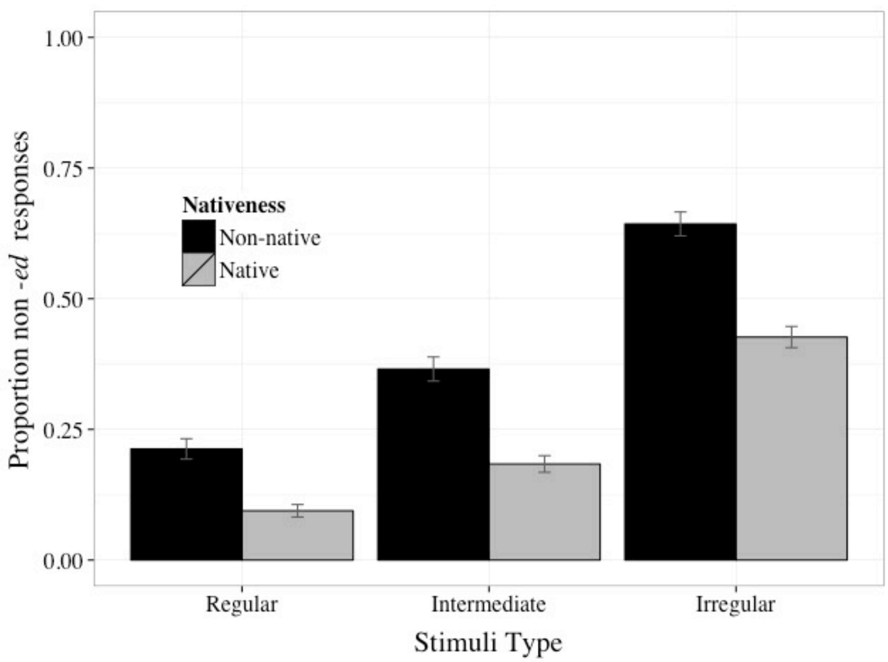


Figure 6: Irregularisation rates of natives vs non-natives in terms of stimuli type. Bars represent Standard Error.

CIs and significance based on Wald's Z (after Jaeger, 2008).

Regular items increased the likelihood of a regular response drastically, and intermediate items more than doubled the odds of a regular form (Table 5). Re-leveling the model for Experiment 2 also showed a significant differ-

Table 5: Summary of fixed effects in mixed logit model for Experiment 2 (N=3112, log likelihood = -1505). The intercept represents the log-odds of an irregular response for the reference values (in this case,native participant with an regular item). The estimate or β coefficient represents the increased (positive) log-odds or decreased (negative) log-odds of an irregular response. SE and CI represent the standard error and confidence interval of the β value. The Wald's Z and p-values are obtained by dividing the β estimate over the SE, providing a normal distribution from which the p-values are derived. These values represent the probability of obtaining the observed estimate or a more extreme one, given the true estimate is 0 (i.e., given the null hypothesis that a change in nativeness or item category has no effect on the regularity or irregularity of response). The OR column indicates the Odds Ratio, an exponent of the β coefficient.

Predictor	β Coef.	SE	CI (95%)		Wald's Z	p	OR
			2.5%	97.5%			
Intercept	-2.87	(0.289)	-1.34	-0.22	-9.92	<0.001	0.06
Non-native	1.20	(0.194)	0.29	2.10	6.17	<0.001	3.31
Irregular	2.46	(0.361)	1.56	3.35	6.814	<0.001	11.67
Intermediate	0.945	(0.360)	0.89	1.00	2.628	<0.01	2.57

ence between intermediate items and irregular items (an intermediate item gave reduced odds of a regular response, β = -0.94 (SE = 0.359), CI (95%) = -1.65 – -0.25, Wald's Z = -2.16, p = 0.009). In other words, intermediates acted somewhere between the two other word types, increasing the odds of an irregular response over regular stimuli, but still giving lower odds of an irregular response than irregular stimuli.

Nativeness also played an influential role in the likelihood of regularisation. Natives were more likely to provide a regular response than non-natives, reflective of the higher irregularisation rate for non-natives overall (Figure 6). This higher irregularisation rate among non-natives is also reflected in the number of irregular forms provided per participant. Figure 7 shows the overall number of irregularisations per participant in terms of nativeness. Native speakers peak around 1-2 irregularisations across all 15 items, while non-natives peak around 6 irregularisations across all 15 items.

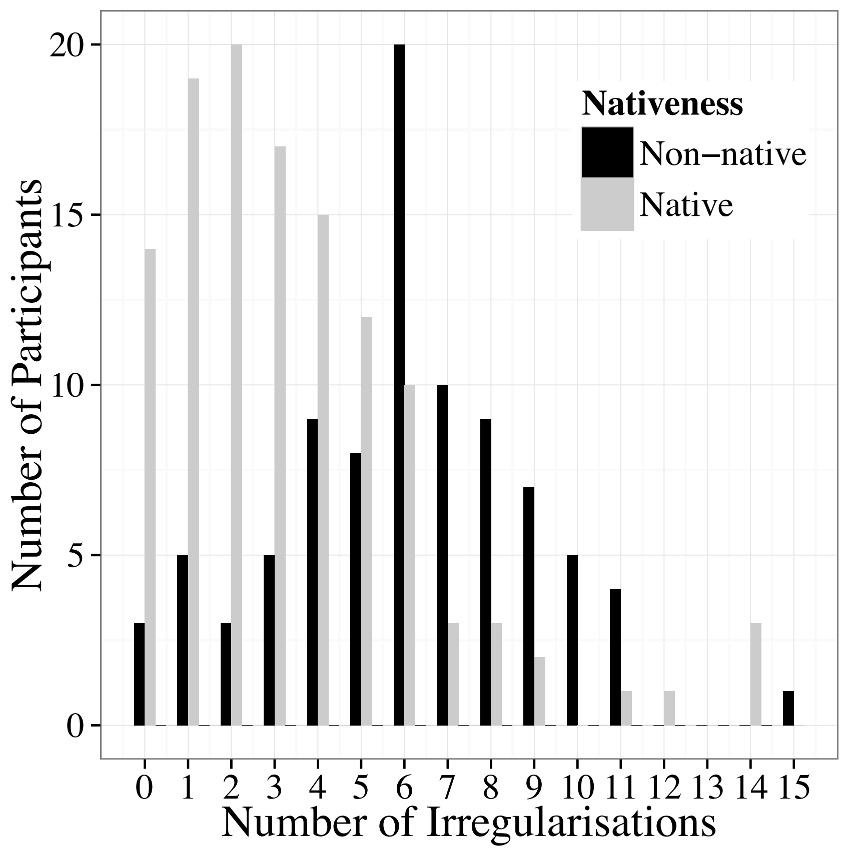


Figure 7: Number of irregularisations per participant by nativeness.

While in Experiment 1 proficiency was an influential factor in irregularisation among non-natives, Age of Acquisition showed similar significant effects in Experiment 2 (see Appendix Appendix C for full model and comparison with proficiency). We ran a model using data from non-native participants only (N=89) with response type as the outcome variable and AoA and item type (regular, intermediate, irregular) as predictors. In this experiment, each unit increase in AoA (i.e., each year later that a participant reported having started to study English) resulted in a decreased likelihood of providing a regular form (β = -0.08, SE = 0.029, z = -2.632, $p < 0.01$ OR = 0.93). In other words, the older a non-native participant was when they started learning English, the more likely they were to provide an irregular form (Figure 8).

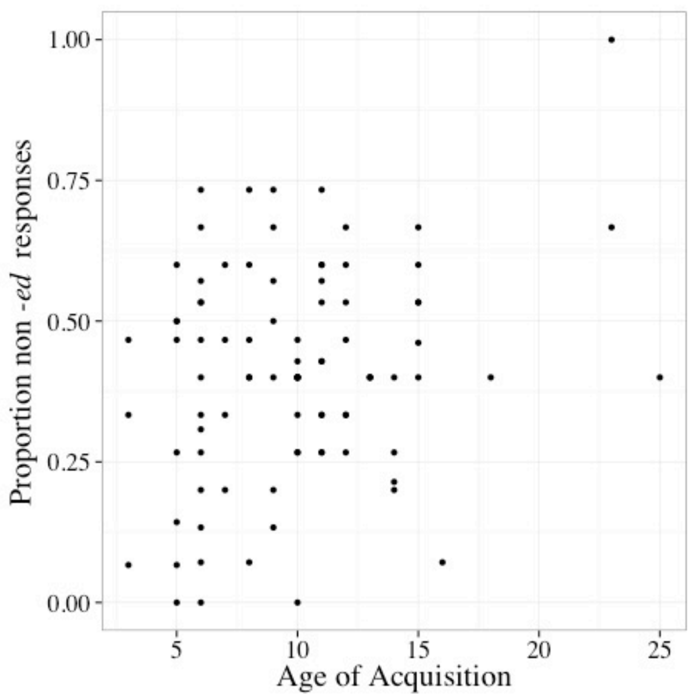


Figure 8: Age of acquisition and rate of irregularisation. Participants who reported learning English later in Expeirment 2 had higher rates of irregularisation.

Given that each participant responded to all items in this experiment, there was greater potential for the data to show specific effects of different first languages represented among the non-native speakers. In an attempt to examine this, we categorised different first languages based on particular features found in the WALS database (Dryer et al., 2005) including the presence or absence of past tense, degree of suffixation, the use of suppletion in verb forms, and the presence or absence of multiple forms of regularity. These features were not predictive of irregularisation rates in our data. For some features, this was likely due to almost total homogeneity in the represented langauges (e.g., all but two of the langauges, Indonesian and Chinese, have past tense). For others, this was likely due to missing data; for example, verb suppletion seemed a promising feature, but data on this feature was missing for six of the languages in our sample.

The reduced number of stimuli in Experiment 2 allow for a more informative qualitative analysis of irregularisations. In other words, we can take a close look at exactly what participants are doing when they provide a non *-ed* past tense form for a novel verb. The first observation is that the large majority of irregularisations across all participants and stimuli categories adhered to recognisable irregular "rules" present in English, as described in Table 6.

Table 6: Summary of types of irregular categories observed in responses and their equivalents in English.

Category	e.g., English	e.g., Experiment	Count	% of Non -ed Responses
Vowel Change	sing-sang	sleen-slen	590	0.62
Level	cut-cut	sleen-sleen	181	0.19
Vowel Change + d	hear-heard	sleen - slinned	52	0.05
Vowel Change + t	dream-dreamt	sleen-slent	28	0.03
Weak	dwell-dwelt	sleen-sleent	36	0.04
Ruckumlaut	teach-taught	sleen - slaught	13	0.01
Other	—	sleen-slonk	54	0.06

In other words, participants' irregularisations were not completely random, and do not introduce much new variation. Most non -ed responses involved verb-internal vowel changes as found in many English irregular verbs (*bear, feed, hide*, etc.), likely largely due to the fact that for most stimuli, the nearest neighbor irregular form involved a vowel change. The majority of the remaining irregularisations could also be categorised according to other patterns found in English irregular verbs. Figure 9 shows that the contributions to these sub-categories were distributed across participants; in other words, non *-ed* forms were not introduced by some small minority of participants, rather, many participants contributed to the most productive non *-ed* categories[7].

The few exceptions (54 responses in total, classified as "other" from now on), represent only about 1.5% of all responses and 5% of irregular responses. Of the "other" responses, several were mistaken irregular past participle forms rather than simple past tense forms (e.g., *cluse/clusen*, following the irregular past participle pattern in e.g., *prove/proven*), while others were "true" irregulars following patterns generally not found in English verb inflection (e.g., *thring/thronk*). This use of "irregular rules" was evident not only across participants, but within participants. Figure 10 demonstrates this by showing that although the number of non-native irregularisations peaked at 6 (see Figure 7), the number of irregular *categories* for non-natives

[7]There may be some natural concern regarding the level category, which involves no change to the non-word stimuli provided, and could indicate that a participant was simply ignoring the task. However, note that this is a past-tense formation strategy in English (potentially even a productive one in the recent past, see Cuskley et al., 2014), and did not proliferate within an individual participant (i.e., most participants provided only 1-2 level responses). Furthermore, although the level category was fairly prolific overall, our main findings (effects of stimuli type and nativeness) survive the removal of these items entirely.

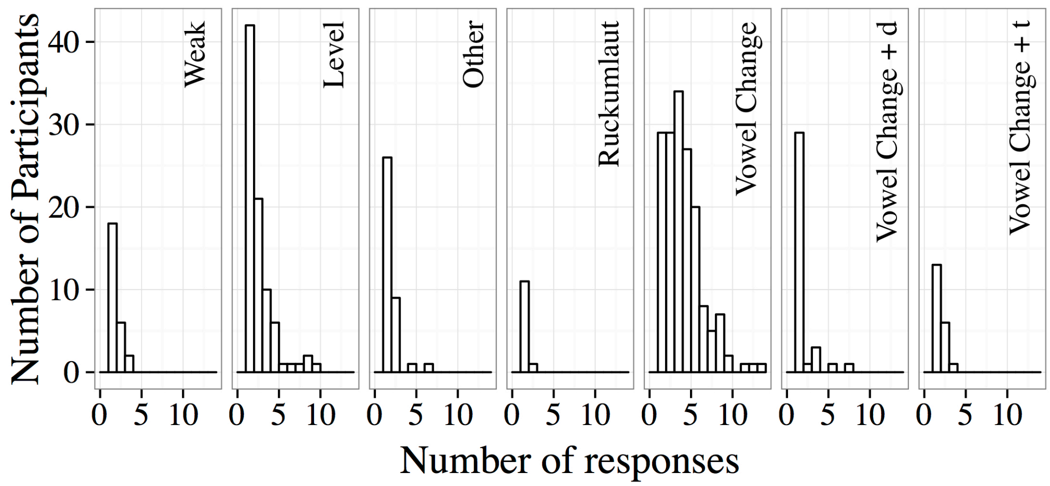


Figure 9: Histogram showing how many participants contributed a given number of responses to each non *-ed* category. This demonstrates that categories do not appear productive due to the contribution of a few individuals, and some categories (e.g., level, vowel change) appear to be productive both across and within participants.

peaks at 2 (Figure 10).

This indicates that a given participant applies sub-rules across irregular responses. In other words, although they are not following "the" regular rule, their non-ed responses are largely governed by sub-rules. The contrast between the number of irregularisations and the number of irregular categories provides a preliminary representation of this: Figure 11a shows that the number of categories used by a participant is sub-linear with respect to the number of irregularisations. However, with fewer irregularisations overall, the opportunity for natives to apply sub-rules to irregulars is generally reduced. To account for this, an entropy measure, S, for each participant j, was defined as follows: Given n_j as the total number of irregularisations for

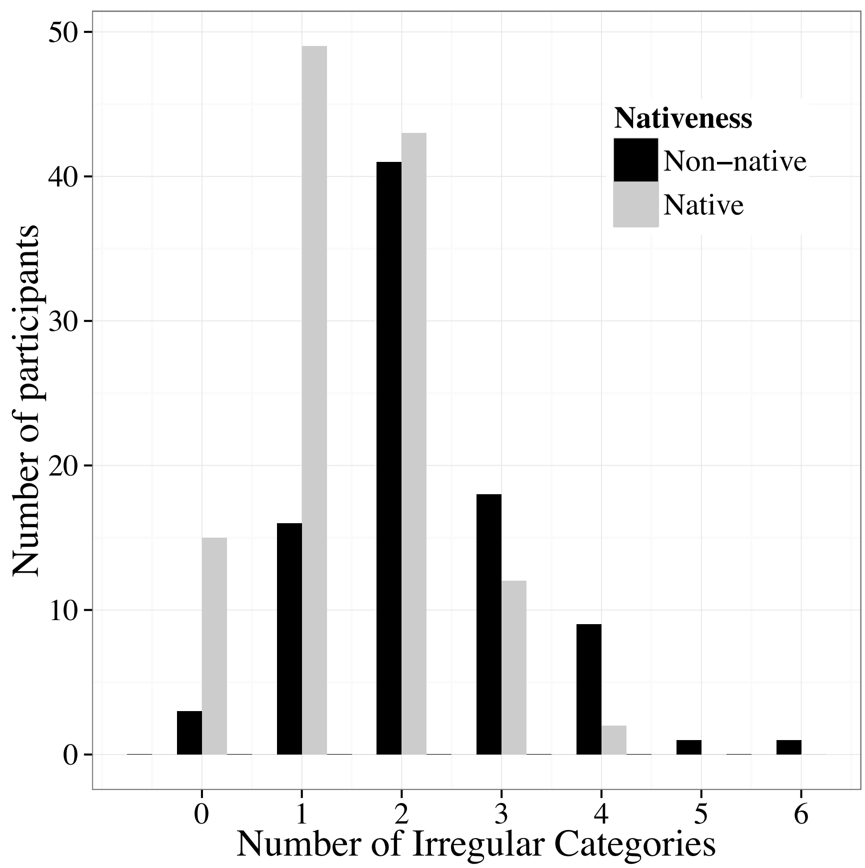


Figure 10: Number of distinct non -ed categories per participant. Native participants peak at 1 category, in line with the general peak of 1 irregularisation for natives overall (Figure 7). Non-natives peak at two categories, despite peaking at 6 non -ed responses overall, indicating the use of sub-rules across irregular forms.

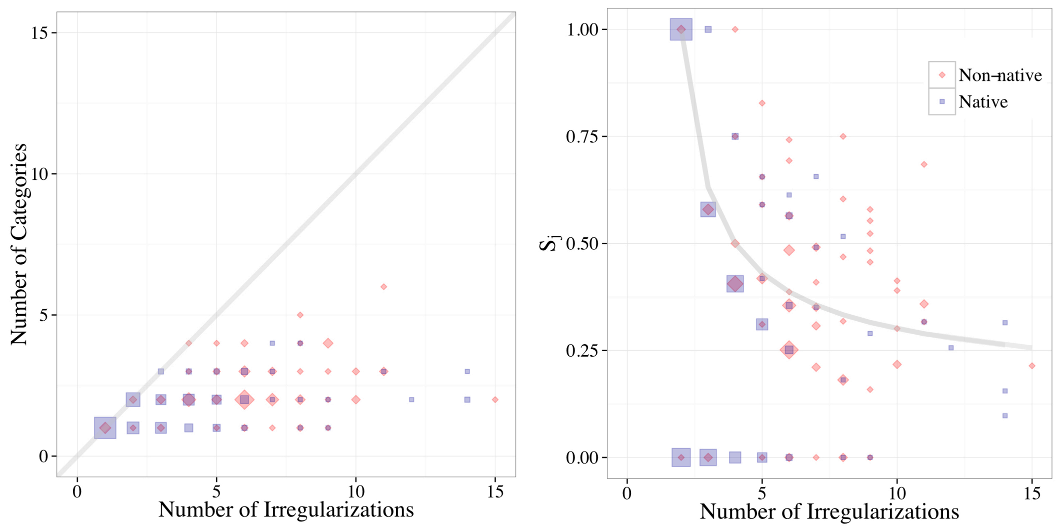
a participant, we define p_i^j as the fraction of irregularisations adhering each sub-rule i adopted by the participant. Given this, S_j is defined as:

$$S_j = \frac{-\sum_{i=1}^{n_j} p_i^j \log_2 p_i^j}{\log_2 n_j} \tag{1}$$

with the normalisation given by the maximal value of the numerator, $\log_2 n_j$, which is acquired when $p_i^j = 1/n_j$ for all the adopted sub-rules. In this way, the normalised quantity S_j provides a value for each participant ranging between 0 and 1. $S_j = 0$ means that the participant always made use of the same sub-rule across his/her irregularisations[8]. Small values of S_j indicate that the participant had fewer ways of irregularising than irregularisations. On the other hand, values of S_j close to 1 indicate that the participant did not apply sub-rules across irregularisations; in other words, provided uniquely irregular forms for each irregularisation. The S_j measure also reflects the distribution of rules, such that S_j is lower if the distribution of rules is skewed (the grey line in 11 reflects the S_j value for each value of n_j if two categories were used evenly across all regularisations) .Figure 11b shows the S_j value for each participant against their total number of irregularisations, n_j. This plot shows that participants who provided more irregular responses are not introducing new variation, since their responses tended to adhere to "sub-rules" already present in English. In particular, the values and range of S_j decrease as the value of n_j increases. In other words, participants who provide more non-ed forms also tend to use sub-rules in a less uniform way, i.e., they tend to prefer a limited set of sub-rules.

There are also interesting patterns across participants. Based on research from the ALL literature testing input-output overlap in particular (Wonnacott et al., 2013), we sought to compare the distribution of response types from our participants to the distribution of real irregulars from corpus data (Cuskley et al., 2014). In other words, by taking a corpus to be at least

[8]Participants who provided only one irregularisation were excluded from this analysis.



(a)

(b)

Figure 11: (a) Plot of the number of irregularisatoins against the number of irregular categories for each participant. The size of squares/diamonds represents the number of participants clustered at a given point. This shows that the number of categories does not increase with the number of irregularisations (the grey line would represent linear growth), demonstratng the use of sub-rules across irregular forms. (b) Plot of the total number of irregularisations (n_j) for each participant vs their S_j value, which represents the use of sub-rules. Towards 0, the S_j value indicates that the participant employed few distinct sub-rules across irregularisations in a non-distributed way. For example, if two rules were used across 8 irregularisations, the S_j is lower if one rule was used for seven irregularisations and another for one, than if each rule was applied to four irregularisations. $S_j = 0$ indicates that a single rule was used across all non *-ed* responses. As the S_j value increases towards 1, this indicates the of usage sub-rules was more uniform across irregularisations. $S_j = 1$ when the number of irregularisations is equal to the number of irregular categories (which, in this case, do not qualify as true sub-rules). The grey curve represents where participants would fall if sub-rules were evenly distributed across irregularizations. Most participants fall below this curve, showing that some sub-rules were more broadly applied than others.

broadly representative of learner input, we examine how input compares in particular to responses in the Wug-task. Results from ALL studies show that adult participants (Wonnacott et al., 2013) and some children (Boyd and Goldberg, 2012) reproduce the proportion of irregularity found in their input. In this sense, the generally high frequency of irregular verbs (Bybee, 2001; Cuskley et al., 2014) may influence irregularisation rates of novel verbs. In other words, non-natives are more likely to receive input that favours irregulars more extremely, exhibiting a token-based preference for irregularity. On the other hand, native speakers are more likely to have broader input including the 'long tail' of regular verb types (e.g., see Cuskley et al., 2014), and thus exhibit a type-preference for regularity[9].

To examine this, we plotted the proportion of different types of responses (regular and different categories of irregular) against the actual distribution of regulars and irregulars from the 1980-1989 decade of the Corpus of Historical American English (CoHA; Davies, 2012)[10]. Figure 12 shows that non-natives systematically underestimate the regular category relative to natives, over-estimating the vowel change (e.g., *blow/blew*) and level (e.g., *cut/cut*) categories in particular.

Overall, Experiment 2 replicated the findings of Experiment 1, showing that non-native English speakers provide non-ed past tense forms at a significantly higher rate than native speakers, and both groups are sensitive to phonological similarity between non-words and existing regular or irregular verbs. While "irregular" responses by definition did not follow the type-dominant "add -*ed*" rule, they generally followed existing sub-rules governing

[9]Note that we did not find any effects of nearest neighbor frequency in terms of specific proximate real verbs, though these frequencies are reported in Appendix B.

[10]This decade of CoHA was used in lieu of more recent corpora (e.g., Corpus of Contemporary American English, Davies, 2013) because of the detailed measurements of irregularity available from the study reported in Cuskley et al. (2014). These measurements mean these frequencies are not a sum of the frequencies of all verbs which involve e.g., a vowel change in the irregular past tense form, rather, they are actual frequencies of past tense vowel change tokens (i.e., tokens like *sneaked* are excluded).

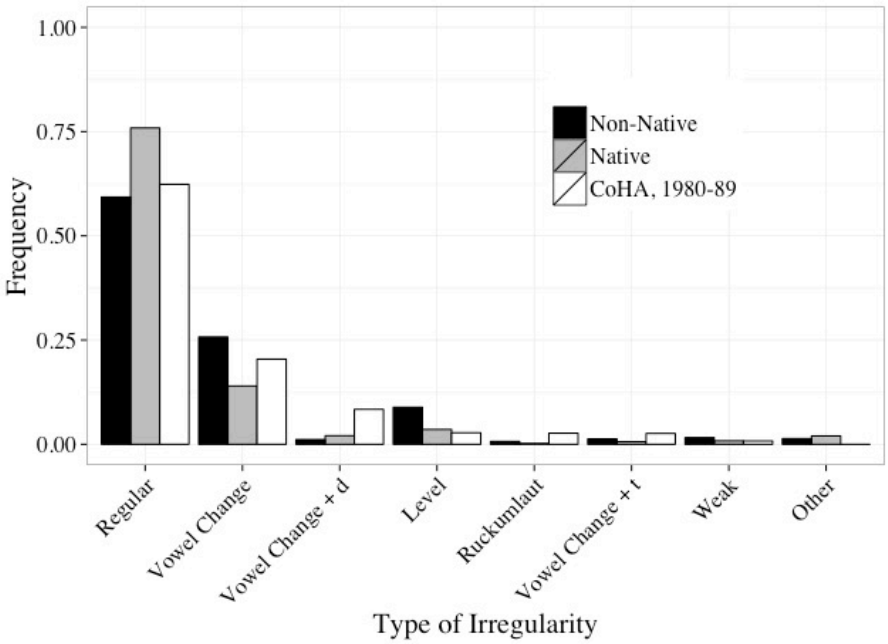


Figure 12: Distribution of types of irregular responses in the experiment, for natives, non-natives, and all participants contrasted with the distribution of irregular types from the 1980-1989 decade of CoHA. Each irregular category is described in Table 6 and represents a more general collapse of the specific classes found in Cuskley et al. (2014). The reported CoHA frequencies exclude the highly frequent irregular suppletive forms for *be, have, go*, and *do*; note that this makes the Regular category for CoHA appear particularly distorted (the actual percentage of regular verbs in CoHA is closer to 40%).

English irregular forms. Broadly, the distribution of regulars and irregulars among responses mirrored the actual distribution found in an English corpus.

However, specific important differences were evident: non-natives systematically underestimated the proportion of regulars, and over-estimated the proportion of vowel changes and levelled forms. For the vowel change category, the over-estimation is likely due at least in part to the non-word input: for 14 of the 15 novel verbs presented in Experiment 2, the closest irregular form involved a vowel change (see Appendix B). Regardless, the pattern of responses is still informative. Participants seem to treat word internal vowel changes as a very broad category, often not sensitive to the specific shift a proximate existing verb would dictate. For example, the novel verb *sleen*, by strict analogy with its closest irregular *sling*, should have taken the form *slun*; but across 43 vowel change irregularisations for this novel verb, there were only four occurrences of *slun* specifically, with the forms *slen* and *sloon* being more frequent among a wide range of different vowel changes. The class of verbs exemplified by *sling* specifically is generally considered to be the strongest class of the "strong" verbs (Bybee and Moder, 1983); but rather than prompting a very specific irregular form, it seems more likely that a general rule roughly summarised as "change an internal vowel" is at work.

Less expected given the input was the tendency among non-natives to over-estimate the level category (e.g., *cut-cut*). Only one novel verb had a proximate irregular which falls into this category (*quet* was closest to *quit*). While *quet* formed a large portion of the level irregularisations, *drust* and *slaide* also had high rates of levelling, both for native and non-native participants. The level category also appears to be expanding at least in the recent past (Cuskley et al., 2014) and represents the ultimate in deflexion (i.e., a total loss of the past tense inflection), perhaps indicating a shift away from a marked past-tense inflection altogether.

3. General Discussion

Experiments 1 and 2 contrasted the behaviour of native and non-native speakers in a simple past-tense Wug-task. The broader goal of these experiments was to examine differences between native and non-native speakers in how they apply linguistic rules to novel tokens. The first experiment showed that non-native speakers are more likely than native speakers to inflect novel verbs irregularly, and both groups are sensitive to the phonological distance between novel verbs and existing verbs. Detailed results of experiment 2 showed that for both native and non-native speakers, irregularisations of novel verbs do not introduce new complexity per-se, but rather, participants' irregularisations build on existing past tense sub-rules in English. The pattern of these sub-rules largely reproduces patterns of irregular sub-rules found in a corpus representative of input, with an interesting deviation: non-natives over estimate the prevalence of vowel changes (e.g., *hide-hid*) and levelled past tense forms (e.g., *quit-quit*) in particular.

In experiment 1, proficiency was a predictor of regularity among non-natives, while AoA was more influential in the second experiment. As coarse self-reported measures which form a proxy for overall nativeness, these measures seem to indicate that more native-like English is associated with a lower rate of irregularisation. Future investigations should aim to make more direct measures of nativeness, perhaps including more objective measures of proficiency and exposure (e.g., vocabulary size), as well as duration of exposure in addition to AoA.

In part, non-natives may engage in more irregularisation in an effort to reproduce their input, as many ALL studies might predict. The influence of high frequency irregulars may be out-sized in non-natives for two reasons. First, since non-natives presumably encounter less input data in terms of sheer tokens, they lack the long tail of regular verb types more likely to be known by learners with more exposure, with natives having the most comprehensive exposure to the language. This is further supported by the results that increased proficiency (Experiment 1) and earlier AoA (Experiment 2) were associated with reduced irregularisation rates, since both of these likely relate at least somewhat to increased exposure to the language; future studies should examine duration of exposure (a measure we did not collect) more explicitly. Even in terms of relatively comprehensive exposure to tokens, approximately 60-65% of past tense verb tokens in English are irregular (Cuskley et al., 2014). Thus, non-natives may be more likely to underestimate the overall amount of regularity (i.e., the -ed "rule") they should reproduce. Second, L2 learning emphasises irregular forms not only because of their frequency, but also because of their markedness, often explicitly dividing irregular forms into sub-rules to facilitate learning (Greenbaum and Quirk, 1996). Overall, these results can be used to inform hypotheses regarding how changes in social structure lead to changes in language. At first glance, the higher irregularisation rates among non-natives may seem to contravene hypotheses about language structure and social structure predicting that non-natives reduce complexity. But our results also show that *how* non-natives irregularise is consistent with a non-native preference for rules.

Analysis of which categories of irregularity participants over-estimated is informative. In terms of overestimation of the level category, this result may be an indication that the level category is on the rise. This suggestion is reinforced by the fact that the level category has drawn new irregulars in the past hundred or so years; namely *wed* and *quit* have become more irregular in the period covered by CoHA Cuskley et al. (2014). Interestingly, the only other irregular category which exhibits growth in CoHA is a particular group of vowel changes of the *hide-hid* variety, and vowel changes were also notably over-estimated by non-natives in Experiment 2, although the over-estimation of this category was likely heavily influenced by the non-word input. Yet, the pattern of vowel-change responses indicates a potential generalisation over different kinds of vowel changes, and an indication that several similar sub-rules may be collapsed. If true, this would signify a reduction in the complexity of the rule set, as hypotheses regarding how non-native learners affect language would predict. Similarly, the proliferation of the level category - despite technically being an "irregularisation" - also reinforces general hypotheses suggesting that adult non-natives may drive deflexion. This is perhaps an indication that non-native speakers are driving a shift to a more lexical strategy for the past tense in favour of an inflectional strategy (e.g., markers like "yesterday" or "earlier" to indicate temporal information, incidentally common among "basic variety" language forms of beginner-level adult speakers; Noyau, 2002) .

Overall, our results support the broad hypothesis that non-native, adult learners seem to have a preference for rules over exceptions (Wray, 2007), and simplicity over complexity (Lupyan and Dale, 2010). However, as the first explicit contrast of how natives and non-natives implement inflection in production, we uncovered some suprising results in terms of *how* this preference manifests. Rather than presenting as a straightforward non-native preference for the type-dominant *-ed* rule, the preference for rules was more nuanced in the context of production with novel verbs. Non-natives tended to proliferate sub-rules which have high token-frequency in input (since irregular verbs are generally more frequent), perhaps finding a sort of local optimum of rule simplification. In other words, it is possible that in an overall trend of decreasing complexity as non-native influence increases, a collapse of sub-rules may precede a preference for the type-dominant *-ed* rule. The over-estimation of the level category among non-natives in particular may indicate a shift towards a null morpheme for the past tense, which would ultimately be a more extreme simplification than the elimination of past-tense irregularity in favour of the *-ed* form.

These results are particularly surprising in light of earlier experiments indicating that native speakers who have increased contact with non-natives have a marked preference for regular past-tense inflection (Dale and Lupyan, 2012). Two specific methodological differences may account for these results. First, Dale and Lupyan (2012) used actual English verbs which have potentially ambiguous past tense forms (e.g., *speed* → *speeded/sped*) rather than non-words, and rated the acceptability of multiple forms rather than generating a preferred form. It is possible that particularly for known words with additional semantic information (Patterson et al., 2001), and particularly given the explicit choice of a regular form, non-natives would demonstrate a clearer preference for the regular rule. However, given open-ended production as in our task, the preference for structure plays out in a less predictable way. Second, it is possible that language contact has different effects on the preferences of native and non-native speakers. In showing a preference for the regular form among natives, our results are in line with Dale and Lupyan (2012), given that their experiment only tested the preferences of native speakers. Perhaps an influx of non-natives pushes native production towards simpler forms as an audience accommodation effect. Indeed, changes in non-native production to accommodate non-native speakers have been found in studies of prosody (Smith, 2007) and humor (Bell, 2007), indicating that such a strategy could play a role in inflection.

4. Conclusions

Our results confirm the broad strokes of previous theories regarding language structure and social structure, and provide new detail regarding how different learner profiles may affect the rule set of a language. The specific mechanisms underlying co-morbid changes in language structure and social structure are still largely unexplored, but our experiments indicate that adult non-native speakers proliferate existing rules in language in a complex way: rather than simply reproducing the regular rule, they also extend existing irregular rules. Future studies should focus more specifically on how non-native learners reproduce and generalise over irregular sub-rules, and how non-native audience effects may alter native inflectional strategies. Another open area of inquiry could focus more specifically on how differences in input frequency may effect irregularisation behaviour in natives and non-natives (an extension of existing ALL work in this area, e.g., Wonnacott et al., 2013; Reali and Griffiths, 2009); this would provide a more concrete bridge between work on social structure and language structure and existing work focusing on child and adult learner profiles.

A final area that warrants further investigation is the potential for specific effects of non-natives' first languages. Our analysis did not reveal any specific L1 effects, but this does not provide definitive evidence of a lack of L1 effects. The absence of any L1 effects may simply be indicative of a limited set of L1s represented among our participants, and a skew towards certain languages (e.g., Romance languages in Experiment 2). However, given the strong results in both experiments, each with diverse L1 samples, it is also possible that this effect is generally robust across all learners regardless of their specific L1. Future studies could consider in more specific detail how different substrate L1s among non-native English speakers might affect regularisation behaviour.

The broad theory that social structure and learner profiles are a potentially influential factor in language structure drew most direct evidence from historical linguistics and corpus data, and here we have provided evidence from a production task which broadly supports these theories. Our results provide a stepping stone to future work examining exactly how natives and non-natives realise linguistic rules differently, and draw attention to broader questions regarding the complex relationship between social structure, learner profiles, and language structure.

Acknowledgements

This work was supported by the European Science Foundation as part of the DRUST project, a EUROCORES EuroUnderstanding programme.

Appendix A. Method for generating non-words

To generate the large set of non-words from which the final set of 68 items was for Experiment 1 was taken, we first took all the segments occurring in the set of irregular English verbs and the 500 most frequent regular verbs. Each phonological segment in the set was rated in terms of presence/absence/ or a inapplicability in terms of 12 features: [1] consonant, [2] voiced, [3] approximant, [4] sonorant, [5] continuant, [6] labial, [7] dorsal, [8] front, [9] back, [10] high, [11] low and [12] round. Features 2-7 applied only to consonants and 8-12 only to vowels. The distance between segments thus depends on a feature being shared (e.g., two voiced consonants, incurring a cost of 0), opposite (a voiced and voiceless consonant, incurring a cost of 1), or simply unshared (a voiced consonant and vowel, incurring a cost of 0.5). This is analogous to the procedure in Levenshtein edit distance wherein a substitution is considered twice the cost of an insertion or deletion (given that a substitution operation involves both a deletion and an insertion; Nerbonne & Herringa, 1997). In this case, unshared features are considered deletions, where opposing features are considered substitutions. In order to calculate distances at the word-level, a Levenshtein edit distance was calculated where substitution cost was defined by phonological segment distance, and insertion or deletion was defined as half of the average substitution cost across the entire phone set. This distance was normalised for word length given the generally shorter length of high frequency words (Piantadosi, 2014), and the fact that irregular verbs are generally higher frequency.

Appendix B. Non-word materials

Appendix B.1. Table B.1: Words only in Experiment 1

Appendix B.2. Table B.2: Words from Experiment 1 & 2

Appendix C. Model Details

Appendix C.1. Experiment 1

Appendix C.1.1. Table C.1.1: Main model selection

Appendix C.1.2. Table C.1.2: Proficiency Model Selection

Appendix C.1.3. Table C.1.3: Proficiency Model Summary

Appendix C.2. Experiment 2

Appendix C.2.1. Table C.2.1: Main model selection

Appendix C.2.2. Table C.2.2: AoA Model Selection

Appendix C.2.3. Table C.2.3: AoA Model Summary

Albright, A., Hayes, B., 2003. Rules vs. analogy in english past tenses: A computational/experimental study. Cognition 90, 119–161.

Allen, C., 2003. Deflexion and the development of the genitive in english. English Language and Linguistics 7 (1), 1–28.

Arnon, I., 2010. Rethinking child difficulty: The effect of np type on children's processing of relative clauses in hebrew. Journal of Child Language 37 (1), 27–57.

Baayen, R., Davidson, D., Bates, D., 2008. Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language 59, 390–412.

Bakker, P., Matras, Y., 2013. Contact Languages: A Comprehensive Guide. Language contact and bilingualism. De Gruyter Mouton.

Bell, N., 2007. How native and non-native english speakers adapt to humor in intercultural interaction. International Journal of Humor Research 20 (1), 27–48.

Bergen, B., 2001. Nativization processes in l1 esperanto. Journal of Child Language 28 (3), 575–595.

Berko, J., 1958. The child's learning of english morphology. Word 14, 150–177.

Boyd, J., Goldberg, A., 2012. Young children fail to fully generalize a novel argument structure construction when exposed to the same input as oler learners. Journal of Child Language 39 (3), 457–481.

Bybee, J., 2001. Phonology and language use. Cambridge, Cambridge University Press.

Bybee, J. L., Moder, C. L., 1983. Morphological classes as natural categories. Language 59 (2), 251–271.

Carrol, R., Svare, R., Salmons, J., 2012. Quantifying the evolutionary dynamics of German verbs. Journal of Historical Linguistics 2 (2), 153–172.

Christiansen, M., Chater, N., 2008. Language as shaped by the brain. Behavioral and Brain Sciences 31 (5), 489–509.

Cicali, C., Tria, F., Servedio, V., Gravino, P., Loreto, V., Warglien, M., Paolacci, G., 2011. Experimental tribe: A general platform for web gaming and social computation. In: Proceedings of NIPS Workshop on Computational Social Science and the Wisdom of Crowds. pp. 1–5.

Clahsen, H., Felser, C., 2006. How native-like is non-native language processing. Trends in Cognitive Sciences 10 (12), 564–570.

Clahsen, H., Felser, C., Neubauer, K., Sato, M., Sliva, R., 2010. Morphological structure in native and non-native language processing. Language Learning 60 (1), 21–43.

Cornish, H., 2010. Investigating how cultural transmission leads to the appearance of design without a designer in human communication systems. Interaction Studies 11 (1), 112–137.

Cuskley, C., 2013. Mappings between linguistic sound and motion. Public Journal of Semiotics 5 (1), 39–62.

Cuskley, C., Pugliese, M., Castellano, C., Colaiori, F., Loreto, V., Tria, F., 2014. Internal and external dynamics in language: Evidence from verb regularity in a historical corpus of English. PLoS ONE 9 (8), e102882.

Dale, R., Lupyan, G., 2012. Understanding the origins of morphological diversity: The linguistic niche hypothesis. Advances in Complex systems 15 (3), 1–16.

Davies, M., 2012. Corpus of historical american english: 400 million words from 1810-2009.
URL `http://corpus.byu.edu`

Davies, M., 2014. Corpus of contemporary american english: 450 million words, 1990-present.
URL `http://corpus.byu.edu`

Dryer, M., Gil, D., Comrie, B., Jung, H., Schmidt, C., 2005. The world atlas of language structures. www.wals.info.

Gershkoff-Stowe, L., Thelen, E., 2004. U-shaped changes in behavior: a dynamic systems perspective. Journal of Cognition and Development 5 (1), 11–36.

Greenbaum, S., Quirk, R., 1996. A student's grammar of the english language. Longman, London.

Hickey, R. (Ed.), 2010. The handbook of language contact. John Wiley & Sons, New York.

Hockett, C., 1960. The origin of speech. Scientific American 203, 88–96.

Hudson Kam, C., Newport, E., 2005. Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. Language Learning and Development 1 (2), 151–195.

Jaeger, F., 2008. Categorical data analysis: Away from anovas and towards logit mixed models. Journal of Memory and Language 59, 434–436.

Kam, C. L. H., Newport, E. L., 2009. Getting it right by getting it wrong: When learners change languages. Cognitive Psychology 59 (1), 30 – 66.
URL `http://www.sciencedirect.com/science/article/pii/S0010028509000048`

Kersten, A. W., Earles, J. L., 2001. Less really is more for adults learning a miniature artificial language. Journal of Memory and Language 44 (2), 250–273.

Kirby, S., Cornish, H., Smith, K., 2008. Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. Proceedings of the National Academy of Sciences 105 (31), 10681–10686.

Lass, R., 1992. Phonology and morphology. In: Blake, N. (Ed.), The Cambridge History of the English Language, Vol. II: 1066-1476. Cambridge University Press, Cambridge, UK, pp. 23–155.

Lieberman, E., Michel, J.-B., Jackson, J., Tang, T., Nowak, M. A., 2007. Quantifying the evolutionary dynamics of language. Nature 449 (2007), 713–716.

Lupyan, G., Dale, R., 2010. Language structure is partly determined by social structure. PLoS ONE 5 (1), e8559.

Maslen, R., Theakston, A., Lieven, E., Tomasello, M., 2004. A dense corpus study of past tense and plural over-regularization in english. Journal of Speech, Language, and Hearing Research 37, 1319–1333.

Michaelis, S. M., Maurer, P., Haspelmath, M., Huber, M. (Eds.), 2013. APiCS Online. Max Planck Institute for Evolutionary Anthropology, Leipzig.
URL `http://apics-online.info/`

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, T. G. B., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., Aiden, E. L., 2011. Quantitative analysis of culture using millions of digitized books. Science 331 (6014), 176–182. URL `http://www.sciencemag.org/content/331/6014/176.abstract`

Monaghan, P., 2014. Age of acquisition predicts rate of lexical evolution. Cognition 133, 530–534.

Nerbonne, J., Heeringa, W., 1997. Measuring dialect distance phonetically. In: Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology. pp. 11–18.

Noyau, C., 2002. Temporal relations in learner varieties: Grammaticalization and discourse construction. In: Slaberry, R., Sirai, Y. (Eds.), The L2 Acquisition of Tense-aspect morphology. John Benjamins Publishing Company, pp. 107–127.

Paolacci, G., Chandler, J., 2014. Inside the Turk: Understanding Mechanical Turk as a participant pool. Current Directions in Psychological Science 23 (3), 184–188.

Patterson, K., Lambon Ralph, M., Hodges, J., McClelland, J., 2001. Deficits in irregular past-tense verb morphology associated with degraded semantic knowledge. Neuropsychologia 39 (7), 709–724.

Piantadosi, S. T., 2014. Zipf's word frequency law in natural language: A critical review and future directions. Psychonomic Bulletin & Review 21 (5), 1112–1130.

Prasada, S., Pinker, S., 1993. Generalisation of regular and irregular morphological patterns. Language and cognitive processes 8 (1), 1–56.

Reali, F., Griffiths, T. L., 2009. The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. Cognition 111 (3), 317–328.

Roberts, S., Winters, J., 2012. Social structure and language structure: The new nomothetic approach. Psychology of Language and Communication 16 (2), 89–112.

Seidlhofer, B., 2001. Closing a conceptual gap: The case for a description of english as a lingua franca. International Journal of Applied Linguistics 11 (2), 133–158.

Smith, C., 2007. Prosodic accommodation by french speakers to a non-native interlocutor. In: Proceedings of the XVIth International Congress of Phonetic Sciences. pp. 313–348.

Smith, K., Wonnacott, E., 2010. Eliminating unpredictable variation through iterated learning. Cognition 116 (3), 444–449.

Thomason, S., 2001. Language Contact. Edinburgh University Press.

Trudgill, P., 2010. Investigations in sociohistorical linguistics: stories of colonisation and contact. Cambridge University Press, Cambridge, UK.

Weinreich, U., 1963. Languages in contact: findings and problems. Publications. Mouton.

Wonnacott, E., Brown, H. E., Nation, K., June 2013. Comparing generalisation in children and adults learning an artificial language. In: Child Language Seminar. Manchester, UK.

Wray, A., 2007. 'needs only' analysis in linguistic ontogeny and phylogeny. In: Lyon, C., Nehaniv, C. L., Cangelosi, A. (Eds.), Emergence of Communication and Language. Springer, New York, pp. 53–70.

Wright, S., 2006. French as a lingua franca. Annual review of applied linguistics 26, 35–60.

Table B.1: Non-word stimuli used only in Experiment 1, with closest (proximate) real verbs and irregularisation rates. The Type column represents the category of the non-word, defined by its closest real verb (IN: intermediate, I: irregular, R: regular). Each the frequency bin of each verb, taken from the Corpus of Contemporary English Davies (2014), is provided in parentheses (a bin of e.g., 10^{-4} indicates a frequency $0.0001 \leq f < 0.001$). The ρ_I indicates the regularisation rate. Because of different numbers of native and non-native participants (described in detail in the methods section of Experiments 1 and 2), note that the total ρ_I is not the midpoint between native and non-native values of ρ_I. Note that the audio files were generated to match the IPA description of each non-word used to calculate edit distances, rather than the orthographic written form. Therefore, in many cases, the written form used to generate the audio file differs from the written form presented to participants. For example, to generate the form /kwoʊk/ using the Mac text to speech feature, the written form *kwoke* was used, but the more standard spelling of *quoke* was used for presentation to participants.

Target	Type	Prox. R (f bin)	Prox. I (f bin)	Total ρ_I	Native ρ_I	Non-native ρ_I
hend	IN	handle (10^{-5})	send (10^{-4})	0.32	0.14	0.63
shenk	IN	thank (10^{-4})	think (10^{-3})	0.24	0.10	0.50
spleem	IN	scream (10^{-5})	spring (10^{-5})	0.17	0.17	0.40
stip	IN	step (10^{-4})	stick (10^{-5})	0.05	0.10	0.20
thail	IN	sign (10^{-4})	shine (10^{-5})	0.14	0.21	0.33
chauze	I	study (10^{-4})	choose (10^{-4})	0.03	0.29	0.00
choove	I	prove (10^{-4})	choose (10^{-4})	0.35	0.08	0.50
chune	I	ensure (10^{-5})	choose (10^{-4})	0.10	0.05	0.13
dwal	I	yell (10^{-5})	dwell (10^{-6})	0.17	0.14	0.33
dweel	I	yell (10^{-5})	dwell (10^{-6})	0.23	0.30	0.38
dweer	I	yell (10^{-5})	dwell (10^{-6})	0.32	0.36	0.40
dwen	I	dress (10^{-5})	dwell (10^{-6})	0.37	0.13	0.40
dwill	I	yell (10^{-5})	dwell (10^{-6})	0.25	0.07	0.44
fring	I	plan (10^{-4})	sling (10^{-6})	0.67	0.15	0.50
queeke	I	treat (10^{-4})	quit (10^{-5})	0.19	0.18	0.22
queep	I	treat (10^{-4})	quit (10^{-5})	0.23	0.33	0.40
slin	I	plan (10^{-4})	sling (10^{-6})	0.30	0.25	0.50
spang	I	expand (10^{-5})	span (10^{-6})	0.29	0.18	0.38
speem	I	estimate (10^{-5})	spin (10^{-5})	0.11	0.42	0.17
speeze	I	kiss (10^{-5})	spin (10^{-5})	0.31	0.33	0.15
spid	I	step (10^{-4})	spit (10^{-5})	0.42	0.37	0.56
spim	I	estimate (10^{-5})	spin (10^{-5})	0.39	0.71	0.50
sping	I	stir (10^{-5})	sting (10^{-6})	0.67	0.35	0.57

Target	Type	Prox. R (f bin)	Prox. I (f bin)	Total ρ_I	Native ρ_I	Non-native ρ_I
splew	I	explore (10^{-5})	strew (10^{-6})	0.29	0.25	0.42
sprew	I	explore (10^{-5})	strew (10^{-6})	0.21	0.00	0.20
sweave	I	slip (10^{-5})	swim (10^{-5})	0.46	0.69	0.71
threen	I	plan (10^{-4})	sling (10^{-6})	0.17	0.33	0.10
threeng	I	slip (10^{-5})	fling (10^{-6})	0.42	0.43	0.75
thrin	I	plan (10^{-4})	sling (10^{-6})	0.43	0.71	0.43
blop	R	drop (10^{-4})	blow (10^{-5})	0.10	0.07	0.13
brop	R	drop (10^{-4})	blow (10^{-5})	0.13	0.05	0.22
clote	R	close (10^{-4})	throw (10^{-4})	0.13	0.00	0.33
cluve	R	prove (10^{-4})	grow (10^{-4})	0.11	0.04	0.00
crey	R	pray (10^{-5})	slide (10^{-5})	0.10	0.16	0.25
croose	R	cross (10^{-5})	grow (10^{-4})	0.17	0.26	0.20
croze	R	close (10^{-4})	throw (10^{-4})	0.30	0.23	0.38
cruve	R	prove (10^{-4})	grow (10^{-4})	0.15	0.31	0.00
drup	R	drop (10^{-4})	thrust (10^{-6})	0.11	0.30	0.33
flug	R	shrug (10^{-5})	tread (10^{-5})	0.17	0.16	0.38
fluve	R	prove (10^{-4})	fling (10^{-6})	0.17	0.20	0.20
fote	R	vote (10^{-5})	cost (10^{-5})	0.34	0.71	0.67
fruve	R	prove (10^{-4})	fling (10^{-6})	0.27	0.26	0.44
greel	R	clear (10^{-5})	grind (10^{-5})	0.31	0.05	0.40
grop	R	drop (10^{-4})	grow (10^{-4})	0.11	0.18	0.29
hoke	R	hope (10^{-4})	cost (10^{-5})	0.13	0.20	0.11
kaist	R	taste (10^{-5})	take (10^{-3})	0.29	0.22	1.00
metch	R	match (10^{-5})	knit (10^{-6})	0.29	0.23	0.50
nast	R	last (10^{-5})	thrust (10^{-6})	0.25	0.18	0.33
spaull	R	score (10^{-5})	spend (10^{-4})	0.22	0.08	0.33
spop	R	stop (10^{-4})	spit (10^{-5})	0.28	0.21	0.40
stot	R	stop (10^{-4})	shut (10^{-5})	0.14	0.37	0.18
throg	R	shrug (10^{-5})	tread (10^{-6})	0.17	0.19	0.22
wutch	R	watch (10^{-4})	wet (10^{-6})	0.13	0.07	0.22

Table B.2: Non-word stimuli used in Experiments 1 and 2, with closest real verbs and irregularisation rates.

				Experiment 1 ρ_I			Experiment 2 ρ_I		
Target	Type	Prox. R (f)	Prox. I (f)	Total	Native	Non native	Total	Native	Non native
bleen	IN	breathe (10^{-5})	bring (10^{-4})	0.28	0.09	0.44	0.34	0.17	0.58
dake	IN	bake (10^{-5})	take (10^{-3})	0.33	0.16	0.38	0.26	0.17	0.38
drust	IN	trust (10^{-5})	thrust (10^{-6})	0.29	0.07	0.20	0.24	0.22	0.28
slaide	IN	trade (10^{-5})	slide (10^{-5})	0.25	0.29	0.50	0.32	0.29	0.36
waip	IN	wait (10^{-4})	wake (10^{-5})	0.18	0.07	0.17	0.14	0.08	0.23
quet	I	treat (10^{-4})	quit (10^{-5})	0.43	0.07	0.67	0.37	0.29	0.48
sleen	I	plan (10^{-4})	sling (10^{-6})	0.42	0.23	0.67	0.39	0.32	0.49
spink	I	thank (10^{-4})	stink (10^{-6})	0.48	0.11	0.83	0.52	0.38	0.70
swin	I	switch (10^{-5})	swim (10^{-5})	0.70	0.00	0.71	0.62	0.46	0.83
thring	I	slip (10^{-5})	fling (10^{-6})	0.74	0.15	0.80	0.69	0.68	0.72
cluse	R	cross (10^{-5})	grow (10^{-4})	0.06	0.15	0.17	0.10	0.09	0.10
drock	R	drop (10^{-4})	draw (10^{-4})	0.21	0.05	0.30	0.23	0.10	0.39
plal	R	plan (10^{-4})	draw (10^{-4})	0.18	0.10	0.20	0.18	0.09	0.30
puve	R	prove (10^{-4})	blow (10^{-5})	0.19	0.18	0.43	0.11	0.11	0.11
quoke	R	quote (10^{-5})	throw (10^{-4})	0.18	0.18	0.38	0.11	0.08	0.16

Table C.1.1: **Main model selection**.The table below provides Bayesian Information Criterion (BIC), Alkali Information Criterion (AIC), and log-likelihood (logLik) for several potential models fit to the data for Experiment 1. For all models, the glmer() call was Response ~[Fixed effects]+(1|Participant)+(1|Item)+(1|ResponseNumber), and fit a binomial model (i.e., all models used the same outcome variable and random effects). Model selection was accomplished by comparing information criterion and log-likelihood for different potential models, as well as comparing models for significant differences in fit. Generally, the lower the AIC and BIC values, and the lower the absolute value of the log likelihood, the better the fit of the model. We also report chi-square values comparing each model to the final model reported (model A) using an ANOVA. Where two models displayed comparable fit, the simpler model was preferred (e.g., model A was not significantly different from model B, but the model without interactions was preferred for its simplicity). Although self-reported proficiency as a predictor provided a better fit and slightly lower AIC/BIC and log likelihood values than nativeness (models C and D), this measure had ceiling effects since natives were automatically rated at full proficiency, and some non-natives also rated themselves at full proficiency (mean proficiency = 92.9, SD=15.6). Thus, the nativeness model was preferred for the overall data and proficiency was examined in more detail among non-natives only (see below).

Model	Fixed effects	AIC	BIC	logLik	ANOVA with A	Pref. Mod.
A	Nativeness + ItemType	1752	1790	-868.9	——	——
B	Nativeness x ItemType	1751	1800	-866.5	$\chi^2 = 4.9, p = 0.09$	A
C	Proficiency x ItemType	1742	1791	-862.2	$\chi^2 = 13.5, p < 0.01$	C
D	Proficiency + ItemType	1739	1777	-862.6	$\chi^2 = 12.6, p < 0.01$	D
E	AoA + ItemType	1767	1805	-876.3	$\chi^2 = 0, p = 1$	A
F	AoA x ItemType	1771	1820	-876.3	$\chi^2 = 0, p = 1$	A
G	Nativeness	1772	1799	-880.8	$\chi^2 = 23.8, p < 0.001$	A
H	ItemType	1778	1811	-883.1	$\chi^2 = 23.4, p < 0.001$	A
I	Proficiency	1759	1787	-874.7	$\chi^2 = 11.5, p < 0.01$	A
J	AoA	1757	1815	-888.6	$\chi^2 = 39.4, p < 0.001$	A

Table C.1.2: **Proficiency Model Selection**. The table below provides Bayesian Information Criterion (BIC), Alkali Information Criterion (AIC), and log-likelihood (logLik) for several potential models fit for non-native participants in Experiment 1. For all models, the glmer() call was Response ~[Fixed effects]+(1|Participant)+(1|Item)+(1|ResponseNumber), and fit a binomial model (i.e., all models used the same outcome variable and random effects). Model selection was accomplished as in the case of the main model.

Model	Fixed effects	AIC	BIC	logLik	ANOVA with A	Pref. Mod.
A	Proficiency + ItemType	641	671	-313.9	–	–
B	Proficiency x ItemType	647	681	-315.5	$\chi^2 = 0.64, p = 0.73$	A
C	AoA + Proficiency + ItemType	644	674	-315.2	$\chi^2 = 1.36, p = 0.24$	A
D	AoA + ItemType	657	682	-322.6	$\chi^2 = 0, p = 1$	A

Table C.1.3: Summary of fixed effects in mixed logit model for proficiency among non-natives in Experiment 1 (N=512, log likelihood = -313.9). Reference values for the intercept are a regular response with the overall mean proficiency (76.4).

Predictor	β Coef.	SE	CI (95%)		Wald's Z	p	OR
			2.5%	97.5%			
Intercept	0.904	(0.556)	0.04	1.77	1.628	0.104	2.4
Irregular	0.867	(0.280)	-0.03	1.75	3.13	<0.001	2.4
Intermediate	0.501	(0.388)	-0.41	1.41	1.295	0.195	1.6
Proficiency	-0.025	(0.007)	-0.08	-0.03	-3.616	<0.001	0.97

Table C.2.1: **Main model selection**The table below provides Bayesian Information Criterion (BIC), Alkali Information Criterion (AIC), and log-likelihood (logLik) for several potential models fit to the data for Experiment 2. For all models, the glmer() call was Response ~[Fixed effects]+(1|Participant)+(1|Item)), and fit a binomial model (i.e., all models used the same outcome variable and random effects). Model selection was accomplished by comparing information criterion and log-likelihood for different potential models, as well as comparing models for significant differences in fit. Generally, the lower the AIC and BIC values, and the lower the absolute value of the log likelihood, the better the fit of the model. We also report chi-square values comparing each model to the final model reported (model A) using an ANOVA. Where two models displayed comparable fit, the simpler model was preferred.

Model	Fixed effects	AIC	BIC	logLik	ANOVA with A	Pref. Mod.
A	Nativeness + ItemType	3021	3057	-1505	——	——
B	Nativeness x ItemType	3025	3073	-1504	$\chi^2 = 0.3, p = 0.86$	A
C	Proficiency x ItemType	3044	3093	-1514	$\chi^2 = 0, p = 1$	A
D	Proficiency + ItemType	3040	3077	-1514	$\chi^2 = 0, p = 1$	A
E	AoA + ItemType	3114	3150	-1551	$\chi^2 = 0, p = 1$	A
F	AoA x ItemType	3117	3165	-1550	$\chi^2 = 0, p = 1$	A
G	Nativeness	3039	3063	-1515	$\chi^2 = 21.7, p < 0.001$	A
H	ItemType	3055	3085	-1522	$\chi^2 = 35.4, p < 0.001$	A
I	Proficiency	3058	3083	-1525	$\chi^2 = 41.47, p < 0.01$	A
J	AoA	3035	3059	-1514	$\chi^2 = 18.0, p < 0.001$	A

Table C.2.2: **AoA Model Selection**. The table below provides Bayesian Information Criterion (BIC), Alkali Information Criterion (AIC), and log-likelihood (logLik) for several potential models fit for non-native participants in Experiment 2. For all models, the glmer() call was Response ~[Fixed effects]+(1|Participant)+(1|Item)+(1|ResponseNumber), and fit a binomial model (i.e., all models used the same outcome variable and random effects). Model selection was accomplished as in the case of the main model.

Model	Fixed effects	AIC	BIC	logLik	ANOVA with A	Pref. Mod.
A	AoA + ItemType	1470	1501	-729	——	——
B	AoA x ItemType	1470	1511	-727	$\chi^2 = 3.72, p = 0.15$	A
C	AoA + Proficiency + ItemType	1471	1507	-728.5	$\chi^2 = 0.92, p = 0.34$	A
D	Proficiency + ItemType	1476.7	1507.8	-732.4	$\chi^2 = 0, p = 1$	A

Table C.2.3: Summary of fixed effects in mixed logit model for age of acquisition (AoA) among non-natives in Experiment 2 (N=89, log likelihood = -729). Reference values for the intercept are a regular response with the overall mean AoA (9.9 years).

Predictor	β Coef.	SE	CI (95%)		Wald's Z	p	OR
			2.5%	97.5%			
Intercept	-2.42	(0.44)	-0.86	0.86	0.004	0.996	1.00
Regular	2.42	(0.463)	1.51	3.32	5.23	<0.001	11.26
Intermediate	1.45	(0.455)	0.56	2.34	3.18	<0.01	4.26
AoA	-0.076	(0.029)	-0.13	-0.02	-2.63	<0.01	0.93