Research Article

Stefano Rastelli*

Intra-language: the study of L2 morpheme productivity as within-item variance

https://doi.org/10.1515/iral-2020.0082
Received June 20, 2020; accepted March 19, 2021; published online April 1, 2021

Abstract: This article suggests a method to appraise L2 morphemes productivity in longitudinal learner data. Traditionally, morpheme productivity is believed to depend on type frequency and on proportion between inflected and uninflected lexemes. However, such measures cannot distinguish between rote-learning and rule-learning of target-like forms. In contrast, the association measure $\Delta P$ (delta pi) can quantify the extent to which a morpheme is contingent upon a limited number of lexemes. Decreasing contingency might parallel learners’ increasing awareness of asymmetrical morpheme-lexeme distribution in the input and this might be a cue of developing L2 grammatical competence beyond appearances. The paper presents the rationale and procedure for analyzing within-item variance – or the ‘intra-language’ – and illustrates a case-study concerning the perfective morpheme in L2 Italian.

Keywords: association measures; contingency learning; morpheme productivity; second language acquisition

1 Topic: morpheme productivity beyond face value

In almost 50 years since Selinker’s (1972) seminal article, interlanguage studies have focused on changes occurring in learners’ speech and writing over time (for a recent collection of studies, see Han and Tarone 2014). A considerable body of work has shown how morphosyntactic form – function mappings gradually approximate those of the target language (TL). However, becoming target-like is not the end of the story; development may continue even when forms are

*Corresponding author: Stefano Rastelli, Humanities, University of Pavia, C.so Strada Nuova 65, Pavia, Italy, E-mail: stefano.rastelli@unipv.it

Open Access. © 2021 Stefano Rastelli, published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.
indistinguishable from TL. What at first may seem an unlikely claim is understandable once it is appreciated that learners (like native speakers, or NSs) can process target-like items in two ways: statistically (as chunks) or grammatically (as projections of abstract features). Indeed, the same morphosyntactic item can be learned as a modification of a template (or productive schema-construction) stored in memory (e.g., Culicover et al. 2017; Jackendoff and Audring 2019) or as the product of generation by a rule (e.g., Yang et al. 2017). One learning mechanism is statistical learning (SL); the other is grammatical learning (GL). Either learning mechanism has its own way to operate, for example, in the acquisition of the subject-predicate agreement rule underlying the English sentence ‘the man walks’.

SL determines how well the learner can recall from memory their previous encounters with the combination ‘the man’ ~ ‘walks’. GL determines how well the same learner knows a rule like “add an –s to any words (frequent or infrequent) falling within the categories verb, singular, and third person.”. SL is a developmental process that culminates in the acquisition of chunks. Chunks are a subset of formulaic language and multiword expressions. They are any adjacent, non-recursive combination of \( n \) (\( n > 1 \)) tokens which (i) have high absolute joint frequency and (ii) co-occur with exceptional frequency and dispersion. Psychologically, chunks are considered “processing units” (Myles 2016). They are sequences of tokens which occur so often that the speaker automatizes, perceives, stores, retrieves, and executes them as single units at every level of linguistic organization.

GL is learning by labels (e.g., Bauke and Blümel 2017; Hornstein 2009). Labels are frequency-independent grouping variables that describe a closed subclass of tokens, e.g., Tense, Mood, Subject, Object, Telic, Singular, Unergative. Labels may further host sublevels and even clusters of properties. For example, the label “Unergative” includes [+Agentive] and [−Telic]. A label turns a sequenced string of words or chunks into a hierarchically organized cognitive unit, regardless of both the joint or disjoint frequencies of its components. Labeling the chunk also ensures that it can be treated as a building block available for combining with other chunks.

This paper proposes a novel use of an already-known contingency-based, unidirectional association measure called \( \Delta P \) (delta pi) as a feasible, handy way to analyze morpheme productivity in target-like forms. Since these can be represented and processed by learners as both analyzed in stem + affix via GL and as unanalyzed wholes or chunks via SL, the study of within-item morpheme productivity implies that – at least in principle – one can distinguish between statistically and grammatically generated L2 forms. The procedure for assessing L2 morpheme productivity – meant as the study of within-form variance – in this paper is dubbed the study of the intra-language.
2 Background: the dual source of learners’ morphosyntactic competence

2.1 Definition and measurement of morpheme\(^1\) productivity

Morpheme productivity\(^2\) is an affix’s ability to combine with different lexemes. The more an affix participates in different lexemes, the more productive it is. In the literature, morpheme productivity is usually considered a direct function of the type frequency of relevant constructions (e.g., Croft and Cruse 2004: 309; Gries and Ellis 2015: 234). The measure of type frequency is captured by the formula \(V(C;N)\), i.e., the type count \(V\) of the members of a morphological category \(C\) in a corpus of \(N\) tokens (Desagulier 2016: 179). For example, the type frequency of the past perfective construction in Italian (the passato prossimo) is the number of different verb lexemes with which the passato prossimo co-occurs in a given corpus. According to such measure, when the perfective construction’s type frequency is low, the perfective morpheme is unlikely to be represented in a L2 Italian learner’s competence. This is because L2 learners can represent the perfective morpheme only if they conceive of it as separate and independent from the various features of the specific lexemes in which the morpheme occurs.

---

1 For limits of space, the current paper restrains from debating the notion of morpheme, the place of morphology in the architecture of language, whether it is independent from other components of the grammar and which are its basic units of analysis (for a recent review see Audring and Masini 2019). We accept the traditional view of morpheme as ‘the smallest meaningful constituent of words’ (Haspelmath and Sims 2010: 3), being aware that such definition ‘is unequally prominent in all languages’ (ib). As to Second Language Acquisition (SLA), we assume that acquiring functional morphology or ‘morphosyntax’ – the level that houses the properties that cannot be subsumed under phonology or lexical semantics, e.g., the categories of Tense and Aspect (Jackendoff and Audring 2019) – is especially difficult in adulthood (Hawkins 2001: 34; Slabakova 2016: 176). Finally, we agree that measures of language complexity based on morphological richness can be used as cues of L2 attainment (Brezina and Pallotti 2019).

2 In this article, I adopt ‘productivity’ as a cover term that includes ‘morpheme analyzability’ and ‘morpheme decomposability’. While the latter expression concern the process, the former concerns the outcome of the process. Here we can only briefly recall that there are two views on the definition of ‘morpheme productivity’. One considers productivity as the learner ability to create a new word involving a morpheme. The other deduces productivity from the likelihood that new word types with a morpheme will be found in a corpus as it increases in size. The two views differ because the likelihood of finding a morpheme attached to a word depends not only on the instantiation (in a learner’s mind) of the cognitive procedure that can generate such morpheme, but also on language uses and on the peculiarities of the corpus that is used as a reference for counting the inflected forms. Since word frequencies fluctuate with the topic of discussion, such count will depend strongly on the characteristics of the corpus.
A different measure of morpheme productivity is ‘hapax-based’ productivity (Baayen 1992). The idea underlying this approach is that the more productive affixes are, the more often they produce rare or even unique tokens. Finally, according to Bybee (1985), the process of productivity implies the shift from ‘analyzability’ to ‘autonomy.’ Constructions begin being productive when learners analyze them into stem + affix combinations. As an exemplar of a productive (analyzed) construction reaches a certain token-frequency threshold, it is stored again as an autonomous unit, after which the construction’s productivity may diminish.

2.2 Morpheme productivity as within-item variance: interlanguage and ‘intra-language’

In almost 50 years of interlanguage (IL) research, the use of distance-based metrics – measuring convergence or divergence of IL and TL – has predominated. Despite words of caution (e.g., Ortega 2014: 197; Sorace 1996: 386; White 2003: 26) and warnings against the risk of committing the “comparative fallacy” (e.g., Bley-Vroman 1983), assessing the IL from the TL standpoint was the standard, regardless of the paradigm. Distance-based metrics also innervated the debate on emergence criteria. This debate concerned the relationship between the first appearance of linguistic structures of TL in a learner’s IL and the moment when they are acquired. According to some, morphological productivity depends on the percentage of target-like forms learners produce. The higher the percentage, the more the TL and IL systems are held to converge. In the SLA literature, this percentage has varied from 30 to 90% (see Pallotti 2007 for a review). According to others, the difference between correct and incorrect forms should not be taken at face value, and IL development should not be confounded with mastery or accuracy. First, the measurement of accuracy cannot easily account for well-known nonlinear or curvilinear (e.g., U-shaped) learner behavior (Norris and Ortega 2003: 737). Second, percentages could be meaningless because they do not distinguish between formulaic and productive uses of target-like forms. For this reason, Pienemann (1998: 191, 2015: 133) proposed shifting attention from pairwise differences (correct vs. incorrect) to systematicity variance in both the lexeme and

---

3 Sorace (1996) wrote that – when reconstructing the learner’s grammar – “the evaluation of the distance between native and non-native grammars becomes an irrelevant criterion” (p. 386).
4 “The issue then is whether the interlanguage representation is a possible grammar, not whether it is identical to the L2 grammar” (White 2003: 26).
morpheme. This view, too, was criticized because learners may produce correct form–function associations randomly, using a form with no clear or consistent association to a given function (e.g., Pallotti 2007: 375). In sum, whether it is about percentage correct, variance in stem–affix alternations, or in form–function mappings, SLA literature has often been paired L2 development with both systematic and TL-directed changes in a learner’s production. As noted by Ortega (2014), while in principle few would disagree with Selinker (1972) that the IL should be studied as a system in its own right, in practice most researchers have assessed native-like attainment by looking at “the isomorphic conformity with idealized target-like norms” (Ortega 2014: 197).

A less cited line of SLA research shifted the attention from pairwise differences (correct vs. incorrect) to target-like items in learners’ production. In what sense studying target-like forms – and not only deviant forms – is important for SLA? It is because – as stated before – target-like forms can be represented and processed in two ways, statistically and grammatically (Section 1). Whereas the study the IL targets how forms approximate forms of TL (‘between items variance’), the study of the intra-language concerns how target-like forms are generated. By the expression ‘within-item variance’ one means variance across different processes behind the same linguistic item.

In SLA research, knowing when a given sequence has been generated by the grammar and when it has been produced as an unanalyzed whole is difficult (e.g., Myles 2004: 140). Although learners may know more than what they can do (VanPatten et al. 2019: 80), it may also be true that learners know less than what they seem already capable of doing. The literature shows cases where grammatical performance may temporarily occur without grammatical competence. Dulay et al. (1982: 232) observed that initial learners often produce unanalyzed stretches of speech, which are far beyond their developing L2 rule systems. Such forms stand out because they are target like from the onset, while learning curves for other items are gradual. Norris and Ortega (2003: 737) suggested that – to distinguish between productive and formulaic uses of a form – one should balance the frequency of form suppliance in the expected functional contexts and accuracy. Unfortunately, the operationalization of frequency and accuracy was not provided.

---

5 “Ensuring that a morpheme appears with different lexical items and that a given item is also used with a different morpheme is one measure to ensure that formulaic chunks are not counted as instances of morpheme insertion” (Pienemann 2015: 133).

6 Pallotti (2007) proposed using chi-squared tests to evaluate the independence of the use of a form from a specific functional context.
2.3 Intra-language and SLA

Since its inception, SLA has been concerned with whether forms are stored in memory or generated by a rule. For example, in Processability Theory, SL (the “formula stage”) developmentally precedes GL (the “categorical stage”; Pienemann 1998) and an emergence criterion is established to ensure that formulaic chunks are not counted as instances of morpheme insertion (Pienemann 2015: 133). According to a recent version of the Fundamental Difference Hypothesis (Bley-Vroman 2009), L2 learners resort to chunks when UG fails or is missing. According to the Shallow Structure Hypothesis, “there is more than one way of processing and mentally representing morphologically complex words… Two (or more) different processing routes are assumed to operate in parallel, one of which involves creating a detailed grammatical representation of the input and the other one involving the heuristically driven construction of a “rough-and-ready” representation that lacks grammatical detail” (Clahsen and Felser 2018: 698). Also, the existence of a developmental shift from chunks to productive forms has long been acknowledged and debated in SLA research. For example, an idea is that learning whole chunks precedes learning their parts, and that these chunks feed into grammar as learners decompose these chunks into their subcomponents (stems, affixes) (Myles et al. 1998). A similar view claims that language learning is the learning of formulaic sequences and that developmental sequences develop from fixed formulas to productive schematic patterns (Wagner-Gough and Hatch 1976; Wong Fillmore 1976). In the MOGUL framework, complex items can be accessed as wholes or component units, so their processing is always “a race between whole forms and decomposition” (Sharwood Smith and Truscott 2014: 120) because both memory-based activity and rule-governed activity are necessary and “make-up the two halves of processing” (p. 124). Another, related question concerns why learners should learn grammatically what they have already learned statistically. One possible answer is that chunks and grammatical forms are part of a dual learning/processing mechanism, but they are not redundant, as they produce the same result in two ways. Learners may benefit from learning the same item of the TL twice. Different processing circumstances may require different kinds of underlying knowledge. One kind is the basic knowledge that certain units are associated with certain other units. The other is the knowledge of what is being associated and why some units associate and not others (Eubank and Gregg 2002: 239). The former kind of knowledge is statistical; the latter is grammatical. Learners should resort to either chunks or grammatical forms by following the same criteria that one uses when deciding to use “predictive text technology” software when writing a text message on a cell phone rather than typing the word oneself. Usually,
people use this technology not because they are ignoring their internal grammar, but because it is faster and works better for routines. The fact that it is less suitable for writing poems, for example, does not make the statistical technology redundant. Even weird sentences like *colorless green ideas sleep furiously* – once a certain frequency threshold is passed – can be incorporated into the technology and be processed faster, effortlessly, and without errors.

### 2.4 The discontinuity model: the shift between statistical and grammatical learning

The discontinuity model (DM) proposes that, in adult SLA, like in native speakers’ competence, statistical and grammatical learning – which language theory often considers opposed – integrate and superpose (Rastelli 2014, 2019). According to DM, learners’ implicit and gradual appraisal of the distributional asymmetry (skewness) of the lexeme and the morpheme in morphological forms is what disables contingency learning and eventually triggers the shift from unanalyzed, morphosyntactic wholes to productive forms. Importantly, once grammatical learning occurs, statistically learned morphosyntactic forms do not disappear: they remain active and parallel their grammatical counterparts. Even if the passage from statistical to grammatical learning is developmentally moderated, the former could be a permanent part of a learner’s toolkit. In fact, advanced learners – like native speakers – can either retrieve target-like morphosyntactic items as chunks or generate them by rules. The superposition of frequency-based and grammatical learning devices represents both the steady-state in a native speaker’s competence and the end state in L2 development. Ontogenetically, while in L1 acquisition, the natural endowment for language constrains statistical learners’ capacity by narrowing the hypothesis space, in adult SLA, statistics can reopen the window of opportunity for grammar and drive adult learners to derive part of L2 morphosyntax statistically. The DM proposes a computational and psycholinguistic model of how this might occur. In short, skewness between transition probabilities (TP) represents the triggering factor in both L1 and L2 acquisition. As a fluctuation in TP between adjacent words drives children to individuate the word boundaries in a speech stream, so skewness between TP of morphemes and lexemes (learner’s awareness that morpheme and lexemes have different distributions) is what drives adult learners to individuate the grammatical features – the morphemes and their function – that are hidden in morphosyntactic chunks. In SLA, there is no competition between statistical and grammatical learning because the former precedes and prepares the latter. Statistical and grammatical learning apply to the same language domain (morphosyntax), albeit with different timings and perhaps
also under different processing circumstances. The DM is a neurologically plausible model. Although learners' capacity for statistical language learning might decline with age, adults can still memorize, extract and associate n-grams to efficiently form chunks. If any, difficulties in decomposing stem and affix in chunks may arise because procedural memory is less available in adulthood (Ullman 2005: 151), not because implicit statistical learning is impaired. The DM assumes that statistically-learned (memory-based) representations are processed with statistical processing mechanisms, and that grammatically-learned representations are processed with grammatical processing mechanisms. Although it is not logically necessary that there is a one-to-one mapping between memory-based generation of forms and kind of learning mechanism, there seems to be a tight connection between kind of memory system – declarative or procedural – and kind of learning. The declarative memory system is the associative network of facts and events, which is subserved by medial temporal lobe regions (hippocampal regions, ento-rhinal and peri-rhinal cortices, para hippocampal cortex) and parieto-temporal neocortical regions (Eichenbaum 2012; Squire and Wixted 2011). In language, declarative memory supervises both the implicit and explicit learning, storage, and use of arbitrary (non-derivable, idiosyncratic) information; it underlies aspects of the mental lexicon and word learning. On the other hand, procedural memory, which is subserved by a network of subcortical (basal ganglia) and cortical (frontal cortex) structures, is assumed to underlie the combinatorial aspects of mental grammar, such as the rules of simple past formation in English (e.g., walk + -ed = walked; Ullman 2004). Some models maintain that the lexicon is learned by declarative memory, whereas the grammar is learned by procedural memory. However, learning morphosyntactic chunks requires learners to rely on both declarative and procedural memory circuits. Unlike isolated words, morphosyntactic chunks have a complex internal structure (see also Embick and Marantz 2005: 245). When learning morphosyntactic chunks, learners do not just associate a form with a meaning. They must also index the implicit information about the combinations and the skewed distribution of the stem and affix. Therefore, it is likely that for morphosyntactic chunks to be formed, stored, and retrieved, not only is associative (declarative) memory necessary, but also the combinatorial skills sustained by procedural memory.

Finally, the DM recognizes that – given the current state of our knowledge – it is impossible to establish exactly whether a given lexeme-morpheme combination in a given context was produced by general rules of grammatical encoding or by memory-based retrieval. However, observing morpheme distribution in longitudinal data (whether lexeme-morpheme associations become more diversified over time) can provide cues that the memory-based, lexical retrieval strategy gradually yields to more grammatical, generative processing. This paper proposes a
probabilistic way to assess whether the distribution of morphemes may lead one to hypothesize that something changed in the way the morphemes are represented by the learner.

3 Rationale of the study: contingency and distributional asymmetry as measures of morpheme productivity

Variance in the number of tokens and the number of types over time is insufficient for revealing whether and to what extent a morpheme is productive in a learner’s competence. The simple count of inflected forms cannot discriminate whether learners represent and process such morphosyntactic forms statistically or grammatically. Recent statistical studies suggest that – in general – frequency comparisons are not informative enough for language development (e.g. Baayen 2010). Items less frequent in absolute could be acquired earlier, more easily, and more stably than more frequent items. Other than raw frequency, other measures can inform us on how distribution in the input impacts L2 acquisition: dispersion, predictability and surprisal, recency, salience and association/contingency (e.g., Ellis 2016; Gries 2015a; Gries and Ellis 2015).

To discriminate between the statistical and the grammatical nature of target-like morphosyntactic forms, this study uses the notion of ‘contingency’. Contingency learning (CL) is a probability-based mechanism for learning whether relations between events are causal or non-causal (Beckers et al. 2007: 289; Shanks 2007). Subjects tend to label co-occurring events as causal when the probability \( P \) of getting a response \( R \) (e.g., thunder) given a cue \( C \) (e.g., lightning) \( P (R|C) \) is high. Causal cue–outcome relationships trigger category formation (Gluck and Bower 1988) based on events’ (i.e., words’) adjacency and resemblance (Reeder et al. 2010). In L1 and L2 acquisition, CL might support category formation by exploiting the frequent co-occurrence between a lexeme and a grammatical construction. Be ‘lexeme’ an Italian telic predicate and ‘grammatical construction’ the Italian perfective past (the passato prossimo). Telic predicates like uscire ‘exit’ – e.g., morire ‘die’, cadere ‘fall’, arrivare ‘arrive’ – have an inherent endpoint or culmination in their semantic template (Dowty 1979; Krifka 1992; Mourelatos 1978). The presence of an endpoint means that, at the perfective, the event of uscire ‘exit’ is true only if the resulting state of ‘being out’ is attained. The endpoint makes telic predicates more compatible than atelic ones with the perfective morpheme, because the perfective in Italian is used to depict past events as bounded. Both boundedness and telicity delimit events; boundedness does so overtly, by
visualizing the boundaries of the event in the discourse through perfective morphology, while the inherent (covert) lexical meaning of telic predicates suggests that the resulting state of the event is relevant, independent of perfective or imperfective morphology. If L2 learners, driven by the congruence between perfectivity and telicity, realize that telicity is a strong cue for the perfective morpheme, they might characterize the whole perfective category as contingent upon telicity. Contingency between telic lexemes and the perfective morpheme provides the learner with a conceptual support for identifying the morpheme and then generalizing its functions to other members of the category. This support is only temporary though. In fact, the most relevant feature of CL for language acquisition and development is its transiency. The contingency of lexeme–morpheme associations is expected to diminish as learning progresses. As long as the strength of association between a cue and the response is the only factor driving the learning process, category membership will be restricted to a few perfectly matching exemplars. Generalizations at later stages of learning will break prior association by introducing new, less prototypical members (e.g., not all perfectives are telic and vice versa). In initial learner data, a given morpheme is likely contingent upon an exceptionally limited number of lexemes. ‘Exceptionally’ means that the morpheme is not distributed as one would expect given the distribution of all other morphemes in the corpus. This abnormal morpheme-to-lexeme reliance indicates that the morpheme is not (yet) productive and that the corresponding morphosyntactic form – although being perfectly target-like – is probably an unanalyzed chunk. As a learner becomes proficient, morpheme-lexeme reliance is expected to diminish because learners conceive the morpheme independently of the lexeme and starts to extend it to the available lexicon.

4 The method: ΔP

This section introduces the unidirectional, contingency-based association scale ΔP as a mean to analyze morpheme productivity in longitudinal data. Fluctuations of ΔP scores provide probabilistic cues to discriminate between statistical and grammatical learning by identifying how lexeme–morpheme contingency changes within the same target-like form over time. ΔP is a one-way dependency statistic developed by Allan (1980) (for a description see Desagulier 2016; Ellis 2006, 2007; Ellis and Ferreira-Junior 2009: 198; Schmid and Küchenhoff 2013). Its application to L2 data was first discussed by Ellis (2007) and exemplified by Ellis and Ferreira-Junior (2009). Ellis (2007: 11) defines ΔP as in (1) and below:
\[ \Delta P = p(\text{Response} | \text{Cue present}) - p(\text{Response} | \text{Cue} = \text{absent}) \]  

(1)

\( \Delta P \) is the probability of the outcome given the cue \( P(R|C) \) minus the probability of the outcome in the absence of the cue \( P(R|\neg C) \). 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 \( \Delta P = 0 \). \( \Delta P \) approaches 1.0 as the presence of the cue increases the likelihood of the response and approaches −1.0 as the cue decreases the chance of the response. Unlike bidirectional association measures – such as log-likelihood, T-score or Mutual Information – and unlike chi-square or Fisher’s-exact test, \( \Delta P \) can separately assess each item’s contribution to the overall strength of association by comparing two kinds of transition probabilities between two items, such as the lexeme (a verb) and the morpheme (e.g. the Italian perfective). The first transition 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 transition probability is attraction. It compares the relative frequency of the lexeme with the morpheme to the relative frequency of the lexeme without the morpheme.\(^7\) The starting point for calculating \( \Delta P \) is a contingency table like Table 1, 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 given learner corpus:

<table>
<thead>
<tr>
<th>Presence of morpheme ( y ) (response)</th>
<th>Absence of morpheme ( y ) (no response)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presence of lexeme ( x ) (cue)</td>
<td>( a )</td>
</tr>
<tr>
<td>Absence of lexeme ( x ) (no cue)</td>
<td>( b )</td>
</tr>
<tr>
<td></td>
<td>( c )</td>
</tr>
<tr>
<td></td>
<td>( d )</td>
</tr>
</tbody>
</table>

Table 1: A contingency table.

\(^7\) Schmid (2000) attributed the terms ‘attraction’ and ‘reliance’ to the morpheme and to the lexeme, respectively: “\( \Delta P \) measures the contingent probability of a given construction attracting a given lexeme (\( \Delta P \) construction \( \rightarrow \) word; henceforth, ‘\( \Delta P \) Attraction’) and of a given lexeme relying on a given construction (\( \Delta P \) word \( \rightarrow \) construction; henceforth, ‘\( \Delta P \) Reliance’)” (pp. 54–55).
reliance, treats the lexeme as the cue and the perfective morpheme as the response; it shows the difference between the frequency of the perfective morpheme with and without the lexeme. The second half of the formula \( \frac{a}{a + c} - \frac{b}{b + d} \), which reflects attraction, treats the perfective morpheme as the cue and the lexeme as the response; it shows the difference between the frequency of the lexeme with and without the perfective morpheme. \( \Delta P \) is a scale (based on proportions), not a test of significance, so there is no minimal threshold value (like the p-value, Ellis 2012: 28). The relevance of \( \Delta P \) in the case-study lies just in the comparative information it provides among the lexeme–morpheme contingency across L2 developmental periods.

Three advantages of using \( \Delta P \) are highlighted by Gries (2015b). First, \( \Delta P \) is 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, \( \Delta P \) 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: 3). Second, unlike many other statistics, \( \Delta P \) is easy to understand because it involves nothing but a mere difference in proportion. Third, \( \Delta P \) has received experimental support in psychological studies because of the importance the feature of lexeme-morpheme asymmetry for learning (the fact that lexemes are many and morphemes are few). As we have seen, this latter point is relevant for the current study. In longitudinal learner data, \( \Delta P \) can not only identify strong morpheme-lexeme collocations, but also changes in the asymmetries among their subcomponents (the morpheme and lexeme). A decrease in \( \Delta P \) means not only that the morpheme is no more contingent upon the lexeme, but also that the learner perceives that their distribution is skewed. \( \Delta P \) has already been used in SLA research (e.g. Ellis and Ferreira-Junior 2009; Ellis et al. 2014; Tracy-Ventura and Cuesta Medina 2018; Wulff et al. 2009), but not in order to measure morpheme productivity.

5 Case-study: contingency learning and perfective morpheme productivity in L2 Italian

5.1 The perfective morpheme in Italian

The case-study described in this section illustrates an intra-language approach to the analysis of morpheme productivity in longitudinal data. The study focuses in particular on changes in morpheme-to-lexeme reliance and lexeme-to-morpheme
attraction over time. Such changes may indicate that target-like morphosyntactic forms can actually be represented and processed differently by L2 learners at different stages of language development. All perfective morphemes in this case study are found in the passato prossimo, an Italian past tense. In Italian past tenses, the [-perfective] distinction is encoded in morphology. Having evolved from an original present perfect value and having absorbed the aoristic (punctual) value of a competing simple past tense (the passato remoto), the modern Italian passato prossimo (e.g., Mario ha giocato a tennis ‘Mario played tennis’) is a compound tense, which encodes the perfective past. This form may also express the meaning of a present perfect in English. By contrast, the Italian imperfetto (e.g., Mario giocava a tennis ‘Mario used to play/was playing/played tennis’) encodes the imperfective value and conveys unbounded, habitual, or ongoing past events. The Italian perfective morpheme in the passato prossimo expresses also the features of person and number, whose interaction has been excluded from this study for the reason of space.8

5.2 The data

We explored lexeme–morpheme contingency in the Corpus Pavia. This is the largest and best known longitudinal learner Italian corpus to date (~700,000 tokens, 19,000 types overall). The corpus Pavia it is publicly available and it represents the result of fine-grained sampling procedures. 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 h of oral interviews of 22 Italian L2 learners from 11 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

8 The current paper is not concerned with the acquisition of the Italian perfective past (the passato prossimo) even though in this study’s dataset, all perfective morphemes occur in that form. There are two reasons for this choice. First, the passato prossimo comprises an auxiliary (the inflected form of avere ‘have’ or essere ‘be’) and a past participle that hosts the inflected perfective morpheme. In Italian learner corpora, about 30% of total auxiliaries – especially in early production – are missing or wrong (Rastelli 2007). Second, the acquisition of the auxiliary and acquisition of the perfective morpheme are treated separately in the literature. Some authors have connected auxiliary selection with lexical semantics and the unaccusative/unergative split (e.g. Sorace 2000), but these theories do not connect auxiliary selection the emergence of the perfective morpheme. Finally, the paper is not concerned with the aoristic-perfective (the passato remoto), which is nearly absent in Italian learner corpora.
Italian language courses before arrival. 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 a description of pictures and oral retelling of picture-stories and video excerpts from the film *Modern Times*. In Supplementary material, the procedure for formatting and coding data is described. Supplementary material also describes the Corpus Pavia’s composition in terms of learner- and data-related dimensions. After lemmatization and POS-tagging with Sketch Engine, each learner’s file comprised between 124 and 2,887 finite verb forms (983 on average), totaling 22,109 verb tokens and 5,540 perfective tokens stemming from 304 perfective types. For this study, we selected a sample of 39 perfectives types with frequency ≥ 25. The cutoff point was chosen after visual inspection of the Zipfian curve with the purpose of guaranteeing the sample’s homogeneity, manageability, representativeness, and the density of data (see below). After that threshold, the absolute frequencies of perfective predicates drop considerably, and the Zipfian curve approaches its inflection point. The resulting sample contained 3,940 perfective tokens stemming from 39 perfective predicates. In Supplementary material, the raw frequency of the sampled 39 perfective predicates in the Corpus Pavia is listed. To minimize data sparseness – the fact that perfective types and tokens were unevenly distributed across both interviews and learners (see Section 5.4) – longitudinal data in this study were aggregated. The period of interview was chosen as the most comprehensive and neutral criterion for aggregating data across learners (see below). For each learner, interviews were grouped into two periods: early and late. Each period was balanced within-learner for number of interviews (from 1 to 6), but the time-span between periods (ranging from one week to four months) and the overall number of verb tokens in each period varied.

5.3 Pros and cons of aggregating learner data

This study identified changing patterns of lexeme-perfective associations in longitudinal learner data. Data sparseness – the fact that perfective types and tokens

---

9 https://auth.sketchengine.eu.

10 There is at least one caveat with trimming off infrequent items. Some metrics of productivity rely on the tail of the Zipfian distribution, so one cannot exclude *a priori* that implementation decision does not have problematic effects on the results.
were unevenly distributed across both interviews and learners – might have undermined this goal. For example, although the telic perfective *venuto* ‘come’ has high absolute frequency in the corpus Pavia, some learners used it only at early interviews, others used it only at late interviews and some did not use it at all. Having many, few or zero-occurrence of *venuto* at early interviews might not depend on the main independent variable in our study (learner’s sensitivity to morpheme-lexeme contingency) but on the kind of task and on the topic of interviews, whose distribution did not follow a predictable pattern across interviews. To minimize this contextual bias, longitudinal data in this study were aggregated. Period of interview was chosen as the most comprehensive and neutral criterion for aggregating data across learners. For each learner, interviews were grouped into three periods: ‘early’, ‘mid’, and ‘late’. Each period was balanced within-learner for number of interviews (from 1 to 6). Although the overall number of verb tokens changed across periods, the time-span (ranging from one week to four months) in within-learner interviews was kept constant. The choice of aggregating data from different learners across three equally distant periods is methodologically questionable. This seems the kind of sampling that Gries and Stoll (2009) have argued against for acquisition data by proposing a Variability-based Neighbor Clustering (VNC) approach and that Stoll and Gries (2009) have argued against by proposing regression-based approaches that determine the temporal stages from the behavior of the regression models. In fact, cutoffs on the temporal continuum (in this study, ‘early’, ‘mid’ and ‘late’ interviews) are arbitrary and that the existence of developmental stages is assumed *a priori* instead of emerging ‘bottom-up’ from the data (Gries and Stoll 2009: 223). However, there are cases when arbitrariness does not stem from theoretical preconceptions (as it is claimed by Gries and Stoll 2009: 219) but it is a necessary expedient to minimize the impact of intervening variables that cannot be controlled for. To elicit spoken production in the corpus Pavia, researchers used an open repertoire of tasks and topics. The choice of whether or not – and of when – to use a pre-determined topic/task was left to interviewers’ decision. For example, some learners were asked to describe how they arrived in Italy or to tell a scene of Modern Times – two pre-determined topics that respectively likely elicit or dissuade the use of *venire* ‘come’ at the perfective past – indifferently at early, mid or late interviews. Other learners were instead asked just to talk about their plans for the future. Since topics and tasks were either intermittent or interspersed randomly across interviews, a clustering algorithm that agglomerate temporal adjacent values of the perfective predicates based on a similarity metrics – like the VNC – would risk exchanging the results of a biased elicitation procedure for the presence or absence of cues of a learner’s developing competence over time. Sampling data from large intervals – rather than from adjacent points – and aggregating different within- and between-learners
interviews minimize the impact of the heterogeneity of the elicitation tasks and topics. The logic underlying our choice is simple: the larger the sampling intervals, the more topics and tasks they include and the least their differences will impact the outcomes. If the analysis of ΔP can actually surface patterns of lexeme-morpheme contingency over time, these will have emerged in spite of – and not because of – the heterogeneity of tasks and topics. While bearing in mind the risks that may come with this methodological choice, we considered that in this case data aggregation is not a ‘bug’, but rather a ‘feature’ of the current study and that it represents the lesser evil. A possible objection to aggregating learner data is that early and late interviews might not reflect – respectively – low and high proficiency levels in general. The data from each learner were divided into an early and a late period disregarding the fact some learners might have been at a more advanced stage in their early interviews than other learners were in their late interviews. Since instances of the same verbs may occur in only early or late interviews for different learners, there is no guarantee that the instances of a particular verb in the ‘early’ category are really from an earlier stage in development than those found in the ‘late’ category. However, this shortcoming has a limited impact, if one considers how learners’ proficiency levels were distributed in the sample. In fact, corpus editors rated learners’ proficiency based on Klein and Perdue (1993). ‘Prebasic’ learners used mainly or exclusively pragmatic criteria (topic-comment, given-new) to build utterances. ‘Basic’ learners could go beyond information structure and consider 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). By looking at the data, one can see that there is extreme uniformity in the sample: out of 22 subjects, 19 were eventually classified as post-basic, 1 pre-basic and 2 were classified as basic. This means that – although the risk of confounding the variables ‘period of interview’ and ‘proficiency’ cannot be eliminated – it is at least minimized.

5.4 Results

Table 2 reports the complete ΔP values of Reliance and attraction of the 39 sampled perfectives of the Corpus Pavia, at early and late interviews:

One can observe four things. First, reliance and attraction values patterned differently. As predicted given the transiency of CL in longitudinal data, reliance values of nearly all 39 predicates in the sample declined from early to late interviews, with the only exception of the telic predicates caduto ‘fallen’ and uscito
Table 2: ΔP values of Reliance and attraction in early and late interviews. The values refer to all inflected forms of the Italian perfective past (the Passato prossimo).

<table>
<thead>
<tr>
<th>Perfective</th>
<th>English</th>
<th>reliance_early interviews</th>
<th>reliance_late interviews</th>
<th>attraction_early interviews</th>
<th>attraction_late interviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>dimenticato</td>
<td>Forgotten</td>
<td>0.828</td>
<td>0.649</td>
<td>0.014</td>
<td>0.016</td>
</tr>
<tr>
<td>sbagliato</td>
<td>Mistaken</td>
<td>0.743</td>
<td>0.397</td>
<td>0.011</td>
<td>0.006</td>
</tr>
<tr>
<td>perso</td>
<td>Lost</td>
<td>0.664</td>
<td>0.439</td>
<td>0.006</td>
<td>0.016</td>
</tr>
<tr>
<td>finito</td>
<td>Finished</td>
<td>0.645</td>
<td>0.284</td>
<td>0.025</td>
<td>0.012</td>
</tr>
<tr>
<td>portato</td>
<td>Brought</td>
<td>0.539</td>
<td>0.283</td>
<td>0.013</td>
<td>0.014</td>
</tr>
<tr>
<td>morto</td>
<td>Died</td>
<td>0.533</td>
<td>0.218</td>
<td>0.009</td>
<td>0.007</td>
</tr>
<tr>
<td>dato</td>
<td>Given</td>
<td>0.523</td>
<td>0.016</td>
<td>0.012</td>
<td>0.001</td>
</tr>
<tr>
<td>incontrato</td>
<td>Met</td>
<td>0.474</td>
<td>0.38</td>
<td>0.008</td>
<td>0.005</td>
</tr>
<tr>
<td>arrivato</td>
<td>Arrived</td>
<td>0.469</td>
<td>0.267</td>
<td>0.05</td>
<td>0.028</td>
</tr>
<tr>
<td>capitò</td>
<td>Understood</td>
<td>0.458</td>
<td>0.15</td>
<td>0.057</td>
<td>0.02</td>
</tr>
<tr>
<td>caduto</td>
<td>Fallen</td>
<td>0.41</td>
<td>0.41</td>
<td>0.009</td>
<td>0.01</td>
</tr>
<tr>
<td>tornato</td>
<td>Returned</td>
<td>0.386</td>
<td>-0.026</td>
<td>0.004</td>
<td>-0.001</td>
</tr>
<tr>
<td>lasciato</td>
<td>Left</td>
<td>0.371</td>
<td>0.075</td>
<td>0.012</td>
<td>0.003</td>
</tr>
<tr>
<td>imparato</td>
<td>Learned</td>
<td>0.325</td>
<td>-0.122</td>
<td>0.021</td>
<td>-0.05</td>
</tr>
<tr>
<td>passato</td>
<td>Past</td>
<td>0.32</td>
<td>0.277</td>
<td>0.01</td>
<td>0.007</td>
</tr>
<tr>
<td>trovato</td>
<td>Found</td>
<td>0.308</td>
<td>0.015</td>
<td>0.022</td>
<td>0.002</td>
</tr>
<tr>
<td>sposato</td>
<td>Got married</td>
<td>0.295</td>
<td>0.216</td>
<td>0.006</td>
<td>0.01</td>
</tr>
<tr>
<td>venuto</td>
<td>Come</td>
<td>0.274</td>
<td>-0.113</td>
<td>0.025</td>
<td>-0.012</td>
</tr>
<tr>
<td>uscito</td>
<td>Exit</td>
<td>0.247</td>
<td>0.302</td>
<td>0.009</td>
<td>0.008</td>
</tr>
<tr>
<td>fatto</td>
<td>Done</td>
<td>0.239</td>
<td>-0.05</td>
<td>0.101</td>
<td>-0.021</td>
</tr>
<tr>
<td>sentito</td>
<td>Heard</td>
<td>0.234</td>
<td>0.155</td>
<td>0.011</td>
<td>0.009</td>
</tr>
<tr>
<td>telefonato</td>
<td>Phoned</td>
<td>0.219</td>
<td>0.134</td>
<td>0.002</td>
<td>0.005</td>
</tr>
<tr>
<td>visto</td>
<td>Seen</td>
<td>0.206</td>
<td>0.118</td>
<td>0.093</td>
<td>0.026</td>
</tr>
<tr>
<td>detto</td>
<td>Said</td>
<td>0.178</td>
<td>0.003</td>
<td>0.062</td>
<td>0.001</td>
</tr>
<tr>
<td>scritto</td>
<td>Written</td>
<td>0.175</td>
<td>0.029</td>
<td>0.008</td>
<td>0.002</td>
</tr>
</tbody>
</table>
Table 2: (continued)

<table>
<thead>
<tr>
<th>Perfective</th>
<th>English</th>
<th>reliance_early interviews</th>
<th>reliance_late interviews</th>
<th>attraction_early interviews</th>
<th>attraction_late interviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>chiesto</td>
<td>Asked</td>
<td>0.164</td>
<td>0.07</td>
<td>0.006</td>
<td>0.005</td>
</tr>
<tr>
<td>cambiato</td>
<td>Changed</td>
<td>0.145</td>
<td>0.037</td>
<td>0.004</td>
<td>0.001</td>
</tr>
<tr>
<td>andato</td>
<td>Gone</td>
<td>0.141</td>
<td>−0.06</td>
<td>0.063</td>
<td>−0.027</td>
</tr>
<tr>
<td>comprato</td>
<td>Bought</td>
<td>0.126</td>
<td>−0.068</td>
<td>0.006</td>
<td>−0.001</td>
</tr>
<tr>
<td>parlato</td>
<td>Talked</td>
<td>0.097</td>
<td>−0.127</td>
<td>0.013</td>
<td>−0.017</td>
</tr>
<tr>
<td>lavorato</td>
<td>Worked</td>
<td>0.071</td>
<td>−0.217</td>
<td>0.008</td>
<td>−0.016</td>
</tr>
<tr>
<td>studiato</td>
<td>Studied</td>
<td>0.059</td>
<td>−0.111</td>
<td>0.005</td>
<td>−0.004</td>
</tr>
<tr>
<td>preso</td>
<td>Taken</td>
<td>0.049</td>
<td>0.098</td>
<td>0.004</td>
<td>0.017</td>
</tr>
<tr>
<td>dormito</td>
<td>Slept</td>
<td>0.041</td>
<td>0.09</td>
<td>0.001</td>
<td>0.002</td>
</tr>
<tr>
<td>pensato</td>
<td>Thought</td>
<td>0.036</td>
<td>−0.133</td>
<td>0.005</td>
<td>−0.015</td>
</tr>
<tr>
<td>mangiato</td>
<td>Eaten</td>
<td>0.03</td>
<td>−0.173</td>
<td>0.003</td>
<td>−0.015</td>
</tr>
<tr>
<td>piaciuto</td>
<td>Liked</td>
<td>−0.005</td>
<td>−0.191</td>
<td>0</td>
<td>−0.015</td>
</tr>
<tr>
<td>messo</td>
<td>Put</td>
<td>−0.01</td>
<td>−0.052</td>
<td>0</td>
<td>−0.003</td>
</tr>
<tr>
<td>voluto</td>
<td>Wanted</td>
<td>−0.04</td>
<td>−0.339</td>
<td>−0.007</td>
<td>−0.087</td>
</tr>
</tbody>
</table>
‘exited’. In contrast, attraction values of all predicates clustered around zero, showing no covariation between cue and response. Also, fluctuations of attraction over time were negligible (except fatto ‘done’, which goes down from 0.1 to 0.021). Recall (Section 4) that the formula for attraction (\(\Delta P_{\text{cue}x} = \frac{a}{a+c} - \frac{b}{b+d}\)) calculates the difference between the proportion of a given lexeme–perfective association out of the total perfectives in a corpus \(\frac{a}{a+c}\) and the proportion of the occurrences of the lexeme out of all other verb constructions \(\frac{b}{b+d}\). Attraction values around zero mean that there was no difference between the frequency of the sampled lexemes with and without the perfective morpheme. In contrast, mid-to-high values of reliance (\(\geq 0.4\)) suggest that – at early interviews – the presence of the perfective morpheme was contingent upon the presence of a limited number (11) of lexemes in the corpus (especially dimenticato ‘forgotten’, sbagliato ‘mistaken’, perso ‘lost’, finito ‘finished’, and portato ‘taken’). It should also be observed that the morpheme-to-lexeme reliance consistently diminished from early to late interviews. This perhaps represents the most relevant result of this study, which indicates that at late interviews the perfective morpheme was probably no more contingent upon a restricted number of lexemes.

Second, attraction and reliance became increasingly correlated over time. Unlike in early interviews, at late interviews, the more a lexeme attracted the perfective, the more the perfective relied on that lexeme. Table 3 shows the changing relationship between attraction and reliance in longitudinal data as measured by Pearson’s correlation.

The correlation between attraction and reliance can be appreciated by comparing the \(\Delta P\) values plotted in Figure 1 relative to early versus late interviews. To make interpretation easier, we used the English translation of Italian perfectives.

As one can see, at late interviews, higher reliance perfectives tend to have also higher values of attraction. Although in the literature there is no indication that the correlation between attraction and reliance might be another cue of morpheme productivity – as a function of learners’ increased awareness of the independent

<table>
<thead>
<tr>
<th>Interview</th>
<th>Pearson’s r</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early</td>
<td>0.10</td>
<td>0.34</td>
</tr>
<tr>
<td>Late</td>
<td>0.70</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

11 Fluctuations in the values of preso ‘taken’ and dormito ‘slept’ are negligible since they approach zero.
nature of the morpheme – as a matter of fact collinearity characterizes also lexeme–perfective morpheme contingency in the normalized occurrences of four corpora of spoken and written contemporary L1 Italian. Figure 2 visualizes the collinearity of attraction and reliance in L1 Italian.

12 ITTenTen16 is a 4.9 billion-word web corpus (downloaded by SpiderLing from May to August 2016) of texts collected from the Internet. The corpus is a part of the TenTen corpus family, a set of web corpora built using the same method. The Perugia Corpus is a 26 million-word corpus of
The third observation is that reliance scores and type frequency inversely correlated. Desagulier (2016: 175) observed that $\Delta P$ alone cannot be used to measure productivity because it is based on token frequencies, not type frequencies. However, our data showed a relationship between reliance and type frequency. Namely, reliance values decreased from early to late interviews as type frequency increased. It is important to establish which deviating value(s) in the contingency table (Section 4; Table 1) might have been most responsible for the decline in reliance scores from early to late interviews. As Table 4 shows, the relative frequency (density) of the perfectives compared to other verbal forms (cells $a + b$ of the contingency table) did not change from early to late interviews, nor did the overall number of predicates minus $a$ and $b$ (cell $d$ of the contingency table). Rather, the number of types (different perfective predicates) increased from early to mid interview period, indicating an expansion of the repertoire of lexical verb types available to the learners.

It is perhaps possible to assume that the decrease of reliance scores inversely correlates with the count of perfective verb types and not with the count of perfective verb tokens or with the expansion of the verbal lexical repertoire available to the learner. To say it differently, the pattern of productivity of the written and spoken Italian distributed across 10 textual genres. The CLIP (Corpora e Lessico di Italiano Parlato) is a 342,000-word corpus of 100 h of spoken Italian divided into five subcorpora (e.g., dialog, TV broadcast, phone conversations). The LIP (Lessico dell’Italiano Parlato) is a 490,000-word corpus of spoken Italian made up of 58 h of monologic and dialogic conversations recorded in five Italian cities in the early 1990s.
perfective morpheme might not follow the pattern of growth of the verbal lexicon (the fact that a learner produces more predicates and more perfectives in general).

6 Discussion

6.1 Summary of findings

The current study proposes that the association score $\Delta P$ measuring fluctuations in lexeme-morpheme contingency over time can be used to assess morpheme productivity. The results of the case-study showed that a generalized and consistent decrease in morpheme-to-lexeme reliance of perfective predicates occurred over time, going from early to late interviews. In contrast, the values of attraction were negligible. Two further cues suggested that a decline in morpheme-to-lexeme reliance is developmentally moderated. The first cue is that attraction and reliance value scores become collinear over time. Collinearity could be a feature of the asymmetry of the lexeme–morpheme distribution in a mature language, even if more studies are needed before this can be confirmed (see Section 6.3). The second cue is that the values of reliance are inversely proportional to type frequency, which in the relevant SLA literature is held to be the hallmark of morpheme productivity in second language development (e.g. Baayen 1992, 2010) (Section 6.4).

6.2 A measure of significance for $\Delta P$

According to the literature, unidirectional measure $\Delta P$ still lacks a proper measure of significance, so that any comparison between reliance values at early and late interviews is at risk of being rather impressionistic. One ideal solution would be to derive the null distribution of $\Delta P$ and to perform the usual null hypothesis significance testing. The derivation of such a distribution, however, seems difficult (Natalia Levshina, p.c). Another, more complex solution would be to test the significance of the difference in $\Delta P$ between two or more groups based on

<table>
<thead>
<tr>
<th>Interview Period</th>
<th>Total perfectives (cell $a + b$)</th>
<th>Total verbs (cell $d$)</th>
<th>Perfectives as Fraction of total (%)</th>
<th>Number of Different perfectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early</td>
<td>1,508</td>
<td>6,500</td>
<td>23</td>
<td>116</td>
</tr>
<tr>
<td>Late</td>
<td>2,169</td>
<td>8,639</td>
<td>25</td>
<td>145</td>
</tr>
</tbody>
</table>

Table 4: Density of perfective forms in the verbal lexicon per period of interview.
bootstrapping, along the line of Murakami and Alexopoulou (2016). This is exactly what Rastelli and Murakami (forthcoming) did to compare L1 and L2 perfectives from different corpora. The procedure consisted of randomly selecting learners from different groups (single learners and single native speakers can be selected multiple times) and of calculating $\Delta P$ scores for those sampled learners. The differences between $\Delta P$ were calculated a large number of times, resulting in a large number of difference scores. Finally, the 95% range of the distribution of the repeated difference scores obtained was calculated. If the range excluded 0, then one could conclude that the difference in $\Delta P$ scores between the two groups is unlikely to be due to chance (i.e., it is significant). In other terms, Rastelli and Murakami (forthcoming) ran a computational simulation, in which they built a large number of regression models based on sets of $\Delta P$ values arguably representing sampling variability and examined the variability in the estimated parameter values. If the variability of the slope parameter was small and excluded 0 in most of the models, then one could conclude that L1 $\Delta P$ values were significantly associated with L2 $\Delta P$ values. What the authors found was that 26 of 39 perfectives (67%) showed significant differences between L1 and L2 values in reliance, while 24 perfectives (62%) demonstrate significant differences in attraction. In sum, $\Delta P$ values differed between L1 and L2 in most target perfectives. L1 $\Delta P$ values were larger than L2 $\Delta P$ values, but while some predicate marked higher $\Delta P$ values in L2 than in L1, other predicates exhibited a reverse pattern. Would the same analysis be suitable for comparing $\Delta P$ scores of L2 learners within the corpus Pavia? Probably the size of the corpus and the choice of aggregating data by period of interview would undermine the reliability of the procedure described above. Indeed, the dataset in the current paper is much smaller than in Rastelli and Murakami (forthcoming), the perfective predicates were unevenly distributed across periods of interview and – as we have noticed – their occurrence strongly depended on the topic of interviews (word frequencies fluctuate a lot with the topic of discussion). Since bootstrapping the $\Delta P$ values could not adequately represent the sampling variability, it is unlikely that ‘time of interview’ would turn out as factor accounting for a great deal of variance in our data.

6.3 Why reliance and not attraction?

As we have said (Section 4), $\Delta P$ can distinguish between the situation in which the perfective exceptionally relies upon the lexeme and situation in which that lexeme attracts exceptionally the perfective. We recall that reliance compares the relative frequency of the morpheme with the lexeme to the relative frequency of the morpheme without the lexeme, while attraction compares the relative frequency of
the lexeme with the morpheme to the relative frequency of the lexeme without the morpheme. Our data showed that most L2 reliance values decreased consistently over time, whereas morpheme-to-lexeme attraction values did not pattern consistently and were negligible across interviews, clustering around zero. One possible explanation is that the sources of such behaviors are different. Decreasing reliance values can be due to the semantics of the lexeme, namely, its lexical aspect. One of the most famous hypotheses in SLA is that early L2 perfectives are telic (Andersen and Shirai 1994) and that the biased L1 distribution reinforces the telic-perfective association at initial stages of acquisition. However, lexical aspect is not the only possible explanation. High reliance values in the corpus Pavia can be either to the fact that initial L2 perfectives are contingent upon lexemes telicity or to the fact that telic lexemes are highly specific to the corpus Pavia (see below).

In contrast, the difference between lexeme-to-morpheme attraction in L1 and L2 data could be due to different L2 learners’ and native speakers’ sensitivity to formulaic uses of high frequency perfective predicates. If we compare Figures 1 and 2, we can see that there is a group of lexemes that attract the perfective morpheme in L1 but not in L2 Italian. These are general-purpose lexemes which are very frequent in both L1 and L2, such as fare, do’, prendere ‘take’, mettere ‘put’, dare ‘give’. These Italian verb lexemes often occur in light-verb constructions (such as fare colazione ‘have breakfast’, fare una passeggiata ‘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. In L1 Italian, constructions with general-purpose predicates are often found in the past perfective. The exceptional lexeme-to-morpheme attraction in our data might indicate that in L1 corpora – but not in the corpus Pavia – those verbs occur especially in formulaic expressions and light verb constructions in the past. It is worth investigating why such perfectives – although the frequency of lexemes in L1 and L2 corpora is comparable – attract the perfective morpheme in L1, but not in L2 Italian. It is possible that L2 learners are less sensitive to collocational uses of high frequency past perfectives and lack a native speaker’s capacity of processing (understanding and producing) collocations in real time. As Figure 1 shows – the most attracting predicates in the corpus Pavia were neither the most frequent nor the most generic in meaning. Rather, they were very specific and low frequency predicates such as arrivato ‘arrived’, capito ‘understood’ sbagliato ‘mistaken’ and sposato ‘got married’. These occurrences are highly corpus-specific. Unlike in L1 data, in the corpus Pavia the predicates that attract the perfective are very often those prompted by the interlocutor or requested by the elicitation task. (e.g. picture telling and personal narratives in the past) or by the topics of interactions used to elicit the data. For example, the higher attraction value of the perfective sposato ‘got married’ in L2 data is due to the scene used for
the re-telling task (a clip taken from the Charlie Chaplin film *Modern Times*), the higher attraction value of the perfective *arrivato* ‘arrived’ derives from personal narratives where learners were telling (in the past tense) how they arrived in Italy. Many other motion verbs in the corpus Pavia (e.g. *venire* ‘come’, *andare* ‘go’) probably attract exceptionally the perfective morpheme not because learners know the meaning and function of the perfective past tense, but because all participants were asked to tell how they arrived in Italy as immigrants and those verbs were reprised directly from the interviewer’s question (*come sei arrivato qui* ‘How did you get here?’). The lexemes *vedere* ‘see’ and *capire* ‘understand’ too attract the perfective morpheme, probably because they work as specific markers in the turn-taking system (e.g. *non ho capito* ‘I did not understand’) that holds between the L2 learner and native interviewers.

6.4 What could collinearity between reliance and attraction mean?

Our data showed that there might be a tendency (in a group of learners, but possibly in individual learners as well) for lexeme-morpheme associations to become more diversified over time. This may point to the fact that the memory-based, lexical retrieval strategy gradually yields to more grammatical, generative processing as a learner’s proficiency increases. Of course, the decrease of reliance scores over time cannot be used as a litmus test to establish whether a single inflected form has been ‘produced’ or ‘retrieved’. In contrast, one can use our data to suggest the presence of a more general developmental trend, but not to determine cutoff points in ΔP scores that may discriminate between grammatical and statistical representation/processing for single morphosyntactic items. The results also showed that the distribution of perfective morphemes is asymmetrical for native speakers, too. One may ask what the learners’ expected trajectory would be, for example, whether the expected endpoint could correspond to a distribution similar to that of native speakers. Here, the end state acquisition could be described as a gradual reduction in asymmetry between reliance and attraction, which we labeled as ‘increasing collinearity’. Collinearity between the values of reliance and attraction that we found at late interviews does not mean that both reliance and attraction tend to zero. The idea is worth exploring that – in both advanced interlanguages and eventually in mature languages – there is such a configuration or balance between variable (lexemes) and invariable (morphemes) items so that the more a morpheme relies on a given lexeme the more that lexeme attracts the morpheme. A viable working hypothesis – given the current state of our knowledge – is that in a native speaker’s competence, the perfective morpheme is
unevenly distributed across all verbal lexicon available to the native speaker. In fact, the perfective morpheme could become specialized for certain meanings and/or for certain classes of predicates. This would also represent the end state configuration of a L2 learner’s competence. All these hypotheses should be tested with larger longitudinal learner corpora than those currently available and on other morphosyntactic features.

References


Gries, Stefan. 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.


**Supplementary Material:** The online version of this article offers supplementary material (https://doi.org/10.1515/iral-2020.0082).