Apparently identical verbs can be represented differently. Comparing L1-L2 inflection with contingency-based measure ΔP

(Figure 1 corrected on 20 December, 2020)

Stefano Rastelli (University of Pavia)
Akira Murakami (University of Birmingham)

Abstract

We propose a method that models unidirectional, contingency-based association scale ΔP in order to analyse the different degrees of morpheme productivity in apparently identical L1-L2 inflected pairs. The method has the potential to uncover differences in how in L1-L2 inflected items are represented by L2 learners and native speakers. Such differences are at risk of remaining invisible if one considers only frequency, distribution and rank of predicates.

1 Introduction: the comparison of L1-L2 items in SLA

When observing an inflected item produced by L2 learners – e.g., a verb – one may wonder if it is the same verb that a native speaker of that language uses and comprehends. What at first may seem an unlikely claim is understandable once it is appreciated that apparently indistinguishable inflected forms produced by native speakers and L2 learners can be generated and represented in different ways because native speakers and L2 learners may be differently aware of verb morphology. For example, a learner might just be mimicking the verb from the L1 input – after having roughly inferred its meaning from the native interlocutor – without necessarily being capable of distinguishing between its invariable and variable subcomponents, the stem and the affix. The issue of whether a target-like L2 inflected form is rule-generated or is represented as an unanalysed chunk is as old as SLA research (e.g. Dulay et al. 1982: 232; Pienemann 1998). Cases abound in the SLA literature where L2 morphosyntactic use achieves high accuracy without the corresponding morphosyntactic competence (Norris & Ortega 2003: 737). In learner corpus research, frequency comparison between native speakers' and L2 learners' language use is sometimes taken as a way to assess L2 development (Kilgarrif 2001; 2005). The idea is that the comparable frequency of a linguistic structure between native speakers' and L2 learners' production is considered to indicate that the structure has been acquired by the L2 learners. This crucially depends on two assumptions. The first one is that items that are more frequent in the L1 input are learned earlier and more stably than less frequent ones. So when L1 and L2 normalized frequencies of the same item are similar one can be driven to conclude that L2 usage is modelled after the target-language and that L1-L2 underlying representations too are comparable. Three considerations prevent from drawing such conclusion. First, the observed frequencies are always relative to a corpus, but any corpus – no matter its size – can be considered at most an acceptable approximation of the L1 input learners may happen to be exposed to[1]. Second, many L1 and L2 occurrences can be highly corpus-specific: rather than L1 uses, they might reflect the kinds of task (e.g. picture telling, narratives) and the topics of interactions used to elicit the data. Third, there is no agreement on which statistical test of significance one should adopt when comparing L1 and L2 frequencies. Given that comparing L1-L2 frequencies is at least problematic, one may resort to comparing L1-L2 rankings because such comparison may appear more immune from

[1] The issue of corpus representativeness is as old as corpus-based research. In short: a corpus cannot realistically be representative of *all features* of language use. A big part of this issue concerns corpus design, namely, how to address bias in sampling, whether it is stratificational sampling or probabilistic sampling. Whether representativeness is a direct function of corpus size is also debated (see for instance special issue of *CogniTextes*, Revue de l'association Française de Linguistique Cognitive, vol 19, 2019, https://doi.org/10.4000/cognitextes.1311). A separate issue – which is relevant here – is to what extent the linguistic information that is available in large L1 corpora can truly reflect the amount of information that is actually available to learners.

the bias deriving from outliers (although they are sensitive to other sources of bias). Table 1 reports the number of occurrences (normalized per million) and the rank (in brackets) of the ten most frequent past perfectives (the *passato prossimo*) in the corpus Pavia (the best-known Italian learner corpus, Section 6) with the number of occurrences and the rank of the same past perfectives in ItTenTen, the largest L1 Italian corpus to date. The rightmost columns report also the frequency and the rank of the corresponding lemmas – respectively – in L1 and L2 corpora.

Table 1: Number of occurrences and rank (in brackets) in the Pavia corpus (L2) and in ItTenTen (L1) of the ten most frequent past perfectives and of the corresponding lemmas

perfective	English	pastpL2	pastpL1	lemmaL2	lemmaL1
detto	said	477 (1)	384 (2)	1430 (3)	1562 (6)
fatto	done	437 (2)	767 (1)	1284 (5)	3877 (2)
andato	gone	420 (3)	154 (14)	1394 (4)	1177 (8)
visto	seen	323 (4)	68 (58)	620 (17)	887 (12)
capito	understood	215 (5)	66 (35)	386 (12)	337 (52)
arrivato	arrived	184 (6)	148 (17)	297 (20)	537 (20)
preso	taken	138 (7)	201 (7)	424 (9)	560 (19)
trovato	found	129 (8)	179 (10)	323 (16)	1044 (9)
parlato	talked	99 (9)	60 (22)	408 (11)	579 (16)
venuto	come	96 (10)	78 (60)	304 (18)	1601 (37)

The normalized frequencies of perfectives and lemmas in L1 and L2 Italian are very different, although both distributions are Zipfian. On the contrary, L1 and L2 rankings appear more similar. Particularly, L1 and L2 rankings of lemmas resemble more than L1 and L2 rankings of the perfectives (respectively, Kendall $\tau = 0.47$; p-value = 0.07255; Kendall $\tau = 0.36$; p-value = 0.1557). This difference increases in the low-frequency band of the rankings. In fact, if one scrolls down the L2 frequency list beyond the 1-10 positions and take – say – the ten least frequent L2 perfectives in our sample (see below), one finds that the difference between lemmas and inflected forms becomes greater (correlation of L1-L2 lemmas: Kendall $\tau = 0.60$, p-value = 0.01; correlation of L1-L2 perfectives Kendall $\tau = 0.24$; p-value = 0.38). The fact that distributional patterns of L1 vs L2 inflected forms differ more than the distributional patterns of L1 vs L2 verbal lemmas might suggest that – at least for initial (lower proficiency) L2 learners of Italian – the lack of similarity in frequency between L1 and L2 perfectives is not entirely attributable to the choice of verbs, which would presumably be reflected on the distribution of lemmas, and that L2 learners' production may be affected more by encountering lemmas in the L1 input than by encountering L1 perfectives. Whatever the effect that frequency may have on L2 acquisition, it seems that it facilitates learners' access to the lexical meaning of predicates rather than increasing learners' sensitivity to how L2 verb morphology works. This is somehow reflected by the well-known phenomenon of the low rate of morphemes supplied by initial and intermediate learners. The corpus Pavia too – like many other learner corpora – abund of mandatory past tense contexts in which low proficiency learners used the bare infinitive of the proper lemmas (e.g. *andare* 'go', *venire* 'come') or the basic present form rather than the expected past morpheme *-to* (e.g. *anda-to* 'gone', *venu-to* 'come'). For example, in sentence (1), when a Tygrinian learner of Italian is asked whether people felt good with the last emperor of Ethiopia, she answers by using the present rather than the past tense, as it would be expected:

(1) [Interviewer] *Quando c'era Hailè Selassiè si stava meglio*?
Did one live better when Hailè Selassiè was in power?

[Learner] *Sì perché non sappiamo niente*
Yes because (we) not knowPRES nothing
Yes because we did not know anything

The second, even more apparently undisputable assumption that often motivates L1-L2 frequency comparison in L2 research is that learners' production reflects their acquisition. Said it differently, a learner's exposure to target-like exemplars is expected to impact directly on the acquisition and – consequently also – on the production of target-like forms. In the past, many SLA researchers doubted that even very high (e.g. ≥ 90) percentages of target-like forms in learners' production necessarily indicates the presence of underlying morphological competence (see Pallotti 2007 for a review). This also echoes the traditional issue in SLA of how much of the L1 'input' can become L2 'intake'. It was proposed that encountering L1 forms does not necessarily entail the production or acquisition of those forms (Corder 1967, see also VanPatten & Benati, 2015, p 131; Gass, 2013, p. 340). This is because it is not L1 frequency in itself that counts for L2 acquisition, but how much of L1 frequency can filter down to a L2 competence, given such factors as the learner's stage of acquisition or readiness. Yet, it is possible that, especially at initial stages of acquisition, L1-L2 identical perfectives pairs could be more different than they appear at face value because they are represented and used differently by L2 learners and native speakers of Italian. Since comparing L1-L2 frequencies cannot answer the question of whether identical L1-L2 morphosyntactic forms are actually represented in the same manner, this paper focuses on an alternative way, which is described in the following sections.

2 Comparing probabilities is more telling than comparing frequencies in SLA

In the last years the attention of SLA researchers gradually shifted from frequency to probability as both an explanatory factor and a marker of language development. Frequency is a property of items in isolation (whether single words or *n*-grams), whereas probability refers to the company of words and reflects speakers' (and learners') implicit expectations about the likelihood that a word is preceded or followed by another word (Brezina et al 2015; Glablasova et al 2017). On the one hand it was suggested that learners' sensitivity to association and to transition probabilities among items in the L1 input impact L2 acquisition more than raw frequency (e.g., Baayen 2010; Gries 2015; Ellis 2016). On the other hand, it was proposed that learning based on contingency between items (henceforth contingency learning, CL) can measure the stages of L2 morphosyntactic development. CL is a probability-based mechanism for learning whether relations between events are causal or noncausal. Subjects tend to label co-occurring events as causal when the probability P of getting an outcome O (e.g., thunder) given a cue C (e.g., lightning) (P(O|C)) is very high. (Beckers et al., 2007, p. 289; Shanks, 2007). Causal cue–outcome relationships, once identified, trigger category formation. CL was found to impact on SLA: items that are highly expected by L2 learners in a given context seem to be acquired earlier and more easily than items that are more frequent but less contingent (e.g. Ellis & Ferreira Junior 2009) . A complementary idea is that a learner's proficiency increases, contingency should lose its developmental importance, as it will be clearer in Section 4. Mismatches or similarities between L1-L2 inflected forms should be interpreted in the light of a learner's developing sensitivity to co-occurrence probabilities. L2 learners are equipped with a much more sophisticated acquisitional device than a mere 'counter in the head'. This device keeps track also of the company of words, not only of the number of hits (Baayen 2010). In this article we focus on a developmental factor that mediates between the distributional features of the L1 input and the developing L2 morphosyntax. This factor is lexeme-morpheme contingency. Contingency-based association score ΔP (delta P) can be used to discriminate the degree of productivity-formulaicity of identical surface forms in L1-L2 morphosyntactic pairs.

3 Productivity and formulaicity

Degree of productivity and kind of formulaicity are two developmentally moderated dimensions that differentiate between L1 and L2 apparently identical forms, regardless of the frequency of these forms in the L1 input. Morpheme productivity is an affix's ability to generate new forms systematically. The more an affix participates in different lexemes, the more productive it is. In the literature, morpheme productivity is considered a direct function of the type frequency of the relevant constructions (e.g., Croft & Cruse, 2004, p. 309). The simplest formula that captures type frequency is *V (C;N)*, i.e., the type count *V* of the members of a morphological category *C* in a corpus of *N* tokens (e.g. Desagulier, 2016, p. 179). However, no morpheme can be said 'fully' productive. In fact, no one expects that speakers of a language with rich morphology – such as Italian – will encounter the full inflectional paradigm of a lexeme during their whole experience with that language. Since Zipf's law also applies to inflected forms (e.g. Bentz, 2014), we expect that in any corpus the number of lexemes that appear in all their forms is very small and approaches zero as the size of the paradigm expands (Janda & Tyers 2018).

'Formulaicity' is the opposite of productivity. Formulaic language is whatever multiword expression that is 'stored and retrieved whole from memory [...] rather than being subject to generation or analysis' (Wray, 2002, 9). If an inflected item is analysed into 'stem + affix', then the item is productive. If the item is not analysed, then it is a formula. Actually developmental linguists distinguish between 'initial formulaiticy' and 'competent formulaicity'. Initially, learners are sensitive only to the joint frequency of adjacent items, which ensures the storage and retrieval of unanalysed chunks and acquisitional formulas (Bardovi-Harlig 2009, p. 757; Erman, 2009: 326). As their proficiency increases, learners become sensitive also to the probabilities of association between items. Frequently co-occurring items form abstract representational schema or 'constructions' in learners' mind. Such constructions – unlike chunks and formulas – may include open slots, which are fixed positions that can be filled with items belonging to the same category (Wulff et al., 2009; Bartning et al. 2012). In a native speaker's competence, competent formulaicity and productivity are gradient-like properties that cohabit and may feed into each other cyclically. For example, according to Bybee (1985), a formula becomes a construction when speakers analyse it, that is, when they can recognize - in a given grammatical construction (in our case, the past perfective) - the fixed and the movable parts, respectively, the stem and the suffix. When the construction reaches a certain token-frequency threshold, it is stored again as an autonomous unit (a formula), after which the construction's productivity may diminish. Formulaicity in the competence of low proficiency L2 learners may have a different status though. While in mature languages formulaicity contributes to faster processing (Wiechman & Kerz, 2016) and does not indicate a defective morphological competence, initial formulaicity in SLA – when it concerns learners' use of inflected forms – may signal that the mechanism of contingency learning may be operating and temporarily constraining morpheme productivity. Indeed, formulaicity and productivity in this paper concern how stems and inflectional morphemes (in our case, the perfective) are represented in L2 Italian learners' competence. While 'initial formulaicity' characterizes the stage in development when inflected forms are still unanalysed, 'productivity' characterizes the stage when the morpheme is represented independently of the lexical entry (Rastelli 2020). This is the topic of next section.

4 Contingency as the inverse of morpheme productivity

As we have seen in Section 3, CL is a probability-based mechanism for learning whether relations between events are causal or noncausal). In L1 and L2 acquisition, CL supports category formation by exploiting the higher-than-average frequency of co-occurrence between a lexeme and a grammatical construction (e.g., the perfective morpheme). As long as CL drives acquisition, we expect that the morpheme is spread to limited number of lexemes, although the vocabulary at a learner's disposal may be larger. This restriction is the natural reaction of human cognition facing the paramount task of learning a new language. As long as the strength of association between a cue (e.g. a lexeme) and the outcome

(e.g. a morpheme) is one factor affecting the learning process (probably among other factors), category membership (e.g., the perfective) will be likely restricted to a few exemplars, even though one cannot exclude that learners will encounter also rare associations (less contingent lexeme-morpheme pairs). Paradoxically, such restriction to exemplars facilitates rather than hinders category learning. If early members of a category are very few and very similar to each other, it will be easier for learners to establish a category membership based on resemblance. At later stages of learning, a richer experience with the L1 and generalizations based on analogy will help L2 learners to break prior associations by introducing new members (e.g., new lexemes) in the category (e.g., the perfective). Therefore, as a learner becomes proficient, also morpheme-lexeme contingency is expected to diminish because the learners start to conceive the morpheme independently of the lexeme and starts to extend it to the available lexicon. In the remainder of the article, by 'morpheme productivity' we will mean precisely the inverse function of morpheme-to-lexeme contingency. The more a form is productive, the less the morpheme is expected to be contingent upon a restricted number of lexemes.

5 ΔP (Delta P)

Unidirectional, contingency-based association scale ΔP can be used to analyse the different degrees of morpheme productivity-formulaicity in apparently identical L1-L2 inflected pairs. ΔP is a one-way dependency statistic developed by Allan (1980) and already utilized in SLA (Ellis, 2006; ; Ellis & Ferreira-Junior, 2009; Ellis et al 2014; Wulff et al 2009). To our knowledge, ΔP scores have never been used to compare L1-L2 inflected forms. Ellis (2006: 11) defines ΔP as in (2):

(2) $\Delta P = p\,(\text{Outcome} \mid \text{Cue} = \text{present}) - p\,(\text{Outcome} \mid \text{Cue} = \text{absent})$

ΔP is the probability of the outcome given the cue (P(O|C)) minus the probability of the outcome in the absence of the cue (P(O|-C)). When these are the same (when the outcome is just as likely when the cue is present as when it is not) there is no covariation between the two events and ΔP = 0. ΔP approaches 1.0 as the presence of the cue increases the likelihood of the outcome and approaches −1.0 as the cue decreases the chance of the outcome. Unlike bidirectional association measures – such as log-likelihood. T-score or Mutual Information – and unlike chi-squared or Fisher-exact test, ΔP can separately assess each item's contribution to the overall strength of association by comparing two kinds of conditional probabilities between two items, such as the lexeme (in our case, a verb) and the morpheme (in our case, the perfective). The first conditional probability is reliance. It compares the relative frequency of the morpheme with the lexeme to the relative frequency of the morpheme without the lexeme. The second conditional probability is attraction. It compares the relative frequency of the lexeme with the morpheme to the relative frequency of the lexeme without the morpheme. The starting point for the calculation of ΔP is a contingency table like Table 2, where values *a* through *d* correspond – for example – to the co-occurrence frequencies between a verbal lexeme (x) and the perfective morpheme (y) in a learner corpus:

Table 2: A contingency table

	Presence of *y* (response)	**Absence of *y*** (no response)
Presence of *x* (cue)	a	b
Absence of *x* (no cue)	c	d

Cell (a) in Table 2 above corresponds to the number of responses (e.g. the perfective) given the cue (e.g., a verb lexeme). Cell (b) corresponds to all cues (e.g., instances of the lexeme *a*) without the response (the perfective). Cell (c) corresponds to the number of responses without the cue (all perfectives in the interviews except those of cell *a*); and cell (d) corresponds to all predicates uttered by the learner in the

whole corpus, not including the values (a) and (b). The formula for calculating ΔP has two halves (2a) and (2b):

(2a) [a/(a+b)] – [c/(c+d)]
(2b) [a/(a+c)] – [b/(b+d)].

The first half of the formula targets morpheme-to-lexeme reliance. It treats the lexeme as the Cue and the perfective morpheme as the Outcome (or 'response'). The formula shows the difference between the frequency of the perfective morpheme with and without the lexeme. The second half of the formula targets lexeme-to-morpheme 'attraction'. It treats the perfective morpheme as the Cue and the lexeme as the Outcome. This part of the formula shows the difference between the frequency of the lexeme with and without the perfective morpheme. High ΔP values of reliance indicate that the probability of getting the perfective given the lexeme is higher than the probability of getting the same perfective with different lexemes. High ΔP values of attraction indicate that the probability of getting the lexeme given the perfective morpheme is higher than the probability of getting the same lexeme with other morphemes.

Unlike bidirectional association scores, ΔP can distinguish between the situation in which the perfective relies exceptionally upon the lexeme and the situation in which that lexeme attracts the exceptionally high frequency of the perfective. Two further advantages of using ΔP are highlighted by Gries (2015b). First, ΔP is very easy to compute: unlike many traditional measures it makes no distributional assumptions (normality, variance homogeneity, etc.), it involves neither complicated formulae nor computationally intensive exact tests. Moreover, since it is proportion-based, ΔP does not depend on the corpus size, so one does not run the risk of conflating the frequency of tokens co-occurring in a corpus and the effect size of such association (Gries 2019, p. 3). Second, unlike many other statistics, ΔP is easy to understand because it involves nothing but a mere difference of percentages. Δ*P* is a scale (based on proportions), not a test of significance, so there is no minimal threshold value for the measure to be meaningful (like the *p*-value, Ellis, 2012, p. 28). Therefore one may wonder how the significance of ΔP differences can be tested. We face this problem in the next section.

6 Case study: Rationale of the method and the dataset

In this section, we illustrate the utility of ΔP for SLA research through the empirical analyses of corpus data. In the previous sections, we illustrated the importance of ΔP scores for studying language development. In this and the following sections we propose a method to correlate L1 and L2 ΔP scores. The rationale of the method is that ΔP scores can measure contingency, e.g. the degree of association between the morpheme and the lexeme. Studying contingency is important because a difference in contingency scores may indicate that L2 learners and native speakers represent identical forms in different ways. High contingency scores may indicate that the morpheme is not yet represented as being independent of the lexeme. Low contingency scores may indicate that the morpheme is conceived as independent of the lexical entry and is spread to all the available verbal lexicon. Specifically, we explored lexeme–morpheme contingency in the corpora Pavia and in ItTenTen. The Corpus Pavia is the largest and best-known longitudinal learner Italian corpus to date (~700,000 tokens, 19,000 types overall). The corpus was collected between the mid-1980s and the late 1990s in Northern Italy, at the University of Pavia, Bergamo and Milan (Giacalone Ramat, 2003). It contains transcriptions of about 120 hours of oral interviews of 22 Italian L2 learners from 11 different L1 backgrounds spanning five typological families. All learners (aged 12-48 years, mostly 20-30 years) were residents in Italy at the time of interview, with different amounts of instruction and lengths of residence. All but five learners had attended Italian language courses prior to arrival. Corpus editors rated learners' proficiency basing on Klein and Perdue (1993). 'Prebasic' learners utilized mainly or exclusively pragmatic criteria (topic-comment, given-new) to build utterances. 'Basic' learners could go beyond information structure and take into account the

argument structure of predicates and theta-roles. Finally, 'postbasic±' learners were aware of the fundamental morphological oppositions in Italian (e.g., perfective vs. imperfective, singular vs. plural). Nineteen learners were rated as post-basic, one as pre-basic, and two were rated as basic. During the interviews, learners engaged in spontaneous and semi-structured conversations and tasks with Italian interviewers. Conversation topics varied across both learners and interviews and included everyday life, cultural differences, countries of origin, leisure activities, interpersonal relations, and features of Italian. Supervised elicitation tasks were also used at times and included description of pictures and oral retelling of picture-stories and video excerpts from the film Modern Times. Table 3 reports the complete breakdown of the corpus Pavia.

Table 3: The corpus Pavia. Language-related and learner-related variables

Learner	N. Token	Standardised type/token ratio	N. verb token	N. perfective token	Proficiency level	L1	Age	Number of interviews	Time span of interviews (months)	Months in Italy at first interview
AB	59,000	26	1778	489	Post-basic +	Tigrinya	21	12	7	12
AL	33,000	26	820	180	Post-basic+	French	21	5	2	30
AN	39,000	26	1192	203	Post-basic+	German	20	5	7	2
CH	72,000	24	1519	268	Post-basic-	Wú Chinese	17	19	16	11
EO	28,000	29	603	220	Post-basic+	English	27	11	5	4
FA	28,000	26	680	150	Post-basic+	Arabic	29	4	2	24
FD	6,000	28	108	10	Post-basic-	Mandarin		4	3	24
FI	48,000	27	1184	410	Post-basic+	English	26	13	5	5
FR	52,000	24	2887	773	Post-basic+	German	48	12	14	7
HG	19,000	21	124	32	Pre-basic	Tygrinia	15	7	5	1
JO	35,000	26	330	97	Post-basic-	English	20	9	5	1
MA	26,000	29	651	71	Post-basic+	Chichewa	18	4	1	8
MH	19,000	32	1006	269	Post-basic+	Albanian	19	6	5	3
MI	25,000	29	732	84	Post-basic+	Chichewa	18	4	1	8
MK	46,000	24	1446	438	Basic	Tygrinia	20	12	7	1
MT	17,000	32	1104	347	Post-basic+	German	22	8	10	1
PE	27,000	24	740	197	Basic	Malaysian	25	17	8	1
ST	7,000	27	234	62	Post-basic-	Mandarin	16	5	6	1
TU	12,000	30	643	177	Post-basic-	Wú Chinese	45	11	7	56
UL	20,000	31	1272	315	Post-basic+	German	33	8	10	3
WZ	8,000	28	200	31	Post-basic+	Wú Chinese	38	7	6	24
XI	92,000	24	2856	603	Post-basic+	Wú Chinese	12	18	12	18

After lemmatization and part-of-speech-tagging with Sketch Engine,[2] each learner's file comprised between 124 and 2,887 finite verb forms (983 on average), totalling 22,109 verb tokens and 5,540 perfective tokens stemming from 304 perfective types. For the present study, we selected a sample of 39 perfectives with raw frequency ≥ 25. The resulting sample contained 3,940 perfective tokens. The most frequent perfective token in the sample was *detto* 'said', which occurred 276 times. The least frequent perfective was *comprato* 'bought', which occurred 12 times. Table 4 reports the complete list of the sampled perfective predicates

ITTenTen16[3] is a 4.9 billion-word web corpus (downloaded by SpiderLing[4] from May to August 2016) made up of texts collected from the Internet (. The corpus is a part of the TenTen corpus family, a set of

[2] https://auth.sketchengine.eu
[3] https://www.sketchengine.eu/ittenten-italian-corpus/
[4] SpiderLing — a web spider for linguistics — is software for obtaining text from the web for building text corpora. See http://corpus.tools/wiki/SpiderLing

the web corpora built using the same method. It includes a wide range of registers and text types, and thus makes it possible to examine a considerable amount of highly heterogeneous data. The resulting sample contained 21,484,341 perfective tokens.

Table 4: Raw frequency and rank of the sampled 39 perfective predicates in the corpus Pavia

Rank	Verb	Frequency	Rank	Verb	Frequency
1	*detto* 'said'	477	21	*dimenticato* 'forgot'	53
2	*fatto* 'done'	437	22	*sposato* 'married'	51
3	*andato* 'gone'	420	23	*imparato* 'learned'	51
4	*visto* 'seen'	323	24	*caduto* 'fallen'	50
5	*capito* 'understood'	215	25	*uscito* 'exited'	46
6	*arrivato* 'arrived'	184	26	*lasciato* 'left'	44
7	*preso* 'taken'	138	27	*passato* 'passed'	43
8	*trovato* 'found'	129	28	*tornato* 'returned'	42
9	*parlato* 'talked'	99	29	*lavorato* 'worked'	42
10	*venuto* 'come'	96	30	*voluto* 'wanted'	42
11	*finito* 'finished'	92	31	*telefonato* 'phoned'	41
12	*sentito* 'heard'	77	32	*cambiato* 'changed'	38
13	*pensato* 'thought'	69	33	*piaciuto* 'liked'	36
14	*scritto* 'written'	69	34	*messo* 'put'	35
15	*mangiato* 'eaten'	65	35	*sbagliato* 'mistaken'	34
16	*portato* 'brought	61	36	*studiato* 'studied'	31
17	*morto* 'died'	58	37	*dormito* 'slept'	30
18	*chiesto* 'asked'	58	38	*incontrato* 'met'	30
19	*dato* 'given'	55	39	*comprato* 'bought'	25
20	*perso* 'lost'	54			

7 Regressing L2 ΔP scores against L1 ΔP scores

The first distinctive feature of our method is that it allows assessing whether the ΔP values of the 39 perfectives in the L1 corpus (ItTenTen) are systematically related to those in the L2 corpus (Pavia). An intuitive way to approach it is to build a simple regression model predicting L2 ΔP values based on L1 ΔP values. Indeed, the analysis shows that the two are significantly and positively associated in both reliance (B = 0.98; SE = 0.29; $t[37] = 3.33$; $p = .002$; $R^2 = 0.23$) and attraction (B = 0.39; SE = 0.10; $t[37] = 3.90$; $p < .001$; $R^2 = 0.29$). A potential issue with this approach, however, is that it disregards the differences in the uncertainty between individual data points (see also Murakami, 2020). For instance, the denominator of the first term of reliance computation (i.e., a + b in Table 2) in L2 ranges from 43 to 1,430 in our data set, depending on the overall frequency of the perfective (i.e., *sbagliato* 'mistaken' vs *detto* 'said', respectively). This suggests that the reliance of *sbagliato* is less reliable and is likely to exhibit larger sampling variability than that of *detto*. This difference in uncertainty between perfectives should ideally be reflected on the uncertainty in the estimated parameter values (e.g., slopes of regression models) in order to fully exploit the information in the data. One way to incorporate the information is the simulation-based method described below. The data and R code used for the analysis are available at https://osf.io/sqcjk/.

Our analysis is based on computational simulation, in which we built a large number of regression models based on the sets of ΔP values arguably representing sampling variability, and examined the variability in the estimated parameter values. If the variability of the slope parameter is small and excludes 0 in most of the models, then we can conclude that L1 ΔP values are significantly associated with L2 ΔP values. Specifically, in order to evaluate the reliability of reliance (i.e., $\frac{a}{a+b} - \frac{c}{c+d}$ in Table 2), the nominator of each term (i.e., a and c), or the count of perfectives, was assumed to be distributed

binomially with the corresponding denominator (i.e., a + b and c + d) as the number of trials.[5] A ΔP value is basically a difference between the two success probabilities of the two binomial distributions. Based on the assumptions above, we followed the procedure below:

1. For each term of reliance in each perfective in each of L1 and L2, one value was randomly drawn from a standard normal distribution.
2. The limit of the Wilson score interval (Wilson, 1927), which is one form of confidence intervals, of binomial distribution was calculated with the value selected above used as the z-score[6] to determine confidence levels. For example, if the value selected in Step 1 happened to be 1.96, the upper limit of the 95% confidence interval of the corresponding term (i.e., $\frac{a}{a+b}$ or $\frac{c}{c+d}$) was calculated.
3. In each perfective in each of L1 and L2, simulated reliance values were computed based on 2.
4. We fit a regression model predicting L2 reliance based on L1 reliance and recorded the model's estimated intercept, slope, and the p-value of the slope.
5. The above process was repeated 1,000 times, resulting in 1,000 regression models and associated statistical values.

In the above process, a perfective with smaller denominator values (e.g., *sbagliato*) invites larger variability across the 1,000 iterations in their simulated reliance values because the confidence intervals (CIs) of the component proportions of the perfective are larger. For example, the word *sbagliato* has 43 as the denominator of the first term, and 34 as the nominator (i.e., Cell a in Table 2) in the original data set. The 95% CIs of the proportion (i.e., 34/43, or 0.79) run from 0.65 to 0.89, with the span of 0.24. On the other hand, iteration-by-iteration variability is much smaller when the denominator is large. In the case of the word *detto*, the denominator is 1,430 while the nominator is 477. The 95% CIs of the proportion (i.e., 477/1,430, or 0.33) are much narrower and run from 0.31 to 0.36, with the span of 0.05. This difference in the uncertainty (as represented by the width of CIs) is directly reflected on the uncertainty in the resulting simulated reliance values. Reliance values with large uncertainty, in turn, contribute to larger iteration-by-iteration variation in parameter estimates. The distribution of the parameter estimates, therefore, reflects uncertainty in each perfective. A similar procedure was followed in attraction (i.e., $\frac{a}{a+c} - \frac{b}{b+d}$) as well, except that the nominators are now Cells a and b in Table 2 and the denominators are a + c and b + d.

In order to examine the validity of the above analytical procedure, we performed two computational experiments. The first experiment was conducted to test whether our method accurately rejects the cases where there is no true (i.e., population-level) relationship between L1 and L2 ΔP values, while the second one investigated potential bias (i.e., difference between the true value and the estimated value) in the estimation of two focal parameters in the regression models, the intercepts and slopes. Specifically, in the first experiment examining the false-positive rate of our method, we randomized the pairings of L1 and L2 ΔP values. In other words, in the data analysis reported in the main text, the ΔP value of a particular perfective in L2 was naturally matched with that of the same perfective in L1 (e.g., reliance of *detto* in L2 was paired with reliance of *detto* in L1). Here, we broke the relationship, and ΔP values of perfectives in L2 were matched with those of randomized perfectives in L1 (e.g., reliance of

[5] Binomial distribution represents the number of occurrences of a certain outcome in a series of what is called Bernoulli trials, which are a series of independent trials with two possible outcomes (e.g., head and tail in coin tosses) with a certain probability of success (e.g., head). In the calculation of reliance in this study, two outcomes correspond to whether a perfective was used or not in the target lemma (a in Table 2) or in any other verbs (c), and a trial corresponds to each occurrence of the lemma (a + b) or each occurrence of the other verbs (c + d).

[6] The z-score represents the number of standard deviations away from the mean.

detto in L2 might have been paired with reliance of *venuto*). The idea is that since the relationship between L1 and L2 ΔP values was randomized, there should not be any true systematic pattern between L1 and L2 ΔP values. We then applied the statistical method explained in the main text to the randomized data set and examined if it correctly fails to identify systematic patterns. Specifically, the above randomization procedure was repeated 1,000 times, and for each iteration, we calculated the proportion in which the p-value of the regression slope representing the relationship between L1 and L2 ΔP values is significant at the significance level of $\alpha = .05$. The results showed that the proportion of the significant slope parameter was 0.032 for reliance and 0.060 for attraction. Although the value of attraction is slightly higher than the nominal significance level ($\alpha = .05$), it is fair to consider the false-positive rate of the method used in our study to be within an acceptable range.

Now that we have confirmed that our method is likely to accurately detect true patterns, we next investigated to what extent there is a bias in the estimated parameter values. If the bias in the estimation of the slope parameter is large, we should not make a strong statement on how much change of ΔP in L2 is associated with a change of a certain value (e.g., 0.3) of ΔP in L1. The way we estimate potential bias is through a parameter recovery experiment, in which we create a fake dataset with known true parameter values and investigate whether our method correctly recovers the true values. The details of the simulation are in the Appendix. The results of the simulation showed that the mean absolute bias (i.e., difference between the mean estimated parameter value and the true parameter value) was 0.02 with the 95% quantile of the bias ranging from -0.05 to 0.05 in the intercept parameter of reliance. Similarly, the mean absolute bias was 0.03 with the 95% quantile ranging from -0.08 to 0.09 in the slope parameter of reliance. The corresponding values in attraction were identical to the above within the reported digits. The bias in parameter estimation is, therefore, fairly small (< 0.10) in so far as location parameters (i.e., intercept and slope) are concerned. The two validation experiments above, therefore, demonstrate that our analytical method rarely results in false positives (i.e., misidentifying the cases where there is no relationship between L1 and L2 ΔP values as the ones where there is) and that bias in the estimates of the intercept and the slope is within an acceptable range (i.e., the difference between the estimated values and true values is small in the intercept and the slope).

The results of the simulation-based analysis are summarised in Table 5. All the slopes, both in reliance and attraction, were statistically significant, indicating that L1 and L2 ΔP values are positively associated in both reliance and attraction. Specifically, one-unit change in L1 reliance is associated with the change of 0.97 in L2 reliance. The relationship is slightly weaker in attraction, and one-unit change in L1 attraction is associated with the change of 0.39 in L2 attraction.[7] The intercepts were non-significant at all in reliance and were only significant in 42% of the cases in attraction. This suggests that there is no significant difference between L1 and L2 ΔP values.

8 Pairwise comparison of L1-L2 perfectives

The second distinctive feature of the method is that it can identify exactly which perfectives are similar between L1 and L2 in their ΔP values and which ones instead exhibit significantly different values. We have shown above that, overall, no significant difference is observed in our data set between L1 and L2 ΔP values. This, however, pertains to the overall pattern and does not mean that no significant difference is found in any of the 39 target perfectives (see Table 4). In order to perform by-perfective pairwise comparison between L1 and L2 ΔP values, we employed a similar simulation technique to the one used above. Specifically, the following procedure was used.

[7] Figure 3 points to the possibility that the slope in attraction is largely driven by an outlier (i.e., *fatto* 'done' at the top right corner). To assess the potential impact of the outlier, we excluded the word and repeated the same procedure. The results largely remained the same, except that 1.7% of all the iterations returned non-significant slope coefficients. Given the small percentage, however, it is still safe to conclude that L1 and L2 attraction values are significantly positively associated.

Table 5. Summary of the Simulation-Based Analysis of Δ*P*

ΔP	Parameter	Mean Estimate	95% Quantiles	Mean SE	Mean t-value	Mean p-value
Reliance						
	Intercept	0.071	[0.046, 0.096]	0.066	1.073	0.298
	Slope	0.971*	[0.837, 1.095]	0.298	3.267	0.003
Attraction						
	Intercept	0.007	[0.006, 0.008]	0.004	1.995	0.056
	Slope	0.385*	[0.336, 0.439]	0.100	3.863	< 0.001

Note. Asterisks (*) indicate that all the 1,000 iterations returned statistically significant results with α = .05.

1. The same procedure as Steps 1 through 3 in the earlier analysis was followed here as well. Namely, simulated reliance (or attraction) values were computed based on the CIs of binomial distribution. The limits of the intervals were determined based on the values randomly drawn from a standard normal distribution.
2. The difference between L1 and L2 ΔP values was computed.
3. The above two steps were repeated 1,000 times, resulting in 1,000 differences in ΔP values in each perfective.
4. For each perfective, the proportion in which those differences exceeded 0 or were below 0 (depending on the direction of the difference between observed L1 and L2 values) was calculated and multiplied by two for the sake of two-tailed tests. The resulting value was considered as the p-value indicating the significance of the difference between the two.
5. The resulting p-values were finally adjusted for multiple comparisons via Hommel's (1988) method within each ΔP type (i.e., reliance or attraction).

We tested the validity of the above procedure by examining its false-positive rates and statistical power. Specifically, in order to test the false-positive rates, we generated L1 and L2 fake data based on the same underlying Δ*P* values and examined whether our method correctly fails to identify the difference as significant. The concrete procedure is explained in the Appendix. The results showed that the mean false-positive rate was 0.050 when the denominators of reliance were used as the number of trials, whereas it was 0.048 when those of attraction were used. They were both smaller than the nominal significance level, which suggests that the false-positive rate of our method is within an acceptable range.

Now that we have confirmed that the significant results we find through the method we employed are likely to reflect true underlying differences between L1 and L2 ΔP values, we now test the reverse, that is, whether we can accurately identify differences when there are differences at the population level. The procedure of the validation is similar to the above, except that some values were added to L1 Δ*P* values in the creation of underlying L2 Δ*P* values. Specifically, the values were randomly selected from a uniform distribution between 0.1 and 0.4, and in 20 out of the 39 perfectives, the values were subtracted from L1 Δ*P* values, while in the remaining 19 cases, they were added to L1 values. The results showed that the difference was successfully identified through the simulation-based method employed in our study at the rate of 98% when the denominators of reliance were used in the simulation, and 100% when those of attraction were used. The statistical power of our method, therefore, can be considered as high when the difference between L1 and L2 Δ*P* values is 0.1 or larger. Our validation experiments, therefore, suggest that both Type I and Type II error rates are reasonably low.

The results of our analysis indicated that 26 out of 39 perfectives (67%) show significant differences between L1 and L2 values in reliance, while 24 perfectives (62%) demonstrate significant differences in attraction. In other words, ΔP values differ between L1 and L2 in the majority of the target perfectives. The stark contrast between the findings here and those reported in the earlier section suggests that there is no consistent difference between L1 and L2 ΔP values (e.g., L1 ΔP values are generally larger than L2 ΔP values) but that some words mark higher ΔP values in L2 than in L1, while other words exhibit the reverse pattern. Figure 1 illustrates in which perfectives significant differences were observed.

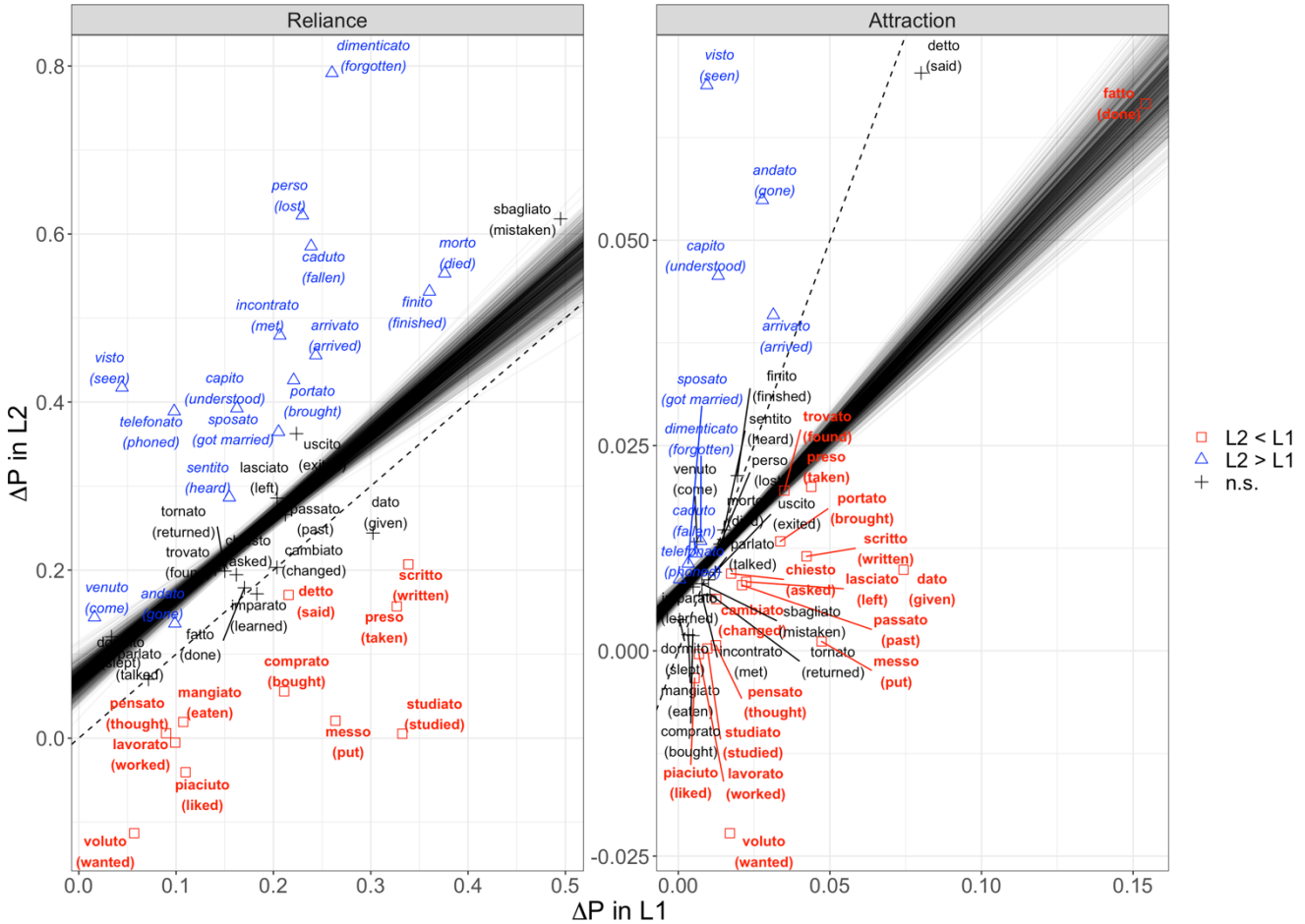

Figure 1. ΔP values in L1 and L2 and whether their values significantly differ from each other. The dashed line in each panel represents the values where L1 and L2 ΔP values are identical. Thin lines are the 1,000 regression lines estimated in the earlier analysis. Bold face represents the words in which L1 ΔP values are significantly larger than L2 ΔP values, while italics represents those with the reverse pattern (i.e., L2 ΔP values > L1 ΔP values).

9 Interpretation of results

The method presented above is aimed to discriminate between L1 and L2 perfectives that have comparable contingency values and those which differ significantly. The results illustrated by Figure 1 bear directly on SLA theory. For reasons of space, we limit ourselves to three short observations about how the data can be interpreted. First, the eight L2 perfectives that show the most exceptional reliance values (e.g. *dimenticato* 'forgotten', *perso* 'lost', *morto* 'died', *caduto* 'fallen', *incontrato* 'met', *finito* 'finished', *arrivato* 'arrived', *capito* 'understood') are all telic lexemes. Proponents of the Aspect Hypothesis (AH, Andersen & Shirai, 1994) would probably claim that the reason for such exceptional morpheme-to-lexeme reliance resides in the semantics of the lexeme, namely, in its lexical aspect, since

most early perfectives are telic (although the AH predicts that the same pattern should be observed in L1, for a recent review see Comajoan & Bardovi-Harlig 2020). Moreover, according to the AH, such morpheme-lexeme association would characterize only the early stages of acquisition. Therefore, the reliance values are expected to decrease as learners' proficiency increases, which seems in fact to be the case (Rastelli, 2020). Interestingly, rank correlation analysis shows that the most reliable L2 predicates do not correspond to those predicates having similar L1 vs L2 frequency (Kendall $\tau = 0.21$; p-value = 0.407). Therefore, high reliance of L2 predicates seems independent of the frequency of the corresponding perfective predicates in the L1 input. Second, there is a group of lexemes that attract the perfective morpheme in L1 but not in L2 Italian. These are general-purpose lexemes that are very frequent in both L1 and L2, such as *fare*, do', *prendere* 'take', *mettere* 'put', *dare* 'give'. These predicates often occur in light-verb constructions (such as *fare colazione* 'have breakfast', *fare una passe*ggiata 'go for a stroll', *fare la doccia* 'have a shower') and enter also many multi-word units (e.g. *prendere alla lettera* 'take literally', *prendere a cuore* 'take to heart') just because of their flexibility and genericity of meaning. Exceptional lexeme-to-morpheme attraction signals that those verbs occur especially in formulaic expressions in the past. It is worth investigating whether and why such perfectives – although their frequency in L1 and L2 corpora is comparable – are formulaic in L1, but not in L2 Italian. The method can probably unveil cases of 'competent formulaicity' which is still not available to initial L2 learners. As we have seen (Section 3), competent formulaicity refers to a native speaker's capacity of processing (understanding and producing) collocations in real time (Section 3). Such capacity can be defective in early interlanguages. Finally, the method can be useful also to assess whether inflected forms are actually analysed by the learners or they are rather unanalysed 'acquisitional formulas' (§4), maybe prompted by the interlocutor or requested by the elicitation task. For example, many motion verbs in the corpus Pavia (*venire* 'come', *arrivare* 'arrive', *andare* 'go') probably rely exceptionally on the perfective morpheme not because learners know the perfective past tense, but because all participants were asked to tell how they arrived to Italy as immigrants and those verbs were reprised directly from the interviewer's question (*come <u>sei arrivato</u> qui* 'How did you get here?'). Also the lexemes *vedere* 'see' and *capire* 'understand' rely to the perfective morpheme probably not because of their semantics, but because they work as routinary markers in the turn-taking system (e.g. *non ho capito* 'I did not understand') that holds between the L2 learner and the native interviewer.

10 Domains of application (among many others)

Besides the examples described above, the method presented in this article is replicable in other SLA domains. One may concern the study of L2 acquisition of free and bound morphemes. For example, the analysis of ΔP may reveal if the choice between auxiliaries *avere* 'have' and *essere* 'be' in the past compound in Italian – especially at early stages of acquisition – is contingent upon a restricted number of telic or of strongly agentive predicates, as predicted by the Auxiliary Selection Hierarchy (Sorace 2004). Also article-noun agreement in the DP and noun-adjective agreement in the NP may be subject to lexeme-morpheme contingency, so that one should not expect that – even when the rule seems acquired – a learner can spread it to all the available lexicon. Another domain of application of the method is the study of grammatical constructions. As we have seen (§3), initial (low proficiency) learners store unanalysed chunks and acquisitional formulas. With increasing proficiency, their competence can encompass also abstract representational schema or 'constructions'. Such constructions – unlike chunks and formulas – may include open slots, which are fixed positions that can be filled with items belonging to the same category . The method can allow researchers to link a decrease of contingency score to the developmental shift from chunks to schemas with open slots. In these and in other cases, the method has the potential to uncover the presence of many kinds of asymmetries in how in L1-L2 inflected items are represented. Such asymmetries are at risk of remaining totally invisible if one considers only frequency, distribution and rank of L2 predicates.

References

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, 15(3), 147–149.

Andersen, Roger & Shirai, Yasuhiro. (1994). Discourse motivations for some cognitive operating principles. *Studies in Second Language Acquisition*, 16(2), 133-156.

Baayen, H. 2010. Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon*, 5(3), 436-461

Bardovi-Harlig, K. 2009. Conventional Expressions as a Pragmalinguistic Resource: Recognition and Production of Conventional Expressions in L2 Pragmatics. *Language Learning* 59/4, 755-795.

Bardovi-Harlig, K. & Comajoan-Colomé, L. 2020. The Aspect Hypothesis and the acquisition of L2 past morphology in the last 20 years. A state-of-the-scholarship review. *Studies in Second Language Acquisition*, doi:10.1017/S0272263120000194

Bartning, I., Forsberg Lundell, F. & Hancock, V. 2012. On the role of linguistic contextual factors for morphosyntactic stabilization in high-level L2 French. *Studies in Second Language Acquisition* 34, 243-267.

Beckers, Tom, De Houwer, Jan, & Matute Helena. 2007. Editorial: Human contingency learning, *The Quarterly Journal of Experimental Psychology*, 60 (3). 289–290.

Bentz, C. 2014. Zipf's law and the grammar of languages: A quantitative study of Old and Modern English parallel texts. *Corpus Linguistics and Lingustic Theory*. DOI: 10.1515/cllt-2014-000

Brezina, V., McEnery, T., Wattam, S. 2015. Collocations in context. A new perspective on collocation networks, in *International Journal of Corpus Linguistics*. 20, 2, p. 139-173.

Bybee, J. 1985. *Morphology: A study of the relation between meaning and form*. Amsterdam: John Benjamins.

Corder, P. 1967. The significance of learners' errors. *International Review of Applied Linguistics – RILA*, 5, pp.161-170

Croft, W. & Cruse, D. A. 2004. *Cognitive linguistics*. Cambridge–New York: Cambridge University Press.

Desagulier, G. 2016. A lesson from associative learning: Asymmetry and productivity in multiple slot constructions. *Corpus Linguistics and Linguistic Theory*, 2(2), 173–219.

Dulay, H., Burt, M., & Krashen, S. 1982. *Language Two*. Oxford: Oxford University Press.

Ellis, N. 2006. Language acquisition as rational contingency learning. *Applied Linguistics*, 27(1). 1-24.

Ellis, N. 2012. Formulaic Language and Second Language Acquisition: Zipf and the Phrasal Teddy Bear. *Annual Review of Applied Linguistics* 32. 17-44.

Ellis, N. C. 2016. Cognition, corpora, and computing: triangulating research in usage-based language learning. *Language Learning*, 67(51), 40–65.

Ellis, N. & Ferreira-Junior, F. 2009. Construction learning as a function of frequency, frequency distribution, and function. *The Modern Language Journal*, 93(3). 370-385.

Ellis, N., O'Donnell, M.B. & Römer, U. 2014. Second language verb-argument constructions are sensitive to form, function, frequency, contingency, and prototypicality. *Linguistic Approaches to Bilingualism*. 4(4), 405-431.

Erman, B. 2009. Formulaic Language from a learner perspective. What the learner needs to know. In R. Corrigan - E.A. Moravcsik - H. Ouali - K. M. Wheatly (eds.), *Formulaic Language, vol. 2*, Amsterdam-Philadelphia, John Benjamins Publishing Company, 323-346.

Gass, S. 2013. *Second Language Acquisition. An Introductory Course*. New York – Londn, Routledge.

Giacalone Ramat, Anna. (Ed.) 2003. *Verso l'italiano. Percorsi e strategie di acquisizione*. Roma: Carocci.

Gablasova, D., Brezina, V., McEnery, A.M. 2017. Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence, *Language Learning*. 67, p. 130-154.

Gluck, M. & Bower, G. 1988. From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology*, 117(3). 227–247.

Gries, S. Th. 2015a. Statistics for learner corpus research. In G. Gilquin, S. Granger, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 159-182). Cambridge: Cambridge University Press.

Gries, S. Th. 2015b. 50-something years of work on collocations. What is or should be next. In S. Hoffmann, B. Fischer-Starcke & A. Sand (eds.) *Current Issues in phraseology*, Amsterdam-Philadelphia, John Benjamins Publishing Company.

Gries, S. Th. 2019. 15 years of collostructions: some long overdue additions/corrections (to/of actually all sorts of corpus-linguistics measures). In S. Hunston and F. Perek (eds.) *Constructions in Applied Linguistics,* Special Issue of *International Journal of Corpus Linguistics* 24(3), 385-412.

Hommel, G. 1988. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, *75*(2), 383–386. doi: 10.2307/2336190.

Janda, A. L., & Tyers, M. F. 2018. Less is more: why all paradigms are defective, and why that is a good thing. *Corpus Linguistics and Linguistic Theory*. doi: 10.1515/cllt-2018-0031

Kilgarriff, A. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6/1, 97-133.

Kilgarriff, A. 2005. Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 12 (2005), 263275

Murakami, A. (2020). On the sample size required to identify the longitudinal L2 development of complexity and accuracy indices. In W. Lowie, M. Michel, A. Rousse-Malpat, M. Keijzer, & R. Steinkrauss (Eds.), *Usage-based dynamics in second language development* (pp. 20–49). Bristol: Multilingual Matters.

Norris, J., & Ortega, L. 2003. Defining and measuring SLA. In C. Doughty & M. Long (Eds.), *The Handbook of Second Language Acquisition* (pp. 716–761). London: Wiley-Blackwell.

Pallotti, G. 2007. An operational definition of the emergence criterion. *Applied Linguistics*, 28(3), 361–382.

Pienemann, M. 1998. *Language processing and second language development: Processability theory*. Amsterdam: Benjamins.

Rastelli, S. 2020. Contingency learning and perfective morpheme productivity in L2 Italian: A study on lexeme–morpheme associations with ΔP, *Corpus Linguistics and Linguistic Theory*, https://doi.org/10.1515/cllt-2019-0071

Shanks, David. 2007. Associationism and cognition: Human contingency learning at 25. *Quarterly Journal of Experimental Psychology*, 60(3). 291–309.

Sorace A., 2004. Gradience at the Lexicon-Syntax Interface: Evidence from Auxiliary Selection and Implications for Unaccusativity. In Alexiadou, A., Anagnostopoulou, E. & Everaert, M., (Eds.), *The Unaccusativity Puzzle. Exploration of the Syntax-Lexicon Interface*, Oxford - New York: Oxford University Press, pp. 243-268.

VanPatten, B. & Benati, A. 2015. Key Terms in Second Language Acquisition. London, Bloomsbury

Wiechman, D. & Kerz, E. 2016. Formulaicity as a determinant of processing efficiency: investigating clause ordering in complex sentences. *English Language and Linguistics* 20.3: 421–437

Wilson, E. B. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*(158), 209–212. doi:10.1080/01621459.1927.10502953

Wray, A. 2002. *Formulaic Language and the Lexicon*, Cambridge, Cambridge University Press

Wulff, S., Ellis, N., Bardovi-Harlig, K., Leblanc, C. J., & Römer, U. 2009. The acquisition of tense-aspect: Converging evidence from corpora and telicity ratings. *The Modern Language Journal*, 93(3), 354–369.

APPENDIX

Validation of the Statistical Methods Employed in the Study

Bias in Parameter Estimation

Below is the procedure we took in order to examine the extent to which bias in parameter estimation is introduced in our first simulation-based method, which we used to investigate whether L1 and L2 ΔP values are significantly associated with each other.

1. The true values of the parameters of regression models were randomly selected in the following manner:
    a. The ΔP values of 39 perfectives in L1 were randomly selected from a uniform distribution between -1 and +1.
    b. The intercept parameter and the slope parameter were each randomly drawn from a normal distribution with the mean of 0 and the standard deviation (SD) of 0.5.
    c. The SD parameter was randomly drawn from a normal distribution with the mean of 0 and the SDs of 0.2. The absolute value was taken in order to avoid negative values.
    d. The ΔP values in L2 were computed by $\alpha + \beta \times$ L1 ΔP values + noise, where α represents the intercept, β represents the slope, and noise was randomly drawn from a normal distribution with the mean of 0 and the SD determined in the Step 1-c above.
    e. Whether all the 39 L2 ΔP values determined above fall within their theoretical limits (i.e., between -1 and +1) was checked, and if not, Steps 1-b through 1-d were repeated until all the L2 ΔP values fell between -1 and +1.
2. A fake data set based on the true values determined above was created in the following manner:
    a. For each of L1 and L2, two values (θ_1 and θ_2) representing the true probabilities of the two terms of ΔP calculation (e.g., $\frac{a}{a+b}$ and $\frac{c}{c+d}$ in reliance) were selected for each perfective. Their values were randomly selected from uniform distributions with the condition that they satisfy the definition of ΔP (e.g., reliance = $\theta_1 - \theta_2$) determined above.
    b. For each of L1 and L2, the nominators involved in ΔP calculation (e.g., a and c in reliance) were randomly selected based on binomial distributions with denominators in the original data set (e.g., a + b or c + d in reliance) as the number of trials and the true probability above (θ_1 or θ_2) as the success rates.
3. The simulation-based statistical analysis described in the main text was applied to the data set with the new nominators selected above, and the difference between the mean parameter value in the 1,000 iterations and the true value selected in Step 1-b was recorded.
4. Steps 1 through 3 were repeated 1,000 times for each of reliance and attraction.

The idea is to select true parameter values within their likely ranges (Step 1) and to identify bias in parameter estimation by comparing those true values and the values that were estimated through our simulation-based method based on the fake data set created with those true values (Steps 2 and 3). The entire process was repeated a large number of times to ensure the reliability of this validation procedure (Step 4). Figure A1 shows the extent to which estimates are biased across different true parameter values. The figure indicates that the bias does not generally exhibit systematic patterns, which suggests that bias and true values are independent and that our analysis is generally credible regardless of the true value. Note that slightly larger variance of bias around the true value of 0 in the slope parameter is presumably because it is where more sampling took place (see Step 1-b above).

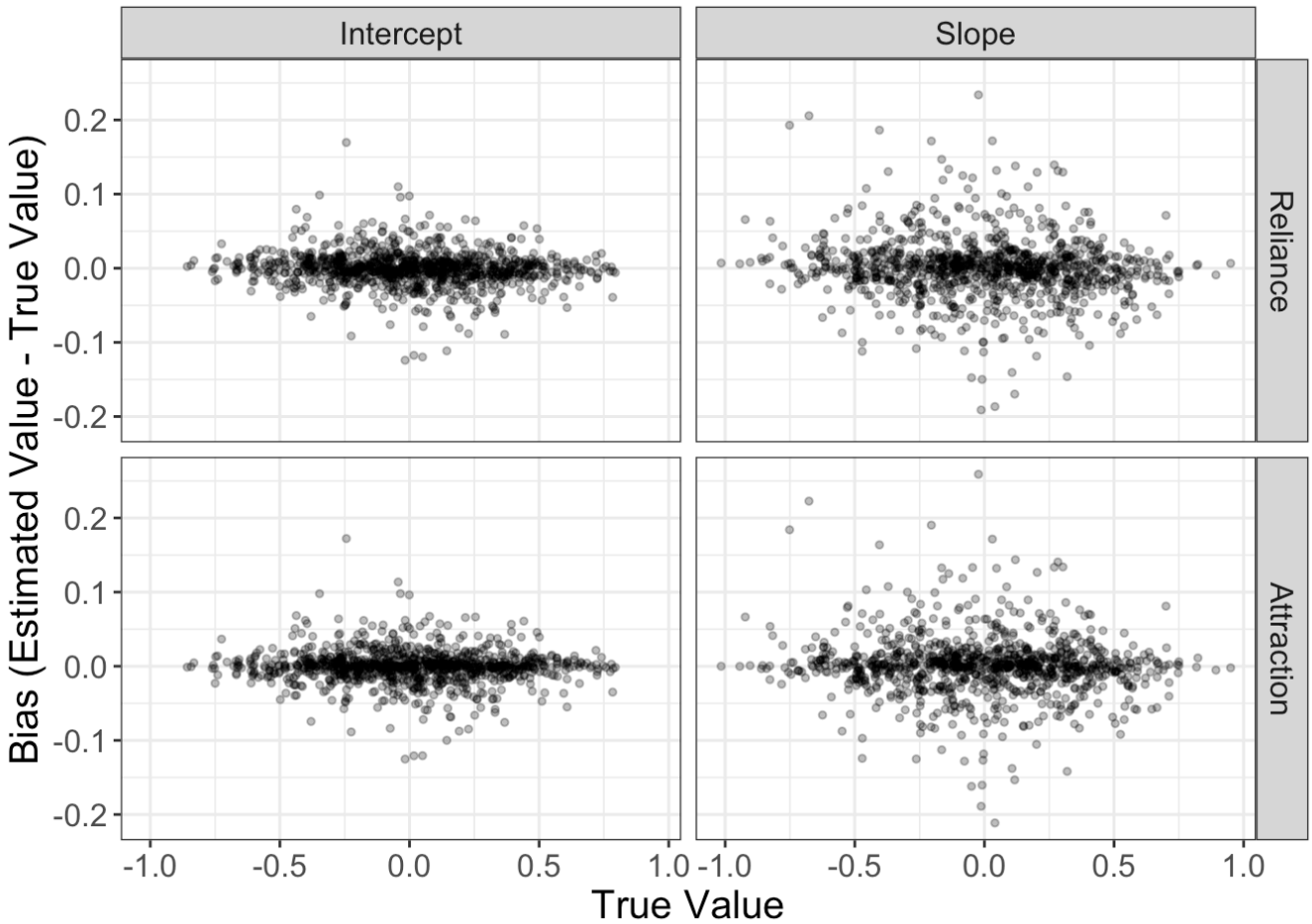

Figure A1. Scatter plots between true values and the magnitude of bias in each simulation

False-Positive Rate in By-Perfective Analysis

In the examination of false-positive rates in our by-perfective analysis, we took the following steps for each of reliance and attraction.

1. Thirty-nine ΔP values were randomly selected from a uniform distribution spanning from -1 to +1.
2. Based on those ΔP values, a fake data set was created, following Step 2 (both 2-a and 2-b) in the validation experiment reported above. This was carried out once for L1 data and once for L2 data, thereby generating data for both L1 and L2 with the same set of ΔP values.
3. The same by-perfective analysis as the one reported in the main text was conducted on the fake data set, and based on the results, we recorded the proportion of non-significant differences in the 39 perfectives with the significance level of $\alpha = 0.05$. No adjustment for multiple comparison was made for the ease of interpretation.
4. Steps 1 through 3 were repeated 1,000 times for the sake of the reliability of this validation procedure.